MEDICINAL CHEMISTRY

A Series of Monographs

EDITED BY

GEORGE DESTEVENS

Department of Chemistry
Drew University
Madison, NJ

Volume 1. GEORGE DESTEVENS. Diuretics: Chemistry and Pharmacology. 1963

Volume 2. RODOLFO PAOLETTI (ED.). Lipid Pharmacology. Volume I. 1964
RODOLFO PAOLETTI AND CHARLES J. GLUECK (EDS.). Volume II. 1976

Volume 3. E. J. ARIËNS (ED.). Molecular Pharmacology: The Mode of Action of Biologically Active Compounds. (In two volumes.) 1964

Volume 4. MAXWELL GORDON (ED.). Psychopharmacological Agents. Volume I. 1964. Volume II. 1967. Volume III. 1974. Volume IV. 1976

Volume 5. GEORGE DESTEVENS (ED.). Analgetics. 1965

Volume 6. ROLAND H. THORP AND LEONARD B. COBBIN. Cardiac Stimulant Substances. 1967

Volume 7. EMIL SCHLITTLER (ED.). Antihypertensive Agents. 1967

Volume 8. U. S. VON EULER AND RUNE ELIASSON. Prostaglandins. 1967

Volume 9. G. D. CAMPBELL (ED.). Oral Hypoglycaemic Agents: Pharmacology and Therapeutics. 1969

Volume 10. LEMONT B. KIER. Molecular Orbital Theory in Drug Research. 1971

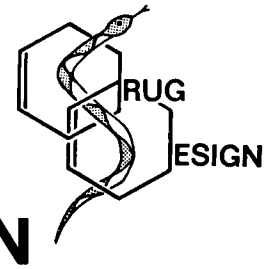
Volume 11. E. J. ARIËNS (ED.). Drug Design. Volumes I and II. 1971. Volume III. 1972. Volume IV. 1973. Volumes V and VI. 1975. Volume VII. 1976. Volume VIII. 1978. Volume IX. 1980

Volume 12. PAUL E. THOMPSON AND LESLIE M. WERBEL. Antimalarial Agents: Chemistry and Pharmacology. 1972

Volume 13. ROBERT A. SCHERRER AND MICHAEL W. WHITEHOUSE (Eds.). Antiinflammatory Agents: Chemistry and Pharmacology. (In two volumes.) 1974

Volume 14. LEMONT B. KIER AND LOWELL H. HALL. Molecular Connectivity in Chemistry and Drug Research. 1976

Volume 15. JULIUS A. VIDA. Anticonvulsants. 1977



DRUG DESIGN

Edited by E. J. Ariëns

DEPARTMENT OF PHARMACOLOGY
UNIVERSITY OF NIJMEGEN
NIJMEGEN, THE NETHERLANDS

VOLUME IX

ACADEMIC PRESS 1980
A Subsidiary of Harcourt Brace Jovanovich, Publishers
New York London Toronto Sydney San Francisco

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data
Main entry under title:

Drug design.

(Medicinal chemistry; a series of monographs, v. 11)
Includes bibliographies.
1. Pharmacology--Collected works. 2. Chemistry. Pharmaceutical--Collected works. I. Ariens, Everhardus Jacobus, ed. II. Series.
RM300.D74 615'.1 72-127678
ISBN 0-12-060309-8 (v. 9)

PRINTED IN THE UNITED STATES OF AMERICA

80 81 82 83 9 8 7 6 5 4 3 2 1

List of Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

E. J. ARIËNS (1), Institute of Pharmacology and Toxicology, Medical Faculty and Faculty of Sciences, Geert Grooteplein N21, 6500 HB Nijmegen, The Netherlands

C. J. DE BLAEY* (237), Department of Pharmaceutics, University of Leiden, Leiden, The Netherlands

V. E. GOLENDER (299), Institute of Organic Synthesis, Riga 226006, USSR

PETER P. MAGER (187), Bereich Pharmakologie, Sektion Pharmazie, Fakultaet fuer Naturwissenschaften, Martin Luther Universitaet, GDR 402 Halle, DDR

EDGAR F. MEYER, JR. (267), Department of Biochemistry and Biophysics, Texas A & M University, College Station, Texas 77843

WALTER MOROZOWICH (121), Division for Pharmaceutical Research, The Upjohn Company, Kalamazoo, Michigan 49001

J. POLDERMAN (237), Department of Pharmaceutics, University of Leiden, Leiden, The Netherlands

A. B. ROZENBLIT (299), Institute of Organic Synthesis, Riga 226006, USSR

STEFAN H. UNGER (47), Syntex Research, Division of Syntex, Palo Alto, California 94304

SAMUEL H. YALKOWSKY (121), Division for Pharmaceutical Research, The Upjohn Company, Kalamazoo, Michigan 49001

* PRESENT ADDRESS: Department of Pharmaceutics, Pharmaceutical Laboratory, University of Utrecht, Utrecht, The Netherlands.

Preface

The positive response to Volumes I–VIII of *Drug Design* and the rapid developments in this field warrant continuation of this series.

Chapter 1 of Volume IX elucidates efforts to avoid toxicity, not only of drugs, pesticides, and food additives but also of chemicals in general. Various aspects of the development of bioactive agents, including the optimalization of existing agents by the development of more efficient prodrugs, e.g., transport forms, and of special delivery forms are presented in Chapters 3 and 5 respectively.

More theoretical approaches to drug design also receive attention: Hansch's paradigm is applied to industrial practice in Chapter 2, multivariate statistics is applied to pharmacochemistry in Chapter 4, and computer-assisted drug design is described in Chapter 7. Chapter 6 presents a new and promising approach to the study of spatial arrangements in bioactive molecules that is especially important in the analysis of structure–activity relationships. The aim of the authors has been to present the reader with insight into both promising and actual developments in the field of drug design. The topics are presented in an informative, concise, systematic, and thought-provoking manner, in which speculations and new perspectives are encouraged.

The presentations in Volumes I–VIII as well as those in Volume IX indicate the wide scope of *Drug Design*. It is hoped that the reader will find the interdisciplinary approach fruitful.

E. J. ARIËNS

Contents of Other Volumes

VOLUME I

VOLUME II

VOLUME III

VOLUME IV

VOLUME V

VOLUME VI

VOLUME VII

VOLUME VIII

VOLUME X (*In preparation*)

Chapter 1 Design of Safer Chemicals

E. J. Ariëns

I. Introduction

In the past the primary concern of both the chemical industry and the individual chemist was the economical production of agents that served a particular purpose in an optimal way. This resulted in a wealth of chemical products (e.g., plastics, insecticides, weedkillers, food additives, drugs, and dyes) touching practically every aspect of life. After the fulfilment of the primary objectives, and with a growing awareness of the hitherto often unrecognized or neglected risks inherent in chemical

ISBN 0-12-060309-8

agents, the emphasis is more and more on safety. Since the Softenon disaster in 1961, this has become clearly manifest for drugs and food additives. The recognition and reevaluation of the impact of pesticides on ecological systems has resulted in a more critical approach in that field. The detection of long-term risks, such as carcinogenesis and mutagenesis caused by, for example, monomers in plastic manufacturing, has resulted in legislation with regard to protection against chemical health risks in general. TOSCA, the Toxic Substances Control Act, introduced recently in the United States, is an example of things to come.

For new chemicals, whatever purpose they may serve, the balance between advantages and disadvantages, among which health risks will be highly significant, will have to be assessed before acceptance for application. The goals are not to cure but to prevent, implying efforts to design safer chemicals. The term "design" indicates that the new agents will be developed on as rational a basis as possible, reducing the trial-and-error factor to a minimum and thereby avoiding situations where major or minor disasters are needed to point up the problem. Design involves control of potentially toxic actions of chemical agents by molecular manipulation, which requires an insight into the chemical mechanisms of toxic action, and therewith an insight into the relationship between structure and toxic action.

II. Toxic Action

A complex sequence of processes constitutes the basis for toxic action. It can, however, be split up in three main phases (Fig. 1) (5).

1. The exposure phase. This is composed of the factors or processes that are risk-determining in the handling of or exposure to potentially toxic substances. The most toxic agents, i.e., agents that are toxic at extremely low dosages, are by no means the most dangerous ones. The risks as a rule are mainly determined by the chance of contact with the toxon, the method of contact (i.e., the type of han-

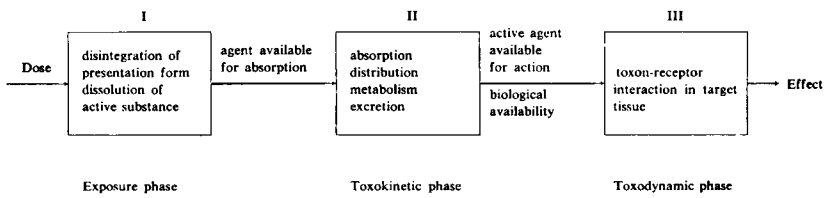


Fig. 1. Schematic representation of the main phases of toxic action.

dling), and thus the extent of absorption. The degree or efficacy of exposure as a function of time, the "exposure profile" is a determining factor. These profiles depend on, for instance, the presentation form of the toxon, the formulation, the concentration, the route of exposure (skin, oral, etc.), and the time course of exposure.

2. The toxokinetic phase. This involves those processes involved in absorption, distribution, excretion, and metabolic conversion of the active agent. The fraction of the dose that reaches the general circulation is a measure of the biological or systemic availability. The plasma concentration as a function of the time provides the "biological availability profile." The concentration of the toxon in the target tissue, related to the time, gives the "physiological availability profile."
3. The toxodynamic phase. This covers the processes involved in the interaction of the toxic agent and its molecular sites of action (receptors). This interaction results in the induction of a stimulus that initiates a sequence of biochemical and biophysical events finally leading to the effect. The characteristics of the toxodynamic phase form the basis for the toxicological classification of agents.

Various aspects of the main phases of toxic action, especially with regard to the molecular mechanisms, will be discussed as a basis for the design of safer agents, which will then be exemplified.

III. The Exposure Phase

Preventive factors in this phase are the proper labeling and instruction for the use of potentially toxic agents and the use of safe containers for handling. An example of the latter is the "child-resistant" packaging (*77*) not only of drugs but of household chemicals. With regard to industrial hygiene, it should be emphasized that giving the less intelligent or less educated worker the dirty, often risky, work indicates a lack of either social conscience or intelligence on the part of those in charge. An understanding by those exposed of the risks involved is a prerequisite to safe handling.

The uptake of a chemical by the organism is highly dependent on both the degree and the rate at which the substance in an absorbable (as a rule the molecularly dispersed) form comes into contact with an absorbing surface of the organism. In case of occupational poisoning as well as air pollution, the respiratory system is the primary route. In occupational poisoning, the skin is also an important path. Oral ingestion is practically restricted to toxon residues in food and to accidental poisoning.

The particle size, the relative lipid–water solubility, and the metabolic stability are determinant factors for both the persistence or the accumulation in the environment and in biological systems. For instance, the half-life in soil of the insecticide diflubenzuron is ½–1 week for a particle size of 2 microns and 8–16 weeks for a particle size of 19 microns (*71*).

Lipid solubility is an important factor for possible penetration through intact skin. Hydrophilic agents such as strongly ionized molecules can barely pass. This holds true for biological membranes (usually composite membranes, such as the intestinal epithelium) in general. Physiological conditions also play a role. For penetration through the skin, humidity of the skin, temperature, and the type of contact (e.g., via clothing soaked with the agent) influence absorption (Fig. 2) (*28*). For retention in case of

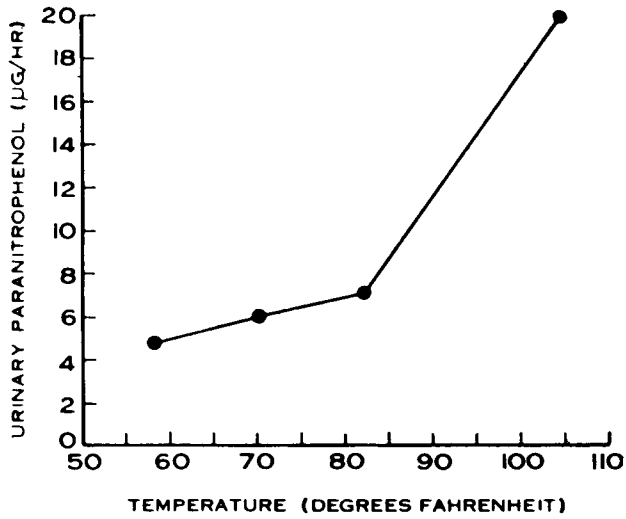


Fig. 2. The average hourly rates at which *p*-nitrophenol was excreted in urine for 41 hours following dermal exposure at different temperatures to 5 grams of a 2% parathion dust by three volunteers (Funckes, *28*).

inhalation, particle size and respiratory depth and volume per unit of time—which in their turn are dependent on, e.g., physical exertion, humidity, and temperature—play a role.

Control of toxicity or enhanced selectivity in action can be obtained by molecular manipulation. In the synthesis of dyes, mainly azo dyes, organic amines are important intermediates. These amines penetrate the skin readily and are potentially toxic; an example is β-naphthylamine which is a strong carcinogen. The extreme dangers associated with the handling of such substances could be brought under control by alteration

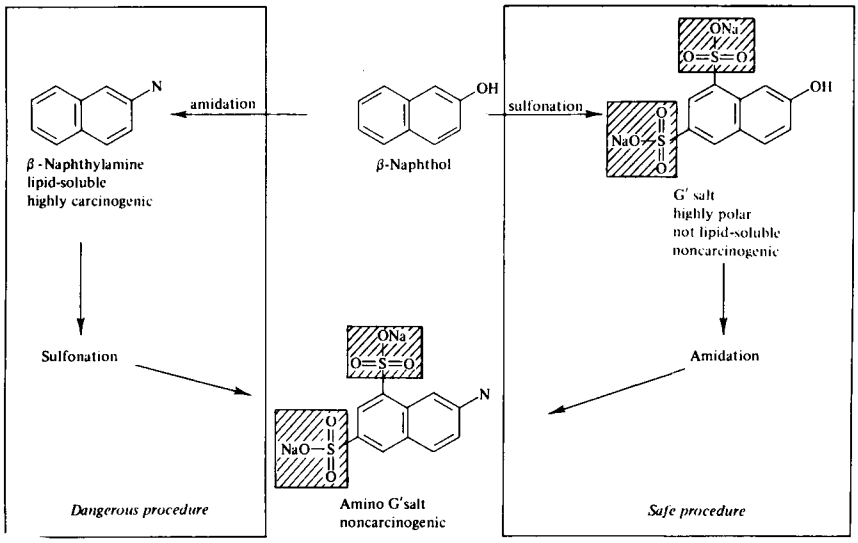


Fig. 3. Safeguarding of an industrial chemical process by adaptation of the route of synthesis, namely early introduction of toxicity-reducing distribution-restricting moieties (after Scott, *75*).

of the synthetic pathway. As shown in Fig. 3, introduction early in the synthetic procedure of the highly hydrophilic sulfonic acid groups, necessary in the final product anyway, produces a safe procedure by avoiding the potentially toxic lipophilic aromatic amines (*75*). Control of absorption by introduction of highly ionized groups is also realized in the development of selectivity in action, and thus reduction in adverse effects, of weedkillers (Fig. 4) (Crafts, 1957) (*18*). Plants with extensive foliage are exceptionally vulnerable to lipid-soluble weedkillers. Because of the large surface and the waxy character of that surface, hydrophilic compounds cannot penetrate. Such compounds can, however, be taken up by plant roots. Superficially rooting plants will be more vulnerable than deep rooted plants, trees, etc., as a result of the dilution and degradation involved in the process of penetration to deeper soil layers. Besides making use of high polarity, therewith restricting lipid solubility and thus absorption, as discussed before, there is the possibility of making use of polymeric agents which, due to their size, are not taken up by the biological systems and which are therefore restricted in their distribution (*27*). This approach holds true for chemicals in general, such as pigments, additives to plastics, and preservatives for wood. As far as protection against insects is concerned, the latter may be made digestible for insects, so that the toxon is liberated and a target-directed systemic action is

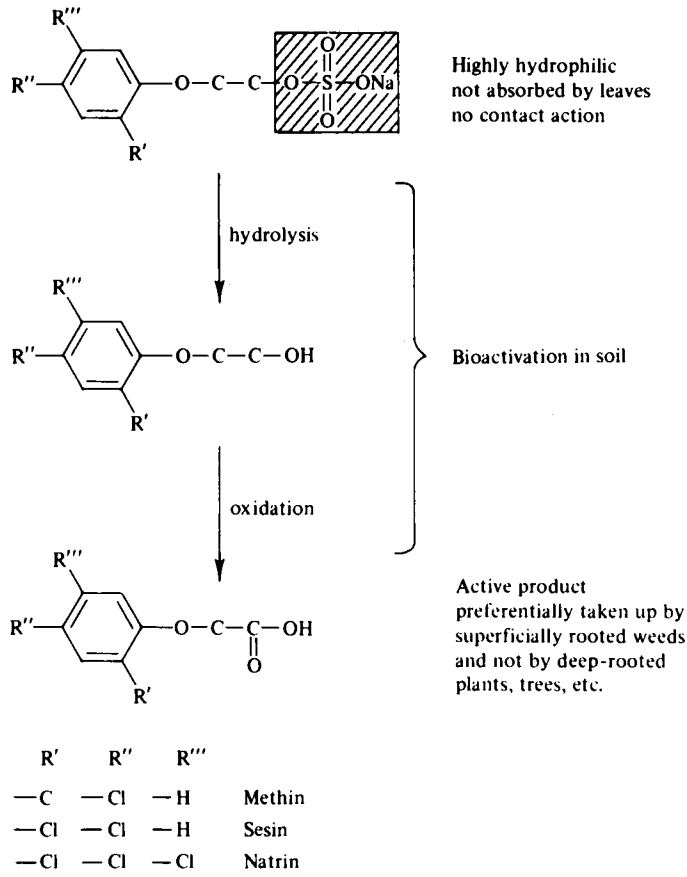


Fig. 4. Selectivity in action of weedkillers obtained by introduction of highly hydrophilic, disposable restricting moieties in a precursor compound (after Crafts, *18*).

obtained. This insolubilization approach also holds true for food additives such as colorants and even sweeteners. The fact that certain bioactive agents, e.g., insulin, remain bioactive after irreversible binding to a polymeric carrier such as sephadex and the fact that proteins such as monellin and thaumatin, which have such a molecular size that they cannot penetrate cells, have an intensely sweet taste, indicate that as long as the sites of action for such agents are located on the cell surface, this approach may be feasible. In this respect, polymers with chelating qualities may also be mentioned; such polymers are suitable, for instance, for extraction of metals from the effluent of sewage clearing plants and are even to be used as oral antidotes in the case of metal poisoning. The use of

insoluble polymers will facilitate both recovery and recycling and will restrict dissipation in the environment. A disadvantage might be no or poor biodegradability.

IV. The Toxokinetic Phase

The main aspects of this phase are the transport, especially via lipid membranes (involved in absorption, distribution, and excretion), and the metabolic conversion of the chemical agent.

A. MODULATION OF TRANSPORT

Passive transport via biological membranes is strongly dependent on relative lipid–water solubility and therewith on the partition coefficient of the agent. Introduction of strongly hydrophilic groups will restrict not only absorption, but also cell penetration; the compound, as far as absorbed, tends to stay in the extracellular fluid where it is readily available for excretion in the urine, for which hydrophilicity is advantageous.

Figure 5 shows how introduction of a quaternary onium group in an organic phosphate (irreversible acetylcholinesterase inhibitor) restricts the distribution of the agent to the extracellular compartment (*59*). For insecticidal action, lipophilicity is a requirement, since otherwise the compound will not be absorbed sufficiently. In clinical use, cell penetration, and particularly penetration of the blood–brain barrier with concomitant interference with central nervous system action, is to be avoided. The quaternary derivatives thus find application in the therapeutic treatment of glaucoma.

The use of azo dyes as food colorants has had serious consequences. The azo dye butter yellow, used to give winter butter the appearance of spring and summer butter, is for instance, a strongly carcinogenic agent. It is now substituted by lipid-soluble carotinoids. The incorporation of highly hydrophilic sulfonic acid groups into the azo dyes provides agents that are hardly if at all absorbed from the intestinal tract and which, when absorbed, are limited in their distribution mainly to the extracellular fluid and which are readily excreted in the urine. A characteristic of the azo and many other types of food colorants accepted for usance by the WHO (*89*), is the presence of such highly ionized groups. Since the intestinal flora are capable of reducing the azo link, one must be sure that each of the moieties in the molecule linked by azo groups has been safeguarded by sulfonic acid groups (Fig. 6) (*4*).

A counterpart to the poor absorption and rapid elimination of strongly

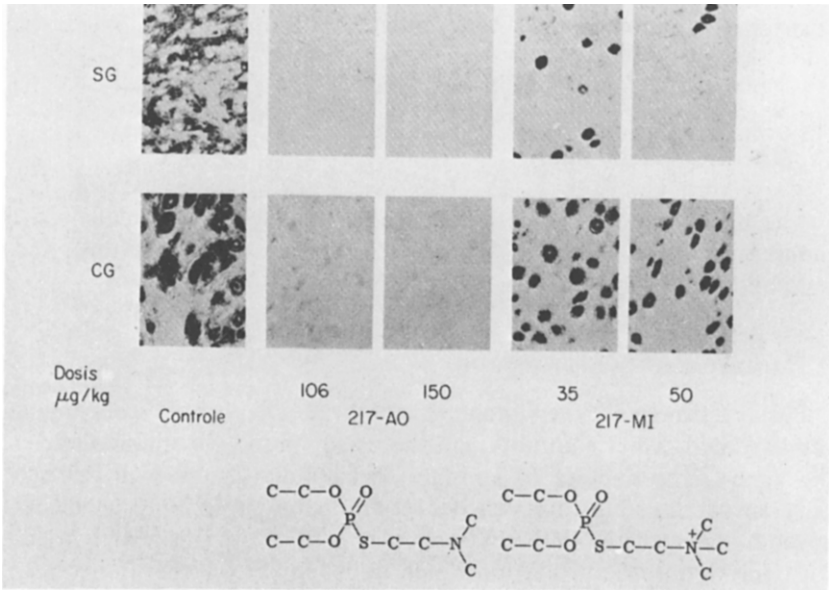


Fig. 5. The distribution of two organophosphates, a lipid-soluble tertiary base (217-AO) and a hydrophilic quaternary ammonium compound (217-MI) in the ganglion stellatum (SG) and ganglion ciliare (CG) of the cat. Note: the esterase activity (dark areas) is blocked intra- and extracellularly by the lipid-soluble compound and only in the extracellular space by the quaternary ammonium compound (McIsaac *et al.*, *59*).

hydrophilic agents is the easy absorption and accumulation tendency in the biological organisms of highly lipophilic, metabolically stable agents, of which DDT is an example. This compound and its also strongly lipophilic and metabolically stable dehydrochlorination product DDE show a high accumulation factor in biological organisms and an extremely long half-life (*64*) (Fig. 7), with as a result a strong accumulation along food chains (*19*) (Fig. 8).

Another example of lipid solubility and metabolic stability resulting in persistence and accumulation is that of the plasticizers abundantly present in a variety of plastic containers, among which are the disposable plastic blood-transfusion sets. Plasticizers are mostly phthalic acid esters; one of the best known is diethylphthalate (DEHP), which is resistant to the various body esterases. This highly lipophilic agent migrates not only to lipid food stuffs (Table I) (*84*) but also to the blood, when blood comes into contact with plastics rich in plasticizers (Fig. 9) (*46*). This results in a persistent accumulation of DEHP in the tissues of patients after blood

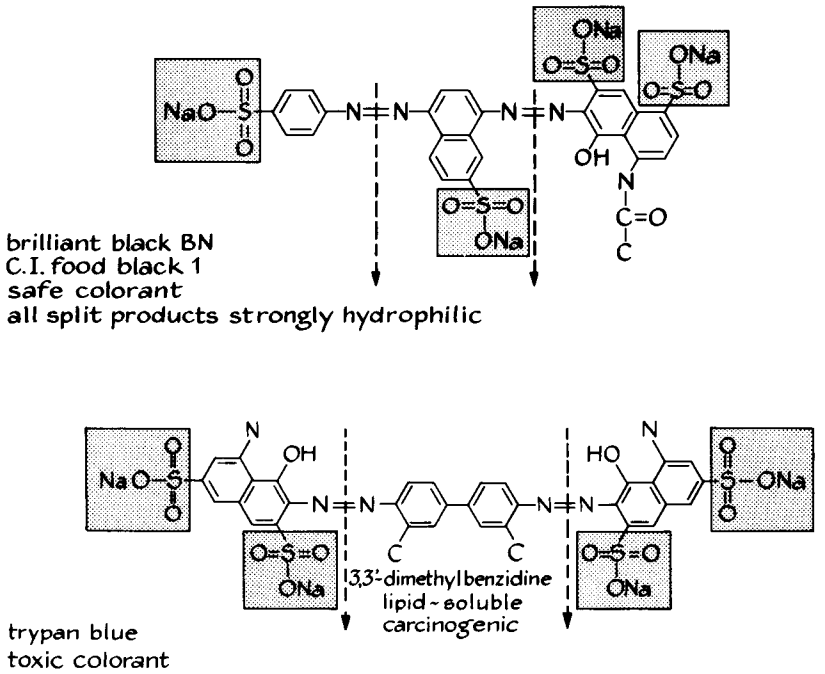


Fig. 6. Two azo dyes bearing strongly ionized sulfonic acid groups. If used as food colorant, one must take the azo reduction (arrows) by intestinal microorganisms into account. In case of trypan blue, the reduction results in the formation of a lipid-soluble carcinogenic benzidine derivative. In brilliant black, correctly, all three moieties linked by azo groups are safeguarded by strongly ionized groups reducing toxicological risks.

transfusions. In experiments with monkeys, a persistence of up to 14 months in liver and adipose tissue is observed (*3*, *45*, *46*).

In the cases of both the insecticide DDT and the plasticizer DEHP, one might consider the development of more hydrophilic agents. The possibilities in this respect are, however, very restricted since the absorption of the contact insecticide by the insect as well as the incorporation of the plasticizer into the plastics require a relatively high lipid solubility. The solution to this problem has been found in a reduction of metabolic stability, and it will be discussed in the section on modulation of metabolic conversions (IV,B).

An intermediate position, as far as lipid–water solubility coefficients are concerned, is taken by weak organic acids and bases. Such agents are, in the ionized form, restricted in membrane penetration and thus in both initial biological absorption and reabsorption in the kidney; they do pass freely in the nonionized form. This means that the pH in the environment

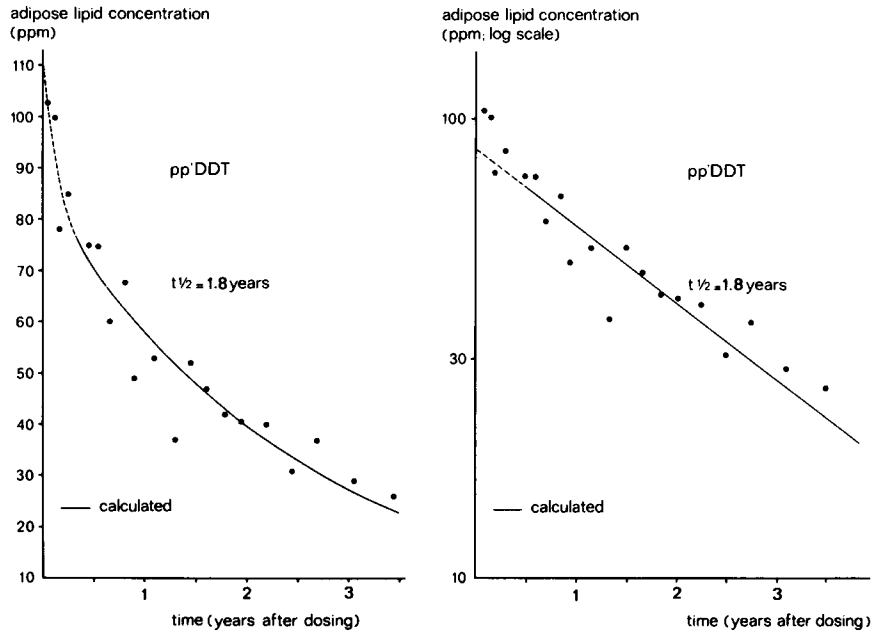


Fig. 7. The decline in the level of DDT in adipose tissue in an individual exposed to the insecticide at zero time. The disappearance of DDT follows an exponential course (left) such that on a semilog scale a straight line is obtained (right) (Morgan *et al.*, *64*).

TABLE I

MIGRATION OF DIETHYLPHTHALATE (DEP) FROM PLASTIC CONTAINER INTO CONTENT[a]

Product[b]	Quantity (gm)	Migration of DEP (mg/dm²)
Beans	676	0.4
Pudding powder	496	3.5
Salt	1199	0.8
Mustard	760	2.3
Sago	667	0.7
Lentils	783	1.8
Skim-milk powder	288	2.0

[a] Wandel and Tengler (*84*).

[b] The product (quantity in gm) was kept in contact with the plastic surface (64 dm²) at 40°C.

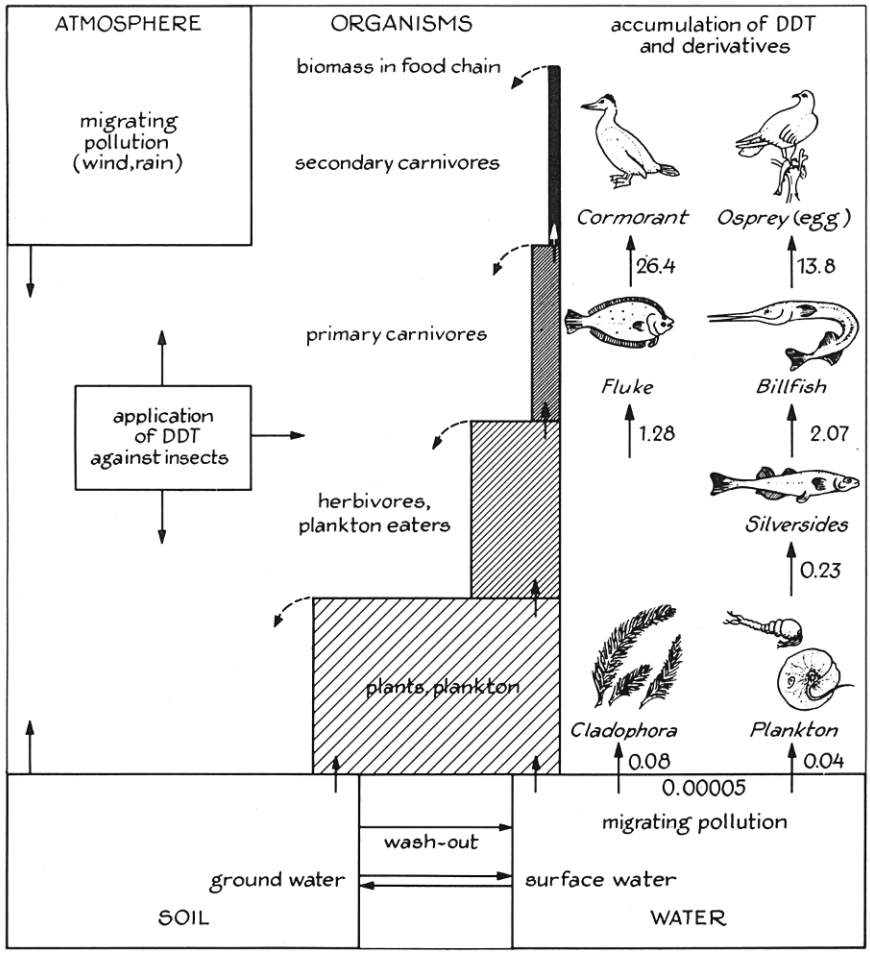


Fig. 8. Schematic representation of the accumulation of DDT and its metabolite DDE, highly lipophilic agents resistant to oxidative degradation along a food chain. The surfaces approximately represent the relationship between the biomasses while the shade represents the concentration of the agents. The species at the end of the food chain (birds) are exposed to excessively high concentrations due to the accumulation (Dewaide, *19*).

at the site of absorption, can be a determinant factor. This should be taken into account in the various toxicological tests on the exposure of fish or other water animals to organic agents and other effluents from industries. Figure 10 demonstrates how the uptake, and therewith the steady state body concentration of a pesticide (a phenol), increases with decreasing pH (*36*). For organic amines, the accumulation increases with increase of the

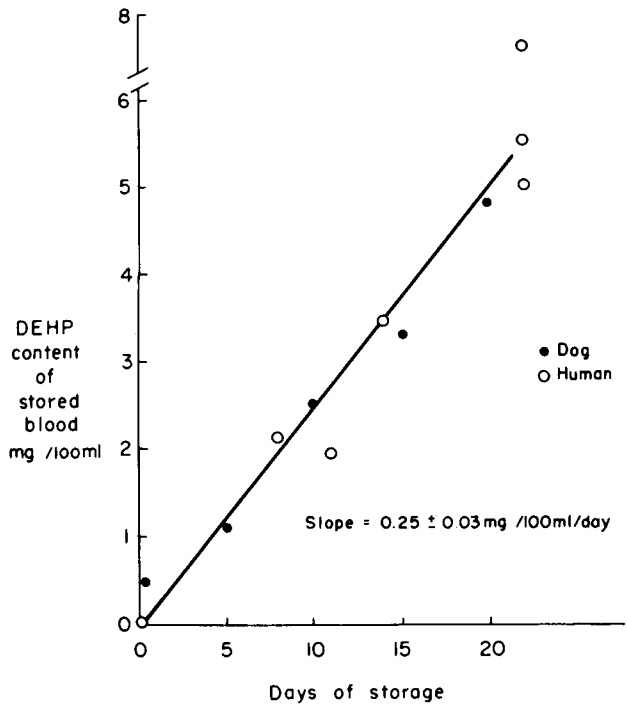


Fig. 9. The migration of DEHP from polyvinyl chloride in blood as a function of time (Jaeger *et al.*, *46*).

pH. This should be taken into account in determining the limits of allowed concentrations and in demonstrating that the ecological risks, with respect to accumulation in biological systems, can be dependent upon and possibly controlled by the environmental pH.

The dependence upon the passage of biological membranes by weak acids and bases also implies that the potential risks of exposure to such compounds as organic amines may be greatly enhanced by alkalinizing agents, for instance, by the use of such simple over-the-counter medications as sodium bicarbonate (present in many stomach preparations). Since such preparations cause an increase in the urinary pH, and thus an enhanced reabsorption of the amines, the urinary excretion thereof is reduced, resulting in higher, possibly toxic steady-state levels in the organism. In this respect, the best example is that of the racing cyclists who use amphetamine (pep pills) in combination with sodium bicarbonate. This alkalinizing agent buffers lactic acid formed during intensive muscular exertion, for instance in the sprint, thus controlling the acidosis (due to

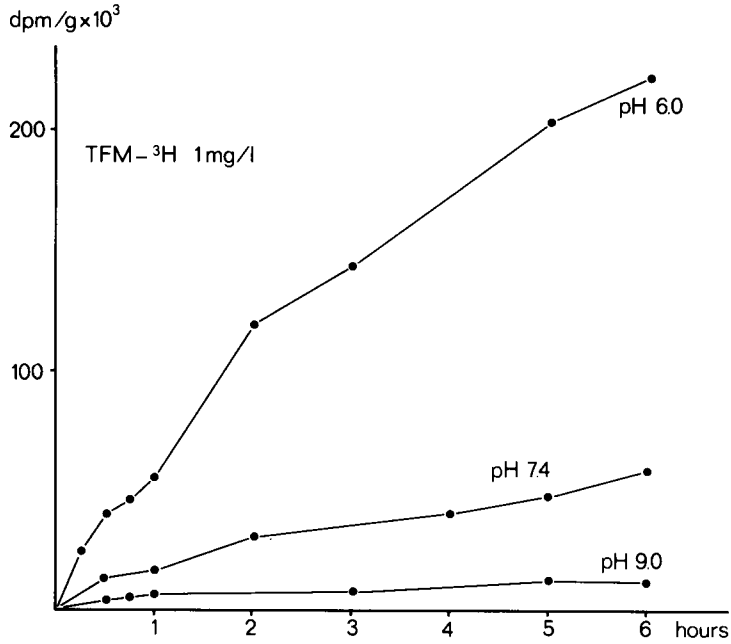


Fig. 10. Uptake by the rainbow trout of 3-trifluoromethyl-4-nitrophenol (TFM-^{3}H), a lampricide, from water at different pH values. The uptake of the weakly acidic pollutant is enhanced at low and reduced at high pH values (Hunn *et al., 36*).

lactacidaemia), the main cause of symptoms of fatigue. As a result of this "combined doping," nontoxic doses of amphetamine, now excreted very slowly (Fig. 11) (*7*), result in toxic plasma levels. What is true for amphetamine holds for potentially toxic organic amines in general. For these agents also, accumulation, and thus surpassing of toxic concentrations, can be brought about by the alkalinizing agents mentioned. Combination of exposure to potentially toxic amines and alkalinizing agents should be avoided to avoid unnecessary risks.

B. Modulation of Metabolic Conversion

Metabolic conversion, usually indicated as biodegradation and biodetoxification, is a coin with two sides. Indeed, metabolic conversion usually results in the formation of inactive, water-soluble, readily excretable products, and thus biodetoxification (Fig. 12) (*5*). The two main steps in metabolic conversion are a) the *oxidation,* especially, of relatively lipophilic agents and thus their conversion to more polar, less lipophilic agents and b) their *conjugation* with, as a rule, highly water-soluble, natur-

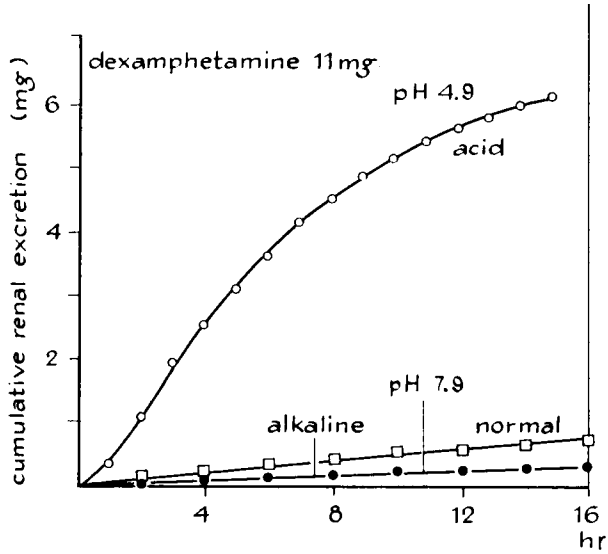


Fig. 11. Cumulative excretion of dexamphetamine in the urine during production of normal, alkaline and acidic urine after the administration of 11 mg dexamphetamine, orally, to volunteers. The excretion of dexamphetamine is enhanced at low and delayed at high values of the urine pH (Beckett *et al.*, 7).

ally produced molecules such as glucuronic acid and sulfate, resulting in highly water-soluble end products suitable for rapid renal excretion. The goals are detoxification and elimination.

In far from exceptional cases, however, the metabolic conversion takes place via chemically reactive, toxic, intermediate products such that the term biotoxification is more appropriate. In the field of metabolism, an understanding of "body-foreign" agents (xenobiotics) is essential to an insight into toxic mechanisms on a molecular level, especially with regard to such worrysome actions as carcinogenesis, mutagenesis (*5, 13, 14, 31, 32, 34, 38, 53, 57, 78, 82, 86, 88*), and allergic sensitization (*26, 55, 72, 85*).

In a discussion of the metabolism of xenobiotics in relation to the possible control of toxic action by the design of safer chemicals, the two main aspects are (1) control of undesired toxokinetics, especially accumulation, by modulating the half-life of the agent through introduction of biochemically vulnerable groups into the molecule, making it accessible for conversion into more water-soluble and thus more rapidly excreted products. The vulnerable moieties should result in a metabolic conversion not proceeding via intermediate biotoxification. Such moieties can be indicated as "safe" metabolic handles; (2) control of biotoxification by modulation of xenobiotic metabolism by molecular manipulation. The objective will

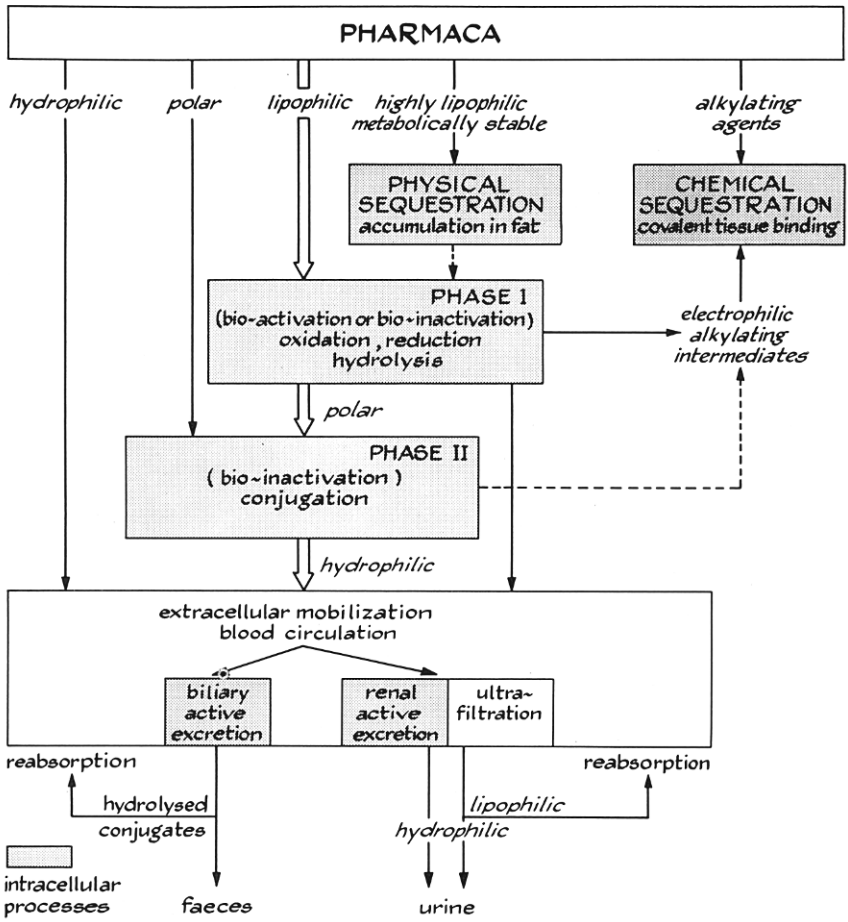


Fig. 12. Schematic representation of the major aspects of pharmacon (toxon) metabolism. The general trend is a transformation of lipophilic substances into more hydrophilic substances, reducing toxicity and enhancing urinary excretion.

be stabilization or elimination of the risky metabolic handles—vulnerable moieties involved in biotoxification—and/or introduction into the molecule of safe metabolic handles, diverting the metabolic conversion along a less risky pathway.

1. *Control of Undesired Toxokinetics by Enhancing Biodetoxification through Molecular Manipulation*

One of the predominant, adverse aspects of toxokinetics is the accumulation tendency of chemicals in man and in living organisms in general

(e.g., along food chains, as well as in the environment as such). Important factors are the stability of the agents, i.e., resistance to decomposition or biodegradation, and the lipid solubility, especially as far as accumulation in the biosphere is concerned. A possible solution to the problem involves the introduction of groups that make the agents accessible to biodegrading enzymes, hydrolases, or oxidases. Furthermore, insofar as this is compatible with the required action of the agents, a decrease in lipophilicity may be considered. In case of insecticides, weedkillers, plasticizers, etc. This as a rule will be acceptable only to a very limited degree.

Figure 13 (5) illustrates in a schematic way the role of biodegradability, and therewith the half-life, of chemicals in the process of environmental, in this case water, pollution. With a continuous, daily input of a polluting effluent, into for instance a lake, a higher or lower steady-state concentration will be reached, depending on the rate of biodegradation and therewith on the fraction of the pollutant eliminated per unit of time. This steady-state level may, especially if the chemical agent is only slowly

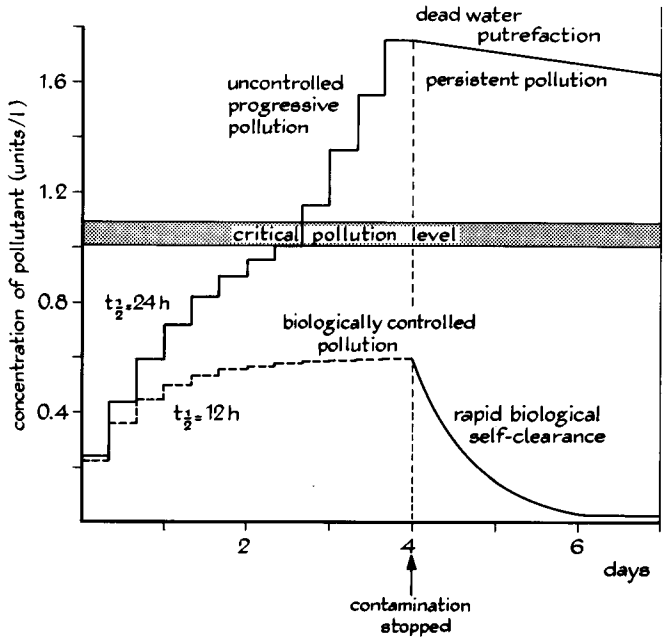


Fig. 13. The influence of the biological half-life of pollutants on the degree of water pollution. At a given input rate, the concentration in the environment will be higher for compounds with a longer biological half-life. At the critical pollution level, the biological systems taking care of the biodegradation are damaged producing a delay in biodegradation, prolongation of the half-life, and uncontrolled progressive pollution with a loss of the capacity for self-purification with, as a consequence, persistent pollution.

biodegraded, reach the critical pollution level with, as a consequence, uncontrolled progressive pollution, ending in persistent pollution and dead water. Table II (*3, 5, 42*) illustrates the conversion of persistent water pollutants, in this case "hard" detergents, into acceptable "soft" detergents by eliminating the alkyl chain branching, which interferes with oxidative biodegradation. The soft detergents (straight-chain) are biodegraded very rapidly via an oxidative pathway. This is an example of the control of environmental toxokinetics by molecular manipulation.

DDT, a persistent insecticide, can be similarly converted to much less persisting but still active derivatives by the introduction of vulnerable moieties, groups suitable for biochemical attack, into the molecule (Table III) (*61*).

Another example is the avoidance of accumulation of the plasticizer DEHP by the elimination of the ethyl groups in the alcohol, which stabilize the ester moiety, and by the introduction of a vulnerable moiety in the form of an extra ester group (Table IV). The plasticizer thus obtained (*3, 70*) is rapidly hydrolyzed with the formation of water-soluble products that can be excreted and which are accessible to oxidative biodegradation. Table V shows the practical results of this procedure.

In addition to accumulation and pollution, a lack in selectivity of action constitutes yet another aspect of toxicology (especially with respect to pesticides). Examples are the organophosphates, irreversible inhibitors of acetylcholinesterase, an essential enzyme for the proper functioning of the nervous system both in mammals and insects. As a consequence, organophosphates do not differentiate between the uneconomic species, the insects to be eliminated, and the economic species, cattle, chicken, etc., including man himself. Organophosphates were not originally designed as pesticides but as "homicidal" war gasses.

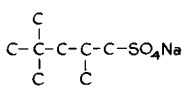
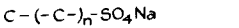
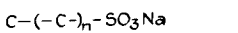
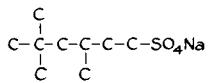
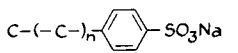
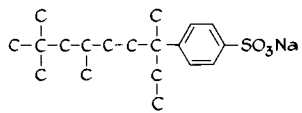
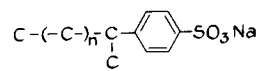
Mammals and birds have a relatively high capacity for hydrolyzing carboxy esters; insects are poor in this respect. Introduction of a carboxy ester moiety into the organophosphate results in rapid hydrolysis and thus in inactivation in mammals and birds and, as a result of a much longer half-life, in full toxic action in insects. In this way a relatively high degree of selectivity and therewith a reduction of toxicity for the economic species is obtained (Fig. 14) (*3, 67, 68*). This is an example of control of toxokinetics in the biosphere.

2. *Avoidance of Biotoxification by Molecular Manipulation*

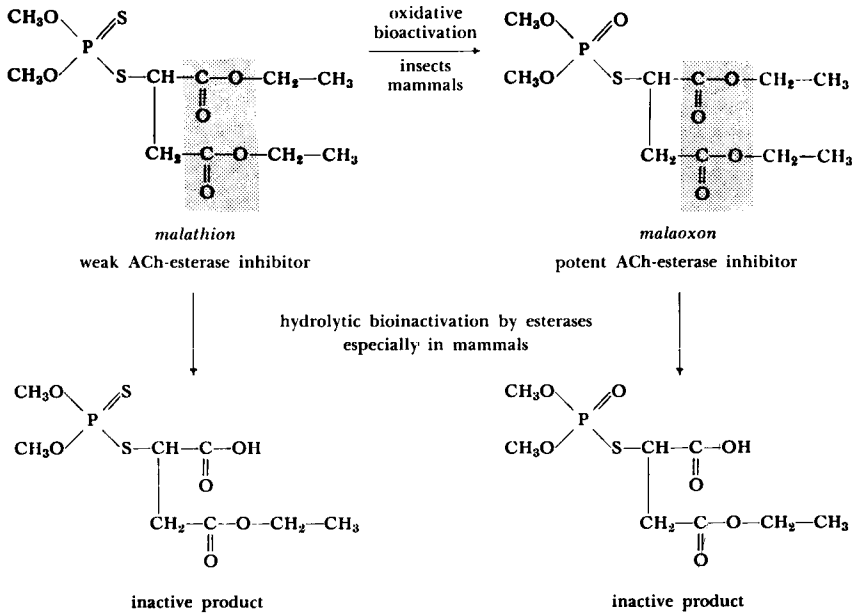
Certain types of toxic actions, such as carcinogenic, mutagenic, and teratogenic action and allergic sensitization, with few exceptions, require a chemical interaction, a covalent-bond formation, between toxon and biological polymers. For the first three types of action mentioned, the

TABLE II

A COMPARISON OF HARD AND SOFT DETERGENTS[a]

Hard	% Residue	Days	Soft	% Residue	Days
	100	18		0	1–3
				0	3
	100	18			
				0	4
	97	28			
				0	5–11

[a] Storage test in river water 20°C, dark, open bottles. The rate of biodegredation of detergents with branched alkyl chains (hard) is very low while that for detergents with unbranched alkyl chains (soft) is high, such that they are readily eliminated from the environment (Huyser, *42*).



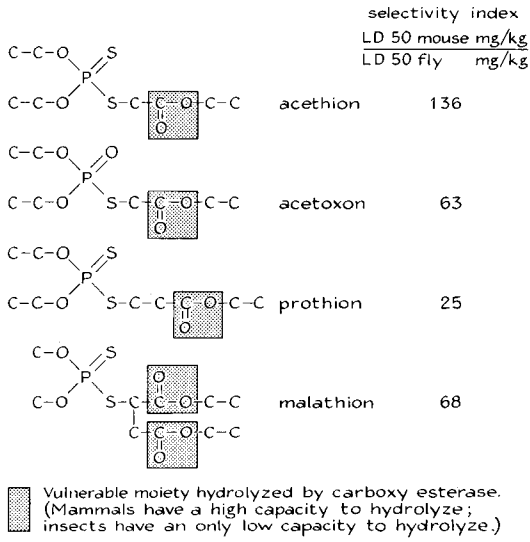
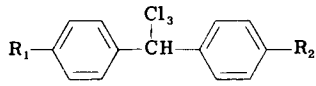


Fig. 14. Species selectivity in organophosphates obtained by introduction of carboxy ester groups into the molecule. In mammals rapid hydrolysis, and therewith inactivation of the pesticide, takes place in contrast to the insects which have only a limited capacity for carboxy ester hydrolysis (after O'Brien, *67*).

TABLE III

SUBSTITUTION OF THE 4,4'-CHLOROATOMS IN DDT BY "VULNERABLE MOIETIES"[a]



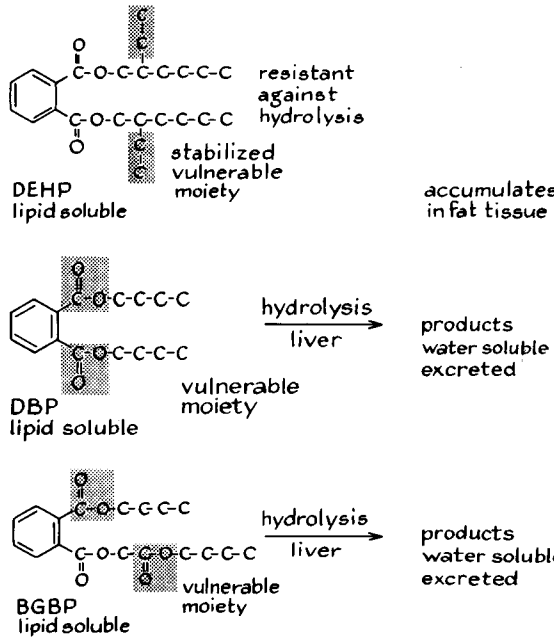
($R_1 = R_2$)

	Biodegradability	Accumulation factor	LD_{50} ppm (Culex pipens)
-Cl(DDT)	0.015	84500	0.07
-OCH_3(methoxychlor)	0.94	1545	0.07
-CH_3	7.14	140	0.06
-SCH_3	47.00	5.5	0.21

[a] This substitution in DDT enhances biodegradation and reduces ecological accumulation with maintenance of action. For details see Metcalf (*61*).

TABLE IV

INFLUENCE OF THE STABILITY OF THE VULNERABLE MOIETIES IN PLASTICIZERS ON THEIR PHARMACOKINETICS[a]



[a] DEHP is resistant to biodegradation and accumulates in adipose tissue. DBP and BGBF are hydrolyzed in the liver to hydrophilic metabolites that can be readily eliminated (Ariëns *et al.*, *5*).

TABLE V
PERSISTING POLLUTANTS[a]

Pollutant	Mississippi delta		Gulf of Mexico
	Surface water[b]	Sediment[c]	Biota (fish, crab, shrimp, etc)[c]
DDT's	1.7	4.2	11.4
PCB's	2.45	18.7	29.6
DBP[d]	95	13	< 0.1
DEHP[d]	70	69	4.5

[a] After Giam *et al.* (*31a*).
[b] ng/l.
[c] ng/g.
[d] Plasticizer.

polymer is DNA (*11, 14, 37, 63, 78, 79, 86, 87*). In the case of allergic sensitization, covalent binding to proteins is involved (*26, 55*). These reactions result in an interference with the proper functioning of the biopolymers which is then indicated as chemical lesion. Such chemical lesions are detrimental if located in critical positions in biopolymers essential for healthy life, i.e., in critical biopolymers. Most of the well-known carcinogens, mutagens, and allergens, however, are chemically rather stable compounds, with practically no tendency to react chemically with the biopolymers mentioned. It is now understood that, with the exception of those agents that have direct biologically alkylating properties, in fact most carcinogens, including carcinogenic azo dyes and polycyclic hydrocarbons such as benzpyrene, are strictly taken not carcinogenic but are converted into carcinogens by the oxidative enzymes—microsomal mixed-function oxidases—involved in the biochemical conversion of chemicals in the organism (*13, 32, 37, 62, 86*).

In this process, electrophilic, biologically alkylating intermediates, such as epoxides and *N*-hydroxyarylamines, that can covalently bind to nucleophilic groups in DNA and proteins, are formed. Thus, chemical lesions are brought about in DNA with, as a consequence, a carcinogenic or mutagenic action. If the chemical lesions are brought about in proteins, these are converted into "body-foreign" proteins against which antibodies are developed, with concomitant allergic sensitization (*26, 55, 72*). Further chemical reactions with various cell constituents can result in cell degeneration and necrosis, especially in the liver where most of the reactive intermediates are generated (*33*). Although with regard to biotoxification, emphasis is usually put on the oxidative conversion processes, conjugation reactions may also occasionally result in the formation of biologically alkylating or acylating products (*10*).

In fact, one can distinguish between two types of sequestration of xenobiotics in the organism: a) *Physical sequestration* is based on an accumulation of highly lipid-soluble metabolically stable agents, e.g., DDT especially in adipose tissue. In this case extraction by, for instance, organic solvents is possible. b) *Chemical sequestration* is based on covalent tissue binding. In this case solvent extraction is not possible. Figure 15 (*8*,

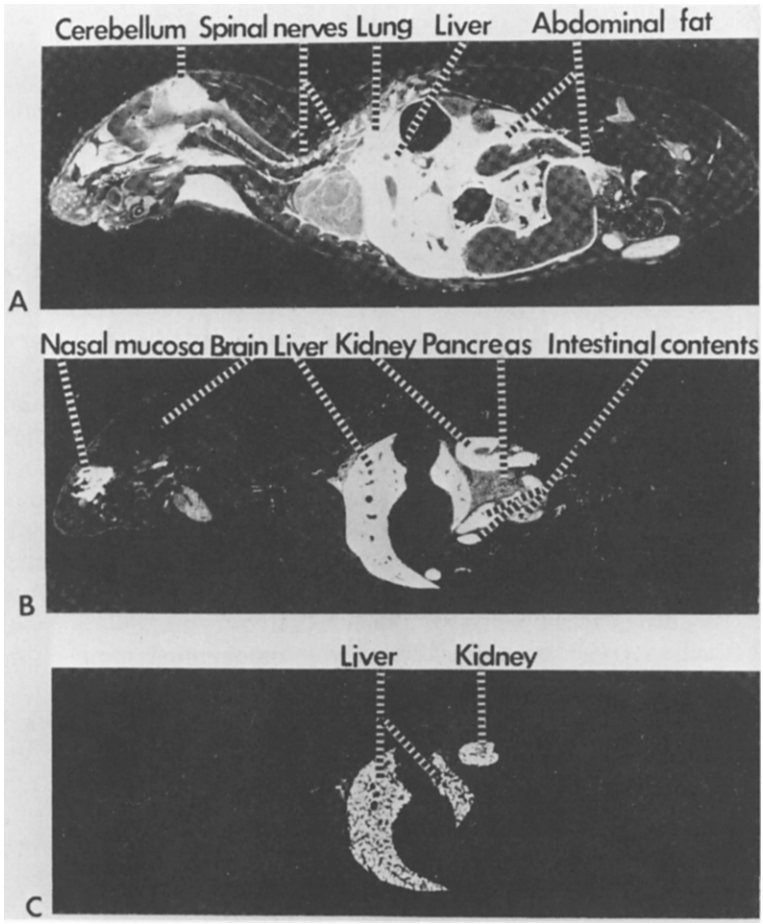


Fig. 15. Autoradiograms of a mouse two hours after inhalation of [^{14}C]chloroform. A) exposure of a hemisection at −80°C. The distribution of the nonmetabolized chloroform and its metabolites is registered. B) exposure of a tape-fastened section at −20°C. The distribution of only the nonvolatile metabolites is registered. C) exposure of an organic solvent-extracted tape-fastened section at −20°C. The distribution of chemically sequestered metabolites, covalently bound to biopolymers, is registered (Bergman *et al.*, *9*).

9) gives an example of this phenomenon. Figure 12 (5) summarizes in a schematic way the main aspects of pharmacon metabolism, including the two types of sequestration.

The organism is not defenseless against the risks involved in biodegradation of body-foreign chemicals. Glutathione and glutathione transferase act as a scavenging system for the electrophilic alkylating intermediates and the possibly formed free radicals. Glutathione *S*-conjugates are formed. Furthermore, the enzyme epoxide hydratase takes care of a rapid conversion of the biologically alkylating epoxides into dihydrols, which can be further converted to phenols which in their turn can be conjugated with sulfate and excreted in the urine. The glutathione conjugation products appear in the urine as mercapturic acid derivatives (Fig. 16) (*3, 13, 31, 32, 34, 37, 47, 50, 60, 69, 78, 86*). Figure 17 (*4, 5*) summarizes in a schematic way these aspects of biotoxification and biodetoxification.

a. Avoidance of Epoxide Formation-Based Biotoxification. In structure–activity studies for potentially carcinogenic and mutagenic agents with the involvement of epoxide formation, one must take into account the posi-

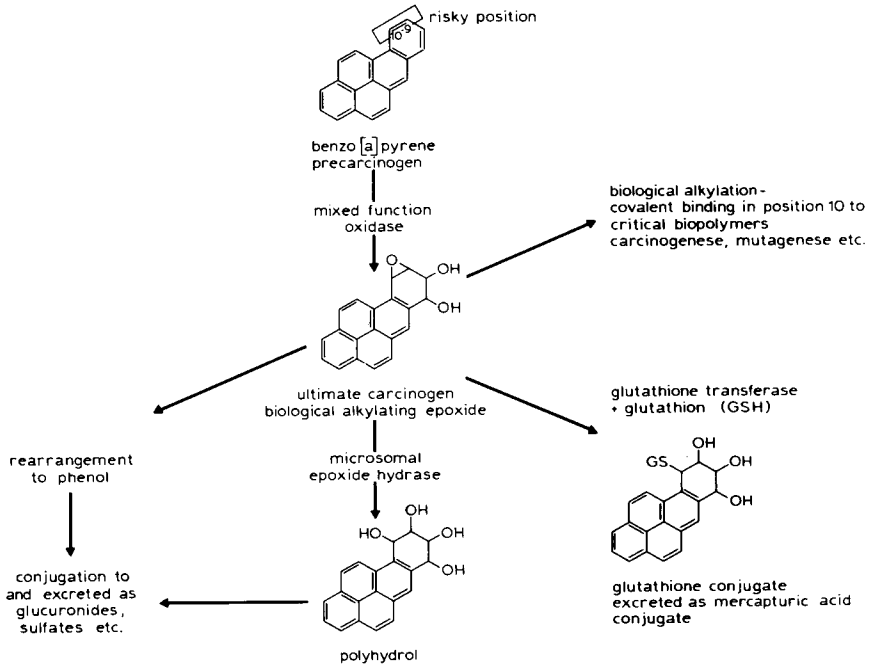


Fig. 16. Schematic representation of the conversion of benzo(a)pyrene to a biologically alkylating epoxide and the conversion of this toxic intermediate, the ultimate carcinogen, to nontoxic, water-soluble metabolites suitable for urinary excretion.

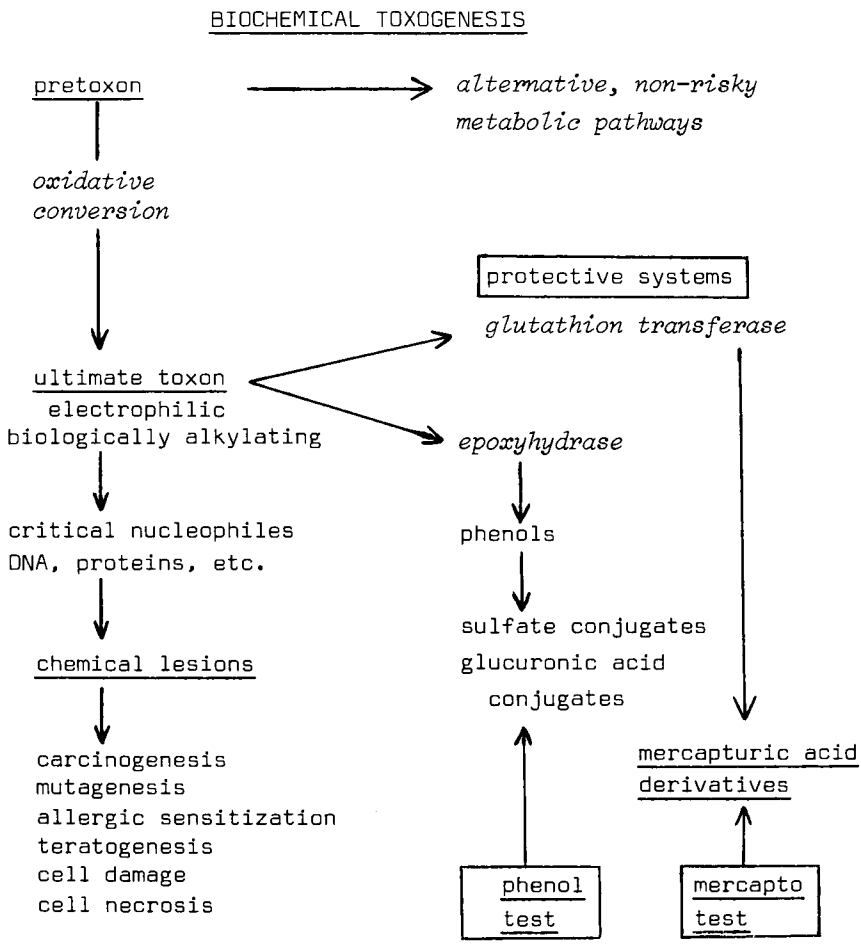


Fig. 17. Summary of biochemical toxicogenesis.

tion in the molecule prone to oxidative conversion to epoxides, the degree of electrophilicity of the epoxide, and the capacity for carbonium ion formation. The properties of the epoxide group as a matter of fact depend on the position in the molecule, and on the type and position of substituents in the molecule (*41, 78*). Highly electrophilic epoxides will react with the plentiful, available water to form dihydrol derivatives. Poorly electrophilic epoxides may be so stable that they do not even react, under biological conditions, with the nucleophilic groups present in the biopolymers. The intermediate class of epoxides will be stable enough to avoid reaction with water and reactive enough to alkylate nucleophilic

groups such as HS–R and H_2N–R present in critical biopolymers such as DNA, RNA, and functional proteins, and HS–R supplied by the glutathione–glutathione transferase system scavenging the alkylating agents.

A straightforward approach to the avoidance of toxic epoxide formation is the design of chemicals that are resistant to those metabolic conversions that proceed via epoxides capable of bioalkylation. This implies a stabilization or elimination of risky metabolic handles in the molecule. In carcinogenic polycyclic hydrocarbons, the double bond in the phenyl ring next to the "bay region" or, in certain cases, the double bond of the "k-region" is the site of introduction of the biologically alkylating epoxy group involved in the carcinogenic action (*41, 49, 51*). Benz(a)pyrene is bound in such a position (carbon atom 10) to DNA (Fig. 18) (*48, 65*). The risky epoxide formation is not necessarily the first oxida-

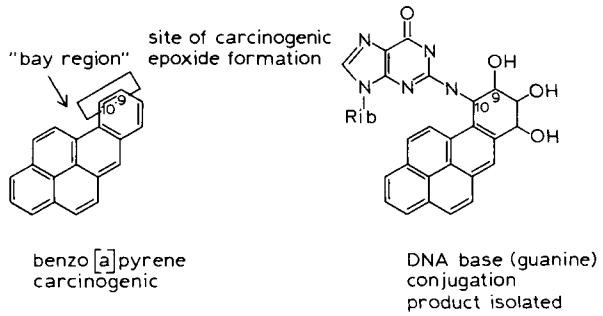


Fig. 18. A benzo(a)pyrene–ribonucleic acid conjugation product formed *in vivo* by benzo(a)pyrene metabolism (after Nakanishi *et al., 65*).

tive step. In a number of cases formation of a dihydrol takes place first via an epoxide that is so highly reactive that it reacts with water, followed by a second oxidative attack generating the alkylating epoxy group. The dihydrol group can play a role in the electrophilicity and in the capacity for carbonium ion formation from the risky epoxide (*41, 51, 78*).

One of the ways to "cure" a carcinogenic molecule is substitution of hydrogens in critical positions, especially on the risky metabolic handle, by fluorine because this atom is so tightly bound to carbon that an oxidative attack in that position is practically excluded. As shown in Fig. 19, this procedure is indeed effective as a result of a shift in the charge distribution and/or blockade of the initial conversion to a dihydrol. Substitution of fluorine in the vicinity of the risky position is also effective. On the other hand, introduction of fluorine, an electron-drawing group, further away in the conjugated double bond system can enhance the carcinogenic action (Fig. 19) (*49, 51*). Since the fluorine atom is practically

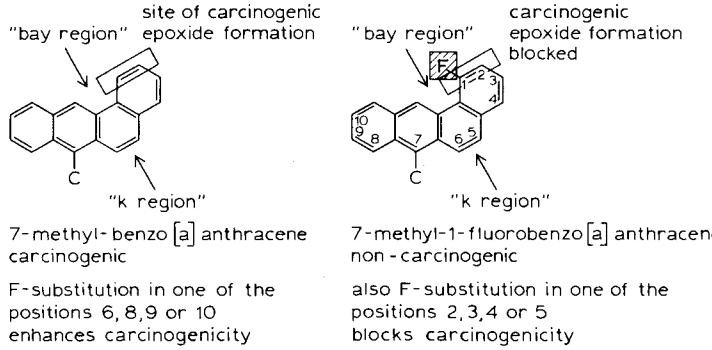


Fig. 19. The avoidance of the carcinogenic action in benzo(a)anthracene by blocking the site of carcinogenic epoxide formation (after Jerina *et al.*, *51*).

isosteric with the hydrogen atom, in many cases the substitution will be compatible with the required properties of the compound. This, as a matter of fact, is dependent on the general structure–activity relationship for the compound concerned.

Another example of the "curing" a carcinogen is the introduction of a methyl group in *F*-norsteranthrene, a carcinogen (Fig. 20). In this case, the alkyl substituent serves as a group blocking the risky position (*54*).

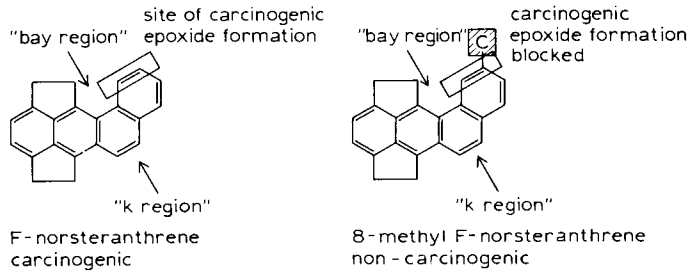


Fig. 20. Elimination of carcinogenic action by molecular manipulation. The methyl group blocks the metabolic conversion to a carcinogenic epoxide (after Lacassagne *et al.*, *54*).

An extension and systematization of the information on epoxide formation in aromatic systems and alkene and alkyne groups (such as vinyl, ethinyl and allyl groups; in general, unsaturated moieties) in various positions will help to understand structure–activity relationships (SAR) in this field and thus will help to reduce the toxicological risks.

b. Avoidance of N-Hydroxyarylamine Formation-Based Toxicity. Also, for the carcinogenic and mutagenic action of aromatic amines, a biotoxification, primarily based on oxidative conversion to *N*-hydroxyarylamines by

the mixed-function oxidase is an essential step (*33, 38, 52, 53, 82*). Often the aromatic amines (arylamines) are generated from azo compounds, e.g., azo dyes, by biochemical reduction, for instance, by the intestinal flora.

The *N*-hydroxyarylamines have alkylating properties, especially at pH values of about 6 where highly electrophilic nitrenium ions are formed. They can be N-conjugated to glucuronides, resulting in relatively stable nonreactive products from which, as a result of enzymatic hydrolysis in the urine, a regeneration of the carcinogenic *N*-hydroxyarylamines takes place, with as a consequence a preferential location of the tumor in the bladder. Aromatic amines often preferentially show bladder cancer development (*33, 52*).

Also, the conjugation products, namely the esters of *N*-hydroxyarylamines and sulfate, can be strongly electrophilic and thus alkylate nucleophilic groups in biopolymers (*43, 53, 88*). In the alkylation reaction, the covalent link is usually located on a carbon atom in the aromatic ring next to the nucleophilic group (the biochemically converted amino function) in the biopolymer (Fig. 21).

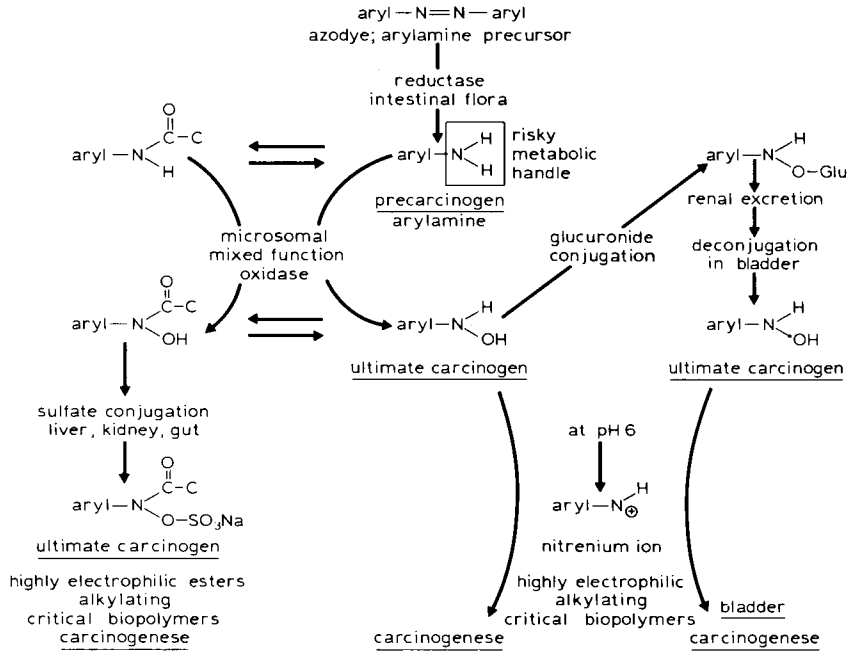
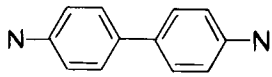


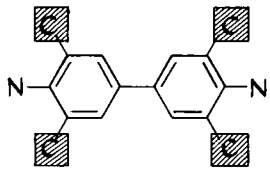
Fig. 21. Schematic representation of the conversion of arylamines to carcinogens, this in relation to the tumor location [after Kadlubar *et al.* (*52*) and Kriek (*53*)].

To control biotoxification of the aromatic amines, a blockade of the oxidative attack on the amino function, by for instance adjacent alkyl substitution ("packing" of the amino function), opens perspectives. Suitable substituents in the aromatic ring may also interfere with the covalent bond formation. Figure 22 and Tables VI and VII give examples of successful efforts to "cure" the carcinogen benzidine (*29, 40*).



benzidine (precarcinogen) converted by oxidative attack on amino group to ultimate carcinogen

used as specific reagent for detection of blood. dangerous to handle



3,3′,5,5′-tetramethylbenzidine[a] noncarcinogenic, metabolic conversion to ultimate carcinogen blocked

used as specific reagent for detection of blood. safe to handle

Fig. 22. Elimination of carcinogenic action by molecular manipulation. "Packing" of the aromatic amino function blocks the metabolic conversion to a carcinogenic N-oxidation product. [a]Aldrich products, technical information 117, product no. 86033-6.

TABLE VI
MUTAGENICITY OF BENZIDINE (B) AND 3,3′,5,5′-TETRAMETHYLBENZIDINE (TMB)[a]

	Mutation rate (*Salm. typhimurium* TA1538)[b]					
	0[c]		50[c]		100[c]	
Liver microsomes	−	+	−	+	−	+
B			5	430	15	640
TMB			5	15	9	15
Control	8	16				

[a] After Garner *et al.* (*29*).
[b] Histidine revertants/plate.
[c] μg/plate.

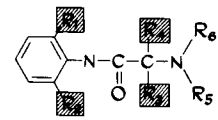
TABLE VII
CARCINOGENICITY OF BENZIDINE (B) AND 3,3',5,5'-TETRAMETHYLBENZIDINE (TMB)[a]

	Cumulative dose (g/kg)	Days	Number of rats with tumors
B	0.75	(150)	20 of 22
TMB	4.15	(224)	3 of 12
Control	0	(224)	1 of 12

[a] After Holland *et al.* (*40*).

Another example is given in Table VIII. Here the hydrolysis of the amide function is a primary requirement to make an oxidative attack on the aniline amino function possible. This implies that a blockade of the hydrolysis by "packing" of the amide function, as well as "packing" of the arylamino function, is effective (*5*).

TABLE VIII
CONTROL OF HEMATOTOXICITY OF AMINOACYLANILIDES BY MOLECULAR MANIPULATION[a]

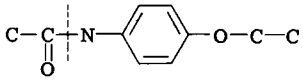
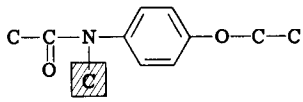
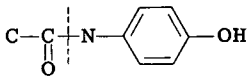
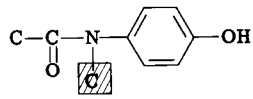


R_1	R_2	R_3	R_4	R_5	R_6	Rate of hydrolysis by cat liver slices nmol gm^{-1} min^{-1}	% Methemoglobin in cats 2 hrs after iv inf. 20 mg/kg (20 min)
C	H	H	H	C–C	C–C	59	28
C	H	C	H	H	C–C–C–C	46	21
C(1)	H	C	C	H	C–C–C	0	< 1
H	H	C	H	H	C–C–C	25	27
C(2)	C	H	H	H	C–C C	31	< 1
C(3)	C	H	H	C–C	C–C	20	< 1

[a] The toxic action can be eliminated by reducing the rate of hydrolysis by "packing" of the amide group (1) or by packing of the aromatic amino group in the anilide (2 and 3), in both cases avoiding formation of the toxic N-oxidation products (after Åkerman *et al.*, *1*).

The hepato- and nephrotoxic actions of phenacetin and acetaminophen are well known (*60*). Stabilization of the acetamido function in these agents by *N*-methyl substitution results in an elimination of methemoglobin formation and hepatic necrosis (studied only for acetaminophen) (Table IX) (*66*). It is not clear yet, whether the analgesic action, which is ascribed to the deacetylation products, will not be lost too. "Packing" of the arylamino function possibly opens more perspective for elimination of

TABLE IX
MASKING OF TOXOGENIC GROUP (BLOCKING OF ARYLAMINE FORMATION)[a]

	Dosage (mg/kg ip)	Methemoglobin in mice (%)	Liver necrosis in mice (mortality)
 Phenacetin	400	31.5 ± 2.1	ND[b]
 *N*-Methylphenacetin	800	4.8 ± 0.5	ND[b]
 Acetaminophen	800	ND[b]	13 of 20 (5)
 *N*-Methylacetaminophen	800	ND[b]	1 of 20 (2)

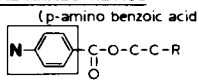
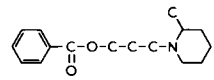
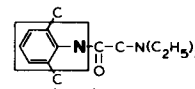
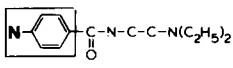
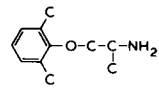
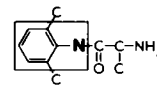
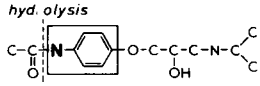
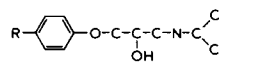
[a] S. D. Nelson *et al.* (*66*).
[b] ND = not determined.

toxicity under maintenance of action in this case. Table X gives further examples of efforts to "cure" toxic agents.

c. Avoidance of Toxic Intermediate Metabolite Formation by Deviating Metabolism via Safer Metabolic Pathways. A further possibility for controlling toxicity is the introduction into the molecule of a safe metabolic handle, readily available for biochemical attack, thus diverting the metabolic conversion along a safer pathway. Carboxy ester and alkyl groups might serve the purpose. They must be located in the molecule in suitable, exposed positions. This approach is preferable if a relatively short half-life is required. On the other hand, the half-life, and therewith the tendency to biological accumulation, can be reduced by a controlled manipulation of the lipid–water solubility coefficient, which usually will not be a great problem for the molecular designer, the pharmacochemist. An example of the reduction of a potentially risky oxidative attack on an aromatic ring system by diverting the oxidative attack by means of suitable alkyl substituents in the molecule is given in Fig. 23. The reduction in the risky aromatic ring oxidation by introduction of suitable, metabolic pathway-

TABLE X

CONTROL OF TOXICITY BY CHEMICAL MODIFICATION OF DRUGS

Aniline derivatives (arylamines)[a]	Arylamine-free analogs[b]	Analogs with "packed" arylamine group[b]
LOCAL ANAESTHETICS (p-amino benzoic acid type)		
 benzocaine R = -H	 piperocaine	 lidocaine
ANTI-ARRHYTHMICS		
 procainamide	 mexiletine	 tocaidine
β_1-ADRENERGIC BLOCKERS		
hyd. olysis  practolol	 metoprolol R = -C-C-O-C atenolol R = -C-C-NH$_2$	
SULFONAMIDES 	(p-amino group essential for action)	(no packing tolerated)

[a] Agents with immuno response-type side effects due to biotoxification on the aniline (arylamine) group.

[b] Elimination of immuno response-type side effects by elimination or packing of the risky metabolic handle.

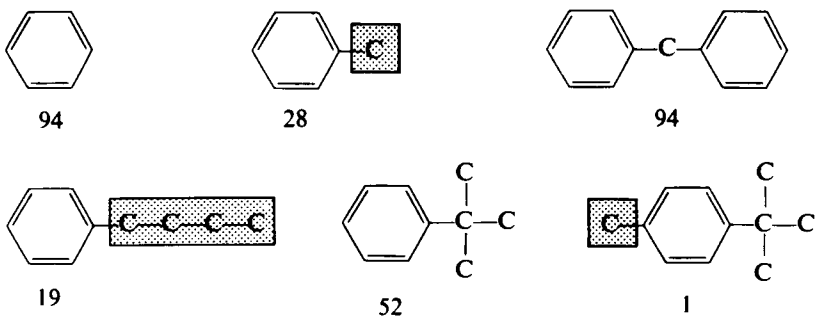


Fig. 23. The shift from potentially toxic ring oxidation to nontoxic side-chain oxidation. The figures indicate the percentage of organic sulfate in the sulfate excretion. The ring oxidation finally leads to the formation of phenols excreted as organic sulfates. Unbranched alkyl chains serve as alternative safe metabolic handles, easily oxidized to carboxylic acid groups suitable for renal excretion. Branched side chains are not suitable to divert the oxidative pathway because of their resistance to oxidative attack (after Gerarde, *30*).

diverting groups is manifested by a reduction in phenol, and therewith organic sulfate, excretion. Branched alkyl groups, not suitable for a rapid oxidative conversion, are, therefore unsuitable for this purpose (*3, 30*). Compare benzene, with its leukemiogenic action (*81, 83*), with toluene, a relatively safe solvent. In a study of chemical sequestration after exposure to benzene and toluene based on autoradiography and organic solvent extraction, Bergman (1979) observed that exposure to benzene resulted in nonextractable residues, especially in the liver, which was not the case with toluene. The risky oxidative attack on the aromatic ring is diverted in toluene to the methyl group (*9*). A study of the potentials of methyl, ethyl, etc. groups as alternative metabolic handles is promising. One has to count with the possibility that conjugation products can also have alkylating properties (Fig. 21) (*10, 64a*).

The reduction of metabolic conversion by molecular manipulation has particular advantages in drug design. Since, for instance, species differences are to a large extent based on differences in metabolic capacity, fewer species differences will be found after metabolic stabilization; thus, data obtained in animals can be more readily transferred to the human situation. Moreover, lower dosages will be needed. The number and quantity of metabolites formed, and therewith the complexicity of biological and toxicological evaluation, will be reduced. A more direct and thus clearer relationship between structure and activity, the basis for an optimal design of new agents, is a further advantage (*4*).

d. Reduction in Toxicity by Metabolic Stabilization. As can be concluded from the foregoing sections, the control of drug metabolism—be it either in the sense of developing metabolically stable agents or in developing agents metabolized along controlled pathways, avoiding risky metabolic handles, and making use of safe, nontoxicogenic, metabolic handles—may reduce the toxicological risks (*5, 6*).

The metabolically stable agents designed will not only reduce the risk of formation of reactive, biologically alkylating intermediates, and therewith the risk of a carcinogenic, mutagenic, teratogenic, allergenic, irritating, or cell-damaging action in general, but will have further advantages. For instance, for such agents a better correlation between animal and human data may be expected, since species differences are mainly based on differences in drug-metabolizing capacity. Moreover, with such agents fewer patient-to-patient differences in potency and fewer disturbing interactions will occur, since in many of these cases drug metabolism is involved. The dosages required for stable agents may often be reduced to a fraction of what is required for metabolically instable agents, while lower frequencies of dosage or applications will suffice. As a result of first-path elimination by metabolic degradation in the liver after oral appli-

cation, a large fraction of the dose applied is wasted before action, whereas rapid metabolic inactivation in general implies only a short action and thus requires more frequent application. The reduction in quantity or dose needed also implies a reduction in the load on the environment.

The fact that chemical agents are stable with regard to the enzymes involved in metabolic conversion—inactivation and degradation—in animals and plants, in no way means that such agents would also be resistant against the microflora in the soil and in sewage and sewage clearing installations. If drugs with a short half-life are required, even for the metabolically stable agents a relatively rapid clearance from the body can be obtained by proper balancing of the lipid–water solubility, thus guaranteeing sufficiently rapid renal excretion. This appears to be compatible with action on the CNS, which requires a certain degree of lipophilicity.

On the other hand, there is a possibility of making use of controlled metabolic pathways by introduction into the molecule of labile but safe metabolic handles such as, under circumstances suitable, alkyl substituents and ester groups [see Fig. 23 (*30*) and Fig. 24].

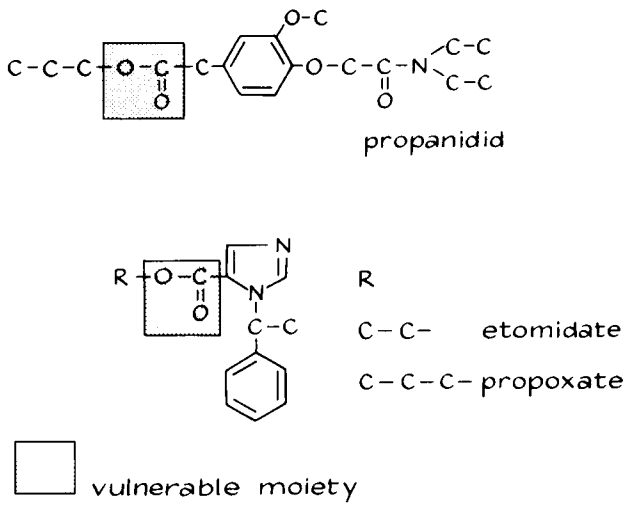


Fig. 24. Introduction of carboxy ester-type vulnerable moieties, considered to be safe metabolic handles in order to obtain ultra short-acting iv anesthetics.

C. PREVENTIVE EARLY DETECTION OF EXPOSURE TO POTENTIALLY TOXIC AGENTS

Another aspect of the growing insight into the molecular basis of toxic actions, based on the induction of chemical lesions, such as carcinogenesis, mutagenesis, etc., is the possibility of an early detection of

risky exposure. As indicated, part of the reactive intermediate metabolic products appears in the urine as mercapturic acid derivatives. As a consequence, an increase in the excretion of mercapturic acid derivatives over the normal value indicates that the individual concerned is exposed to chemicals that are suspicious since they are metabolized via biologically alkylating intermediate products. Thus, measurement of mercapturic acid excretion ("mercapto test") can give an early exposure warning, before unacceptable damage has occurred.

A positive mercapto test is an indication of the fact that the individual is "in defense" against a risky exposure to chemicals [see Figs. 25 and 26,

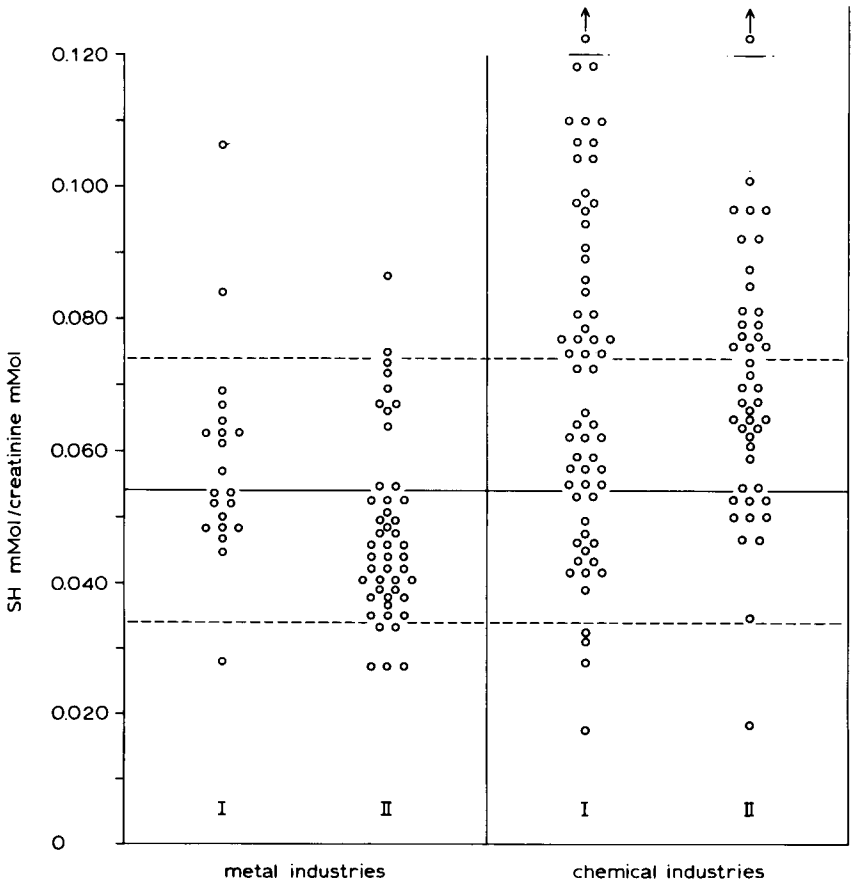


Fig. 25. Comparison of urinary levels of mercapturic acid derivatives and other thioethers in groups of individuals working in different metal or chemical factories. The horizontal, unbroken, and dotted lines indicate the average background value ± S.D. obtained with fifty students (Seutter-Berlage *et al.*, *76*).

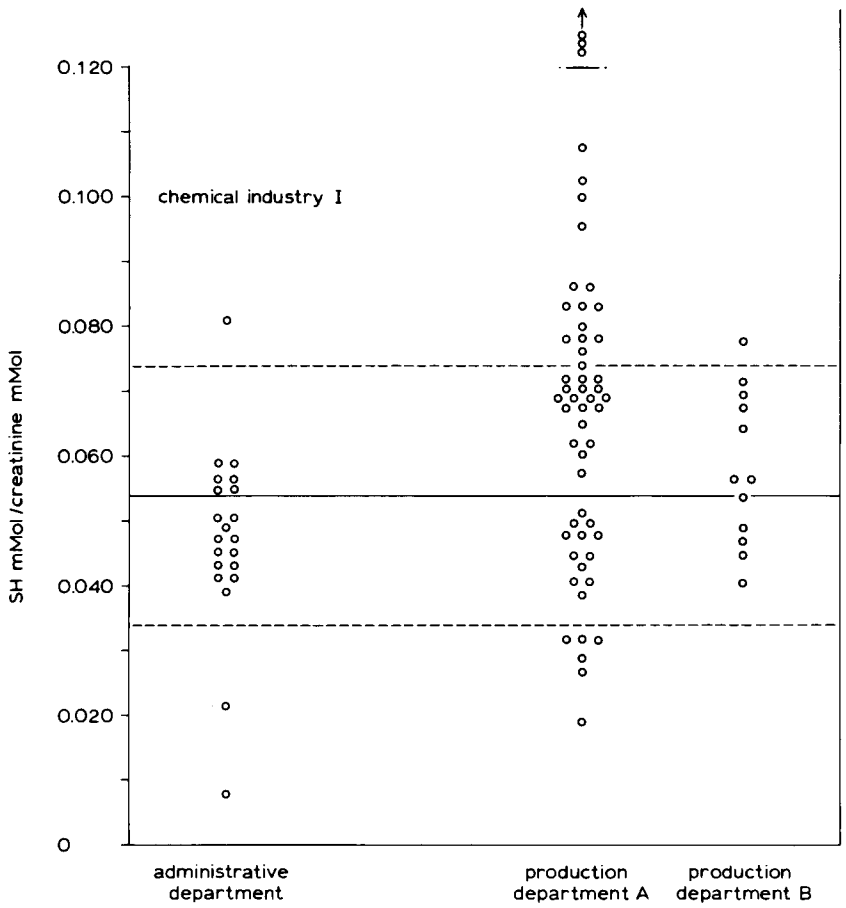


Fig. 26. Comparison of the urinary excretion of mercapturic acid derivatives and other thioether compounds of employees engaged in an administrative department and two different production departments of a chemical industry. Department A: possible exposure to a mixture of different chemicals, among which the following compounds may be suspect: dimethylformamide, acrylonitrile, diethyleneglycol, oleic acid, biphenyl, hydroquinone monomethylether. Department B: possible exposure to a wide variety of other chemicals, including paint removers and organophosphorus compounds. The horizontal, unbroken, and dotted lines indicate the average background value ± S.D. obtained with fifty students (Seutter-Berlage *et al.*, *76*).

TABLE XI
URINARY THIOETHER EXCRETION[a]

Male[b]	Clerks	Condensator workers	Rubber workers
	30 ± 7 (20)	28 ± 7 (8)	51 ± 17 (3)
Female[b]	Clerks	Crossply-tire builders	Radial-tire builders
	37 ± 9 (16)	37 ± 8 (12)	63 ± 14 (16)

[a] μmol/mmol creatinine ± SD with number of subjects in brackets. Vainio *et al.* (*82a*).
[b] Nonsmokers, no medication.

Seutter-Berlage *et al.* (*76*) and Table XI (*82a*)]. An advantage is the general character, the "nonselectivity" of the test; it does not indicate exposure to one particular substance but gives an indication of inacceptable exposure in general and is thus a signal to take action. It is not a test meant for the identification of metabolites, nor is it a test for exposure to particular carcinogens that may be active in extremely low dosages and thus not detectable as mercapturic acid derivatives in the urine. Its primary significance is as an alarm signal for industrial and environmental hygienists. A negative test cannot be regarded as an indication that everything is all right.

Certain epoxides, for instance, may be eliminated predominantly via the epoxide hydratase pathway; this results in an increased phenol excretion which, taking into account the relatively large amounts and the variability in the normal phenol excretion, is hardly detectable. The mercapto test, which also detects organic "mercapto conjugates" formed by biological alkylation of SH-groups different from those in glutathione (e.g., cysteine SH-groups), can possibly also serve as an indicator for potential toxicity in the search for new, safer compounds and thereby contribute to their design.

V. The Toxodynamic Phase

The toxodynamic phase consists of the processes leading from the activation or blockade of specific receptors or from the induction of the chemical lesion in critical biopolymers via a sequence of biochemical and biophysical steps to the final effect observed. Only the first step, the interaction with the specific receptors or the induction of the chemical lesion, is dependent on the chemical structure of the active agent.

A. REDUCTION OF TOXICITY BY INCREASE IN POTENCY

With regard to the final induction of the effect, an increase in selectivity of action by definition reduces toxicity. This, as a rule, requires a high affinity of the agents concerned for their specific receptors, sites of action, such that the concentration required for action is low. The chance that one molecule fulfils the highly specific requirements for high affinity to two or more different types of receptors is small. The concentrations required for the highly active agents are so low that actions on sites with low affinity are practically excluded. Inherent in this situation is the fact that only small dosages or quantities of the agents concerned are required, especially if by proper design rapid biochemical degradation—in fact often a wastage—is reduced to a minimum.

Since the toxic effects based on biological alkylation are the result of an irreversible binding, a reduction in the quantity of the chemical involved will more or less proportionally reduce the extent of the chemical lesions. Not the concentration as such, but much more the quantity of the agent involved is in the end determinant for the effect.

The reduction in the quantity of the agent required reduces the load on environment and biosphere in general, thus reducing the risk that toxicologically active concentrations are reached.

The often heard reasoning that a drug active at, for example, a 0.1 mg dose has no advantage over one for which 50 mg is required, since the patient can as easily take (swallow) 50 mg as 0.1 mg, does not hold true.

With regard to minimizing the quantities of chemical agents required, the use of pheromones or pheromonomimetics, highly selective and active in usually extreme low concentrations, can be mentioned in the control of insects. Substitution by such agents for the classical organophosphate- or DDT-type agents, which have a more indifferent type of action and require relatively high dosages, will undoubtedly be helpful in reducing toxicological risks (*74*).

VI. Relationship Between Structure and Toxic Action

In a discussion of the relationship between chemical structure and toxic action, it makes sense to differentiate among different types of toxic action.

1. Toxic actions based on an interaction of the toxic agent and specific sites of action with specific receptors, which implies a relatively strict structure–activity relationship. This type of action is usually

reversible. Most of the pharmacological actions made use of in therapeutics are of this nature. One could indicate these types of toxic actions as "undesired pharmacological actions." In many cases the question whether such actions are desired or not is greatly dependent on the circumstances. Effects such as inhibition of salivation, mydriasis, inhibition of gastric secretion and intestinal motility, as caused by atropine-like agents, will be regarded as either therapeutic or adverse toxic effects, depending on the pathological condition for which these drugs are used.

2. Toxic actions based on irreversible interaction of the active agent with critical biopolymers in the organism such as DNA, enzymes, and proteins in general, with as a consequence chemical lesions and the inherent actions, are regarded as toxic under practically all circumstances. This also holds true if such biologically alkylating agents are used as, for instance, cytostatics in the treatment of cancer. These toxic effects are inherent in their mechanisms of action. This type of toxic action is not, or is to a much lesser degree, based on interactions of, or a specific fit between, the active agent and specific receptors. It is a much more general type of chemical reaction—alkylation or acylation—of the target molecule by the toxic agent. As a consequence, a structure–activity relationship for this type of action is not manifested in the structure of the molecule as a whole, but much more by the presence in the molecule of particular reactive—biologically alkylating—groups or precursors thereof, indicated as toxicogenic groups. This makes it understandable that these types of toxic action, namely carcinogenesis, mutagenesis, allergic sensitization, etc. are properties of compounds totally different in nature, both from the biological and from the chemical point of view. This implies that in efforts to eliminate such a toxic action, identification and classification of toxicogenic groups are essential. It must be taken into account that, in many cases, the ultimate toxicogenic group, the biologically alkylating group, is generated by metabolic conversion of the toxicogenic group. On basis of the information thus obtained, the design of new and the adaptation of existing chemical agents, aimed at avoidance or elimination of various worrisome toxic actions, will be possible.
3. Toxic actions related to physical sequestration and accumulation in the body of mostly inert agents. This, although not clearly toxic in a strict sense, is considered as undesirable from a general hygienic, and especially from an environmentally hygienic, point of view. Possibilities for avoidance thereof by suitable design of chemicals requires insight, especially in the relationship between structure and

metabolic conversion of xenobiotics. Especially important will be a classification of the various metabolic handles in relation to their metabolic stability or vulnerability and to the possible generation of ultimate toxicogenic groups in the process of their metabolic conversion.

The discussion of the relationship between structure and toxic action given here will be restricted to the actions based on the induction of chemical lesions.

In the efforts to study and systematize SAR for chemical carcinogenesis, mutagenesis, allergic sensitization, etc. and in general for chemicogenic diseases, including iatrogenic diseases—based on chemical lesions—a distinction should be made between the directly acting toxons, compounds which have biologically alkylating properties as such (*16, 20, 21, 22, 23, 24*), and indirectly acting toxons, whose toxic action is based on the formation of reactive intermediate metabolic products (*21, 37, 49, 63*).

For the directly acting toxons, one can distinguish between biological alkylation (an attack of electrophilic groups in the toxon on nucleophilic groups in biopolymers) and biological acylation (coupling of the biopolymer, usually a protein with an acid moiety in which β-lactams, phenol esters, and acid anhydrides may be involved). The directly alkylating and acylating agents may be recognized on basis of their chemical structures or chemical properties. Figures 27 and 28 give examples. One should be well aware of the fact that, in certain cases, impurities, and not the compounds as such, are involved in the toxic actions (*15*). Besides

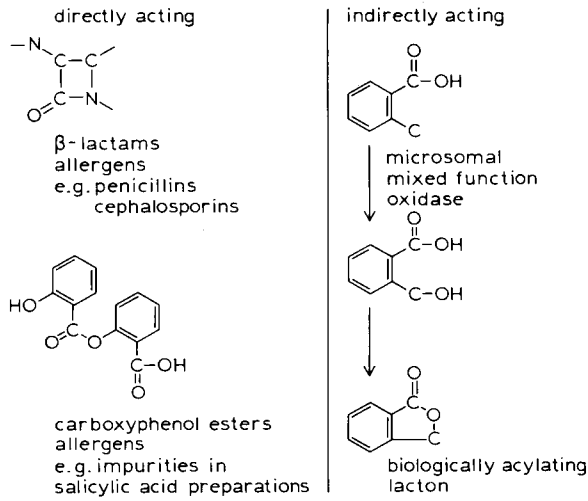


Fig. 27. Structures with potential biological acylating action.

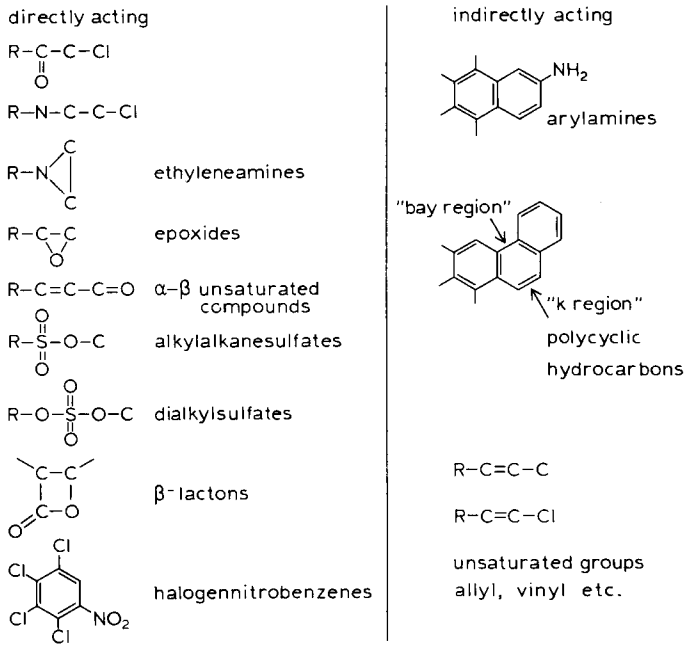


Fig. 28. Structures with potential biological alkylating action.

the types of agents mentioned in these tables, a wide variety of other substances (*17*) are converted to glutathione conjugates and mercapturic acid derivatives, indicating potential alkylating properties. In the design of new chemical agents, and from a concern with industrial hygiene, especially in the chemical industry and including laboratories, the suspect character of particular structures should be taken into account.

Incubation of agents suspected of a direct alkylating or acylating action *in vitro,* at the proper pH, with suitable target molecules such as cysteamine, amino acids, small polypeptides, and eventually proteins, may give useful information (*15*). For the indirectly acting agents, a host- or microsome-mediated (mixed-function oxidase) test will be required. An example is the Ames test, where mutagenicity is tested on the basis of back mutation of particular strains of bacteria exposed to the chemical to be investigated, in the presence of liver microsomal preparations (*2, 58, 73*). Early detection and, where possible, prediction of mutagenic and therewith probable carcinogenic action are the aims.

The possibility of understanding and predicting the mutagenic and therewith potential carcinogenic action of compounds will be more within reach with simple than with complex molecules. An example of the

former are the halogenated saturated and unsaturated alkyl compounds such as vinyl chloride (world production in 1973 was 1050×10^4 tons) vinyliden chloride (1950×10^4 tons), trichloroethylene (101×10^4 tons), and 1,1,1-trichloroethane (48×10^4 tons), widely used in the synthesis of plastics, as solvents, and as additives to a large variety of household chemicals. The toxicity of this group of compounds has been systematically studied by Greim *et al.* (*30*) and Henschler (*39*). The mutagenic action, from a modified Ames test, requires the presence of microsomal mixed function oxidase, which indicates an indirect action in which reactive intermediate metabolic products, in this case epoxides, are involved. For the compounds, 1-chloroethylene (vinyl chloride), 1,1-dichloroethylene (vinyliden chloride), and trichloroethylene, the mutagenic action has been well-established (Fig. 29) (*39*). For the former two com-

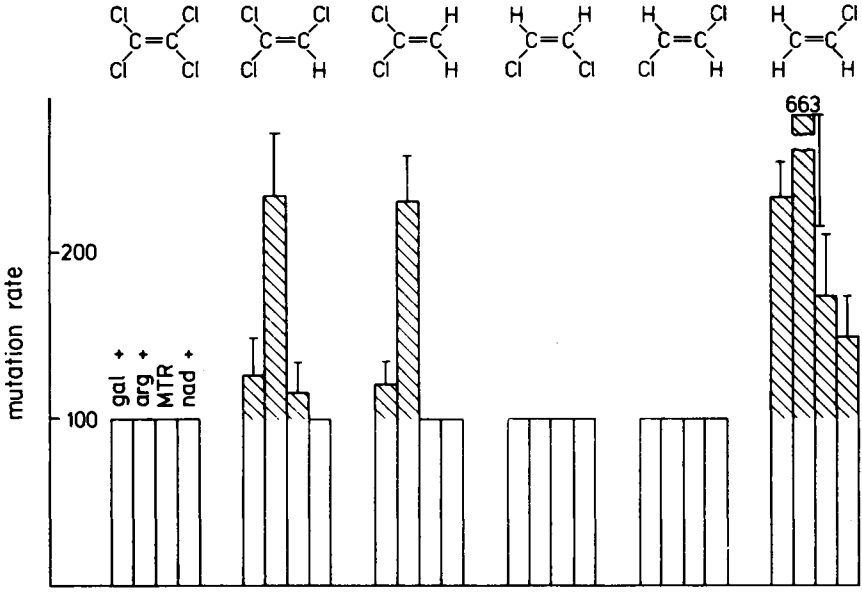


Fig. 29. Mutagenicity of chlorinated ethylenes tested in a modified Ames test on *E. coli* K 12 after incubation with mouse liver microsomes. Given is the increase of the mutation rate for three back (gal^+, arg^+ and nad^+) and one forward (MTR) mutation system. The rate at control conditions is 100%. Note the high mutation rate with vinyl chloride (1-chloroethylene) and the absence of a mutagenic action in the "symmetric" compounds in which the chlorine atoms are partitioned equally on both carbon atoms (after Henschler, *39*).

pounds, a carcinogenic action is definitely observed; the experiments with trichloroethylene are positive, but not yet definitive. The epoxides formed from the symmetric chloroethylene derivatives appear

relatively stable and nonmutagenic while the epoxides formed from the asymmetric derivatives are unstable and mutagenic (Fig. 30) (*39*). In the asymmetric Cl-substituted epoxides, the asymmetry in the charge distribution enhances the electrophilicity and therewith the mutagenic action.

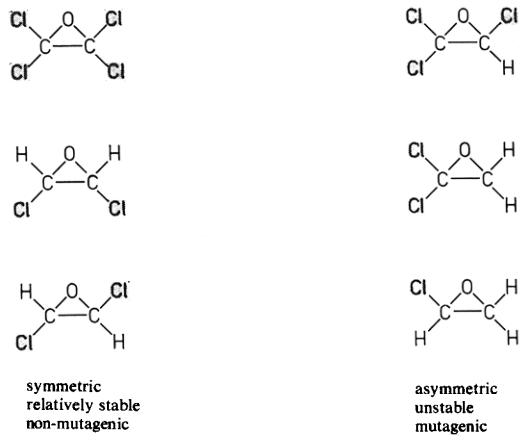


Fig. 30. Relationship between chemical structure reactivity and mutagenicity of epoxides (oxiranes) formed from various chlorinated ethylenes (after Henschler, *39*).

The principles previously outlined imply that classical structure–action relationship studies, which predominantly involve the pharmacodynamic aspects of the agents, have to be extended to a study of the relationship between structure and metabolic conversion. A special aim will be the recognition and classification of toxicogenic groups that are converted via biologically alkylating intermediates, leading to the induction of biochemical lesions.

For detection of the latter, a number of short-term *in vitro* tests for rapid screening of compounds are available. They detect mutagenic and/or carcinogenic actions, between which there is a large overlap (*2*). In addition to the Ames test, the cell transformation test, the test for chromosomal aberrations and, related therewith, the sister chromatid exchange test, the DNA repair-synthesis test, and the reaction with nucleic acids, can be mentioned. With suitable combinations of such tests, potentially risky agents can be picked up with a close to 100% yield (*2, 12, 14, 56, 73, 80*).

These screening tests are therefore promising as guides to safe chemicals if applied very early in the procedure of development of new chemical agents, including bioactive agents such as drugs, pesticides, etc. They will

give an early indication as to whether the type of structure used contains toxicogenic groups, so that safer structures can be given preference. Such studies will undoubtedly contribute to the insight into the relationship between chemical structure and toxic action.

This does not yet mean that, on basis of such *in vitro* studies, definite conclusions can be drawn as far as admission of agents to the market is concerned. For the time being, long-term animal studies are still regarded as necessary to reach optimal safety in this respect. On the other hand, the use of short-term tests, as indicated, will undoubtedly drastically reduce the risk that in the end, after elaborate, expensive and time-consuming investigations, the agents selected as most promising for further development will have to be discarded because of failure in animal studies for carcinogenic, mutagenic, etc. action.

The fact that the *in vitro* tests thus applied may sometimes give false positives should not be taken too seriously. The relatively large degree of freedom in structure–action relationships observed for many if not most types of biologically active agents—and this holds true for technical aims for new chemical products too—implies that although a few potentially useful agents may be discarded, this seldom will be prohibitive as far as reaching the final goal: the development of safe, optimally effective chemical agents, is concerned.

VII. Conclusion

There are good arguments that 60 to possibly even 90% of cancers are preventable (*25, 90*). The same may hold true for chemicogenic diseases in general. This implies that proper analysis of structure–action relationships in the field of carcinogenicity and the related fields of mutagenicity, allergic sensitization etc., and application of the insights gained into the design of safer chemicals, be it drugs, pesticides, household chemicals, or whatever, will undoubtedly contribute to that prevention. Asking for 100% safety is as unwise as neglecting the potential risks of chemistry and its products is unwarantable. The practice will have to be a compromise, a sound balance, between higher safety and technological and economic feasibility.

Although there is still a long way to go, there are perspectives for the development of safer chemicals. This is essential since life without chemicals is impossible, taking into account the large world population, its urbanization and high standards of living, with the inherent necessity of mass production and conservation of food, massive substitution of scarce natural products, and progressive technology.

ACKNOWLEDGMENT

The author wants to thank Mrs. A. Wigmans for her efficient help in the preparation of the manuscript.

References

1. S. B. A. Åkerman, S. B. Ross, and A. Tĕlc, *Acta Pharmacol. Toxicol.* **32,** 88 (1973).
2. B. N. Ames and K. Hooper, *Nature (London)* **274,** 19 (1978).
3. E. J. Ariëns, "Drug Design" Vol. 2, pp. 68, 97, 98, 100, 114. Academic Press, New York, 1971.
4. E. J. Ariëns and A. M. Simonis, "Topics in Current Chemistry" Vol. 52, p. 3. Springer-Verlag, Berlin and New York, 1974.
5. E. J. Ariëns, A. M. Simonis, and J. Offermeier, "Introduction to General Toxicology," 2nd ed. Academic Press, New York, 1978.
6. E. J. Ariëns and A. M. Simonis, *In* "Drug Design and Adverse Reactions," (H. Bundgaard, P. Juul, and H. Kofod, eds.) p. 317. Munksgaard, Copenhagen and Academic Press, New York, 1977.
7. A. H. Beckett and M. Rowland, *J. Pharm. Pharmacol.* **17,** 628 (1965).
8. K. Bergman and H. Tjälve, *Acta Pharmacol. Tox.* **41,** Suppl. 1, 22 (1977).
9. K. Bergman, *Scand. J. Work, Environm. Health,* **5,** Suppl. 1, (1979).
10. K. W. Bock, *Arch. Toxicol.* **39,** 77 (1977).
11. E. Boyland, *Brit. Med. Bull.* **20,** 121 (1964).
12. B. A. Bridges, *Nature (London)* **261,** 195 (1976).
13. B. B. Brodie, *In* "Pharmacology and the Future of Man," Vol. 2, p. 48. Karger, Basel, 1973.
14. P. Brookes, *Life Sciences* **16,** 331 (1975).
15. H. Bundgaard, *In* "Drug Design and Adverse Reactions," (H. Bundgaard, P. Juul, and H. Kofod, eds.) p. 165. Munksgaard, Copenhagen and Academic Press, New York, 1977.
16. M. Calvin, *Naturwissenschaften* **62,** 405 (1975).
17. L. F. Chasseaud, *In* "Glutathione: Metabolism and Function" (I. M. Arias and W. B. Jakoby, eds.) p. 77. Raven Press, New York, 1976.
18. A. S. Crafts, *Advan. Pest Control Res.* **1,** 39 (1957).
19. J. H. Dewaide, "Metabolism of Xenobiotics". Dissertation, Univ. Nijmegen. Drukkerij Leijn, Nijmegen, 1971.
20. H. Druckrey, R. Preussmann, N. Nashed, and S. Ivankovic, *Z. Krebsforsch.* **68,** 103 (1966).
21. H. Druckrey, R. Preussmann, S. Ivankovic, and D. Schmähl, *Z. Krebsforsch.* **69,** 103 (1967).
22. H. Druckrey, R. Preussmann, J. Bücheler, and S. Ivankovic, *Z. Krebsforsch.* **71,** 105 (1968).
23. H. Druckrey, H. Kruse, R. Preussmann, S. Ivankovic, and Ch. Landschütz, *Z. Krebsforsch.* **74,** 241 (1970a).
24. H. Druckrey, H. Kruse, R. Preussmann, S. Ivankovic, Ch. Landschütz, and J. Gimmy, *Z. Krebsforsch.* **75,** 69 (1970b).
25. Editorial, *Lancet* **1,** 685 (1977).
26. H. N. Eisen, "Cellular and Humoral Aspects of the Hypersensitive States." P. B. Hoeber, New York, 1959.

27. J. Feijen and S. W. Kim, *In* "Drug Design" (E. J. Ariëns, ed.) Vol. 10. Academic Press, New York, in preparation.
28. A. J. Funckes, *J. Agric. Food Chem.* **11,** 455 (1963).
29. D. D. Garner, K. M. Cano, R. S. Peimer, and T. E. Yeshion, *J. For. Sci.* **21,** 816 (1976).
30. H. W. Gerarde, "Toxicology and Biochemistry of Aromatic Hydrocarbons." Elsevier, Amsterdam, 1960.
31. J. R. Gillette, *In* "The Poisoned Patient: the Role of the Laboratory." Excerpta Medica, Amsterdam, 1974.
31a. C. S. Giam, H. S. Chan, G. S. Neff, and E. L. Atlas, *Science* **199,** 419 (1978).
32. J. R. Gillette, J. R. Mitchell, and B. B. Brodie, *Annu. Rev. Pharmacol.* **14,** 271 (1974).
33. J. R. Gillette, *In* "Pathogenesis and Mechanisms of Liver Cell Necrosis," (D. Keppler, ed.), p. 239. MTP Press, Lancaster, 1975.
34. J. R. Gillette and L. R. Pohl, *J. Toxicol. Environ. Health* **2,** 849 (1977).
35. H. Greim, G. Bonse, Z. Radwan, D. Reichert, and D. Henschler, *Biochem. Pharmacol.* **24,** 2013 (1975).
36. J. B. Hunn and J. L. Allen, *Annu. Rev. Pharmacol.* **14,** 47 (1974).
37. Ch. Heidelberger, *Fed. Proc.* **32,** 2154 (1973).
38. Ch. Heidelberger, *Annu. Rev. Biochem.* **44,** 79 (1975).
39. D. Henschler, *Arzneim.-Forsch.* **27,** 1827 (1977).
40. V. R. Holland, B. C. Saunders, F. L. Rose, and A. L. Walpole, *Tetrahedron* **30,** 3299 (1974).
41. P. B. Hulbert, *Nature (London)* **256,** 146 (1975).
42. H. W. Huyser, Original Lectures 3rd. Int. Congress of Surface Activity, Cologne, Vol. III, Section C, 295, 1960.
43. Ch. C. Irving, "Proc. Symposium on The Biological Oxidation of Nitrogen in Organic Molecules." (J. W. Gorrod, ed.) p. 75. Taylor and Francis, London, 1971.
45. M. S. Jacobson, S. V. Kevy, and R. J. Grand, *J. Lab. Clin. Med.* **89,** 1066 (1977).
46. R. J. Jaeger and R. J. Rubin, *New Engl. J. Med.* **287,** 1114 (1977).
47. W. B. Jakoby and J. H. Keen, *TIBS* **2,** 229 (1977).
48. A. M. Jeffrey, I. B. Weinstein, K. W. Jennette, K. Grzeskowiak, K. Nakanishi, R. G. Harvey, H. Autrup, and C. Harris, *Nature (London)* **269,** 348 (1977).
49. D. M. Jerina and J. W. Daly, *In* "Drug Metabolism" (D. V. Parke and R. L. Smith, eds.) Taylor and Francis, London, 1976.
50. D. M. Jerina and J. R. Bend, *In* "Biological Reactive Intermediates" (D. J. Jollow, J. J. Kocsis, R. Snyder, and H. Vainio, eds.) p. 207. Plenum Press, New York, 1977a.
51. D. M. Jerina, H. Yagi, W. Levin, and A. H. Conney, *In* "Drug Design and Adverse Reactions" (H. Bundgaard, P. Juul, and H. Kofod, eds.) p. 261. Munksgaard, Copenhagen and Academic Press, New York, 1977b.
52. F. F. Kadluber, J. A. Miller, and E. C. Miller, *Cancer Res.* **37,** 805 (1977).
53. E. Kriek, *Biochim. Biophys. Acta* **355,** 177 (1974).
54. A. Lacassagne, F. Zajdela, and N. P. Buu-Hoi, *Naturwissenschaften* **53,** 583 (1966).
55. K. Landsteiner, "The Specificity of Serological Reactions." Harvard University Press, Cambridge, Mass., 1945.
56. P. N. Magee, *In* "Prevention and Detection of Cancer" (H. E. Nieburgs, ed.) Part I, Vol. 2, p. 1935. Marcel Dekker, New York, 1977.
57. T. H. Maugh, *Science* **183,** 940 (1974).
58. J. McCann, E. Choi, E. Yamasaki, and B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.* **72,** 5135 (1975).
59. R. J. McIsaac and G. B. Koelle, *J. Pharmacol. Exp. Ther.* **126,** 9 (1959).

60. R. J. McMurtry, W. R. Snodgrass, and J. R. Mitchell, *Toxicol. Appl. Pharmacol.* **46,** 87 (1978).
81. R. L. Metcalf, *In* "The Future for Insecticides" (R. L. Metcalf and J. J. McKelvey, eds.) p. 223. John Wiley and Sons, New York, 1976.
62. E. C. Miller, J. A. Miller, and H. A. Hartman, *Cancer Res.* **21,** 815 (1961).
63. J. A. Miller, *Cancer Res.* **30,** 559 (1970).
64. D. P. Morgan and C. C. Roan, *In* "Essays in Toxicology" (W. J. Hayes, ed.) Vol. 5, p. 39. Academic Press, New York, 1974.
64a. G. J. Mulder, J. A. Hinson, and J. R. Gillette, *Biochem. Pharmacol.* **26,** 189 (1977).
65. K. Nakanishi, H. Kasai, H. Cho, R. G. Harvey, A. M. Jeffrey, K. W. Jennette, and I. B. Weinstein, *J. Am. Chem. Soc.* **99,** 258 (1977).
66. S. D. Nelson, A. J. Forte, and R. J. McMurty, *Res. Commun. Chem. Pathol. Pharmacol.* **22,** 61 (1978).
67. R. D. O'Brien, "Toxic Phosphorus Esters." Academic Press, New York, 1960.
68. R. D. O'Brien, *Ann. N.Y. Acad. Sci.* **123,** 156 (1965).
69. F. Oesch, *Arzneim.-Forsch.* **27,** 1832 (1977).
70. Y. Omori, *Environ. Hlth. Perspect.* **17,** 203 (1976).
71. F. J. Oppenoorth and A. F. H. Besemer, *Chem. Weekblad* m99 (1977).
72. Ch. W. Parker, *New Engl. J. Med.* **292,** 511 (1975).
73. I. F. H. Purchase, E. Longstaff, J. Ashby, J. A. Styles, D. Anderson, P. A. Lefevre, and F. R. Westwood, *Nature (London)* **264,** 624 (1976).
74. F. J. Ritter and C. J. Persoons, *In* "Drug Design" (E. J. Ariëns, ed.) Vol. 7, p. 20. Academic Press, New York, 1976.
75. T. S. Scott, "Carcinogenic and Chronic Toxic Hazards of Aromatic Amines." Elsevier, Amsterdam, 1962.
76. F. Seutter-Berlage, H. L. van Dorp, H. G. J. Kosse, and P. Th. Henderson, *Int. Arch. Occup. Environ. Hlth.* **39,** 45 (1977).
77. J. R. Sibert, A. W. Craft, and R. H. Jackson, *Lancet* **2,** 289 (1977).
78. P. Sims and P. L. Grover, *Adv. Cancer Res.* **20,** 165 (1974).
79. B. Singer, *TIBS* **2,** 180 (1977).
80. A. D. Tates, *T. Soc. Geneesk.* **55,** 140 (1977).
81. J. A. Timbrell and J. R. Mitchell, *Xenobiotica* **7,** 415 (1977).
82. H. Uehleke, *In* "Pharmacology and the Future of Man." Vol. 2, p. 124. Karger, Basel, 1973.
82a. H. Vainio, H. Savolainen, and I. Kilpikari, *Br. J. Ind. Med.* **35,** 232 (1978).
83. E. C. Vigliani and G. Saita, *New Engl. J. Med.* **271,** 872 (1964).
84. M. Wandel and H. Tengler, *Deutsche Lebensm. Rundschau* **11,** 326 (1963).
85. A. L. de Weck, *Naturwissenschaften* **62,** 162 (1975).
86. E. K. Weisburger, *Ann. Rev. Pharmacol. Toxicol.* **18,** 395 (1978).
87. J. H. Weisburger and E. K. Weisburger, "Biochemical Formation and Pharmacological and Pathological Properties of Hydroxylamines and Hydroxamic Acids." Williams and Wilkins, Baltimore, 1973a.
88. J. H. Weisburger and E. K. Weisburger, *Pharmacol. Rev.* **25,** 1 (1973b).
89. World Health Organization, "Specifications for Identity and Purity of Food Additives," Vol. 2. FAO, Rome, 1963.
90. E. L. Wynder, *Nature (London)* **268,** 284 (1977).

Chapter 2 Consequences of the Hansch Paradigm for the Pharmaceutical Industry

Stefan H. Unger

ISBN 0-12-060309-8

I. Introduction

The usual meaning of paradigm is "accepted model or pattern." However, T. S. Kuhn (*1*) has recently used the term in a much broader and more powerful sense: A paradigm is not just a simple model but an entire system of thought which allows one to solve problems "scarcely imagined and never undertaken" in its absence. Without agreeing with all of Kuhn's conclusions, it is still useful to consider the "Hansch paradigm" for drug design*, as there are few who would deny that the early work of Hansch and Fujita (*2a–c*)—demonstrating the application of extrathermodynamic concepts and linear regression analysis—represented a turning point in quantitative structure–activity relationships, QSAR. Although many other workers laid the ground work by providing important seminal ideas and techniques, it was the Pomona group who put these ideas into an *operational* state and who most fully exploited the consequences of this paradigm. While there have been numerous excellent review articles on the "how to" aspects of QSAR (*3a–e*), surprisingly little has been written on the "why" or "what of it". It is these aspects that I wish to cover; and I hope to indicate that, indeed, problems "scarcely imagined" can not only be conceived but also solved. The reader is referred to the many excellent reviews for technical details on the methods cited.

Before continuing, it might be useful to outline some of the assumptions underlying the paradigm:

1. Biological activity is a function of the structure of the drug.
2. The structure of a drug implies certain global properties such as lipophilicity, net charge, solubility, etc., and certain local properties such as distribution of lipophilicity, charge, and bulk at particular positions in the molecule.
3. These global and local properties can be quantified by extrathermodynamic parameters (*4a–g*) such as π, σ, MR, E_s etc. and by various molecular orbital (MO) parameters, indicator variables, and substructural parameters.
4. A statistical model relating changes in biological response to changes in global and local properties of the drug *exists,* although it may not be either a simple model or simple to discover.

* I have taken the liberty of renaming the "QSAR paradigm" (cf. *26e*).

As a consequence of this paradigm, it is possible to make predictions, test hypotheses concerning causal factors, and study relationships between compounds and bioassays. All this would be possible with any model, but the significance of the Hansch paradigm is that it is both *quantitative* and *statistical.* No statistical model can be any better than the quality of the data upon which it is based; nor can the use of a statistical model guarantee skill in interpretation or application.

It has been about 17 years since the first papers appeared (*2a–c*), and the field has undergone considerable maturation. There are a large number of improved lipophilic, electronic and steric parameters to consider, several complicated, nonlinear models for relating biological response to lipophilicity have been proposed, and the problems of conformation are just starting to be examined in a quantitative manner. QSAR has evolved into a complicated subspeciality and there are a number of full-time practitioners in pharmaceutical companies throughout the world. In addition to special sessions at various scientific meetings, a Gordon Research Conference, "QSAR in Biology," has been held biannually since 1975, with a high percentage of industrial attendees.

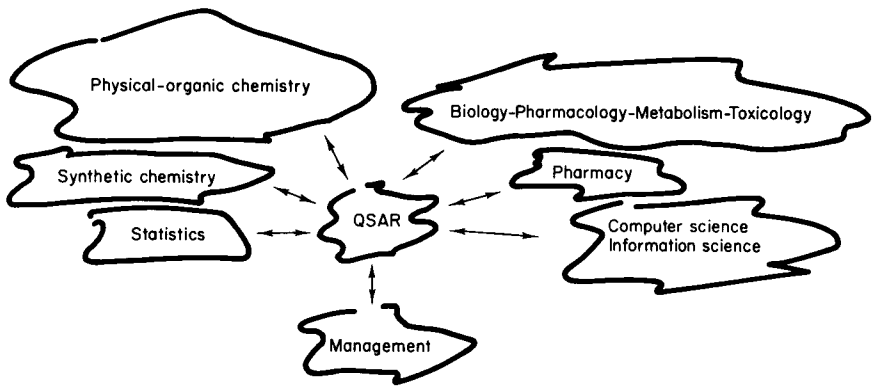


Fig. 1. Central nature of QSAR.

In the following I have in no manner attempted to survey the literature regarding either techniques or applications of QSAR in industrial laboratories, but rather I have focused mainly on activities at Syntex Research during the approximately six years that I have been there. The purpose of this article is to illustrate the many types of activities in which a QSAR specialist can become involved as a consequence of the Hansch paradigm. This outlook places the drug designer in a central position, as indicated in Fig. 1.

II. Design of Biological Experiments

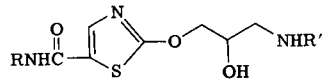
A. Desirability of Fixed Response Data

A statistical model of the change in biological response (BR) for a series of analogs requires that the biological response be expressed in a suitable manner. Since these are extrathermodynamic models, it has often been assumed (*3a–e*) that the BR should be in a form proportional to free energy, thus the familiar log $1/C$, where C is the concentration of drug producing a specified response, e.g., 50%. The reciprocal is used because of the convention that the BR be arranged such that a more positive number is a more desirable characteristic (for toxicity, log C should be used).

Unfortunately, such factors as difficulty of synthesis, time or cost of biological assay, and lack of understanding of the requirements of QSAR often mitigate against successful QSAR work. The biologist* may be interested in finding either the most active compound or the one with the best overall profile. He might look upon each analog as a potential candidate for development, rather than as a piece in a puzzle providing information that may eventually lead to an optimal compound. Much data is of the fixed dose screening type, which is easier to obtain but more difficult to use in QSAR studies. Part of this attitude may have arisen from the apparent lack of order that seemed to govern structure–activity studies in the past. QSAR is slowly providing this order and demonstrating that it is worth the extra effort to provide fixed response data.

Example: Thiazole Beta-Blockers

We have been working as a team (*5a,b*) directed toward optimization of a series of alkyl carbamoyl substituted thiazolyloxy beta-blockers (**I**), one



(I)

member of which (R = 2-Me-hexyl, R′ = *t*Bu) was selected for development as a cardioselective beta-blocker shortly before the QSAR work started (*6a*). The series was developed from the mixed myocardial beta-stimulant/beta-blocker tazolol (*6b*), which lacked the alkyl carbamoyl side

* In the interests of euphony, biologist or biology will be used to mean the whole cluster of disciplines involved with experimentation or understanding of living systems, e.g., biochemistry, pharmacology, metabolism, toxicology, etc.

chain. The group R had been varied over a wide range of linear and branched alkyl, and some cycloalkyl, alkenyl, and alkynyl groups, with R′ mainly *t*Bu and *i*Pr. The team felt that improvements in profile might be possible, but that an activity barrier had been reached with compounds of about the activity of practolol. Data had been obtained in an elegant dog preparation (vagotomized, anesthetized, iv) that allowed simultaneous determination of activity, depressivity, and selectivity. Initially, only peripheral blood pressure was determined, but later tracheal overpressure could also be determined. The dose response relationship of an "ideal" cardioselective, nondepressive beta-blocker is shown in Fig. 2. After some discussion with the pharmacologist as to the most appropriate response levels, the definitions shown in Table I were agreed upon. For example, 20% depression was selected because it was usually achieved and felt to be significant; 50% depression of cardiac function was usually

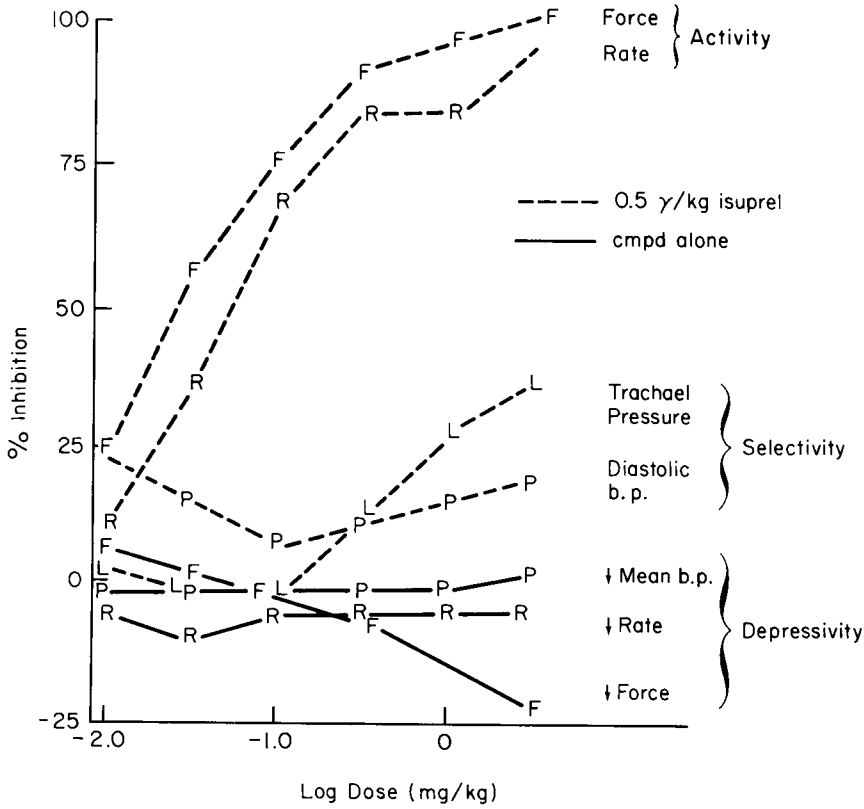


Fig. 2. Dose–response relationship for a cardioselective, nondepressive beta-blocker (IX) in the anesthetized, vagotomized dog, iv. Terms are defined in Table I.

TABLE I
DEFINITIONS OF BETA-BLOCKING PARAMETERS[a]

Parameter	Definition	Explanation
Activity	$pED_{50} = \log(1/ED_{50})$	Dose causing 50% block of isoproterenol-induced chronotropic (rate) or inotropic (right ventricular force) response
Depressivity	$-pED_{-20} = \log ED_{-20}$	Dose causing 20% drop in rate, force, or mean blood pressure; effect of compound alone
Selectivity	$-pED_{20} = \log ED_{20}$	Dose causing 20% decrease in isoproterenol-induced hypotension (diastolic blood pressure); dose causing 20% block of isoproterenol protection against histamine (dog) or methacholine (rat) increases in pulmonary overpressure (dog) or dynamic compliance (rat)

[a] All ED values based on cumulative molar doses. Anesthetized, vagotomized dog, iv; anesthetized rat in plethysmograph, iv. Dog data from A. Strosberg; rat data from R. Weissberg (*79*).

not achieved, or the animal was near death at the time. Statistical programs were written to supply exact values and confidence limits based on cumulative doses, and the bulk of the historical data was converted to ED values. Subsequently, these values were used in reports to chemists (both mg/kg and molar values), although for simplicity, only force activity, depressivity, and selectivity were actually given. The quantification of this data thereby allowed QSAR work to proceed and gave everyone concerned a better feeling for the total profile of our compounds and their intercomparisons with standards.

B. A GLOBAL PICTURE

Without a model, biological data become a mere jumble of numbers, each unrelated to the next. Each analog can be tested in a number of different ways and a compound can be selected for further development, but one does not know the answer to two critically important questions:

1. *Are the bioassays unique or are some redundant?* If the assays are redundant, there is much less information present than one is led to believe. Good sense dictates that drugs should be subjected to a battery of tests before being developed, but these tests must be independent measures of effects in order to contribute true informa-

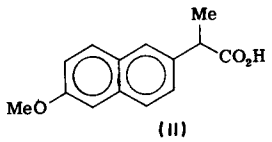
tion. Whereas it is easy to detect simple correlation between two bioassays, it is next to impossible to detect multiple correlation without statistical methodology, e.g., factor analysis (*7ab, 8, 9*).

2. *Has the optimum compound been made?* There is no way to know if optimum activity has been achieved, without relying on some model. While it is difficult to optimize one activity, it is virtually impossible to undertake multiple optimizations without the help of a quantitative, statistical model.

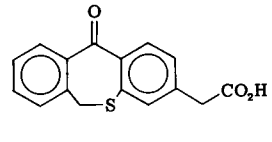
Therefore, QSAR is of great help to the biologist in understanding his data in a systematic, global manner. His training is eminently suited to optimizing test parameters, but it has tended to neglect techniques that would help in examining entire series of compounds as probes of his test systems. Of course, this need not necessarily be the case; Janssen (*8, 10ab*) has applied a more statistical, global approach since the 1960s.

Example: Nonhormonal Anti-Inflammatories

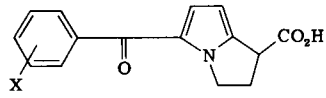
Syntex has had a long involvement in the area of nonhormonal anti-inflammatories [NHAs; e.g., naproxen (**II**) and tiopinac (**III**)] and a recent series of interest (*11a*) has been the pyrrolo[1,2-*a*]-pyrrole compounds (**IV**). A number of bioassays can now be performed in the evaluation of



(II)



(III)



(IV)

NHAs and as part of our study of **IV** we noted the similarity in QSAR between the mouse phenylquinone writhing (QW) and rat carrageenan paw (CP) assays (putative measures of analgetic vs. anti-inflammatory activity). This was confirmed directly by the following correlation:

$$A_{QW} = 1.3(\pm 0.2) + 0.6(\pm 0.2)\, A_{CP} \qquad (1)$$

$$n = 27 \quad s = 0.27 \quad r = 0.808 \quad F_{1,25} = 47.1$$

A = log (activity relative to standard, mg/kg), aspirin was the standard in QW and phenylbutazone was standard in CP, n = number of compounds,

s = standard deviation of regression, r = correlation coefficient and F is the overall F-test for significance of regression. Values in parentheses are 95% confidence limits on the regression coefficients.

This is a highly statistically significant equation. It means that about 65% of the variation in QW can be explained by the variation in CP. (Whether this means that 65% of the activity measured by QW is *due to* anti-inflammatory activity is certainly not established by this regression!) We next checked a series of analogs of tiopinac (**III**) (*11b*) and obtained:

$$A_{QW} = 0.2(\pm 0.1) + 0.5(\pm 0.2)\, A_{CP} \qquad (2)$$

$$n = 17 \quad s = 0.26 \quad r = 0.860 \quad F_{1,15} = 42.42$$

Note that the intercept is much smaller than for **IV**, but that the slope is the same. On the other hand, a noncarboxylic acid series, the 2-amino-7-oxa-3-thia-1-azaspiro[5.5]undec-1-enes (*12*), did not show a significant correlation between yeast inflamed paw and QW.

Because of the apparent redundency of these two assays we next looked at three additional bioassays. A squared correlation matrix is given in Table II for nine analogs of **IV** in the following assays: CP (oral), QW

TABLE II
SQUARED CORRELATION MATRIX FOR A SET OF NHA ASSAYS ON NINE ANALOGS OF **IV**[a]

	CP	QW	HPC	BSV	COL
CP	1.0	0.67	0.51	0.33	0.06
QW		1.0	0.83	0.56	0.02
HPC			1.0	0.49	0.07
BSV				1.0	0.01
COL					1.0

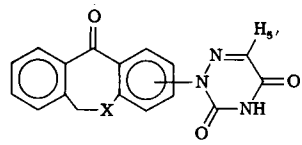
[a] See text for abbreviations. CP and QW data from W. Rooks, HPC and BSV from D. V. K. Murthy, and COL from J. Bruno.

(oral), inhibition of human platelet cyclooxygenase (*in vitro*, HPC), inhibition of bull seminal vesicle PGE_2 (*in vitro*, BSV), and inhibition of collagen-induced platelet aggregation (*in vitro*, COL). A principal component analysis of this correlation matrix gave the following eigenvalues: 3.25, 1.16, 0.427, 0.0891, and 0.0781. Thus, there are only *two* underlying factors which account for 88% of the information in the data set. A factor analysis [using BMDP4M (*13*)] shows that the two factors can be labeled "PG inhibition" and "platelet aggregation inhibition." In other words, it is possible that two specific types of agents may be possible within this

structural class. However, it is important to realize that factor analysis detects the *main* effects of the data set. The correlation between QW and CP is 65% and not 100%, which means that there are compounds that are outliers, much more active (or inactive) than expected from the behavior of the entire class. These may be identified and one may try either to exploit such leads or to discover the reasons behind their "extra" activity.

C. MODELS VS ERROR

The drug designer supplies the mathematical tools for examining data and optimizing activity. When data are of high quality (equiresponse data such as ED_{50}), regression analysis may be used (*14*). If, because of the limitations mentioned above, only percent-response at fixed dose is available, then discriminant analysis (*15*) can be used. However, the quality of the data determines the quality of the information that one can get out of the data. A collection of data is really a collection of information and the "noise" in this information package is composed of the inherent experimental error of the test system and the lack of fit error to the "true" QSAR. The latter arises from such factors as poor solubility, impurities, and improper sample preparation for a *particular* analog, while the former arises from the usual sources of biological variation over the entire series. For example, in a study of some tricyclic, aryl-substituted anticoccidial azauracils (*16*) (V; X = CH, O, S), the QSAR was left at a relatively



(V)

simple observational stage because of the difficulty of obtaining good data for all analogs in a chemically difficult series.

D. BIOLOGICAL RESPONSE AND LIPOPHILICITY

Another useful function involved in the design of biological experiments is an understanding the relationship between BR and lipophilicity. If the BR is a simple function of lipophilicity, it is possible that the proper time parameters for observation have not been found. Dearden and Townsend (*17*) have shown that the concentration of drug in a bioreceptor (implies BR) can have considerable variation with time in an analog series. If bioassay is performed too soon or too late, stereoelectronic factors may

not have an opportunity to express themselves. The result would be a false "optimum" compound.

E. Who Is Man's Best Friend?

A final study illustrates the usefulness of cluster analysis in the area of metabolism. Table III gives the metabolic profile (see Section III,F) of

TABLE III
METABOLIC PROFILE OF AMPHETAMINE IN FIVE MAMMALS[a]

Species	Rate excretion in urine (%/24 hr)	Excretion product[b] (% dose)				
		I	II	III	IV	V
Rat	84	13	60	0	0	3
Rabbit	81	4	7	22	8	27
Man	66	30	3	3	0	20
Dog	89	38	7	3	2	32
Horse	84	9	8	29	12	29

[a] From Parke (*18*).
[b] I = amphetamine, II = *p*-OH-amphetamine, III = benzylmethylketone, IV = phenylpropanol, V = benzoic acid.

amphetamine in various mammals (*18*). Which mammal tested most closely resembles man and therefore would be most suitable for extensive metabolic studies? A simple hierarchical cluster analysis using CLU (Section VI) gives the dendogram (*19, 20*) shown in Fig. 3. This represents the course of the clustering procedure. The two closest profiles are those of the rabbit and the horse; the average of these profiles replaces the individuals and all distances are redetermined. Next, man and dog are clustered; rabbit–horse and man–dog are then merged and rat is seen as an outlier in this case. Therefore, dog *is* man's best friend for this particular collection of data. This case is meant to be illustrative and not conclusive; the distances at which clustering occur are not close in this analysis, but we seek *nearest* neighbors. Clearly many uses can be envisioned for such cluster analyses.

III. Search for New Leads: The Impossible Dream

Many people associate QSAR with analog development or "lead optimization" and therefore assume that it has little to offer in the realm of "lead generation." This is a serious misconception.

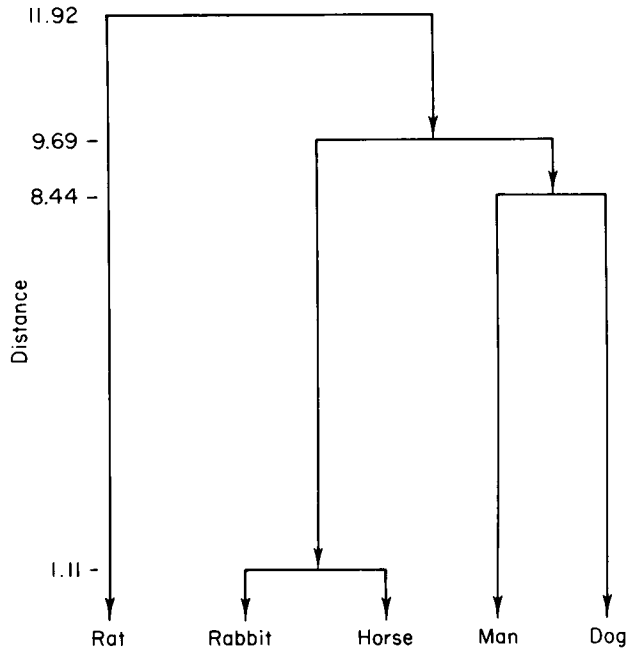


Fig. 3. Hierarchical cluster analysis of comparative metabolic data on amphetamine (see Table III).

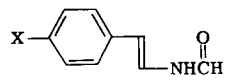
A. PHYSICOCHEMICAL VS GRAPHICAL APPROACH

Lead optimization can sometimes be lead generation, in the sense that unusual substituents might be suggested because of physicochemical reasons (*20*) that might not have occurred to the chemist and that more likely than not represent a unique system to the patent attorneys. The graphical system of chemical notation is a different paradigm than the physicochemical paradigm and therefore frequently leads to different types of suggestions.

Example: Tuberins

A series of tuberins (**VI**) (*21*) active against *Mycobacterium phlei* gave the following QSAR [Eq. (3)]:

$$\begin{aligned} pA = {} & -0.3(\pm 1.4) + 2.9(\pm 1.1)\log P - 0.5(\pm 0.2)(\log P)^2 \\ & -2.6(\pm 0.9)\mathcal{F}^2 - 1.7(\pm 0.5)\mathcal{R}^2 + 0.4(\pm 0.3)D_{ha} \\ & n = 19 \quad s = 0.161 \quad r = 0.952 \quad F_{5,13} = 25 \end{aligned} \tag{3}$$



(VI)

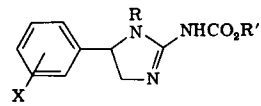
$pA = \log 1/MIC_{molar}$, $\log P$ was determined by our HPLC procedure (*22*; Section VI,A), $\mathcal{F}$ and $\mathcal{R}$ are the corrected Swain and Lupton electronic component constants (*4a*) and $D_{ha} = 1$ if the substituent X is a hydrogen-bond acceptor. No linear terms in $\mathcal{F}$ or $\mathcal{R}$ could be added, and therefore it was concluded that the aromatic moiety may not be critical, as long as the net electronic effect was similar to that of a phenyl. Indeed, replacing the phenyl by linear or cyclic alkyl groups of the proper lipophilicity (log $P_0 = 2.87$) maintained high activity. Both phenyl and alkyl groups are slightly electron donating by $\mathcal{R}$, but differ in $\mathcal{F}$.

B. Importance of Outliers

Without a mathematical model, there is no way of telling if a given analog is an outlier from the rest of the members of its class. One is often most interested in identifying outliers because they may represent the presence of a special effect that can provide large boosts in activity. Compounds that do not fit the model are usually more interesting than those that do. They require more thought and can eventually be incorporated into a more global model.

Example: Phenyl Substituted Imidazoline Alkyl Carbamates

The phenyl substituted imidazoline alkyl carbamates (**VII**) (*23*) show an



(VII)

interesting spectrum of activities. Despite a number of difficulties with the biodata (large biovariation and small range of activity within the series) and the limits on the initial range of substituents which had to be worked on, we derived QSAR for antidepressant (reversal of reserpine hypothermia in mice, po), anticonvulsant (antagonism of maximal electroshock in mice, ip), muscle relaxant (antagonism of etonitizene induced rigidity in mice, ip), and depressant (loss of righting reflex in mice, ip) activities. The general pattern of QSAR is shown in Fig. 4 as a function of lipophilicity. Hydrophilic, electron-withdrawing substituents were suggested in order to enhance antidepressant activity while maintaining rea-

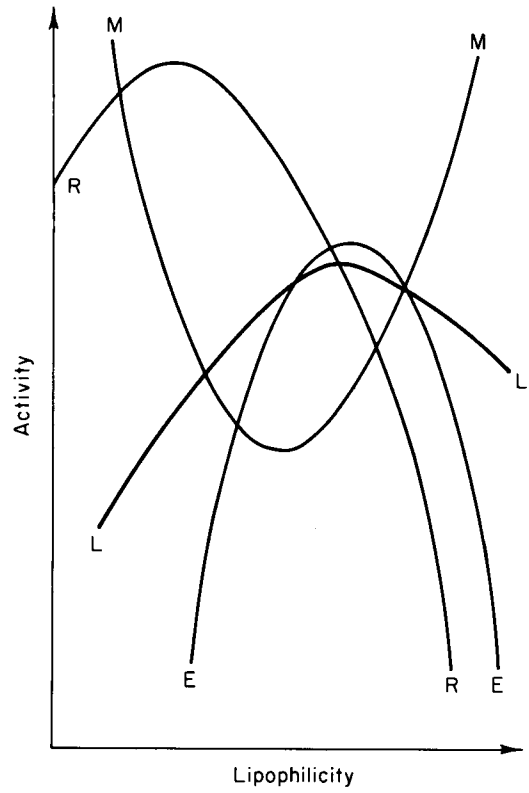


Fig. 4. Representation of the QSAR for four CNS activities in the series (**VII**). QSAR were individually developed for R = reserpine reversal (anti-depressant); E = maximal electroshock (anti-convulsant); L = loss of righting (general depressant); M = etonitizene rigidity antagonism (muscle relaxant). Average values for nonlipophilic terms were used in plotting graphs. Minimum in activity for M could arise, for example, by a competing process which removes compound in an optimal manner with increasing lipophilicity. R, E and M from M. Wallach; L from A. P. Roszkowski.

sonable depressant effects; however, the bioassay showed such substituents to have very low activity. Was this a failure? No, because further investigation showed that these hydrophilic, electron-withdrawing substituents were also the first in the series that were also hydrogen-bonding, and we were finally able to obtain a good fit of these outliers by including an indicator variable for hydrogen-bonding effects. Polar substituents could not be tolerated; this was reverse lead generation, since we could eliminate large classes of substituents from consideration. We were also able to determine that these substituents were not simply *too* hydrophilic. If this were the case, they would not really be outliers with respect to the

rest of the class. Another important result of the QSAR study was the realization of the trade-off in activity between antidepressant and other activities.

C. Bioisosterism

Lead generation can also stem from the redefined concept of bioisosterism (*20*). The main difficulty here is that in QSAR one is always aware of the importance of understanding the *relevant* factors determining activity (*20, 24*). Two substituents are (isometrically) bioisosteric if they produce the same bioresponse because they are essentially identical with respect to the relevant physicochemical parameters. This implies that the receptor does not experience any difference between the two substituents. While they may differ in some respects (no two substituents are identical in all respects), this difference is not relevant to the particular receptor. A second type of bioisosterism (nonisometric) results from the compensation effect. If $BR = aX + bY + \ldots$, then if $X_1 = X_2$, $Y_1 = Y_2$ etc., then $BR_1 = BR_2$ (isometric); however, X and Y may also compensate in such a manner that $BR_1 = BR_2$ even if $X_1 \neq X_2$. Tables of bioisosteric substituents are given in Ref. *20* and provided in updated form through the Pomona College Medicinal Chemistry Project (*4b*) for various combinations of parameters. These lists provide many leads for (isometrically) bioisosteric substitution.

D. "QSAR Insight"

There is an area of lead generation that can be referred to as "QSAR insight." A person trained in a particular speciality tends to see the world in a certain way. A drug designer sees molecules as a framework for charge, lipophilicity, and bulk, and if these can be duplicated in essential features using different atoms, equal responses should result. The major difficulty, of course, is in identifying the relevant features. There are many examples in the biochemical literature in which a change in BR due to substituents will be assigned to a particular factor, with complete disregard for the remaining changes which such a substitution must bring about. If nothing else, one must be especially careful to establish the underlying effect of changes in lipophilicity.

Example: QSAR of Enzyme Inhibitors

One example of QSAR insight is that provided in the long series of papers by the Pomona group on the quantitative correlation of B. R. Baker's (*25*) extensive investigations into the design of active-site-directed

irreversible enzyme inhibitors (*26*). Baker provided useful *qualitative* models of the factors which could be involved in receptor site binding, and he exploited the concept of an exoreceptor (peripheral region adjacent to active site) as an appropriate target for the introduction of specificity into enzyme inhibitors. However, the Pomona group provided relatively simple *quantitative* models which could explain the activity of up to hundreds of compounds in a single, highly significant equation. The most significant general findings of these QSAR studies are that:

1. Complicated substituents may in fact have a very simple function (bridging) and can therefore be handled by indicator variables (*26b,c,d*).
2. Lipophilicity and molar refractivity differ, one being a measure of loss of solvation and the other a polarizability (or bulk, at times) phenomenon (*26a,d*).
3. Stereoisomers may be handled by QSAR techniques (*26a*).
4. Receptors may be "mapped" by studying substitutions in many positions of the substrate (*26b,d*).

Such enzyme studies can be the starting point for directed synthesis programs (*26e*).

Example: Thiazole Beta-Blockers

Returning to the thiazole beta-blocker series, **I** provides another example of QSAR insight. Levy (*27*) had shown that cardiac depressivity (the ability of the compound alone to reduce rate or force of contraction of the heart) was a nonspecific effect unrelated to beta-blocking activity, chloroform–water partition coefficient, or pK_a. Whereas activity in rat and dog was uncorrelated (Fig. 5), depressivity was linear between the two species (Fig. 6). This tended to confirm the notion that depressivity was a nonspecific effect, unrelated to beta-blocking activity. We therefore reasoned that if it were a membrane-stabilizing property [often called local anesthetic activity, (*27*)], a bulky substituent should sit on or fit into the membrane much less efficiently than an isolipophilic open-chain compound. This turned out to be the case, as shown in Table IV.

From a study of the adenylate cyclase inhibiting activity (Fig. 7), we developed the idea that the lung and heart β-receptors differed in that the lung possessed an exoreceptor cleft. Whereas the heart activity appeared to be smoothly increasing with lipophilicity, the lung activity actually decreased. Therefore, we postulated that "shorter" substituents of higher lipophilicity, such as adamantylethyl, might be very potent and highly selective, since most of the bulk would fall into the exoreceptor cleft at the lung β_2-receptor. Bulky substituents would also be less depressive, as

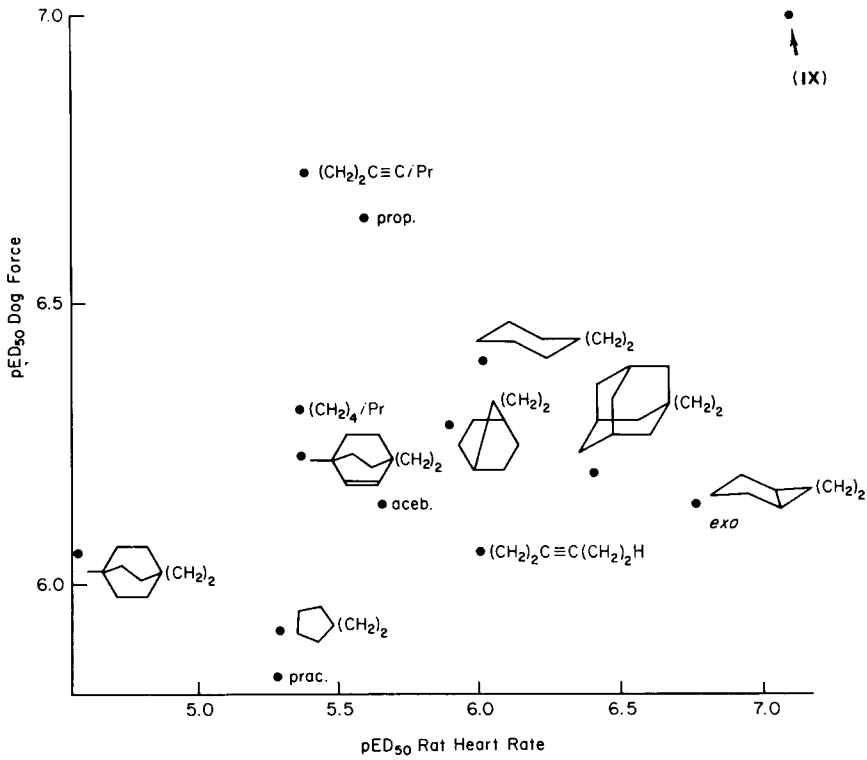


Fig. 5. Beta-blocking activity of I (R indicated) in dog and rat models.

postulated. Although the initial concept proved correct in all essential details, it was further complicated by the discovery of additional factors which contributed significantly to activity (see Section IV,E). We were able to derive QSAR which unified the profiles of over 40 analogs, to handle these factors. Therefore, lead generation stemmed directly from attempts at lead optimization, since the newer substituents were novel structures with respect to the types of alkyl groups previously synthesized.

E. Receptor Fit

The concept of designing a drug specifically to fit into a receptor is becoming practical as more enzyme crystal structures are determined (*26e*). The Wellcome group (*28a,b*) designed inhibitors of hemoglobin based on "drug design by receptor fit," using knowledge of the conforma-

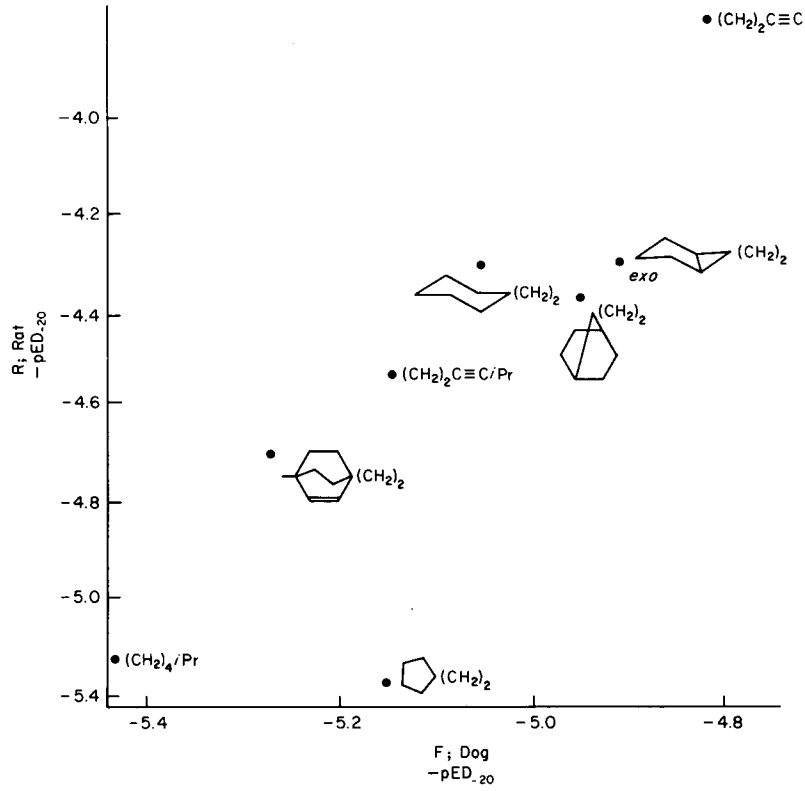


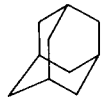
Fig. 6. Depressivity of beta-blocker I (R indicated) in dog and rat models.

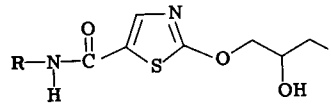
tional transition which hemoglobin experiences when binding to oxygen. The Squibb group (*28c*) successfully used the putative analogy in structure between angiotensin-converting enzyme (target) and carboxypeptidase A (known) to design SQ14,225 (captopril), an orally active peptide with effective antihypertensive activity. The book by Martin (*3b*) outlines other methods of designing leads based on enzymes.

F. MULTIDIMENSIONAL VIEW OF THE WORLD

An important source of new leads comes from a multidimensional view of data. An ordered collection of numbers is a vector; therefore, if we arrange a matrix of data for a collection of compounds (rows) over a number of different bioassays (columns), each column can be plotted in a multidimensional vector space (each compound being an orthogonal axis)

TABLE IV
DEPRESSIVITY AS A FUNCTION OF BULK OF ALKYL CARBAMOYL GROUP[a]

R	Log k'	$-pED^{E}_{-20}$	
$(CH_2)_9H$	1.31	− 6.07	
 $(CH_2)_2$	1.26	− 5.06	10.2×
$(CH_2)_7H$	0.86	− 5.77	
$(CH_2)_2$ 	0.81	− 4.95	6.6×

[a] Data is for  NHR', R' = *t*Bu.

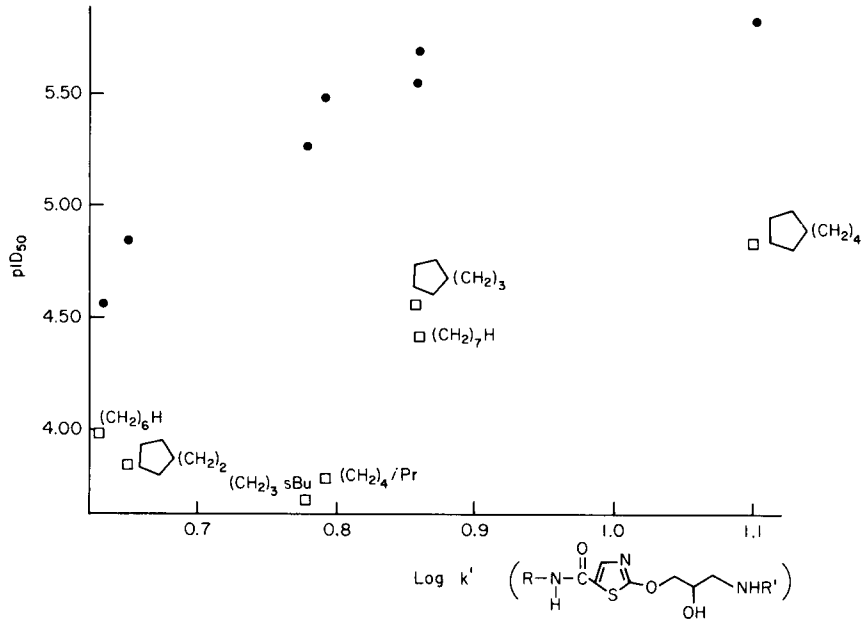


Fig. 7a. Adenylate cyclase (Guinea pig: •heart □lung). Inhibition as function of lipophilicity. Historical R' = *t*Bu. Data from R. Alvarez.

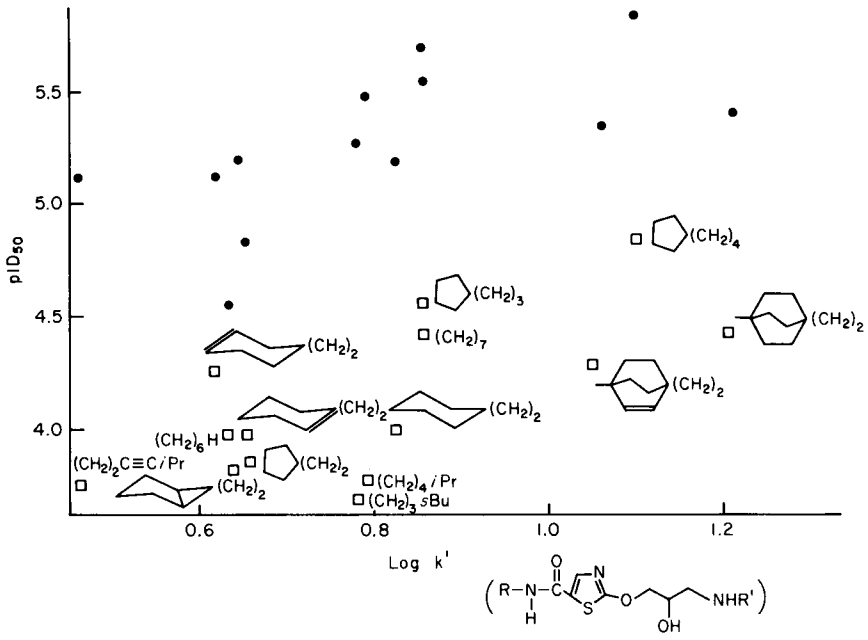


Fig. 7b. Adenylate cyclase (guinea pig •heart □lung). Inhibition as a function of lipophilicity. Current R' = *t*Bu. Data from R. Alvarez.

or each row can be plotted with each assay being an orthogonal axis. Thus, a point in this space can represent either a particular bioassay or a particular compound (see Fig. 8). Such plots are a conceptional starting point for various cluster or factor analytical methods (*3a,b, 8a,b, 10, 20, 29a,b, 38a*).

We have already discussed a factor analytical study on the pyrrolopyrrole series (**IV**), a cluster analysis on metabolic data (Fig. 3) in order to determine the laboratory animal most similar to man, and have referred to a cluster analysis (*20*) on substituent constants in the area of bioisosteric modifications. Chen, Horváth and Bertino (*38b*) discuss an integrated approach using several multivariate techniques.

Originally used in the classification of people with respect to their scores on a battery of psychological tests and more recently with EEG-evoked potentials (*30*) or in the area of numerical taxonomy (*19*), the concept of *profile* is also useful in drug design. Compounds can be compared over several bioassays (Fig. 8) by calculating the distance [Eq. (4)] between each compound in the bioassay space:

$$d = [\Sigma(X_{ij} - X_{ik})^2]^{\frac{1}{2}} \tag{4}$$

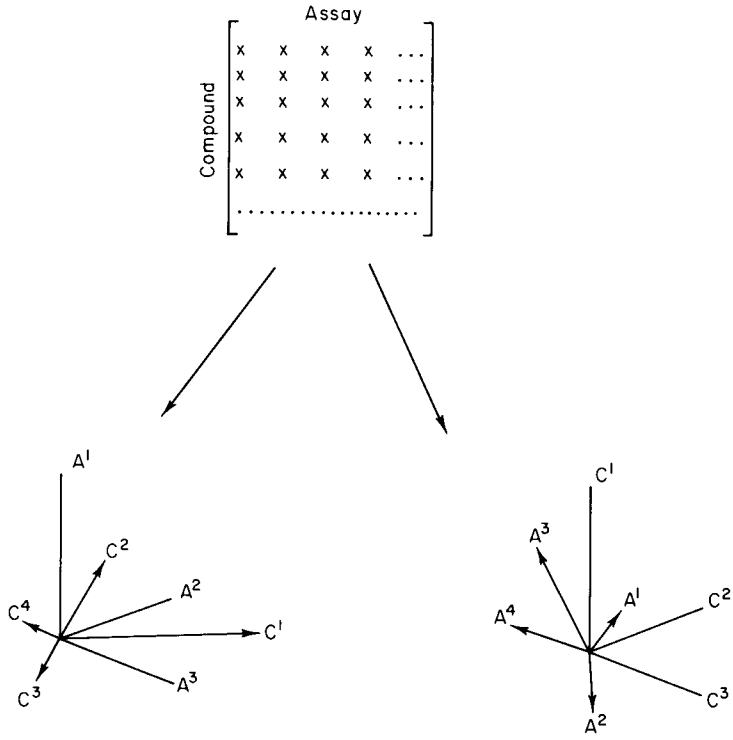


Fig. 8. Graphical representation of a data matrix. If assays are plotted on the axes, then each compound is a separate point and *vice versa* if compounds are the axes. In this example, only three axes are pictured.

This distance can be used in one of two ways, both of which assume that "you know a person by the company he keeps" and that "the whole can be represented by something less than the sum of its parts." (That is, the closer two objects with respect to the bioassays examined, the closer in general with respect to all bioassays; and that a small number of critical, relevant factors can characterize the whole.) The first method is the k-nearest-neighbor (kNN) approach which finds d between an unknown and all members of the data set; the k (usually 3) nearest neighbors are then examined. The second method is to use a hierarchical clustering algorithm, such as CLU (Section VI), BMDP2M (*13*), CLUSANL (*20*), etc. (*19*), and to determine into which cluster the unknown(s) fall. New leads are then either compounds which share a similar profile to known reference compounds or which fall into "active" regions (see Section III,F).

Example: Mouse Behavior Screen

The mouse behavior and symptomalogy screen (Table V) is a general screening procedure for CNS compounds (cf. *8, 10*). Groups of three male mice are dosed ip with one of about six increasing doses of compound and the dose at which each of a battery of 36 standardized behavior tests

TABLE V
MOUSE BEHAVIOR AND SYMPTOMATOLOGY SCREEN[a]

1.	No. acute deaths (test period)
2.	No. delayed deaths (24 hours)
3.	Writhing response
4.	Locomotion, increase
5.	Induced activity, increase
6.	Locomotion, decrease
7.	Induced activity, decrease
8.	Pinna reflex
9.	Cornea reflex
10.	Passivity
11.	Analgesia
12.	Abnormal posture or gait
13.	Grip strength, increase (wire mesh)
14.	Grip strength, decrease (wire mesh)
15.	Orientation (inverted glass tube)
16.	Ataxia (tilt board)
17.	Righting reflex
18.	Muscle tone, increase
19.	Muscle tone, decrease
20.	Bizarre reaction
21.	Straub tail
22.	Tremor
23.	Twitching
24.	Convulsions
25.	Rectal temperature, increase
26.	Rectal temperature, decrease
27.	Pupillary size, increase
28.	Pupillary size, decrease
29.	Ptosis
30.	Exophthalmia
31.	Lacrimation
32.	Salivation
33.	Piloerection
34.	Defecation
35.	Urination
36.	Skin color

[a] A. P. Roszkowski and M. Schuler.

either occurs or fails to occur in $\frac{2}{3}$ of the mice is recorded. The resulting profile of dose codes (1 = 0, 2 = 1, 3 = 3, 4 = 10, 5 = 30, 6 = 100, 7 = 300, 8 = 1000 mg/kg) for 176 standard drug assays (120 different compounds) can then be compared with each other (by cluster and factor analysis) and unknowns can be compared to standards by kNN, discriminant analysis, or factor analysis techniques. The goal of this procedure is a more sensitive classification than that provided by visual inspection of the data sheet using certain empirical rules. The procedure is complementary and one can assign weights [i.e., d or a similarity $s = 1 - (d/\max d)$] as an additional aid in classifying unknowns.

Since the doses are already standardized on a uniform scale (from 1 to 8), we performed a CLUSANL (*20*) clustering on this raw data directly. The standard compounds were augmented by a small number of Syntex compounds, which did not have a material effect on the clustering (see Table VI). The major clusters at 16 and 45 clusters were selected by plotting the "information loss" (i.e., essentially distance) vs. cluster numbers and finding the breakpoints. Repeat assays are given only if they fall in different major clusters. The hierarchical procedure (cf. Fig. 3) represents a continuous loss of information as more and more diverse clusters are merged. Therefore, the 16-level clusters represent rather broad regions of the data space and show the amalgamation of general activity types. True nearest neighbors are shown at the 45-level clusters. A number of points should be made:

1. Biological variation is relatively wide as seen in the repeat assays.
2. Clustering is on the basis of the information provided and not additional information from supplementary assays.
3. The main purpose of this procedure is to be suggestive of further testing; it is meant to be supplemental in that compounds not otherwise tested in secondary assays might be examined further.
4. Many compounds show slight-to-strong depressant effects and the procedure is not well-suited to sorting these out, although there are some interesting subclusters for muscle relaxants (e.g., cluster VI).

Despite these cautions, there are a number of suggestive observations that can be made. While there is no direct measure of antidepressant activity in the protocol, the tricyclic antidepressants tend to cluster together (clusters V,VIII), as do the major tranquillizers (clusters III,IV,IX). We have examined the antidepressant data more carefully. When desipramine is given as an unknown in a kNN procedure imipramine $s = 0.87$ and $s = 0.83$ are two nearest neighbors (others are atropine $s = 0.84$, procaineamide $s = 0.84$, pentylenetetrazole $s = 0.83$, glutethimide $s = 0.82$ and cocaine $s = 0.82$); amitryptyline gives imip-

TABLE VI
HIERARCHICAL CLUSTERING ON MOUSE SCREEN DATA:
SIXTEEN AND FORTY-FIVE CLUSTER LEVELS

I	[aspirin,levodopa,β-hydroxybutryic acid,cocaine,azacyclonol,nalorphine]
	[5-HT]
	[acetylcholine,pentylenetetrazole,SCH 12679]
	[N-aminorhodanine,propoxyphene,procaine,dibenamine,pargyline, naloxone,ethyl alcohol,2-methylglutamic acid,pentazocine]
	[meprobamate,trimethadione,dibutyryladenosine,lithium carbonate, nialamide,chlorphentermine,menadione sodium bisulfite,pyrithione sodium]
	[phenmetrazine,morphine]
II	[aspirin,pemoline magnesium,amphenazole, 3,5-cAMP,aminophylline, caffeine,tetramisole]
	[imipramine,dimenhydrinate,cyclizine,ephedrine]
	[phenobarbital,aminoglutethimide,phensuximide,α-d-glucochloralose, methylphenidate,chlorpheniramine,viloxazine,codeine,ethylmorphine]
III	[chlordiazepoxide,methaqualone,zoxazolamine,pimozide,ethinamate, mephenytoin,phenacetylurea,lidocaine,methohexital,hexobarbital, promazine,trifluoperazine]
	[phenobarbital,pentobarbital sodium,glutethimide,phenaglycodol]
	[fluanisone,propranolol,burimamide,BAY 7272]
IV	[thioridazine,clozapine,prochlorperazine]
	[chlorimipramine,diphenidol]
	[procaine,propranolol,dextromethorphan,mebutamate,quinidine sulfate,methysergide,suramine]
	[PGE_2]
V	[desipramine,promethazine,pentylenetetrazole,pilocarpine, amitryptyline,imipramine,nortryptyline,atropine,fenfluramine]
	[scopolamine]
VI	[perinaphthenone,meprobamate,mephenesin,carisoprodol,gluethimide, metaxalone,oxazepam,phenobarbital,methocarbamol,chlorzoxazone]
VII	[haloperidol,perphenazine]
	[chlorpromazine,xylazine]
	[barbital sodium,chlordiazepoxide,diazepam]
	[gallamine triethiodide]
VIII	[amitryptyline,orphenadrin,pentobarbital sodium,doxepin]
	[morphine]
IX	[chlorpromazine,mesoridazine]
	[trifluoperazine,PGE_1]
X	[dextroamphetamine,tetrahydronaphthylamine]
	[tranylcypromine]
XI	[strychnine]
	[oubain,decamethonium]
XII	[oxotremorine,neostigmine]
XIII	[reserpine,tetrabenazine]
XIV	[strychnine (4 assays)]
XV	[oxotremorine (3 assays)]
XVI	[reserpine]

ramine $s = 0.93$ and nortryptyline $s = 0.91$ (others are lidocaine $s = 0.86$, chlordiazepoxide $s = 0.87$, ethinamate $s = 0.87$, and phenaglycodol $s = 0.86$). Most interesting is that the nontricyclic antidepressant, viloxazine, gives imipramine $s = 0.83$ as a nearest neighbor (however, chlorpheniramine $s = 0.89$, codeine $s = 0.88$, ethylmorphine $s = 0.87$, amphenazole $s = 0.84$, aminoglutethimide $s = 0.83$, BAY 7272 $s = 0.83$, and pentylenetetrazole $s = 0.83$ are also given).

To understand the data set better, a factor analysis was performed on the bioassays. Out of 36 assays, the following 11 eigenvalues were obtained: 6.156, 4.459, 2.534, 1.950, 1.769, 1.740, 1.337, 1.293, 1.213, 1.104, and 1.035. These 11 eigenvalues account for 68.3% of the total information in the data set (176 assays with 36 observations). There is a high degree of redundency in the 36 assays. An orthogonal rotation was performed and factor scores recomputed from the rotated factor loadings and standard scores [BMDP4M (*13*)]. One method of identifying the nature of the principal factors is to examine the sorted rotated factor loadings to determine which assays are the dominant components of the factor. However, in this case we obtained factor score averages for each of five biological classes, as shown in Table VII. It can be seen that antidepressants possess a unique pattern over the 11 eigenvalues. Therefore, it is not surprising that one can classify antidepressants using a general mouse behavior and symptomatology screen which does not contain any direct measure of antidepressant action. This novel approach to classifying compounds will be pursued further and reported more fully at a later time.

G. "FALLACY OF PROBABILITIES"?

Cramer, Redl, and Berkoff (*31 a,b*) proposed a computerized screening method based on an analysis of the SmithKline substructural code. This system has not been pursued (*31c*); however, Hodes (*32*) has derived a better method of calculating statistical weighting factors and has applied his procedure to anticancer drugs. Essentially, one calculates a weight for each substructure based on the experience table for actives and inactives, a total of some 280,000 compounds. If there are n active compounds and a substructure whose incidence in the total set is p, then the expected incidence of the feature in the active set is np, even if the feature has nothing to do with activity. The weighting factor is determined by finding "how far the actual number of actives with the feature differs from its expected value" and using the probability P_i. The score for a compound is then $\log(1/\text{II}\, P_i)$. While this method is an improvement over Cramer's procedure, both suffer from what might be called the "fallacy of probabilities," that by gluing together fragments

TABLE VII

FACTOR SCORE AVERAGES FOR ELEVEN FACTORS BY BROAD BIOLOGICAL ACTIVITY TYPE[a]

Class[b]	No./No. cmpd[c]	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Antidepressants	13/7	0.6	−0.04	0.04	−0.7	0.4	−1.	−0.3	−0.3	0.00	0.00	0.2
Stimulants	21/11	−0.6	0.00	0.7	−0.2	0.2	0.4	−0.7	−0.2	−0.7	0.3	0.00
Sedatives	13/6	0.5	−0.4	−1.2	0.2	0.3	0.00	−0.3	0.2	0.3	0.00	0.2
Minor tranq.	12/4	0.3	−0.7	−0.6	0.1	1.	0.3	0.5	0.6	0.4	−0.2	0.2
Major tranq.	19/12	0.3	−1.0	0.1	−0.3	−0.1	−0.8	0.4	−0.3	0.1	0.1	0.2

[a] Averages differing significantly from zero (t-test) are underlined (≧ 95% C.L.).

[b] *Antidepressants:* desipramine, amitryptyline, nortryptyline, imipramine, doxepin, chlorimipramine, viloxazine. *Stimulants:* amphetamine, phenmetrazine, amphenazole, methylphenidate, pemoline, caffeine, fenfluramine, pentylenetetrazole, strychnine, tranylcypromine, nialamide. *Sedatives:* hexobarbital, methaqualone, glutethimide, barbital sodium, phenobarbital, pentobarbital. *Minor tranquillizers:* oxazepam, meprobamate, diazepam, chlordiazepoxide. *Major tranquillizers:* thiroidazine, chlorpromazine, promazine, trifluoperazine, prochlorperazine, chlorprothixene, mesoridazine, clozapine, pimozide, haloperidol, fluanisone, azacyclonol.

[c] Number assays/number compounds.

with high probabilities an active moiety will result. This obviously neglects global properties of molecules, such as lipophilicity, and any interactions between pharmacophores.

We attempted to correct for these effects in a study directed toward the identification of leads in an antihypertensive program. Hodes' procedure could not be used directly because we had no meaningful inactive class. From the large literature on antihypertensives, we culled a list of 40 diverse active structures. We coded the 40 actives for 14 features we thought might be important; each feature was redundantly coded by giving one point for each occurrence (F). For each compound, the frequency of each feature in each molecule was determined (if there were 4 features, each occurring once, then each was given a value of 0.25). Finally, the "importance" (I) was calculated as the average of the frequencies in the 40 actives. The absence of groups was not taken into account. The importance values were then used to calculate a score [$\Sigma I \cdot F$] for the candidates. Whereas a reasonable percentage of the unknowns submitted for screening showed some activity, there is no real way to judge the actual success rate for this experiment since the compounds were already preselected because they "looked like antihypertensives".

H. PATTERN RECOGNITION

Pattern recognition techniques were much discussed (*33a*) in the early 1970s, especially as a cure all for drug design (*33b,c*); however, there were serious drawbacks in the methods used (*24a*). One of the most difficult tasks in any SAR study is to determine the significance of the results (*14, 24*). An obviously trivial result is one in which, for example, the author has overlooked a simple classifying rule such as all active contain sulfur (*24a, 35,*), for his data set (*34a–c*). A trivial result which is nonobvious would be one in which the author has "discovered" that all actives are soluble, yet he is unaware of it because he has not correctly interpreted the model equation. If he were to "discover" this from his equation, and it had not previously been suspected, then the result would probably be considered significant. A second difficulty is in the construction of the training sets. If the descriptors are simply substructures, then it would seem unlikely that a novel active structure would result, because of the fallacy of probabilities mentioned above. A third problem is that while pattern recognition has been successful with single source data, QSAR involves multisource data (*24b*). There is a fundamental difference between experiments in which each feature is simply one out of hundreds of frequencies and experiments in which each feature might be the result of a different generating algorithm, might have a different scale and origin,

might be binary, ordinal or cardinal, and might be highly collinear with other features. In the former experiments, each feature is conceptually equivalent, and therefore the features can be treated with some abandon; the data space can be twisted, stretched, shrunk, or convoluted. In many instances we know there is an answer; we must only find it. This is rarely the case in QSAR work. Not only do we not always know *if* an answer exists (that the classes can be separated by the features used), but we do not always know if we are even asking the right questions (disagreement or ambiguities in defining classes). Thus, construction of the universe of features becomes a major part of the experiment.

Despite these cautions, there have been a number of pattern recognition studies that deserve careful consideration. Jurs (*36a,b,e*) studied the classification of 5,5′-barbiturates based on duration of effect and a classification of psychotropic drugs as sedatives or tranquillizers (*36c,e*), while Chu (*37a–c*) did a number of preliminary studies using various techniques.

One key to obtaining novel structures is to include physicochemical parameters along with structural features in the analysis (*29*). In this way regions of active compounds within the data space are defined by open-ended features that may be combined in useful ways or used to compensate for global properties.

I. OTHER METHODS

Further sources of new leads are to be found in the open and patent literature, by general screening, and by a review of all compounds submitted for bioassay, all perused with "QSAR insight". Finally, there are invaluable discussions with peers, be they chemists or biologists. On a number of occasions, biologists involved in exploratory work have discussed the types of compounds that should be screened in a new bioassay for which no or very few leads exist. In such cases there are two basic approaches. If a few leads exist, one may use the various methods previously discussed to develop physicochemically equivalent leads (e.g., bioisosteric substitutions) or one of the multidimensional lead generation techniques. If no leads exist, one may try to develop objective screening procedures for searching the compound file, perhaps based on some physicochemical criteria such as lipophilicity.

IV. Optimization of Lead: An Interactive Approach

Most people associate the Hansch paradigm with lead optimization. The reader has seen from the previous sections that this is but one of

several types of activities made possible by this approach. A lead may be generated by any of several mechanisms from random screening to receptor fit, but what then? The historical approach has been to make a large number of analogs of a lead—possibly with some notions of the active site—and essentially to stop when a compound good enough for development has been found. A compound is then "selected for development," which means advanced chemistry, pharmacology, toxicology (including mutagenicity, teratogenicity, carcinogenicity, etc., as the case may warrant), pharmacy (dosage design), and consideration of clinical protocols. The decision is essentially one of committing large amounts of capital to a single compound. Therefore, the hope is that QSAR can help to provide a more rational understanding of the factors involved in bioactivity, to predict the most active or most desirable compound, and to make sure "we haven't missed anything." In a sense, QSAR is an insurance policy. However, several caveats are in order:

1. The QSAR model may be incorrect for any of several reasons (*14*): the data may be too poor for an accurate model, the mathematical equation or parameters examined may be incorrect, etc.
2. A statistical error may be made in choosing between alternative models (see e.g., *39a,b*).
3. An error in interpretation of the QSAR may be made.
4. The bioassay may be incorrect.

Although there are some contributions that can be made by the QSAR specialist (see Section II) towards point 4, it is generally outside the purview of his skills to select relevant bioassays. QSAR is based on the data provided, and extrapolation to other experimental conditions and species (especially to man) are no better than any other procedure and are outside the bounds the paradigm. Success for the drug designer is successful analysis of the data presented, although one always hopes that these successes will translate into useful medicinals. Drug design is a process of continual hypothesis making, testing, and reformulating as new information is obtained (*78*). This is the normal course of all science.

Compounds are suggested for two basic reasons: to try to obtain the optimum compound, and/or to improve the statistical validity of the model (leading eventually to the optimum compound, or greater confidence in the predictions). Important information to help in optimizing the QSAR can be obtained from weakly or even inactive compounds.

A. SPANNING SUBSTITUENT SPACE

The place to start in any optimization program is at the very beginning, before any (or too many) analogs have been synthesized. Once a lead compound has been discovered, there are usually two lines of development: certain gross structural changes might be made in order to define further the most active pharmacophoric groups or to provide novelty; and/or more detailed analog studies of the lead might be planned in order to optimize activities or properties. Given sufficient prior knowledge in the QSAR literature, it is possible to help in the former by providing "QSAR insight," bioisosteric modifications, etc., as previously discussed. Eventually most projects will proceed to the latter course, and here a number of schemes have been proposed for rapidly optimizing the series. These are based on the paradigm, the quantitative, statistical model. One of the most well-known is the so-called "Topliss Tree" (*40a–c*), wherein a decision is made as to the next compound to be synthesized according to simple rules, once the report of activity is in hand. The method assumes that the biological data are both readily forthcoming and infinitely accurate and that the QSAR is relatively simple. None of these assumptions is usually true. There is no provision for such special effects as metabolism of particular groups, special differences due to hydrogen-bonding substituents, etc. If these effects were present, one would be misled, since there is no method of detecting such effects (*78*). The compounds would be of lower or higher activity than they should be and the wrong branch of the "tree" would be taken, leading to a false optimum.

The simplex optimization method of Darvas (*41a,b*), as modified by Gilliom, Purcell and Bosin (*42*), avoids some of the difficulty of experimental error, since provision is made for repeating active compounds. However, special effects would also throw the optimization off the track if they were not parameterized. Both of these so-called "rapid optimization" methods suffer from a number of additional problems:

1. Only a very small number of biological assays are, in practice, sufficiently rapid or accurate to allow one to postpone decisions as to synthesis until all previous results are in.
2. Some chemists may be unwilling to follow such a scheme; once they have a "hot" lead they are very anxious to find out its worth as soon as possible. There are frequently time constraints on projects. Starting materials often take weeks to be delivered, which means the next compound would not be synthesized for several weeks after the results are in on the first compound. Biologists often wait for a group of compounds before running an assay.

3. Most of the QSAR selected to illustrate rapid optimization are, perforce, known and simple, while in practice, most QSAR are both unknown (initially) and frequently complex.
4. If the lead turns out to be a good one, there is usually a decision to spend a certain amount of time on the series and eventually tens to hundreds of compounds might be made.
5. Compounds are often tested in many additional random or directed screens. Even if the original lead fails, frequently other activities might be found and different compounds in the series might be optimal in these new bioassays. Diversity is therefore useful.
6. Assuming one were to follow rigorously the Topliss or Darvas procedures, one would often be left with a multicollinear set of derivatives that could not be easily analyzed by the usual multiple regression techniques. Thus, further derivatives would have to be made in order to break the multicollinearity.

For these reasons, we favor a more general approach to planning a series of derivatives that avoids these pitfalls and, in practice, need not take appreciably longer to reach an optimum compound. The key to this approach is to recognize that in order to *fully* exploit and explore a new lead, one must "span substituent space" (*14, 43*); in short, one must make examples of all types of substituents so that a range of electronic, steric, lipophilic, and hydrogen-bonding substituents is included. With such a basis set included, even if the plan is not followed exactly and even if more than the minimum basis set is made, most of the pitfalls are avoided. Once the basis set has been evaluated, one can use a number of different QSAR techniques for optimizing activity (including simplex optimization, using the basis set as an initial set). An attraction of the Topliss and Darvas procedures is that they require neither statistical nor computer expertise. One may use our bioisosteric cluster procedure (*20*) to find substituents similar to the most active compounds (*78*).

Wootton, Cranfield, Sheppey, and Goodford (*44*) have published a modification of an idea by Kennard and Stone (*45*) on selecting a well-spread set of derivatives. This algorithm states that compounds are selected such that they are both at least a minimum distance from each point already included in the design, and also as close to the center of gravity of the included points as possible. The user specifies the minimum distance and the initial points to be included; the program (see Section VI) applies the algorithm until no more points can be included. The obvious weak points of such a procedure are that no precise method of defining or determining the minimum distance is specified, so that there is no control over the final number of points included except by trial and error. (One should note that

the $\mathcal{F}$ values used (*46*) by Wootton *et al.* (*44*) are incorrect by a factor of 1.65, as noted in Hansch *et al.* (*4a*), Swain and Lupton (*47*) neglected to include this factor in converting σ' to $\mathcal{F}$). Because of the minimum distance criteria, the design will most likely have sufficient variance; however, there is no direct control over covariance (multicollinearity, *48*) and so the design may be multicollinear to some extent.

Hansch, Unger, and Forsythe (*20*) suggested that whereas clustering would provide good variance, it did not guarantee lack of multicollinearity, and that the selected points should be checked by evaluating the determinant of the variance-covariance matrix, det $[(\mathbf{X}'\mathbf{X})\sigma^2]$, or simply det $(\mathbf{X}'\mathbf{X})$, since σ^2, the variance of the fitted equation, is a constant. Unger (*49a*) has presented the results of simulation studies using random selection of points to indicate the magnitude of det $(\mathbf{X}'\mathbf{X})$ which would be considered acceptable for various combinations of parameters (*49b*). While it is thus possible to assign a single number to a given design, until now it has not been feasible to use the det $(\mathbf{X}'\mathbf{X})$ criteria to arrive at an *ab initio* design. This is due to the enormous amount of calculation involved [the det $(\mathbf{X}'\mathbf{X})$ for every combination of N selected from M substituents].

The use of det $(\mathbf{X}'\mathbf{X})$ in experimental design, well known in the statistical literature, is called "D-optimal" design and originated from an idea of Box and Draper (*50, 51*). If Eq. (5) is

$$Y = \mathbf{X}b + e \tag{5}$$

the standard linear regression model, then the regression coefficients b are given by Eq. (6) and the error in these is given by Eq. (7), where s is the standard deviation of the regression.

$$b = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'Y \tag{6}$$

$$V(b) = (\mathbf{X}'\mathbf{X})^{-1}s^2 \tag{7}$$

Thus maximizing det $(\mathbf{X}'\mathbf{X})$ has the effect of minimizing the error in b, in other words, obtaining the best estimates from the regression coefficients. Since the goal of QSAR is to obtain accurate predictions by virtue of an accurate mode, the det $(\mathbf{X}'\mathbf{X})$ criterion is appropriate [alternative criteria, such as maximizing the trace of $(\mathbf{X}'\mathbf{X})$, are also acceptable]. Clearly the procedure is model dependent in that various powers and cross-products must be included in the matrix $\mathbf{X}$ if these are in the final model. Since one usually does not know the model *a priori,* one must make an educated guess as to the relevant parameters. This is *also* required in the Wootton *et al.* procedure since one must include the relevant parameters in order to define the parameter space. Although the simplex method (*41, 42*) is sometimes claimed to be model independent, in fact, if relevant parameters are omitted, one may not reach a true optimum, except by

chance. Therefore, all procedures assume either a knowledge, or an educated guess, of the relevant parameters.

As mentioned, det (X'X) is easy to evaluate for any particular set of substituents and parameters, but would be inefficient to use in an *ab initio* search for a "best" set. A fortuitous discussion with Dr. Jerry Johnson (Syntex Biostatistics) revealed that he, in fact, had an algorithm which allows the det (X'X) criteria to be used in arriving at constrained experimental designs (*52a,b*). Since we program in a matrix-oriented computer language [APL (*8, 53a,b*], we undertook to write programs that could be used for drug design using these algorithms (see Section VI,B).

If $\mathbf{A} = \mathbf{X}'\mathbf{X}$, a is an arbitrary row of the existing matrix X (substituents along rows, parameters along columns), and b is a row vector representing a candidate for inclusion in the design, then Eq. (8) represents a *stepup* procedure for augmenting an existing design (five compounds have been synthesized, what would be the next best compound to make in order to obtain the best possible QSAR?). Equation (9) represents a *stepdown* procedure for eliminating candidates (given a list of 30 possible substituents, what would be the best 10 to actually be made?). And finally, Eq. (10) is a *swapping* procedure for checking on the optimality of the final design, since these algorithms are "sub-optimal" in that they do not examine every possible combination. In practice, the swapping algorithm rarely improves the det A since stepup and stepdown are actually rather efficient.

$$\det \mathbf{A}_+ = (1 + b'\mathbf{A}^{-1}b)(\det \mathbf{A}) \tag{8}$$

$$\det \mathbf{A}_- = (1 - a_i\mathbf{A}^{-1}a_i)(\det \mathbf{A}) \tag{9}$$

$$\det \mathbf{A}_0 = [(1 + b'\mathbf{A}^{-1}b)(1 - a_i\mathbf{A}^{-1}a_i) + (a_i\mathbf{A}^{-1}b)^2](\det \mathbf{A}) \tag{10}$$

Maximizing det A has the effect of maximizing variance and minimizing covariance; in other words, the substituents are well spread out and not multicollinear. The easiest way to do this is to pick points at the extrema of a symmetrical polygon. The det A would do this if it could choose any points in the parameter space. However, it is well adapted to the "constrained" nature of parameter space in drug design. Only a limited number of substituents exist and these represent a rather uneven distribution throughout parameter space. One may fix certain substituents in the design so that they remain in the design (a number of analogs may have already been made) and the algorithms will choose the best substituents to "repair the response surface" (*54*). If it is important to include substituents away from the extrema, then one may also include square or cross-product terms in X.

The STEPDOWN procedure (see Section VI,B) was used to find the best 20-substituent design in the collection of 171 aromatic substituents given in Table VIII. The parameters were π^2, π, $\mathcal{F}$, $\mathcal{R}$, MR, HA, and HD and were unstandardized. The final design could not be SWAPped. The value of det **A** = 3.35×10^{11} was 100 times larger than the largest value obtained in 100 simulations using 20 randomly selected substituents (2.73×10^{9}). This choice is better than about 99% of the random choices.

TABLE VIII
"AROMATIC" SUBSTITUENT CONSTANTS

		π^2	π	σ_m	σ_p	$\mathcal{F}$	$\mathcal{R}$	MR	HA	HD
1	1-TETRAZO	1.08	-1.04	0.52	0.50	0.52	0.02	18.33	1.00	0.00
2	2-BZTHIAZ	4.54	2.13	0.27	0.29	0.25	0.06	38.88	1.00	0.00
3	2-THIENYL	2.59	1.61	0.09	0.05	0.10	0.04	24.04	0.00	0.00
4	25WEPYRR×	3.80	1.95	0.49	0.38	0.52	-0.10	28.80	1.00	0.00
5	3-THIENYL	3.28	1.81	0.03	-0.02	0.04	-0.06	24.04	0.00	0.00
6	34(CH)4	1.74	1.32	0.04	0.04	0.03	0.01	17.47	0.00	0.00
7	34(CH2)3	1.44	1.20	-0.26	-0.26	-0.27	-0.01	13.94	0.00	0.00
8	34(CH2)4	1.93	1.39	-0.48	-0.48	-0.49	-0.03	18.59	0.00	0.00
9	5-CLTETRZ	0.42	-0.65	0.60	0.61	0.58	0.07	23.16	1.00	0.00
10	(OCH2O)	0.00	-0.05	-0.16	-0.16	-0.17	0.00	8.96	1.00	0.00
11	(PH)	3.84	1.96	0.06	-0.01	0.08	-0.08	25.36	0.00	0.00
12	ADAMANTYL	10.89	3.30	-0.12	-0.13	-0.12	-0.02	40.63	0.00	0.00
13	AM ×	7.13	2.67	-0.08	-0.15	-0.06	-0.09	24.25	0.00	0.00
14	B(OH)2	0.30	-0.55	-0.01	0.12	-0.07	0.18	11.04	1.00	1.00
15	BR	0.74	0.86	0.39	0.23	0.44	-0.17	8.88	0.00	0.00
16	BU ×	4.20	2.05	-0.08	-0.16	-0.06	-0.11	19.59	0.00	0.00
17	C-BU ×	2.92	1.71	-0.07	-0.15	-0.05	-0.10	17.88	0.00	0.00
18	C-HX ×	6.30	2.51	-0.07	-0.22	-0.02	-0.20	26.69	0.00	0.00
19	C-PN ×	4.58	2.14	-0.07	-0.02	-0.10	0.07	22.02	0.00	0.00
20	C-PR ×	1.30	1.14	-0.07	-0.21	-0.03	-0.19	13.53	0.00	0.00
21	C=C(CN)2	0.00	0.05	0.66	0.83	0.58	0.30	19.37	1.00	0.00
22	C=C(PH) T	7.18	2.68	0.03	-0.07	0.06	-0.12	34.17	0.00	0.00
23	C=CCO(PH)	0.90	0.95	0.18	0.05	0.22	-0.15	40.25	1.00	0.00
24	C=CCOET	0.23	0.48	0.21	-0.01	0.28	-0.27	25.75	1.00	0.00
25	C=CCOME	0.00	-0.06	0.21	-0.01	0.28	-0.27	21.10	1.00	0.00
26	C=CCOPR ×	1.04	1.02	0.21	-0.01	0.28	-0.27	30.40	1.00	0.00
27	C=CO2ET×	0.74	0.86	0.19	0.03	0.24	-0.19	27.21	1.00	0.00
28	C=CO2H	0.00	0.00	0.14	0.90	-0.15	1.04	17.91	1.00	1.00
29	C=CO2ME×	0.24	0.49	0.19	0.03	0.24	-0.19	22.56	1.00	0.00
30	C=CO2PR×	1.96	1.40	0.19	0.03	0.24	-0.19	31.86	1.00	0.00
31	C=CH2	0.67	0.82	0.05	-0.02	0.07	-0.08	10.99	0.00	0.00
32	C=CHCN	0.03	-0.17	0.24	0.17	0.26	-0.07	16.23	1.00	0.00
33	C=CNO2 T	0.01	0.11	0.32	0.26	0.33	-0.05	16.42	1.00	0.00
34	C=N(PH)	0.08	-0.29	0.35	0.42	0.31	0.13	33.01	1.00	0.00
35	C=NNCO(PH	0.18	0.43	0.38	0.50	0.33	0.20	42.07	1.00	1.00
36	C=NNCONN	1.74	-1.32	0.22	0.16	0.23	-0.05	24.22	1.00	1.00
37	C=NNCSNH	0.07	-0.27	0.45	0.40	0.46	-0.02	29.62	1.00	1.00
38	C=NOBU ×	4.08	2.02	0.37	0.30	0.39	-0.06	29.68	1.00	0.00
39	C=NOET ×	0.88	0.94	0.37	0.30	0.39	-0.06	20.38	1.00	0.00
40	C=NOH	0.14	-0.38	0.22	0.10	0.25	-0.13	10.28	1.00	1.00
41	C=NOME	0.16	0.40	0.37	0.30	0.39	-0.06	15.73	1.00	0.00
42	C=NOPR ×	2.19	1.48	0.37	0.30	0.39	-0.06	25.03	1.00	0.00
43	C≠C(PH)	7.02	2.65	0.14	0.16	0.12	0.05	33.21	0.00	0.00
44	C≠CH	0.16	0.40	0.21	0.23	0.19	0.05	9.55	0.00	1.00
45	CBR3	2.28	1.51	0.28	0.29	0.27	0.04	28.81	0.00	0.00
46	CCCCO2H ×	0.06	0.25	-0.03	-0.07	-0.02	-0.05	21.17	1.00	1.00
47	CCCO2H	0.08	-0.29	-0.03	-0.07	-0.02	-0.05	16.52	1.00	1.00
48	CCL3	1.72	1.31	0.32	0.33	0.31	0.05	20.12	0.00	0.00
49	CF2CF3	3.57	1.89	0.47	0.52	0.44	0.11	9.23	0.00	0.00
50	CF3	0.77	0.88	0.43	0.54	0.38	0.19	5.02	0.00	0.00
51	CHO	0.42	-0.65	0.35	0.42	0.31	0.13	6.88	1.00	0.00
52	CH2(PH)	4.04	2.01	-0.08	-0.09	-0.08	-0.01	30.01	0.00	0.00

(*Continued*)

TABLE VIII—*Continued*

53	CH2BR		0.62	0.79	0.12	0.14	0.10	0.05	13.39	0.00	0.00
54	CH2CL		0.03	0.17	0.11	0.12	0.10	0.03	10.49	0.00	0.00
55	CH2CN		0.32	-0.57	0.16	0.01	0.21	-0.18	10.11	1.00	0.00
56	CH2I		2.25	1.50	0.10	0.11	0.09	0.03	18.60	0.00	0.00
57	CH2O(PH)		2.76	1.66	0.06	0.07	0.05	0.03	31.77	1.00	0.00
58	CH2OBU	x	0.71	0.84	0.02	0.03	0.01	0.02	26.02	1.00	0.00
59	CH2OET	x	0.06	-0.24	0.02	0.03	0.01	0.02	16.72	1.00	0.00
60	CH2OH		1.06	-1.03	0.00	0.00	0.00	0.00	7.19	1.00	1.00
61	CH2OME		0.61	-0.78	0.02	0.03	0.01	0.02	12.07	1.00	0.00
62	CH2OPR	x	0.09	0.30	0.02	0.03	0.01	0.02	21.37	1.00	0.00
63	CL		0.50	0.71	0.37	0.23	0.41	-0.15	6.03	0.00	0.00
64	CN		0.32	-0.57	0.56	0.66	0.51	0.19	6.33	1.00	0.00
65	CO(PH)		1.10	1.05	0.34	0.43	0.30	0.16	30.33	1.00	0.00
66	COET	x	0.00	0.06	0.37	0.49	0.32	0.20	15.83	1.00	0.00
67	COH(CF3)2		1.64	1.28	0.29	0.31	0.28	0.05	15.18	1.00	1.00
68	COME		0.30	-0.55	0.38	0.50	0.32	0.20	11.18	1.00	0.00
69	CONH2		2.22	-1.49	0.28	0.36	0.24	0.14	9.81	1.00	1.00
70	CONHET	x	0.53	-0.73	0.35	0.36	0.34	0.05	19.22	1.00	1.00
71	CONHME		1.61	-1.27	0.35	0.36	0.34	0.05	14.57	1.00	1.00
72	CONHPR	x	0.04	-0.19	0.35	0.36	0.34	0.05	23.87	1.00	1.00
73	COO-		19.01	-4.36	-0.10	0.00	-0.15	0.13	6.05	1.00	0.00
74	COOET		0.26	0.51	0.37	0.45	0.33	0.15	17.47	1.00	0.00
75	COOH		0.10	-0.32	0.37	0.45	0.33	0.15	6.93	1.00	1.00
76	COPR	x	0.28	0.53	0.37	0.49	0.32	0.20	20.48	1.00	0.00
77	CO2(PH)		2.13	1.46	0.37	0.43	0.33	0.13	32.31	1.00	0.00
78	CO2ME		0.00	-0.01	0.37	0.45	0.33	0.15	12.87	1.00	0.00
79	CO2PR	x	1.14	1.07	0.37	0.45	0.33	0.15	22.17	1.00	0.00
80	CSI(ET)3	x	10.63	3.26	-0.16	-0.21	-0.15	-0.07	43.56	0.00	0.00
81	CSNHME		0.10	-0.32	0.30	0.34	0.27	0.09	22.33	1.00	1.00
82	ET		1.04	1.02	-0.07	-0.15	-0.05	-0.10	10.30	0.00	0.00
83	FERROCEN		6.05	2.46	-0.15	-0.18	-0.15	-0.04	48.24	0.00	0.00
84	F		0.02	0.14	0.34	0.06	0.43	-0.34	0.92	0.00	0.00
85	H		0.00	0.00	0.00	0.00	0.00	0.00	1.03	0.00	0.00
86	I-PR		2.34	1.53	-0.07	-0.15	-0.05	-0.10	14.96	0.00	0.00
87	IO2		11.87	3.46	0.68	0.78	0.63	0.20	63.51	1.00	0.00
88	I		1.25	1.12	0.35	0.18	0.40	-0.19	13.94	0.00	0.00
89	ME		0.31	0.56	-0.07	-0.17	-0.04	-0.13	5.65	0.00	0.00
90	N(ET)2		1.39	1.18	-0.23	-0.90	0.01	-0.91	24.85	1.00	0.00
91	N(ME)2		0.03	0.18	-0.15	-0.83	0.10	-0.92	15.55	1.00	0.00
92	N(PH)2		12.25	3.50	0.00	-0.22	0.07	-0.29	54.96	1.00	0.00
93	N=C=S		1.32	1.15	0.48	0.38	0.51	-0.09	17.24	1.00	0.00
94	N=CH(PH)		0.08	-0.29	-0.08	-0.55	0.09	-0.63	33.01	1.00	0.00
95	N=CCL2		0.17	0.41	0.21	0.13	0.23	-0.08	18.35	0.00	0.00
96	N=N(PH)		2.86	1.69	0.32	0.39	0.28	0.13	31.31	0.00	0.00
97	NHCSNHET		0.50	-0.71	0.30	0.07	0.38	-0.28	31.66	1.00	1.00
98	NH(PH)		1.88	1.37	-0.12	-0.40	-0.02	-0.38	30.04	1.00	1.00
99	NHBU		1.35	1.16	-0.34	-0.51	-0.28	-0.25	24.26	1.00	1.00
100	NHCHO		0.96	-0.98	0.19	0.00	0.25	-0.23	10.31	1.00	1.00
101	NHCN		0.07	-0.26	0.21	0.06	0.26	-0.18	10.14	1.00	1.00
102	NHCO2ET		0.03	0.17	0.07	-0.15	0.14	-0.28	21.18	1.00	1.00
103	NHCO2ME		0.14	-0.37	0.07	-0.15	0.14	-0.28	16.53	1.00	1.00
104	NHCO2PR	x	0.50	0.71	0.07	-0.15	0.14	-0.28	25.83	1.00	1.00
105	NHCO(PH)		0.24	0.49	0.02	-0.19	0.09	-0.27	34.64	1.00	1.00
106	NHCOCCL		0.25	-0.50	0.17	-0.03	0.23	-0.25	19.77	1.00	1.00
107	NHCOCF3		0.01	0.08	0.30	0.12	0.36	-0.21	14.30	1.00	1.00
108	NHCOET	x	0.18	-0.43	0.21	0.00	0.28	-0.26	19.58	1.00	1.00
109	NHCOME		0.94	-0.97	0.21	0.00	0.28	-0.26	14.93	1.00	1.00
110	NHCONH2		1.69	-1.30	-0.03	-0.24	0.04	-0.28	13.72	1.00	1.00
111	NHCOPR	x	0.01	0.11	0.21	0.00	0.28	-0.26	24.23	1.00	1.00
112	NHCSET	x	0.01	0.12	0.24	0.12	0.27	-0.13	28.05	1.00	1.00
113	NHCSME		0.18	0.42	0.24	0.12	0.27	-0.13	23.40	1.00	1.00
114	NHCSNH2		1.96	-1.40	0.22	0.16	0.23	-0.05	22.19	1.00	1.00
115	NHCSPR	x	0.44	0.66	0.24	0.12	0.27	-0.13	32.70	1.00	1.00
116	NHET		0.01	0.08	-0.24	-0.61	-0.11	-0.51	14.98	1.00	1.00
117	NHME		0.22	-0.47	-0.30	-0.84	-0.11	-0.74	10.33	1.00	1.00
118	NHNH2		0.77	-0.88	-0.02	-0.55	0.17	-0.71	8.44	1.00	1.00
119	NHOH		1.80	-1.34	-0.04	-0.34	0.06	-0.40	7.22	1.00	1.00
120	NHPR	x	0.38	0.62	-0.24	-0.61	-0.11	-0.51	19.63	1.00	1.00
121	NHSO2(PH)		0.20	0.45	0.16	0.01	0.21	-0.18	37.88	1.00	1.00

TABLE VIII—*Continued*

122	*NHSO2ET*	x	0.41	−0.64	0.20	0.03	0.25	−0.20	22.82	1.00	1.00
123	*NHSO2ME*		1.39	−1.18	0.20	0.03	0.25	−0.20	18.17	1.00	1.00
124	*NHSO2PR*	x	0.01	−0.10	0.20	0.03	0.25	−0.20	27.47	1.00	1.00
125	*NH2*		1.51	−1.23	−0.16	−0.66	0.02	−0.68	5.42	1.00	1.00
126	*NNN*		0.21	0.46	0.27	0.15	0.30	−0.13	10.20	0.00	0.00
127	*NO2*		0.08	−0.28	0.71	0.78	0.67	0.16	7.36	1.00	0.00
128	*O(PH)*		4.33	2.08	0.25	−0.32	0.46	−0.74	27.68	1.00	0.00
129	*O-I-PR*		0.13	0.36	0.10	−0.45	0.30	−0.72	17.06	1.00	0.00
130	*OBU*		1.77	1.33	0.10	−0.32	0.25	−0.55	21.66	1.00	0.00
131	*OCF3*		1.08	1.04	0.38	0.35	0.38	0.00	7.86	1.00	0.00
132	*OCO(PH)*		2.13	1.46	0.21	0.13	0.23	−0.08	32.33	1.00	0.00
133	*OCOET*	x	0.01	−0.10	0.39	0.31	0.41	−0.07	17.12	1.00	0.00
134	*OCOME*		0.41	−0.64	0.39	0.31	0.41	−0.07	12.47	1.00	0.00
135	*OCOPR*	x	0.19	0.44	0.39	0.31	0.41	−0.07	21.77	1.00	0.00
136	*OET*		0.14	0.38	0.10	−0.24	0.22	−0.44	12.47	1.00	0.00
137	*OH*		0.45	−0.67	0.12	−0.37	0.29	−0.64	2.85	1.00	1.00
138	*OME*		0.00	−0.02	0.12	−0.27	0.26	−0.51	7.87	1.00	0.00
139	*OPR*		1.10	1.05	0.10	−0.25	0.22	−0.45	17.06	1.00	0.00
140	*OSO2(PH)*		0.86	0.93	0.36	0.33	0.36	0.00	36.70	1.00	0.00
141	*OSO2ME*		0.77	−0.88	0.39	0.36	0.39	0.00	16.99	1.00	0.00
142	*P(ET)2*	x	1.64	1.28	0.03	0.31	−0.08	0.39	30.49	0.00	0.00
143	*P(ME)2*		0.19	0.44	0.03	0.31	−0.08	0.39	21.19	0.00	0.00
144	*PO(OET)2*	x	0.03	−0.18	0.55	0.60	0.52	0.12	28.52	1.00	0.00
145	*PO(OME)2*	x	1.39	−1.18	0.42	0.53	0.37	0.19	21.87	1.00	0.00
146	*PO(PH)2*		0.49	0.70	0.38	0.53	0.31	0.24	59.29	1.00	0.00
147	*PR*		2.40	1.55	−0.07	−0.13	−0.06	−0.08	14.96	0.00	0.00
148	*PYRRYL*		0.90	0.95	0.47	0.37	0.50	−0.09	19.51	1.00	0.00
149	*S=C=N*		0.17	0.41	0.41	0.52	0.36	0.19	13.40	1.00	0.00
150	*SCF3*		2.07	1.44	0.40	0.50	0.56	0.01	13.81	0.00	0.00
151	*SCOET*	x	0.41	0.64	0.39	0.44	0.36	0.11	23.07	1.00	0.00
152	*SCOME*		0.01	0.10	0.39	0.44	0.36	0.11	18.42	1.00	0.00
153	*SCOPR*	x	1.39	1.18	0.39	0.44	0.36	0.11	27.72	1.00	0.00
154	*SEME*		0.55	0.74	0.10	0.00	0.13	−0.12	17.03	0.00	0.00
155	*SET*		1.14	1.07	0.18	0.03	0.23	−0.18	18.42	0.00	0.00
156	*SF5*		1.51	1.23	0.61	0.68	0.57	0.15	9.89	0.00	0.00
157	*SH*		0.15	0.39	0.25	0.15	0.28	−0.11	9.22	0.00	1.00
158	*SI(ME)3*		6.71	2.59	−0.04	−0.07	−0.04	−0.04	24.96	0.00	0.00
159	*SME*		0.37	0.61	0.15	0.00	0.20	−0.18	13.82	0.00	0.00
160	*SOET*	x	1.08	−1.04	0.52	0.49	0.52	0.01	18.35	1.00	0.00
161	*SOME*		2.50	−1.58	0.52	0.49	0.52	0.01	13.70	1.00	0.00
162	*SOPR*	x	0.25	−0.50	0.52	0.49	0.52	0.01	23.00	1.00	0.00
163	*SO2(PH)*		0.07	0.27	0.61	0.70	0.56	0.18	33.20	1.00	0.00
164	*SO2CF3*		0.30	0.55	0.79	0.93	0.73	0.26	12.86	1.00	0.00
165	*SO2ET*	x	1.19	−1.09	0.60	0.72	0.54	0.22	18.14	1.00	0.00
166	*SO2F*		0.00	0.05	0.80	0.91	0.75	0.22	8.65	1.00	0.00
167	*SO2ME*		2.66	−1.63	0.60	0.72	0.54	0.22	13.49	1.00	0.00
168	*SO2NH2*		3.31	−1.82	0.46	0.57	0.41	0.19	12.28	1.00	1.00
169	*SO2PR*	x	0.30	−0.55	0.60	0.72	0.54	0.22	22.79	1.00	0.00
170	*SPR*	x	2.59	1.61	0.18	0.03	0.23	−0.18	23.07	0.00	0.00
171	*T-BU*		3.92	1.98	−0.10	−0.20	−0.07	−0.13	19.62	0.00	0.00

[a] All parameters defined as in Ref. *4a* or Pomona College Medicinal Chemistry Project (1978) parameter listing (*4b*). HA = hydrogen bond acceptor, HD = hydrogen-bond donor. x = estimated from homologs. Explanations of names in table: **1** 1-tetrazolyl, **2** 2-benzthiazolyl, **4** 1-(2,5-dimethyl-pyrryl), **9** 5-Cl-1-tetrazolyl.

Another way of assessing the quality of the choice is by examining the eigenvalues. These would ideally be co-equal for a perfectly noncollinear set. For this design we have 1.76, 1.57, 1.08, 1.07, 0.776, 0.499, and 0.243. Thus, there are 4 significant eigenvalues out of 7; the smallest is 3.7% of the total variance. In 20 simulations using random choices, only one (5%) gave a smallest eigenvalue with $\geqq$ 3% of the total variance, and only 10% had 4 eigenvalues $\geqq$ 1.0. The task of choosing 20 substituents in 7 dimen-

sions, such that there is maximum variance and minimum co-variance, would be totally hopeless by any noncomputerized technique. The substituents are given in Table IX, along with Farrar and Glauber's (*48*) omega vector and t-matrix (see Section VI,B) to evaluate further multicollinearity. Note that there are no significant multiple correlations, and that the largest single correlation, between π^2 and MR, is only 17% and is insignificant ($t = 2.006$; $t_{13}^{\alpha=0.05} = 2.16$). This example is meant to be illustrative of the results obtained; it is not synthetically useful. One might use cluster analysis or table of bioisosteric substituents (*4b, 9*) to pick alternatives.

Wootton *et al.* (*44*) presented the results of a search for a useful set of para analogs. We have corrected their f$\mathcal{F}$ values (*46*) and present the full data matrix in Table X. Using our DISTPLAN (see Section VI,B) version of their algorithm and standardizing the 35 substituent matrix using π, f$\mathcal{F}$, r$\mathcal{R}$ and MR, we arrive at a slightly different design: H (fixed), 4-Et, 4-OEt, 4-COOMe, 4-Cl, 4-CN, 4-OH, 4-SO_2NH_2, 4-NH_2, 4-NMe_2, 4-O-*n*Bu, and 4-*n*Bu. Unfortunately, it took a number of attempts at guessing a cutoff distance that would yield 12 substituents. This proved more expensive than using STEPDOWN directly, which yielded the following design: H (fixed), 4-*n*Bu, 4-*t*Bu, 4-CF_3, 4-O-*n*Am, 4-OPh, 4-NH_2, 4-NMe_2, 4-NO_2, 4-SO_2Me, 4-SO_2NH_2, and 4-F. The eigenvalues are very similar for both designs: 1.81, 1.13, 0.675, and 0.389 for DISTPLAN design and 1.94, 0.986, 0.728, and 0.346 for STEPDOWN plan. The values of det $\mathbf{A} = 2.68 \times 10^4$ and 1.38×10^5 respectively (note that the matrix **X** contains a column of 1's for this evaluation, since the linear model contains an intercept). Neither design shows any significant multicollinearity by Farrar and Glauber's (*48*) criteria. Aside from slightly greater variance in the STEPDOWN design and an inefficient method of specifying the number of substituents in DISTPLAN, there is little to distinguish between the two methods in the present test.

Greater differences are obtained when a larger number of substituents and parameters are considered. When the full 95 substituents of Table X are used (standardized by mean and standard deviation), with π^2, π, f$\mathcal{F}$, r$\mathcal{R}$, MR, HA, and HD, the STEPDOWN procedure gives in a single (relatively expensive) pass: H (fixed), 2-*n*Bu, 2-OPh, 2-NHAc, 2-F, 4-SO_2Me, 4-Ph, 4-NH_2, 4-NMe_2, 4-NO_2, and 4-SO_2NH_2. DISTPLAN gives the following design, but only after a considerable number of passes to determine the proper cutoff: H (fixed), 3-SMe, 3-OMe, 4-F, 2-CN, 2-O-*i*Pr, 3-OH, 3-O-*n*Bu, 2-*i*Pr, 3-$CONH_2$, and 2-NH_2. The fact that STEPDOWN (with one substituent SWAPped) does not give any meta substituents is of no relevance; this could have been specified in the design, if necessary. The eigenvalues are: 2.52, 1.42, 1.25, 1.22, 0.25,

TABLE IX

"BEST" SET OF 20 SUBSTITUENTS SELECTED FROM TABLE VIII BY STEPDOWN PROCEDURE USING π^2, π, $\mathcal{F}$, $\mathcal{R}$, MR, HA, AND HD[a]

3,4-$(CH_2)_4$
(—OCH_2O—)
adamantyl
CH=CHCOOH
CH=NNHCOPh
C≡CH
COO^-
CH_2SiEt_3
F
IO_2
NMe_2
NPh_2
NHMe
NH_2
OPh
PMe_2
$POPh_2$
SH
SO_2CF_3
SO_2F

[a] Farrer and Glauber's (*48*) measures of multicollinearity.

omega ($F_{6,13}^{\alpha=0.05} = 2.28$):

1.71 1.66 0.61 0.31 1.78 0.87 0.67

r^2:

1.0	0.01	0.03	0.00	0.17	0.00	0.15
	1.0	0.02	0.02	0.08	0.14	0.02
		1.0	0.00	0.01	0.09	0.01
			1.0	0.03	0.00	0.00
				1.0	0.02	0.08
					1.0	0.00
						1.0

t (diagonal undefined; $t_{13}^{\alpha=0.05} = 2.16$):

x	−1.5	−1.3	−.53	2.0	−.4	−1.5
	x	−.8	−1.0	1.9	−1.7	−.7
		x	−.1	.8	.7	−.8
			x	1.1	−.8	.04
				x	1.1	.01
					x	.1
						x

TABLE X

CORRECTED SUBSTITUENT CONSTANTS OF NORRINGTON *et al.*[a]

		π^2	π	π^2_-	π_-	$f\mathcal{F}$	$r\mathcal{R}$	MR	HA	HD
1	*H*	0.00	0.00	0.00	0.00	0.00	0.00	1.03	0.00	0.00
2	*2-ME*	0.71	0.84	0.24	0.49	−0.04	−0.12	5.73	0.00	0.00
3	*2-ET*	1.93	1.39	0.98	0.99	−0.05	−0.10	10.43	0.00	0.00
4	*2-N-PR*	3.57	1.89	2.22	1.49	−0.02	−0.10	15.03	0.00	0.00
5	*2-I-PR*	3.13	1.77	1.88	1.37	−0.05	−0.10	15.03	0.00	0.00
6	*2-N-BU*	5.71	2.39	3.96	1.99	−0.04	−0.11	19.73	0.00	0.00
7	*2-T-BU*	4.71	2.17	3.13	1.77	−0.08	−0.12	19.53	0.00	0.00
8	*2-CF3*	1.08	1.04	1.80	1.34	0.48	0.16	5.03	0.00	0.00
9	*2-OH*	0.17	−0.41	0.34	−0.58	0.37	−0.56	2.53	1.00	1.00
10	*2-OME*	0.11	−0.33	0.02	−0.13	0.32	−0.43	7.53	1.00	0.00
11	*2-OET*	0.03	0.17	0.14	0.37	0.27	−0.38	12.33	1.00	0.00
12	*2-O-N-PR*	0.45	0.67	0.76	0.87	0.28	−0.39	16.93	1.00	0.00
13	*2-O-I-PR*	0.30	0.55	0.56	0.75	0.37	−0.63	17.03	1.00	0.00
14	*2-O-N-BU*	1.37	1.17	1.88	1.37	0.31	−0.48	21.73	1.00	0.00
15	*2-O-N-AM*	2.79	1.67	3.50	1.87	0.32	−0.50	26.33	1.00	0.00
16	*2-OPH*	0.94	0.97	0.66	0.81	0.56	−0.64	27.63	1.00	0.00
17	*2-OAC*	0.34	−0.58	1.04	−1.02	0.52	−0.06	12.63	1.00	0.00
18	*2-NH2*	1.96	−1.40	0.71	−0.84	0.03	−0.58	5.23	1.00	1.00
19	*2-NME2*	0.03	0.16	0.23	−0.48	0.02	−0.73	15.43	1.00	0.00
20	*2-NHAC*	0.02	−0.14	0.55	−0.74	0.36	−0.24	15.63	1.00	1.00
21	*2-CHO*	0.18	−0.43	0.06	0.24	0.51	−0.13	6.33	1.00	0.00
22	*2-SME*	0.76	0.87	0.09	0.30	0.25	−0.16	14.03	0.00	0.00
23	*2-CN*	0.11	−0.33	0.02	0.13	0.64	0.16	6.23	1.00	0.00
24	*2-F*	0.00	0.00	0.06	0.25	0.53	−0.29	0.63	0.00	0.00
25	*2-CL*	0.58	0.76	0.48	0.69	0.52	−0.14	5.83	0.00	0.00
26	*2-BR*	0.71	0.84	0.79	0.89	0.55	−0.15	8.63	0.00	0.00
27	*2-I*	0.86	0.93	1.42	1.19	0.51	−0.17	13.83	0.00	0.00
28	*3-ME*	0.27	0.52	0.25	0.50	−0.03	−0.05	5.73	0.00	0.00
29	*3-ET*	0.98	0.99	0.88	0.94	−0.04	−0.04	10.43	0.00	0.00
30	*3-N-PR*	2.10	1.45	2.07	1.44	−0.02	−0.04	15.03	0.00	0.00
31	*3-I-PR*	1.77	1.33	1.74	1.32	−0.04	−0.04	15.03	0.00	0.00
32	*3-N-BU*	3.69	1.92	3.76	1.94	−0.04	−0.05	19.73	0.00	0.00
33	*3-T-BU*	2.89	1.70	2.96	1.72	−0.06	−0.05	19.53	0.00	0.00
34	*3-PH*	3.69	1.92	3.13	1.77	0.08	−0.03	25.33	0.00	0.00
35	*3-CF3*	1.21	1.10	2.22	1.49	0.38	0.07	5.03	0.00	0.00
36	*3-OH*	0.25	−0.50	0.44	−0.66	0.29	−0.22	2.53	1.00	1.00
37	*3-OME*	0.01	0.12	0.01	0.12	0.25	−0.17	7.53	1.00	0.00
38	*3-OET*	0.38	0.62	0.38	0.62	0.22	−0.15	12.33	1.00	0.00
39	*3-O-N-PR*	1.25	1.12	1.25	1.12	0.22	−0.16	16.93	1.00	0.00
40	*3-O-I-PR*	1.00	1.00	1.00	1.00	0.29	−0.25	17.03	1.00	0.00
41	*3-O-N-BU*	2.62	1.62	2.62	1.62	0.24	−0.19	21.73	1.00	0.00
42	*3-O-N-AM*	4.49	2.12	4.49	2.12	0.25	−0.20	26.33	1.00	0.00
43	*3-OPH*	2.43	1.56	2.43	1.56	0.44	−0.26	27.63	1.00	0.00
44	*3-OAC*	0.36	−0.60	0.05	−0.23	0.41	−0.03	12.63	1.00	0.00
45	*3-NH2*	1.66	−1.29	1.66	−1.29	0.02	−0.24	5.23	1.00	1.00
46	*3-NME2*	0.01	0.11	0.01	0.10	0.02	−0.29	15.43	1.00	0.00
47	*3-NHAC*	0.61	−0.78	0.53	−0.73	0.28	−0.10	15.63	1.00	1.00
48	*3-NO2*	0.01	0.11	0.29	0.54	0.66	0.05	7.03	1.00	0.00
49	*3-CHO*	0.22	−0.47	0.01	−0.08	0.40	−0.05	6.33	1.00	0.00
50	*3-AC*	0.08	−0.28	0.00	−0.07	0.32	0.07	10.93	1.00	0.00
51	*3-COOME*	0.00	−0.04	0.18	0.43	0.27	0.07	12.43	1.00	0.00
52	*3-COOET*	0.21	0.46	0.86	0.93	0.33	0.05	17.23	1.00	0.00
53	*3-CONH2*	2.28	−1.51	0.32	−0.57	0.24	0.05	9.83	1.00	1.00
54	*3-SME*	0.41	0.64	0.30	0.55	0.20	−0.07	14.03	0.00	0.00
55	*3-SO2ME*	1.56	−1.25	1.04	−1.02	0.53	0.08	13.53	1.00	0.00
56	*3-SO2NH2*	3.46	−1.86	4.41	−2.10	0.41	0.07	12.33	1.00	1.00
57	*3-CN*	0.10	−0.31	0.06	0.24	0.50	0.06	6.23	1.00	0.00
58	*3-F*	0.05	0.22	0.22	0.47	0.42	−0.12	0.63	0.00	0.00
59	*3-CL*	0.59	0.77	1.08	1.04	0.41	−0.06	5.83	0.00	0.00
60	*3-BR*	0.92	0.96	1.37	1.17	0.43	−0.06	8.63	0.00	0.00
61	*3-I*	1.39	1.18	2.16	1.47	0.40	−0.07	13.83	0.00	0.00
62	*4-ME*	0.36	0.60	0.23	0.48	−0.03	−0.14	5.73	0.00	0.00
63	*4-ET*	1.21	1.10	0.96	0.98	−0.04	−0.11	10.43	0.00	0.00
64	*4-N-PR*	2.56	1.60	2.19	1.48	−0.02	−0.11	15.03	0.00	0.00
65	*4-I-PR*	2.04	1.43	1.85	1.36	−0.04	−0.11	15.03	0.00	0.00

TABLE X—*Continued*

66	4-*N-BU*	4.41	2.10	3.92	1.98	−0.04	−0.13	19.73	0.00	0.00
67	4-*T-BU*	3.53	1.88	3.10	1.76	−0.06	−0.14	19.53	0.00	0.00
68	4-*PH*	3.03	1.74	3.03	1.74	0.08	−0.09	25.33	0.00	0.00
69	4-*CF3*	1.08	1.04	1.10	1.05	0.38	0.19	5.03	0.00	0.00
70	4-*OH*	0.37	−0.61	0.76	−0.87	0.30	−0.64	2.53	1.00	1.00
71	4-*OME*	0.00	−0.03	0.01	−0.12	0.25	−0.50	7.53	1.00	0.00
72	4-*OET*	0.22	0.47	0.12	0.35	0.22	−0.44	12.33	1.00	0.00
73	4-*O-N-PR*	0.94	0.97	0.72	0.85	0.22	−0.46	16.93	1.00	0.00
74	4-*O-I-PR*	0.72	0.85	0.53	0.73	0.30	−0.72	17.03	1.00	0.00
75	4-*O-N-BU*	2.16	1.47	1.82	1.35	0.25	−0.55	21.73	1.00	0.00
76	4-*O-N-AM*	3.86	1.97	3.42	1.85	0.25	−0.58	26.33	1.00	0.00
77	4-*OPH*	1.80	1.34	2.13	1.46	0.45	−0.74	27.63	1.00	0.00
78	4-*OAC*	0.34	−0.58	1.12	−1.06	0.41	−0.07	12.63	1.00	0.00
79	4-*NH2*	1.69	−1.30	2.02	−1.42	0.02	−0.68	5.23	1.00	1.00
80	4-*NME2*	0.01	−0.08	0.48	−0.69	0.02	−0.85	15.43	1.00	0.00
81	4-*NHAC*	0.31	−0.56	1.46	−1.21	0.28	−0.27	15.63	1.00	1.00
82	4-*NO2*	0.05	0.22	0.20	0.45	0.67	0.16	7.03	1.00	0.00
83	4-*CHO*	0.22	−0.47	0.00	−0.06	0.41	−0.15	6.33	1.00	0.00
84	4-*AC*	0.15	−0.39	0.01	−0.11	0.32	0.20	10.93	1.00	0.00
85	4-*COOME*	0.00	−0.04	0.25	0.50	0.27	0.19	12.43	1.00	0.00
86	4-*COOET*	0.21	0.46	1.00	1.00	0.33	0.14	17.23	1.00	0.00
87	4-*CONH2*	2.28	−1.51	0.77	−0.88	0.25	0.14	9.83	1.00	1.00
88	4-*SME*	0.76	0.87	0.10	0.32	0.20	−0.19	14.03	0.00	0.00
89	4-*SO2ME*	1.44	−1.20	1.04	−1.02	0.55	0.22	13.53	1.00	0.00
90	4-*SO2NH2*	3.46	−1.86	2.25	−1.50	0.41	0.19	12.33	1.00	1.00
91	4-*CN*	0.11	−0.33	0.02	0.14	0.52	0.18	6.23	1.00	0.00
92	4-*F*	0.02	0.15	0.10	0.31	0.43	−0.34	0.63	0.00	0.00
93	4-*CL*	0.53	0.73	0.86	0.93	0.42	−0.16	5.83	0.00	0.00
94	4-*BR*	1.42	1.19	1.28	1.13	0.44	−0.18	8.63	0.00	0.00
95	4-*I*	2.04	1.43	2.10	1.45	0.41	−0.20	13.83	0.00	0.00

[a] Norrington *et al.* (*46*).

0.20, and 0.11 for STEPDOWN and 2.40, 1.82, 1.34, 0.96, 0.22, 0.15, and 0.11 for DISTPLAN. The det A = 1.53×10^8 and 3.7×10^6, respectively. Finally, analysis by multicollinearity (*48*) shows that π and HA have the largest, although insignificant, correlation with $r^2 = 36\%$ ($t = 1.57$; $t_4^{\alpha=0.05} = 2.78$) for the STEPDOWN design, while π and HD are more highly correlated, although still not significantly, in the DISTPLAN design, $r^2 = 58\%$ ($t = 1.97$). Therefore, STEPDOWN appears to have a slight advantage over the DISTPLAN algorithm.

B. STATISTICAL MODELING

With a well-planned set of analogs in hand, one then attempts to derive a QSAR using any of the appropriate techniques (*3*), depending upon the quality of the data, goals of the study, etc. The discussion here will focus on their uses, implications, and interrelations.

Regression analysis seeks a quantitative relation between the BR and physicochemical or property variables. There are both linear and nonlinear models. The main difficulty, as in most of QSAR work, is in finding the relevant subset of parameters. Whereas this candidate list is relatively small, all regressions may be examined. However, with a larger list of

candidates this becomes impractical. Since all combinations of n parameters goes up as $2^n - 1$, one rapidly increases the cost of calculation. Therefore, several sub-optimal procedures are used. Stepwise addition or deletion of parameters generally gives satisfactory results. The popular BMD programs (*13*) contain a very flexible algorithm (BMDP2R) with four modes of selection of parameters based on F-to-enter and F-to-remove. A second approach is that called LEAPS AND BOUNDS (*55*) (also available as BMDP9R), a modified all-regression algorithm that follows a critical path through all regressions. This is based on the fact that it is impossible to reduce the residual sum of squares for any regression by removing variables. This is a useful program especially in making the investigator aware that a number of combinations of parameters can frequently give the same goodness-of-fit. [Chemical and biological understanding must be used to select the most desirable equation (*14*).] A third method may be used for parameter selection, but is most useful in studying the "best" equation for multicollinearity (*48*). This is called RIDGE regression. The least squares solution [Eq. (6)] gives best predictions when all parameters are orthogonal. Since this is usually not entirely true, a number of studies (*56*) have shown that improved predictions are obtained by "relaxing" the least squares condition by adding bias, as in Eq. (11).

$$b = [(\mathbf{X}'\mathbf{X}) + k\mathbf{I}]^{-1}\mathbf{X}'Y \tag{11}$$

where **I** is a diagonal identity matrix, and k is a positive number ($0 \leqq k \leqq 1$) which may be approximated by Eq. (12) (*39b*):

$$k = ps^2/b'b \tag{12}$$

where p is the number of parameters. The k may also be varied continuously and the regression coefficients b plotted vs. k to visually find the stable point.

One seeks the best equation in which all parameters are as nearly orthogonal as possible, each accepted parameter is significant by a t-test or stepwise F-test at $\geqq$ 90% C.L., the overall F is significant, and the parameters in combination make chemical and biological sense. In our studies on the tuberins (**VI**) (*21*), we employed RIDGE regression in an attempt to deconvolute the high correlation between $\log P$ and $(\log P)^2$; however, $\log P_0$ did not change significantly. We have found that unless $k \leqq 0.1$, the QSAR is usually very difficult to work with.

A study by Topliss and Costello (*57*) has caused much confusion and provoked a lot of thought. These authors studied "chance correlations in structure–activity studies using multiple regression analysis" in order to determine the effect of the number of parameters *examined* on the possibil-

ity of obtaining chance correlations.* Note that they did *not* study the effect of the number of parameters *included* in the regression equation (this is adequately covered by the overall F-test), but the number of parameters *studied*. The obvious conclusion was that the greater the number of parameters studied, the greater the role chance might play in finding "significant" correlations. Topliss and Costello recommended, approximately, the following ratio of compounds to parameters examined (No. parameters examined/compounds per parameter): 5/6, 10/5, 20/3.2, and 30/2.8. Thus if one has a list of 10 candidate parameters, one should have at least 50 compounds. The implication is that if one has fewer than 50 compounds, one should simply examine fewer candidate parameters! This is, of course, a misinterpretation of Topliss and Costello's meaning and results. One must examine enough parameters to convince oneself that the model is correct, while being cognizant of the possibility of chance correlations. The only solution to this apparant dilemma is to test the model by testing its predictions. One must also be careful to watch for good model building criteria (*14, 51*).

Cluster analysis has some use in lead optimization by helping to span substituent space (*20, 43*). Dunn (*59a,b*) has used this approach in the design, synthesis, and QSAR analysis of analogs of the antitumor 1-phenyl-3-benzyl-3-methyltriazines and 1-phenyl-3,3-dimethyltriazines. Coats (*60*) has studied the 7-x-4-hydroxyquinoline-3-carboxylic acids as inhibitors of cellular respiration using cluster analysis. The tuberins (**VI**) (*21*) were well planned, which facilitated the QSAR study.

When the BR can only be classified by rough activity groups, discriminant analysis (*13, 15a,b*), or pattern recognition (*24a, 34–38, 58a–d*) techniques can be used. One can only predict broad activity classes and may not necessarily achieve optimization [cf. SIMCA (*29, 38*)]. The technique is therefore somewhat more useful for lead generation than for optimization. However, one expects the optimum compound to lie in the most active class, and it would be predicted qualitatively but not quantitatively. Factor analysis is of greatest use in studying interrelations of compounds or tests and can be used to detect outliers. In this sense, it is of some use in lead optimization.

C. QUANTUM AND MOLECULAR MECHANICS

There has been considerable interest in the application of quantum chemical calculations to drug design (*61a–c*). The cost of such calculations

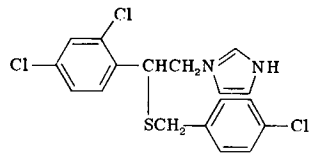
* Unpublished observations by the author confirm the general conclusions of Topliss and Costello with respect to discriminant analysis.

does not seem to have been justified by the level of success. However, with increasing knowledge of QSAR and the multidimensional approach to drug design, there is more hope for success. Calculation of gas-phase properties has little relevance to the biophase; yet, if these properties are proportional within a series, there is some usefulness for the technique in combination with measures of steric and lipophilic effects. There are some attempts at calculating substrate–receptor interactions, in which case these latter effects are automatically accounted for. QM calculations can be used either as an end in themselves or as input for multiple regression or other QSAR techniques (*62a,b*).

Because of the cost of QM calculations, the molecular mechanics approach, such as CAMSEQ* (*63a,b*), has generated considerable interest. Hopfinger and Battershell (*64*) have successfully applied CAMSEQ to a study of some pyrethroid insecticides by correctly predicting conformational effects. One of the most serious problems with such an approach is the question of the relevance of the results with respect to the active site. For example, some enzymes, such as hemoglobin, undergo a conformational transition upon complexation with substrates; others may twist the substrate, without themselves twisting (*65a–c*). This must be dealt with on a case by case basis. Cramer (*66*) has made some progress in applying CAMSEQ to QSAR by calculating point force fields (lattice model) about each of the preferred conformations determined by CAMSEQ. The field is composed of steric, dispersion, and electrostatic effects. Such an approach generates an enormous amount of data, which, after dimension reduction by factor analysis, is used as input in QSAR regressions.

D. Stop or Continue? Example: Sulconazole Analogs

A very important function of QSAR is to help answer the question "stop or continue"? For example, in a study of the *in vivo* activity of antifungal compounds related to sulconazole (**VIII**), we found that the



(VIII)

most active compounds being screened were also the most lipophilic and that the trend was linear. Therefore, even more lipophilic analogs were

* CAMSEQ is available through the NIH/EPA/CIS network on Interactive Sciences Corp., Braintree, Massachusetts.

tested in order to define the maximum activity. In another case, involving beta-blockers, we found very little change in activity over a relatively broad range of lipophilicity and structural type and suggested dropping the series. In a CNS series, we encouraged full investigation of the anomaly of the hydrophilic substituents, and this greatly simplified future work.

E. MULTIPLE OPTIMIZATION

If one BR can be studied by QSAR, then several can be studied simultaneously. These parameters can be activities, toxicities, or such other properties as solubility and taste. One then wishes to optimize various absolute activities and/or therapeutic ratios. If each BR is a function of several factors, then, without the aid of a computer, this is clearly a hopeless task. The most direct approach is to treat each BR separately and then to examine the QSAR to see how each may be optimized. It is very important to understand the interrelationships of the various bioassays, if one attempts a multiple optimization. If the bioassays are uncorrelated, one might be optimistic about separating effects; if the bioassays are correlated, the task of separating activities becomes more and more hopeless, as the correlation increases. Assuming the bioassays are independent, one can often find a certain position of substitution, type of substituent, lipophilicity, etc. that confers specificity (*25, 26, 28b*).

Example: Phenyl Substituted Imidazoline Alkyl Carbamates

In series **VII**, mentioned above, we noted that antidepressant activity increased as the other "depressant" activities decreased, as a function of lipophilicity (see Fig. 4). However, stereoelectronic effects allowed one to gain significant increases in activity. One could thus predict the *profile* of these compounds with reasonable accuracy.

Example: Thiazole Beta-Blockers

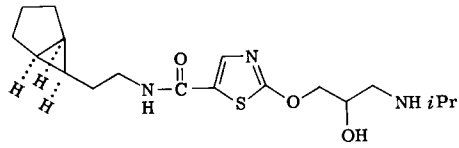
The alkyl carbamoyl thiazole beta-blockers (**I**) provide a second example of multiple optimization. First, we noted that the chronotropic and inotropic activities in dog models (Table I) were highly correlated for a diverse group of analogs:

$$pED_{50}^{rate} = 0.8(\pm 0.7) + 0.9(\pm 0.1)pED_{50}^{force}$$

$$n = 62 \quad s = 0.16 \quad r = 0.898 \quad F_{1,60} = 250 \qquad (13)$$

In order to reduce the amount of work, we concentrated on force measurements, because these seemed more reliable. We have already de-

scribed (Section III,D) how some of the preliminary ideas concerning activity, selectivity, and depressivity arose. These concepts came mainly from a combination of what we have called "QSAR insight," and from a study of the interrelationships of various data sets. Although the initial concept concerning activity and selectivity came from the study of a small number of compounds in the adenylate cyclase assay (Fig. 7a), inclusion of additional compounds to test these hypotheses showed that activity was much more complicated than originally thought (Fig. 7b). Bulky substituents, such as adamantylethyl, were only slightly more selective than expected, but they were relatively nondepressive. It appeared that these compounds were too lipophilic or bulky for high activity. On the other hand, the cyclohexylethyl possessed a fairly good profile. The chemist (K. Untch) then proposed the shorter, bulky, less lipophilic *endo*-bicyclo[3.1.0]hexylethyl analog (**IX**), which fit all of the design criteria and was synthetically available:



(IX)

The side chain proved to be quite exceptional. As shown in Fig. 5, not only was **IX** extremely active in the dog model (pED_{50} = 7.01), but it was *equiactive* in the rat model, while all other analogs were totally uncorrelated in activity. This suggested that the *endo*-bicyclo[3.1.0]hexylethyl hit an exoreceptor near the β_1-adrenergic receptor in the heart quite strongly, and also that this exoreceptor was common to both dog and rat. Furthermore, the compound was about 45 times more selective for heart than lung β-receptors *in vivo*, and potency was 141 times a 20% depressive dose (i.e., comparing pED_{50}^{force} with $-pED_{-20}^{force}$). As shown in Table XI, the compound possessed a better profile than propranolol, practolol, acebutolol, and atenolol and was considerably better than (**I**: R = 2-Me-hexyl, R′ = *t*Bu) on all counts. The hypothesis seemed proved; however, **IX** was actually *more* active than expected.

Up to the time **IX** was produced, we had been working mainly with simple plots of parameter vs. lipophilicity and with a preliminary QSAR for activity and selectivity. It was quite apparent that subtle steric properties were responsible for both of these effects. Unfortunately, such effects as steric shape were not easy to parameterize. Verloop's (*4d*) steric constants were not yet available, and other common steric parameters such as

TABLE XI

OPTIMIZATION OF ALKYL CARBAMOYL SIDE CHAIN IN THIAZOLYLOXY BETA-BLOCKERS (I)[a]

	Dog[b]			Rat[c]		
	pED_{50}^{force}	$-pED_{-20}^{force}$	$-pED_{20}^{po}$	pED_{50}^{rate}	$-pED_{-20}^{rate}$	$-pED_{20}^{c}$
I; (R = 2-Me-hexyl, R′ = *t*Bu)	0.5	0.5	255	0.6	101	97
IX	2.3	≃ 3	362	31	v.lg.	v.lg.
Propranolol	(1.0)	(1.0)[d]	(1.0)	(1.0)	(1.0)	(1.0)
Practolol	0.2	> 4	173	0.5	v.lg.	206
Acebutolol	0.3	2	179	1.1	246	18
Atenolol	1.1	≃ 1[e]	7	1.4	3.4	74

[a] Relative to propranolol; po = pulmonary overpressure; c = dynamic compliance; molar comparisons.

[b] Vagotomized, anaesthetized, iv, isoproterenol and histamine challanges (Strosberg, *et al.*).

[c] Anaesthetized in plethysmograph, iv, isoproterenol and methacholine challanges (Weissberg, *et al.*).

[d] Difficult to compute due to sympathetic drive, at ED_{50}.

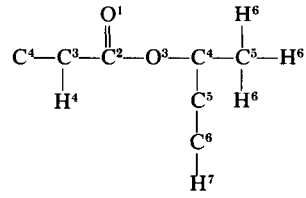
[e] No dose–response.

E_s and MR were of little use. We therefore borrowed the concept of Newman's six number (*67a, 4c*), which has been used (*67b*), for example, to study alkaline hydrolysis of alkyl acetates and the chemical shift of methyl protons in alkyl acetates (relative to methyl acetate):

$$\log k = 1.4 + 0.7\sigma^* + 0.6\,E_s^c + 0.05\,(\Delta 6\#)$$
$$n = 9 \quad s = 0.07 \quad r = 0.997 \tag{14}$$

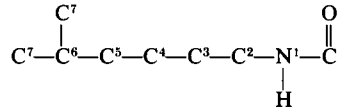
$$\text{rcps} = 0.9 + 20.4\sigma^* + 1.1\,(\text{C6}\#)$$
$$n = 18 \quad s = 0.66 \quad r = 0.943 \tag{15}$$

where σ^* is Taft's inductive parameter, E_s^c is Hancock's corrected steric parameter, $\Delta 6\#$ is the change in 6# between the left- and right-hand side of the molecule (see below) and C6# is the number of carbons in the sixth position. To define the 6#, the following numbering scheme is used:



We used this concept, however, only in its topological sense and renumbered the carbamoyl nitrogen as "one". The numbering scheme for

2-Me-hexyl is as follows:



The following types of descriptors are typical of those derived:

(C @ 5) #C at fifth position (=1)
(H @ 8) #H at eighth position (=0)
(C @ 8) #C at eighth position (=0)
(C @ ≧ 7) #C at seventh and higher position (=2)
[(C ≧ 6) ≧ 3] are #C at sixth and higher positions greater than or equal to 3? (=1, i.e., yes)

Together with measured HPLC log k' values (C18/35 w% CH_3CN/pH 7.0 buffer) these topological parameters were useful in obtaining preliminary QSAR. For example, lack of activity in the lung (dog, iv) at 1 mg/kg could be predicted by:

$$I = 0.06 + 0.14(\text{H @ 6}) \tag{16}$$

where $I = 1$ if no effect and 0 otherwise. This equation correctly predicted that the *endo*-bicyclo[3.1.0]hexylethyl side chain would be "inactive" in the lung, while the *endo*-bicyclo[3.1.0]hex-3-enylethyl side chain would be "active" in the lung. Since $0.5 > 0.06 + 0.14(\text{H@6})$, then $(\text{H@6}) < 3$ should give lung inactivity; the saturated analog has (H@6) = 4, while the unsaturated analog has (H@6) = 3. Experimentally, the former gave 24% vs. 42% for the latter at 1 mg/kg. This model was for (**I**; R′ = *t*Bu).

When we attempted to derive a full QSAR based on continuous variables, we obtained statistically significant equations; however, they did not have very high correlation coefficients and we felt that something was missing from our parameter set. Further, **IX** was underpredicted in activity. Recalling the excellent profile of this analog, we decided to *assume* that the *endo*-bicyclo[3.1.0]hexylethyl group was in fact the *ideal* shape and to try to parameterize all other groups according to their similarity to it. At the time, we did not have sophisiticated computer graphics or molecular modeling-programs available; thus we simply constructed Dreiding models of all 40 side chains and put them in "allowed" conformations which would fit into the putative cleft defined by the *endo*-bicyclo[3.1.0]hexylethyl group. We followed an objective protocol in doing this and did not change any parameters once they were determined. Figure 9 shows how the maximum distance D_{max}, the C-2–C-6 distance $D_{2,6}$, the depth D (including hydrogens), and the width (including hydrogens) were defined. We also counted numbers of carbons on the "face" of

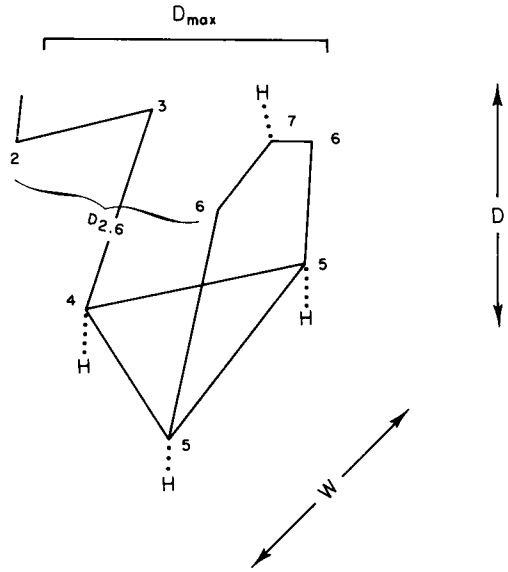


Fig. 9. *endo*-Bicyclo[3.1.0]hexylethyl group showing topological numbering scheme and various topographical parameters.

the cleft (C-5, C-6, C-7 area) in two modes, maximum contact (any carbons possibly in contact) and minimum contact (only those in intimate contact). Some substituents (e.g., alkynyl) would only fit into the cleft in a "down" conformation $D_{\downarrow}$ and we made the interesting observation that certain side chains that had shown high activity all appeared to have a point of unsaturation in the fourth position D" @ 4 (for the *endo*-bicyclo[3.1.0]hexylethyl this was the cyclopropyl moiety, while for others it was a double or triple bond, e.g., 2-Me-hex-3-ynyl or *cis*-hept-4-enyl).

With this extended set of geometric, topological, measured lipophilicity and fractional lipophilicity (*f*) (*4f*) constants, we were able to derive QSAR of higher precision and greater intuitive satisfaction. Results were all obtained by objective use of stepwise, ridge, and leaps and bounds regression techniques; no parameters were forced into any equation and all terms are significant at 95% by *t*-tests, except as noted.

For the guinea pig adenylate cyclase data, we obtained Eqs. (17) and (18):

$$-\mathrm{pID}_{50}^{\mathrm{lung}} = -4.5 + 0.4(\mathrm{C}\ @\ 5) + 0.9(\mathrm{C}\ @\ 8) - 0.5(\mathrm{H}\ @\ 8) + 0.1(C_{\mathrm{min}}^{\mathrm{face}}) - 0.5D_{\mathrm{cycle}}$$

$$n = 19 \quad s = 0.14 \quad r = 0.924 \quad F_{5,13} = 15.1 \qquad (17)$$

$$pID_{50}^{heart} = 4.9 + 0.4((C \geqq 6) \geqq 3) + 0.2(C\ @\ 7) - 0.1(C_{max}^{face}) - 0.2D_{=\equiv} - 0.5D_{iPr} + 0.6D_{iPr}D_{cycle}$$

$$n = 19 \quad s = 0.14 \quad r = 0.947 \quad F_{6,12} = 17.5 \tag{18}$$

$D_{cycle} = 1$ if cycloalkyl, $D_{=\equiv} = 1$ if double or triple bond, $D_{iPr} = 1$ if R′ = *i*Pr. C_{max}^{face} is significant at the 90% C.L.. Note that lipophilicity is not important in the *in vitro* data set, but that detailed topographical features are quite important. Lung activity is diminished by bulk along the face, particularly by carbons at C-5 and C-8. Heart activity is mainly increased by bulk at the end of the side chain (C_6 and higher), but the side chain should be cyclic rather than linear olefin or acetylene. This agrees with the hypothesis that lung selectivity is due to "short" substituents.

This pattern will be repeated in the *in vivo* data; however, there are species differences which complicate matters, as we have already seen in Figs. 5 and 6.

For the dog, iv, mode, we obtained:

$$pED_{50}^{force} = 5.8 + 0.4(D^{\cdot\cdot}\ @\ 4) - 0.4D_{\downarrow} + 0.3(\Sigma f_{\geqq 5}) + 0.2[(C \geqq 6) = 3] - 0.1(H\ @\ 4) - 0.1(f_{8,9}) - 0.4DW$$

$$n = 40 \quad s = 0.12 \quad r = 0.918 \quad F_{7,32} = 24.6 \tag{19}$$

$$-pED_{-20}^{force} = -4.4 - 1.4W + 0.4DW + 0.4D_{iPr}$$

$$n = 35 \quad s = 0.31 \quad r = 0.783 \quad F_{3,31} = 16.4 \tag{20}$$

$$-pED_{20}^{po} = -5.0 + 1.1(C\ @\ 8) + 0.2(H\ @\ 5) - 0.8 \log k' - 0.3[(H\ @\ 4) = 1]$$

$$n = 14 \quad s = 0.14 \quad r = 0.908 \quad F_{4,9} = 10.6 \tag{21}$$

Plots of the fit of these equations are given in Fig. 10. Note that intrinsic sympathomimetic activity.

Finally, the rat, iv, model data is correlated by:

$$pED_{50}^{rate} = 3.3 + 0.7(C_{max}^{face}) + 0.5D_{=\equiv}$$

$$n = 11 \quad s = 0.31 \quad r = 0.927 \quad F_{2,8} = 24.3 \tag{22}$$

$$-pED_{-20}^{rate} = 5.9 + 0.6(C\ @\ 6) - 5.8D_{2,6} - 3.6W - 2.5 \log k'$$

$$n = 8 \quad s = 0.24 \quad r = 0.940 \quad F_{4,3} = 5.66 \tag{23}$$

$$-pED_{20}^{c} = -5.4 + 0.7f_5 + 1.1D^{\cdot\cdot}\ @\ 4$$

$$n = 8 \quad s = 0.08 \quad r = 0.987 \quad F_{2,5} = 92.4 \tag{24}$$

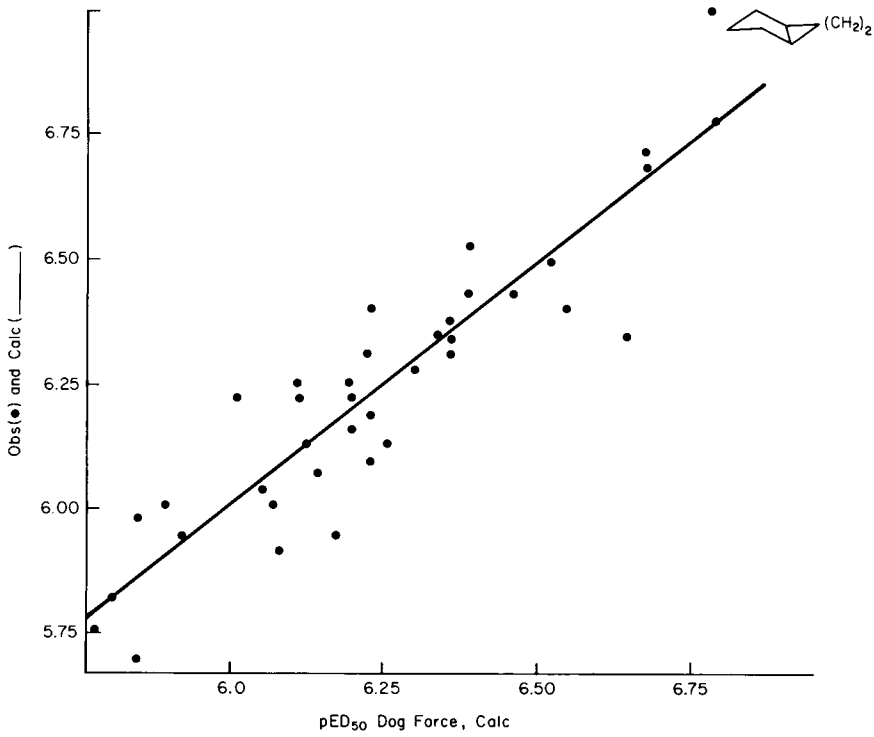


Fig. 10a. Plot of observed and calculated values according to Eq. (19), dog iv.

Only rate was determined in the rat model; c is dynamic compliance of lung after methacholine challenge. Activity is due to unsaturation at the fourth position, maximum contact with the face; however, the cross section (DW) should not be too large. As seen in the QSAR for depressivity [Eqs. (20) and (23)], depth but not width is desired, and hydrophilic substituents are less depressive. Lung activity is decreased by hydrophilic substituents with unsaturation at the fourth position and bulk at the fifth position.

Combining all of these objective QSAR models confirms our original hypothesis that short, bulky substituents would be selective and nondepressive. The additional activity of the special steric shape of the bicyclo[3.1.0]hexylethyl side chain is seen to be due mainly to the presence of unsaturation at the fourth position and to bulk along the face of the putative enzyme exoreceptor cleft. The quantitative model explains the activity of a wide range of linear and (poly)cyclic alkyl, alkenyl, and alkynyl

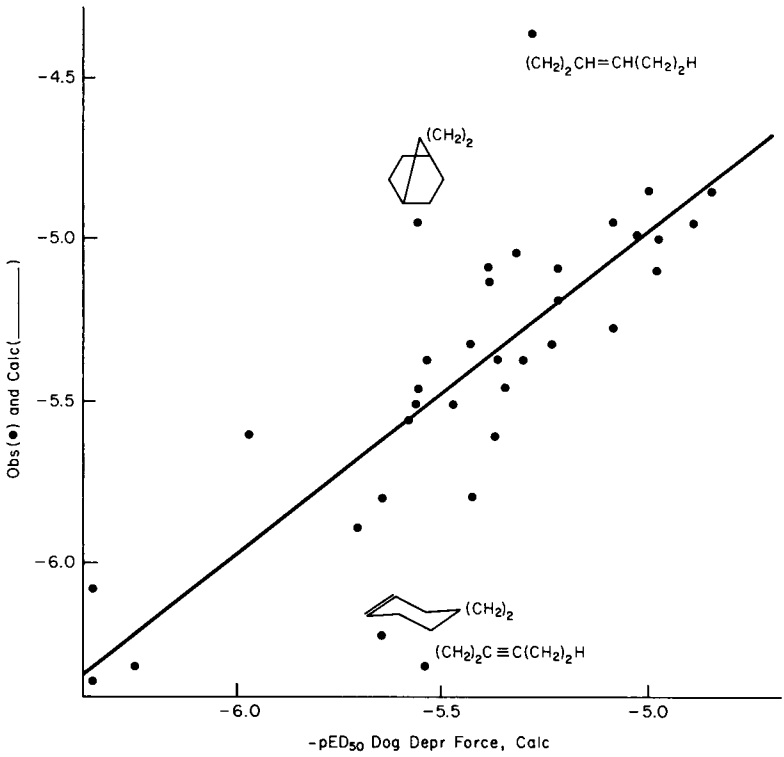


Fig. 10b. Plot of observed and calculated values according to Eq. (20), dog iv.

groups as well as **IX.** Lipophilicity has been shown to be relatively unimportant, possibly because the heart is directly attached to the site of dosage, i.e., intravenous injection. Only rat depressivity and dog lung activity is sensitive to lipophilicity. The topographical effects are emphasized in Table XII, which compares the *endo* and *exo* analogs of **IX.**

In this example, QSAR has been useful in a supportive role; however, the final models have allowed prediction of new side chains and have provided considerable understanding of the nature of the β-adrenergic receptors. It can be seen that this model agrees in general with some of the aspects of Davies' model [c.f., Fig. 11; (*68*)] but delineates the nature of the region of "steric freedom" near the amide group. For example, it had long been apparent to the team that substitution at C-1 or C-2 brought about a considerable decrease in activity.

One of the major difficulties in working with the series was that bulky alkyl groups have not been parameterized. Neither steric nor lipophilic constants exist for adamantylethyl or norbornylethyl, for example. We

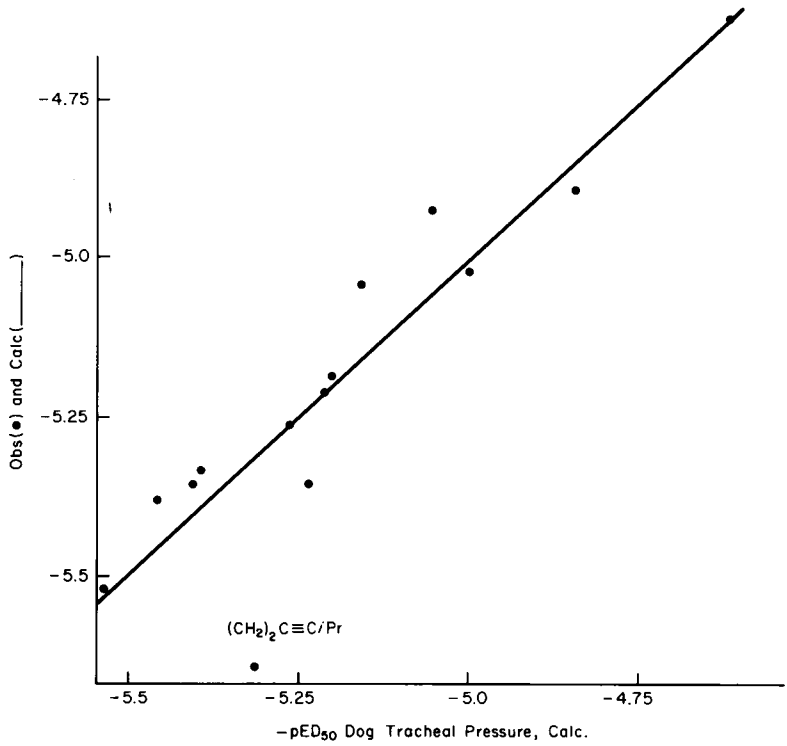
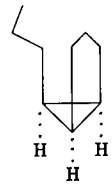
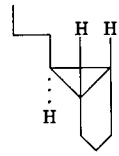


Fig. 10c. Plot of observed and calculated values according to Eq. (21), dog iv.

therefore measured all compounds by HPLC using C18-Corasil and 35w% CH_3CN in pH 7.0 buffer. Since the series differs only in hydrocarbon fragments, this was deemed satisfactory (*22*). While looking for a more hydrophilic, short, bulky side chain, the chemists prepared a number of benzamides RNHC(O)Ph for lipophilicity determination. Having established a correlation between benzamides and the corresponding beta-blocker (r = 0.98), we were therefore able to reduce somewhat the synthetic effort involved. The beta-blockers could not be studied by the octanol HPLC procedure (*22*) due to very poor peak shapes. Table XIII gives some classical physicochemical constants for propranolol (**I**; R = 2-Me-hexyl, R′ = tBu) and **IX**. Note that **IX** is isolipophilic with propranolol as the free base.

Finally, we present data in Table XIV for a number of analogs of **IX** in which the thiazole and carbamoyl groups have been replaced. This further supports the contention that the *endo*-bicyclo[3.1.0]hexylethyl side chain is near optimal (*79*).

TABLE XII
STERIC EFFECTS ON BETA-BLOCKER PROFILE OF I

R	Log k'	Dog (iv)			Enzyme		Rat (iv)		
		pED^{E}_{50}	$-pED^{E}_{-20}$	$-pED^{HR}_{20}$	pID^{H}_{50}	$-pID^{E}_{50}$	pED^{R}_{50}	$-pED^{R}_{-20}$	$-pED^{C}_{20}$[a]
endo-Bicyclo[3.1.0]hexylethyl	.54	7.01	−5.02	−5.35	5.37	−4.00	7.09	> −5.6	> −5.6
									
exo-Bicyclo[3.1.0]hexylethyl	.56	6.13	−4.92	> −5.0	5.04	−3.30	6.79	−4.30	−3.71
									
Fold Comparison:		7.6	−1.3		2.1	−5.0	2.0		

[a] Dynamic compliance is measured in rat model since rat is in plethysmograph.

Compound **IX** was selected for clinical evaluation as a nondepressive, cardioselective beta-blocker.

V. Drug Designer as "Interlocutor"

Because of the central nature of QSAR studies (see Fig. 1), a drug designer is sometimes regarded as the most readily available person to interact between groups. The drug designer tends to know many people throughout the company and can often direct information, literature, or questions to the proper person and can, in some ways, expedite the scientific work of the company.

Since QSAR implies computer work, there is also the possibility of becoming involved with setting up data systems for one's own or particular scientists' use, or for consulting on company-wide data retrieval systems. For example, a number of chemists have discussed the possibility of retrieving reaction conditions (offered by Derwent Publications), searching the Aldrich catalogue (offered by Aldrich), computer aided synthesis (cf., Corey and Wipke, *69*), and establishing "personal" data bases. We have established a full information retrieval system for the mouse screen

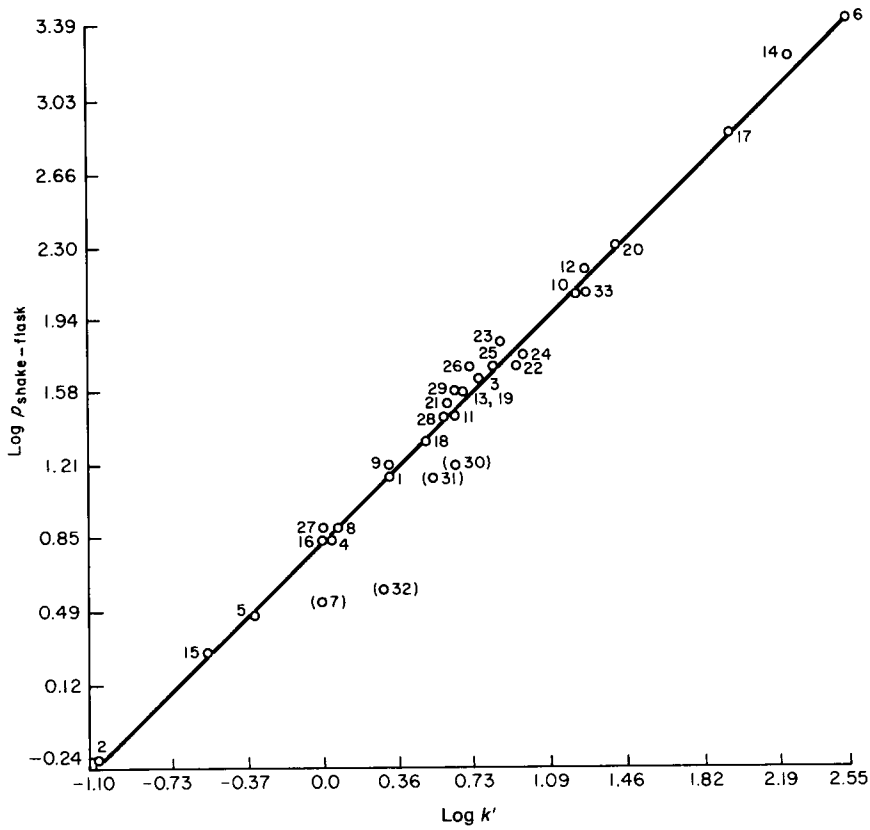
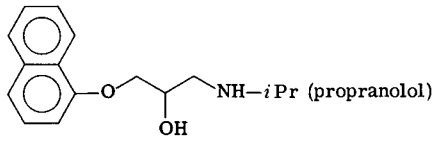
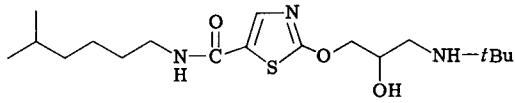
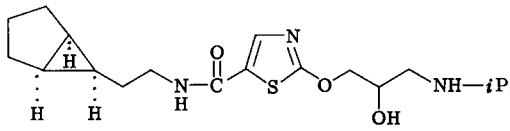


Fig. 11. Comparison of $\log P_{\text{shake-flask}}$ with $\log k'$ determined by octanol C18 procedure. All compounds related to 50-cm column by regressions using overlapping series. Note the excellent agreement over 3.5 log P units, except for unhindered pyridines, which can be attributed to specific binding to residual free silinol sites. Correlation is for Eq. (26). Mnemonics are: 1 = acetanilide; 2 = acetone; 3 = acetophenone; 4 = 2-acetylpyridine; 5 = 4-acetylpyridine; 6 = acridine; 7 = 2-amino-4-picoline; 8 = aniline; 9 = *o*-anisidine; 10 = anisole; 11 = benzaldehyde; 12 = benzene; 13 = benzonitrile; 14 = benzophenone; 15 = 2-butanone; 16 = catechol; 17 = chlorobenzene; 18 = 2 = chloropyridine; 19 = 4-cyanophenol; 20 = *N,N*-dimethylaniline; 21 = m-dinitrobenzene; 22 = 2-ethylpyridine; 23 = 4-fluorophenol; 24 = hexaldehyde; 25 = 2,6-lutidine; 26 = 4-nitrophenol; 27 = 2-pentanone; 28 = phenol; 29 = phenylacetonitrile; 30 = 3-picoline; 31 = 4-picoline; 32 = pyridine; 33 = quinoline.

data (see Section III) using the INFO system on our APL language computer service,* and we have consulted on establishing a computer-based retrieval system for our in-house compound file.

* Proprietary Computer Systems, Inc., Van Nuys, California.

TABLE XIII
PHYSICOCHEMICAL PROPERTIES OF BETA-BLOCKERS

	Log k' [a]	Log P_T [b]	pK_a^c
	0.48	3.44	9.34 ± 0.03
	0.79	4.18	9.49 ± 0.03
	0.54	3.48	9.29 ± 0.03

[a] Log k' = persilated C18-Corasil/35 w% CH_3CN/pH = 7.00.
[b] Log P_T = log P_{True}.
[c] pK_a = 25 + 0.5.

VI. Methodology

A. HPLC Partition Coefficients

Three systems have been proposed for calculating octanol–water partition coefficients. Fujita, Iwasa, and Hansch (*2b*) based the original system on replacement constants:

$$\pi_X = \log P_{RX} - \log P_{RH}$$

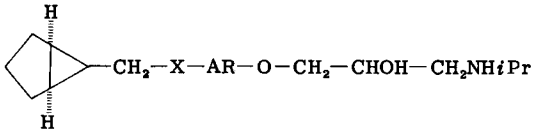
Rekker (*4e*) based his fragment constants on a statistical analysis of a collection of data

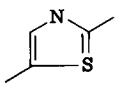
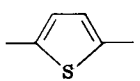
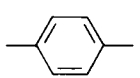
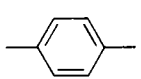
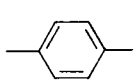
$$\log P = \Sigma f_i$$

and a similar system was proposed by Leo, Jow, Silipo, and Hansch (*4f*) using fragment constants derived from measurements on important fundamental molecules (H_2, CH_4, C_2H_4, etc.). Whereas these methods all give reasonably good results for most normal molecules, they often require a plethora of special interaction constants and factors to make the more unusual molecules come out correctly. The best method for determining lipophilicities is still to measure the value experimentally, assum-

TABLE XIV

ANALOGS OF IX IN DOG, IV MODEL



X	AR	pED^{f}_{50}	$-pED^{f}_{-20}$	$-pED^{po}_{20}$
CH_2NHCO	 (IX)	7.01	−5.02	−5.35
R = 2-Me-hexyl,	R′ = *t*Bu (I)	6.29	−5.45	−5.21
CH_2NHCO		6.49	−5.10	−5.32
CH_2NHCO		6.76	−4.80	−5.59
CH_2CONH		7.06	−4.80	−6.05
NHCONH		7.11	> −5.0	−6.06

ing sufficient compound is available. Unfortunately, this is a completely impractical task for anyone working in the pharmaceutical industry because of the large number of compounds produced in a typical project and the tediousness of the shake-flask procedure.

Considerable interest has been shown in any procedure which promises enhanced throughput. The first of these was reversed-phase TLC (*70, 71a,b*). Standardization of plates to obtain daily reproducibility and the necessity of running solvent series to obtain extrapolation to 100% aqueous values (which also detects hydrogen-bonding), however, makes this procedure of limited practical use.

The availability of chemically bonded octadecylsilane C18 (*72a*) for HPLC sparked the next wave of interest. However, C18 is *alkanelike* and not *alcohollike*. Use of mixed solvents such as 30% methanol–water or acetonitrile–water further clouds the picture.

The next proposal (*73a–c*) seemed very close to the mark: Coat silica gel HPLC columns with octanol and use octanol-saturated buffers. Unfortunately, this is not a very stable system, and in practice, it is very difficult to obtain stable baselines. We therefore thought that if octanol could be coated on C18 columns, we might obtain a stable system, since octanol itself is rather lipophilic ($\log P = 3.15$). The C18 packing would therefore act as a reservoir system for the octanol. This simple combination of ideas indeed gave a very reliable, reproducible, and useful system that has allowed us to make determinations, of $\log P$, for hundreds of compounds per year (*22*). [It turns out that similar systems using pentanol on C2, C8, and C18 had been proposed by Wahlund (*74a–f*), but for analytical separations in order to study ion-pair chromatography.]

The method, in brief, is to persilate (*72b*) C18 Corasil, 37–50 microns, and dry pack in stainless steel tubing 2 mm i.d. in various lengths (e.g., 3, 10, and 50 cm). The column is rapidly coated by flushing with octanol-saturated water or buffer (all solutions are kept at 0.15 ionic strength, and passed through millipore filters to reduce possible contamination or clogging of the column). An HPLC with variable wavelength UV-detector, constant-flow pump, and syringe injection are most useful. Samples are dissolved in buffer, methanol, or DMF, the latter two serving as unretained markers for determining t_0, the dead time. Elution times (t_x) are measured with a stopwatch (or by data system). Then, $k' = (t_x - t_0)/t_0$. Sample concentrations are adjusted so that relative peak areas remain approximately constant. Column length and flow rate (ca. 1–3 ml/min) are selected for analytical convenience. Experiments are performed at ambient temperature (25 ± 1°C) and the pH of eluate is determined.

Daily standards of known $\log P$ or $\log D$ (α is degree of ionization) are run and the regression between their $\log k'$ and

$$D = P\,(1 - \alpha) \tag{25}$$

log P or log D values is used to convert unknowns. Figure 11 shows the correlation obtained for 33 standards (for this correlation all log k' were related by overlapping series to the data on the 50-cm columns). The correlation is given in Eq. (26), omitting

$$\log P = 1.01(\pm 0.02)\log k' + 0.857(\pm 0.02)$$
$$n = 27 \quad s = 0.04 \quad r = 0.999 \quad F_{1,25} = 10{,}533 \tag{26}$$

small, unhindered, basic pyridines which deviate, presumably, due to binding to residual silinol sites. One must be careful to work in idential solvents for both shake-flask and HPLC when comparing the two methods, especially with ionizable compounds. Some shake-flask and HPLC comparisons are given in Table XV.

TABLE XV
AGREEMENT BETWEEN CLASSICAL AND HPLC LOG P AND pK_a

Compound	pK_a^c	pK_a^{ref}	log P	log P^{ref}
VI; X = H			1.95 ± 0.03[a]	1.93 ± 0.01[a]
VI; X = OMe			1.94 ± 0.03[a]	1.92 ± 0.02[a]
VI; X = CN			1.62 ± 0.03[a]	1.58 ± 0.01[a]
VI; X = MeCOO			1.32 ± 0.03[a]	1.40 ± 0.01[a]
II	4.28 ± 0.02	4.53 ± 0.06[c,e]	3.21 ± 0.01[c]	3.18[b]
	4.21 ± 0.02[d]	4.39 ± 0.02[d,e]	3.20 ± 0.01[d]	
III	3.83 ± 0.02	3.85[c,f]	3.00 ± 0.01[c,d]	3.13[c,g]
	3.80 ± 0.02[d]			
IV; X = H	3.52 ± 0.01	3.54[c,e]		
	3.49 ± 0.02[d]		2.72 ± 0.02[d]	
Thiophenoyl-(IV)	3.51 ± 0.03	3.5[c,e]3.36[f]		
	3.52 ± 0.01[d]		2.64 ± 0.01[d]	(2.57 @ pH = 1.00)
Benzoic acid	4.33 ± 0.02	4.20[c,h]	1.78 ± 0.01	1.87[b]
	4.38 ± 0.03[d]	4.18[h]	1.77 ± 0.01[d]	
Salicylic acid	3.52 ± 0.03	3.00[h]	2.00 ± 0.01	2.23 ± 0.03[b]
	3.29 ± 0.03[d]		2.18 ± 0.01[d]	
p-Toluic acid	4.30 ± 0.09	4.37[h]	2.22 ± 0.02	2.27[b]
	4.41 ± 0.01[d]		2.26 ± 0.01[d]	
X			1.15 ± 0.01	1.10 ± .06 @ pH = 5.82[c]
			3.05 ± 0.01	3.07 ± 0.17 @ pH = 2.42[c]

[a] pH = 7.00 (0.01M) phosphate buffer; ambient temperature.
[b] Ref. (4b).
[c] μ = 0.1.
[d] μ = 0.01.
[e] Spectrophotometric (D. Johnson).
[f] Solubility (J. Kent, M. Yost).
[g] J. Kent and M. Yost.
[h] G. Kortüm, W. Vogel, K. Andrussow, "Dissociation Constants of Organic Acids in Aqueous Solution" Butterworths, London, 1961.

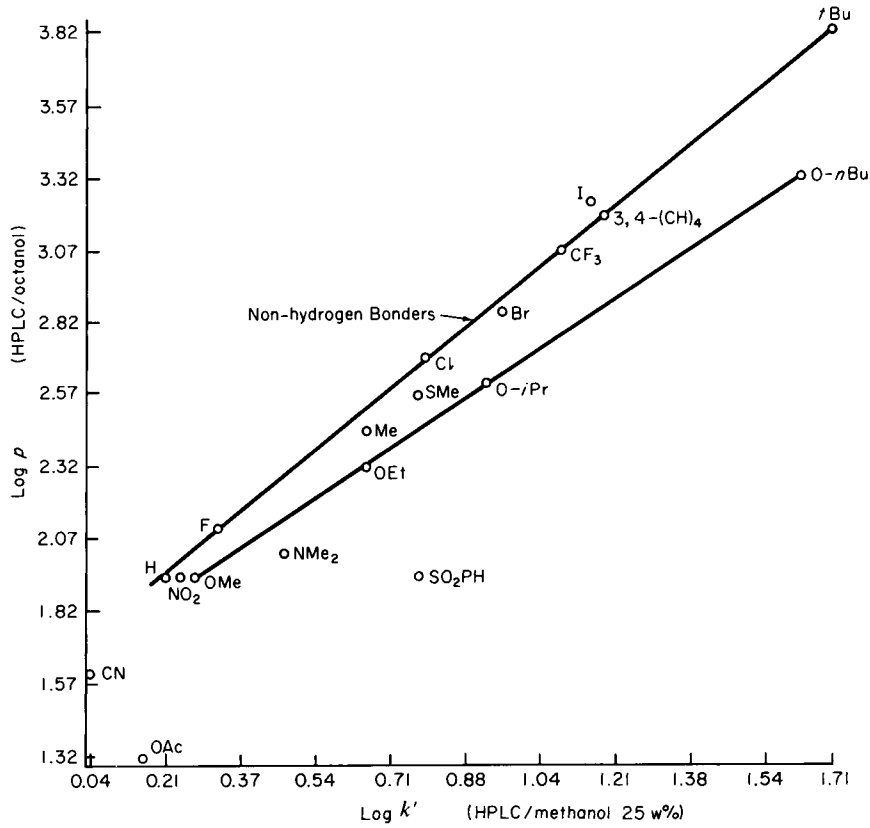


Fig. 12. Partitioning behavior of tuberins (**VI**) in two HPLC systems. Vertical axis gives log P determined on octanol-coated C18 using octanol saturated pH 7.00 (0.01M) phosphate buffer as mobile phase. Horizontal axis gives log k' determined on the same C18 columns, but without octanol, using 25 w% methanol in pH 7.00 (0.001M) phosphate buffer. Note the deviation of hydrogen-bonded substituents between the two systems.

The difference between alkane and alcohol partitioning can be seen in Fig. 12, in which a series of tuberins (**VI**) (*21*) has been studied on the same column, both in the absence and presence of octanol. Nonhydrogen-bonding substituents fall on a line which differs from that of hydrogen-bonding substituents. Therefore, C18 is not a good model for octanol partitioning, which is generally considered a better model for average biological tissue (*3*). (However, in the QSAR study of **VI**, we concluded that alkane partitioning may be a better model for highly lipophilic bacteria, such as mycobacteria.)

The agreement between shake-flask and HPLC procedures is excellent over the approximately 4 log P unit range examined so far.

A further advantage of the HPLC procedure is that it allows determination of apparent pK_a and the detection of ion-pairing under the same conditions as log P is determined. Equation (25) is valid only if the neutral species partitions exclusively into the octanol phase at all pH values. It is easily shown (*22*) that:

$$D = P - (D/H)K_a \qquad \text{acid } H \geqq K_a \tag{27a}$$

$$D = P - (DH)/K_a^b \qquad \text{base } H \leqq K_a^b \tag{27b}$$

the values of H are the eluant $[H^+]$ and D is obtained by extrapolation from the standards curve. P and K are then obtained by ordinary least squares analysis of D vs either DH or D/H for bases or acids, respectively. The log forms of Eq. (27) may be fit by nonlinear least squares and are valid over a somewhat wider pH range:

$$\log D = \log P - \log[(K_a/H) + 1] \qquad \text{acid} \tag{28a}$$

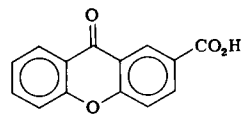
$$\log D = \log P - \log[(H/K_a^b) + 1] \qquad \text{base} \tag{28b}$$

Horváth (*75*) and Wahlund (*74*) have derived equations for the case of both ionized and neutral species partitioning into the lipoid phase, and also for multiple ionizations or zwitterionic partitioning. In the case of ion-pair partitioning (which occurs mainly with the more lipophilic acids and bases), the relevant equations (*75a,b*) are (P' is partition coefficient for ion pair):

$$\log D = \log[P + P'(K_a/H)] - \log[1 + (K_a/H)] \qquad \text{acid} \tag{29a}$$

$$\log D = \log[P + P'(H/K_a^b)] - \log[1 + (H/K_a^b)] \qquad \text{base} \tag{29b}$$

An example of ion-pair partitioning provided by xanthone-2-carboxylic acid (**X**) which is the parent of a series of 7-substituted-xanthone-2-



(X)

carboxylic acids studied as mast cell inhibitors (*76a*), and also for inhibition of phosphodiesterase in several tissue systems (*76b,c*). The data collected at 7 pH values in the pH operating range of the C18 column was well fit to the standard model [Eq. (27a)] with $r = 0.990$ and, in fact, gave insignificant P' when fitted to Eq. (29a). However, we noted that the

points at the most basic pH's deviated systematically. Additional points were taken at still more basic pH (7.32 and 9.42) by classical shake-flask procedures in order to better define the "tail". The data could then be fit very well to Eq. (29a). A comparison of the two models, with and without ion-pair partitioning, is shown in Fig. 13. The derived constants are

$$\log P = 3.12 \pm 0.07$$

$$\log P' = -0.21 \pm 0.07 \qquad n = 9 \quad s = 0.086 \quad r = 0.998$$

$$pK_a = 3.73 \pm 0.10$$

The data in Table XVI for 18 xanthones (X) was collected in about a month and illustrates both the advantages and disadvantages of the HPLC method. For most of the analogs, less than 1 mg of material was used,

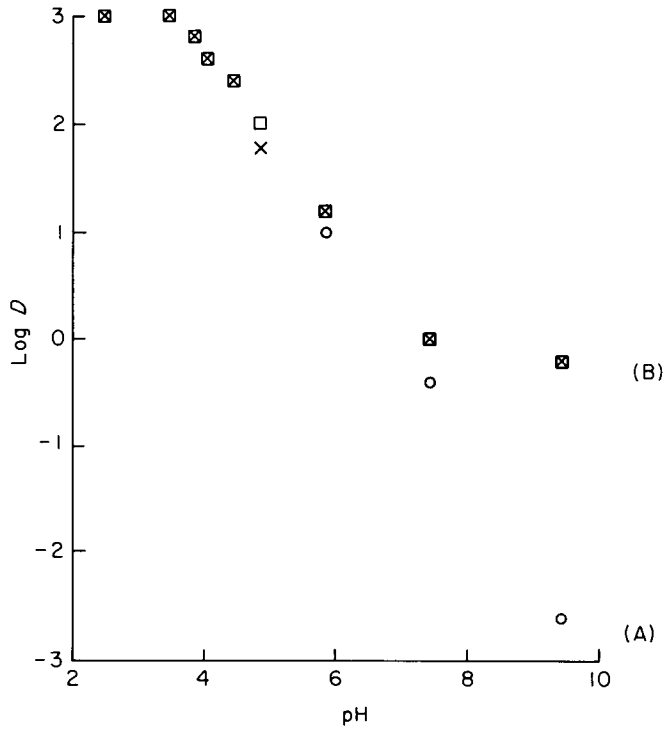


Fig. 13. Comparison of neutral species partitioning (A) with neutral plus anion partitioning (B) for an acid (X) with $pK_a = 3.73$, $\log P = 3.12$ and $\log P' = -0.21$. These constants were obtained by nonlinear fit of Eq. (29a). All points determined by HPLC, except pH = 7.32 and 9.42, which were determined by classical shake-flask.

TABLE XVI
PHYSICOCHEMICAL DATA ON *X* DETERMINED BY HPLC TECHNIQUES

X	log P^d	log P'^h	(*X*) $pK_a{}^i$	n^a	s^b
H	3.12	−0.21	3.73	9[c]	0.086
Me	3.67	0.70	3.77	8	0.052
Et	4.25	1.60	3.75	6	0.133
Pr	(4.65)	(2.15)	(3.73)	2	
*i*Pr	(4.65)	(2.02)	(3.73)	2	
*i*Bu	(5.07)	(2.53)	(3.73)	1	
Am	(5.61)	(3.29)	(3.73)	1	
OH	3.12	0.68	3.67	7	0.074
O*i*Pr	4.32	1.56	3.67	6	0.065
$OCH_2CH{=}CH_2$	3.56	0.56	3.81[e]	8	0.056
SMe	4.19	0.95	3.44[e]	6	0.065
SOMe	(1.70)[f]	(−0.50)[f]	(3.49)[e,f]	5[c]	0.069
Br	3.94	1.38	3.71	7	0.081
CF_3	3.93	1.38	3.77	6	0.136
COMe	2.78	0.44	3.51[e]	7	0.067
C (OH) Me	2.70	0.14	3.60	7	0.074
C (OMe) Me	3.42	0.11	3.75	8	0.098
(—$SCH_2CH_2CH_2CH$—) |	(5.34)[g]	(2.98)	(3.73)	1	

[a] Number of pH values.
[b] Standard deviation of fit to Eq. (29a).
[c] Shake-flask values added.
[d] ( ) from Eq. (30).
[e] $p > 95\%$ differs from 3.73.
[f] See text.
[g] Since log *D* is average of *i*Bu and Am, log *P* was set equal to average of log *P* values for *i*Bu and Am.
[h] ( ) from log *D*, log *P* and pK_a in Eq. (29a), rearranged.
[i] ( ) from 3.73 ± 0.10 for 9 neutral substituents.

compared to about 3 mg of sample per pH value for shake-flask determinations. Because of the pH limitations of the column (pH = 2 to 7.5), the data had to be supplemented by shake-flask determinations at certain points. For example, X = SOMe was too hydrophilic to be studied at pH values above the pK_a with the longest column (50 cm) then available. With X = H, as mentioned above, we could not obtain a well defined tail (Fig. 13) below pH 7.5. The very lipophilic analogs could only be studied at pH > pK_a and log *P* could therefore not be estimated directly. In order to fill in the table, we made the following observations and approximations. The pK_a values for nine neutral substituents were essentially constant at

3.73 ± 0.10; only X = $OCH_2CH{=}CH_2$, SMe, and COMe gave significant deviations. X = SOMe had not yet been evaluated for pK_a by HPLC; however, its pK_a by aqueous titration (M. Alton) was 3.72 ± 0.02. pK_a and log P for octanol–water were assigned as described below.

In order to assign log P values to the more lipophilic compounds, we noted the following correlation for X = H, Me, Et, OH, *i*PrO, CF_3, COMe, SMe, and Br:

$$\pi = 0.15(\pm 0.21) + 0.85(\pm 0.27)\pi_\Phi - 0.81(\pm 0.53)\mathcal{R}$$
$$n = 9 \quad s = 0.192 \quad r = 0.956 \quad F_{2,6} = 31.88 \tag{30}$$

where $\pi = \log P_X - \log P_{X=H}$ and π_Φ is the phenyl-based Hansch constant (Table VIII). This allows calculation of log P for Pr, *i*Pr, *i*Bu, and Am. Since log D for tetrahydrothiophenyl was the average of the *i*Bu and Am values, log P for this substituent was taken as the average of the log P values of *i*Bu and Am. Log P for X = SOMe was obtained by combining the shake-flask value at pH 1.18 (log D = 1.76), with HPLC values at pH 2.42, 2.85, 3.51, and 3.74 (log D = 1.63, 1.58, 1.35, 1.30). Since all points were taken at $[H] \geqq K_a$, the data was fit to Eq. (28a) by nonlinear least squares. This gave log P = 1.70 ± 0.05 and pK_a = 3.49 ± 0.12 (n = 5, s = 0.07).

The log P' value for X = SOMe was estimated from the following correlation:

$$\log P' = 0.19(\pm 0.35) + 1.18(\pm 0.40)\pi_\Phi + 1.17(\pm 0.62)D_{HA}$$
$$n = 12 \quad s = 0.258 \quad r = 0.923 \quad F_{2,9} = 25.88 \tag{31}$$

Note that log P' for the ion pair is a function of lipophilicity and hydrogen-bond acceptor properties, while log P [Eq. (30)] is a function of lipophilicity and resonance effects. Thus, one would expect that log P' would not be very highly correlated with log P. This is the case, since for the same 12 analogs, s = 0.361 and r = 0.824 for a correlation between log P' and log P. In other words, one cannot predict log P' from log P very accurately. Since an ion pair is by definition ionized, it is reasonable that electronic effects or pK_a does not influence lipophilicity and also that hydrogen-bonding properties of X do influence lipophilicity beyond that of their normal π contribution. The $\mathcal{R}$ effect in the free acid may be thought of as influencing electron density at the *para* oxygen, which in turn may influence the *para* carboxy group (since pK_a is sensitive to large $\mathcal{R}$ effects, this is plausible).

In another case, we studied a lipophilic (log P = 3.2) antihypertensive agent with several sites of protonation. The data was not well fit by Horváth's equation for dibasic compounds; however, ion-pair partitioning

could be detected by curvefitting and by the observation that buffers of constant ionic strength with different anions gave different log D values. Log $P' = 1.9$ for the phosphate ion pair, which is only 1.3 log units drop upon ionization, compared to a 3.9 unit drop for acridine (*80*).

In determining log P and pK_a with a buffer series, one must be very careful to check for ion-pair partitioning and to verify that the particular buffer does not have an untoward effect. Since one can only adjust column length, flow, and pH, one must stick to a log D range of about -0.5 to 3.5. Some of these limitations may be overcome by packing less material in the smallest column (for more lipophilic compounds) by using less lipophilic alcohols such as pentanol or by using C8 reversed-phase packing materials. Very long columns may help extend the range to more hydrophilic columns, but these must be packed with care.

B. COMPUTER PROGRAMMING IN APL

Computer work is performed on either an in-house IBM 370/158 computer (used mainly for nonlinear curvefitting and listing of tapes) or a commercial APL language (*53a*) time-sharing system.* Lewi (*8*) has described an APL language system applied to drug design. Our APL system provides access to high-level APL-SV language enhancements (*53b*), to many specialized programs (e.g., data base management), and to a remote batch facility through which we run BMD programs (*13*) and other FORTRAN programs. In addition, we have access to the NIH–EPA–CIS (National Institutes of Health–Environmental Protection Agency–Chemical Information System) system for access to a number of special data bases and to the CAMSEQ molecular mechanics program of Hopfinger.

APL is a very convenient language for drug design work because it allows the user maximum creativity and rapid data manipulation, rather than burdensome computer programming. APL is vector and matrix oriented and is thus ideally suited to drug design (*8*). Below we describe briefly some of our more unique APL programs, which are listed in Fig. 14. The two functions, NIN and CIN, are special programs* that allow numeric or character input on the same line as the question prompt following their use. They may be replaced by any convenient data entry statement.

The three experimental design programs STEPUP [Eq. (8)], STEPDOWN [Eq. (9)], and SWAP [Eq. (10)] use an existing matrix EA and a design matrix of candidates DA. STEPUP augments the EA with the case which maximizes det $\mathbf{A}_+$ and cycles until the final

* Proprietary Computer Systems, Inc., Van Nuys, California.

```
    ∇ Z←DA STEPUP EA;NUM;LIST;DAA;D;S
 [1]  NUM←NIN,'FINAL NUMBER OF POINTS?:'
 [2]  LIST←⍳1↑⍴DAA←DA←1,DA
 [3]  EA←1,EA
 [4] LL:IEA←⌹(⍉EA)+.×EA
 [5]  D←+/DA×⍉IEA+.×⍉DA
 [6]  DA←DAA[LIST←(LIST≠S←⊃⍋LIST[D⍳⌊/D])/LIST;]
 [7]  EA←EA,[1] DAA[S;]
 [8]  →LL×⍳(0≠⍴LIST)∧(NUM>1↑⍴EA)
 [9]  Z←((⍳1↓⍴EA)≠1)/EA
    ∇
    ∇ Z←STEPDOWN EA;NUM;F;LIST;EAA;S;D;SD
 [1]  NUM←NIN,'FINAL NUMBER OF POINTS?(>',(⍕1↓⍴EA←1,EA),'):'
 [2]  F←NIN,'FREEZE POINTS NUMBER:'
 [3]  LIST←⍳1↑⍴EAA←EA
 [4] LL:D←+/EA×⍉(⌹(⍉EA)+.×EA)+.×⍉EA
 [5] L2:→L1×⍳0=+/F∈S←⊃⍋LIST[SD←D⍳⌊/D]
 [6]  (⍕S),' NOT ELIM., FROZEN'
 [7]  D[SD]←1000
 [8]  →L2
 [9] L1:EA←EAA[LIST←(LIST≠S)/LIST;]
[10]  →LL×⍳(NUM<⍴LIST)
[11]  Z←((⍳1↓⍴EA)≠1)/EA
[12]  LIST
    ∇


    ∇ Z←DA SWAP EA;I;IEA;J;D;JM;IM;A;B;C;DEA;Y;ID;F
 [1]  F←NIN,'FREEZE POINTS NUMBER:'
 [2]  DA←1,DA
 [3]  EA←1,EA
 [4]  I←DEA←1
 [5]  JM←1↑⍴EA
 [6]  IM←1↑⍴DA
 [7]  ID←DET(⍉EA)+.×EA
 [8] LL:IEA←⌹(⍉EA)+.×EA
 [9]  J←1
[10]  D←⍳0
[11] L:A←1+DA[I;]+.×IEA+.×DA[I;]
[12] L1:B←1-EA[J;]+.×IEA+.×⍉EA[J;]
[13]  C←⍳EA[J;]+.×IEA+.×DA[I;])*2
[14]  D←D,,(A×B)+C
[15]  →L1×⍳JM≥J←J+1
[16] L6:→L3×⍳0=1<D[Y←D⍳⌊/D]
[17]  →L5×⍳0=+/F∈Y
[18]  (⍕Y),' NOT SWAPPED WITH ',(⍕I),' ...FROZEN'
[19]  D[Y]←0
[20]  →L6
[21] L5:DEA←DEA×D[Y]
[22]  'REPLACE  ',(⍕Y),'  WITH  ',(⍕I),' : CRITERION = ',⍕D[Y]
[23]  EA[Y;]←DA[I;]
[24] L3:→LL×⍳IM≥I←I+1
[25] L4:Z←((⍳1↓⍴EA)≠1)/EA
[26]  'INITIAL DET = ',⍕ID
[27]  'FINAL DET = ',⍕DEA×ID
[28]  'FACTOR = ', 15 3 ⍕DEA
    ∇
```

Fig. 14. APL programs.

```
    ∇ Z←MXSNT1 M
[1]   Z←(COLAV M)⍪(⍴M)⍴STDEV M
    ∇


    ∇ Z←STDEV M
[1]   Z←((+⌿(COLAV M)*2)÷¯1+1↑⍴M)*0.5
    ∇


    ∇ Z←COLAV M
[1]   Z←M-(⍴M)⍴(+⌿M)÷1↑⍴M
    ∇


    ∇ IP←DISTPLAN M;N;D;DMD;DCG;IPP
 [1] ⍝WELLCOME
 [2]  →L×⍳'N'=1↑,CIN 'STANDARDIZE?'
 [3]  M←MXSNT1 M
 [4] L:IP←,NIN 'INITIAL POINTS?'
 [5]  N←1↑⍴M
 [6]  'AVE D:',▼(+/,D)÷+/,(D←RD2 M)≠0
 [7]  DMD←D≥,NIN 'CUTOFF D?'
 [8]  DMD[;IP]←0
 [9] LL:DCG←+/(M-(⍴M)⍴(+⌿M[IP;])÷⍴IP)*2
[10]  →0×⍳N<IPP←DCG⍳⌊/(∧⌿DMD[IP;])/DCG
[11]  IP←IP,IPP
[12]  DMD[;IPP]←0
[13]  →LL×⍳N≥⍴IP
    ∇


    ∇ TOM M;P;N;R;RII;RIIT;RIJ;U;V
[1]   'OMEGA:'
[2]   (¯1+RIIT← 1 1 ⍉RII←⊞R←CM M)×(V←(N←1↑⍴M)-P)÷U←¯1+P←1↓⍴M
[3]   'F WITH ',(▼U,V),' DOF'
[4]   'R*2:'
[5]   R*2
[6]   RIJ←-RII÷(RIIT∘.×⍉RIIT)*0.5
[7]   'T, DIAG.UNDEF., T(',(▼V),') = ',(▼TM V),':'
[8]   (RIJ×(N-P)*0.5)÷(((⍳P)∘.=⍳P)+1-R*2)*0.5
    ∇


    ∇ R←CM X;V
[1]   R←R÷(V∘.×V←(1 1)⍉R←(⍉R)+.×R←X-(⍴X)⍴(+⌿X)÷(⍴X)[1])*0.5
    ∇


    ∇ Z←TM N
[1]   Z←*(N∘.*- 0 1 2 3)+.× 0.673019 1.200524 0.8139429 ¯0.1453636
    ∇
```

Fig. 14. (*continued*)

```
    ∇ CLU M;MN;D;S;AV;ZZ;NAM;MS;I;ML;MST;N;MM
 [1]  'ROWS (LOWER SUPERCEDES, REPLACED BY AVERAGE)'
 [2]  NAM←⍳N←1↑⍴M
 [3]  ZZ←⍳0
 [4]  MM←(1,N)⍴0
 [5]  MST←ML←MN←2↑MINIDX D←RD2 M←MXSNT1 M
 [6] L:M[1↑MN;]←AV←(+⌿M[MN;])÷2
 [7]  D[;1↑MN]←D[1↑MN;]←+/(M-(⍴M)⍴AV)*2
 [8]  S←(1↑⍴D)⍴1
 [9]  S[1↓MN]←0
[10]  NAM←S/NAM
[11]  ML←ML,(~MS∈ML)/MS←NAM[MN←2↑MINIDX D←S/S≠D]
[12]  MST←MST,MS
[13]  M←S≠M
[14]  →L×⍳2<1↑⍴M
[15]  3 0 ⍎ML
[16]  ''
[17]  MST←((¯1+N),2)⍴MST
[18]  I←0
[19] LL:MM←MM,[1] ML∈MST[I←I+1;]
[20]  →((N-1)>I)/LL
[21]  (((2×N)⍴3 0),10 2)⍎MM[1↓⍳N;],ZZ
    ∇
```

```
    ∇ D←RD2 M;N;I;L
 [1]  D←(2⍴N←1↑⍴M)⍴0
 [2]  I←1
 [3]  L←(N⍴L),0
 [4] L:D[I;]←+/(M-(⍴M)⍴M[I;])*2
 [5]  →L[I←I+1]
    ∇
```

```
    ∇ Z←MINIDX M;A;C;D;E;F
 [1]  Z←(⌈F÷C),(C|¯1+(F←A+⌈(A←E⍳D←⌊/E←((C*2)⍴0,C⍴1)/,M)÷(C←1↑⍴M)))+1
 [2]  ZZ←ZZ,D
    ∇
```

Fig. 14. (*continued*)

number of cases NUM is reached or the candidate list is exhausted. STEPDOWN discards one case at a time, that which makes the least contribution to det $\mathbf{A}_-$. SWAP will attempt to replace each point in EA sequentially with each point in DA but will only select the one case where det $\mathbf{A}_0$ is maximally increased. If successful, the following cases must improve over the new det $\mathbf{A}_0$. Both STEPDOWN and SWAP allow cases to be frozen in the design. All three programs assume that the model to be fit includes an intercept term c, *viz* $Z = aX + bY + c$.

The function of MXSNT1 (and associated COLAV and STDEV) is to achieve mean and standard deviation standardization by columns. MXSNT1 can be used to standardize the combined DA and EA before using one of the design programs in order to equalize the different types of constants, which are on different scales.

DISTPLAN is our version of the Wootton *et al.* (*44*) algorithm for designing starting sets of substituents. Starting point(s) must be specified, along with a cutoff distance (actually distance squared is used throughout in order to save some computation time). An average distance is printed to help in selecting this cutoff. A distance matrix Eq. (4) is determined between all cases and the center of gravity of the included points IP, namely DCG. DMD is a symmetrical distance matrix, determined by the service program RD2, which stores all those distances greater than or equal to the cutoff distance as a 0 or 1 (no,yes). In line [10] (Fig. 14), any member of DCG that is greater than the cutoff of the IP is determined and then only the smallest is selected and added to the list of IP. Thus, the point which is both greater than the cutoff for each individual IP and also closest to the center of gravity is selected. The cycle is repeated until no further points can be selected.

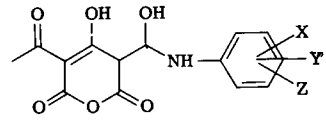
TOM is Farrar and Gluaber's (*48*) algorithm for detecting multicollinearity. Omega vector (which parameters might be multicollinear, F distribution), r^2 matrix (simple correlation), and a t matrix [parameters actually (multi) colinear] are determined. The function *TM* calculates the 95% t values by an approximation:

$$t^{\alpha=0.05} = e^{[0.673019 + 1.200524(1/f) + 0.8139429(1/f^2) - 0.1453636(1/f^3)]} \quad (32)$$

Finally, the program CLU calculates a hierarchical cluster analysis on rows (i.e., cases, usually compounds) be replacing the two closest cases (squared distance from RD2), with their center of gravity and redetermining all squared distances. The cycle is repeated until only one cluster results. Center of gravity cases are treated as all other cases.

VII. Conclusions

A recent "QSAR success story" by Cramer *et al.* (*78*) at SmithKline illustrates many of the points made in this article, and in many ways parallels the thiazole beta-blocker example discussed herein. A series of pyranenamine (**XI**) antiallergy compounds was optimized in several stages. In a pre-QSAR stage the historical data (19 compounds) was analyzed. Most of the derivatives had been synthesized according to the



(XI)

Topliss tree method, but serious deficiencies were found and the current optimal compounds would not have been obtained by rigorous application of this method. A number of bioisosteres of the most active analog were then prepared; this yielded compounds of activity approximately equal to the most active compound, which was about 8 times more active than the parent. The next phase consisted of a QSAR analysis of the compounds by graphical analysis (cf. tuberins (**VI**) (*21*) and thiazole beta-blockers), which suggested the synthesis of more hydrophilic and electronically neutral compounds. The resulting derivatives were about 2.5 times more potent than those obtained in the first stage. Thereafter, a large number of hydrophilic analogs were prepared, but with an eye toward fleshing out other aspects of the QSAR, such as hydrogen-bonding and bulk effects. Regression analyses were then performed and certain deviations were noted. From the QSAR it was suggested that, for example, acyl derivatives might be subject to hydrolysis. A final phase of synthesis occurred in which the QSAR was further tidied up by directed synthesis to answer specific hypotheses. The final result was a compound that was structurally unique, that confirmed the original hypotheses (from graphical analysis), and that was approximately 1000-fold more active then the initial lead. Although this example did not become available until this manuscript was nearly completed, it nicely reinforces many of the conclusions as to method, the necessity of interactive hypothesis testing, and the power of the Hansch paradigm.

The terms QSAR and rational drug design are often used interchangeably. Perhaps the connection between a quantitative, statistical model and the rational development of drugs will be misinterpreted; because techniques are labeled rational or statistical in no way guarantees that the results will be necessarily readily forthcoming, straightforward, or even useful. The paradigm provides only a clue, a method for understanding an extremely complex phenomenon; but it is only a tool and not an answer.

In evaluating the usefulness of a QSAR program, it is very important to compare the results in a suitable manner. One often hears "that would have been done anyway." This may, of course, be true; but with QSAR, there is a reason beyond obviousness or availability of materials, namely hypothesis testing by a statistical or, at least, objective paradigm.

ACKNOWLEDGMENT

I would like to thank most sincerely Dr. John Edwards, Director of the Institute of Organic Chemistry, Syntex Research, for continued support and encouragement as well as all of my colleagues throughout Syntex Research for hundreds of hours of stimulating discussion, cooperation, and friendship. In particular, I thank the following scientists who have contributed to parts of the studies reported herein: Bob Alvarez, Belig Berkoz, John Bruno, Ian Harrison, Al Holstein, Dave Johnson, Jerry Johnson, John Kent, Art Kluge, Michael Marx, Tom Matthews, Joe Muchowski, Krishna Murthy, John Patterson, Jurg Pfister, Wendell Rooks, Peter Roszkowski, Art Strosberg, Karl Untch, Al VanHorn, Keith Walker, Marshall Wallach, Klaus Weinhardt, and Bob Weissberg. Special thanks to my assistants Jim Cook, for much of the development work on the octanol HPLC technique, and Tony Feuerman, for laboratory work and some creative APL programming. This work is contribution No. 524 from the Institute of Organic Chemistry, Syntex Research.

References

1. T. S. Kuhn, "The Structure of Scientific Revolutions." 2nd ed., International Encyclopedia of Unified Science (O. Neurath, ed.), Vol. 2, No. 2. Univ. of Chicago Press, Chicago, 1970.
2a. C. Hansch, R. M. Muir, T. Fujita, P. Maloney, E. Geiger, and M. Streich, *J. Am. Chem. Soc.* **85,** 2817 (1963).
2b. T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.* **86,** 5175 (1964).
2c. C. Hansch and A. R. Steward, *J. Med. Chem.* **7,** 691 (1964).
3a. Y. C. Martin, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 8, p. 1. Academic Press, New York, 1978.
3b. Y. C. Martin, "Quantitative Drug Design." Dekker, New York, 1978.
3c. G. Redl, R. D. Cramer III, and C. E. Berkhoff, *Chem. Soc. Rev.* 273 (1974).
3d. C. Hansch, *In* "Structure–Activity Relationships" (C. J. Cavallito, ed.), Vol. 1. p. 75. Pergamon, Oxford, 1973.
3e. R. Franke and P. Oehme, *Pharmazie* **28,** 489 (1973).
4a. C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikaitani, and E. Lien, *J. Med. Chem.* **16,** 1207 (1973).
4b. A. Leo and C. Hansch, Pomona College Medicinal Chemistry Project, Claremont, CA. Log P and parameter listings, 1978.
4c. S. H. Unger and C. Hansch, *In* "Progress in Physical Organic Chemistry" (R. W. Taft, Jr., ed.), Vol. 12, p. 91, Wiley, New York, 1976.
4d. A. Verloop, W. Hoogenstraaten, and J. Tipker, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 7, p. 165. Academic Press, New York, 1976.
4e. R. F. Rekker, "The Hydrophobic Fragmental Constant: Its Derivation and Application. A Means of Characterizing Membrane Systems." Elsevier, Amsterdam, 1977.
4f. A. Leo, P. Y. C. Jow, C. Silipo, and C. Hansch, *J. Med. Chem.* **18,** 865 (1975).
4g. C. Hansch, S. D. Rockwell, P. Y. C. Jow, A. Leo, and E. E. Steller, *J. Med. Chem.* **20,** 304 (1977).
5a. Chemistry: J. Edwards, K. Untch, B. Berkoz, B. Lewis. Biology: A. Strosberg, R. Weissberg, R. Alvarez, J. Bruno, L. Miller, L. Johnson, J. Marian, L. A. Taylor.
5b. U.S. Patents 4,064,135, 4,064,257, 4,064,258, and 4,064,259 (1977).
6a. B. Berkoz, J. E. Edwards, and B. Lewis, 172nd Meet. Am. Chem. Soc., Abstr. No. MEDI 014, San Francisco, 1976.

6b. J. A. Edwards, B. Berkoz, G. S. Lewis, O. Halpern, J. H. Fried, A. M. Strosberg, L. M. Miller, S. Urich, F. Liu, and A. P. Roszkowski, *J. Med. Chem.* **17,** 200 (1974).
7a. M. L. Weiner and P. H. Weiner, *J. Med. Chem.* **16,** 655 (1973).
7b. J. P. Fernandez, G. Grassy, A. Terol, Y. Robbe, J. P. Chapat, R. Granger, and H. Sentenac-Roumanou, *Eur. J. Med. Chem.* **13,** 245 (1978).
8a. P. J. Lewi, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 7, p. 209. Academic Press, New York, 1976.
8b. P. J. Lewi, *Arzneim. Forsch.* **26,** 1295 (1976).
8c. P. J. Lewi, *Pharmac. Ther. B,* **3,** 481 (1978).
9. R. J. Rummel, "Applied Factor Analysis." Northwestern University Press, Evanston, 1970.
10a. P. A. J. Janssen, C. J. E. Niemegeers, and K. H. L. Schellenkens, *Arzneim.-Forsch.* **15,** 104 (1965).
10b. L. K. C. Desmedt, C. J. E. Niemegeers, P. J. Lewi, and P. A. J. Janssen, *Arzneim.-Forsch.* **26,** 1592 (1976).
11a. J. M. Muchowski and A. F. Kluge, U.S. Patents 4,087,539, 4,089,969, and 4,097,579, (1978).
11b. J. Ackrell, Y. Antonio, F. Franco, R. Landeros, A. Leon, J. M. Muchowski, M. L. Maddox, P. H. Nelson, W. H. Rooks, A. P. Roszkowski, and M. B. Wallach, *J. Med. Chem.* **21,** 1035 (1978).
12. N. Cohen, B. L. Banner, R. J. Lopresti, and H. W. Baruth, *J. Med. Chem.* **21,** 895 (1978).
13. W. J. Dixon and M. B. Brown, eds., "Biomedical Computer Programs, P-Series." Uni. California Press, Los Angeles, 1977.
14. S. H. Unger and C. Hansch, *J. Med. Chem.* **16,** 745 (1973).
15a. Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff, *J. Med. Chem.* **17,** 409 (1974).
15b. D. R. Henry and J. H. Block, *J. Med. Chem.* **22,** 465 (1979).
16. A. F. Kluge, J. M. Caroon, S. H. Unger, and J. F. Ryley, *J. Med. Chem.* **21,** 529 (1978).
17. J. Dearden and M. S. Townsend, *J. Pharm. Pharmacol* **28s,** 13 (1976).
18. D. V. Parke, *Chemistry and Industry* 380 (1976).
19. P. H. A. Sneath and R. R. Sokal, "Numerical Taxonomy." Freeman, San Francisco, 1973.
20. C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.* **16,** 1217 (1973).
21. I. T. Harrison, W. Kurz, I. J. Massey, and S. H. Unger, *J. Med. Chem.* **21,** 588 (1978).
22. S. H. Unger, J. Cook, and J. Hollenberg, *J. Pharm. Sci.* **67,** 1364 (1978).
23. A. P. Roszkowski, C. C. Beard, C. Dvorak and K. Weinhardt, U.S. Patent 4,088,771, (1978).
24a. S. H. Unger, *Cancer Chemotherapy Reports, Part 2* **4,** 45 (1974).
24b. S. H. Unger, Proc. Workshop on Chemical Appl. Pattern Recognition, p. 62. Washington, D.C., 1975.
25. B. R. Baker, "Design of Active-Site-Directed Irreversible Enzyme Inhibitors." Wiley, New York, 1967.
26a. C. Hansch, C. Grieco, C. Silipo, and A. Vittoria, *J. Med. Chem.* **20,** 1420 (1977).
26b. C. Hansch, J. Y. Fukunaga, P. Y. C. Jow, and J. B. Hynes, *J. Med. Chem.* **20,** 96 (1977).
26c. C. Silipo and C. Hansch, *J. Med. Chem.* **19,** 62 (1976).
26d. M. Yoshimoto and C. Hansch, *J. Med. Chem.* **19,** 71 (1976).
26e. C. Hansch, *J. Med. Chem.* **19,** 1 (1976).
27. J. Levy, *Eur. J. Pharmacol.* **2,** 250 (1968).

28a. F. F. Brown and P. J. Goodford, *Br. J. Pharmacol.* **60,** 337 (1977).
28b. C. R. Beddell, P. J. Goodford, F. E. Norrington, S. Wilkinson, and R. Wootton, *Br. J. Pharmacol.* **57,** 201 (1976).
28c. M. A. Ondetti, B. Rubin, and D. W. Cushman, *Science* **196,** 441 (1977).
29a. W. J. Dunn III, S. Wold, and Y. C. Martin, *J. Med. Chem.* **21,** 922 (1978).
29b. W. J. Dunn III and S. Wold, *J. Med. Chem.* **21,** 1001 (1978).
30. E. R. John, B. Z. Karmel, W. C. Corning, P. Easton, D. Brown, H. Ahn, M. John, T. Harmony, L. Prichep, A. Toro, I. Gerson, F. Bartlett, R. Thatcher, H. Kaye, P. Valdes, and E. Schwartz, *Science* **196,** 1393 (1977).
31a. R. D. Cramer III, G. Redl, and C. E. Berkoff, *J. Med. Chem.*, **17,** 533 (1974).
31b. *Chem. Eng. News,* p. 33, April 8, 1974.
31c. R. D. Cramer III, personal communication, 1978.
32. L. Hodes, G. F. Hazard, R. I. Geran, and S. Richman, *J. Med. Chem.* **20,** 469 (1977).
33a. *Chem. Eng. News,* p. 14, August 28, 1972.
33b. *Chem. Eng. News,* p. 19, February 18, 1974.
33c. *Anal. Chem.* **46,** 417A (1974).
34a. K. L. H. Ting, R. C. Lee, G. W. A. Milne, M. Shapiro, and A. M. Buarino, *Science* **180,** 417 (1973).
34b. C. L. Perrin, *Science* **183,** 551 (1974).
34c. J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia* (Buenos Aires) **27,** 639 (1973).
35. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **96,** 916 (1974).
36a. A. J. Stuper, W. E. Brugger, and P. C. Jurs, *In* "Chemometrics: Theory and Application" (B. R. Kowalski, ed.). A.C.S. Symposium Series, Vol. 52, p. 165, 1977.
36b. A. J. Stuper and P. C. Jurs, *J. Pharm. Sci.* **67,** 745 (1978).
36c. A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.* **97,** 182 (1975).
36d. P. C. Jurs, J. T. Chou and M. Yuan, *J. Med. Chem.* **22,** 476 (1979).
36e. A. J. Stuper, W. E. Brügger and P. C. Jurs, "Computer Assisted Studies of Chemical Structure and Biological Function." Wiley, New York, 1979.
37a. K. C. Chu, *Anal. Chem.* **46,** 1181 (1974).
37b. K. C. Chu, R. J. Feldmann, M. B. Shapiro, G. F. Hazard, Jr., and R. I. Geran, *J. Med. Chem.* **18,** 539 (1975).
37c. K. C. Chu, D. E. Goldender, and A. B. Rosonblit, *Comp. Biomed. Res.* **6,** 411 (1973).
38a. S. Wold and M. Sjöström, *In* "Chemometrics: Theory and Application" (B. R. Kowalski, ed.), A.C.S. Symposium Series, Vol. 52, p. 243, 1977.
38b. B.-K. Chen, C. Horváth, and J. R. Bertino, *J. Med. Chem.* **22,** 483 (1979).
39a. A. E. Hoerl and R. W. Kennard, *Technometrics* **12,** 55 (1970).
39b. A. E. Hoerl, R. W. Kennard, and K. F. Baldwin, *Commun. Statistics* **4,** 105 (1975).
40a. J. G. Topliss, *J. Med. Chem.* **15,** 1006 (1972).
40b. J. G. Topliss and Y. C. Martin, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 5, p. 1, Academic Press, New York, 1975.
40c. J. G. Topliss, *J. Med. Chem.* **20,** 463 (1977).
41a. F. Darvas, *J. Med. Chem.* **17,** 799 (1974).
41b. F. Darvas and A. Eöry, in press.
42. R. D. Gilliom, W. P. Purcell, and T. R. Bosin, *Eur. J. Med. Chem.* **12,** 187 (1977).
43. S. H. Unger, Doctoral Dissertation, M.I.T., Cambridge, Massachusetts, 1970.
44. R. Wootton, R. Cranfield, G. C. Sheppey, and P. J. Goodford, *J. Med. Chem.* **18,** 607 (1975).
45. R. W. Kennard and L. A. Stone, *Technometrics* **11,** 137 (1969).
46. F. E. Norrington, R. M. Hyde, S. G. Williams, and R. Wootton, *J. Med. Chem.* **18,** 604 (1975).

47. C. G. Swain and E. C. Lupton, *J. Am. Chem. Soc.* **90,** 4328 (1968).
48. D. E. Farrar and R. R. Glauber, *Rev. Econ. Stat.* **49,** 92 (1969).
49a. S. H. Unger and C. Hansch, 167th Meet. Am. Chem. Soc., Abstr. No. CHLT 004, Los Angeles, 1974.
49b. K. H. Kim, C. Hansch, J. Y. Fumunaga, E. R. Steller, P. Y. C. Jow, P. N. Craig and J. Page, *J. Med. Chem.* **22,** 366 (1979).
50. M. J. Box and N. R. Draper, *Technometrics* **13,** 682 (1971).
51. N. R. Draper and H. Smith, "Applied Regression Analysis." Wiley, New York, 1966.
52a. J. D. Johnson and K. W. Last, Presented at "Symposium on Design and Linear Models," Ft. Collins, Colorado, 1973.
52b. O. Dykstra, *Technometrics* **13,** 682 (1971).
53a. K. E. Iverson, "A Programming Language." Wiley, New York, 1962.
53b. IBM, "APL Shared Variables System, User's Manual," SH20-1460, IBM, New York, 1974.
54. T. L. Hebble and T. J. Mitchell, *Technometrics* **14,** 767 (1972).
55. G. M. Furnival and R. W. Wilson, Jr., *Technometrics* **16,** 499 (1974).
56. A. P. Dempster, M. Schatzoff, and N. Wermuth, *J. Amer. Stat. Assn.* **72,** 77 (1977).
57. J. G. Topliss and R. J. Costello, *J. Med. Chem.* **15,** 1066 (1972).
58a. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **94,** 5632 (1972).
58b. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **95,** 686 (1973).
58c. A. Cammarata and G. K. Menon, *J. Med. Chem.* **19,** 739 (1976).
58d. G. K. Menon and A. Cammarata, *J. Pharm. Sci.* **66,** 304 (1977).
59a. W. J. Dunn III and M. J. Greenberg, *J. Pharm. Sci.* **66,** 1416 (1977).
59b. W. J. Dunn IH, M. J. Greenberg, and S. S. Callejas, *J. Med. Chem.* **19,** 129 (1976).
60. K. J. Shah and E. A. Coats, *J. Med. Chem.* **20,** 1001 (1977).
61a. J. J. Kaufman and W. S. Koski, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 5, p. 252. Academic Press, New York, 1975.
61b. L. B. Kier, "Molecular Orbital Theory in Drug Research." Academic Press, New York, 1971.
61c. B. Pullman and A. Pullman, "Quantum Biochemistry." Wiley, New York, 1961.
62a. B. V. Cheney, J. B. Wright, C. M. Hall, H. G. Johnson, and R. E. Christoffersen, *J. Med. Chem.* **21,** 936 (1978).
62b. G. H. Loew, B. S. Sudhindra, J. M. Walker, C. C. Sigman, and H. L. Johnson, in press.
63a. R. Potenzone, Jr., E. Cavicchi, H. J. R. Weintraub, and A. J. Hopfinger, *Computers in Chemistry,* **1,** 187, 1977.
63b. H. J. R. Weintraub and A. J. Hopfinger, *In* "Molecular and Quantum Pharmacology" (E. D. Bergmann and B. Pullman, eds.), pp. 131–152, Reidel, Dordrecht, 1974.
64. A. J. Hopfinger and R. D. Battershell, Proc. of I.U.P.A.C. Conf. Pest. Sci., Pergamon, London.
65a. A. Levitzki, *In* "A Guide to Molecular Pharmacology–Toxicology" (R. M. Featherstone, ed.) Part I, pp. 305–322, Dekker, New York, 1973.
65b. D. E. Koshland, Jr., *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed. Vol. 1, pp. 341–395, Academic Press, New York, 1970.
65c. R. N. Lindquist, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. 5, p. 24. Academic Press, New York, 1975.
66. R. D. Cramer III and M. Milne, personal communication, 1978.
67a. M. S. Newman, *J. Am. Chem. Soc.* **72,** 4783 (1950).
67b. O. Rosado-Lojo, C. K. Hancock, and A. Danti, *J. Org. Chem.* **31,** 1899 (1966).
68. R. H. Davies, *Int. J. Quant. Biol.: Quantum Biology Symp.* **4,** 413 (1977).

69. E. J. Corey and W. T. Wipke, *Science* **166,** 178 (1969).
70. G. L. Biagi, A. M. Barbaro, M. F. Bamba, and M. C. Guerra, *J. Chromatogr.* **41,** 371 (1969).
71a. E. Tomlinson, *J. Chromatogr.* **113,** 1 (1975).
71b. A. Hulshoff and J. H. Perrin, *J. Chromatogr.* **129,** 263 (1976).
72a. W. J. Haggerty, Jr. and E. A. Murrill, *Res. Dev.* **25,** 30 (1974).
72b. J. M. McCall, *J. Med. Chem.* **18,** 549 (1975).
73a. M. S. Mirrless, S. J. Moulton, C. T. Murphy, and P. J. Taylor, *J. Med. Chem.* **17,** 409 (1976).
73b. D. Henry, J. H. Block, J. L. Anderson, and G. R. C. Carlson, *J. Med. Chem.* **19,** 619 (1976).
73c. K. Miyake and H. Terada, *J. Chromatogr.* **157,** 386 (1978).
74a. K. G. Wahlund, *J. Chromatogr.* **115,** 411 (1975).
74b. K. G. Wahlund and U. Lund, *J. Chromatogr.* **122,** 269 (1976).
74c. B. Fransson, K. G. Wahlund, I. M. Johansson, and G. Schill, *J. Chromatogr.* **125,** 327 (1976).
74d. I. M. Johansson and K. G. Wahlund, *Acta Pharm. Suec.* **14,** 459 (1977).
74e. K. G. Wahlund and I. Beijersten, *J. Chromatogr.* **149,** 313 (1978).
74f. K. G. Wahlund and A. Sokolowski, *J. Chromatogr.* **151,** 299 (1978).
75a. C. Horváth, W. Melander, and I. Molnár, *Anal. Chem.* **49,** 142 (1977).
75b. C. Horváth and W. Melander, *J. Chromatogr. Sci.* **15,** 393 (1977).
76a. J. R. Pfister, R. W. Ferraresi, I. T. Harrison, W. H. Rooks, A. P. Roszkowski, and A. Van Horn, *J. Med. Chem.*, **15,** 1032 (1972).
76b. S. H. Unger and T. F. Feuerman, *J. Chromatogr.* **176,** 426 (1979).
76c. S. H. Unger, Proc. E.P.A. Seminar on Structure Activity–Relations, November 8–9, 1978, Annapolis, Md.
77. A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.* **71,** 525 (1971).
78. R. D. Cramer III, K. M. Snader, C. R. Willis, L. W. Chakrin, J. Thomas, and B. M. Sutton, *J. Med. Chem.* **22,** 714 (1979).
79. S. H. Unger, K. Untch, B. Lewis, B. Berkoz, J. Edwards, A. Strosberg, R. Weissberg, and R. Alvarez, Proc. Third Congress of the Hungarian Pharmacological Society, August 22–25, 1979, Budapest, Hungary, in press.

Chapter 3 A Physical Chemical Basis for the Design of Orally Active Prodrugs

Samuel H. Yalkowsky and Walter Morozowich

ISBN 0-12-060309-8

I. Introduction

Drug Efficiency

It is a well-known fact that many drugs that show good activity when administered parenterally are either weakly active or totally inactive when given orally. In most cases, poor oral activity results from the inability of the orally administered drug to reach its desired receptor in sufficient quantities to elicit a response. This inefficiency in reaching the receptor is invariably due to one or more of three causes: inefficient release of the drug from its formulation; inefficient transport of the drug across the limiting gastrointestinal membrane; and poor stability of the drug in the gastrointestinal tract (including the gut wall and the liver).

In this chapter, we will be concerned with improving the efficiency with which an orally administered drug can be introduced into the general circulation. For the purposes of this chapter, we define efficiency as the ratio of the oral to parenteral activity of a drug. Thus, a rapid dissolving, well absorbed, stable drug will have an efficiency of 1.0. The term efficiency is to be distinguished from the broader term efficacy in that the former is not explicitly concerned with altering the therapeutic index of the drug. Improving the absorption and/or availability (efficiency) of a drug will, except in rare instances, have no effect on its therapeutic index. It will, on the other hand, reduce the quantity of drug that must be administered to get a desired biological effect, i.e., it will increase the efficiency with which the drug passes from its dosage form into the systemic circulation.

Usually, a two- or threefold increase in the extent of absorption may not drastically reduce the dose and/or cost of a drug, and will have little effect upon the drug's systemic therapeutic index. In general, a more completely absorbed drug will be more uniformly absorbed than a poorly absorbed drug. Thus, day-to-day and patient-to-patient variation will be minimized. Also, orally administered drugs that show local side effects (irritation of the gastrointestinal mucosa) may produce fewer side effects if they are efficiently absorbed.

The decision to attempt drug modification will depend upon the size of the dose, the cost of the drug, its sales potential, etc., as well as the efficiency of drug uptake.

If the processes of dissolution and absorption are to be controlled by the design of new drugs, it is necessary to understand these processes in terms of drug structure and properties. In this chapter, we will attempt to show that drug dissolution and absorption can be correlated quite success-

fully with the solubility and partition coefficient of the drug, and that both of the latter properties can be predicted with reasonable reliability from chemical structure.

II. Factors Governing Drug Efficiency

In order for an orally administered drug to become available to its systemic receptor, it must first be dissolved in the gastrointestinal fluid and then transported across the gastrointestinal membrane, after which it can be picked up and distributed by the systemic circulation. For many drugs, availability to the receptor is limited by poor dissolution and/or poor absorption (transport).

In this section, we will consider the processes of dissolution and transport and their dependence on the physical-chemical properties of drugs. We will show that these processes are primarily dependent on the aqueous solubility and the membrane–water partition coefficient of the drug. We will also attempt to provide guidelines for assessing the most desirable combination of solubilities and partition coefficients for efficient transfer of drugs from their formulations to the systemic circulation. Particular emphasis will be given to the prodrug approach, which has been found to offer great potential for improving drug efficiency (*1–12*).

A. DISSOLUTION

The rate of dissolution R of a drug in an aqueous medium is known to be directly proportional to the solubility S of the drug in the medium:

$$R = KS \tag{1}$$

where K is a constant that reflects the surface area of the drug particles and the degree of agitation of the medium in the region of the drug particles. The initial rate of dissolution of a wide variety of drugs in aqueous media is dramatically illustrated by the data of Hamlin *et al.* (*13*) which is presented in Fig. 1. Note that the data covers over four orders of magnitude.

B. TRANSPORT

On the basis of a straightforward diffusional model [first presented by Zwolinski, Eyring, and Reese in 1949 (*14*) and more recently expanded

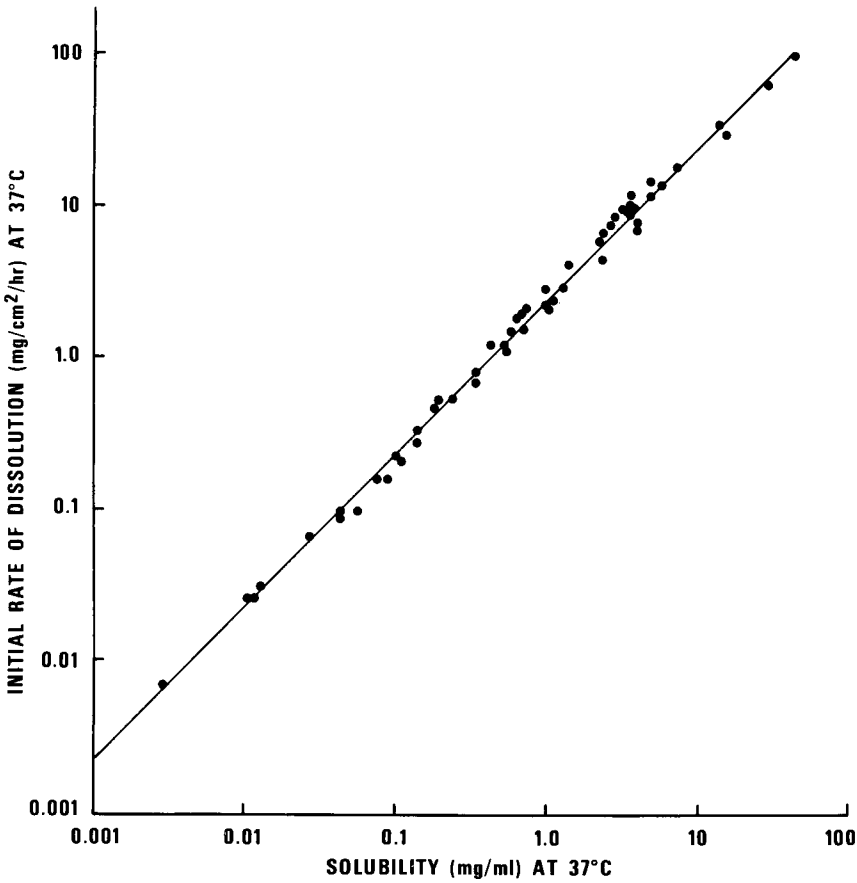


Fig. 1. Dependence of dissolution rate on solubility for a number of drugs (*13*).

upon by Higuchi, Ho, Stehle, Flynn, and Yalkowsky (*15–20*)], the rate of transport of a drug across a biological membrane adjacent to one or more aqueous compartments is

$$F = \frac{\Delta C}{(R_m/PC) + R_{aq}} \tag{2}$$

where ΔC is the concentration differential across the membrane and aqueous region(s); PC is the membrane–water partition coefficient; and R_m and R_{aq} are the resistances of the membrane and aqueous regions, respectively, to diffusion of the drug. These resistances depend primarily on the

thickness and viscosity of the phases and only slightly on the structure of the drug. For this reason we can regard R_m and R_{aq} as constants for the transport of any series of drugs across a particular membrane.

If we assume that the concentration of the drug on the receptor side of the membrane is negligible, the term ΔC becomes simply the concentration of drug in the donor compartment, i.e., the dose divided by the compartment volume.

Equation (2) predicts that for any applied phase concentration, the transport rate will increase exponentially with increasing partition coefficient until PC is sufficiently large that $R_m \gg R_{aq}/PC$ and the curve levels off to a plateau. Increasing the applied phase concentration (i.e., the dose) results in a parallel line, as shown in Fig. 2.

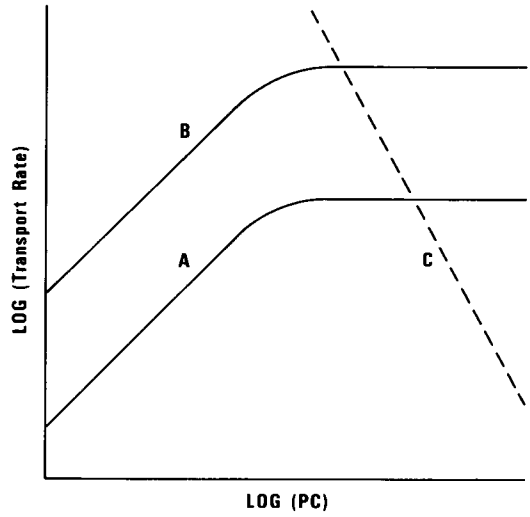


Fig. 2. Dependence of transport rate on partition coefficient. A–low dose, B–high dose, C–saturated.

Whereas it is not possible, from the data available in the literature, to obtain a very accurate optimal value of log PC for intestinal absorption, it appears that a value of 2.0 (on the octanol–water scale) is appropriate. In other words, drugs that have log $PC > 2.0$ are usually well absorbed from the gut. Provided, of course, that they are completely soluble in the GI fluid at the dose given.

The complete form of Eq. (2) is that of a hyperbola, as shown by the solid lines of Fig. 2. Note that increasing the concentration differential

(which is equivalent to increasing the dose of the drug) results in a parallel upward shift of the curve.

A hyperbola as described here differs from a parabola as described by Hansch (*21–24*) in several important respects.

A parabola of the form $y = ax^2 + bx + c$ has a maximum value at x_0 and is symetrical about that value, whereas a hyperbola of the form $y = a/(b + c/x)$ does not have a maximum value at any x but approaches a plateau as x becomes very large (see Fig. 2). It is thus not symmetrical about any value of x. Also at low values of x, a hyperbola can be approximated by a straight line while a parabola of the above form (contrary to what has been stated in the literature) is never linear except when it approaches a line of infinite slope.

Since Eq. (2) asymptotically approaches a limiting value, no single optimum partition coefficient for maximum transport can be defined (except when additional phenomena are involved, as are now described). It is possible, however, to define the minimum partition coefficient required to produce a flux equal to 90% or more of the asymptotic value. By defining the flux at infinite partition coefficient as 100% of maximum, we can easily show that at 90% of maximum:

$$PC_{\geq 90} \geq 0.9\, R_m/R_{aq} \tag{3}$$

Now since for any barrier, R_m and R_{aq} are independent of the drug and its partition coefficient, the term $0.9\, R_m/R_{aq}$ is a constant for that barrier. Thus, there is a single minimum partition coefficient required to get 90% of maximal transport across any biological barrier.

The significance of Eq. (3) lies not only in the fact that it predicts a linear $\log F$ vs $\log PC$ region, whereas parabolic equations do not, but in the fact that it does not predict a single optimum partition coefficient PC for maximum transport. On the basis of transport theory alone, there is no reason to expect a decrease in flux (or biological activity, which is dependent upon flux) with increasing partition coefficient. There is especially no reason to expect such a decrease to be symmetrical with respect to the increase observed at lower partition coefficients.

Many small, polar compounds are known to be well absorbed. Since polar compounds would not be expected to pass rapidly through lipoidal membranes, it is postulated that they pass through pores in the membrane. The incorporation of the concept of aqueous pores would change Eq. (2) to:

$$F = \frac{\Delta C}{R_m/PC + R_{aq}} + \frac{\Delta C}{R_p} \tag{4}$$

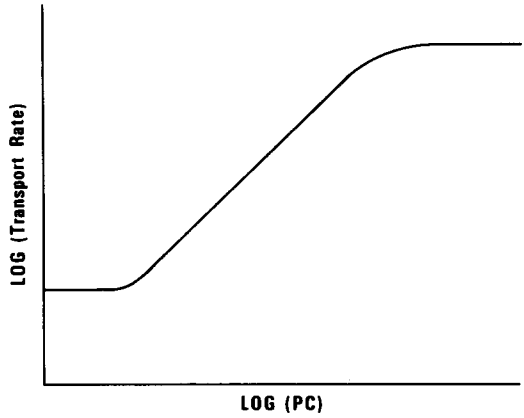


Fig. 3. Dependence of transport rate on partition coefficient when pore transport is not negligible.

where R_p is the total resistance of the aqueous pores. The significance of the pore pathway in membrane transport is discussed more completely by Ho *et al.* (*16*).

The effect of the presence of pores on the overall structure–activity curve is to convert the hyperbolic relationship into a sigmoidal one, as shown in Fig. 3.

Note that because pore transport is dependent only on molecular size, it is independent of partition coefficient. In general, the pore pathway is only important for low-molecular-weight compounds. For larger drugs (MW > 200), this factor can be ignored. The importance of pore transport for solutes of various sizes is shown in Fig. 4.

C. ACTIVITY-LIMITING PHENOMENA AND SOLUBILITY

We know from experience that there are situations in which activity decreases with increasing partition coefficient. Since this cannot be explained by simple transport theory, we must consider what other factors might be responsible for the decrease, and what role these factors play in determining the shape of the overall structure–activity curve for a series of drugs.

These factors can be broadly classified as either biological or physical-chemical. The former would include such factors as receptor site fit, binding to inert receptors, and uptake by lipoidal tissues. Since these factors do not generally affect the transport of a drug across its primary

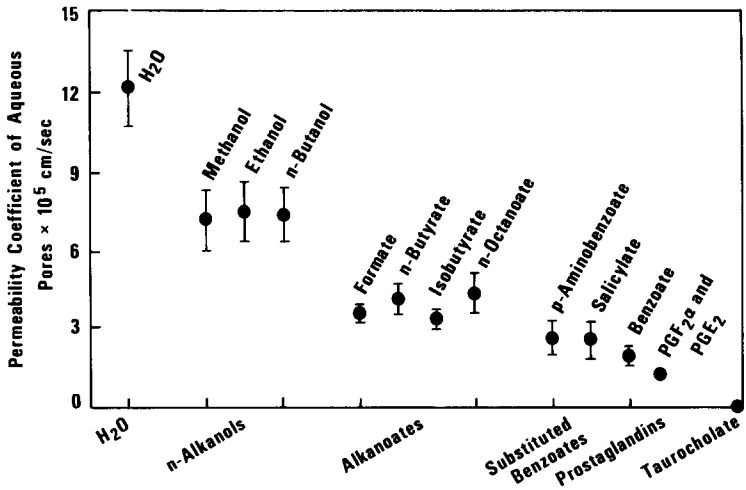


Fig. 4. Dependence of permeability coefficient on solute size (X-axis is arbitrary) (*16*).

absorptive barrier, they will be considered as beyond the scope of this report.

Physical-chemical factors are those which alter the availability of the drug at the transport-limiting membrane and are thus an integral part of the structure transport picture. Since it is usually the free, un-ionized, monomeric drug that is most efficiently transported across biological barriers; any factor that limits the concentration (or more appropriately, the thermodynamic activity) of the un-ionized monomeric species will limit transport and decrease activity.

The role of phenomena that restrict the thermodynamic activity of drugs in structure–activity relationships was first recognized by Ferguson (*25*) in 1939. Although the importance of these phenomena was recognized by other workers (*26–29*), they were largely ignored by much of the medicinal chemical community.

The most important flux-limiting, physical-chemical factor is solubility. Other factors, such as micellization or complexation, can be treated in a manner similar to the way we will treat solubility, since they too limit the concentration of free monomeric drug in solution. Mathematically, the effect of solubility is to set a limit on the maximum attainable value of ΔC (i.e., $\Delta C \leq S$).

The combined effects of aqueous solubility and membrane–water partition coefficient on absorption have been quantitatively described by Flynn

and Yalkowsky (*18–20*). By reasoning that the maximum value of the trans-membrane concentration differential is limited to the solubility S of the drug, they have shown that the maximum flux occurs from saturated donor phase solutions, i.e., suspensions, and is equal to:

$$F_{max} = \frac{S}{R_{aq} + R_m/PC} \tag{5}$$

The effects of the solubility limitation on transport are shown schematically by the descending line of Fig. 2. In their discussion of transport from equimolar and saturated solutions, Flynn and Yalkowsky (*18–20*) have shown that flux can be dependent on concentration (or dose) and partition coefficient, concentration alone, solubility alone, or the product of solubility and partition coefficient.

In order to simplify the discussion of the roles of solubility and partition coefficient in transport, we will confine the remainder of this section to homologous series.

For most homologous series, the partition coefficient increases threefold, while the solubility decreases by a factor of four for each methylene group, so that:

$$\log PC = \log PC_0 + 0.5\, n \tag{6}$$

and

$$\log S = \log S_0 - 0.6\, n \tag{7}$$

When these equations are incorporated into Eq. (4), we get four possibilities for chain length flux relationships. These have been discussed in great detail by Flynn (*18*) and Yalkowsky (*19, 20, 30*). In the most common instance, the solubility limitation results in a decrease in flux (or activity), as shown by the bold line in Fig. 2.

Because the observed biological activity of the members of a series of compounds can be dependent upon PC and/or S and dose, it would be a gross oversimplification to expect a single parameter to be sufficient to predict the complete structure–activity curve. It is, therefore, inappropriate to believe that there is single optimum partition coefficient for maximum activity in a series of compounds. The nonconstancy of the optimum partition coefficient for biological activity is illustrated in Fig. 5. All three plots represent alkyl *p*-aminobenzoate-induced narcosis of goldfish (*31*). The first curve is observed when the esters are administered in suspensions and has an "optimum" log (partition coeff.) of 1.65. The middle curve was obtained for a constant dose of 10^{-3} moles/liter of bathing

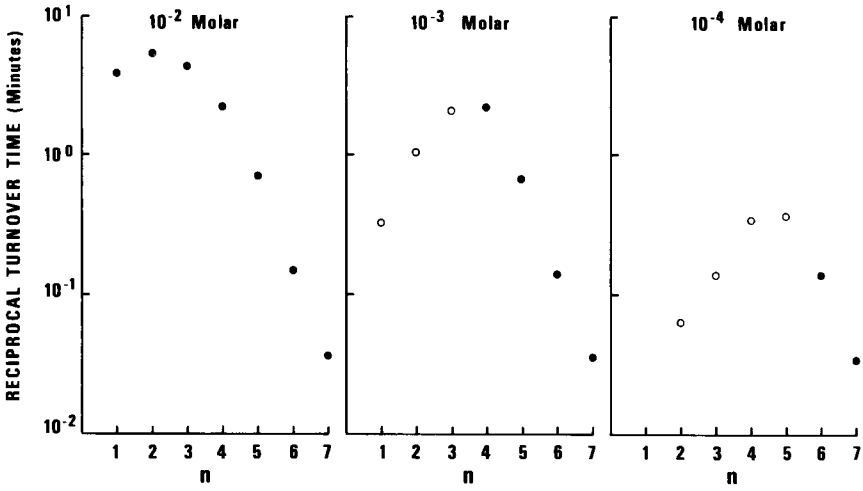


Fig. 5. Dependence of goldfish turnover time on alkyl-*p*-aminobenzoate chain length and on drug concentration.

solution and has an apparent "optimum" log (partition coefficient) of 2.44. The third curve, for a dose of 10^{-4} moles/liter, has an "optimum" log *PC* of 2.97.

The differences among the three curves can easily be explained in terms of the solubility and dose of each ester. In all three curves, the descending portion is composed of points for which the dose exceeds the solubility of the ester. These points are indicated by filled circles, whereas the points representing true solutions are designated by open circles. The curves also illustrate the fact that there cannot be a single criterion for solubility, without regard for the dose administered.

D. Key Physical Parameters

From this discussion, it is clear that both solubility and partition coefficient are important factors in drug uptake after oral administration. Drugs that are too polar often exhibit poor transport properties, whereas those that are too nonpolar frequently have low availabilities because of their poor solubility and dissolution characteristics. The designer of efficient prodrugs must take cognizance of both of these phenomena, otherwise gains in transport will be offset by losses in availability. The next section of this chapter will deal with the interrelationships among solubility, partitioning, and chemical structure, and will provide guidelines for controlling these parameters in drug design.

As just stated, a drug must first dissolve in the gastrointestinal fluid and then be transported across some portion of the gastrointestinal membrane. The first process, dissolution, is dependent upon the solubility in the gastric or intestinal fluid while the second process, transport, is dependent upon the drug's gastrointestinal membrane–lumen partition coefficient as well as its solubility in the appropriate GI fluid.

For reasons of convenience, we generally study solubility in water or aqueous buffer rather than in real or simulated gastric or intestinal fluid. For the same reasons, we usually study organic solvent–water (or buffer) partition coefficients rather than membrane–aqueous systems. Traditionally, *n*-octanol is used as the organic phase. Although the meaningfullness of these *in vitro* systems representing complex biological mileau is always subject to some question, such systems have been shown to be extremely useful for correlating and modeling both dissolution and transport phenomena.

In spite of having reasonable *in vitro* systems for modeling solubility and partition coefficient, we are not able to assign, definitively, optimal values for these two parameters. Part of the reason for this is that bioavailability can be dependent on S and not PC, PC and not S, S and PC, or neither, depending on the experimental conditions of the study (*18–20*). On the basis of the available data, we can say with a reasonable degree of certainty that drugs having octanol–water partition coefficients of 100 or more ($\log PC \geq 2$) are well absorbed, provided that they are given in doses that do not exceed their solubility in the gastric medium.

The problem with selecting an optimum solubility for good bioavailability is implicit in the preceding sentence, i.e., it depends on the dose. A solubility of 1 mg/ml for a steroid derivative presents no problem, but a much greater solubility would be needed to assure the availability of a tetracycline derivative.

In the absence of more definitive information, we can generally strive for a prodrug solubility equal to or greater than that of the parent drug. This will insure that efficiency will not be reduced as a result of poor prodrug solubility. As an alternative guideline, we can aim for a prodrug solubility equal to the parent drug dose divided by 100. Finally, if we do not have any dose for the parent drug, we can attempt to obtain an aqueous solubility of 10μg/ml.

If we do not have any means of knowing the expected dose of the prodrug, we can apply the following factors:

1. Transport rate increases with partition coefficient until $\log PC \simeq 2$ and then tends to level off.
2. Dissolution rates, as well as transport from saturated solutions, decrease as solubility decreases.

3. Solubility is frequently inversely proportional to partition coefficient. (This will be discussed at length in the next section).

From these factors we can reason that there is likely to be little benefit to transport from increasing log *PC* above 2, and that because this will likely reduce solubility, there is a possibility for a loss of efficiency. This is believed to be the reason that for many series of compounds, maximum activity is observed when log $PC \simeq 2$. Some examples of this are given later.

Lien (*31*) has shown that maximum absorption rate of several series of drugs in the GI tract of the rat occurs for derivatives or analogs with log *PC* near 2. The optimum log *PC* values (or log P_o) corresponding to the apex of a parabolic fit to the data are shown in Table I. The average log P_o value is 1.8.

TABLE I
LOG P_0 FOR ABSORPTION IN RATS[a]

Compounds	Absorption Site		Log P_0
Acids	Stomach		1.97
Barbiturates	Stomach		2.01
Bases	Intestine		1.39
Sulfonamides	Intestine		2.41
Bases	Colon		1.32
		(Average)	1.82

[a] From Lien (*31*).

In cases where the diffusional model has been applied to homologous compounds, such as alkanols, the limiting plateau in the absorption rate vs chain length relationship is just entered in the range of C_5–C_7 alkanols. This corresponds to an octanol/water log *PC* of 1.4–2.4, and the average value (2.0) is in accord with the above (*16*). Steroid absorption rate constants vs log *PC* indicate entrance into the plateau region at log *PC* 2.0–2.5.

Additional support for an optimal lipophilicity is obtained from the antibiotic area where the main reason for poor oral absorption is attributed to low lipophilicity. Pore transport of antibiotics should be negligible and absorption via the lipoidal regions of the intestinal tract should be the dominant pathway. Table II gives a listing of antibiotics and antibacterial agents grouped according to the isoamyl alcohol/water log *PC* as obtained from the solubility ratio (*32, 33*). The compounds with log *PC* less than zero in general show poor oral absorption, whereas those with log *PC*

TABLE II
ANTIBIOTICS TABULATED ACCORDING TO ISOAMYL ALCOHOL/WATER LOG *PC*[a]

Above zero	log *PC*	Below zero	log *PC*
Tyrothricin	0.04	Ampherocetin B	− 0.06
Anisomycin	> 0.54	Nystatin	− 0.22
Chloramphenicol	0.59	Hygromycin A	− 0.32
Erythromycin	0.66	Oleandomycin PO_4	− 0.50
Pyramycin	0.73	Oxytetracycline	− 0.84
Tetracycline	0.92	Bacitracin	− 1.08
Novobiocin acid	1.32	Streptohydrazide	− 1.19
Carbamycin	1.44	Neomycin B · HCl	− 1.65
Penicillin V acid	1.70	Vancomycin acetate	− 1.75
Erythromycin		Cycloserine	− 1.81
ethylsuccinate	> 1.83	Kanamycin base	− 2.11
Triacetyloleandomycin	> 1.90	Soframycin	− 2.19
Dihydronovobiocin	1.91	Streptomycin	− 2.23
Gramicidin	2.00	Ristocetin	− 2.45

[a] Log *PC* given by solubility ratio (*32, 33*).

above zero show good oral absorption in the human (*34*). Isoamyl alcohol /water log *PC* values approximate the *n*-octanol/water log *PC* values, and the proposed cut-off point for intestinal absorption is thus at log $P \simeq 0$. The proposed lipophilicity cut-off point for absorption is not sharp, as shown by a few compounds near this critical lipophilicity. Tyrothricin, for example, is in the group with log $PC > 0$, and thus absorption is indicated; but in fact, absorption is poor in the human. Another apparent exception is oxytetracycline, which is placed in the list with log $PC < 0$, where absorption is not indicated. Oxytetracycline is absorbed upon oral administration in the human, although incompletely. Cycloserine is absorbed orally in spite of the low isoamyl alcohol/water partition coefficient. However, this can be rationalized on the basis of aqueous pore transport due to low molecular weight.

Based on this analysis, adoption of log $PC \simeq 2$ as the preferred lipophilicity for improving the absorption of drugs is reasonable. Again, this applies to large, bulky molecules, where the aqueous pore permeability coefficient is negligible. *N*-Hydroxyurea, with estimated *n*-octanol/water log $PC = -2.54$, would not be expected to be absorbed, on the basis of its hydrophilic nature. The compound, however, is well absorbed upon oral administration (*35*), and this is attributed to facile aqueous pore transport resulting from its low molecular weight. 5-Fluorouracil, with an *n*-octanol/water log *PC* of − 0.95 (*36*), is absorbed

upon oral administration in humans (*37*) in spite of its unfavorably low lipophilicity. Absorption presumably occurs via the aqueous pore pathway, because of the relatively small size of the drug.

III. Approaches to Improved Drug Efficiency

There are three basic approaches to the problem of improving drug absorption: the formulation approach, the prodrug (reversible derivative) approach, and the analog (irreversible derivative) approach. Each of these approaches should be contemplated whenever there is a problem with drug efficiency.

A. The Formulation Approach

Since it involves no synthesis and only minimal IND amending, the formulation approach is the quickest, easiest, and least expensive of the three; it should therefore be considered first. By altering the formulation, it is usually possible to increase or decrease the drug's apparent aqueous solubility and/or dissolution rate. Micronization, milling, cosolvents, buffers, micelles, complexing agents, coprecipitates, etc. can be used to increase solubility and dissolution. Various insoluble matrices, controlled crystallization, buffers, etc. can be used to retard dissolution. The formulation approach is of almost no value for increasing membrane transport, and it cannot always be used to alter drug stability. The enteric coating of acid-labile drugs is an exception. Because these techniques are generally regarded as part of the art of formulation, it is difficult to obtain patent coverage of a modified formulation for an existing drug. In general, the formulation approach, when applied to drugs which are insoluble (or too soluble) or which are degraded in the stomach, offers a high probability of success for a comparatively small investment.

B. The Analog Approach

Analogs are generally defined as that type of molecular modification consisting of a skeletal transformation or a substituent group synthesis. Thus, introduction of a methyl or other functional group, such as a hydroxyl or a nitro, is considered as analog formation.

These irreversible derivatives are synthesized with the intent of altering or improving intrinsic activity. Because they often interact differently with the receptor, analogs frequently have a different spectrum of ac-

tivities and side effects than the parent drug. For these reasons the analog approach is not normally used to improve absorption although, in cases of metabolism limited bioavailability, this approach is worthy of consideration.

C. THE PRODRUG APPROACH

The prodrug approach lies somewhere between the formulation and analog approaches with respect to both time and cost of development. The term prodrug is employed for that class of drug derivative that *in vivo* is hydrolyzed or converted by other mechanisms to the parent compound. The important distinction between analogs and prodrugs is that the former is biologically active per se whereas the latter requires *in vivo* conversion to the parent drug for eliciting biological activity. The primary utility of analogs is to improve potency and to achieve specificity of action, whereas the prodrug is used to improve pharmaceutical or biological properties. The latter properties are collectively termed biopharmaceutical properties. Because it does not alter the primary structure of the parent drug, prodrug synthesis is usually much less difficult than analog synthesis and the probability of a prodrug being active is much greater than for an analog. In the case of the most common prodrugs, i.e., simple esters and amides, it might be possible to speed up the development and registration process by using an abbreviated preclinical testing program.

One of the important applications of prodrugs is in improving oral absorption. The unique feature of the prodrug is that the physicochemical properties of the resulting derivative can be carefully tailored by means of structural modification of the pro-moiety. The intrinsic activity of the parent drug is assured through *in vivo* cleavage of the prodrug.

The prodrug approach has been successfully applied to a wide variety of drugs (*1–10*). It is most effective when an undesirable characteristic of the parent drug needs to be eliminated, especially if that characteristic can be related to a physical-chemical property such as melting point, boiling point, solubility, or partition coefficient. In these instances, the characteristic to be altered is related as quantitativity as possible to a physical-chemical property. (For example, taste can be related to solubility for a series of drugs.) The physical-chemical property is then related to chemical structure (e.g., solubility and alkyl chain length). Finally, prodrugs are synthesized that have the desired physical chemical property and are, it is hoped, free of the undesirable biological property.

The design of prodrugs is subject to a number of economic and biological constraints. Some of these are now listed. The choice of the derivative or pro-moiety depends upon a number of factors such as:

1. The economics of synthesizing the prodrug should be such that the cost is not significantly higher than that of the parent drug. Single-step prodrug synthesis using inexpensive pro-moieties is preferred.
2. The pro-moiety selected should yield a fragment, upon cleavage of the prodrug, that is free of toxicity or side effects and preferably free of biological activity. The choice of the pro-moiety is also dependent upon the nature of the parent drug. For example, the pro-moiety chosen for a parent drug with a dose of 1 mg may not be acceptable with another parent drug given at a high dosage (such as 500 mg) because of either the unacceptably high weight of an equivalent dose of the prodrug or the toxicity or other side effects of the pro-moiety at this high dosage level.
3. The pro-moiety should be rapidly cleaved once the prodrug is absorbed in order to elicit a pharmacokinetic profile similar to that of the parent drug. A prodrug with both improved absorption and prolonged action through chemical sustained release requires that the pro-moiety be carefully tailored with respect to lability. Only brief mention of the latter will be made since the theoretical principles underlying chemical sustained release have been developed elsewhere (*11, 12*).
4. The intact drug should not produce any unusual drug distribution patterns that could lead to unfavorable tissue distribution.

IV. Controlling Physical Properties of Nonelectrolytes

In the first section, we discussed the roles of solubility and partition coefficient in determining the absorption and dissolution rates of drugs. It was found that a prodrug which has a log *PC* above 2 and which has an aqueous solubility equal to or greater than that of the parent drug is likely to show improved efficiency. In this section, we will attempt to provide some useful guidelines for designing prodrugs that meet these requirements.

If drug design is to become a reality, we must learn to control both of these properties simultaneously, along with such other key parameters as melting point, dissociation constant, chemical stability, and enzymatic lability.

In most cases, a prodrug is made by linking a hydrocarbon group to the parent drug by an ester, amide, or ether linkage. However, in certain cases other chemical moieties are utilized. The selection of the optimum chemical linkage will not be discussed in this chapter. This subject has been discussed in great detail by Sinkula and others (*1–10*). We will first

consider the effects of the more common structural modifications on partition coefficient and then the effects of these modifications on solubility and melting point.

A. PARTITION COEFFICIENTS

Next to melting point and aqueous solubility, partition coefficient is probably the most frequently reported physical property of drugs. Although partition coefficients have been recorded in nearly a hundred partitioning solvent systems, more than 25% of the available data refers to the octanol–water system (*38*). This system has become the standard reference system for correlating biological data.

Fortunately, the partition coefficients of most drugs can be predicted reasonably well by any of several group contribution approaches (*36, 38–43*) or by correlation with other molecular properties (*44–50*). The concept of predicting partition coefficients from group contribution values was first proposed by Hansch (*39–41*) who (by analogy to the Hammett and Taft approaches) showed that:

$$\log PC_{RG} - \log PC_{RH} = \pi_G \tag{8}$$

where π_G is a constant which is characteristic for any given atom or group. Some of the more important group π-values are given in Table III. Recently, Nys and Rekker (*42*) proposed a new system of f-values for the calculation of partition coefficients. The Nys and Rekker f-(or fragment) values are defined so that:

$$\log PC_{\text{molecule}} = \sum_{\text{all G}} f_G \tag{9}$$

Also the f-values were derived from statistical analysis of a much larger data base than was available in the early stages of π-value development.

The f-values of some of the most commonly used structural units are listed in Table IV. Because the actual contribution of a particular group to the overall molecular partition coefficient is dependent upon its molecular environment, the f-values listed are averages based on a large number of comparisons.

More recently, Leo (*43*) derived a new set of f-values which is claimed to be more accurate than those of Nys and Rekker. Unfortunately, the increase in accuracy claimed for the Leo f-values is more than offset by the difficulty or inconvenience of using the combination of group and bond contributions. These values are listed in Table V. The π-values and both sets of f-values were obtained by fitting the partition coefficients of a large number of compounds to linear functions of their constituent groups. The only fundamental difference among the three systems is the manner in

TABLE III
HANSCH π-VALUES FOR SOME COMMON ATOMS AND GROUPS[a]

Substituent	π-Value	
Hydrocarbon		
CH_3	0.50	
CH_2	0.50	
CH (saturated)	0.50	
CH (unsaturated)	0.35	
C (saturated)	0.50	
C (unsaturated)	0.35	
H	0.00	
$CH{=}CH_2$	0.70	
C_6H_5	2.15	
C_6H_4	2.13	
C_6H_3	2.13	
C_5H_4N	0.65	
Nonhydrocarbon	**Aliphatic**	**Aromatic**
F	−0.17	0.13
Cl	0.39	0.76
Br	0.66	0.94
I	1.00	1.15
S	−0.05	1.12
SH	.	.
O	−0.98	−0.52
OH	−1.16	−0.67
OCH_3	−0.47	−0.02
COO	−0.77	−0.55
COOH	−1.26	−0.28
C=O	−1.21	−1.05
N	.	.
NH	.	.
NH_2	−1.19	−1.23
NO_2	−0.82	−0.28
$CONH_2$	−1.71	−1.49
C=N	−0.84	−0.57

[a] Corrections: Branching −0.20; Ring formation −0.09.

TABLE IV
NYS AND REKKER *f*-VALUES FOR SOME COMMON ATOMS AND GROUPS

Substituent	*f*-Value	
Hydrocarbon		
CH_3	0.70	
CH_2	0.53	
CH (saturated)	0.24	
CH (unsaturated)	0.36	
C (saturated)	0.14	
C (unsaturated)	0.16[a]	
H	0.20	
$CH{=}CH_2$	0.93	
C_6H_5	1.90	
C_6H_4	1.73	
C_6H_3	1.48	
C_5H_4N	1.22	
Nonhydrocarbon	**Aliphatic**	**Aromatic**
F	−0.41	0.43
Cl	0.06	0.93
Br	0.24	1.17
I	0.59	1.46
S	−0.51	0.14
SH	0.00	0.62
O	−1.54	−0.46
OH	−1.44	−0.37
OCH_3	−0.83	0.24
COO	−1.28	−0.40
COOH	−1.00	0.00
C=O	−1.69	−0.99
N	−2.13	−1.07
NH	−1.86	−0.93
NH_2	−1.38	−0.91
NO_2	−1.06	−0.09
$CONH_2$	−1.99	−1.26
C=N	−1.13	−0.20

[a] 0.30 for carbon atom shared by 2 aromatic rings.

TABLE V
Leo *f*-VALUES FOR SOME COMMON ATOMS AND GROUPS[a]

Substituent	*f*-Value	
Hydrocarbon		
CH_3	0.89	
CH_2	0.66	
CH (saturated)	0.43	
CH (unsaturated)	0.36	
C (saturated)	0.20	
C (unsaturated)	0.13	
H	0.23	
$CH{=}CH_2$	0.73	
C_6H_5	1.90	
C_6H_4	1.67	
C_6H_3	1.44	
C_5H_4N		
Nonhydrocarbon	Aliphatic	Aromatic
F	−0.38	0.37
Cl	0.06	0.94
Br	0.20	1.09
I	0.60	1.35
S	0.79	0.03
SH	.	.
O	−1.81	−0.57
OH	−1.64	−0.40
OCH_3	.	.
COO	−1.49	−0.56
COOH	−1.09	−0.03
C=O	−1.90	−0.32
N	.	.
NH	−2.11	−1.03
NH_2	−1.54	−1.00
NO_2	−1.26	−0.02
$CONH_2$	−2.18	−1.26
C=N	−1.28	−0.34

[a] Corrections: single bonds (non-ring) after the first − 0.12; single bonds (ring) after the first − 0.09; chain branching − 0.13; group branching − 0.22.

which hydrocarbons are treated. It is, therefore, not surprising that π- and f-values for nonhydrocarbon groups are readily interrelatable. For aromatic substituents:

$$\pi - 0.20 \cong f_{\text{Nys}} \cong f_{\text{Leo}} - 0.06 \tag{10}$$

and for aliphatic substituents:

$$\pi - 0.39 \cong f_{\text{Nys}} \cong f_{\text{Leo}} + 0.14 \tag{11}$$

Any differences beyond this are mostly statistical, being due to differences in data selection. The treatment of hydrocarbons by these schemes is summarized in Table VI. Note that f_{Nys} is usually lower than f_{Leo} and higher than π so that the differences in nonfunctional group values are compensated for by the differences in functional group values. In other words, the Hansch π-values give more weight to functional groups and less weight to hydrocarbon groups than Nys and Rekker f-values, whereas Leo f-values give more weight to hydrophobic groups and less weight to

TABLE VI
COMPARISON OF π- AND f-VALUES FOR SOME HYDROCARBON MOIETIES

Moiety	π	f_{Nys}	f_{Leo}
Aliphatic			
CH_3	0.50	0.70	0.83
CH_2	0.50[a]	0.53	0.54[b]
CH	0.35[a]	0.24	0.24[b]
C	0.20[a]	0.14	−0.06[b]
Olefinic			
$CH{=}CH_2$	0.73	0.93	1.15
CH—CH	0.73	0.74	0.98
Aromatic			
C_6H_5	2.13	1.90	1.90
C_6H_6	2.13	1.73	1.67
C_6H_3	2.13	1.48	1.44

[a] Corrected for chain branching but not group branching or ring formation.

[b] Corrected for single bonds and chain branching but not group branching or ring formation.

polar groups. The net result is that for most compounds the three calculations give nearly equivalent results.

The use of the above techniques is illustrated by the examples of Table VII. Note that in most cases, the three calculations are in rather good agreement with one another. Certainly, the agreement among the calculated values is better than the agreement between the experimental values reported by different workers (cf. toluene and butylamine). We have chosen the Nys and Rekker f-values in this report primarily because of their greater ease of use.

Table VIII gives the logarithms of the octanol–water partition coefficients of over 60 mono- and di-substituted benzenes calculated from the f-values of Table III along with the reported observed values. In most, but not all cases, the calculated values are in good agreement with the experimentally observed values.

The effect of a prodrug modification such as converting a phenolic hydroxyl group to an acetate group is determined from the difference in f-values for the aromatic OH and the acetate (COO + CH_3). Note that it is not necessary to know the partition coefficient of either the parent drug or the prodrug in order to calculate the probable change in log PC, $\Delta\log PC$. The values of $\Delta\log PC$ for the most common prodrug modifications are given in Table IX.

The use of Tables III–IV or the more extensive tables of Nys and Rekker (*42*). Hansch (*41*), and Leo (*43*) enable the medicinal chemist to design prodrugs of nearly any poorly absorbed parent drug that would have partition coefficients in the desired 10^{-2}–10^{-3} range. For example, a compound such as salicylic acid, with $\log PC = 1.36$, could be converted to its acetate, methyl ether, or methyl ester to produce compounds having log PC values of around 2. The choice of modification should depend largely on enzymatic, synthetic, and other chemical factors. It should also be dependent upon the effects of the various structural modifications on the drug's melting point and aqueous solubility.

It is important to realize that although group contribution values give good estimates of log PC for most compounds, they cannot account for interactions between nonbonded atoms or groups. Therefore, for certain ortho-substituted compounds and for certain "bulky" or flexible molecules, estimated values can be expected to differ somewhat from experimentally determined partition coefficients. Tute (*49–50*) has reviewed some of the cases for which major discrepancies exist. Fortunately, these errors are usually not great enough to cause significant problems in structure–activity correlations. In fact, the use of f-values usually gives partition coefficients that are more accurate than the biological data with which they will be correlated.

TABLE VII
COMPARISON OF TECHNIQUES FOR CALCULATING PARTITION COEFFICIENTS

Drug	Substitutent	Hansch π-value	Nys and Rekker f-value	Leo f-value	Reported values (*38*)
Toluene	C_6H_5	2.13	1.90	1.90	
	CH_3	0.50	0.70	0.89	
	Bonds				
	Toluene	2.63	2.60	2.79	2.69, 2.73, 2.11, 2.80
1-Butylamine	CH_3	0.50	0.70	0.89	
	$(CH_2)_3$	1.50	1.59	1.98	
	NH_2	−1.19	−1.38	−1.54	
	Bonds			−0.36	
	1-Butylamine	0.81	0.91	0.97	0.97, 0.88, 0.81, 0.68, 1.02
2-Butanone	$(CH_3)_2$	1.00	1.40	1.88	
	CH_2	0.50	0.53	0.66	
	C=O	−1.21	−1.69	−1.90	
	Bonds			−0.24	
	Branching			−0.22	
	2-Butanone	0.29	0.24	0.18	0.26, 0.29
Aspirin	C_6H_4	2.13	1.73	1.67	
	COOH (arom)	−0.28	0.00	−0.03	
	COO (aliph)	−0.77	−1.28	−0.56	
	CH_3	0.50	0.70	0.89	
	Bonds			−0.48	
	Branching				
	Aspirin	1.58	1.15	1.47	1.19, 1.23
Cyclohexanol	$(CH_2)_5$	2.50	2.65	3.30	
	CH	0.50	0.24	0.43	
	OH	−1.16	−1.44	−1.64	
	Bond	−		−0.12	
	Branching	−0.20		−0.22	
	Ring	−0.09		−0.45	
	Cyclohexanol	1.55	1.45	1.30	1.23
Phenethanol	C_6H_5	2.13	1.90	1.90	
	$2CH_2$	1.00	1.06	1.32	
	OH	−1.16	−1.44	−1.64	
	2 bonds			−0.24	
	Phenethanol	1.97	1.52	1.34	+1.36, −0.39

TABLE VIII

CALCULATED AND OBSERVED LOG (PARTITION COEFFICIENTS) OF SOME SUBSTITUTED BENZENES

		H		CH_3		Cl		NO_2		NH_2		OH		COOH	
		calc	obs	calc	obs	calc	obs	calc	obs	calc	obs	calc	obs	calc	obs
H		2.08	2.13	2.60	2.70	2.83	2.84	1.82	1.87	0.98	0.90	1.52	1.48	1.90	1.87
	o			3.13	3.12	3.36	3.42	2.34	2.30	1.52	1.29	2.06	1.95	2.34	—
CH_3	*m*			3.13	3.20	3.36	3.28	2.34	2.42	1.52	1.41	2.06	1.96	2.34	3.27
	p			3.13	3.15	3.36	3.33	2.34	2.40	1.52	1.39	2.06	1.94	2.34	2.27
	o					3.59	3.38	2.57	2.24	1.75	—	2.29	0.97	2.66	1.98
Cl	*m*					3.59	—	2.57	2.43	1.75	1.89	2.29	2.48	2.66	1.83
	p					3.59	3.39	2.57	2.40	1.75	1.83	2.29	2.39	2.39	1.89
	o							1.55	1.58	0.73	1.79	1.26	1.76	1.64	—
NO_2	*m*							1.55	1.49	0.73	1.37	1.26	2.00	1.64	1.83
	p							1.55	1.47	0.73	1.39	1.26	1.91	1.64	(0.68)(1.89)
	o									−0.09	0.15	0.45	0.57	0.82	—
NH_2	*m*									−0.09	0.93	0.45	0.16	0.82	—
	p									−0.09	—	0.45	0.04	0.82	0.68
	o											0.98	0.95	1.36	2.25
OH	*m*											0.98	0.79	1.36	1.50
	p											0.98	0.55	1.36	1.58
	o													1.73	—
COOH	*m*													1.73	1.66
	p													1.73	—

TABLE IX
VALUES OF ΔLOG *PC* FOR SOME PRODRUG MODIFICATIONS

	Arom	Aliph
$COOH \rightarrow COOCH_3$	0.30	0.42
$COOC_2H_5$	0.83	0.94
$COOC_6H_5$	1.50	1.62
$CONH_2$	− 1.26	− 0.99
$OH \rightarrow OCH_3$	0.61	0.60
OC_2H_5	1.14	1.13
$OCOCH_3$	0.67	0.86
$OCOC_6H_5$	1.87	1.16
$NH_2 \rightarrow NHCOCH_3$	0.21	—
$NHCOC_6H_5$	0.09	—

B. MELTING POINTS

In spite of the fact that it is the most frequently determined and recorded property of chemical compounds, there are no available means of predicting melting point from chemical structure. The effect that a particular chemical substituent will have on melting point is highly dependent upon a variety of factors that are too complex to be analyzed here. (The reader is referred to references *47*, *51*, and *52* for discussions of these factors.) We can, however, evaluate the effects of the major chemical modifications involved in prodrug formation, i.e., those listed in Table X.

TABLE X
VALUES OF ΔMP FOR SOME PRODRUG MODIFICATIONS

	Arom	Aliph
$COOH \rightarrow COOCH_3$	− 100	− 70
$COOC_2H_5$	− 150	− 70
$COOC_6H_5$	70	—
$CONH_2$	50	20
$OH \rightarrow OCH_3$	− 100	− 30
OC_2H_5	− 100	− 50
$OCOCH_3$	− 20	− 20
$OCOC_6H_3$	0	0
$NH_2 \rightarrow NHCOCH_3$	60	—
$NHCOC_6H_5$	60	—

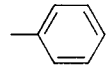
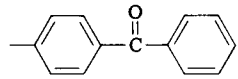
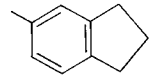
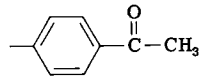
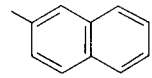
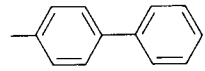
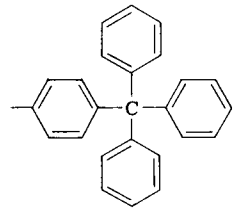
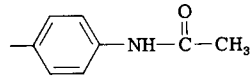
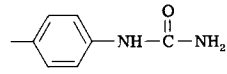
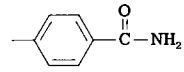
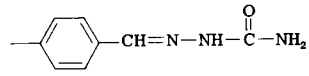
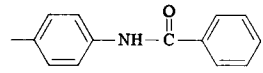
In order to do this we have compared melting point data for a variety of COOH-, OH-, and NH_2-containing compounds and for some of their derivatives. These values were mostly taken from Weast (*53*). It is obvious from the data that there is no truly constant increment in melting point, ΔMP, that results from a particular chemical modification. In spite of the rather broad scatter in the ΔMP values, there are certain fairly reliable generalizations that can be made. Most of the ΔMP values reported for a particular modification do not differ by more than 50 degrees from the values given in Table X. Because of the lack of true constancy of these values, it is more appropriate to refer to them as expected values than as calculated values. The values listed are in qualitative agreement with what would be expected from a consideration of dipole moment and hydrogen bond formation. Replacing an acidic proton with an alkyl group consistently produces a decrease in melting point. Methyl and ethyl groups are nearly equivalent in their effect on melting point.

The effects of more elaborate chemical modifications are generally more difficult to quantitate than the ones just described. Because the melting point of a compound is dependent upon its overall size, compactness, and symmetry, as well as its ability to form intramolecular and intermolecular hydrogen bonds, most substituents will have different effects when added to different parent compounds. In the case of prostaglandin prodrugs, certain generalizations can be made: Pro-moieties that are high melting tend to produce high-melting prodrugs [e.g., a carboxylic acid ester of *p*-benzamidophenol (which melts at 218°C) will probably have a much higher melting point than the parent free acid (*54*) (see Table XI)]. The effects of a series of substituents on the melting points of a compound will parallel the effects of that series of substituents on a related parent compound (i.e., the effect of a series of substituents on the *MP* of benzoic acid, testosterone, prostaglandin $F_{2\alpha}$, or lincomycin will be similar to the effect of that same series on phenylacetic acid, norandrolone, prostaglandin E_2, or clindamycin, respectively).

Hydrophilic groups that interfere with the symmetry or compactness of the parent drug will often reduce the melting point. This is especially true if the parent compound is highly crystalline. Groups such as isopropyl, cyclohexyl, cyclopentyl, and cyclopentoxy, which are not coplanar with the parent molecule, are particularly effective in reducing the melting point.

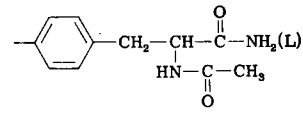
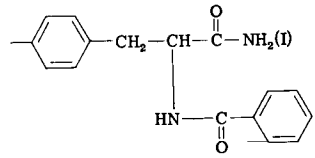
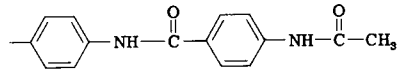
The use of hydrophobic groups that reduce the melting point of very high-melting drugs often result in a fortuitous set of circumstances; i.e., they increase partition coefficient and, as we will see below, can also increase aqueous solubility.

TABLE XI
MELTING POINTS OF PGE_2 C_1 ESTERS AND THEIR PARENT PHENOLS[a]

Compound number	PGE_2 C_1 ester and phenol	MP (°C) ester	MP (°C) phenol
1		< 25	43
2		< 25	131
3		46	52
4		77	109
5		80	123
6		92	165
7		97	283
8		103	168
9		106	171
10		107	197
11		126	224
12		134	218

(*Continued*)

TABLE XI—*Continued*

Compound number	PGE_2 C_1 ester and phenol	MP (°C) ester	MP (°C) phenol
13		139	224
14		139	224
15		152	275

[a] From Morozowich *et al.* (*54*).

C. Aqueous Solubility

Although aqueous solubility is one of the most important of the physical properties that affect biological activity, its importance is frequently overlooked. Very few medicinal chemists have attempted to design drugs that have a specified aqueous solubility. This is because solubility is a more difficult property to understand and predict than is partition coefficient (*57–59*). The major problem in understanding aqueous solubility is the fact that it is dependent upon interactions stabilized by the geometry of the crystal lattice as well as interactions between molecules in pure liquid and in solution.

An understanding of solubility can be facilitated by breaking down the solution process into one component which is dependent only on the crystal lattice energy and another which is only dependent upon the interactions of the liquid drug and the liquid solute. The latter term can be studied by considering only the aqueous solubility of liquids. Then the solubility of crystalline compounds can be handled by adjusting liquid solubility by the appropriate crystal energy term.

1. *Liquids*

Since the partition coefficient of a drug is approximately equivalent to the ratio of the solubilities of the drug in oil and in water (*55*); and since most semipolar liquids are completely miscible with octanol (*56*), it is not surprising that for semipolar liquids there is a linear relationship between the octanol–water partition coefficient and the aqueous solubility of liquids. Hansch correlated the solubilities of some 150 liquid nonelectro-

lytes with their partition coefficients, and found (*57*):

$$\log S_w = -1.339 \log PC + 0.978 \quad (12)$$
$$r = 0.935, \quad s = 0.472$$

Unfortunately, this correlation was based primarily on calculated partition coefficients, and all of the hydrocarbon partition coefficients used in the study are low by a factor of 6.3 (0.8 in log units)* (*47*). Correcting the hydrocarbon values will significantly increase the coefficient of log *PC* (i.e., make it less negative) and improve the correlation.

Using more recently obtained solubility and partition coefficient values (*56*), it has been found that an equation of the form

$$\log S = -1.07 \log PC + 0.67 \quad (13)$$
$$r = 0.954, \quad s = 0.344$$

is more accurate. The fact that the coefficient of log *PC* is so close to − 1.0 implies that the product of $S \times PC$ is approximately constant for liquids. Consequently, any structural modification that increases the partition coefficient of a liquid and that does not raise the melting point above room temperature will result in a comparable loss in aqueous solubility.

In terms of absorption and dissolution, this means that any chemical modification that produces an increase in partition coefficient, and thus absorption rate, is likely to produce a decrease in solubility and dissolution rate. If the parent drug is very soluble with respect to its dose, this loss in solubility is of no consequence; but if the parent drug is poorly soluble in water, the gain in absorption can be completely offset by the loss in dissolution rate.

2. *Crystals*

On the basis of a semi-empirical derivation, Yalkowsky, Valvani, and Roseman (*55, 56, 60*) found that the aqueous room temperature solubility of crystalline nonelectrolytes, which do not have long flexible polymethylene chains, can be estimated by

$$\log S^c = -\log PC - 0.01\, MP + 0.5 \quad (14)$$

In using this equation for substances that are liquid at room temperature, it is necessary to replace *MP* by 25°C and revert to

$$\log S^L = -\log PC + 0.25 \quad (15)$$

which has the general form of Eq. (12) and which fits the data quite well.

* C. Hansch, personal communication.

The significance of the coefficients of log *PC* and *MP* in Eq. (14) are nicely illustrated by its application to hydrocarbons. Table XII gives the values of log *PC*, *MP*/100, and − log *S* for a variety of aliphatic and aromatic hydrocarbons. Note that the values of − log *S* calculated by Eq. (9) are in very good agreement with the observed values. The significance of log *PC* is readily seen from the data for the liquids, for which *MP*/

TABLE XII
CALCULATION OF AQUEOUS SOLUBILITIES OF SOME HYDROCARBONS

Hydrocarbon	log *PC*[a]	*MP*/100	− log *S*	
			calc[b]	obs
Butadiene	1.99	0.25	1.74	1.85
Butyne	1.46	0.25	1.21	1.28
1-Butene	2.38	0.25	2.12	2.40
Butane	2.89	0.25	2.64	2.60
Pentadiene 1,4	1.48	0.25	1.23	1.65
Pentyne	1.98	0.25	1.73	1.67
1-Pentene	(2.69)	0.25	(2.44)	2.68
n-Pentane	3.30	0.25	3.05	3.27
Neopentane	3.11	0.25	2.86	3.35
Benzene	2.13	0.25	2.55	2.26
1,4-Cyclohexadiene	2.30	0.25	2.05	1.95
Hexadiene 1,5	2.45	0.25	2.20	2.69
Hexyne	(2.51)	0.25	(2.26)	2.39
Cyclohexene	2.86	0.25	2.61	2.60
1-Hexene	(3.33)	0.25	(2.98)	3.26
Cyclohexane	3.44	0.25	3.09	3.07
n-Hexane	(3.52)	0.25	(3.27)	3.80
Neohexane	3.82	0.25	3.58	3.61
2,3-Dimethylbutane	3.85	0.25	3.60	3.62
Toluene	2.80	0.25	2.55	2.26
Cycloheptane	(3.76)	0.25	(3.51)	3.52
Heptene	(3.75)	0.25	(3.50)	3.52
Heptane	(4.09)	0.25	(3.84)	3.53
Styrene	2.95	0.25	2.60	2.54
o-Xylene	3.12	0.25	2.87	2.78
m-Xylene	3.17	0.25	2.92	2.83
p-Xylene	3.30	0.25	3.05	2.83
Ethylbenzene	3.15	0.25	2.90	2.81
α-Methylstyrene	(3.48)	0.25	(3.23)	3.01
n-Propylbenzene	3.63	0.25	3.38	3.00
l-Propylbenzene	3.66	0.25	3.41	3.16
Mesitylene	3.42	0.25	3.17	3.20

TABLE XII—*Continued*

Hydrocarbon	log *PC*[a]	*MP*/100	− log *S*	
			calc[b]	obs
Naphthalene	3.37	0.80	3.67	3.63
n-Butylbenzene	(4.19)	0.25	(3.84)	3.45
t-Butylbenzene	(4.11)	0.25	(3.91)	3.65
p-Cymene	(4.07)	0.25	(3.82)	2.59
Durene	4.00	0.80	4.30	4.20
1-Methylnaphthalene	(3.70)	0.25	3.45	3.74
2-Methylnaphthalene	(3.70)	0.37	3.57	3.76
n-Pentylbenzene	(4.72)	0.25	4.47	
t-Pentylbenzene	(4.67)	0.25	4.42	4.15
Biphenyl	4.05	0.70	4.25	4.31
2,3 Dimethylnaphthalene	(4.06)	1.04	(4.60)	4.80
2,6 Dimethylnaphthalene	(4.06)	1.10	(4.66)	5.08
1,5 Dimethylnaphthalene	(4.06)	0.83	(4.39)	4.74
2-Ethylnaphthalene	(4.24)	0.25	(3.99)	4.29
Hexamethylbenzene	4.31			
Fluorene	4.18	1.16	4.84	4.91
Diphenylmethane	4.14	0.25	3.89	4.07
Anthracene	4.45	2.18	6.13	6.30
Phenanthrene	4.54	1.00	5.04	4.99
trans-Stilbene	(4.48)	1.24	(5.22)	5.79
Acenaphthene	(4.32)	0.96	(4.78)	4.50
Pyrene	4.88	1.56	5.94	6.09
Chrysene	(5.91)	2.54	(7.94)	7.72
1.2-Benzanthracene	(5.91)	1.62	(7.03)	7.34
Naphthacene	(5.91)	3.35	(8.76)	8.18
Triphenylene	(5.16)	1.99	(6.65)	6.73
Picene	7.19	3.66	(10.35)	8.05

[a] Log *PC* values in parentheses are calculated from *f*-values, all others are from (*1*).

[b] −Log *S* values in parentheses are calculated from calculated log *PC* values.

100 = 0.25. For these compounds the solubility is a direct function of log *PC*. Also, naphthalene and 1,5-dimethylnaphthalene have similar melting points but very different log *PC* and log *S* values.

The role of melting point can be seen by comparing the solubilities of compounds having similar partition coefficients. Anthracene and phenanthrene both have log *PC* values of about 4.5; but anthracene, which melts over 100 degrees higher than phenanthrene, is more than 10 times less

soluble in water. The same is true for 1,2-benzanthracene and naphthacene and for naphthylene and *p*-xylene.

This treatment is strictly applicable only to nonelectrolytes. However, it can be empirically extended to cover the intrinsic* solubilities of weak electrolytes as follows:

$$\log S_w = -\log PC - 0.01\,MP + 0.5 + A \tag{16}$$

where $A = 1.0$ if the molecule has an acidic proton and $A = 0$ if it does not. Based on a fairly limited sampling, it does not appear necessary to account for basic groups.

The calculation of the solubilities of a number of disubstituted benzenes from melting points and partition coefficients, by Eq. (16), is illustrated in Table XIII. Note that this equation provides a reasonable (within a factor of three, or one-half of a log unit) estimate of solubility in most cases. This agreement is particularly good in view of the fact that reported values of both S and PC are frequently only reliable to within a factor of two, or one-third of a log unit.

Besides errors in reported solubility and partitioning values, errors in the melting point value can contribute to the error in the calculations. Polymorphism, hydration, and solid–solid phase transitions can render the foregoing treatment invalid if they are not taken into account.

The application of Eq. (16) to a number of drugs and their derivatives is illustrated in Table XIV.

The agreement between the observed and calculated values of Tables XII, XIII, and XIV, while not always spectacular, is sufficient to help provide a basis for the design of prodrugs with optimum solubility properties. In most instances, the solubilities calculated by Eq. (16) are more reliable than those calculated by the $\Delta\log S$ values. This is expected since the former utilizes the drug's true melting point and is thus not subject to the errors inherent in the melting point estimation. The latter, while less accurate, is more applicable to designing drugs because it can be applied to nonexistent compounds.

It should by now be clear that the significance of controlling melting point lies not only in controlling ease of handling, but also in controlling solubility. Equation (16) states that if partition coefficient is held constant, a 100 degree increase in melting point will decrease solubility 10-fold, a 200 degree increase will decrease solubility 100-fold, and a 300 degree increase will result in a 1000-fold decrease in solubility. This concept provides a means of assessing the effect of polymorphic transition on aqueous solubility.

* The intrinsic solubility of a weak electrolyte is that of the free unionized species. See the next section.

TABLE XIII

CALCULATED AND OBSERVED LOG (SOLUBILITIES) OF SOME SUBSTITUTED BENZENES

		H		CH_3		Cl		NO_2		NH_2		OH		COOH	
		calc	obs	calc	obs	calc	obs	calc	obs	calc	obs	calc	obs	calc	obs
H		1.76	1.64	2.33	2.26	2.59	2.36	1.62	2.00	0.64	0.44	0.41	0.05	1.59	1.55
	o			2.70	2.78	3.18	2.53	2.05	2.32	1.15	0.82	0.72	0.62	—	2.05
CH_3	*m*			2.95	2.87	3.03	2.53	2.18	2.44	1.05	0.80	0.71	0.80	2.00	2.14
	p			2.90	2.80	3.08		1.92	2.39	1.33	1.05	0.79	0.70	2.57	2.55
	o					2.88	2.92	2.05		1.66	1.35	0.92	0.70	1.87	1.89
Cl	*m*					2.88	3.02	2.18	1.29	1.64	1.42	1.32	0.63	2.52	2.70
	p					3.41	3.26	2.43	3.26	2.03		1.32	0.63	3.50	3.52
	o							2.26	2.06	1.90	2.06	0.69	0.96	—	1.36
NO_2	*m*							1.89	2.18	2.01	2.18	1.49	1.01	1.75	1.70
	p							2.80	2.38	2.36	2.38	1.54	1.59	2.81	3.06
	o									0.69	0.71	0.80	0.72	1.16	1.36
NH_2	*m*									—	0.70	− 0.11	0.56	0.90	1.22
	p									—	1.41	0.38	0.81	1.07	1.22
	o											− 0.49	− 0.62	2.35	1.80
OH	*m*											− 0.62	− 0.81	2.00	1.26
	p											− 0.25	0.19	2.50	1.24
	o													—	1.39
COOH	*m*													2.46	3.03
	p													—	—

TABLE XIV

OBSERVED AND CALCULATED SOLUBILITIES OF SOME DRUGS AND THEIR DERIVATIVES

Parent compound derivative[a]	− log *S*	
	obs	calc[b]
Benzoic acid	1.55	1.59
benzamide	0.92	0.60
methyl benzoate	1.83	1.76
ethyl benzoate	2.17	2.47
Salicylic acid	1.80	2.35
salicylamide	1.82	2.18
aspirin	1.55	1.06
o-methoxybenzoic acid	1.80	2.35
p-Hydroxybenzoic acid	1.24	2.50
methyl *p*-hydroxybenzoate	1.78	1.80
ethyl *p*-hydroxybenzoate	2.21	2.15
p-methoxybenzoic acid	2.83	2.30
p-Nitrobenzoic acid	3.06	2.81
methyl *p*-nitrobenzoate	3.00	(2.65)
ethyl *p*-nitrobenzoate	2.95	(2.53)
Phthalic acid	1.39	(2.46)
methyl phthalate	2.11	2.10
ethyl phthalate	2.86	2.14
Nicotinic acid	0.87	1.40
nicotinamide	− 0.61	− 0.91
p-Aminobenzoic acid	1.22	1.07
methyl *p*-aminobenzoate	1.89	1.49
benzocaine	2.32	1.79
Butyric acid	0.85	0.67
methyl butyrate	0.79	1.00
ethyl butyrate	1.28	1.61
Succinic acid	0.15	− 0.23
succinamide	− 0.70	− 0.80
methyl succinate	0.96	1.00
Hexanoic acid	0.99	0.90
ethyl hexanoate	2.31	2.52
Phenol	0.05	0.41
anisole	1.01	0.85
phenethol	2.02	2.26
Catechol	0.62	0.49
guaiacol	0.91	0.10
o-Nitrophenol	0.96	0.69
o-nitroanisole	0.94	1.07

TABLE XIV (*Continued*)

Parent compound derivative[a]	− log *S*	
	obs	calc[b]
p-Nitrophenol	1.54	1.59
p-nitroanisole	1.96	2.12
Picric acid	1.30	1.73
2,4,6-trinitroanisole	2.41	2.82
2,4,6-trinitrophenethol	3.07	3.19
Methylparaben	1.24	2.50
p-methoxybenzoic acid methyl ester	2.41	2.48
Acetaminophen	1.05	1.00
methoxyacetanilide	1.15	0.89
phenacetin	1.39	0.91
acetoxyacetanilide	1.90	
o-Aminophenol	0.72	0.80
o-anisidine	0.94	1.07
Morphine	2.28	1.79
codeine	1.57	(2.39)
ethyl morphine	2.19	(3.64)
heroin	2.79	(3.49)
Testosterone	4.00	4.37
testosterone acetate	5.50	(5.09)
Prednisolone		
prednisolone acetate	4.30	(4.16)
Aniline	0.44	− 0.64
acetanilide	− 1.35	− 1.00
4-Aminophenol	0.81	0.44
acetaminophen	− 1.09	− 1.10
benzamidophenol		
Toluidine	1.05	1.33
acetotoluide	− 1.15	− 1.53
Ansidine	− 0.95	− 0.70
acetaniside	− 1.15	− 1.02
o-Nitroaniline	− 2.06	1.90
o-nitroacetanilide	− 1.91	1.43
p-Nitroaniline	− 2.38	2.36
p-nitroacetanilide	− 1.91	− 2.41

[a] The indented entries are derivatives.

[b] − Log *S* values in parenthesis are calculated from calculated log *PC* values.

It must be remembered that these guidelines for predicting partition coefficients, melting points, and solubilities are *just guidelines*. They work quite well in most cases, but they cannot be indiscriminantly relied upon, especially for classes of compounds for which their applicability has not been verified.

V. Controlling Physical Properties of Weak Electrolytes

We have so far considered the effects of nonelectrolyte structure on solubility, partition coefficient, and to some extent, melting point. Structural modifications of weak electrolytes can produce even more dramatic changes in physical properties. In many cases, the primary property-determining effect of a structural modification is the alteration of the dissociation constant of the parent drug. Before discussing the effects of structural modifications on dissociation constants of weak acids and bases, we will briefly review the relationship between pH and pK and between the properties that most directly relate to drug efficiency, i.e., solubility and partitioning.

A. Solubility and Dissociation Constants

Weak electrolytes differ from nonelectrolytes in that they can exist in either an ionized and/or an unionized form in solution and can be crystallized as either a neutral molecule or a salt. The relative concentrations of each species in solution is related to the pK of the drug and the pH of the solution by the Henderson–Hasselbalch equation:

$$\mathrm{pH} - \mathrm{p}K_a = \log[[A^-]/[HA]] \tag{17}$$

for weak acids; and

$$\mathrm{pH} - \mathrm{p}K_a = \log[B/[BH^+]] \tag{18}$$

for weak bases. The observed solubility of a weak acid is

$$s_{obs} = [A^-] + [HA]$$

If the solution is saturated with respect to the free acid $[HA] = S_{HA}$ and

$$S_{obs} = S_{HA}[1 + K_a/[H^+]] \tag{19}$$

similarly, for weak bases

$$S_{obs} = S_B\,[1 + [H^+]/K] \tag{20}$$

The ionized forms A^- and HB^+ will almost always have much higher solubilities and much lower partition coefficients than their corresponding

neutral species HA and B. The differences in solubility are normally so great that, in most cases, we can regard the ionized form as having nearly infinite solubility. Thus, we can assume that most weak electrolytes will have solubilities that are sufficiently high ($\log S \gg 0$) to insure complete solution at any conceivable dose, and partition coefficients close enough to zero ($\log P \ll 0$) to render them completely unabsorbable, except through pores. For the case in which protonated and unprotonated species have similar solubilities, see Kramer and Flynn (*60*).

The relationship between solubility and pH for a weak acid [i.e., Eq. (19)] is illustrated graphically in Fig. 6. Note that: at $pH \ll pK_a$ the observed solubility is equal to the intrinsic solubility of the free acid; at $pH = pK$ the observed solubility is equal to twice the intrinsic solubility; and as the pH is increased beyond the pK_a, the solubility increases 10-fold for each pH unit until the solubility of the anion is reached.

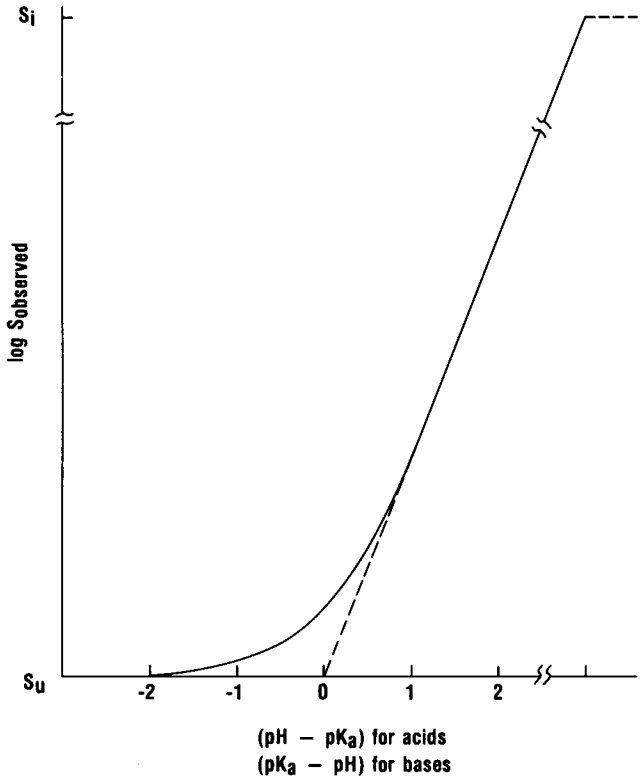


Fig. 6. Dependence of solubility upon pH for weak acids and bases. S_u–solubility of unionized species, S_i–solubility of ionized species.

B. Partition Coefficients and Dissociation Constants

Just as a free acid and its anion have different solubilities, so have they different partition coefficients. The apparent or observed partition coefficient of a weak acid is dependent upon the intrinsic partition coefficient of each form and the fraction, f, of the drug in that form is shown by the equation

$$PC_{obs} = f_{A^-} PC_{A^-} + f_{AH} PC_{AH} \tag{21}$$

Now since PC_A is usually negligible

$$PC_{obs} = f_{AH} PC_{AH} = (1 - H/K)\, PC_{AH} \tag{22}$$

Figure 7 shows the dependence of PC_E on pH for a weak acid. Note the inverse relationship between the dependence of solubility and partitioning on pH.

The logarithmic relationship between the observed and intrinsic parti-

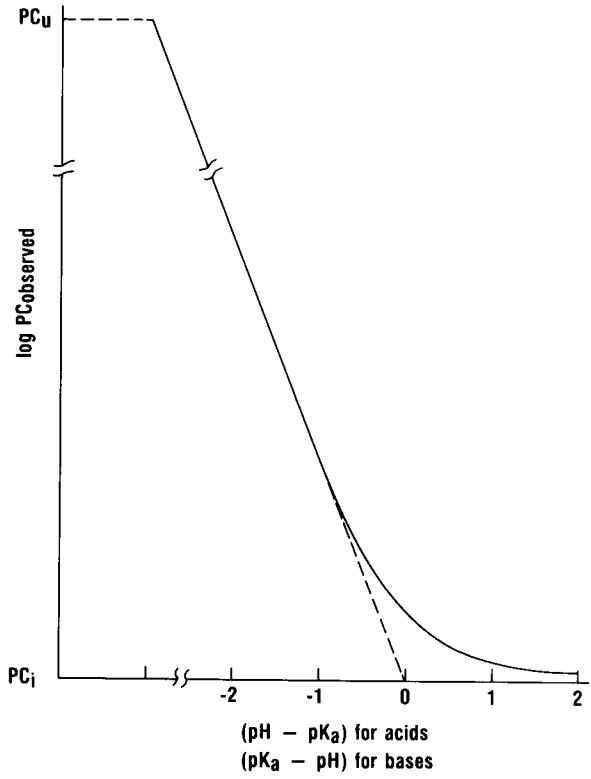


Fig. 7. Dependence of partition coefficient upon pH for weak acids and bases. PC_u–partition coefficient of unionized species, PC_i–partition coefficient of ionized species.

tion coefficients of a weak acid is

$$\log PC_{obs} = \log PC_I - \log (1 + 10^{pH-pK_a}) \quad (23)$$

Thus, when $pH - pK < -1$, $\log PC_E$ is essentially equal to $\log PC_I$. Conversely when $pH - pK_a > 1$, Eq. (19) is approximated by

$$\log PC_{obs} = \log PC_I + pK_a - pH \quad (24)$$

In the case of basic drugs, similar treatment leads to analogous expressions. These are summarized in Table XV.

Figures 8 and 9 show the relationship between pH and partition coefficient for some acidic and basic drugs, respectively.

C. pK_a VALUES OF IONIZABLE PRO-MOIETIES

The introduction of an ionizable pro-moiety to a parent nonionic drug is a useful means of altering both the solubility and the partition coefficient

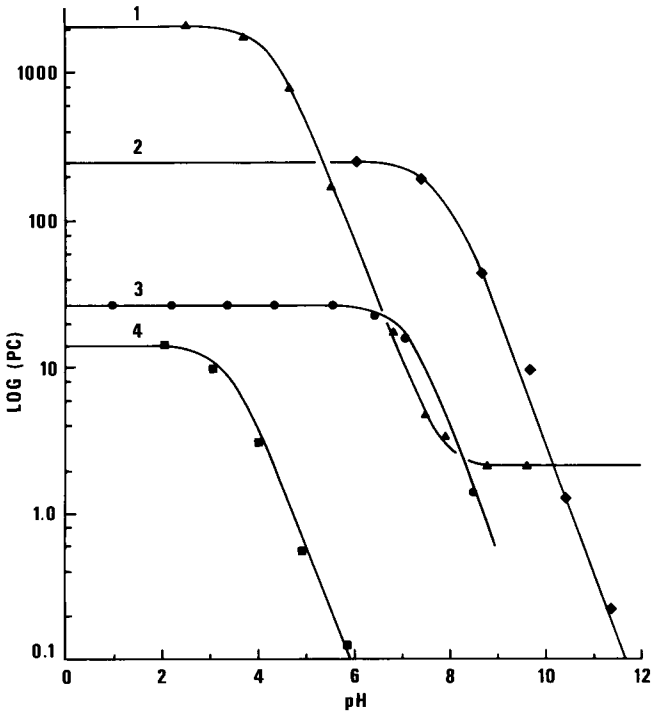


Fig. 8. Dependence of partition coefficient on pH for some weak bases (octanol–water at 37°C). 1–dioxypromethazine, 2–chlordiazepoxide, 3–aminophenazone, 4–promethazine, 5–caffeine. Adapted from G. LePetit (*61*).

TABLE XV

EXPRESSIONS FOR OBSERVED PARTITION COEFFICIENTS OF ACIDS AND BASES

	Acids	Bases
Exact expression	$\log PC_{obs} = \log PC_{HA} - \log(1 + 10^{pH-pK_a})$	$\log PC_{obs} = \log PC_B - \log(1 + 10^{pK_a-p})$
Approximate expression	$\log PC_{obs} \simeq \log PC_{HA} - pH + pK_a$	$\log PC_{obs} \simeq \log PC_B + pH - pK_a$
condition for above	$pH - pK_a > 1$	$pK_a - pH > 1$
Approximate expression	$\log PC_{obs} \simeq \log PC_{HA}$	$\log PC_{obs} \simeq \log PC_{HA}$
condition for above	$pH - pK_a < -1$	$pK_a - pH < -1$
At $pH = pK_a$	$\log PC_{obs} = \log PC_{HA} - 0.31$	$\log PC_{obs} = \log PC_B - 0.31$

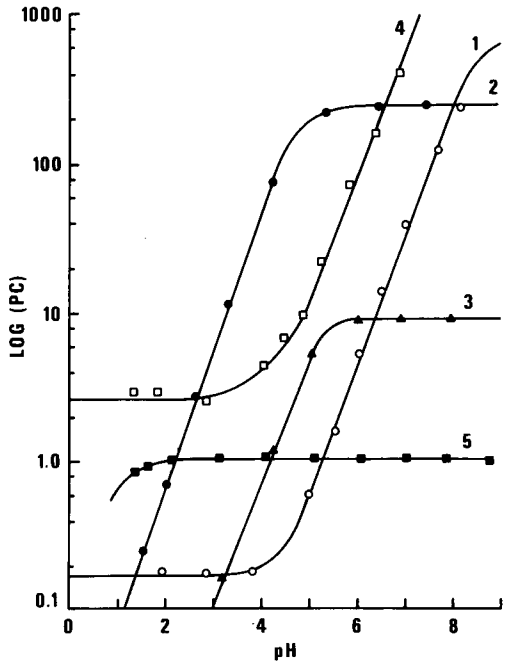


Fig. 9. Dependence of partition coefficient on pH for some weak acids (octanol–water at 37°C). 1–phenylbutazone, 2–phenytoin, 3–phenobarbital, 4–acetylsalicylic acid. Adapted from G. LePetit (*61*).

of the parent drug. As described previously, the purpose of this section is to briefly outline the approaches to estimation of the influence of structure on pK_a of use to prodrug design. Further details on pK_a prediction can be found in Barlin and Perrin (*62*), Clark and Perrin (*63*), and Barlin and Perrin (*64*), as can pK_a tabulation in Kortum *et al.* (*65*), Albert *et al.* (*66*), and Charton (*67*).

Since the observed solubility of a weak electrolyte depends upon its intrinsic solubility and pK as well as upon the pH of the solution, as given by Eqs. (19) and (20), a nonionic drug can obviously be solubilized by proper design of an ionizable pro-moiety. Since the lipophilicity is also dependent on pK_a, as discussed, it may be desirable in some cases to provide a prodrug with a given pK_a, and hence a given apparent lipophilicity. Thus, the influence of structure on the pK_a of a pro-moiety is important in designing the solubility and lipophilicity of the resulting prodrug.

1. *Acids*

In the case of aromatic carboxylic acids, the Hammett equation [Eq. (25)] gives the relationship between the ionization constant of substituted benzoic acid analogs (log k), the ionization constant of benzoic acid (log k_{HA}), the sigma value (σ) of the substituent in the benzoic acid analog, and the reaction sensitivity factor or the rho (ρ) value.

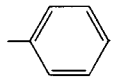
$$\log(k/k_{HA}) = \rho\sigma \tag{25}$$

Since, by definition, $\rho = 1$ for the ionization of benzoic acids and the pK_a of benzoic acid is 4.20

$$pK_a = 4.20 - \sigma \tag{26}$$

The pK_a's of meta- and para-substituted benzoic acids can be calculated from the sum of the appropriate substituents constants given in Table

TABLE XVI
σ-VALUES OF COMMON SUBSTITUENTS FOR SUBSTITUTED BENZOIC ACIDS[a]

Substituent	σ_{meta}	σ_{para}
—H	0	0
—CH_3	−0.07	−0.17
—nC_4H_9	−0.07	−0.16
	0.06	0.01
—CHO	0.36	0.22
—CF_3	0.47	0.54
—COOH	0.37	0.41
—COO^-	0.01	0.31
—COO—CH_3	0.32	0.39
—$CONH_2$	0.28	0.36
—$COCH_3$	0.38	0.50
—CN	0.61	0.66
—NH_2	−0.04	−0.66
—$N(CH_3)_2$	−0.05	−0.83
—NH_3^+	1.13	1.70
—$N(CH_3)_3^+$	0.88	0.82
—OH	0.10	−0.37
—OCH_3	0.14	−0.32
—$OCOCH_3$	0.39	0.31
—SCH_3	0.15	0.00
—SO_2NH_2	0.46	0.57
—SO_3^-	0.05	0.09
—$NHCOCH_3$	0.21	0.00
—NO_2	0.71	0.78

[a] From Barlin and Perrin (*64*).

XVI, providing that steric effects are not introduced. Convenient compilations of substituent constants can be found in Barlin and Perrin (*64*).

The pK_a values of meta-, para-, and ortho-substituted phenols can be predicted from the relationship

$$pK_a = 9.92 - 2.23 \Sigma \sigma_{phenol} \quad (27)$$

Admittedly, most phenols are too weakly acidic to be of use as a solubilizing pro-moiety. Tables of the phenol σ-values can be found in Clark and Perrin (*63*).

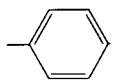
In the case of aliphatic and alicyclic carboxylic acids, the pK_a values can be predicted from the Taft σ^*-value (Table XVII) using

$$\text{RCOOH:} \quad pK_a = 4.66 - 1.62 \sigma^* \quad (28)$$

$$\text{R-CH}_2\text{-COOH:} \quad pK_a = 5.16 - 0.73 \sigma^* \quad (29)$$

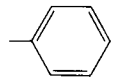
An alternate approach to estimation of pK_a values is given by Barlin and Perrin (*62*) using group contributions (or $-\Delta pK_a$). Thus, for substituted

TABLE XVII
σ^* VALUES FOR COMMON SUBSTITUENTS ON ALIPHATIC SYSTEMS[a]

Substituent	σ^*
—H	0.49
—CH_3	0
—nC_3H_7	−0.115
	0.60
—COOH	2.08
—COO^-	−1.06
—$COOC_2H_5$	2.12
—CH_2COOCH_3	1.06
—Cl	2.96
—CH_2Cl	1.05
—CCl_3	2.65
—$CH_2NHCOCH_3$	0.43
—OC_2H_5	1.68
—ONO_2	3.86
—$COCH_3$	1.65
—CN	3.30
—$CONH_2$	1.68
—NH_2	0.62
—$N(CH_3)^+$	4.55
—OH	1.34
—CF_3	2.61
—SO_3^-	0.81

[a] From Barlin and Perrin (*64*).

TABLE XVIII
ACID STRENGTHENING OR $-\Delta pK_a$ VALUES OF α-SUBSTITUTED ALIPHATIC ACIDS[a]

Substituent	$-\Delta pK_a$
—CH=CHR	0.25
	0.46
—CN	2.29
$-COCH_3$	1.14
$-CONH_2$	1.12
—COOH	1.01
$-COO^-$	− 0.62
—COOR	1.41
$-CF_3$	1.70
$-NH_3^+$	2.41
$-N(CH_3)_3^+$	2.93
$-NHCOCH_3$	1.09
$-NO_2$	3.08
—OH	0.93
—OR	1.11
$-ONO_2$	2.50
—F	2.19
—Cl	1.91
$-SO_3^-$	0.56
$-SO_2R$	2.36

[a] To calculate the pK_a of acetic or higher homologous acids, use 4.80 minus the substituent value ($-\Delta pK_a$) for the appropriate group. If the substituents are on carbon positions further than the α-position, estimates of the pK_a are given by taking one-half of the $-\Delta pK_a$ values given for each carbon beyond the α-carbon. From Barlin and Perrin (*62*).

acetic acids:

$$pK_a = 4.80 - \Delta pK_{a_{substituent}} \tag{30}$$

Table XVIII lists some useful substituents and their ΔpK_a contributions.

2. *Amines*

The pK_a values of primary, secondary, and tertiary aliphatic amines can be estimated (*63*) from the correlative equations:

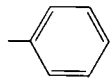
$$\text{Primary Amines:} \qquad pK_a = 13.23 - 3.14 \Sigma \sigma^* \tag{31}$$

$$\text{Secondary Amines:} \quad pK_a = 12.13 - 3.23 \Sigma \sigma^* \quad (32)$$

$$\text{Tertiary Amines:} \quad pK_a = 9.61 - 3.30 \Sigma \sigma^* \quad (33)$$

In an alternate approach to estimation of pK_a values of amines, base-weakening group contributions to the pK_a are employed (*63*). The pK_a values of the typical aliphatic primary, secondary, and tertiary amines are taken as 10.77, 11.15, and 10.5, respectively. The base-weakening contributions ($-\Delta pK_a$) of the substituents (Table XIX) are subtracted from the aforementioned pK_a values. As an example, the pK_a value of *N*,

TABLE XIX
BASE WEAKENING OR $-\Delta pK_a$ VALUES OF SUBSTITUENTS ON ALIPHATIC AMINES[a]

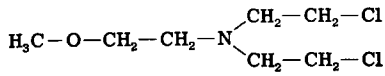
Substituent	$-\Delta pK_a$	
	α-Carbon	β-Carbon
—CH=CHR	~ 1.1	~ 0.5
	1.4	0.8
—CN	5.8	3.0
—COO^-	0.8[b] / −0.1[c]	− 0.2
—COR		1.6
—COOR	3.0	1.3
—$CONH_2$	2.8	1.1
—Cl		1.9
—F		~ 1.6
—NH_2		0.8
—NHR,—NR_2	~ 1.7	0.9
—NH_3^+,—NR_3^+		3.6
—$NHCOCH_3$		1.5
—OH		1.1
—OCH_3,—OR		1.2
—OCOR		~ 1.7

[a] To estimate pK_a values, subtract the $-\Delta pK_a$ value from 10.77 if a primary amine, 11.15 if a secondary amine, or 10.5 if a tertiary amine. Definitions: the α-carbon is the one attached to the amine; the β-carbon is two carbons removed from the amine. If the substituent is further removed, estimates of pK_a values can be obtained by halving the $-\Delta pK_a$ value for each additional carbon removed. From Clark and Perrin (*63*).

[b] Primary or secondary amine.

[c] Tertiary amine.

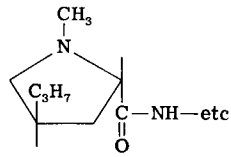
N-bis-β-chloroethyl-β-methoxyethylamine:



can be obtained as follows:

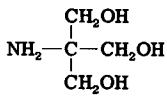
$$pK_a = 10.5 - 2\Delta pK_{a_{-Et-Cl}} - \Delta pK_{a_{-Et-OR}}$$
$$pK_a = 10.5 - 2 \times 1.9 - 1.2$$
$$pK_a = 5.5$$

The estimated pK_a value of 5.5 is in good accord with the observed value of 5.3. Likewise, the pK_a of lincomycin can be predicted from the partial structure:



Lincomycin
partial structure

Here, the carboxamide is alpha to a tertiary amine, and from the ΔpK_a of this group (− 2.8), a pK_a of 7.7 is predicted, which is in good accord with the experimental value of 7.66. In a further example, the pK_a of tromethamine is

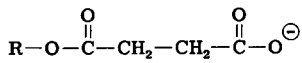


Tromethamine

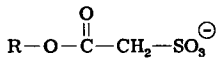
estimated as 7.5, which is in reasonable accord with the observed pK_a of 8.06.

3. *Ionizable Pro-Moieties*

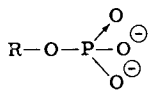
The most commonly used ionizable moieties for solubilizing alcoholic drugs (ROH) are as follows (*68*, *69*):



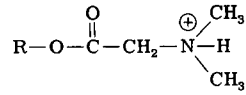
Hemi-succinates
($pK_a \sim 5$)



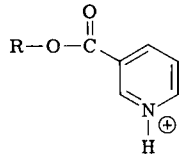
Sulfoacetates
($pK_a \sim 1$)



Phosphates
($pK_a \sim 2$ and ~ 6)

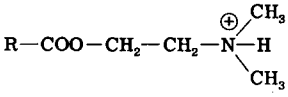


Dimethylamino
acetates ($pK_a \sim 8$)

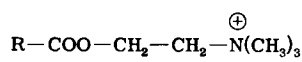


Nicotinyl esters
($pK_a \sim 4$)

For carboxylic acid drugs (R-COOH), the commonly used ionizable pro-moieties are



β-Dimethylaminoethyl
ester ($pk_a \sim 8.8$)



Choline ester

The choice of the pro-moiety will be dictated by a variety of factors including the solubility desired, the pH desired, the dosage form intended (i.e., solid or aqueous solution), and the desirability of absorption of the intact prodrug. High aqueous solubility will be achieved by those prodrugs with the largest difference between the pH (if chosen) and the pK_a of the pro-moiety. In the case of the aforementioned acidic pro-moieties, one would expect the sulfoacetate ester and the phosphate ester to have high aqueous solubility in general. The hemi-succinate ester would have high solubility only at relatively high pH values. All of the amino pro-moieties would have potential for high solubility at pH 1, whereas solubility at pH 7.4 would be expected to be poor.

The only ester with potential for good, long term aqueous stability in solution is the phosphate ester, providing the pH is above about 6.5 (*70*).

VI. Examples of Improved Drug Efficiency

In most instances, drugs showing poor oral availability are large, polar molecules having several polar functional groups. These compounds are too polar to be transported through lipid membrances, too large to be transported through pores, and because of their polyfunctionality, cannot

be transported as ion pairs. Some examples of poorly absorbed drugs include erythromycin, penicillin, cytosine arabinoside, and 6-azuridine.

The oral availability of these compounds can usually be improved greatly by the introduction of hydrophobic moieties that mask one or more of the parent drug's functional groups. The analogs and derivatives shown in Table XX are just some of a large number of successful examples of this approach. Because most of the parent drugs are very soluble in

TABLE XX
EXAMPLES OF PRODRUGS WITH IMPROVED ABSORPTION ATTRIBUTED TO INCREASED LIPOPHILICITY

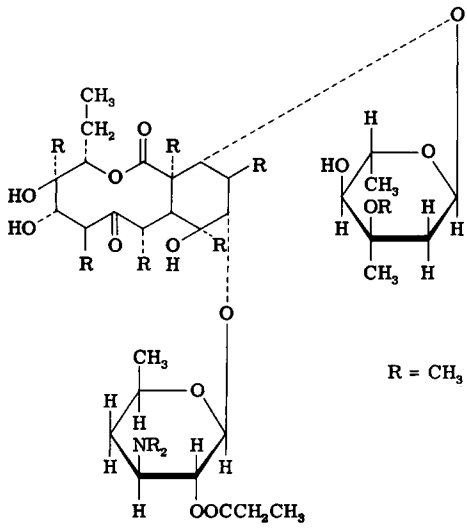
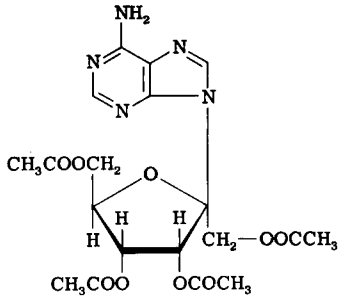
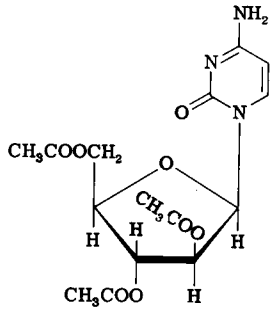
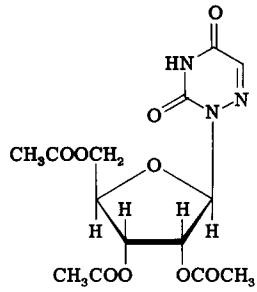
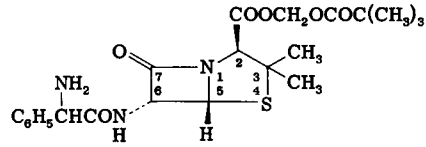
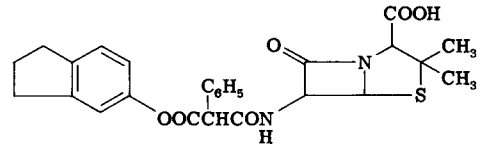
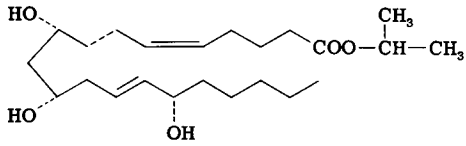
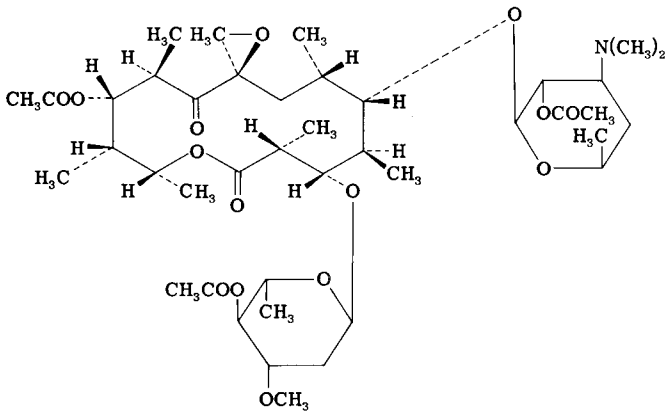
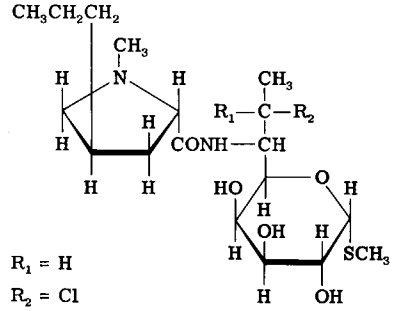
Compound	Reference
 Erythromycin propionate	*71*
 Psicofuranine tetraacetate	*77*

TABLE XX—*Continued*

Compound	Reference
 Cytosine arabinoside triacetate	*78*
 6-Azuridine triacetate	*79*
 Pivampicillin	*73*
 Carindacillin	*74*

(*Continued*)

TABLE XX—*Continued*

Compound	Reference
 PGE$_{2\alpha}$ isopropyl ester	75
 Triacetyloeandomycin	26
 Clindamycin	76

water, there is little danger of reducing solubility below acceptable limits by moderately hydrophobic substitution. Acetates are particularly useful in this regard because they can be multiply substituted without raising the partition coefficient to such an extent that solubility is reduced enough to decrease drug efficiency.

Table XX shows some examples of prodrugs and analogs with improved oral absorption attributed to an increase in lipophilicity over the parent compound. Erythromycin propionate is considerably more lipophilic than the parent compound, erythromycin, due to the reduction in the pK_a and the lipophilic nature of the ester. Calculations show that at pH 6, the log PC_{obs} is -0.13 for erythromycin and $+2.5$ for erythromycin propionate (based on erythromycin $pK_a = 8.8$ and $\log PC_{int} = 2.48$, and erythromycin propionate $pK_a = 6.9$). In humans, erythromycin propionate is more efficiently absorbed than erythromycin (*71*). Higher blood levels, comprising the sum of intact ester and free erythromycin, are attained after dosing with erythromycin propionate, as compared to the blood levels attained after dosing with erythromycin base. About 60% of the compound exists as the intact propionate ester in the blood stream and about 80% of the total ester erythromycin species excreted in the urine occurs as the intact ester (*72*).

Pivampicillin, the pivaloyloxymethyl ester of ampicillin, shows approximately 90% absorption upon oral administration in humans, whereas the parent compound ampicillin shows only about 33% absorption (*73*). The improved absorption of the ester is due to the lipophilicity enhancing effects of (a) the pivaloxyloxymethyl group and (b) conversion of an ionizable carboxyl group to a nonionic species. Using benzyl penicillin as a point of reference, ampicillin is less lipophilic ($\Delta\pi \cong -1.2$) whereas pivampicillin is considerably more lipophilic ($\Delta\pi \cong +1.5$). Similar calculations show that the ester, carindacillin, is more lipophilic than benzylpenicillin ($\Delta\pi \cong +2.8$) whereas the parent dicarboxylic acid, carbenecillin, is less lipophilic than benzylpenicillin ($\Delta\pi = -1.2$). Benzylpenicillin is moderately well absorbed (~60%) in humans; and in accord with this, the less lipophilic compound carbenecillin is poorly absorbed whereas the ester carindacillin shows near quantitative absorption (*74*).

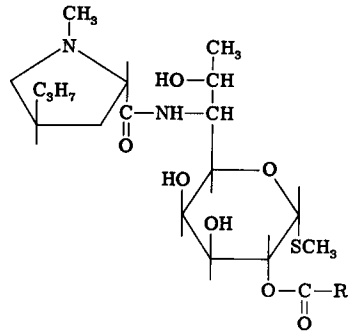
The prostaglandin ester, $PGF_{2\alpha}$ isopropyl ester, shows improved absorption in animals, and this has been attributed to the increased lipophilicity of the ester relative to the parent compound (*75*).

Clindamycin represents the 7-chloro analog of the parent compound lincomycin wherein R^1 = OH and R^2 = H in the latter structure. This substitution of a chloro group for a hydroxyl group produces a substantial increase in lipophilicity. The $\Delta\pi$ between clindamycin and lincomycin ($+1.55$) is equivalent to the lipophilicity of an *n*-propyl group. Studies in

humans show that the more liophilic analog, clindamycin, is absorbed to the extent of 65% whereas lincomycin gives 55% absorption (*76*).

Psicofuranine is not absorbed upon oral administration in humans (*77*). The *n*-octanol/water log*PC* of this compound is − 1.95, and this low value is consistent with poor absorption. The triacetate ester, however, is well absorbed in humans and this is in accord with the estimated log *PC* = + 0.72, which is above the previously mentioned cut-off lipophilicity for absorption, namely, log *PC* ≅ 0. Cytosine arabinoside (with *n*-octanol/water log*PC* = − 2.10) is not absorbed orally whereas the corresponding triacetate ester (log *PC* ≅ + 0.57) is well absorbed (*78*). 6-Azauridine (with *n*-octanol/water log*PC* = − 2.14) again is not absorbed orally in humans whereas the triacetate ester (log PC ≅ + 0.53) is well absorbed (*79*).

The intestinal absorption of the homologous aliphatic esters of lincomycin illustrates the importance of lumenal hydrolysis and aqueous solubility. Table XXI shows the influence of ester chain length on the absorption of lincomycin 2-acyl esters from ligated jejunal sections of the intact rat (*80*).



Lincomycin-2-acyl esters

As lipophilicity or chain length increases, the percentage absorbed decreases. This seemingly contradictory behavior is readily explained on the basis of the competitive lumenal hydrolysis scheme (shown in Fig. 10) and of the protection from hydrolysis resulting from decreased solubility of the higher esters.

The esters were initially placed in the jejunal loops as the soluble hydrochloride salts. Since the solubility of lincomycin 2-esters rapidly decreases at pH 6, the probable pH of the intestine, precipitation of the free base would occur with the longer chain esters. Competitive lumenal hydrolysis of the esters was shown to occur, providing free lincomycin. The apparent decrease in absorption with ester chain length is thus determined

TABLE XXI
ABSORPTION OF LINCOMYCIN ESTERS FROM RAT JEJUNAL LOOPS[a]

Compound	Partition coefficient[b]	Percent absorbed after 2 hours
Lincomycin	0.46	33
Lincomycin 2-propionate	14	64
Lincomycin 2-butyrate	40	54
Lincomycin 2-hexanoate	390	40
Lincomycin 2-octanoate	2750	27
Lincomycin 2-laurate	7800	17

[a] From Fletcher *et al.* (*80*).
[b] Ether/water intrinsic partition coefficient.

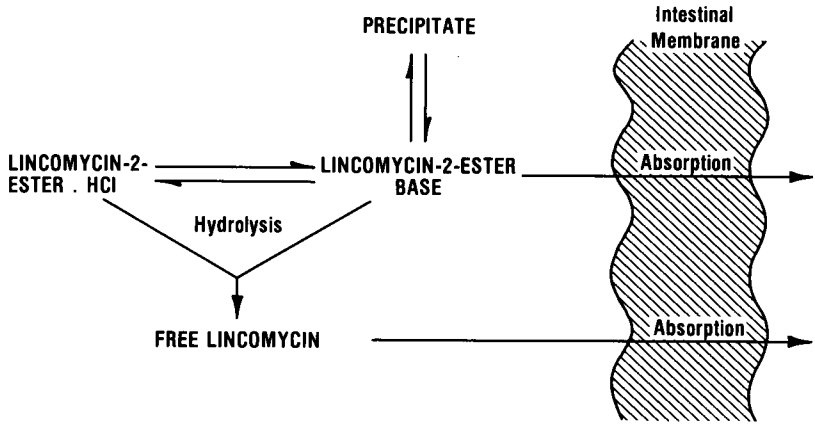


Fig. 10. Model for Intestinal Absorption of Lincomycin Esters.

by the enzyme kinetics, and by the solubility of the esters, resulting in most rapid absorption for the 2-propionate ester. The 7-propionate ester of lincomycin was shown to be absorbed to the extent of 91% in 2 hours as compared to 64% for the 2-propionate ester (Table XXI). Increased absorption of the 7-propionate ester resulted from failure of this ester to undergo intestinal hydrolysis, whereas the 2-propionate ester showed appreciable intestinal hydrolysis.

Blood level studies using intestinally ligated rats, however, showed lower antibacterial blood levels for the 7-propionate as compared to the 2-propionate esters (Fig. 11). Blood samples following dosing of the latter revealed no intact ester, and the resulting area under the blood level curve was significantly higher than that of lincomycin, indicating improved

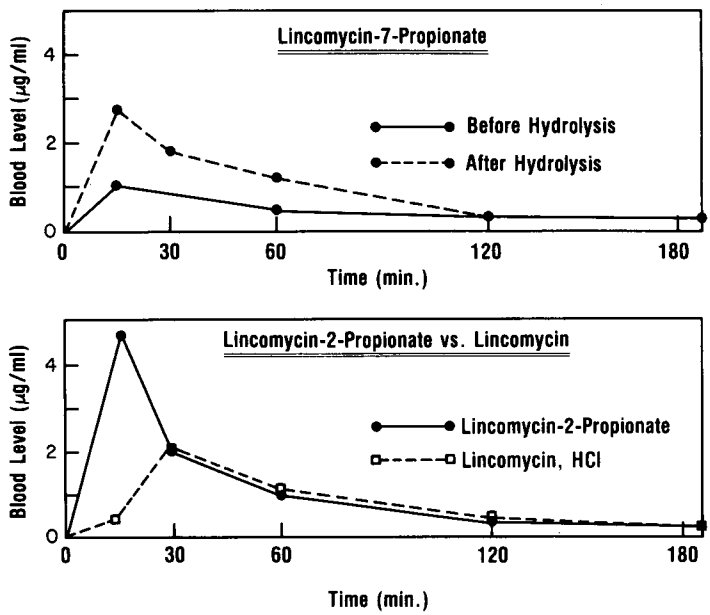


Fig. 11. Antibacterial blood levels following jejunal loop administration of lincomycin 2-propionate (upper) and 7-propionate (lower) in comparison with lincomycin (lower). The lincomycin-2-propionate curve represents the lincomycin level since no intact ester is present (*80*). Hydrolysis was conducted with 0.1 *N* NaOH.

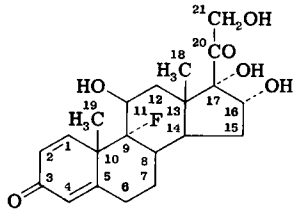
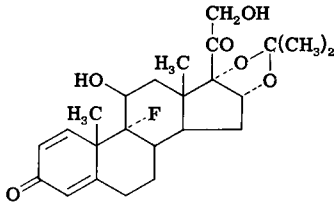
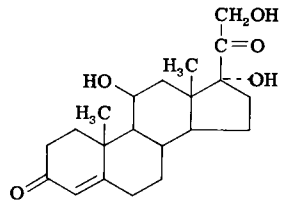
absorption. The blood level studies following administration of the 7-propionate ester showed low levels of free lincomycin. The 7-propionate is virtually inactive in the microbiological assay employed, and as a result, base catalyzed hydrolysis reveals the levels of the sum of the intact ester and free lincomycin. This lower bioavailability of the 7-propionate compared to the 2-propionate, as ascertained by blood level areas, is not consistent with the previously mentioned lumenal absorption studies which showed 91% and 64% absorption with the 7-propionate and 2-propionate esters of lincomycin, respectively. To account for this, it has been proposed that the intact 7-propionate ester undergoes hydrolysis competitively with urinary excretion of the intact ester or metabolism to a biologically inactive species (*12*). The resulting pharmacokinetic model for a prodrug upon oral administration is shown in Fig. 12 where D is the parent drug and P-D is the prodrug.

Obviously, the ideal prodrug for improving oral absorption will have k_{P-D} for intestinal absorption much greater than k_{P-D} for intestinal hydrolysis; and, after absorption, the k_{P-D} for hydrolysis should be much

greater than the $k_{elimination}$ step for both the intact prodrug and the parent drug.

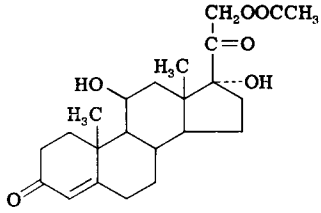
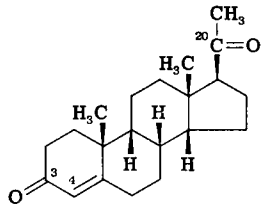
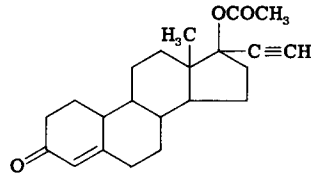
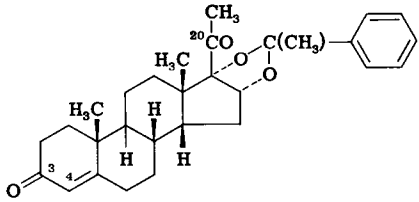
Table XXII shows the influence of lipophilicity on the intestinal absorption of steroids in man using an intubation technique (*81*). The percentage absorbed is given for a 15-cm section of the intestine in a period of 15 minutes. This data shows that absorption is poor when log *PC* approaches zero. However, when log *PC* is above 2, absorption is > 50%, and this is in accord with the previously discussed lipophilicity requirement for absorption.

TABLE XXII
HUMAN ABSORPTION OF STEROIDS USING INTUBATION TECHNIQUES[a]

Steroid	Absorbed (%)	log *PC* (calc)
 Triamcinolone	4	0.15
 Triamcinolone acetonide	24	1.84
 Hydrocortisone	44	1.78

(Continued)

TABLE XXII—*Continued*

Steroid	Absorbed (%)	log *PC* (calc)
 Hydrocortisone 21-acetate	61	2.67
 Progesterone	91	4.60
 Norethindrone acetate	95	4.13
 Progesterone 16,17-acetophenone	99	6.73

[a] From Schedl, (*81*).

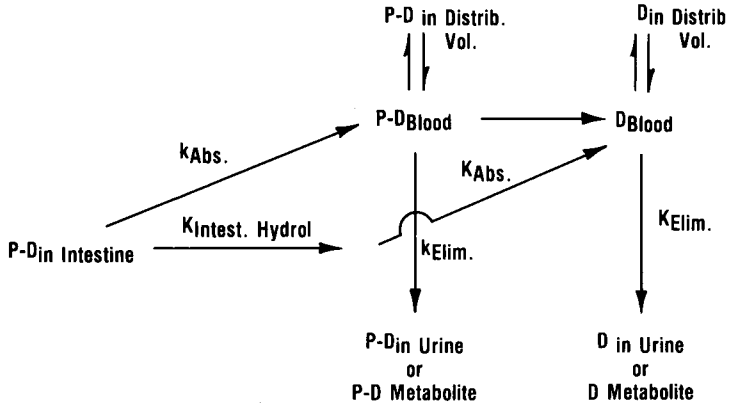
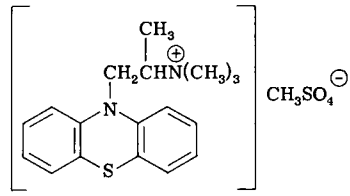


Fig. 12. Prodrug pharmacokinetic model (*12*).

Ionized drugs, such as quaternary ammonium compounds, are not absorbed efficiently upon oral administration due to their low lipid solubility (*82*). Thiazinamium methylsulfate, for example, is absorbed to the extent of only 2–3% upon oral administration in humans (*83*).



Thiazininamium methylsulfate

Ion-pair formation has been used to improve oral absorption of ionic drugs (*84, 85*); although, due to the problem of dilution *in vivo,* this approach does not appear to be as promising as covalent bond formation of prodrugs.

VII. Solubility-Limited Absorption

Poor oral absorption of drugs can occur when the lipophilicity of the drug is high ($\log PC \gg 2$) or if the melting point of the compound is high. It was proposed that drugs with solubilities lower than about 10 μg/ml may show problems with absorption due to dissolution rate control ($K_u \gg$ disso-

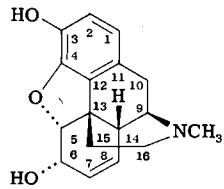
lution rate) (*16*). Studies on the intestinal absorption of hydrocarbons in rats (*86*) show a linear relationship between % absorbed and chain length between C-14 and C-28. At C-14, about 63% absorption occurs whereas at C-28 it is only about 5%. Decreased absorption with increased chain length is attributed to the progressively decreased solubility of the alkanes.

Many highly lipophilic and water-insoluble compounds are absorbed more efficiently when administered as emulsions or as oily solutions. This is exemplified with *n*-hexadecane (*87*), benzopyrene (*88*), α-glycerol mono-oleyl ether, and oleic acid (*89*).

VIII. Metabolism-Limited Bioavailability

Many drugs are metabolized to a significant extent in transit through the intestinal wall (*90, 91*). The blood supply from the intestine passes through the liver, from which it enters the general circulation. The combined "first pass" metabolism of the intestine and liver is so extensive with some drugs that negligible blood levels of the intact species is observed on oral dosing. In this case the problem is one of bioavailability, since the compound may be quantitatively absorbed. It is, therefore, necessary to establish the mechanism responsible for poor efficiency, i.e., first pass metabolism or low lipophilicity, before a synthetic approach to improve bioavailability can be suggested. Initially, this can be done in animals by cannulating the mesenteric blood supply to detect intestinal wall metabolism (*92*).

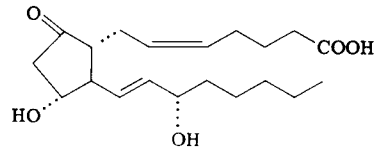
The synthetic approach to improve oral bioavailability of those compounds that undergo extensive first pass intestine and liver metabolism consists of either prodrug formation or analog formation. In both cases, the site of metabolism is masked. Morphine is an example of a drug that undergoes extensive first pass metabolism, with only about 18% of the oral dose reaching the systemic circulation in rats (*93*).



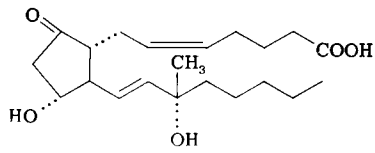
Morphine

Esters at both hydroxyls have been synthesized to improve intestinal bioavailability as summarized by Stella (*94*).

In the case of the analog approach to improve bioavailability, the site of metabolism must be identified. Once this is achieved, the functional group identified is either modified or steric effects are introduced proximally to the group. PGE_2, for example, is a compound that is rapidly metabolized by oxidation at the C-15 hydroxyl group upon oral administration



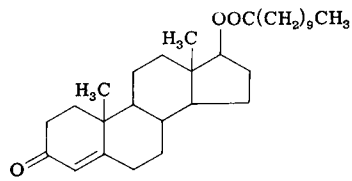
PGE_2



15(*R*)-15-Methyl-PGE_2

The 15-methyl-PGE_2 analog is incapable of metabolism at C-15 and the activity of the parent compound is retained (*95*). The analog 16,16-dimethyl-PGE_2, even though it contains a secondary C-15 alcohol, does not undergo *in vivo* metabolism, presumably due to the steric influence of the adjacent dimethyl substituents (*96*).

An additional method for overcoming intestinal metabolism is by means of shifting the mechanism of absorption from the normal venous route to the lymphatic route. Compounds absorbed via the lymphatics by-pass the liver in their transit to the central venous circulation. Testosterone is an example of a drug that is extensively metabolized upon oral administration; however, the corresponding 17-undecanoate ester is orally effective. The latter ester is highly lipophilic and absorption occurs via the lymphatic system resulting in metabolic stabilization (*97*).



Testosterone
undecanoate

An extremely powerful tool for uncovering gastointestinal metabolism is the use of oral intubation techniques (*98*). Elegant application of this

technique was made by Swahn in studying the absorption and metabolism of two ampicillin esters (*99*). Metabolism in the GI tract is easily ascertained, as well as the site of absorption of the compounds. It is anticipated that oral intubation techniques will be widely used in future studies dealing with metabolism, kinetics of absorption, and physical models.

Some drugs give the appearance of poor oral bioavailability when, in fact, they are efficiently absorbed but undergo secretion into the intestinal tract from the systemic circulation. This can occur through (a) ion-trapping in the stomach with basic drugs such as pethidine (*100*), (b) active transport secretion into the intestine as in the case of erythromycin (*101*), and (c) biliary excretion as in the case of metiapine or 2-methyl-11-(4-methyl-1-piperazinyl)-dibenzo-(1,4)-thiazepine (*102*). It is important to recognize GI and biliary secretion before embarking on a prodrug approach to improve absorption since secretion is difficult to block through prodrug formation.

IX. Design of Prodrug Lability

In addition to attaining the desired lipophilicity and solubility, the prodrug must possess an optimal lability. The prodrug must be stable in the intestinal lumen but should, upon systemic absorption, be rapidly hydrolyzed. The analogy between base-catalyzed and esterase-catalyzed ester hydrolysis has been reviewed (*12*). In the former, nucleophilic attack of the ester occurs by OH^-, in the latter, nucleophilic attack occurs via the nucleophilic group at the active site. Because of this mechanistic similarity, it has been proposed that substituent effects (electronic and steric) be used to tailor the *in vivo* lability of prodrugs. Taft (*103*) has shown that the hydroxide ion-catalyzed hydrolysis of aliphatic acyl esters is given by:

$$\log(k/k_0) = \rho\sigma^* + \delta E_s \tag{34}$$

where: k and k_0 are the second-order rate constants for the substituted and reference acyl ester, ρ the reaction constant, σ^* the substituent constant for the acyl component, δ the proportionality constant and E_s the steric substituent constant for the acyl group. In the case of substituted aromatic esters, the Hammett condition prevails wherein:

$$\log(k/k_0) = \rho\sigma \tag{35}$$

The influence of electronic effects on bioactivity is demonstrated with the 2-acyl esters of lincomycin and the 2-benzoate esters of lincomycin (*12*). In both cases a linear dependency between log antibacterial activity and electronic substituent constant is shown in Figs. 13 and 14. As ex-

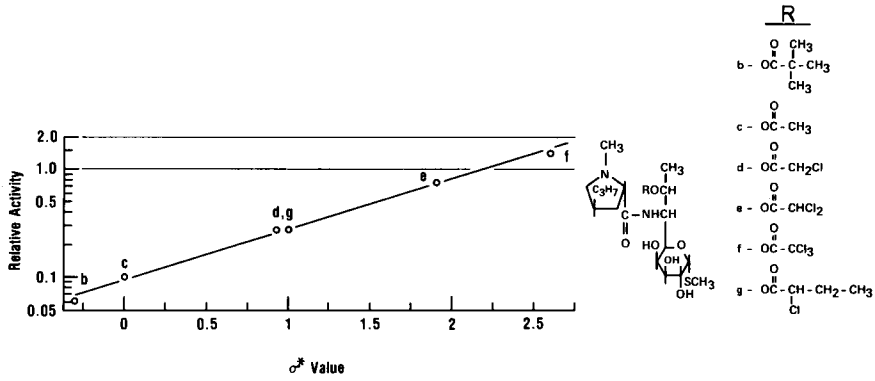


Fig. 13. Relative molar CD_{50} of 7-acyl esters of lincomycin upon subcutaneous injection in the mouse challenged with *s. aureus* (lincomycin = 1) (*12*).

pected, the highly labile or activated esters with large positive σ^*- or σ-values display the highest activity. One should not expect a linear relationship between log activity and substituent constant in all prodrug classes for two reasons. First, the enzyme–substrate interaction within the selected series of prodrugs must be such that the electronic properties are the main variable. Second, the pharmacokinetics of the prodrug series should be similar. For example, an acetate ester of a drug might be biologically inactive *in vivo* if extensive elimination of the intact ester occurs in the urine whereas a more lipophilic ester, if not lost by urinary excretion,

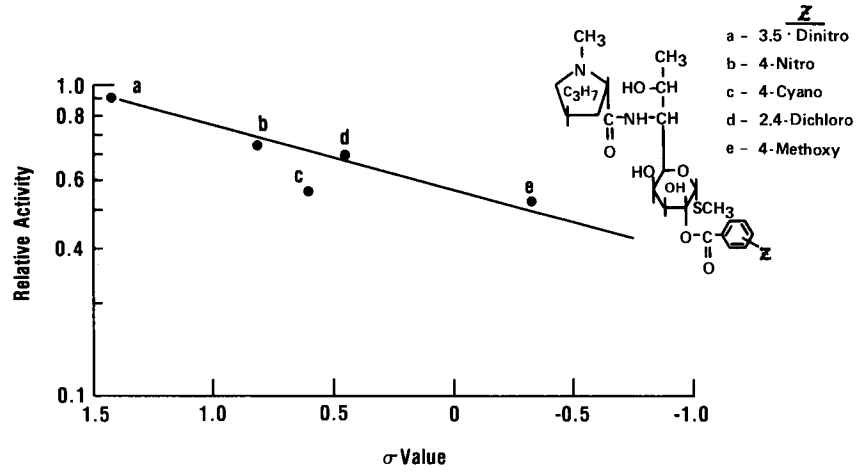


Fig. 14. Relative molar CD_{50} of 2-(benzoate) esters of lincomycin upon oral administration in the mouse challenged with *s. aureus* (lincomycin = 1) (*12*).

could show good biological activity even though the hydrolysis rate constants of the two esters are identical. Further considerations are detailed in Morozowich *et al.* (*12*) along with other examples showing that biologically active prodrugs can be attained by the principle of substituent group activation of the pro-moiety. The problem in designing prodrugs with improved absorption is not one of merely assuring biological activity *in vivo* but of carefully tailoring the lability of the pro-moiety to remain stable in the lumen but hydrolyze rapidly once absorbed systemically. Animal studies are useful in guiding the prodrug design considerations; however, the prodrug selected from animal studies may not have the optimal lability in humans. This is the result of the greater rates of enzymatic reactions in lower animals as compared with man. Monkeys are frequently employed in prodrug evaluations since many of the esterolytic enzymes are kinetically similar to those in humans (*99*).

X. Summary and Overview

The rational design of prodrugs having oral biological activity can be divided into three basic steps: The determination of the physical properties required for maximum efficacy, the selection of a chemical linkage between the parent drug and the pro-moiety that will be cleaved in the desired biological compartment, and the design and synthesis of prodrugs that have the proper physical-chemical properties as well as the proper chemical lability.

The selection of optimum physical-chemical parameters is accomplished through the aid of *in vitro* and *in vivo* experimentation and modeling. If the biological system is clearly understood, it is usually possible to devise models that can give reliable estimates of what physical-chemical parameters are important and of their optimal values. These conclusions can often be verified by studies on the uptake of model compounds into the tissues or circulation of lower animals.

The design of prodrugs that have a specific set of physical-chemical properties is a very rapidly advancing field. We have shown some examples of the success that can be attained by a variety of empirical techniques and that this approach represents another step in the ultimate goal of rational drug design.

References

1. A. A. Sinkula and S. H. Yalkowsky, *J. Pharm. Sci.* **64,** 181 (1975).
2. A. A. Sinkula, *Annu. Rep. Med. Chem.* **10,** 306 (1975).

3. E. B. Roche, ed. "Design of Biopharmaceutical Properties through Prodrugs and Analogs." APhA, Washington, 1977.
4. R. E. Notari, *Pharm. Weekbl.* **110,** 577 (1975).
5. T. Higuchi and V. Stella, eds., "Pro-Drugs as Novel Drug Delivery Systems" ACS Symposium Series Vol. 14, ACS, Washington, 1975.
6. G. A. Digenis and J. V. Swintosky, *In* "Handbook of Experimental Pharmacology, Vol. 28" (O. Eichler *et al.,* eds.). Springer-Verlag, Berlin and New York, 1975.
7. J. N. Harper, *Progr. Drug Res.* **4,** 221 (1962).
8. E. J. Ariëns, ed., "Drug Design." Vol. 4. Academic Press, New York, 1971.
9. R. E. Notari, *J. Pharm. Sci.* **62,** 865 (1973).
10. V. J. Stella, *Aust. J. Pharm. Sci.* **NS2,** 57 (1973).
11. B. K. Martin, *Br. J. Pharmacol. Chemother.* **31,** 420 (1967).
12. W. Morozowich, M. J. Cho and F. Kezdy, *In* "Design of Biopharmaceutical Properties through Prodrugs and Analogs" (E. B. Roche, ed.). APhA, Washington, 1977.
13. W. E. Hamlin, J. I. Northam, and J. G. Wagner, *J. Pharm. Sci.* **54,** 1651 (1965).
14. B. J. Zwolinski, H. Eyring, and C. E. Reese, *J. Phys. Chem.* **53,** 1426 (1949).
15. R. G. Stehle and W. I. Higuchi, *J. Pharm. Sci.* **56,** 1367 (1967).
16. N. F. H. Ho, J. Y. Park, W. Morozowich, and W. I. Higuchi, *In* "Design of Biopharmaceutical Properties through Prodrugs and Analogs" (E. B. Roche, ed.). APhA, Washington, 1977.
17. R. G. Stehle and W. I. Higuchi, *J. Pharm. Sci.* **61,** 1922 (1972).
18. G. L. Flynn and S. H. Yalkowsky, *J. Pharm. Sci.* **61,** 838 (1972).
19. S. H. Yalkowsky and G. L. Flynn, *J. Pharm. Sci.* **62,** 218 (1973).
20. S. H. Yalkowsky and G. L. Flynn, *J. Pharm. Sci.* **63,** 1276 (1974).
21. C. Hansch, *Accounts Chem. Res.* **2,** 232 (1969).
22. J. T. Penniston, L. Beckett, D. L. Bentley, and C. Hansch, *Mol. Pharmacol.* **5,** 333 (1969).
23. C. Hansch and J. M. Clayton, *J. Pharm. Sci.* **62,** 1 (1973).
24. C. Hansch and W. J. Dunn III, *J. Pharm. Sci.* **61,** 1 (1972).
25. J. Ferguson, *Proc. Roy. Soc. London, Ser. B,* **127,** 387 (1939).
26. G. M. Badger, *Nature (London)* 585 (1946).
27. N. A. Allawala and S. Reigelman, *J. Am. Pharm. Assn., Sci. Ed.* **43,** 93 (1954).
28. F. Brink and S. M. Posternack, *J. Cellular Comp. Physiol.* **32,** 211 (1948).
29. J. E. Webb, Symposia, *Soc. Exptl. Biol.* **3,** 143 (1949).
30. S. H. Yalkowsky, T. G. Slunick, and G. L. Flynn, *J. Pharm. Sci.* **63,** 691 (1974).
31. E. J. Lien, *In* "Drug Design" Vol. 5, p. 81. (E. J. Ariens, ed.), Academic Press, New York, 1975.
32. P. J. Weiss, M. L. Andrew, and W. W. Wright, *Antibiot. Chemother. (Washington, D.C.)* **7,** 376 (1957).
33. M. L. Andrew and P. J. Weiss, *Antibiot. Chemother. (Washington, D.C.)* **9,** 277 (1959).
34. W. F. M. Fulton, "Modern Trends in Pharmacology and Therapeutics" Vol. 1, p. 183. Appleton, New York, 1967.
35. Physicians Desk Reference, Medical Economics, p. 1474. Oradell, N.J., 1976.
36. A. J. Leo, C. Hansch, and D. Elkins, *Chem. Rev.* **71,** 525 (1971).
37. J. L. Cohen, L. E. Irwin, G. J. Marshall, H. Darvey, and J. R. Bateman, *Cancer Chemother.* **58,** 723 (1974).
38. Pomona College Medicinal Chemistry Project. Edition, July, 1977. (This is a compilation of reported partition coefficients prepared by C. Hansch, A. Leo, and others.)
39. A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.* **71,** 525 (1971).
40. T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.* **86,** 5175 (1964).

41. C. Hansch, A. Leo, and D. Nikaitani, *J. Org. Chem.* **37**, 3090 (1972).
42. G. G. Nys and R. F. Rekker, *Eur. J. Med. Chem.* **9**, 361 (1974).
43. A. Leo, P. Y. C. Jow, C. Silipo, and C. Hansch, *J. Med. Chem.* **18**, 865 (1975).
44. S. H. Yalkowsky, *J. Med. Chem.* **19**, 727 (1976).
45. J. C. McGowen, *J. Appl. Chem.* **2**, 323 (1952); **4**, 41 (1954).
46. L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research." Academic Press, New York, 1976.
47. S. H. Yalkowsky, *In* "Design of Biopharmaceutical Properties through Prodrugs and Analogs" (E. B. Roche, ed.), p. 392. Academy of Pharmaceutical Sciences, Washington, 1977.
48. S. S. Davis, *Adv. Pharm. Sci.* **4**, 74 (1974).
49. M. S. Tute, *Adv. Drug Res.* **6**, 1 (1971).
50. A. Canas-Rodriguez and M. S. Tute, *In* "Biological Correlations—The Hansch Approach" (W. VanVolkenburg, ed.) ACS, Washington, 1972.
51. A. R. Ubbelohde, "Melting and Crystal Structure." Clearendon, Oxford, 1965.
52. A. Bondi, "Physical Properties of Molecular Crystals, Liquids, and Glasses." Wiley, New York, 1968.
53. R. C. Weast, "Handbook of Chemistry and Physics" 54th Ed. CRC Press, Cleveland, 1973.
54. W. Morozowich, T. O. Oesterling, W. L. Miller, C. F. Lawson, J. R. Weeks, R. G. Stehle and, S. L. Douglas, *J. Pharm. Sci.* **68**, 4313 (1979).
55. T. J. Roseman, S. H. Yalkowsky, and S. C. Valvani, manuscript in preparation.
56. S. C. Valvani, S. H. Yalkowsky, and T. J. Roseman, manuscript in preparation.
57. C. Hansch, J. E. Quinlan, and G. L. Lawrence, *J. Org. Chem.* **33**, 347 (1968).
58. J. H. Hildebrand and R. L. Scott, "Regular Solutions." Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
59. A. N. Martin, "Physical Pharmacy." Lea & Febiger, Philadelphia, 1960.
60. S. F. Kramer and G. L. Flynn, *J. Pharm. Sci.* **61**, 1896 (1972).
61. G. LePetite, *Pharmazie* **32**, 289 (1977).
62. G. B. Barlin and D. D. Perrin, *Quart. Revs Chem. Soc.* **20**, 75 (1966).
63. J. Clark and D. D. Perrin, *Quart. Revs Chem. Soc.* **18**, 295 (1964).
64. G. B. Barlin and D. D. Perrin, "Techniques of Chemistry, Vol. 4" p. 611 (A. Weissberger, ed.) Wiley, New York, (1973).
65. G. Kortum, W. Vogel, and K. Andrussow, "Dissociation Constants of Organic Acids in Aqueous Solution." Butterworth, London, 1961.
66. A. Albert and E. P. Serjeant, "Ionization Constants of Acids and Bases." Wiley, New York, 1962.
67. M. Charton, *J. Org. Chem.* **29**, 1222 (1964).
68. A. A. Sinkula and S. H. Yalkowsky, *J. Pharm. Sci.* **64**, 3259 (1975).
69. V. Stella, *In* "Prodrugs as Novel Delivery Systems" (T. Higuchi and V. Stella, eds.), p. 1. ACS Symposium Series, American Chemical Society, Washington, D.C., 1975.
70. C. A. Bunton, D. R. Llewellyn, and K. G. Oldham, *J. Chem. Soc.* 3574 (1958).
71. V. C. Stephens and J. W. Conine, *Antibiot. Annu.* 346 (1958–1959).
72. H. A. Blough, W. H. Hall and L. Long, *J. Am. Med. Soc.*, 539 (1960).
73. W. Von Daehne, E. Fredericksen, E. Gunderson, F. Lund, P. Morch, H. J. Peterson, K. Roholt, L. Tybring, and W. O. Godtfredsen, *J. Med. Chem.* **13**, 607 (1970).
74. P. Nauman and H. Rosin, *Med. Wochenschr.* **98**, 2200 (1973).
75. W. Morozowich, T. O. Oesterling, W. L. Miller, C. F. Lawson, J. R. Weeks, J. C. Cornett, R. G. Stehle and S. Douglas, p. 150. Abst. APhA Academy of Pharmaceutical Sciences, Atlanta, Ga., 1975.

76. C. Simon, V. Malerczyk, and E. Eidelloth, *Med. Welt.* **27,** 1096 (1976).
77. H. Hoeksema, G. B. Whitfield, and L. E. Rhuland, *Biochim. Biophys. Res. Commun.* **6,** 213 (1961).
78. C. K. Mellet, *Pharmacologist* **12,** 221 (1970).
79. R. E. Handschumaker, W. A. Creasey, M. E. Fink, P. Calabresi, and A. D. Welch, *Cancer Chemother. Rep.* 267 (1962).
80. H. P. Fletcher, H. M. Murray and T. E. Weddon, *J. Pharm. Sci.* **59,** 2100 (1968).
81. H. Schedl, *J. Clin. Endocrinol.* **25,** 1309 (1965).
82. R. A. Levine, *In* "Topics in Medicinal Chemistry, Vol. 4" (J. L. Rabinowitz and R. M. Meyerson, eds.), p. 27. Wiley, New York, 1971.
83. E. Tomlinson and S. S. Davis, *J. Pharm. Pharmacol.* **285,** 75P (1976).
84. J. Nakamura, M. Muranushi, T. Kimura, S. Muranishi, and H. Sezaki, *Chem. Pharm. Bull.* **22,** 851 (1977).
85. J. H. G. Jonkman, J. Wijseek, S. H. Brouwer, and R. A. DeZeeuw, *J. Pharm. Pharmacol.* **605,** 63P (1974).
86. P. W. Albro and L. Fishbein, *Biochim. Biophys. Acta.* **219,** 437 (1970).
87. A. C. Frazer, J. H. Schulman, and H. C. Stewart, *J. Physiol.* **103,** 306 (1944).
88. E. D. Rees, P. Mandelstam, J. Q. Lowry, and H. Lipscomb, *Biochim. Biophys. Acta* **225,** 96 (1971).
89. N. E. Hoffman and V. J. Yeoh, *Biochim. Biophys. Acta* **233,** 49 (1971).
90. W. Erni, *Schweiz. Apoth. Ztg.* **115,** 201 (1977).
91. W. H. Barr and S. Riegelman, *J. Pharm. Sci.* **59,** 164 (1970).
92. T. Suzuki, O. Toshiaki, T. R. Kihisa, and S. Isozaki, *J. Pharm. Sci.* **64,** 896 (1975).
93. C. D. Klassen and K. Iwamoto, *Pharmacologist* **18,** 178 (1976).
94. V. Stella, *In* "Prodrugs as Novel Drug Delivery Systems" (T. Higuchi and V. Stella, eds.), p. 1. ACS Symposium Series, Washington, D.C., 1975.
95. A. Robert, W. E. Magee, O. V. Miller, and J. E. Nezamis, *Biochim. Biophys. Acta* **348,** 269 (1974).
96. G. L. Bundy, E. W. Yankee, J. R. Weeks, and W. L. Miller, *Adv. Biosci.* **9,** 125 (1975).
97. E. Nielschlag, J. Mauss, A. Coert, and P. Kicovic, *Acta. Endocrinol.* **79,** 366 (1975).
98. G. L. Sladen, *In* "Intestinal Absorption in Man" (I. McColl and G. E. Sladen, eds.), p. 1. Academic Press, New York, 1975.
99. A. Swahn, *Europ. J. Clin. Pharmacol.* **9,** 299 (1976).
100. T. Gessner, R. Trudnowski, R. Rico, and J. Rempel, *J. Pharm. Pharmacol.* **28,** 79 (1976).
101. D. R. Holland and J. F. Quay, *J. Pharm. Sci.* **65,** 417 (1976).
102. R. H. Hook, J. M. Williams, and G. J. Wright, *Toxicol. Appl. Pharmacol.* **25,** 445 (1973).

Chapter 4 The Masca Model of Pharmacochemistry: I. Multivariate Statistics

Peter P. Mager

ISBN 0-12-060309-8

I. Introduction

This chapter provides the foundations for understanding of some important statistical techniques in the multivariate analysis in the *Masca* model of Pharmacochemistry, that is, the *m*ultivariate *a*nalysis of *s*tructure–activity relationships in *c*ombination with the multivariate bio*a*ssay (*92, 106–111, 113–118*). Since publication of the first report, considerable progress has taken place and many examples are now available to demonstrate my viewpoint of multivariate analysis in drug design and related fields, including linear free energy relationships in organic chemistry. Scheme 1 gives an overview of the design blocks of the Masca model. It must be emphasized that only an integration of all blocks leads to the desired results.

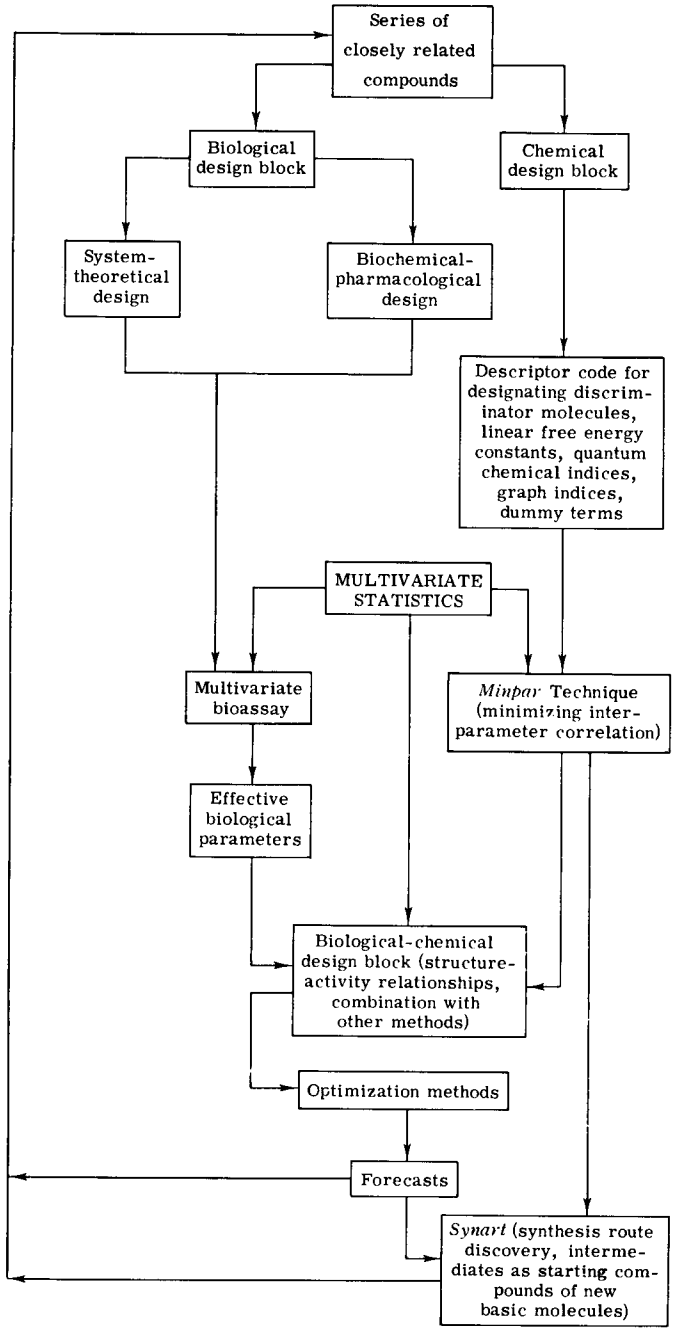
The Biological Design Block consists of the Systemtheoretical and the Biochemical–Pharmacological Blocks. The applied system theory of nonlinear, nonautonomous, nonconservative systems deals with the interrelationships of external and internal signals, biosystems and their environments (such as cellular environments), uncoupling, and the influence of xenobiotics (exogeneous chemicals such as drugs, metabolic inhibitors, and pesticides) and physiologically acting compounds (such as amino acids and endogeneous opioids like endomorphine) on the spatiotemporal structure of biosystems (*122–128, 131–133*). It has been pointed out that hysteresis, oscillations, and spatial changes are common properties of coupled systems (from biological polyelectrolytes to ecosystems, from cells to the heart), and bioactive compounds influence these properties. There is evidence that diseases are characterized by a change of coupling relations. The general rules of biochemical pharmacology deal with the structure of xenobiotics and their pharmacokinetic behavior (absorption, bonding to plasma proteins, metabolism, and transport through mem-

branes like stomach mucosa, the blood–brain barrier, or intracellular structures, elimination of drugs, and so forth), and there are many possibilities for avoiding the formation of metabolically unstable compounds, epoxidation, and long-term toxicity of drugs (*71, 96, 100–105, 117, 118, 136c*) before a compound is synthesized.

The Chemical Design Block consists of a large chemical data bank of linear free energy parameters and extrathermodynamic variables of 450 substituents (*135*), a descriptor code for designating discriminator molecules, and (in preparation) the *Synart* technique (*syn*thesis route discovery by using a reaction library and methods of *art*ificial intelligence). The *Minpar* technique, that is, *min*imizing inter*par*ameter correlation of the physicochemical variables, leads then to a good range and spread of physicochemical parameter values.

Multivariate Statistics is the background for the statistical analysis of biological data. The effective biological values are selected from the computer output (such as significant biological variables, time points of biological events, effective doses, significant treatment groups, and so forth). The effective variables are regressed against the chemical *Minpar* parameters in order to get initial information about the role of electronic, steric, and lipophilic features in the biological action. The Optimization Block (extremal value method, gradient technique, surface response method, Box-Wilson technique) leads to forecasts of xenobiotics with desired biological activities, for instance, a compromise between effectivity and toxicity or, if possible, the discrimination between toxicity and effectivity. Examples are given later. It is clear that some predicted compounds can only be synthesized if a new synthethic route is discovered. In addition, the discovery of already extant routes of series of closely related compounds can lead to novel series (new basic molecule). The reason is that intermediates are often biologically active and can be tested as starting compounds of a novel series. An example is the opioids, involving a change from rigid to more flexible molecules. The common basis is a sigmoidal shaped bridge with a cationic nitrogen head and stereochemically determined rules (Vol. X of Drug Design).

Without a systematic means for selecting compounds in chemistry, drug design, and biochemistry, it is unlikely that classical methods will be able to find optimal compounds. For instance, consider a benzene ring with two sites for substitution and assume the introduction of a second phenyl group via a simple bridge such as —CONH— that increases the possibility of combinations enormously. If we limit the number of substituents to well-selected functions, for example, we have 165 substituents. The total number of all combinations is 7.4119×10^8. Making 74,000 molecules would provide a statistically significant representative sample size. Few



Scheme 1. Integration of methods used in the *Masca* model of Pharmacochemistry. (Mutivariate statistics are discussed here.)

research groups are able to make more than one of a thousand compounds for experimental testing. Taking 74 congeners would imply that we have a chance of 10^{-7} of finding the best compound; that is, we have a probability of 0.00001% of finding the best compound. Only theoretical models can solve the problem: Selection of the best compound for synthesis—with a minimum of effort and cost, including ease of synthesis—is a crucial and proper criterion for determining the success of various methods that have been proposed.

Now, let us speculate as to why the usual techniques lead to so few correct predictions. Experimental research in quantitative drug design tends to focus on a single variable, and precision is gained by constant sources of independent measurements. An example is a known biological assay (*12, 28, 37, 164, 170*) and a known structure–activity analysis (*36, 36a*) wherein a single biological variable is correlated against multiple physicochemical variables. However, bioactive compounds induce multidimensional responses that are dependent on several variables. This multidimensionality can be analyzed by using modern methods of biomathematics. The complexity of biosystems requires that present-day models in medical chemistry be generalized in order to analyze the multidimensionality of biological responses, for instance, the time-dependence of the final regression equations in structure–activity relationships. We think that the *Masca* model is a necessary starting point, providing a small beginning for analyzing the great complexity of biological events. From a practical viewpoint, more time and costs are necessary, and this includes a basic knowledge of multivariate statistics, system theory, pharmacochemistry (in a broad sense), and physical chemistry. The final results, however, more than compensate for these drawbacks. Note that we have defined "pharmacochemistry" as the relationship between drug design, biochemical pharmacology, pharmacy, medicine, physical chemistry, and biomathematics (*92*).

II. General Remarks on Multivariate Statistics

A. Probability of Type I and II Error

If many variables need be analyzed, why can many univariate test statistics not be employed? In the conventional (univariate) test theory, the probability of a type I error experimentwise is given by

$$\alpha_e = 1 - (1 - \alpha)^m$$

for more than one independent comparison, $m = 1, \ldots, g$ (*67, 165*) with

a constant significance level α. For example, we want to examine the means of two treatment groups (standard and test group) at each $m = 6$ successive days. We assume a significance level of 5% and that the time points are independently distributed. (It is clear that this condition is seldom satisfied.) Therefore, if a two-sample test is used (as Student's t test), the probability of a type I error, experimentwise, is 26.5%, although the experimenter sees it is 5%. The corresponding probability of a type I error, experimentwise, for dependent comparisons (the true situation in that example) is more difficult to ascertain and details cannot be given here. It is only important that the general principle is known. However, it should be noted that the α_e of dependent comparisons is larger (32.5%).

Now, let α_v be the type I error that controls the deviations between empirical and theoretical distributions. A method for calculating this error is the Monte Carlo technique. Generally, we obtain

$$\alpha^* = \alpha_e + \alpha_v$$

for the sum of type I errors that are actually found if multiple measurement designs are analyzed by known methods. It seems clear that a condition for an exact analysis is that $\alpha^* \leqslant \alpha$. Note that conventional methods must be considered with caution when many variables are analyzed. In each multivariate design we get $\alpha = \alpha_e$ (because of $m = 1$, that is, a single test is used that contains the total information about the sample); and the condition $\alpha_v = 0$ is exactly satisfied if the assumptions of a multivariate normal distribution (*48, 61*) and a homogeneity of error matrices (*81*) are satisfied. Both assumptions can be practically fulfilled if the amount N_j of all objects in each treatment group $j = 1, \ldots, q$ are equal, $n = N_1 = \ldots = N_q$, and if $n_h \leqslant p$ is nearly satisfied (*129*), where n_h is the degree of freedom of any statistical hypothesis (see later) and p denotes the number of variables.

If the number of all objects in each treatment group is different, there are several methods for examining the assumption of a normal distribution of each variable (*75, 88, 89*) and of all variables (*140, 142–144*). In contrast to the univariate nonparametric tests (*77, 80, 98, 119, 130*), the multivariate nonparametric tests (*53, 138, 157, 167, 183*) do not possess an economy with respect to the computing time, and determined transformations (*3, 8, 75, 88, 89, 156*) can be employed in order to obtain the desired normality. On the other hand, if both the conditions just mentioned are approximatively satisfied, the conditions of homogeneity of variance (univariate case) and, respectively, error matrices (multivariate case) are practically satisfied (*32, 41, 43–46, 81, 112, 141, 174, 181, 188*). There are methods to compensate for the lack of measures (missing observations in statistical literature) (*1, 28, 177*).

The probability of a type II error, experimentwise and distributionwise, increases if univariate methods are used in multidimensional design. Some special cases are illustrated in the literature (*148*) and described mathematically (*14, 29*), but, in general, the methods cannot be applied experimentally.

The analysis of more than one variable is of an omnibus nature; one seeks to answer many questions by a single test. Therefore, we have used the so-called multivariate simultaneous *a posteriori* tests (*Sp* test) with statistics that examine the means of treatment effects, doses, times, and variables where the problem of the error rate, experimentwise, can be statistically controlled. *Sp* tests are used after a significant multivariate result. Note that univariate analogs have been described elsewhere (*24, 30, 147*).

Corollary: A modern data analysis in drug design requires that many influences on biosystems be considered. From an exact statistical viewpoint, data with dependences among dimensions can only be analyzed by using methods of multivariate statistics.

B. MULTIVARIATE TEST CRITERIA

Let S_h be a matrix due to any statistical hypothesis, let S_e be an error matrix, and let $S_t = S_h + S_e$. Starting from these matrices, several criteria have been proposed for testing null hypotheses (*55, 67, 69, 107, 129*). Suppose that

$$\det(S_h S_e^{-1} + \lambda E) = 0$$

where det denotes a determinant, $\lambda = (\lambda_1, \ldots, \lambda_s)$ are real and ordered eigenvalues, $s = \min\{n_h, p\}$, and E is an unit matrix. Another starting point is

$$\det(S_h S_t^{-1} - \lambda^* E) = 0$$

The corresponding degrees of freedom of the three matrices are given by n_h, n_e, and n_t. Among the test criteria in multivariate statistics, the following are used:

$$\Lambda = \det(S_e S_t^{-1}) = \prod_{i=1}^{s} (1 - \lambda_i)^{-1} = \prod_{i=1}^{s} (1 - \lambda_i^*)$$

$$V = \operatorname{tr}(S_h S_e^{-1}) = \sum_{i=1}^{s} \lambda_i = \sum_{1=1}^{s} \lambda_i^* (1 - \lambda_i^*)^{-1}$$

$$\theta = \max\{\lambda_i (1 + \lambda_i)^{-1}\} = \max\{\lambda_i^*\}$$

(*69, 108, 129*). Here, tr denotes the trace of a matrix, and det symbolizes the determinant. Any null hypothesis is rejected at a given significance level α if

$$\Lambda \leq c_1$$
$$V \geq c_2$$
$$\theta \geq c_2$$

is satisfied. Here, c_1, c_2, and c_3 are appropriate constants to ensure attainment of a preassignment significance level (*69, 108, 129, 134*). The following notation is important; at the present time more is known of the comparative power of the three test criteria than was known at the time of development of the test criteria. First, all three tests have a monotonically increasing power function, and the power differs at most in the second decimal place. Second, the likelihood criterion Λ is not so sensitive to group differences that tend to fall along a particular dimension or continuum in the multivariate measurement space. When differences among several groups tend to fall along a single continuum, θ will be more powerful than V and Λ.

The points of significance are listed in many tables for determined degrees of freedom (*38, 39, 42, 55, 57, 146, 150–154, 169, 173*). If the tables are not available, the following approximations can be used (*129*):

$$c_1 = \exp(-q\chi^2_{f;\alpha})$$

where χ^2 is adapted from the χ^2 distribution with $f = n_h p$ degrees of freedom and $q = (C_{p,n_h,M;\alpha})/\tilde{m}$ where the correction factor C is adapted from an overall-available table (*152*), with $M = n_e + 1 - p$ and $\tilde{m} = n_h + n_e - \frac{1}{2}(p + n_h + 1)$. For $\tilde{m}(1 + n_h)^{-1} \geq (n_h^2 + p^2)$, C tends to 1. The value of V is given by

$$c_2 = \frac{f_1 s F_{f_1,f_2;\alpha}}{f_2}$$

where F is the significance point adapted from the F distribution with $f_1 = s(2m + s + 1), f_2 = 2(s\tilde{n} + 1)$ degrees of freedom and $m = \frac{1}{2}(|n_h - p| - 1)$, $\tilde{n} = \frac{1}{2}(n_e - \mathrm{p} - 1)$. The critical values of θ must be directly adapted from a table (*39, 55*). For $s = 1$ the critical values are not tabulated and we get

$$c_3 = z(1 + z)^{-1}$$

where $z = F_{f_1,f_2;\alpha}(m + 1)/(\tilde{n} + 1)$, $f_1 = 2(m + 1)$, $f_2 = 2(\tilde{n} + 1)$. In contrast to the usual procedures (*5, 15, 27*), in the computing scheme (*113*) for the *Masca* model, the three test criteria are included (*135*). If each multivariate criterion is significantly different from zero, the *Sp* test can be done.

Corollary: Each multivariate design is based on the calculation of hypothesis and error matrices defined for a special design in different mathematical formalisms. The calculation of the test criteria is based on the same formalisms.

III. Multivariate Bioassay

A. EXPERIMENTAL DESIGN

The first generalization of the usual bioassay was done by employment of the T^2 test of Hotelling (*76*), the second by application of the one-way classification of *Manova,* that is, of the multivariate variance analysis (*101, 102*). A more general model is the two-way classification (*85, 107, 182*) introduced into the bioassay by the author (*107, 118, 129*).

Table I shows the experimental design where $(y_{hijk}) = Y_{ijh}$ is the observed vector with $h = 1, \ldots, n; i = 1, \ldots, m; j = 1, \ldots, q; k = 1, \ldots, p$ as indices.* In contrast to the earlier literature (*48, 109, 118*), we provide a "total" example, including the *Sp* test and the discriminant analysis (*Discra*). Before any further considerations, let us give some examples of application:

h bioobjects (individuals, animals, bacteria, plants, biochemical measurements such as enzyme activities);

i rows (molar doses, effective doses, concentrations of substrates, and animal groups with different ages or weights);

j treatments (standard group and test groups including a control and substituents in structure–activity studies);

k variables (biological variables such as blood pressure, heart frequency, enzyme activity, flow rate of drugs, antihistaminic and anticholinergic response and time-dependent measurements such as reaction rate and time-dependence of analgesic responses). The battery of variables consists of dependent measurements.

Therefore, for example, a simultaneous dose–time–structure analysis can be done. It can be determined which doses, times, treatment groups, and other variables are significantly different from zero (*Manova* and *Sp* test). In addition, a discrimination of drugs can be mathematically described, and a standardization of drugs accomplished. The usual method of ordering of drugs with respect to diseases (hypnotic, antitumor, and so forth) or chemical skeletons (phenothiazines, benzoic acids) has long been recog-

* The indices ($i = 1, \ldots, m$ of Y_{ijh}) should not be exchanged with those of the test criteria ($i = 1, \ldots, s$; m degrees of freedom).

TABLE I
EXPERIMENTAL DESIGN OF MANOVA, KANORA, AND MACOVA

			Treatment Group 1 ··· q							
			First battery		Second battery		First battery		Second battery	
Type	Row	Object	1 ··· p		1 ··· c		1 ··· p		1 ··· c	
Manova	1	1	y_{1111} ··· y_{111p}		0 ··· 0		··· y_{1q11} ··· y_{1q1p}		0 ··· 0	
		n	y_{n111} ··· y_{n11p}		0 ··· 0		··· y_{nq11} ··· y_{nq1p}		0 ··· 0	
	m	1	y_{11m1} ··· y_{11mp}		0 ··· 0		··· y_{1qm1} ··· y_{1qmp}		0 ··· 0	
		n	y_{n1m1} ··· y_{n1mp}		0 ··· 0		··· y_{nqm1} ··· y_{nqmp}		0 ··· 0	
Kanora	1	1	y_{11} ··· y_{1p}	x_{11}	··· x_{1c}	0 0	··· 0	0	··· 0	
		n	y_{n1} ··· y_{np}	x_{n1}	··· x_{nc}	0 0	··· 0	0	··· 0	
	0	0	0 ··· 0	0	··· 0	0 0	··· 0	0	··· 0	
Macova	1	1	y_{1111} ··· y_{111p}	x_{1111}	··· x_{111c}	··· y_{1q11}	··· y_{1q1p}	x_{1q11}	··· x_{1q1c}	
		n	y_{n111} ··· y_{n11p}	x_{n111}	··· x_{n11c}	··· y_{nq11}	··· y_{nq1p}	x_{nq11}	··· x_{nq1c}	
	m	1	y_{11m1} ··· y_{11mp}	x_{11m1}	··· x_{11mc}	··· y_{1qm1}	··· y_{1qmp}	x_{1qm1}	··· x_{1qmc}	
		n	y_{n1m1} ··· y_{n1mp}	x_{n1m1}	··· x_{n1mc}	··· y_{nqm1}	··· y_{nqmp}	x_{nqm1}	··· x_{nqmc}	

nized as unsatisfactory, and a new standardization is possible if the *Clasca* (classification technique) is employed. A very special case of our model ($m = 1, q = 2$) of such assignment procedure has been demonstrated in the literature (*10*).

B. GENERAL MULTIVARIATE MODEL

The generalization of multivariate model theory (190) leads to mean vectors, matrices, and multivariate test criteria. If the test criteria are significant, the *Sp* test can be done and, in addition, the discriminant and classification analysis. If successive time points are analyzed, a profile analysis follows. All results lead to information about the statistical role of

the measurements, and we obtain the so-called effective biological values of the Biological Design Block of Scheme 1. Note that the effect of concomitant variables can be controlled by using the multivariate covariance analysis, and an unbiased estimate of the treatment effects can be used as biological response in the Biological Design Block (see later).

In Table I, the estimations of the mean vectors are defined by:

$$\bar{\mathbf{m}} = (mqn)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{q} \sum_{h=1}^{n} \mathbf{Y}_{ijh}, \bar{\mathbf{m}}_j = (nm)^{-1} \sum_{i=1}^{m} \sum_{h=1}^{n} \mathbf{Y}_{ijh}$$

$$\bar{\mathbf{m}}_{ij} = n^{-1} \sum_{h=1}^{n} \mathbf{Y}_{ijh}, \bar{\mathbf{m}}_i = (nq)^{-1} \sum_{j=1}^{q} \sum_{h=1}^{n} \mathbf{Y}_{ijh}$$

The matrices of hypotheses (columns, rows, interactions = cells) are obtained by:

$$\mathbf{S}_{h1} = \mathbf{B}_1 = nq \sum_{i=1}^{m} (\bar{\mathbf{m}}_i - \bar{\mathbf{m}})(\bar{\mathbf{m}}_i - \bar{\mathbf{m}})'$$

$$\mathbf{S}_{h2} = \mathbf{B}_2 = nm \sum_{j=1}^{q} (\bar{\mathbf{m}}_j - \bar{\mathbf{m}})(\bar{\mathbf{m}}_j - \bar{\mathbf{m}})'$$

$$\mathbf{S}_{h3} = \mathbf{B}_3 = n \sum_{i=1}^{m} \sum_{j=1}^{q} (\bar{\mathbf{m}}_{ij} - \bar{\mathbf{m}}_i - \bar{\mathbf{m}}_j + \bar{\mathbf{m}})(\bar{\mathbf{m}}_{ij} - \bar{\mathbf{m}}_i - \bar{\mathbf{m}}_j + \bar{\mathbf{m}})'$$

with $n_{hi} = f_1 = m^{-1}$, $n_{h2} = f_2 = q^{-1}$, $n_{h3} = f_3 = f_1 f_2$ degrees of freedom. The error matrix is given by:

$$\mathbf{S}_e = \mathbf{W} = \sum_{i=1}^{m} \sum_{j=1}^{q} \sum_{h=1}^{n} (\mathbf{Y}_{ijh} - \bar{\mathbf{m}}_{ij})(\mathbf{Y}_{ijh} - \bar{\mathbf{m}}_{ij})'$$

and the total matrix is

$$\mathbf{S}_t = \mathbf{T} = \sum_{i=1}^{m} \sum_{j=1}^{q} \sum_{h=1}^{n} (\mathbf{Y}_{ijh} - \bar{\mathbf{m}})(\mathbf{Y}_{ijh} - \bar{\mathbf{m}})' = \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3 + \mathbf{W}$$

with $n_e = mq(n - 1)$, $n_t = nmq - 1$. Note that the two-way classification goes over into the univariate two-way classification for $p = 1$, that, $m = 1$ yields the multivariate one-way classification, that $p = m = 1$ leads to the univariate one-way classification, and that $m = 1, q = 2$ is identical with the test statistic of Hotelling's T^2 test (resp. Mehalanobis D^2 discrimination analysis). For $p = m = 1, q = 2$ we get the t test developed by Gosset ("Student"). In the mainline Fortran program, the dimensions in the first column indicate whether the general case (Table I) or the special cases are wanted. Then, the computer reads itself the desired calculations.

C. MULTIVARIATE VARIANCE ANALYSIS (MANOVA)

Model I (fixed effects) is applied most often in our drug design. By a generalization of model II of the *Anova* (univariate variance analysis), random effects can be also analyzed and by an analogy to model III of the *Anova* (*179*), mixed effects can also be investigated. When heterogeneous correlations among the pairs of levels of the fixed factors are suspect, special measures must be taken to insure the validity of test statistics of the null hypotheses. Therefore in our drug design, Model I, where the dependent measurements are the variables and the columns and rows the fixed factors, is used.

Let $\boldsymbol{\alpha}_i$ be the vector of row means estimated by $\bar{\mathbf{m}}_i$, $\boldsymbol{\beta}_j$ is estimated by $\bar{\mathbf{m}}_j$, and $(\boldsymbol{\alpha\beta})_{ij}$ is estimated by $\bar{\mathbf{m}}_{ij}$. The null hypotheses are

$$H_{01}:: = (\boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_m), \quad H_{02}:: = (\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_q),$$

$$H_{03}:: = (\boldsymbol{\alpha\beta})_{11} = \ldots = (\boldsymbol{\alpha\beta})_{mq}$$

where H_{03} has the priority about all the other hypotheses. The null hypotheses are rejected if the corresponding test criteria are significantly different from zero. Inserting the matrices into the determinantal equation, it follows that

$$\det(\mathbf{B}_1\mathbf{W}^{-1} - \boldsymbol{\lambda}_1\mathbf{E}) = 0, \quad \det(\mathbf{B}_2\mathbf{W}^{-1} - \boldsymbol{\lambda}_2\mathbf{E}) = 0,$$

$$\det(\mathbf{B}_3\mathbf{W}^{-1} - \boldsymbol{\lambda}_3\mathbf{E}) = 0$$

leads to the wanted eigenvalues. The technique is also demonstrated by an example.

D. DISCRIMINANT ANALYSIS (*Discra*) AND CLASSIFICATION (*Clasca*)

The result of the discriminant analysis of the two-way classification of *Manova* is the transformation of p original and correlated variables to a smaller set of statistically independent discriminant variables (orthogonalization). Let

$$T_{ri^*} = f(Y_k)$$

be the discriminant function of the three effects ($r = 1,2,3$), let $k \leqslant p$, and let $i^* = 1, \ldots, s^*$ be an index corresponding to the significant eigenvalues. Then, the understandardized discriminant coefficients are obtained from

$$(\mathbf{B}_r\mathbf{W}^{-1} - \boldsymbol{\lambda}_r\mathbf{E})\mathbf{V}_r = 0$$

where $\mathbf{V}_r$ is the matrix of eigenvectors. The null hypothesis (the eigen-

value is statistically zero) is rejected if

$$TS_{ri} = -\tilde{m}_r ln\Lambda_{ir} \geq \chi^2_{f_{ri};\alpha}$$

is satisfied where $i = 1, \ldots, s, f_{ri} = p + f_r - (2i + 1)$, $\Lambda_{ri} = (1 + \lambda_{ri})^{-1}$, and $\tilde{m}_r = f_r + n_e - \frac{1}{2}(f_r + p + 1)$. Note that f_r is defined as mentioned previously (degrees of freedom of hypothesis). Therefore, it is only necessary to determine the eigenvectors whose corresponding eigenvalues are significantly different from zero. The working procedure is demonstrated by a numerical example.

In our *Masca* model, classification using normal probability densities is done. The maximum likelihood estimation is achieved by calculation of the probability ratios,

$$\text{Prob}(ij \,|\, \mathbf{x}) = \frac{\pi_{ij} f(\mathbf{x})^{(ij)}}{\sum_{i=1}^{m} \sum_{j=1}^{q} \pi_{ij} f(\mathbf{x})^{(ij)}}$$

Here, $f(\mathbf{x})$ symbolizes the multivariate normal distribution density associated with the score vector $\mathbf{x}$ into populations and π_{ij} is the *a priori* probability. Assuming they are equal, we can neglect the *a priori* probabilities.

E. MULTIVARIATE *Sp* TEST

The general univariate multiple comparison technique is attributed to Scheffé (*174*). The limits can be computed for $\alpha^* \leq \alpha$ and are always wider than those of other methods (sure decision). The multivariate generalization leads directly to a method for controlling the type I error, experimentwise, for all multiple tests on linear functions of the response means. In addition, the correlations among the variables do not "negatively" influence the test statistic (*166, 168*).

The null hypothesis

$$H_{0r}: : = (\mathbf{a}\boldsymbol{\delta}_r = 0)$$

is rejected for a given choice of α if

$$TS_r = |\mathbf{a}'\mathbf{d}_r| \geq k_r(\mathbf{a}'\mathbf{W}\mathbf{a})^{\frac{1}{2}}$$

is satisfied. Here, $\mathbf{d}_r$ is the mean difference vectors of rows, treatments, or cells, $\boldsymbol{\delta}_r$ is estimated by $\mathbf{d}_r$, $\mathbf{a}' \in (1\ 0\ 0\ \ldots\ 0), \ldots, (0\ 0\ 0\ \ldots\ 1)$ and $k_r^2 = 2c_{3r}/g_r(1 - c_{3r})$; with $c_{3r} = \theta_{s_r,m_r,\tilde{n};\alpha}$ where $s_r = \min\{f_r, p\}$, $m_r = \frac{1}{2}(|f_r - p| - 1)$, $\tilde{n} = \frac{1}{2}(n_e - p - 1)$, $g_1 = qn, g_2 = mn$, and $g_3 = n$. Therefore, multiple comparisons among rows, treatments, or cells can be made, and

the nature of construction of test statistic leads directly to simultaneous confidence bounds. A numerical example demonstrates the working procedure.

F. Profile Analysis

Comparisons on means of dependent measures (profile analysis) have great importance in the multivariate bioassay. There are different methods (*13, 82*) for examining whether p variables are mutally dependent. The usual univariate methods—like carry-over, change-over, or split-plot design (*26, 28, 52, 139, 145, 187*)—require, theoretically, that the null hypothesis on spherical symmetry (the variables have equal variances and zero covariances) not be rejected. The conservative univariate methods (*31, 33*) are based on a reduction of degrees of freedom, but that technique is incorrect from a theoretical viewpoint. Other, multivariate techniques (*6, 7, 9, 17, 25, 34, 35, 47, 56, 57, 67, 78, 84, 86, 155, 159–162*) are constructed without the possibility of calculating simultaneous confidence limts (resp., an *Sp* test). Therefore, we are going out from another method.

Let C be a (p,p − 1) matrix that allows one to obtain the wanted mean differences. For instance, we want to compare the first with the second day, the second with the third day, and so on. Then, we obtain C′ equal to

$$\begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ . & . & . & \dots & . & . \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$

The matrices due to hypothesis and error are $\mathbf{S}_{h1} = \mathbf{C}'\mathbf{B}_1\mathbf{C}$, $\mathbf{S}_{h2} = \mathbf{C}'\mathbf{B}_2\mathbf{C}$, $\mathbf{S}_{h3} = \mathbf{C}'\mathbf{B}_3\mathbf{C}$, and $\mathbf{S}_e = \mathbf{C}'\mathbf{W}\mathbf{C}$. The multivariate test criteria are calculated as mentioned previously. The corresponding *Sp* test is given by

$$TS_r = |\mathbf{a}'(\mathbf{C}\mathbf{d}_r)| \geqslant k_r(\mathbf{a}'\mathbf{S}_e\mathbf{a})^{\frac{1}{2}}$$

where the notations are the same as previously mentioned.

G. Multivariate Covariance Analysis (*Macova*)

To obtain a better examination of biological variables, the one-way classification of multivariate covariance analysis (*Macova*) was applied (*19, 50, 83, 107, 118*). Here, the effects of concomitant variables (different initial conditions of the individuals, heterogeneous factors, different body weights of animals, and so on) are controlled. Therefore, it is possible to

reduce the variability due to experimental error and to obtain unbiased estimates of the treatment effects. Then, the corrected or adjusted values are used for a structure–activity analysis. The *Macova* can be employed after the data have been collected, whereas a stratification of data into homogeneous blocks (rows or columns in Table I) must be carried out at the beginning of an experiment. In contrast to the univariate covariance analysis (*184*), the multivariate concomitant variables can be influenced by the test variables. If the vectors of the test variables and the concomitant variables are represented by **Y** and **X**, the adjustment used in *Macova* is based on the partioned forms of **W**, **B**, and **T**. The experimental design of the two-way classification of *Macova* is given in Table I. We calculate the matrices of hypotheses in the usual way, and by using the partioned matrices, we obtain

$$\mathbf{S}_{hr} = (\mathbf{B}_{yy} - \mathbf{B}_{yx}\mathbf{B}_{xx}^{-1}\mathbf{B}_{xy})_r$$

($r = 1,2,3$) where $\mathbf{B}_{yy}$ is the matrix of hypothesis of **Y**, $\mathbf{B}_{xx}$ is related to **X**, and $\mathbf{B}_{yx} = \mathbf{B}'_{xy}$ contains the interaction elements. In strict analogy, we get

$$\mathbf{S}_e = \mathbf{W}_{yy} - \mathbf{W}_{yx}\mathbf{W}_{xx}^{-1}\mathbf{W}_{xy}$$

and

$$\mathbf{S}_t = \mathbf{T}_{yy} - \mathbf{T}_{yx}\mathbf{T}_{xx}^{-1}\mathbf{T}_{xy} = \mathbf{S}_{h_1} + \mathbf{S}_{h_2} + \mathbf{S}_{h_3} + \mathbf{S}_e$$

where $n_{h1} = f_1 = m - 1$, $n_{h2} = f_2 = q - 1$, $n_{h3} = f_3 = f_1 f_2$, $n_e = mq(n - 1) - c$, and $n_t = mqn - c - 1$ (for notations see Table I).

The multivariate test criteria are calculated from the eigenvalues of the determinal equation $\det(\mathbf{S}_{hr}\mathbf{S}_e^{-1} - \boldsymbol{\lambda}_r\mathbf{E}) = 0$. Significance means that the corrected mean vectors

$$\tilde{\mathbf{y}}_i = \bar{\mathbf{y}}_i - \mathbf{B}_1(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_1)$$

$$\tilde{\mathbf{y}}_j = \bar{\mathbf{y}}_j - \mathbf{B}_2(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_2)$$

$$\tilde{\mathbf{y}}_{ij} = \bar{\mathbf{y}}_{ij} - \mathbf{B}_3(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_3)$$

of the rows, columns, and cells are significantly different from zero. Here, $\mathbf{B}_r = (\mathbf{W}_{yx}\mathbf{W}_{xx})_r$ with $r = 1,2,3$; $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, $\bar{\mathbf{x}}_3$ are the grand mean vectors of the rows, columns, and cells of the test variables (resp. concomitant variables).

If the multivariate null hypotheses are rejected, a multivariate *Sp* test can be computed or, alternatively, a set of simultaneous confidence intervals for the contrasts of all rows, columns, and cells. The null hypotheses are rejected if

$$TS_r = |\mathbf{a}'\mathbf{d}_r| \geq k_r g_r(\mathbf{a}'\mathbf{S}_e\mathbf{a})^{\frac{1}{2}}$$

where $k_r^2 = 2\theta_{s_r,m_r,\bar{n};\alpha}/b_r(1 - \theta_{s_r,m_r,\bar{n};\alpha})$ with $b_1 = qn$, $b_2 = mn$, $b_3 = n$, $s_r =$

$\min\{f_r,p\}$, $m_r = \frac{1}{2}(|f_r - p| - 1)$, $\tilde{n} = \frac{1}{2}(n_e - p - 1)$, and $g_r = 1 + f_r^{-1}\mathbf{a}'\mathbf{B}_{xx(r)}\mathbf{W}_{xx(r)}^{-1}\mathbf{a}$ tending to 1 for a large number of degrees of freedom $n_{h1} = f_1$, $n_{h2} = f_2$, $n_{h3} = f_3$.

Although the discriminant function technique has been employed in many fields of statistical inquiry, no generalization has been published with respect to the covariance analysis. A combination is possible, of course. We are going out from $(\mathbf{S}_{hr}\mathbf{S}_e^{-1} - \boldsymbol{\lambda}_r\mathbf{E})\mathbf{V}_r = 0$, where the matrix $\mathbf{V}_r$ of eigenvectors represents the adjusted discriminant functions. Therefore, a separation of concomitant variables (for example, drug side effects) is possible in the discriminant analysis, and the adjusted variables of the main action of drugs can be used in order to classify new drugs.

H. Working Procedure in the Multivariate Bioassay

Because of the presence of nonnegligible correlations among biological dimensions, the statistical approach for comparing effective doses, times, drug treatments, and so forth is here based on a multivariate overall test of significance (*Manova*). If the *Manova* leads to the conclusion that nonsignificant effects exist, the *Macova* must be done in order to discriminate the possible high variability from the proper biological effects.

If the hypothesis of equal treatment effects, rows, or cells is rejected by using *Manova* or *Macova* (priority of the hypothesis of cell effects), the *Sp* tests can be used to determine which of the effects differ.

The significant effects are regressed against physicochemical constants (multivariate structure–activity analysis) in the following phase of the Masca approach. In addition to the significance testing of doses, treatments, and times (profile analysis), all measurement variables are considered collectively, and the best discriminant functions (significant eigenvectors corresponding to significant eigenvalues) are used in order to store the information in the computer.

Corollary: A univariate approach may be quite misleading if dependent dimensions and individual tests are considered. This, therefore, provides a strong motivation for a multivariate approach in bioassay.

IV. Reduction of Dimensionality (Principal Component Analysis)

There are different methods for reducing p dependent variables into f few variables. First, the cluster analysis (*20, 171, 176, 178*) is used. However, probability theory has a very small role to play in cluster analysis. The consequence is that we cannot set standard errors or confidence bands, and we cannot conduct exact significance tests on any particular

clustering. Furthermore, the reliability of the results is questionable (*50a*). Therefore, in contrast to Hansch (*36b*), we do not believe that the cluster analysis is a general aid in the selection of substituents. Second, the polynomial regression analysis can be used if time-dependent data are analyzed (*149*). The logic is that one can expect that, if there is a time-dependent treatment effect, it will result in a general shift in performance over time. In the Masca approach we use the following methods: (1) the *Discra*, (2) the canonical correlation (see later), and (3) the principal component analysis (*Princo*). The unstandardized form of the *Princo* is going out from

$$(\mathbf{S} - \lambda_i \mathbf{E})\mathbf{v}_i = 0$$

(*2, 18, 21, 40, 82*) where $\mathbf{S} = (n - 1)^{-1}\mathbf{W}$ is the dispersion matrix. In strict analogy to the *Discra*, the first principal component of the variables (for instance, $X_1, \ldots, X_p$) if defined by

$$C_1 = v_{11}x_1 + \ldots + v_{1p}x_p$$

and the second by

$$C_2 = v_{21}X_1 + \ldots + v_{2p}X_p$$

(and so on), $\mathbf{v}_1$, $\mathbf{v}_2$ are associated eigenvectors of the both largest eigenvalues λ_1, λ_2 of the ordered eigenvalues ($\lambda^1 \geq \lambda_2 \ldots \geq \lambda_s$). Note that λ_i is interpretable as the sample variance of C_i, and if $\lambda_i \neq \lambda_j$ ($i,j = 1, \ldots, s; i \neq j$), the components are necessarily orthogonal. The importance of λ_i is statistically estimated (see discriminant analysis), and $\lambda_i/\text{tr}(\mathbf{S})$ indicates visually the importance (the trace of **S** is identical with the sum of all eigenvalues, of course). Note that $v_{ij}(\lambda_{ij})^{\frac{1}{2}}s_{ii}$ is the correlation of the ith variable and the jth component (s_{ii} is the standard deviation). The reduction of dimensionality occurs if those eigenvectors can be extracted whose corresponding eigenvalues are significant and, in addition, if $f < p$.

The problem posed by unequal variances of the original variables can be overcome by transforming each variable to standard form (the **X** matrix goes over into **Z**) or by using the correlation statistics (**R** is the correlation matrix)

$$(\mathbf{R} - \phi_i \mathbf{E})\mathbf{w}_i = 0$$

The transition into orthogonal variables occurs by calculation of $\mathbf{F} = \mathbf{ZB}$ where $\mathbf{B} = \mathbf{L}^{-\frac{1}{2}}\mathbf{V}$ with **L** equal to

$$\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & 0 & \ldots & \lambda_s \end{bmatrix}$$

and **V** = matrix of eigenvalues (*82*). This method was used to transform the matrix of dependent physicochemical constants into uncorrelated constants (*135*) (for more details see Section IV,E).

Second, the *Princo* was employed in order to discriminate the overall response of biological events into pharmacokinetic and pharmacodynamic components (*21a*). Third, we have used the *Princo* to obtain information about the common basic molecule of rigid, semirigid, and flexible analgesics. It was also possible to separate the moiety responsible for analgesia from those molecules responsible for physical dependence and respiratory depression. Therefore, it is possible to discriminate the basic molecules corresponding to biological responses of diverse structural agents. Fourth, Cammarata (*11a*) was able to classify a set of different compounds into major pharmacological categories. Although criticisms have been directed at Cammarata's method, I find such criticisms invalid.

V. Multivariate Structure–Activity Analysis

A. Experimental Design

The design of a multivariate structure–activity analysis (Table I) is based on the multivariate correlation and regression analysis, called *Kanora* (*108, 111, 117, 118*), in which many biological responses are regressed against many chemical properties (*115, 116, 135*).

B. General Multivariate Model

Now, let **S** be the estimation of the dispersion matrix $\mathbf{S} = (n - 1)^{-1}\mathbf{W}$ ($i = j = 1$) of the data in Table I, written in the partioned form (*163*)

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \hline \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{array}\right]$$

where $\mathbf{S}_{xx}$ is the (c,c)-dispersion matrix (see Table I), $\mathbf{S}_{yy}$ is the (p,p)-dispersion matrix, and $\mathbf{S}_{xy}$ is the (e,p)-dispersion matrix (*82, 107, 129*). Let $\mathbf{m}_x = (m_{x1} \ldots m_{xc})'$ and $\mathbf{m}_y = (m_{y1} \ldots m_{yp})'$ be the mean vectors. By a generalization of the known multiple regression analysis (*16, 51, 156*), we obtain the multivariate regression systems

$$\mathbf{Y} = \mathbf{B}_x\mathbf{X}, \quad \mathbf{X} = \mathbf{B}_y\mathbf{Y}$$

(*107*) where both the regression matrices are defined by

$$\mathbf{B}_x = (\mathbf{b}_{xo}\mathbf{B}_{xc}), \quad \mathbf{B}_y = (\mathbf{b}_{yo}\mathbf{B}_{yp})$$

with

$$\mathbf{B}_{xc} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad \mathbf{B}_{yp} = \mathbf{S}_{xy}\mathbf{S}_{yy}^{-1} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$$

and with the intercept vectors

$$\mathbf{b}_{xo} = \mathbf{m}_y - \mathbf{B}_{xc}\mathbf{m}_x, \quad \mathbf{b}_{yo} = \mathbf{m}_x - \mathbf{B}_{yp}\mathbf{m}_y$$

A bias due to ignoring correlations between departures can be avoided if

$$\mathbf{B}_{xc} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}\mathbf{Y}, \quad \mathbf{B}_{yp} = (\mathbf{Y}'\mathbf{S}^{-1}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{S}^{-1}\mathbf{X}$$

is computed. Basing calculations on the regression matrices, the theoretically calculated data are obtained. The matrices due to regression hypothesis are given by

$$\mathbf{S}_{hx} = \mathbf{B}_{xc}\mathbf{S}_{xy}, \quad \mathbf{S}_{hy} = \mathbf{B}_{yp}\mathbf{S}_{yx}$$

and the matrices due to an error are given by

$$\mathbf{S}_{ex} = \mathbf{S}_{tx} - \mathbf{S}_{hx} = \mathbf{S}_{yy} - \mathbf{S}_{hx}, \qquad \mathbf{S}_{ey} = \mathbf{S}_{ty} - \mathbf{S}_{hy} = \mathbf{S}_{xx} - \mathbf{S}_{hy}$$

The degrees of freedom are $n_{hx} = c$, $n_{ex} = n - c - 1$, $n_{tx} = n_{ty} = n - 1$, $n_{hy} = p$, $n_{ey} = n - p - 1$. The determinal equations

$$\det(\mathbf{S}_{hx}\mathbf{S}_{ex}^{-1} - \lambda\mathbf{E}) = 0, \quad \det(\mathbf{S}_{hy}\mathbf{S}_{ey}^{-1} - \lambda\mathbf{E}) = 0$$

lead to the maximal measurement of a multivariate correlation,

$$R_c^2 = \max\{\lambda_i(1 + \lambda_i)^{-1}\}; \quad i = 1, \ldots, s = \min\{p,c\}$$

that is, the correlation between **X** and **Y**. The observed R_c^2 is significantly different from zero if $\Lambda = (1 - R_c^2)$ is significant. The examination was mentioned in Section II,B. If R_c^2 is significant, the multiple correlation coefficients R_t (resp. R_l) are computed, $R_t^2 = C_t/D_t$, $R_l^2 = C_l/D_l$ with $t = 1, \ldots, p$; $l = 1, \ldots, c$; and $(C_1 \ldots C_p)' = \text{diag}(\mathbf{S}_{hx})$, $(D_1 \ldots D_p)' = \text{diag}(\mathbf{S}_{tx})$, $(C_1 \ldots C_c)' = \text{diag}(\mathbf{S}_{hy})$, $(D_1 \ldots D_c)' = \text{diag}(\mathbf{S}_{ty})$. The multiple correlation coefficients are statistically examined by using $\Lambda_t = (1 - R_t^2)$, $\Lambda_l = (1 - R_l^2)$ at $n_{hx} = 1, n_{hy} = 1$. The results of omitting the insignificant equations (corresponding to insignificant multiple correlation coefficients) are examined with respect to the partial regression coefficients by using the *Sp* test.

C. MULTIVARIATE *Sp* TEST

At this time, there is no universally accepted way of dealing with the problem of selecting regressors in multiple regression analysis. In the

multivariate correlation and regression analysis (*Kanora*), we require that a decision as to whether or not a regressor is to be included depend on the precision with which the corresponding partial regression coefficients are significantly different from zero. Therefore, the *Sp* test must be computed. The null hypotheses are rejected if

$$TS_x = |\mathbf{a}\mathbf{B}_{xc}\mathbf{c}'| \geq k[(\mathbf{c}'\mathbf{S}_{ex}\mathbf{c})(\mathbf{a}'\mathbf{S}_{xx}^{-1}\mathbf{a})]^{\frac{1}{2}}$$

$$TS_y = |\mathbf{a}\mathbf{B}_{yp}\mathbf{c}'| \geq k[(\mathbf{c}'\mathbf{S}_{ey}\mathbf{c})(\mathbf{a}'\mathbf{S}_{yy}^{-1}\mathbf{a})]^{\frac{1}{2}}$$

is satisfied. Here, **a** and **c** are non-null vectors selected from an examination of the prominence of regression coefficients, and $k^2 = \theta_{s,m,\tilde{n};\alpha}/(1 - \theta_{s,m,\tilde{n};\alpha})$ with $s = \min\{p,c\}$, $m = \frac{1}{2}(|p - c| - 1)$, and $\tilde{n} = \frac{1}{2}(n - p - c - 2)$. Hence, an elimination of redundant biological and chemical variables is possible; the *Masca* program (*137*) eliminates successively the nonsignificant regressor variables. Then, the working procedure begins once more, and the final regression matrix consists of elements that are *at least* significantly different from zero at the 5% level. In contrast to the univariate (multiple) regression analysis, where the *t* test is used to examine the role of regression coefficients (*22, 189*), the multivariate tests just mentioned are based on the simultaneous test theory. In addition, we must remember that the forward and backward stepwise methods in univariate regression (mostly applied in structure–activity studies) are usually based on the highest multiple correlation coefficient R_{c-k} where *Y* is correlated against $X_1, \ldots, X_c$ chemical variables ($k \leq c$ is the number of the remaining or discarded variables). These two methods may not give the same answer. And if they do, the answer may not be the optimum, and the stepwise methods suspect. Also, these statistics reflect how well the fitted model estimates the observed sample, but not necessarily the adequacy of prediction or the validity of the model for a population of values. Finally, the total number of all possible regressions is very high. This implies a very great type I error, experimentwise, because of the lack of simultaneous test statistics. And, although the researcher frequently finds that one or more *t* tests will yield "significant" results, he may still worry that one or more such results might obtain by chance because of the number of hypotheses tested.

Corollary: From both a theoretical and a practical viewpoint, it seems that multivariate structure–activity relationships give more insight into the mechanistic interpretation and prediction of biological events. As the test statistic previously mentioned is the generalization of the most important tests in multiple regression technique, the multivariate methods can, thus, be employed if only a single biological response is considered. If that is done, the results are, from the viewpoint of prediction, more acceptable.

D. CANONICAL CORRELATION

Another multivariate regression technique, better known in the statistical literature (*23, 54, 62, 158, 180, 186*), is the canonical correlation. One goes out from the matrices by hypothesis and error, computes the total matrix, and determines the eigenvalues and eigenvectors of

$$(\mathbf{S}_h\mathbf{S}_t^{-1} - \lambda_i^*\mathbf{E})\mathbf{a}_i = 0; \quad i = 1, \ldots, s$$

Then, the vector

$$\mathbf{b}_i = \frac{\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{a}_i}{(\lambda_i^*)^{\frac{1}{2}}}$$

is determined. All the correlations between the sets of original variables have been channeled through the s correlations, and the derived variables $\mathbf{a}_i$, $\mathbf{b}_i$ are uncorrelated. If the eigenvectors corresponding to the largest eigenvalue $\max\{\lambda_i^*\} = \lambda_1^*$ are used, the maximal correlation between X^* and Y^* is described by

$$Y^* = \mathbf{a}_1'\mathbf{X}, \quad X^* = \mathbf{b}_1'\mathbf{Y}$$

Note that the simple correlation coefficient between both the variables Y^* and X^* is identical with the multivariate correlation coefficient R_c. Under special circumstances, $\mathbf{a}_1$ and $\mathbf{b}_1$ are the "best" vectors to transform dependent biological and dependent chemical variables into independent biological and independent chemical variables with a high correlation between Y^* and X^*.

E. ADVANTAGES OF AVOIDING MULTICOLLINEARITIES

It is well known that linear dependence of a set of linear equations leads to a vanishing of the coefficient determinant. In the multiple and multivariate regression, one seldom obtains a zero coefficient determinant because, in general there is no linear dependence of *all* variables. However, another problem is related to that phenomenon, the so-called multicollinearity of regressor variables. Multicollinearities occur when some of the regressor variables are themselves well correlated or "explained" by other regressor variables (nonpredictive multicollinearity). The second case is unknown, in general. The predictive multicollinearity occurs when the predicting variable (for instance, the vector $\mathbf{Y}$ in the model $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$) is also involved.

Although the *Sp* test shows that any variables involved in multicollinearities tend to have small test statistics, in general, many variables are likely to be selected because of the multicollinearity and not a zero regression coefficient. However, there are situations in which multicollinearities

lead to overestimations in the result of regression equations, and in which the multivariate and multiple correlation coefficients are too large compared with the true situation. In both cases, some of the variables involved in the best fitted model are different from those obtained by the *Sp* test.

The multicollinearity can be identified by using

$$D_i = 1 - 1/r^{ii}$$

(internal coefficient of determination) where r^{ii} represents the ith diagonal element of the correlation matrix $\mathbf{R}_{xx}^{-1}$ (model $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$) [resp. $\mathbf{R}_{yy}^{-1}$ (model $\mathbf{X} = \mathbf{B}_y\mathbf{Y}$)]. The test statistic is the same as in the multiple correlation.

In contrast to its widespread meaning in literature, the correlation matrix $\mathbf{R}_{xx}$ (resp. $\mathbf{R}_{yy}$) cannot be used in order to identify the variable which may be involved in multicollinearities. A more informative method of examining is to determine the eigenvalues and eigenvectors of $\mathbf{R}_{xx}$ (resp. $\mathbf{R}_{yy}$). This implies that a principal component analysis must be done. For instance, assume that $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ should be examined and obtain

$$(\mathbf{R}_{xx} - \lambda_i\mathbf{E})\mathbf{v}_i = 0$$

where $0 \leq \lambda_i \leq c$, $\text{tr}(\mathbf{R}_{xx}) = \sum_{i=1}^{c} \lambda_i = c$ = number of regressor variables.

Eigenvalues of or near zero indicate a linear relationship among the regressor variables, and large elements of the corresponding eigenvector indicate those particular variables that most influence the multicollinearity.

The predictive multicollinearity involves the regressor and predicting variable. The *Princo* is going out from **A**, **A** is the $(c,p = 1)$ correlation matrix $\mathbf{R}_{y_t x}$ with $t = 1, \ldots, p$. The eigenvalues and eigenvectors of

$$(\mathbf{A} - \lambda_i\mathbf{E})\mathbf{v}_i = 0$$

are interpreted as before.

Corollary: In general, neither a high multivariate and multiple correlation (for the best fitted model), nor the variable selection technique by using the *Sp* test, nor simple inspection of the correlation matrix of regressor variables (in general, $\mathbf{R}_{xx}$) will solve the problem of the best prediction, since some multicollinear variables exist. Thus, recognition of multicollinearities is an important problem in quantitative structure–activity relationships.

There are several possibilities for avoiding the problem of multicollinearities. First, the original variables are transformed into independent variables (**F** matrix) by using the *Princo* (see Section III). Second, a canon-

ical correlation analysis is done. Third, the usual multivariate regression analysis is computed until the analysis is finished (all regression coefficients are significant at the 0.05 level or less), and the selected variables are compared with those that are eliminated by the *Princo* (near zero or zero eigenvalues, large elements of the corresponding eigenvectors). The test of significance of eigenvalues (Section II,D) shows those eigenvalues that tend statistically to zero. The third method is the *Minpar* technique (Scheme 1) which is discussed in the next section.

F. *Minpar* TECHNIQUE

One begins with the assumption that substituents should be predicted without any experimental information, that is, that substituents must be selected that possess a well-spread, uncorrelated set of chemical parameter values and an ease of synthesis. The latter requirement is related to the *Synart* technique (Scheme 1) and is not a matter of concern here. The former deals with the minimizing interparameter correlation of physicochemical data bank (*Minpar* technique, Scheme 1). Although it is impossible to be complete here, it should be noted that the *Princo* serves as an important guide for the synthesis of derivatives. The analysis is based on the Q-type correlation of substituents. In such analyses, the substituents replace the variables, and the correlations between pairs of substituents are computed across the chemical variables. In most other aspects, the obverse *Princo* is mechanically just like any other *Princo*.

G. OPTIMIZATION

There are many optimization methods. With respect to the regression analysis, the response surface method is of interest. For all Y's we are going out from the third-order model

$$\begin{aligned} Y = {} & b_0 + b_1X_1 + b_2X_2 + b_3X_3 \\ & + b_{11}X_1^2 + b_{22}X_2^2 + b_{33}X_3^2 \\ & + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 \end{aligned}$$

It is clear that Y is any biological response. However, the X_i variables (i = 1,2,3) are (1) continuous variables (significant physicochemical variables) and (2) discrete variables (in sense of the response surface method). The first method is identical with usual regression analysis. The second technique is based on the Box-Wilson regression analysis (*11*). In these composite designs, all terms are orthogonal to one another, so that the regression coefficients are estimated by the ratio of the sum of products of

X_i and Y to the squares of X_i. The analysis is going out from center points, low levels, and high levels of Y. First, we require the direction of the steepest ascent. The solution of

$$\frac{\partial Y}{\partial X_1} = b_1 + 2b_{11}X_1 + b_{12}X_2 + b_{13}X_3 = 0$$

$$\frac{\partial Y}{\partial X_2} = b_2 + 2b_{22}X_2 + b_{23}X_3 + b_{12}X_1 = 0$$

$$\frac{\partial Y}{\partial X_3} = b_3 + 2b_{33}X_3 + b_{13}X_1 + b_{23}X_2 = 0$$

leads to the stationary points X_{1S}, X_{2S}, X_{3S}, written as gradient, grad $\mathbf{Y} = (X_{1S} \quad X_{2S} \quad X_{3S})'$. By a suitable change of the origin, it will now be possible to eliminate the linear terms, $X_i^* = X_i - X_{iS}$ $(i = 1,2,3)$. Replacing X_i by X_i^* gives

$$Y = b_0^* + b_{11}^*(X_1^*)^2 + b_{22}^*(X_2^*)^2 + b_{33}^*(X_3^*)^2 + b_{12}^*(X_1^*X_2^*) + b_{13}^*(X_1^*X_3^*) + b_{23}^*(X_2^*X_3^*)$$

where $b_0^* = Y_S$ is the predicted response at the stationary points. The model is simplified by elimination of the cross products. This is done by rotating the axes of the response surface about its new origin. This canonical form is obtained by calculation with the eigenvalues and eigenvectors of $\mathbf{B}^*$ equal to

$$\begin{bmatrix} b_{11}^* & \frac{1}{2}b_{12}^* & \frac{1}{2}b_{13}^* \\ \frac{1}{2}b_{12}^* & b_{22}^* & \frac{1}{2}b_{23}^* \\ \frac{1}{2}b_{13}^* & \frac{1}{2}b_{23}^* & b_{33}^* \end{bmatrix}$$

that is, a *Princo* must be done, $(\mathbf{B}^* - \lambda_i\mathbf{E})\mathbf{v}_i = 0$ $(i = 1,2,3)$. The fitted model becomes

$$Y = b_0^* + \lambda_1 Z_1^2 + \lambda_2 Z_2^2 + \lambda_3 Z_3^2$$

where the new axes are given by

$$Z_i = v_{i1}X_1^* + v_{i2}X_2^* + v_{i3}X_3^*$$

Using the canonical form of the model, we see that a maximum response is achieved when all eigenvalues and eigenvectors are positive, that is, $\forall\ \lambda_i > 0$, $\forall\ v_{ij} > 0$ $(i,j = 1,2,3)$. A minimum response is obtained if $\forall\ \lambda_i < 0$, and all v_{ij} have the same sign. A saddle point is obtained for all other cases. Thus, a series of small experiments may be necessary in order to move toward the optimum region, and the optimum operating levels of the physicochemical constants can be reached in a very short time.

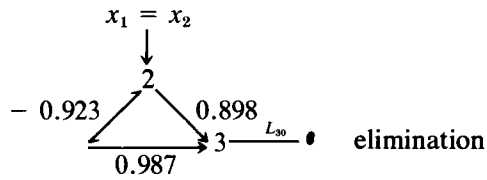
H. INVESTIGATION OF INACTIVE, SLIGHTLY ACTIVE, AND ACTIVE COMPOUNDS

In order to include inactive compounds in structure–activity studies, two possibilities exist. First, assume that zero characterizes the biological response of an inactive compound. Then, $\ln(Y + a)$ is inserted into a regression analysis where Y is the measurement, $a = 1 \vee 2 \vee 3 \vee \ldots (a > 0)$. Second, the *Discra* can be used where the treatment groups (Table I) represent the inactive groups (including the control group), the slightly active group, and different degrees of active groups. However, the statistical condition is that the amount q of groups is larger than the number of physicochemical constants c. If this requirement cannot be satisfied, a considerable bias must be expected.

I. PATH ANALYSIS AND SYSTEM THEORY

The NNN system theory (*123–132*) can be considered as a generalization of the principles of path analysis (*50b, 62a*). Application of NNN system theory to drug design and structure–activity studies yields the result that the coupling coefficients of systems can be interpreted as path coefficients, or, in a statistical sense, as partial correlation coefficients (resp. regression coefficients). Without loss of generality, the methods of interpretation of path coefficients provide an intuitively more appealing orientation in structure–activity relationships, and the diagrammatic representation allows a better mechanistic and statistical interpretation.

For instance, consider a very simple triangular diagram under stationary conditions,



where $Y = X_3 = 3$ represents the biological response, $X_1 = 1$ and $X_2 = 2$ symbolize the electronic and lipophilic properties of $n = 8$ drugs. The system theory leads to the coupling matrix **L** containing the transition or path coefficients (*123*). In general, it follows that

$$Y = \sum_{i=1}^{3} c_{3i} \exp(\lambda_i t)$$

where $\det(\mathbf{L} - \lambda\mathbf{E}) = 0$, $(\mathbf{L} - \lambda_i\mathbf{E})\mathbf{c}_i = 0$ $(i = 1,2,3)$, and t is the time of drug response after administration $(x_1 = x_2)$. Assuming stationary conditions,

the estimated correlation between Y and $\mathbf{X}' = (X_1X_2)$ is $\mathbf{r}(Y,\mathbf{X}) = (0.934 - 0.168)'$, and the correlation matrix of regressors is $\mathbf{R}_{xx}$ equal to

$$\begin{bmatrix} 1 & -0.482 \\ & 1 \end{bmatrix}$$

The path situations are computed by multivariate analysis (partial correlation): The path coefficient L_{23} is obtained if the influence of X_1 on X_2 is eliminated, $L_{23}(X_1,X_3 = Y|X_2) = 0.987$, and L_{13} is obtained if the influence of X_2 on X_1 is eliminated, $L_{13}(X_2,X_3 = Y|X_1) = 0.898$ (both are statistically significant at the 0.01 level). However, the overall test on dependences among regressor variables is significant if

$$TS = -\left(n - 1 - \frac{2c + 5}{6}\right) \ln \det(\mathbf{R}_{xx}) \geq \chi^2_{f;\alpha}$$

with $f = \frac{1}{2}(c^2 - c)$. It follows that TS is nonsignificant and, therefore, the test on nonpredictive multicollinearity is also nonsignificant. However, the predictive multicollinearity is highly significant and we get $L_{12} = L_{21}(X_1,X_2|Y = X_3) = -0.923$ (0.01 level). Thus, X_2 correlates exactly with the part of X_1 that does not correlate with Y. Under real conditions, it is quite clear that Y depends on the time, and that the path coefficients are also time-dependent. Although $\mathbf{R}_{xx}$ alone is constant, the predictive multicollinearity is changed. Therefore, it must be expected that the significance of physicochemical variables depends also on the time point of biological measurement.

J. Multicategory (Dummy) Variables

If no ordinal scale of chemical constants is available, multicategory (dummy) variables are included in regression analyses. The assignment of score values has certain advantages from the viewpoint of interpretation of regression coefficients, and it is quite legitimate to include a combination of continuous variables (X_E = a set of electronic constants, X_H = lipophilic constants, X_B = bonding parameters such as bonding distance, X_S = steric variable) and discrete variables (X_D). For instance, the stereoisomer variable ϵ is used where $\epsilon = +1$ represents an R(+) series and $\epsilon = -1$ represents an S(−) series. Therefore, the role of Pfeifer's rule in a given series can be statistically examined (the corresponding regression coefficient must be significantly different from zero). Second, hydrogen-acceptor substituents (bearing a free electron pair, for instance, CN, CH_2NMe_2, CHO, COMe, COEt, NMe_2, NO_2, OMe, OEt, SO_2Me) are expressed by $n_H = 1$, while the other substituents are denoted by $n_H = 0$. Thus, the importance of hydrogen bonding effects can be investigated.

Third, the Free-Wilson analysis can be done where all continuous variables are neglected, or a combined analysis (continuous variables and discrete variables of substitution effects) can be undertaken.

K. MULTIVARIATE NONCIRCULAR AUTOCORRELATION AND AUTOREGRESSION

The multivariate autocorrelation and autoregression analysis (*Mara*) is based on the noncircular definition (*79, 97, 99*). As in the circular case, the autocorrelation and autoregression are both completely dependent on random variables, a property that does not hold in the noncircular case, consequently, we use the noncircular definition. For example, we have investigated the effects of amino acids on avoidance response and antinociceptive reactions (*68, 73, 74*), and the influence of these compounds on oscillatory properties of biological events can be better explained by using the *Mara* (*63–66, 72*). Thus, we have first proposed that oscillatory properties should be included in structure–activity relationships and in drug design (*107, 111, 114*). The background of autoregressive oscillations is based on the hysteresis of biological phenomena (*70, 90, 91, 113, 120–128, 131–133*).

If the duration of the transmission of any action cannot be neglected, time lags become important. The noncircular serial correlation coefficient $r(k)$ of order k estimates statistically the noncircular autocorrelation

$$r(k) = \frac{\operatorname{cov}(X_l, X_{l+k})}{[\operatorname{var}(X_l) \cdot \operatorname{var}(X_{l+k})]^{1/2}}$$

where $l = 0, \ldots, m$ is the number of time translations, and

$$\operatorname{cov}(X_l, X_{l+k}) = (m-k)^{-1} \sum_{l=0}^{m-k} (X_l, X_{l+k}) - (m-k)^{-2} \left(\sum_{l=0}^{m-k} X_l \right) \left(\sum_{l=0}^{m-k} X_{l+k} \right)$$

$$\operatorname{var}(X_l) = (m-k)^{-1} \sum_{l=0}^{m-k} X_l^2 - (m-k)^{-2} \sum_{l=0}^{m-k} (X_l)^2$$

$$\operatorname{var}(X_{l+k}) = (m-k)^{-1} \sum_{l=0}^{m-k} X_{l+k}^2 - (m-k)^{-2} \sum_{l=0}^{m-k} (X_{l+k})^2$$

The null hypothesis is rejected if $r(k) \geq \varphi(k)$ where $\varphi(k)$ is the product moment correlation coefficient of a population at a given significance level and $f = m\text{-}k$ degrees of freedom. The examination of the order k of an autoregressive process $AR(k)$ is based on the significance of $r(k)$. Now, assume an $AR(1)$. The autoregressive periodicity can be computed by

$$\nu = 2\pi/(\arccos \alpha_1^2/2\alpha_2), \quad \pi = 3.14159 \ldots$$

where

$$\alpha_1 = \frac{r(1) - [1 - r(2)]}{1 - r^2(1)}, \quad \alpha_2 = \frac{r(2) - r^2(1)}{1 - r^2(1)}$$

The appearance of lag periods in kinetic curves of allosteric enzymes was experimentally observed, and time translations of potentials of electrocardiograms after drug administration are well-known. Introduction of methods of time series analysis in structure–activity models and drug design could give insight into problems of spatiotemporal organization of biosystems (*122–128, 131–133*). The spectral density function (Fourier cosine transformation) of an $AR(1)$ is defined by

$$\Gamma(\beta) = \frac{1 - r^2(1)}{1 - 2r(1)\cos\beta + r^2(1)}$$

where $\beta = 0, \ldots, 2\pi$. Assume an $AR(2)$, it follows that

$$\Gamma(\beta) = \frac{(1 - \alpha_2)(1 - \alpha_1^2 + \alpha_2^2 + 2\alpha_2)}{(1 + \alpha_2)(1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_2 + y)}$$

with $y = 2\alpha_1(1 + \alpha_2)\cos\beta + 4\alpha_2\cos^2\beta$. It should be remembered that phase spectra belonging to the coherence give information about the temporal character of rhythmic coordination and coupling processes. Physiological examples (as neuronal oscillations) are well-known (*73, 74, 99*).

L. COROLLARIES

The design of a multivariate structure–activity analysis implies the analysis of a set of physicochemical constants (*115, 116, 135*) and a set of biological responses. There are many possibilities for defining the biological variables (*106–111, 117, 118, 136a,b*). Examples show that a multivariate consideration is more realistic than the usual multiple Hansch analysis (*36, 36a*). We can analyze, for example, (1) the time dependence of optimal values of chemical constants and structure–activity equations; (2) the discrimination of toxicity and biological responses that are of medical interest (analgesia, antiinflammatory action, etc.); (3) the discrimination of biological responses (for instance, the aim is the development of analgesics without tranquilizing and neuroleptic activity); (4) the selectivity in drug metabolism and ligand–macromolelule bonding of diverse basic molecules; (5) autoregressive periodicities in structure–activity relationships (for instance, the dependence of enzymic rhythms on metabolic inhibitors); (6) the prediction of the drug stability; and (7) biochemical reaction mechanisms and their relationship to pharmacological events.

In connection with other methods of the *Masca* model mentioned in the multivariate bioassay, the following aims can be investigated: (1) discrimination of the overall biological response into pharmacokinetic and pharmacodynamic properties; (2) correlation of unbiased biological prop-

erties (obtained by *Macova*) against chemical constants; (3) investigation of common structural parameters of diverse basic molecules of compounds with similiar biological responses (rigid, semirigid, and flexible analgesics); (4) calculation of optimal values of chemical constants with respect to the desired biological response (low toxicity, rapid onset and long duration of action, high activity); (5) transformation of dependent chemical constants (problem of multicollinearity) into orthogonal constants) and (6) prediction of highly active, and inactive compounds; and so on. The other design blocks of the *Masca* model (Scheme 1) must be also included in the study, especially, the system theoretical design and the biochemical–pharmacological design; that, however, is not a matter of concern here (the reviews are in print).

It must again be emphasized that only an integration of all methods leads to practical results, and it must also be emphasized that the proper pharmacochemistry begins with the selection of appropriate substituent groups with a wide range of chemical constants before any synthesis is done. By careful selection of substituents, one can avoid misleading assumptions resulting from inadequate ranges of chemical properties. In addition, it is necessary to make compromises between the desirability of minimizing parameter correlations and the ease of compound synthesis (formalisms: *Minpar*, linear, and nonlinear programming). The selection of the first compounds occurs in the following way: At least two compounds are lying in the center points of activity, at least two compounds with a low activity, at least two compounds with a high activity. A Box-Wilson regression analysis yields a larger sample size, and the repeated procedures tend selfconsistently towards an optimum. Then, the other techniques are included.

VI. Numerical Examples

A. DOSE–TIME-RESPONSE ANALYSIS

Assume that the populations have multivariate normal distribution. Statistical independence of the $p = 4$ time-dependent measurements (at $t = 10, 20, 40, 80$ min) is not assumed. Four treatment groups were compared with a control group (the second group in Table II), hence, $q = 5$ groups are analyzed. The effective doses (ED_5, ED_{10}, ED_{20}, ED_{40}, ED_{80}) are administered ($m = 5$) and the analgesic activity measured ($n = 10$ mice per treatment group and dose). Table III shows the mean vectors and Table IV shows the matrices due to hypotheses and the error matrix.

We obtain (see p. 197) $f_1 = m - 1 = 4, f_2 = q - 1 = 4, f_3 = 16, s_1 = \min\{f_1,p\} = 4, s_2 = \min\{f_2,p\} = 4, s_3 = \min\{f_3,p\} = 4, n_e = mq(n - 1) = 1225, \tilde{n} = \frac{1}{2}(n_e - p - 1) = 110, m_1 = m_2 = -0.5, m_3 = 5.5$. First, we

TABLE II

NUMERICAL EXAMPLE OF A DOSE–TIME-RESPONSE ANALYSIS

		Treatment group 1				Treatment group 2				Treatment group 3				Treatment group 4				Treatment group 5			
Dose	Mice	10[a]	20	40	80	10	20	40	80	10	20	40	80	10	20	40	80	10	20	40	80
ED_5	1	5.5	5.7	17.7	21.2	1.2	1.8	0.9	0.8	5.6	5.8	19.8	21.1	4.2	2.8	12.1	20.1	10.2	22.4	35.2	44.8
	2	5.5	5.8	17.7	21.4	0.8	2.1	0.7	0.7	5.7	5.9	10.1	21.4	3.2	5.9	13.4	19.1	14.1	25.8	44.1	42.1
	3	9.4	7.3	19.8	18.9	0.2	0.4	0.4	0.8	5.8	7.4	17.7	20.0	4.2	4.3	8.7	15.7	18.7	27.7	35.1	38.7
	4	7.4	9.2	10.1	19.5	0.4	0.4	1.2	1.2	6.2	10.0	10.1	18.9	3.8	2.7	9.7	26.7	24.0	30.1	35.8	35.7
	5	6.5	6.9	15.0	20.0	1.8	1.8	1.3	1.4	7.1	6.7	15.2	20.1	1.8	5.8	13.2	27.1	25.3	35.2	37.2	38.1
	6	6.8	8.4	16.2	25.4	1.9	2.7	1.7	0.9	7.8	6.8	17.2	20.2	2.3	5.4	13.8	25.2	18.7	35.7	40.0	40.0
	7	8.2	6.7	15.8	27.1	2.7	0.9	2.1	2.1	8.8	8.4	14.3	26.1	4.2	5.6	12.7	30.7	19.8	36.2	27.1	38.7
	8	8.3	6.7	15.9	25.4	2.4	1.5	0.8	2.0	8.9	6.7	16.3	26.4	0.9	3.9	9.8	19.5	20.5	38.1	35.1	35.3
	9	6.5	5.8	17.1	25.1	0.7	1.6	0.7	1.8	6.5	5.8	15.3	27.1	2.4	3.7	9.7	18.7	12.4	24.2	24.1	28.2
	10	6.6	7.1	14.3	25.1	0.4	0.7	0.8	0.4	6.7	7.4	17.1	27.7	0.9	7.4	4.3	19.8	18.9	24.3	24.8	29.1
ED_{10}	1	7.5	9.8	22.4	30.1	0.4	0.6	1.3	0.4	7.5	7.8	20.1	33.2	0.7	8.7	17.7	30.1	20.8	35.7	47.1	48.2
	2	8.5	10.1	23.1	32.1	0.8	0.9	1.7	0.7	8.2	10.2	21.8	32.2	1.8	5.7	18.3	27.2	25.7	36.8	38.4	38.7
	3	6.6	12.4	20.1	33.2	0.7	1.2	0.8	0.9	8.2	10.1	22.5	30.7	2.7	7.3	10.1	25.2	34.3	37.4	40.8	40.8
	4	10.9	13.7	19.2	29.8	0.7	1.4	0.7	1.2	9.7	13.4	23.0	29.8	0.8	9.2	19.7	29.1	37.4	38.7	45.1	45.1
	5	8.8	14.8	18.5	28.7	0.4	0.7	0.4	1.3	9.6	7.6	19.2	30.1	2.7	8.4	15.8	19.5	38.9	40.1	44.3	50.1
	6	10.9	10.7	17.1	25.1	1.9	1.4	0.8	0.5	10.1	13.0	18.5	33.2	2.7	6.7	17.1	25.2	21.4	40.8	48.7	55.2
	7	11.7	7.8	15.1	27.1	2.7	0.8	1.2	0.7	10.1	12.1	15.5	28.7	0.4	6.8	17.2	30.1	25.7	37.2	38.9	55.3
	8	12.8	15.0	17.8	30.1	2.3	2.7	1.3	0.8	10.2	14.8	17.1	25.6	1.9	9.3	14.9	25.0	18.3	38.1	42.4	47.3
	9	9.5	13.0	20.1	32.1	0.9	1.6	0.4	1.2	8.9	13.0	17.8	27.7	1.9	7.8	14.7	28.7	25.4	34.2	43.5	55.4
	10	8.3	12.1	21.9	33.2	1.8	1.7	1.7	1.4	8.7	14.8	19.1	30.2	2.9	5.8	17.8	29.8	27.1	37.2	44.1	56.1
ED_{20}	1	10.1	25.0	43.1	40.1	0.2	1.8	0.9	0.3	10.2	26.1	29.7	40.1	3.4	7.2	18.2	25.1	38.0	40.1	50.1	60.2
	2	12.4	17.1	32.2	42.2	0.8	1.7	0.4	0.5	12.4	17.8	35.1	45.2	7.2	7.8	21.3	37.1	38.4	47.8	57.8	65.7
	3	13.1	14.2	26.7	45.3	1.2	0.4	1.7	0.7	17.1	19.7	33.2	45.3	1.8	7.5	23.4	25.3	42.7	47.9	60.2	68.9

	4	12.4	19.1	34.7	45.7	0.7	0.5	2.1	1.7	16.2	24.1	29.8	42.7	2.9	8.9	25.7	27.1	43.0	50.1	65.1	70.0
	5	17.8	24.2	35.2	42.3	0.7	1.7	0.7	2.1	17.0	25.1	34.7	50.0	4.3	8.7	19.7	20.8	44.7	41.2	57.2	71.0
	6	15.1	25.3	32.1	50.1	0.7	0.7	0.8	0.8	17.8	24.2	36.8	47.2	2.7	7.5	18.3	25.4	37.0	45.7	58.1	72.0
	7	16.2	27.1	29.8	50.2	0.4	0.5	0.7	0.7	16.2	27.2	24.3	42.3	2.9	9.2	17.7	30.1	38.5	44.8	60.2	71.5
	8	17.1	28.2	35.7	47.1	1.9	0.4	0.9	0.5	13.2	24.1	34.7	43.8	3.4	4.8	15.1	20.1	38.7	47.7	55.1	65.7
	9	16.2	24.1	36.8	42.5	2.4	0.9	1.3	0.7	14.1	15.8	32.2	45.7	3.5	7.9	21.0	27.1	38.8	47.2	58.9	65.7
	10	17.4	25.1	27.9	43.8	1.8	0.7	1.2	0.2	17.1	17.1	29.7	44.8	2.7	12.0	22.0	31.7	42.1	48.2	59.1	64.3
ED_{40}	1	17.4	30.1	44.2	50.1	0.5	0.7	1.2	1.0	21.4	30.2	45.7	48.9	3.4	12.3	25.7	30.8	44.3	50.1	60.2	70.1
	2	19.1	32.0	45.1	42.2	0.7	0.5	1.3	1.1	24.3	32.0	45.8	42.7	5.1	15.7	30.1	34.7	45.2	48.7	65.3	75.2
	3	19.3	31.0	45.2	52.3	0.9	1.2	1.4	1.2	25.3	24.8	42.3	54.4	4.8	12.7	30.2	35.2	44.2	45.7	64.1	79.7
	4	21.4	31.0	42.3	54.4	1.2	1.4	0.8	0.9	24.1	24.7	38.9	50.4	7.9	15.7	34.0	40.7	38.7	55.2	57.2	79.9
	5	21.5	24.0	44.2	64.1	1.4	1.7	0.7	0.4	42.2	30.7	37.4	55.7	5.4	14.8	28.7	45.7	47.1	48.7	58.2	75.0
	6	24.3	28.7	43.2	63.2	1.9	0.4	0.8	1.8	19.8	35.1	44.7	55.8	7.9	14.8	29.6	35.7	48.2	54.0	65.1	80.0
	7	25.1	35.2	44.1	62.1	1.5	0.8	1.3	2.1	19.3	34.2	45.8	45.7	7.8	17.1	31.2	41.2	48.3	57.2	67.8	81.0
	8	24.1	43.3	38.9	64.1	0.3	0.7	1.4	1.2	24.1	27.9	45.7	61.0	4.9	14.8	32.7	37.8	47.1	57.2	68.9	81.1
	9	24.2	27.8	45.8	55.1	0.4	1.6	0.4	1.3	24.3	24.2	38.7	62.8	5.0	17.1	34.1	34.9	50.0	58.1	69.2	75.2
	10	19.8	29.1	46.7	55.4	1.2	1.7	1.7	0.4	21.5	25.7	37.7	50.7	5.2	14.1	30.1	40.7	47.3	55.8	64.1	75.2
ED_{80}	1	19.1	35.2	44.2	60.1	2.3	0.6	1.7	0.5	19.8	40.2	44.8	60.1	8.7	10.2	35.1	41.7	60.1	75.1	80.1	95.7
	2	24.1	35.2	42.7	55.2	0.4	0.5	1.2	0.7	25.8	43.7	38.7	55.1	9.3	14.1	37.1	37.7	62.2	68.7	84.2	95.7
	3	25.8	35.1	45.8	50.2	0.7	1.7	1.5	0.8	25.1	46.7	44.7	45.2	12.0	15.2	36.1	38.7	65.2	72.2	85.2	88.7
	4	25.8	36.1	46.7	48.2	2.7	0.8	2.1	0.7	19.3	45.8	44.8	55.3	10.1	22.1	37.2	40.8	55.8	74.8	89.1	98.7
	5	27.1	27.1	45.1	50.4	2.3	0.9	0.8	0.4	19.2	44.7	47.5	52.7	10.4	14.1	34.8	37.9	56.7	75.7	80.7	100.0
	6	19.2	25.1	38.2	52.2	0.9	1.8	0.4	1.8	24.2	38.3	38.3	56.7	12.4	15.2	35.7	32.3	58.7	75.0	85.4	105.0
	7	26.1	26.7	38.3	56.7	1.8	2.7	0.7	2.1	25.3	37.3	45.1	50.2	15.0	18.8	34.8	34.8	54.0	65.2	87.2	106.0
	8	26.1	19.9	38.3	55.8	1.9	2.5	0.4	2.3	25.7	38.1	37.2	50.1	12.3	19.7	35.7	40.1	55.7	74.0	84.7	100.1
	9	19.1	20.0	27.2	54.7	0.4	1.7	0.5	2.1	19.8	37.2	35.2	54.7	13.7	20.1	27.8	40.2	48.0	75.4	85.7	98.9
	10	19.2	21.7	38.3	53.8	0.5	2.1	1.2	1.4	20.1	40.2	34.1	54.8	12.7	21.2	29.7	42.1	55.7	76.1	88.9	105.7

[a] This horizontal row is timed, in minutes.

TABLE III

MEAN VECTORS OF TABLE II

Vector	Dose	Group	Time (minutes) 10	20	40	80
$\bar{m}_{ij}$	ED_5	1	7.07	6.86	15.97	22.91
		2	1.25	1.45	1.06	1.21
		3	6.91	7.09	15.31	22.90
		4	2.79	4.85	10.74	22.30
		5	18.26	29.97	33.86	37.07
	ED_{10}	1	9.56	11.94	19.53	30.16
		2	1.26	1.30	1.03	0.91
		3	9.12	11.68	19.46	30.14
		4	1.85	7.57	16.33	27.00
		5	27.50	37.62	43.33	49.23
	ED_{20}	1	14.78	22.99	32.52	44.93
		2	1.08	0.93	1.07	0.82
		3	15.13	22.12	32.02	44.71
		4	3.48	8.15	20.24	26.98
		5	40.19	46.07	58.18	67.50
	ED_{40}	1	21.62	30.32	43.97	56.30
		2	1.04	1.07	1.10	1.14
		3	24.63	28.95	42.27	52.81
		4	5.75	14.91	30.64	37.74
		5	46.04	53.07	64.11	77.23
	ED_{80}	1	23.16	28.21	40.48	53.73
		2	1.39	1.53	1.05	1.28
		3	22.63	41.22	41.04	53.49
		4	11.66	17.08	34.40	38.63
		5	57.21	73.22	81.52	99.45
$\bar{m}_{i\cdot}$	ED_5		7.256[a]	10.044	15.388	21.278
	ED_{10}		9.858	14.022	19.936	27.488
	ED_{20}		14.932	20.052	28.806	36.988
	ED_{40}		19.816	25.664	36.418	45.044
	ED_{80}		23.210	32.252	39.698	49.316
$\bar{m}_{\cdot j}$		1	15.238[b]	20.064	30.494	41.606
		2	1.204	1.256	1.062	1.072
		3	15.684	22.212	30.020	40.810
		4	5.106	10.512	22.470	30.530
		5	37.840	47.990	56.200	66.096
$\bar{m}$			15.0144[c]	20.4068	28.0492	36.0228

[a] $(7.07 + 1.25 + \cdots + 18.26)/5 = 7.256$.

[b] $(7.07 + 9.56 + \cdots + 23.16)/5 = 15.238$.

[c] $(7.256 + 9.858 + \cdots + 23.210)/5 = (15.238 + 1.204 + \cdots + 37.840)/5 = 15.0144$.

TABLE IV
MATRICES OF HYPOTHESES AND ERROR, AND MATRIX OF EIGENVECTORS
V_3 OF $(B_3W^{-1} - 2_3E) = 0$

Matrix		Variables			
B_1		8850.560	11783.600	13782.790	15529.370
			15811.310	18235.850	20591.740
				21621.710	24350.420
					27463.740
B_2		40520.500	49662.850	53620.510	61400.110
			61443.460	67562.010	77996.450
				78088.260	92176.940
					110510.800
B_3		5665.637	5568.871	6148.781	8510.125
			7929.621	7234.859	9883.547
				8194.570	10410.280
					14386.840
W		1985.553	396.803	− 173.571	− 84.767
			2216.028	776.400	− 18.031
				3646.043	649.833
					3499.219
V_3		1.535124	0.992770	1.064792	− 0.909885
	10^{-2}	− 1.959683	0.433907	0.876986	− 0.576323
		0.638360	− 1.485254	0.445360	− 0.555444
		− 0.008376	0.183697	1.036364	1.363232
		$\lambda_{31} = 0.76^b$	$\lambda_{32} = 0.20^b$	$\lambda_{33} = 9.8^b$	$\lambda_{34} = 0.09^b$

[a] Note that V_1 and V_2 are without great importance because of the priority of the cell hypothesis.

[b] Rounded eigenvalues.

compute the eigenvalues of $\det(B_3W^{-1} - \lambda_3E) = 0$, because of the priority of the null hypothesis H_{03} (p. 198). It follows that $\lambda_3 = (0.7582329\ 0.1978102\ 9.7841300\ 0.08693654)'$. Hence, the three multivariate test criteria can be calculated (p. 193), and $\Lambda_3 = 0.040506 < \Lambda_{3;0.01} = 0.92043$, $V_3 = 10.827206 > V_{3;0.01} = 0.007510$, $\theta_3 = 0.907271 > \theta_{3;0.01} = 0.003950$ indicate that the null hypothesis H_{03} must be rejected at the 1% level of significance or less. Therefore, it is not necessary to calculate the eigenvalues of $\det(B_1W^{-1} - \lambda_1E) = 0$, $\det(B_2W^{-1} - \lambda_2E) = 0$. In order to demonstrate the working technique, all results are collected in Table V. The global multivariate analysis shows that at least two cell vectors are significantly different from zero at the 0.01 level. In other words, the drug effects interact with doses.

TABLE V

EIGENVALUES OF $\det(\mathbf{B}_r\mathbf{W}^{-1} - \lambda_r\mathbf{E}) = 0$ ($r = 1,2,3$ WITH $r = 1$ = ROWS, $r = 2$ = TREATMENTS, $r = 3$ = CELLS), MULTIVARIATE TEST CRITERIA AND CRITICAL VALUES (0.01 SIGNIFICANCE LEVEL), TEST ON SIGNIFICANCE OF EIGENVALUES OF $\det(\mathbf{B}_3\mathbf{W}^{-1} - \lambda_3\mathbf{E}) = 0$ (PRIORITY OF CELL HYPOTHESIS), AND CENTROIDS OF THE PRINCIPAL DISCRIMINANT ANALYSIS

	Eigenvalues					
Effect	λ_{r1}	λ_{r2}	λ_{r3}	λ_{r4}	Criterion[a]	Level
Rows	0.1412169	0.0001922	20.13392	0.0111715	$\Lambda_1 = 0.0410$	0.8704
					$V_1 = 20.2865$	0.1466
					$\theta_1 = 0.9527$	0.1000
Groups	0.0320765	2.4053540	78.97259	0.0659664	$\Lambda_2 = 0.0033$	0.8704
					$V_2 = 81.4760$	0.1466
					$\theta_2 = 0.9875$	0.1000
Cells	0.7582329	0.1978102	9.78413	0.0869365	$\Lambda_3 = 0.0451$	0.9204
					$V_3 = 10.8272$	0.0075
					$\theta_3 = 0.9075$	0.0040

Eigenvalues[b]	Λ_{3i}	TS_{3i}	f_{3i}	$\chi^2_{f_{3i};0.01}$
9.78413	0.0927280	2948.193[a]	17	33.41
0.75823	0.5687540	694.380[a]	15	30.58
0.19781	0.8348500	222.099[a]	13	27.61
0.08694	0.9200139	102.582[a]	11	24.72

	Treatment group centroids basing on v_{33} (standardized)				
Dose	1	2	3	4	5
ED_5	436.07[c]	39.45	433.14	453.55	940.00
ED_{10}	590.41	35.02	584.24	433.55	1251.93
ED_{20}	938.67	29.99	930.44	480.04	1691.15
ED_{40}	1228.86	34.31	1198.02	715.82	1926.29
ED_{80}	1181.65	41.93	1272.76	809.29	2493.88

[a] All significant at the 0.01 level.

[b] Ordered eigenvalues ($i = 1,2,3,4 = s$).

[c] 7.07(8.43554) + 6.86(7.3413) + 15.97(4.7821) + 22.91(10.8992) = 436.07039.

After the *Manova*, the *Discra* is employed. Because of the significant interaction hypothesis, it is only necessary to compute the matrix $\mathbf{V}_3$ of the eigenvectors of $(\mathbf{B}_3\mathbf{W}^{-1} - \boldsymbol{\lambda}_3\mathbf{E})\mathbf{V}_3 = 0$ (Table IV). The examination of the significance of eigenvalues (p. 194) led to the conclusion that $\lambda_{33} = 9.78413$ explains 90.37% of the total variance. The dispersion matrix $\mathbf{S} = \mathbf{W}/n_e$ is used to normalize the eigenvectors. The diagonal elements of S represent the variances of the sample, and $\mathbf{s}' = [\text{diag}(\mathbf{S})]^{\frac{1}{2}}$ is the vector of

standard deviations. We obtain

$$\mathbf{s}' = (2.9706362 \quad 3.1383137 \quad 4.0254980 \quad 3.9436130),$$

and, therefore (because $\mathbf{a}_{33}$ is the eigenvector belonging to λ_{33}),

$$\mathbf{a}_{33} = \mathbf{v}_{33}\mathbf{s}' = (0.031631 \quad 0.0275226 \quad 0.017928 \quad 0.040870)'$$

Because of $\mathbf{a}_{33}(\mathbf{a}_{33})' = 0.0037498$ we get

$$(0.00374898)^{-1}\mathbf{a}_{33} = (8.4354 \quad 7.3413 \quad 4.7821 \quad 10.8992)'$$

for the coefficients of the discriminant function of the cells. Therefore, the discriminant function

$$T(\text{cells}) = 8.4354\ Y_1 + 7.3413\ Y_2 + 4.7821\ Y_3 + 10.8992\ Y_4$$

reflects the best function for a decision procedure for assigning new drugs among the four investigated groups, and inactive new drugs can be classified into the control group. If more compounds are investigated, the relation to structure–activity analysis is obtained by using the discriminant centroids (Table V) as biological variables (*109, 117, 118*).

The relative magnitude of coefficients can be compared to determine which variable (time point) contributes most to the definition of composite function: Y_4 contains 34.65%, Y_1 contains 26.81%, Y_2 contains 23.34%, and Y_3 contains 15.20%. The examination of the significance of Y_3 is based on the following procedure: Y_3 is eliminated from the data matrix (Table II), and the same formal analysis is done. Then, the multivariate test criteria are statistically compared. If the criteria of the first and second analysis are significantly different, it must be concluded that Y_3 is not significant. Note that the relation of the corresponding test criteria must be itself considered as a test criterion. The analysis shows that Y_3 is still highly significant.

The second principal discriminant function can be computed in an analogous manner, but the eigenvalue 0.75823 shows that this function explains only 7% of the total variance. Although the eigenvalue is significant at the 0.01 level (Table V), the low contribution to the total variance indicates that this function is without large practical importance.

Now, let us examine the differences of means by using the *Sp* test. First, we compute $k_3 = 0.0688294$ (p. 199). The critical values at the 0.01 significance level are $k_3\,(1985.553)^{\frac{1}{2}} = 0.0479$ at Y_1 (10 min), $k_3\,(2216.028)^{\frac{1}{2}} = 0.2834$ at Y_2 (20 min), $k_3\,(3646.043)^{\frac{1}{2}} = 0.35912$ at Y_3 (40 min), $k_3\,(3499.219)^{\frac{1}{2}} = 0.17078$ at Y_4 (80 min). Each absolute differency of the cell means which is equal to or greater than these values is significant at the 0.01 level or less. The differences of the means (the reader should calculate differences since a very large table would be necessary to show all

differences) demonstrate that group 5 is better than all other compounds at the dose ED_5 (and all higher doses) and at the time $t = 10$ min (and all other times). In addition, the successive doses of each compound were compared, and it was found that the ED_5 and ED_{10} of compound 5 and the ED_5 and the ED_{20}, ED_5 and ED_{40}, ED_5 and ED_{80} are significantly different at all times. Similiar significances are obtained for the comparisons of ED_{10} with ED_{20}, ED_{40}, and ED_{80}.

Now, assume that a comparison of successive times is of interest (profile analysis). Therefore, the contrast matrix $\mathbf{C}'$ is

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Therefore, $\mathbf{S}_h = \mathbf{C}'\mathbf{B}_3\mathbf{C}$, $\mathbf{S}_e = \mathbf{C}'\mathbf{W}\mathbf{C}$ must be computed, we get $\mathbf{S}_h$ equal to

$$10^4 \begin{bmatrix} 2.457516 & 4.725328 & -5.712656 \\ -1.274672 & 1.654473 & -0.432978 \\ 0.287344 & -0.439780 & 1.760850 \end{bmatrix}$$

and $\mathbf{S}_e$ equal to

$$10^4 \begin{bmatrix} 17.706727 & -0.869254 & -0.883236 \\ 4.264113 & 4.309271 & -2.201778 \\ -1.682473 & -2.201778 & 5.845595 \end{bmatrix}$$

The same tests as previously mentioned lead to the conclusion that the multivariate test criteria are significant at the 0.01 level. The critical values of the *Sp* test are 24.978 (difference between Y_1 and Y_2), 13.521 (difference between Y_2 and Y_3), and 15.748 (difference between Y_3 and Y_4), and it was found that the significant multivariate result is based on the difference of Y_2 and Y_3 of compounds 1 and 4 at a dose of ED_{40} and of compound 4 at a dose of ED_{80}. However, the differences become more significant if two successive time points are considered (Y_1 against Y_3, Y_2 against Y_4).

B. Time-Dependent Structure–Activity Relationships

A common problem in drug design is to find the optimum substitution on an active lead compound in order to maximize drug potency and minimize toxicologic properties. The pharmacokinetic phase is composed of the various processes determining the concentration of the drug, primarily in the plasma and receptor compartment. The processes involved in this phase are absorption, distribution (including bonding to particular plasma

and tissue components), metabolic transformation, and excretion. As these processes depend on time, the time dependence should be included in structure–activity relationships (*103–105, 108–111, 117, 118, 136, 136a*). The pharmacodynamic response elicited by a drug is determined by its ability to act on the receptor site. Following strictly our model of drug–receptor interactions (*131*), a nonstationary and stationary component should be expected.

Example 1. The antiinflammatory activity of salicylic acid derivatives was measured (3, 4, and 5 hours after drug administration). The substituents, chemical constants, and biological potencies are given elsewhere (*136*). The final regression system is given by $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1\, Y_2\, Y_3)$ is the vector of the analgesic action at 3, 4, and 5 hours, $\mathbf{X}' = (1 X_1\, X_2 X_3)$ with $X_1 = (\sigma - \Delta R^+)$ with σ = Hammett's electronic constant, ΔR^+ = a constant for the transition state, $X_2 = \pi$ = Hansch's lipophilic substitution constant, $X_3 = \pi^2$. We obtained $\mathbf{B}_x$ equals to

$$\begin{bmatrix} 1.09210 & 1.08450 & 0 & 0 \\ 0.86007 & 0.98966 & 0 & 0 \\ 1.09610 & 0.90945 & -0.15096 & -0.17646 \end{bmatrix}$$

where the multivariate correlation coefficient (p. 205) $R_c = 0.80211$ and the multiple correlation coefficients $R_1 = 0.67903$, $R_2 = 0.64351$, $R_3 = 0.88630$ ($n = 15$ substituents) are significantly different from zero at the 0.05 level or less. It is clear that all regression coefficients are significant at the 0.05 level or less (p. 206). The analysis shows that the activity at 3, 4, and 5 hours depends on electron acceptors. In contrast, the lipophilic–hydrophilic balance is significant at 5 hours. In order to investigate that phenomenon, a *Princo* was done. We are going out from $(\mathbf{R}_{yy} - \lambda_i\mathbf{E})\mathbf{v}_i = 0$ where $\mathbf{R}_{yy}$ is the correlation matrix of analgesic responses. We obtain

$$\mathbf{v}_1 = (0.90881 \quad 0.92227 \quad 0.88001)'$$
$$\mathbf{v}_2 = (0.19217 \quad 0.26889 \quad -0.48112)'$$

for the eigenvectors. The group centroids M_1 and M_2 are computed; the first group centroid M_1 is highly correlated with all biological variables and there is a significant correlation with the electronic variable, and we obtain $r(M_1, \mathrm{X}_1) = 0.84707$ (significant at the 0.01 level). The second centroid variable is correlated with the lipophilic constant and we obtain $r(M_2, X_2) = 0.62249$ (significant at the 0.05 level). Therefore, it seems that the first component reflects the time-independent drug–receptor interactions characterized by electronic properties, and it is possible that the second component reflects the time-dependent activity (pharmacokinetic behavior) associated with the lipophilicity of drugs. As lipophilic properties play a role after 5 hours, it must be expected that metabolic processes are

predominantly involved. Note that the *Masca* model led to forecasts of new compounds (*136a*) that do not chemically belong to the category of salicylic acid derivatives (transition to other series of closely related compounds).

Example 2. The antiinflammatory activity of pyrazolones was determined at 3 and 4 hours after edema induction (*136b*). The final regression analysis yields $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1\ Y_2)$ is the biological vector, $\mathbf{X}' = (1X_1X_2)$ is the chemical vector with $X_1 = \sigma$ = Hammett's electronic constant, $X_2 = \xi = (P_i/M_v)_4$ = Mager's relative surface tension with P_i = parachor, M_v = molar volume (Table VI). We obtained $\mathbf{B}_x$ equal to

$$\begin{bmatrix} 244.92 & 0 & -104.280 \\ 177.96 & -22.342 & -62.540 \end{bmatrix}$$

($n = 9$ substituents) where $R_c = 0.93876$, $R_1 = 0.89167$, $R_2 = 0.95894$, and all regression coefficients are significant at the 0.01 level or less. Note that Hansch's lipophilic constant π did not improve the correlation (in contrast to ξ). The internal coefficient of determination (p. 208) is nonsignificant ($D = 0.06042$). The analysis shows that the antiinflammatory activity of pyrazolones depends on the relative surface tension at 3 and 4 hours while electronic interactions play a significant role at 4 hours.

The predicted values of antiinflammatory activities of compounds not included in the derivation of regression matrix led to the conclusion that the forecasts were correctly predicted (*136b*).

Example 3. A novel series of morphine-like analgesics was screened for analgesic and toxicologic activities (*182a*). In the original paper, no quantitative structure–activity study was undertaken. Our results (supplementary material available, Vol. X of Drug Design) shows that the ED_{50} values depend on lipophilic and steric properties. The first run was going out from $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1 \cdot \cdot \cdot Y_7)$ is the vector of time-dependent $\ln(ED_{50})$ values ($t = \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}$, 1, and 2 hours after administration), $\mathbf{X}' = (1X_1X_2)$ where $X_1 = \log P$ = distribution coefficient (phenol/water), $X_2 = (\log P)^2$. The regression matrix led to $\mathbf{B}_x$ equal to

$$\begin{bmatrix} 5.8633 & -5.8962 & 0.69244 \\ 5.6514 & -5.8969 & 0.86484 \\ 5.3720 & -5.9107 & 0.69230 \\ 5.2013 & -5.7046 & 0.65612 \\ 4.9852 & -5.5953 & 0.65206 \\ 5.0894 & -5.1514 & 0.57795 \\ 5.2188 & -4.5831 & 0.49709 \end{bmatrix}$$

TABLE VI[a]

FRAGMENT VALUES OF PARACHOR (P_i), MOLAR VOLUME (M_v), AND MOLAR REFRACTION (MR)[c]. CALCULATION OF MAGER'S RELATIVE SURFACE TENSION (DUE TO A SUBSTITUENT): $\xi = (P_i/M_v)^4$

Fragment	P_i	M_v	MR
Without resonance effect			
cyclic systems with 3 atoms	16.7		
cyclic systems with 4 atoms	11.6		
cyclic systems with 5 atoms	8.5		
cyclic systems with 6 atoms	6.1		
cyclic systems with 7 atoms	4.7		
With resonance			
cyclic systems with 5 atoms	10.2		
cyclic systems with 6 atoms	67.0		
C (aliph)	4.8	3.4	2.56
C═(aliph)		6.0	
C≡(aliph)		12.0	
C (aromat)	4.8	5.5	2.56
O═	43.2	10.9	2.21
—O— (ether)	20.0	3.6	1.64
—O— (hydroxyl)	20.0	3.6	1.53
N (primary)	12.5	17.3	2.32
N (secondary)	12.5	17.3	2.50
N (tertiary)	12.5	17.3	2.84
N (imide)	35.7	17.3	3.78
A═B	23.2		1.73
A≡B	46.6		2.40
H	17.1	5.8	1.03
P	37.7	12.9	9.18
S	48.2	15.5	8.12
F	25.7	12.1	0.92
Cl	54.3	18.7	6.03
Br	68.0	25.4	8.88
I	81.0	25.7	13.94
Si	25.0	12.1	8.01
Ge	35.0	13.6	41.90
Se	65.5	16.5	10.49
CN	63.9	29.3	6.33

[a] Adapted from (*115, 135, 136b*).

[b] Note that M_v, P_i, and ξ are scaled by 1/100, and MR is scaled by 1/10, to make them comparable in size to other physicochemical constants.

[c] Examples. (i) 4-NHCHO: $P_i = 12.5 + 2(17.1) + 4.8 + 43.2 = 94.7$; $M_v = 17.3 + 2(5.8) + 3.4 + 10.9 = 43.2$; $MR = 2.5 + 2(1.03) + 2.56 + 2.21 = 9.33$; $\xi = 23.092175$ dyn/cm. (ii) —C_6H_4—4—OH: $P_i = 6(4.8) + 5(17.1) + 3(23.2) + 1(67) + 1(20) = 270.9$; $M_v = 6(5.5) + 5(5.8) + 3.6 = 65.6$; $MR = 6(2.56) + 5(1.03) + 3(1.73) + 1(1.53) = 27.23$; $\xi = 290.81754$ dyn/cm (values are not scaled).

with $R_c = 0.90210$, $R_1 = 0.94916$, $R_2 = 0.93870$, $R_3 = 0.93816$, $R_4 = 0.93128$, $R_5 = 0.90905$, $R_6 = 0.89021$, $R_7 = 0.82182$ (all correlation and regression coefficients are significant at the 0.05 level and, in general, at the 0.01 level). The time-dependent shift of the optimal values opt(log P) is highly interesting: The onset of action is characterized by opt(log P) = 4.26 with opt(ED_{50}) = $1.25(10^{-3})$ at $t = \frac{1}{32}$ hr, opt(log P) = 4.31 with opt(ED_{50}) = $0.87(10^{-3})$ at $t = \frac{1}{16}$ hr, opt(log P) = 4.27 with opt(ED_{50}) = $0.71(10^{-3})$ at $t = \frac{1}{8}$ hr, opt(log P) = 4.35 with opt(ED_{50}) = $0.75(10^{-3})$ at $t = \frac{1}{4}$ hr, opt(log P) = 4.29 with opt(ED_{50}) = $0.82(10^{-3})$ at $t = \frac{1}{2}$ hr. The mean value is opt(log P) = 4.30. The duration of action is described by opt(log P) = 4.46 with opt(ED_{50}) = $1.68(10^{-3})$ at $t = 1$ hr, opt(log P) = 4.77 with opt(ED_{50}) = $4.77(10^{-3})$ at 2 hr, the mean value of duration of action is opt(log P) = 5.54 (in addition, the steric cross products of two neighboring substituents are significant at the onset of action while the sum of steric constants of the two neighboring substituents is significant at the duration of action).

Example 4. The rate constants k_{12}, k_{21} were determined for *N*-alkylammonium bromides with a three-compartiment model (*61a*), and the concentration $C_i(t)$ was determined at $t = 3$, 10, and ∞ hr. The biological variable $Y_i = -\ln(C_i)$ was regressed against the variable $\xi - 46$ (in order to avoid large variances) where ξ is Mager's relative surface tension. The final regression matrix of the model $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ with $Y' = (Y_1\ Y_2\ Y_3)$, $\mathbf{X}' = (1X_1X_2)$ with $\xi^* = X_1 = \xi - 46$, $X_2 = (\xi - 46)^2$ is given by $\mathbf{B}_x$ equal to

$$\begin{bmatrix} 5.3596 & -6.8867 & 3.3867 \\ 6.6440 & -9.3623 & 3.9318 \\ 6.0642 & -7.4939 & 2.5031 \end{bmatrix}$$

where $R_c = 0.95412$, $R_1 = 0.90557$, $R_2 = 0.93445$, $R_3 = 0.9457$ are significant at the 0.05 level or less. The most important finding is that there is a shift of the optimal values of the relative surface tension of the time–concentration responses; we obtained opt(ξ^*) = 1.0167 with peak(Y_1) = 1.86 at $t = 3$ hr, opt(ξ^*) = 1.1906 with peak(Y_2) = 1.07 at $t = 10$ hr, opt(ξ^*) = 1.4969 with peak(Y_3) = 0.46 at $t = \infty$ hr. Both examples (3 and 4) show that the cut-off point of lipophilicity depends on the time-point of the measured biological response. Therefore, more attention is necessary in the interpretation of such phenomena, and time-dependent phenomena must be included in structure–activity studies.

Example 5. Let Y be the autoregressive periodicity (p. 213) obtained from the hysteresis plane and the amplitudes of rhythms in activity of phenylalanine β-hydroxylase under the influence of phenylalanine

analogs. We obtain $Y = 7.8039 + 1.1489\sigma - 5.2010\sigma^2$, where σ = Hammett's electronic constant. The results indicate the role of substitution effects in rhythmic enzyme activity based on the hysteresis of macromolecules (*111, 120–128, 131–133*).

C. DISCRIMINATION OF TOXICITY AND EFFECTIVITY

Example 6. In Example 3, the lowest ED_{50} values were calculated by polynomial regression analysis (Y_1), and the $LD_{50} = Y_2$ was determined. The first run was going out from $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1Y_2)$, $\mathbf{X}' = (1X_1X_2)$ with $X_1 = \log P$ = distribution coefficient (phenol/water), $X_2 = (\log P)^2$. We obtain $\mathbf{B}_x$ equal to

$$\begin{bmatrix} 5.1207 & -5.7725 & 0.66824 \\ 6.4391 & -1.2608 & 0 \end{bmatrix}$$

with $R_c = 0.85231$, $R_1 = 0.93445$, $R_2 = 0.79514$ (all correlation and regression coefficients are significant at the 0.05 level or less). The optimal value of the analgesic activity is opt(log P) = 4.32, and, because this value corresponds to the mean value of the onset of action, our interpretation (onset and duration) was correct. In contrast, the toxicity shows a departure of the principle of an optimal balance between lipophilicity and hydrophilicity: the higher the lipophilicity, the higher the toxicity. The linearity establishes the fact that the compounds are bound in varying degrees to macromolecules, in the sense of a partition-like process, and that the bonding induces the toxicity. It can be assumed that intracellular macromolecules are involved in the bonding. The second run is going out from $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$, where $\mathbf{Y}$ is as already defined, $\mathbf{X}' = (1X_1X_2)$, where X_1 is the sum of steric properties and X_2 is the cross product. It was found that the analgesia depends on steric terms whereas the toxicity does not.

Example 7. The local anesthetic activity of acyloxypiperidines was measured (Y_2 = ln ED_{50}), and in addition, so was the acute toxicity (Y_1 = LD_{50}). The regression analysis (*136b*) was going out from $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1Y_2)$, $\mathbf{X}' = (1X_1X_2X_3)$ with $X_1 = \sigma_R(R_2 + R_3)$ = resonance constant due to the substituents R_2 and R_3 (see *136b*, additional material available), $X_2 = \pi(R_1 + R_2 + R_3)$, $X_3 = \xi(R_1,R_2,R_3)$ = relative surface tension. We obtain $\mathbf{B}_x$ equal to

$$\begin{bmatrix} 226.8800 & 207.9000 & 0 & 0 \\ 1.4241 & 0 & 0.91314 & 0.012305 \end{bmatrix}$$

with $R_c = 0.90021$, $R_1 = 0.87921$, $R_2 = 0.91712$ (significant at the 0.01 level). The regression coefficient b_{11} is significant at the 0.01 level, b_{22} is

significant at the 0.05 level, and b_{23} is significant at the 0.001 level. No multicollinearity (p. 208) was found and the internal coefficient of determination is $D(\pi,\xi) = 0.10202$ (orthogonality). Therefore, the toxicity depends on electronic interactions with biological macromolecules whereas the anesthetic activity depends on lipophilic properties. Since phospholipids contain ester segments with nucleophilic and nonpolar chains, it is possible that the nucleophilic site is responsible for the toxicity whereas the nonpolar chains are responsible for the activity. In order to maximize the activity and minimize the toxicity, we have predicted fluorocarbon anesthetics of this series (*136b*) with hydrogen bond-breaking properties.

D. Discrimination of Major Pharmacological Effects

Example 8. The toxicity, antihistaminic activity, antispasmodic response, and antitussive action of 2-substituted pyridinyl derivatives were determined (*147a*). The discrimination of the major effects and, in addition, of the major effects and toxicity can be seen from the regression matrix $\mathbf{B}_x$ which is

$$\begin{bmatrix} 5.5299 & 0.44479 & -0.17002 & 0 & 0 & 0 \\ -17.8720 & 4.02230 & 0 & 0 & 0 & 0 \\ -5.1847 & 0 & 0 & -3.5229 & 0.67437 & -1.0248 \\ 3.2154 & 0 & -0.69398 & 0 & 0 & 0 \end{bmatrix}$$

where $R_c = 0.90000$, $R_1 = 0.94568$, $R_2 = 0.93283$, $R_3 = 0.92384$, and $R_4 = 0.74937$ (all correlation and regression coefficients are significant at the 0.01 level). The model is $\mathbf{Y} = \mathbf{B}_x\mathbf{X}$ where $\mathbf{Y}' = (Y_1 \ldots Y_4)$, with $Y_1 = \ln \mathrm{LD}_{50}$ = acute toxicity; $Y_2 = \ln \mathrm{ED}_{50}$ = antihistaminic activity; $Y_3 = \ln \mathrm{ED}_{50}$ = antispasmodic activity; $Y_4 = \ln \mathrm{ED}_{50}$ = antitussive activity; $\mathbf{X}' = (1X_1 \ldots X_5)$, with $X_1 = \sigma_a(R_1)$ = aliphatic electronic constant due to R_1, $X_2 = \sigma_a(R_2)$ = aliphatic electronic constant due to the substituent R_2, $X_3 = \sigma_a(R_1 + R_2)$, and $X_4 = \sigma_a^2(R_1 + R_2)$, $X_5 = \pi(R_1 + R_2)$ = lipophilic substitution constant.

Therefore, it is possible to discriminate the major pharmacological effects: electron acceptors in the R_1 position improve the antihistaminic activity, electron donors in the R_2 position enhance the antitussive action in living animals, the antispasmodic activity depends both on an optimal value of electronic properties of both substituents and on lipophilic properties. In contrast to the antispasmodic activity (both substituents act additively), the toxicity depends on electronic influences that act separately.

E. ACTIVITY–STRUCTURE RELATIONSHIPS

Here, we are going out from $\mathbf{X} = \mathbf{B}_y\mathbf{Y}$ (p. 204). I should like to draw attention to the problem of multicollinearity of regressor variables. A simple example is used in order to demonstrate the working technique.

Example 9. Yeast alcohol dehydrogenase is inhibited by substituted pyridines. Table VII gives the 50% inhibition of the NAD-dependent en-

TABLE VII
EXAMPLE OF AN ACTIVITY—STRUCTURE ANALYSIS: INHIBITION OF ALCOHOL DEHYDROGENASE BY PYRIDINIUM ION DERIVATIVES[a]

Compd	R	X	X_t	Y	Y_t
1	4-Me	− 0.17	− 0.10	− 4.34	− 3.75
2	3-Me	− 0.06	− 0.10	− 4.34	− 4.90
3	H	0	− 0.05	− 5.02	− 5.54
4	3-$CONH_2$	0.28	0.22	− 8.15	− 8.49
5	3-COMe	0.29	0.22	− 8.11	− 8.60
6	3-CO_2Et	0.36	0.43	− 10.56	−9.34
7	3-CHO	0.44	0.48	− 11.11	− 10.18
8	3-CN	0.66	0.53	− 11.68	− 12.50

[a] X (electronic constant = σ) and Y were obtained from the literature (*114, 118, 135*). X_t (from activity–structure analysis) and Y_t (from structure–activity analysis) are theoretically calculated values.

zyme (at mM concentrations of the metabolic inhibitor) in terms of $Y = \ln I_{50}$ (pyridinium ion derivatives). We obtained $Y = -5.5414 - 10.5410\,\sigma$ (the multiple correlation coefficient $R = 0.96632$ is significant at the 0.01 level). The model $\mathbf{X} = \mathbf{B}_y\mathbf{Y}$ led to $\sigma = -0.47593 - 0.08579 \ln Y$ ($R = 0.96632$). We are interested in predicting the parameters. For instance, we want to obtain an activity of $I_{50} = 10^{-10}$ mM. From $\mathbf{B}_y$, it follows that the electronic constant should be $\sigma = -0.47593 - 0.08579 \ln(10^{-10}) = 1.4995 \approx 1.50$. Table number 10f in Mager's physicochemical tables (*135*) shows that 3,3′ − $(POCl_2)_2$ has an electronic constant with a similiar value ($\sigma = 1.66$). Therefore, the predicted value of enzyme inhibition is $I_{50} = \exp[-5.5414 - 10.541(1.66)] = \exp(-23.03946) = 9.865(10^{-11}) \approx 10^{-10}$ mM.

VII. Discussion

In addition to the complexity of chemical structures and, therefore, to the necessity for including chemical multidimensionality into structure–

activity studies by using many physicochemical constants and properties (*36, 36a, 93–95, 115, 116, 136d*), it is also extremely important to include many biological variables in the derivation of structure–activity equations. In contrast to the usual multiple regression analysis (*49*), introduced by Hansch into structure–activity relationships (*36, 36a*), our multivariate viewpoint is more complex, and requires a minimal knowledge of matrix algebra (*87*). The mathematical complexity, however, is based on the complexity of structure–activity relationships.

The most important formalisms used in the *Masca* model are summarized in this chapter. In addition, methods for avoiding the bias in estimation of regression coefficients (*16*) by neglecting of conditions of application become more and more important in applied statistics. Three problems were especially discussed: (1) multicollinearity, (2) error rate per comparison (including the stepwise methods and the calculation of two or more regression equations by using the multiple regression analysis), and (3) that an overal significant result requires that the factors contributing to the overall result must be statistically examined by using the simultaneous test statistics (note that the F test in the multiple regression analysis is also an overall test). Only if these problems are not neglected is there a great probability that tests will not be falsely declared significant. The interpretation of computer outputs is the next problem and, therefore, the final results of some examples were discussed. It was not the aim of this theoretical report to give detailed examples, but it should be noted that detailed supplementary material of these same examples and, in addition, of many other examples is available, and a review is in preparation (physicochemical constants of the chemical structures, experimentally obtained and theoretically calculated biological events, forecasts, and relations to other multivariate methods). Examples dealing with benzoic and salicylic acids are available in the literature (*136a*).

Since publication of the first reports of the *Masca* model (*106–112*), reports have been published by other research groups (*28a*) also dealing with at least two biological variables in structure–activity studies. And an attempt was made to investigate these relationships by multivariate methods such as factor analysis. In contrast to these researchers, we do not believe that the factor analysis (*151, 175, 185*) provides insight into multivariate structure–activity relationships because the statistical background of factor analysis is too small and because too many subjective components influence the final results. The consequence is that the reliability of results is questionable. The spectral mapping (*58–60*) method is indeed simpler than our technique, but probability theory must be introduced in this classification technique. In addition, it appears that some other research groups have also attempted to include multivariate consid-

erations in drug design. It is hoped that the present chapter will help the drug designer to navigate the turbulent waters of statistical problems in multivariate drug design.

References

1. R. L. Anderson, *Biometrics* **2,** 41 (1946).
2. T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," p. 272. Wiley, New York, 1958.
3. J. Berkson, *In* "The Future of Statistics" (D. J. Watts, ed.), p. 175. Academic Press, New York, 1968.
4. C. I. Bliss, "Statistics in Biology," Vol. 2, p. 288. McGraw-Hill, New York, 1970.
5. R. D. Bock, "A Computer Program for Univariate and Multivariate Analysis of Variance." IBM Data Process. Div., White Plains, New York, 1963.
6. R. Bock, *In* "Problems in Measuring Change" (C. W. Harris, ed.), p. 85. Univ. of Wisconsin Press, Madison, 1963.
7. R. D. Bock, *In* "Handbook of Multivariate Experimental Psychology" (R. Cattel, ed.), p. 820. Rand-McNally, Chicago, Illinois, 1966.
8. R. D. Bock and L. V. Jones, "The Measurement and Prediction of Judgment and Choice," pp. 58, 66, 71, 132, and 134. Holden-Day, San Francisco, California, 1968.
9. R. D. Bock and E. H. Haggard, *In* "Handbook of Measurement and Accessment in Behavioral Sciences" (D. K. Whittla, ed.), p. 100. Addison-Wesley, Reading, Massachusetts, 1968.
10. P. R. Bock, B. Pollock, S. Schach, A. Fuchs, and R. Lohaus, *Arzneim.-Forsch.* **26,** 1308 (1976).
11. G. E. P. Box and K. B. Wilson, *J. Roy. Stat. Soc.* **13,** 1 (1951).

11a. G. K. Menon and S. Cammarata, *J. Pharm. Sci.* **66,** 304 (1977).

12. L. Cavalli-Sforza, "Grundbegriffe der Biometrie," p. 136. Fischer, Stuttgart, 1964.
13. I. M. Chakravarti, R. G. Laha, and J. Roy, "Handbook of Methods of Applied Statistics," Vol. 1, p. 424. Wiley, New York, 1967.
14. V. Chew, *J. Am. Stat. Assoc.* **61,** 605 (1966).
15. D. J. Clyde, E. M. Cramer, and R. J. Sherin, "Multivariate Statistical Programs." Biometrika Lab of the Univ. of Miami, Coral Gables, 1966.
16. W. G. Cochran, *In* "Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability" (L. M. LeCam, J. Neyman, and E. L. Scott, eds.), p. 527. Univ. of California Press, Berkeley, 1972.
17. J. W. Cole and J. E. Grizzle, *Biometrics* **22,** 810 (1966).
18. W. W. Cooley and P. K. Lohnes, "Multivariate Procedures for the Behavioral Sciences," p. 157. Wiley, New York, 1962.
19. W. W. Cooley and P. K. Lohnes, "Multivariate Data Analysis," p. 287. Wiley, New York, 1971.
20. R. G. Davies, "Computer Programming in Quantitative Biology," p. 291. Academic Press, New York, 1971.
21. A. P. Dempster, "Continuous Multivariate Analysis," p. 136. Addison-Wesley, Reading, Massachusetts, 1969.

21a. S. Dove and R. Franke, *Wiss. Z. Martin Luther Univ. Halle-Wittenberg* (Engl. Ed.) **42,** 216 (1977).

22. N. R. Draper and H. Smith, "Applied Regression Analysis," p. 107. Wiley, New York, 1966.
23. P. H. DuBois, "Multivariate Correlation Analysis," p. 112. Harper, New York, 1957.
24. O. J. Dunn, *Ann. Math. Stat.* **29,** 1095 (1958).
25. R. C. Elston and J. E. Grizzle, *Biometrics* **18,** 148 (1962).
26. W. T. Federer, "Experimental Design," p. 439. MacMillian, New York, 1955.
27. J. D. Finn, "Mesa 95; Univariate and Multivariate Analysis of Variance and Covariance." Dep. Educ. Stat. Lab. Res. Mem. No. 3, Univ. of Chicago Press, Chicago, Illinois, 1966.
28. D. J. Finney, "Statistical Method in Biological Assay." Griffin, London, 1971.
28a. R. Franke (and several co-workers), Lectures at the Symposium on Chemical Structure–Biological Activity Relationships, October 25–28, 1976, in Suhl (G.D.R.), Akademie-Verlag, Berlin, in press.
29. D. H. Fraser and I. Gultmann, *Ann. Math. Stat.* **27,** 162 (1956).
30. D. H. Fraser, "The Structure of Inference," p. 225. Wiley, New York, 1968.
31. S. Geisser and W. J. Greenhouse, *Ann. Math. Stat.* **29,** 885 (1958).
32. G. V. Glass and J. C. Stanley, "Statistical Methods in Education and Psychology," p. 371. Prentice-Hall, Englewood Cliffs, New Jersey, 1970.
33. W. J. Greenhouse and S. Geisser, *Psychometrika* **24,** 95 (1959).
34. J. E. Grizzle and D. M. Allan, *Biometrics* **25,** 357 (1962).
35. J. E. Grizzle, *In* "Essays in Probability and Statistics" (R. C. Bose, ed.), p. 311. Univ. of North Carolina Press, Chapel Hill, 1970.
36. C. Hansch, *In* "Drug Design" (E. J. Ariëns, ed.), Vo.. I, p. 271. Academic Press, New York, 1971.
36a. C. Hansch, *In* "Structure–Activity Relationships" (C. J. Cavallito, ed.), p. 75. Pergamon, Oxford, 1973.
36b. C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.* **16,** 1217 (1973).
37. F. W. Harpley and G. A. Stewart, *In* "Biostatistics in Pharmacology" (A. L. DeLaunois, ed.), Vol. 2, p. 971, Pergamon, Oxford, 1973.
38. M. L. Hart and A. H. Money, *Biometrika* **63,** 59 (1976).
39. D. L. Heck, *Ann. Math. Stat.* **31,** 625 (1960).
40. K. Hope, "Methods of Multivariate Analysis," p. 40. Univ. of London Press, London, 1968.
41. J. W. Hopkins and P. P. F. Clay, *J. Am. Stat. Assoc.* **58,** 625 (1963).
42. D. T. Hughes and J. J. Sano, *Biometrika* **59,** 224 (1972).
43. K. Ito and W. J. Schull, *Biometrika* **51,** 224 (1964).
44. K. Ito, *In* "Research Papers in Statistics" (F. N. David, ed.), p. 147. Wiley, New York, 1966.
45. K. Ito, *In* "Multivariate Analysis" (P. R. Krishnaiah, ed.), Vol. 2, p. 87. Academic Press, New York, 1969.
46. G. S. James, *Biometrika* **41,** 19 (1954).
47. D. R. Jenssen, *Biometrics* **28,** 39 (1972).
48. O. Kempthorne and L. Folks, "Probability, Statistics and Data Analysis," p. 112. Iowa State Univ., Ames, 1971.
49. M. G. Kendall and A. Stuart, "The Advanced Theory of Statistics," Vol. 2, p. 317. Griffin, London, 1961.
50. M. G. Kendall, "A Course in Multivariate Analysis," p. 68. Griffin, London, 1971.
50a. M. G. Kendall, in "Discriminant Analysis and Applications" (T. Cacoullos, ed.), p. 179. Academic Press, New York, 1973.

50b. M. G. Kendall and C. A. O'Muircheartaigh, "Path Analysis and Model Building," p. 17. Tech. Bull. No. 2, International Statistical Institute, Voorburg, The Hague, 1977.
51. C. G. Khatri, *Ann. Inst. Stat. Math.* (Tokyo) **18,** 75 (1966).
52. R. E. Kirk, "Experimental Design," p. 245. Brooks/Cole, Belmont, California, 1968.
53. C. G. Koch, *Biometrics* **26,** 106 (1970).
54. P. B. Koons, *In* "Computer Applications in the Behavioral Sciences" (H. Borko, ed.), p. 266. Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
55. H. Kres, "Statistische Tafeln zur multivariaten Analysis," pp. 88 and 207. Springer-Verlag, Berlin and New York, 1975.
56. A. M. Kshirsagar, "Multivariate Analysis," pp. 160 and 178. Dekker, New York, 1972.
57. Y. S. Lee, *Biometrika* **95,** 649 (1972).
58. P. J. Lewi, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. V, p. 81. Academic Press, New York, 1975.
59. P. J. Lewi, *Arzneim.-Forsch.* **26,** 1295 (1976).
60. P. J. Lewi, *Int. Pharmacopsychiatry* **11,** 181 (1976).
61. T. O. Lewis and P. L. Odell, "Estimation in Linear Models," p. 164. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.
61a. B. Lippold and G. F. Schneider, *Arzneim.-Forsch.* **25,** 843 (1975).
62. W. A. Love and K. K. Stewart, "Interpreting Canonical Correlations; Theory and Practice." Am. Inst. Res., Palo Alto, California, 1968.
62a. K. I. Macdonal, in "Path Analysis" (C. A. O'Muircheartaigh and C. D. Payne, eds.), Vol. 2, Chap. 3. Wiley, New York, 1977.
63. P. P. Mager, S. Theil, and F. Klingberg, *Acta Biol. Med. Ger.* **28,** 387 (1972).
64. P. P. Mager, S. Theil, and F. Klingberg, *Acta Biol. Med. Ger.* **28,** 507 (1972).
65. P. P. Mager, S. Theil, and F. Klingberg, *Acta Biol. Med. Ger.* **28,** 1081 (1972).
66. P. P. Mager, S. Theil, and F. Klingberg, *Acta Biol. Med. Ger.* **29,** 933 (1972).
67. P. P. Mager, *Act. Nerv. Super.* **15,** 269 (1973).
68. P. P. Mager and F. Klingberg, *Acta Biol. Med. Ger.* **31,** 889 (1973).
69. P. P. Mager, *Morphol. Jahrb.* **119,** 954 (1973).
70. P. P. Mager and R. Mager, *Morphol. Jahrb.* **120,** 529 (1974).
71. P. P. Mager, *Pharma Int., Engl. Ed.* **4,** 23 (1974).
72. P. P. Mager, *Act. Nerv. Super.* **16,** 19 (1974).
73. P. P. Mager, *Experientia* **30,** 154 (1974).
74. P. P. Mager and F. Klingberg, *Experientia* **30,** 1189 (1974).
75. P. P. Mager and S. Theil, *Biom. Z.* **16,** 49 (1974).
76. P. P. Mager and H. Mager, *Biom. Z.* **16,** 253 (1974).
77. P. P. Mager, *Zentralbl. Pharm. Pharmakother. Laboratoriumsdiag.* **113,** 717 (1974).
78. P. P. Mager, *Act. Nerv. Super.* **16,** 141 (1974).
79. P. P. Mager, *Acta Histochem.* **49,** 233 (1974).
80. P. P. Mager, *Zool. Jahrb. Anat.* **93,** 335 (1974).
81. P. P. Mager, *Morphol. Jahrb.* **120,** 127 (1974).
82. P. P. Mager, *Morphol. Jahrb.* **120,** 228 (1974).
83. P. P. Mager, *Morphol. Jahrb.* **120,** 323 (1974).
84. P. P. Mager, *Morphol. Jahrb.* **120,** 485 (1974).
85. P. P. Mager, *Morphol. Jahrb.* **120,** 613 (1974).
86. P. P. Mager, *Morphol. Jahrb.* **120,** 791 (1974).
87. P. P. Mager, *Morphol. Jahrb.* **120,** 800 (1974).
88. P. P. Mager and R. Mager, *Morphol. Jahrb.* **120,** 560 (1974).
89. R. Mager and P. P. Mager, *Morphol. Jahrb.* **120,** 620 (1974).

90. P. P. Mager, *Acta Histochem.* **53,** 302 (1975).
91. P. P. Mager, *Morphol. Jahrb.* **121,** 421 (1975).
92. P. P. Mager, *Sci. Pharm.* **43,** 100 (1975).
93. P. P. Mager, *Sci. Pharm.* **43,** 102 (1975).
94. P. P. Mager, *Sci. Pharm.* **43,** 145 (1975).
95. P. P. Mager, *Sci. Pharm.* **43,** 149 (1975).
96. P. P. Mager, *Pharm. Rundsch.* **1,** 3 (1975).
97. P. P. Mager, *Act. Nerv. Super.* **17,** 149 (1975).
98. P. P. Mager, *Zool. Jahrb. Anat.* **94,** 123 (1974).
99. P. P. Mager, *Acta Biol. Med. Ger.* **34,** 1273 (1975).
100. P. P. Mager, *Internist. Praxis* **15,** 791 (1975).
101. P. P. Mager, *Biom. Z.* **17,** 99 (1975).
102. P. P. Mager and H. Mager, *Biom. Z.* **17,** 325 (1975).
103. P. P. Mager, *Sci. Pharm.* **43,** 223 (1975).
104. P. P. Mager, *Sci. Pharm.* **43,** 234 (1975).
105. P. P. Mager, *Sci. Pharm.* **43,** 260 (1975).
106. P. P. Mager, *Arzneim.-Forsch.* **25,** 1006 (1975).
107. P. P. Mager, *Arzneim.-Forsch.* **25,** 1270 (1975).
108. P. P. Mager, *Arzneim.-Forsch.* **25,** 1355 (1975).
109. P. P. Mager, *Arzneim.-Forsch.* **25,** 1475 (1975).
110. P. P. Mager, *Arzneim.-Forsch.* **25,** 1745 (1975).
111. P. P. Mager, *Arzneim.-Forsch.* **25,** 1864 (1975).
112. P. P. Mager and R. Mager, *Morphol. Jahrb.* **121,** 340 (1975).
113. P. P. Mager, *Arzneim.-Forsch.* **26,** 26 (1976).
114. P. P. Mager, *Arzneim.-Forsch.* **26,** 185 (1976).
115. P. P. Mager, *Sci. Pharm.* **44,** 40 (1976).
116. P. P. Mager, *Sci. Pharm.* **44,** 143 (1976).
117. P. P. Mager, *Sci. Pharm.* **44,** 242 (1976).
118. P. P. Mager, *Sci. Pharm.* **44,** 336 (1976).
119. P. P. Mager, *Zool. Jahrb. Anat.* **95,** 161 (1976).
120. P. P. Mager, *Arzneim.-Forsch.* **26,** 1818 (1976).
121. P. P. Mager, *Arzneim.-Forsch.* **26,** 2164 (1976).
122. P. P. Mager, in "Die Zeit und das Leben; Chronobiologie" (J. H. Scharf and H. von Mayersbach, eds.), p. 85. Barth, Leipzig, 1977 (Nova Acta Leopoldina N.S., Vol. 46, No. 225).
123. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 422 (1977).
124. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 447 (1977).
125. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 459 (1977).
126. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 473 (1977).
127. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 489 (1977).
128. P. P. Mager, *Zool. Jahrb. Physiol.* **81,** 504 (1977).
129. P. P. Mager, *Biom. J.* **19,** 143 (1977).
130. P. P. Mager, *Biom. J.* **19,** 303 (1977).
131. P. P. Mager, *Int. Fed. Autom. Control Symp. Biocybernetics* (Leipzig) **4,** 217 (1977).
132. P. P. Mager, *Morphol. Jahrb.* **123,** 656 (1977).
133. P. P. Mager, *In* "Environmental Biophysics" (K. Unger, ed.), Vol. 2, Akademie-Verlag, Berlin, in press.
134. P. P. Mager, "Significance Points for Tests of Multivariate Hypotheses," Tech. Rep. No. 1 Masca Model, Halle, 1977, unpublished.
135. P. P. Mager, H. Mager, and A. Barth, *Sci. Pharm.* **48,** in press.

136. P. P. Mager, U. Wenzel, R. Hirschelmann, J. Metzner, R. Boehme, and E. Krause, *Wiss. Z. Martin Luther Univ. Halle-Wittenberg* (Engl. Ed.) **42,** 191 (1977).
136a. P. P. Mager and U. Wenzel, *Zentralbl. Pharm. Pharmakother. Laboratoriumsdiag.* **119,** No. 8 (1979).
136b. P. P. Mager, M. Paintz, and U. Wenzel, *Pharmazie* **34,** 175 (1979).
136c. P. P. Mager, *Internist. Praxis* **18,** 349 (1978).
136d. P. P. Mager, *Acta Histochem.* **61,** 20 (1978).
137. P. P. Mager and U. Wenzel, "Computing Scheme and Fortran Program for the Masca Model," Tech. Rep. No. 3, Halle, 1978, unpublished.
138. N. Mantel and R. S. Valand, *Biometrics* **26,** 547 (1970).
139. L. A. Marasculo, "Statistical Methods in Behavioral Science Research," p. 459. McGraw-Hill, New York, 1971.
140. K. V. Mardia, *Biometrika* **57,** 519 (1970).
141. K. V. Mardia, *Biometrika* **58,** 105 (1971).
142. K. V. Mardia, *Sankhya, Ser. B,* **36,** 115 (1974).
143. K. V. Mardia, *Appl. Stat.* **24,** 163 (1975).
144. K. V. Mardia, *Appl. Stat.* **24,** 262 (1975).
145. W. Mendenthal, "Introduction to Linear Models and the Design and Analysis of Experiments," p. 367. Wadsworth, Belmont, California, 1968.
146. T. A. Mijares, "Percentage Points of the Sum $V_1^{(s)}$ of Roots (s = 1, . . ., 50)." The Statistical Center of the Univ. of Philippines, Manila, 1964.
147. R. G. Miller, "Simultaneous Statistical Inference," p. 189. McGraw-Hill, New York, 1966.
147a. S. Miyano, A. Abe, Y. Kasé, Y. Yuizono, K. Tachibana, T. Miyata, and G. Kito, *J. Med. Chem.* **13,** 704 (1976).
148. D. F. Morrison, "Multivariate Statistical Methods," p. 151. McGraw-Hill, New York, 1967.
149. J. E. Overall and C. J. Klett, "Applied Multivariate Analysis," pp. 137, 157, and 314. McGraw-Hill, New York, 1972.
150. D. B. Owen, "Handbook of Statistical Tables," p. 237. Pergamon, Oxford, 1962.
151. D. J. Palumbo, "Statistics in Political and Behavioral Science," p. 290. Appleton, New York, 1969.
152. E. F. Pearson and H. O. Hartley, "Biometrika Tables," Vol. 2, pp. 333, 336, and 342. Cambridge Univ. Press, London and New York, 1972.
153. K. S. C. Pillai, "Statistical Tables for the Tests of Multivariate Hypotheses," pp. 14, 19, and 23. The Statistical Center of the Univ. of Philippines, Manila, 1960.
154. K. C. S. Pillai and A. K. Gupta, *Biometrika* **56,** 109 (1969).
155. R. F. Potthof and S. N. Roy, *Biometrika* **51,** 313 (1964).
156. S. J. Press, "Applied Multivariate Analysis," Chap. 8, and p. 264. Holt, New York, 1972.
157. M. L. Puri and P. K. Sen, "Nonparametric Methods in Multivariate Analysis," p. 252. Wiley, New York, 1971.
158. M. H. Quenouille, "Associated Measurements," p. 203. Butterworth, London, 1952.
159. V. W. Rahlfs and F. K. Bedall, *Int. J. Clin. Pharm.* **5,** 96 (1971).
160. C. R. Rao, *Biometrika* **46,** 49 (1959).
161. C. R. Rao, *Biometrika* **52,** 447 (1965).
162. C. R. Rao, *In* "Multivariate Analysis" (P. R. Krishnaiah, ed.), Vol. 1, p. 87. Academic Press, New York, 1966.
163. C. R. Rao, "Lineare statistische Methoden und ihre Anwendung," pp. 129 and 495. Akademie-Verlag, Berlin, 1973.

164. R. H. Regamy, "Biological Assay Methods," p. 183. Karger, Basel, 1961.
165. B. Roeder, *Biom. Z.* **16,** 287 (1974).
166. S. N. Roy and R. C. Bose, *Ann. Math. Stat.* **24,** 513 (1953).
167. S. N. Roy and S. K. Mitra, "An Introduction to Some Nonparametric Generalizations of Analysis of Variance and Multivariate Analysis." Univ. of North Carolina Press, Chapel Hill, 1955.
168. S. N. Roy, "Some Aspects of Multivariate Analysis," p. 95. Wiley, New York, 1957.
169. S. N. Roy, R. Gnanadesikan, and J. N. Srivastava, "Analysis and Design of Certain Multiresponse Experiments," p. 72. Pergamon, Oxford, 1971.
170. C. L. Rümke and H. de Jonge, *In* "Evaluation of Drug Activities; Pharmacometrics" (D. R. Laurence and A. L. Bacharach, eds.), Vol. 1, p. 47. Academic Press, New York, 1964.
171. P. J. Rulon, V. Tiedeman, M. M. Tatsuoka, and C. R. Langmuir, "Multivariate Statistics for Personal Classification," p. 265. Wiley, New York, 1967.
173. M. Schatzoff, *Biometrika* **53,** 347 (1966).
174. H. Scheffé, "The Analysis of Variance," p. 350. Wiley, New York, 1959.
175. H. Seal, "Multivariate Statistical Analysis for Biologists," p. 153. Methuen, London, 1964.
176. P. H. Sneath, *In* "Numerical Taxonomy" (A. J. Cole, ed.), p. 257. Academic Press, New York, 1969.
177. G. W. Snedecor, "Statistical Methods," Chapters 11.6, 11.7, and 11.10. Iowa State College Press, Ames, 1956.
178. R. R. Sokal and P. H. A. Sneath, "Principles of Numerical Taxonomy," p. 290. Freeman, San Francisco, 1963.
179. R. R. Sokal and F. J. Rohlf, "Biometry," Chapter 8.12. Freeman, San Francisco, 1969.
180. P. Sprent, "Methods in Regression and Related Topics," pp. 63 and 88. Methuen, London, 1969.
181. K. Subrahmaniam and K. Subrahmaniam, *Biometrika* **60,** 107 (1973).
182. M. M. Tatsuoka, "Multivariate Analysis; Techniques for Educational and Psychological Research," p. 198. Wiley, New York, 1971.
182a. W. F. M. van Bever, C. J. E. Niemegeers, K. H. L. Schellekens, and P. A. J. Janssen, *Arzneim.-Forsch.* **26,** 1548 (1976).
183. J. E. Walsh, "Handbook of Nonparametric Statistics," Vol. 3, pp. 203, 287, and 460. Van Nostrand-Reinhold, Princeton, New Jersey, 1968.
184. E. Weber, "Grundriss der biologischen Statistik," p. 403. Fischer, Stuttgart, 1972.
185. E. Weber, "Einführung in die Faktorenanalyse," pp. 5 and 13. Fischer, Stuttgart, 1974.
186. S. S. Wilks, "Mathematical Statistics," p. 587. Wiley, New York, 1962.
187. B. J. Winer, "Statistical Principles in Experimental Design," pp. 109, 191, 298, and 538. McGraw-Hill, New York, 1962.
188. Y. Yao, *Biometrika* **52,** 139 (1965).
189. J. H. Zar, "Biostatistical Analysis," p. 233. Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
190. G. G. Zyskind, "Topics in General Linear Models Theory." Contributions to Research in Analysis of Variance and Related Topics, Aerospace Res. Lab, Tech. Rep. 64–193. Wright-Patterson Air Force Base, Ohio, 1964.

Chapter 5 Rationales in the Design of Rectal and Vaginal Delivery Forms of Drugs

C. J. de Blaey and J. Polderman

I. General Introduction

To attain the full benefit of active drug substances, it is essential that they be supplied in their best possible dosage form. The rapid increase in interest in biopharmaceutics provides lively support for this statement. Research in this field is in progress either to improve the release from well-known dosage forms or to develop completely new types, some of which may enable a more direct delivery to the target organ or may make better use of specific absorption sites, such as the skin, lymphatic vessels, and rectal or vaginal mucosa. However, it is not always kept in mind that biopharmaceutical studies should result in the design of a drug formulation, rather than beginning with it (*49*). Basic studies that emphasize absorption site, drug substance, and dosage form are therefore required

ISBN 0-12-060309-8

for the rational development of a true drug delivery system. This is so for all routes of administration, and the rectal route especially provides an excellent example of the complexity of the problem and of the fact that so little is actually known of the reasons for delivery form design.

The first requirement for the development of a delivery form is a relevant criterion for judging its quality. This is usually found in bioavailability, defined as the rate and extent of absorption from the dosage form as reflected by the time–concentration curve of the unchanged drug in the general circulation. Where necessary this definition has to be adapted, keeping in mind the therapeutic goal, e.g., locally administered products. The application of bioavailability data requires good insight into the determining factors: availability of the drug substance at the site of absorption and membrane transport.

The availability at the absorption site, which may be called pharmaceutical availability, depends mainly on the mechanism of release from the dosage form. For membrane transport the properties of the drug substance are predominant. Well-designed preformulation studies are thus essential for evaluation of the eventual bioavailability, taking into account both the release from the dosage forms and the physicochemical properties of the drug substance in relationship to the physiological conditions of the absorption site. Such studies will start with *in vitro* experiments since the pharmaceutical availability can be worked out in this way, provided the biological requirements are kept in mind. The results obtained *in vitro* then have to be checked *in vivo* to assess their true meaning. Successively, animal and human experiments have to be done. Where clearly the human situation is of interest only, many problems arise, in particular when the physiological conditions differ greatly between animals and man or are not yet known. This undoubtedly holds true for the rectal administration of drugs.

First of all, a great limitation exists in the number and especially the kind of experiments that can be performed in man; e.g., absorption experiments through perfusion (*5*) are very useful, but are much too invasive to permit statistical evaluation of the many variables on which one would wish to gather information. This is all the more so, since the reproducibility of *in vivo* experiments, human or animal, is usually such that the number of experiments has to be large to permit statistically significant conclusions.

Furthermore, the *in vivo* situation is so complex that single variables cannot be studied easily. Even if, for example, only the size of the suppository is varied (as is the case with aminophylline suppositories, marketed in a 2 or 3 g size), a number of events are taking place concurrently. Not only is the concentration of the drug changed, affecting the vehicle

viscosity, but also the spreading behavior may be affected, through the changed viscosity and the direct reaction from the rectum wall muscles provoked by the different volumes inserted. In addition, the chosen suppository base may or may not provoke irritation, resulting in a response of the rectum and in this way obscuring the effect of the variable under investigation. Therefore animal studies, which in part avoid the disadvantages cited, are considered a useful step in studying delivery forms. It goes without saying that their limitations should be kept in mind, especially when extrapolation to man is attempted. This holds even more so for *in vitro* results, where extrapolation is even more difficult. This type of study has one important advantage to offer, namely that single variables can be studied in great depth. Even then, however, one should always consider whether the model used has any significance for man. Nevertheless, this approach seems to us the most appropriate way to find a rationale for the design of delivery forms.

In this chapter we will limit the discussion to factors that might influence both *in vitro* and *in vivo* availability, and we will not include such quality criteria as content uniformity, mechanical properties, and quality of the vehicles, for which we refer the reader to the survey of Senior (*85*).

II. Rectal Delivery Forms

A. Introduction

For several centuries B.C., the use of the rectal route for the administration of drugs was known to Egyptians and Greeks, who used a rhizome dipped in a fat and subsequently in the powdered drug substance. In the eighteenth century a French pharmacist, Baumé, introduced cocoa butter as a vehicle for the preparation of suppositories, and this has made further progress possible. A great variety of fatty vehicles were developed, during and after the second world war that do not show the same pronounced polymorphism and can, therefore, be moulded under less stringent temperature conditions. Also, nonfatty vehicles have become available, and other dosage forms, such as ointments, microenemas, and soft shell gelatin capsules, are in use for rectal administration. The choice between the different forms has to be based on several factors, such as manufacturing equipment, geographical distribution, stability, compatibility, and applicability to human use. These factors are outside the scope of this chapter; more important here are the consequences for bioavailability. Inconvenience to the patient limits the use of capsules and microenemas, whereas the nonfatty vehicles are usually considered irritating.

It seems necessary to deal with the place of rectal delivery forms in modern drug therapy. Some drawbacks are reported: inconvenience to the patient, consequently limiting them to bedridden patients only; technical difficulties in large-scale production; stringent storage conditions, in particular in tropical climates; and often, slow onset of effect and great variation in availability. But some advantages can be seen too: preferrence in cases where oral administration is impossible as in nausea and other gastrointestinal disturbances; postoperative or pediatric use; as a way to possibly avoid drug abuse, e.g., suicides; and as a possible way to decrease first pass metabolism.

Obviously, the rectal route is not an alternative to the oral one except in the situations just outlined, where it can be of much benefit. So far little fundamental work has been done in this area, and it may be that some of the disadvantages mentioned can be overcome or diminished, in particular that the availability can be controlled. In addition, the possibilities should be explored, in more depth, for evaluating its use as a prolonged action preparation, e.g., to sustain a therapeutic level obtained orally during the day, during the night.

The rectal route of administration can be chosen both for local and for systemic effect. Suppositories and ointments for local effect are almost exclusively used for the relief of pain and itching due to hemorrhoids, with such common constituents as astringents, antiseptics, antipruritics, vasoconstrictors, and local anaesthetics. It remains doubtful, however, whether suppositories represent a very rational choice here since they melt or soften in the *ampulla recti,* and spread through the rectum (*32, 76*) and will thus have at least partly passed the target, i.e., the anus.

For laxative purposes, retention enemas or suppositories of glycerin gelled by soap can be given in addition to drugs. For the same purpose, suppositories releasing CO_2 after insertion, and thus stimulating defaecation, have been used. Retention enemas can also be used for local treatment, as in colitis; corticosteroids, either with or without antibiotics, are then dissolved or suspended in a moderately viscous aqueous medium.

Two more recent developments are microenemas (*2, 13, 53, 54*) and soft shell gelatin capsules (*67, 92, 94*). Microenemas are supplied in a plastic container (Rectiole) with an application tube. After insertion of the tube, the container is emptied by compression. This permits the use of either aqueous or oily solutions for systemic application of drugs. The few results available are promising, although we consider this form of only limited applicability, because of patient inconvenience and cost. An example is given in Fig. 1.

Capsules used to achieve a systemic effect are filled with oil or paraffin. Only limited experience has been obtained so far with this dosage form,

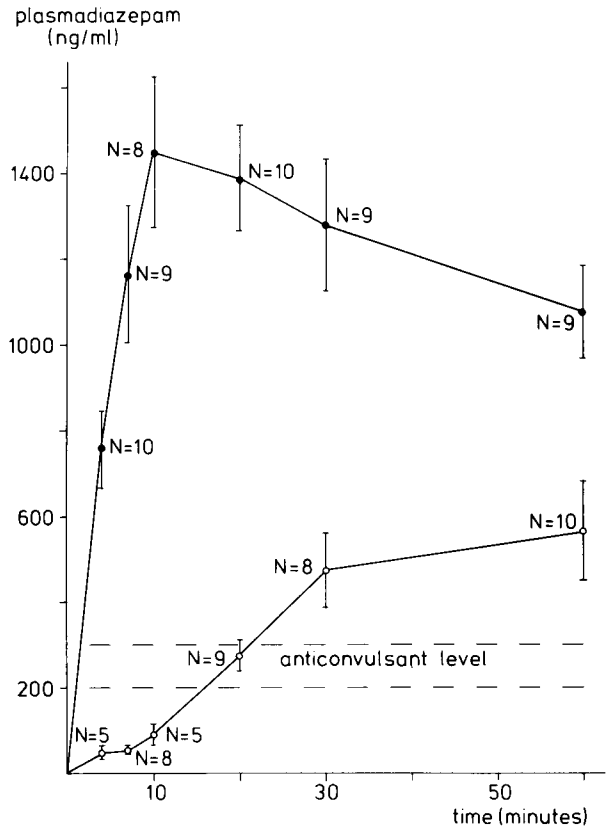


Fig. 1. Plasma diazepam following administration of rectal solution (-●-) and suppositories (-O-) to children (*53*).

but from comparisons made, it seems fair to state that no striking differences exist between suppositories and rectal capsules. Widmann (*94*) reported a somewhat more rapid release of paracetamol, whereas Neuwald and Ackad (*67*) did not find any significant difference for aminophylline. It may be that capsules have some advantages for tropical areas because of their longer shelf life (*40*).

Of all the forms, fatty suppositories containing a suspended drug substance are the most widely used. Commonly used are suspensions of water-soluble substances in cocoa butter or in one of the many semisynthetic vehicles. These are prepared by hydrogenation of vegetable oils, e.g., coconut oil, or by esterification of glycerin with saturated fatty acids isolated from hydrolyzed vegetable fats. Differences between vehicles thus occur (e.g., in hydroxyl number). The actual importance of this

is not yet fully understood, but it seems essential that switching from one vehicle to another one be avoided in order to prevent batch to batch biopharmaceutical variations. The same holds for bases containing extra emulsifying agents.

The aqueous glycerinated gelatin vehicles are an exception and are only used for topical treatments. The macrogols are sometimes used for drugs that are fat-soluble or, e.g., for chloralhydrate. The results obtained are shown in Fig. 2.

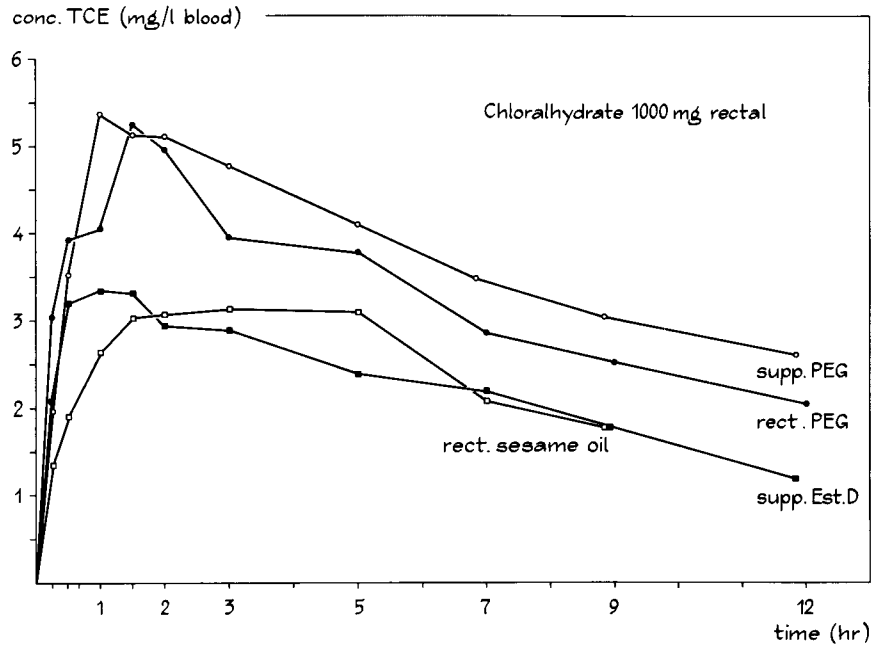


Fig. 2. Average trichloroethanol blood concentration curves after rectal administration of four dosage forms, each containing 1000 mg of chloral hydrate. From (13).

B. Rectal Absorption

Drug uptake from the rectum is not essentially different from that in other parts of the gastrointestinal tract. Differences that occur have to be attributed to anatomy and physiology, rather than to specific membrane properties. Due to the nature of the membranes, passive diffusion is still regarded as the main governing absorption mechanism. Therefore, solubility and partitioning parameters of the drug substance are of paramount importance.

The anatomical situation is described extensively in textbooks (*64*). For this survey it may suffice to mention that the length of the *ampulla recti* is about 15–20 cm. Combined with the absence of *villi* and *microvilli,* this results in a very small absorbing surface (probably < 0.05 m^2) as compared with the small intestine ($\sim$ 70 m^2), even if there is no distinct separation with the sigmoid colon. This is shown, for example, in photographs of the spreading area of suppositories, with X-ray contrast media (*32,* 76). However, these data are disputable, since substances with a high density do not spread evenly with the suppository base, as will be shown later (Section II,E).

Physiologically, the situation is even more complex. Most textbooks deal only superficially with this aspect, and then mostly in discussing defecation. Although the actual situation in man varies a great deal, it may be assumed that normally the rectum is empty, although this should not be taken as absolute (*29*), and largely in rest; that is to say, a regular peristaltic movement is absent. More details concerning this point will be discussed (Section II,E,1).

An important question still under discussion is the possibility of avoiding first pass metabolism through rectal administration of drugs. The lower rectum is surrounded by veins draining directly into the general circulation, whereas the upper and middle rectal veins lead to the *vena porta.* Because of many anastomoses between the two regions, a sharp separation cannot be assumed. Data available indicate that the first pass metabolism may be avoided for some part, but the exact extent is unknown (*12, 14, 32*). It seems therefore advisable to abstain from quantitative rules for the effect.

The absorption step itself seems to follow the pH partition theory, at least in a qualitative sense (*35*). But data obtained by Kakemi and coworkers (*45*), and recently also by Bechgaard (*5*), show that quantitatively serious deviations from this concept occur. Kakemi *et al.* (*45*), studying the small intestine, propose an alternative rate-limiting step, arising from the binding to the mucosal material. Others proposed simultaneous absorption of ionized species (*21, 68*). For drug design, the question is still open as to whether one has to adhere strictly in this case to molecules that are mainly un-ionized in the physiological pH range of 7.5–8.0. At this moment it still seems a good choice, although more evidence is becoming available that drugs that are completely ionized, like quaternary ammonium compounds, are also absorbed to a certain extent after rectal administration (*42*). The amounts are limited (i.e., $\sim$ 10% or even less), thus it seems a long way to go before this will provide an alternative in drug development.

C. Release Mechanisms

Depending on the nature of the vehicle and of the drug substance, different release mechanisms can be envisaged. As was said earlier, rectal delivery forms are solutions or suspensions in fatty or water-soluble vehicles.

In water-soluble vehicles, used mostly for drugs like indomethacine that are poorly soluble in water and that have an appreciable solubility in fat, the drug will be present as a solution in PEG. In that case, the drug substance will be released as soon as contact has been made between vehicle and rectal fluid. Release then takes place by dissolution of the vehicle in the rectal fluid. Absorption will then be limited either by membrane transport or by the diffusion rate in the aqueous contents of the rectum and the partitioning behavior between membranes and contents of the rectum. For drugs such as indomethacine it has been reported that *in vitro* release, and to a somewhat lesser extent *in vivo* absorption, is more rapid from a water-soluble than from a fatty vehicle (*51*). In the case where the drug substance is suspended in the aqueous vehicle, the particle size will be an additional parameter that may be rate-determining. All depends on the relative magnitudes of the rates discussed, and until now, little quantitative information was available to be applied directly. Additionally, it has to be kept in mind that many drug substances are liable to be bound by polyethylene glycols.

The same considerations hold for microenemas, which contain the drug in PEG containing solutions.

But again, these types of delivery forms are only rarely used and most emphasis has to be laid on fatty vehicles and, particularly, on suspensions of relatively highly water-soluble drug substances. Principally, two different release mechanisms are possible (*6*): (a) the suspended drug dissolves in the vehicle, and is released through diffusion; and (b) the suspended drug comes into contact with the aqueous rectal fluid and dissolves directly in it. Higuchi (*34*) has worked out a model for suspension ointments and some authors have been tempted to apply this model to suspension suppositories. Since suppositories melt fairly rapidly after insertion, a rather thin layer of a medium-viscosity suspension ($\sim$ 50 mPa s) is formed. Assuming that the particles spread homogeneously with the vehicle (see Section II,E,3), not all conditions under which the Higuchi treatment applies are met. The particles will not remain distributed homogeneously throughout the vehicle, but will reach an equilibrium position in the vehicle/rectal fluid interface. The reason is found in the viscosity of the melted suppository permitting particles to settle. This is reflected in the observed rate of release from suppositories, which is far more rapid than

predicted by the Higuchi equation. Usually, *in vitro* release is completed within a few hours, whereas only a few percent would have been released according to this equation.

We therefore prefer a release mechanism composed of three steps: (a) approach of the suspended drug particles toward the interface between the melted vehicle and the aqueous rectal fluid; (b) transport through that interface; and (c) dissolution in the aqueous rectal fluid. It should be noted that this descriptive presentation will show more deviations when the solubility in the fatty vehicle is not negligible. The mechanism is now considered in some detail.

This transport process (step a) may take place through two different driving forces. First, sedimentation is possible as soon as the viscosity is lowered during the melting process. This will be all the more important as the spreading area is smaller, and therefore the layer thickness of the melted mass remains larger. Second, the pressure waves occurring in the rectum could induce particle motion in the melted mass and thus collision with the interface, in which the particles will reach an equilibrium position stabilized by the surface forces. Both processes would predict an influence of viscosity of the melted mass and also of the particle size of the drug. This latter observation is in contrast with the Higuchi ointment treatment, where no such influence is predicted. Also, the particle concentration could be a factor, as it influences the viscosity of a melted mass too. Results illustrating these points will be given in Section II,E,2.

Particles will cross the interface (step b) depending on their surface properties, in relation to the same properties of both liquid phases. Parfitt (*73*) has treated the case of cubes arriving at an oil–water interface and has shown that from a thermodynamic point of view, adhesion of the cube takes place spontaneously if the contact angle Θ_{sow} (of aqueous phase on a solid immersed in oil) is smaller than 180°, and immersion if this angle is smaller than 90°. A sphere, however, will take a position in the interface in every case since Θ is always smaller than 180° (*41*). The depth of immersion will increase with decreasing Θ_{sow}. For particles of the size used in suppositories, gravitational forces will not be of any significant influence.

This shows directly the importance of surface properties and the possible role of surface active additives.

The dissolution process (step C) is clearly an intrinsic property of the drug substance and can be derived by using the appropriate equations. There is one problem, however, that has to be considered, i.e., the surface exposed to the dissolving rectal fluid. The situation at the interface is not comparable to the one in which a particle dissolves in a stirred medium. The flow of liquid is negligible and the particle continuously adjusts its position to maintain its equilibrium position, as dictated by the surface

forces. Our preliminary results indicate that the dissolution time of, for example, paracetamol is also linearly related to the radius of the particle. This corresponds to the findings for the dissolution of crystals at a paraffin–water interface (*84*).

An important question now arises as to the use of the proposed mechanism. It is, first of all, relevant to know if one of the steps (a, b, or c) is rate-limiting, and if so, which one. Experiments have been performed in a model system consisting of different substances suspended in paraffin, as the fatty vehicle substitute, and water. For a very soluble substance, steps b and c are very rapid and the release rate is limited by step a, at least in the thickness layer used (1 cm). A less-soluble compound, i.e., paracetamol, showed a more complicate behavior. In addition to the water solubility, this is caused by the agglomeration behavior of the suspension; as a result, the amount released remains constant with increasing concentration (*18*). Thus, by studying the steps (a, b, and c) both separately and in combination, more information that might be useful for development purposes becomes available.

D. Control Methods

1. *In Vitro Techniques*

As argued in the introduction, each biopharmaceutical development study should include in its early stages an evaluation of the release from dosage forms. At the present level of knowledge, however, many different techniques have proved successful only in specific cases, with respect to their predictive capability for *in vivo* performance of the drug. Nevertheless, *in vitro* testing is a useful tool in the first screening of possible alternatives in the development stage.

This implies that such tests should take into account, as far as possible, the most essential conditions prevailing *in vivo*. In the case of rectal delivery forms, these parameters seem to be of importance: temperature, contact area between delivery form and rectal fluid, and such properties of the rectal fluid as surface tension, viscosity, and volume. Stirring conditions also have to be kept in mind. As will be discussed in Section II,E,1, too little information is available to permit a generally successful method to be developed, notwithstanding the many attempts in this direction.

Technically, the temperature provides no problems and it is still somewhat dubious as to whether this is critical. Using alkanes, it was shown clearly that when melting points exceed 37°C, the release at that temperature decreases considerably (*58*). For fatty suppositories with melting

ranges rather than melting points, this is less clear. Some authors did find a reduced release both *in vitro* and *in vivo,* as soon as the melting range exceeded body temperature by a few degrees (*57, 65, 66*). Our own experiences with aminophylline suppositories, with rather extended melting ranges, did show a marked reduction *in vitro* (*10*) but no distinct differences *in vivo* (*11*). Possibly the rheological properties at 37°C are of more importance than the melting behavior.

The contact area *in vivo* cannot be simulated directly, since it may amount to 0.05 m^2, and since more importantly, it may vary between different vehicles. It should at least be standardized *in vitro,* resulting in a kind of intrinsic release rate. In this respect the older methods, reviewed by Wolf and Voigt (*95*), used for suppositories fail; more recently Cox and Breimer (*17*) and Bhavnagri and Speiser (*8*) have overcome this disadvantage by using an open cylindrical tube in which the suppository is retained during and after melting. Thus, a well-defined contact area between melted mass and dissolution medium is used. Schoonen *et al.* (*83*) have used a plate immersed in water, below which the suppository melts and spreads. The surface is thus not constant, but observable.

But even if the *in vitro* contact area is taken equal between experiments, still the *in vitro* results could well be quite different from those obtained *in vivo,* since the spreading behavior between vehicles may be quite different (*80*).

Water or aqueous buffers are mostly in use as dissolution media. With respect to the variables listed in Table II (Section II,E,1), this is certainly a quite unnatural situation. The importance of this depends strongly, however, on the release mechanism prevailing. For formulation research, it is safe to assume that the release from the delivery form is rate-limiting, and therefore working under sink conditions *in vitro* seems the method of choice. For water-soluble vehicles containing compounds of low water solubility, a limited amount of liquid, as in methods with a dialyzing tube, might be more like the *in vivo* situation.

A somewhat lower temperature at the evaporating water surface may result in a lower melting rate (and incomplete melting). This can be avoided by keeping the suppository under water while melting. Floating afterwards is acceptable, since only minor differences in viscosity will be brought about.

Until now only few attempts have been made to include pressure as a feature of *in vitro* testing. Thomas and McCormick (*90*) have used a dialyzing tube, into which the suppository is placed. By means of streaming water, pressure is exerted on the tube. Our own experience in a similar set up, but applying well-defined variations in pressure, proved successful with aminophylline suppositories. Using a static method (*10*), release was

negligible, whereas under pressure (varied from 0–50 g/cm^2, with a cycle lasting ~ 0.5 min every 2 min) complete release took place within 90 minutes, which was more in agreement with the *in vivo* release.

Several release techniques make use of dialyzing membranes for two reasons: to simulate the absorption phenomena and to prevent mixing of the fatty base with the dissolution medium. In our opinion, the former reason is not very sound, since an additional uncertainty with respect to the relevance for the *in vivo* situation is introduced. Moreover such a combined release–absorption test will not easily reveal where the rate-limiting steps are situated. It seems therefore more fruitful to study release and absorption separately and to combine these data afterwards.

The use of semipermeable membranes introduces resistance into the system, the influence of which will depend on the ratio of release to permeation rates. It also creates a situation in which osmotic attraction of water through the membrane toward the suppository may occur. This can be diminished by a proper choice of medium, but an uncertainty remains as to the amount of unstirred water around the suppository. By analogy to release processes governed by diffusion, such a barrier might distort measurement of the real release rate.

In most membrane methods, the interface between suppository and dissolution medium is not well-defined, and since spreading is often limited, the thickness of the mass differs greatly from that occurring *in vivo*.

In conclusion, no satisfactory method is available as yet, although the methods using a constant area and no membrane seem most promising. In addition, the introduction of pressure variations might add a new and useful feature, although in this case the use of a membrane seems unavoidable.

Also with a view to quality control, there is a great need for an *in vitro* test that takes account the *in vivo* phenomena. In particular, for dosage forms like suppositories where so many factors can impair the release of the active substances, a thorough check of the release rate of the active substance in each batch by means of a representative *in vitro* test cannot be spared.

2. *In Vivo Techniques*

As has been said, it is best to separate clearly the release of the active substance and its absorption. Ideally, the release should be determined by *in vitro* tests whereas the absorption rate should be measured by perfusion tests *in situ*. It is evident that none of these tests can yet be performed with satisfactory reliability in all cases. But even if this were possible, we should still want an *in vivo* method at our disposal that would enable us to check the overall effect, i.e., bioavailability in man.

For research purposes such human tests are most impracticable and an intermediate step, i.e., testing in animals, cannot be missed. Only in this way can the number of variables eventually to be tested in man be reduced. This brings forth the problem of a proper choice of animal species. Many considerations have been made, showing where certain laboratory animals fail to represent the human situation. With respect to testing of delivery forms, however, these points focus mainly on differences in size and, sometimes, anatomy.

Much attention should be paid to the difficult problem of preparing representative test batches. Apparently minor variations in manufacturing may result in striking differences in release rates (an example is shown in Table I). The more intensive the mixing, the slower the release of the

TABLE I
In Vitro MORPHINE RELEASE (%) FROM COCOA BUTTER SUPPOSITORIES PREPARED IN TWO DIFFERENT WAYS

Release time (min)	Mortar	Three roller mill
30	48	20
60	62	31
90	68	36

morphine hydrochloride. Accordingly, when the mixture is rubbed for longer times, the release rate is slowed, and thus the reproducibility in making small batches of suppositories poor.

Moreover, the concentration of suspended material should be similar in the animal and human delivery forms. This sometimes provides problems, but may be avoided by comparative trials studying the appropriate concentration range. On a body surface basis, this will result in comparable loading in the animal and in man, using, e.g., a 10% suspended drug in a 50-mg and 3 g-suppository in a 200-g rat and a 70-kg human, respectively. This still might distort the results when capacity limited metabolism or transport processes are involved.

Taking these points into consideration, we have found rats an acceptable animal model for the testing of suppositories. It is mainly a matter of carefully preparing the delivery form in a size that is no more mechanically disturbing to the rat rectum than to that of man. Objections raised as to the small blood volume, which might restrict the sampling, can mostly be overcome today by using sufficiently sensitive analytical techniques, thus permitting reduction of sample volume to ~ 0.3 ml, or by using canulation and/or replacing samples by saline. Our experiences (*80*) were such that, by feeding slack free food, the amount of feces produced was

no longer a disturbing factor. Suppositories of around 50 mg could be used without problems, given the fact that the anus was closed by clips immediately after insertion. Anesthesia to permit blood sampling from the orbital *venous plexus* (behind the eye) did lower the rectal temperature somewhat, but not to such an extent that the already melted suppository would re-solidify, even partially so. Later we did notice, however, some influence of anesthesia on the rectal pressure profiles (*19*); this has to be further evaluated with respect to the spreading behavior.

When the influence of particle size of readily water-soluble drugs suspended in fatty vehicles was studied, good agreement was found between results obtained in rat and in man (See Section II,E,2). This suggests that rats can be a useful intermediate between *in vitro* testing and studies in man, especially after underlying mechanisms (e.g., release) have been studied in detail *in vitro*.

E. Factors Influencing Release *in Vitro* and *in Vivo*

In a summary of the literature concerning the biopharmaceutical aspects of suppositories, Bevernage and Polderman (*6*) have drawn up a list, which is reproduced in Table II in a somewhat adapted form. This

TABLE II
FACTORS INFLUENCING AVAILABILITY FROM SUPPOSITORIES[a]

Rectal fluid	Drug substance	Vehicle
Amount	Solubility	Composition
Composition	Surface properties	Fusion behavior
pH	Particle size	Surface tension
Buffer capacity	Drug concentration	Rheol. behavior
Surface tension	pK_a	
Viscosity		
Luminal pressure		

[a] Adapted from Bevernage and Polderman (*6*).

table will form the basis of our discussion here. To their original list, rectal pressure has been added as a factor, based upon recent experiences with aminophylline suppositories (*10, 11*).

1. *Physiological Factors*

Under this heading, properties of the coating of the absorbing membranes together with the behavior of the rectum itself, will be discussed. These are all more or less given facts, changeable not at all or only with great difficulty, which means that the drug product has to be developed in

such a way that optimal use is made of the given situation. This implies, however, a certain amount of factual knowledge, hardly available today.

The *amount of liquid* is reported as around 3 ml (*31*), which is small compared to the volume available for dissolution in the peroral route. This is a drawback to rectal administration of drugs of low water solubility, and also to rapid absorption in the application of water-soluble vehicles. Not much is to be expected of attempts to increase the available volume—as is, for example, automatically done osmotically by water-soluble vehicles—since this will increase the risk of premature expulsion of the drug. As a consequence, this would limit the rectal administration of drugs of low water solubility, especially when the rate-limiting step is in the dissolution step.

The *composition of the liquid* has drawn only scarce attention (*36, 88*) and for design purposes has to be regarded as unknown. Kakemi (*45, 47, 48*) has postulated the binding of drugs to this material as a possible rate-limiting step in the absorption from the small intestine, thus indicating its possible importance. Its nature seems to be polymeric and for this reason alone it is a possible source of interactions, and thus interference, with absorption.

The *pH* is reported as close to 8 and is not subject to large variations (*5, 9, 30*). This may be important in connection with the pK_a of the drug substance.

The *buffer capacity* is probably at least as important as the pH itself, since it directly determines the pH and therefore the ionized to unionized ratio for any drug substance. Bechgaard (*5*) found that a buffer capacity of 0.01 was sufficient to maintain the pH at 5.9 while perfusing in humans. Withdrawal of the perfusate restored the pH at a rate of about half a pH unit per minute. Our own experiences (*20*) with *in situ* rectal perfusion of rats was such that, depending on the deviation of the perfusate pH from the physiological value, the rectum had an adapted response to restore the pH to around this value. It appeared that a perfused solution containing 10 mg aminophylline in 30 ml water (pH ~ 9) was neutralized in about one hour. Perfusing the same compound in phosphate buffers at various constant pHs revealed that, with increasing aminophylline concentration, the amount absorbed increased linearly. The influence of pH was such that, between pH 7 and 10, only minor differences in the amount absorbed could be detected (between 10 and 15% absorbed in 2 hr), showing that the pH partition theory is not followed quantitatively. Combined with Bechgaard's results on salicylic acid (*5*), (findings that plead against the existence of a fixed pH at the surface of the absorbing mucosa, the so-called virtual pH), it seems rather that the organism tries to keep the pH at the absorbing surface constant by secretion but is successful only after a

certain time period. Smolen (*87*) and Hwang *et al.* (*38*) have stated the surface pH cannot differ significantly from that in the lumen. Desai (*24*) has found, however, that such a difference may exist, which is more in agreement with our findings. Thus, the contribution of ionized species to total absorption cannot yet be concluded.

The *surface tension* of the rectal fluid may have some influence on the spreading behavior of the melted vehicles, although we feel that this is of minor importance since the equilibrium spreading pressures (and thus the ability to spread) are found to be only a few mN/m both for cocoa butter and for Witepsol H15. Of greater consequence would be the potential to *decrease* the contact angle of a suspended drug particle at the melted vehicle/rectal fluid interface, thus directly influencing the contact area with the dissolving rectal fluid. Unfortunately, no such data on contact angles are available for drugs; it is, therefore, almost impossible to predict the influence of additives that might change the interfacial situation. The surface tension of the rectal fluid cannot be changed easily and agents doing so would have to be administered in the suppository itself, making the effect uncertain.

The *viscosity* of the rectal fluids is an important factor in the transport of dissolved molecules and possibly also in the dissolution process of water-soluble vehicles. No attempts seem to have been made to change this parameter by use of mucolytic substances.

A factor that seems to be of great consequence is the *pressure* exerted on the suppository by the rectum wall. As shown earlier, the pressure is such that the *viscosity* of aged suppositories might be overcome. Applying a moderate weight of around 20 g caused aged suppositories to spread *in vitro,* thus increasing the area available for release. This suggests that the pressure encountered *in vivo,* in both man and rat, is large enough to induce spreading. This might be one explanation for the fact that *in vivo* results between different formulations are often much less different than *in vitro* results (*11*). Influencing the rectal pressure has probably been done many times unintentionally by not standardizing the movements of volunteers in experimental trials. The exact consequence is not known as yet, and data are needed to open the possibilities of influencing release in this way, in either a positive or a negative sense.

2. *Factors Related to the Drug Substance*

In Table II these factors are presented in a strictly random fashion, without giving any emphasis to the rank or order given. These factors play an important role in the design of the final dosage form, since they offer many possibilities for the selection of active substances.

Drug solubility. Since only a little fluid is present in the rectum, the first

choice will be substances that are readily soluble in water. Voigt and Falk (*91a*) investigated 35 compounds and found a direct proportionality between release rate and water solubility. Examples are also known for drugs that can be supplied in both a poorly soluble form and as a water-soluble salt, e.g., salicylic acid and sulfonamides or their sodium salts or theophylline as i.a. a choline salt. The *in vitro* consequence is shown in Table III, where a comparison has been made between two suspended

TABLE III
In Vitro PERCENTAGE SALICYLATE RELEASE FROM SUPPOSITORIES[a]

Time	Suspension[b]		Emulsion[c]
	Acid (n = 17)[d]	Na salt (n = 20)[d]	Na salt (n = 21)[d]
50 min	8–18	40– 90	4–14
90 min	12–25	45–100	5–18

[a] From Basart-Westgeest *et al.* (*4*).
[b] 100 mg acid or 115 mg salt in 3 g cocoa butter.
[c] 115 mg salt, 300 mg water, 300 mg *adeps lanae* in 2.4 cocoa butter.
[d] Number of runs.

forms and one emulsified form of salicylic acid. The higher intrinsic dissolution rate favors faster release of the sodium salt. *In vivo,* this may have the same effect as that shown in Fig. 3 for some sulfonamides. It has been shown that the advantage of a higher intrinsic dissolution rate remains, even at unfavourable pHs where transition in, for example, the acid form, will subsequently occur (*63*). This effect, shown in both examples here, has been documented several times in the literature (*23, 26, 81, 82*) and is one of the most firmly established rules for suppository design.

Drug solubility also determines the choice of vehicle type, since the drug should have no great tendency to remain in the vehicle. Thus water-soluble substances are best suspended in fatty vehicles, while substances that show an appreciable solubility in fat are best dispensed in water-soluble vehicles. This has been demonstrated *in vitro* for indomethacin (*51*), although very little release was noticed in all cases. This point, comparing different vehicles *in vivo* (*51*), has been documented to only a very limited extent.

The *chemical form* of the drug can also be considered with respect to *in situ* availability and/or absorption. For substances with a high lipid solubility, rectal administration has little use even if they are sometimes absorbed through the lymphatic vessels (*52*), since this route is too slow to offer any interesting practical possibility. Another possibility lies in the application of prodrugs, the more general aspects of which have been

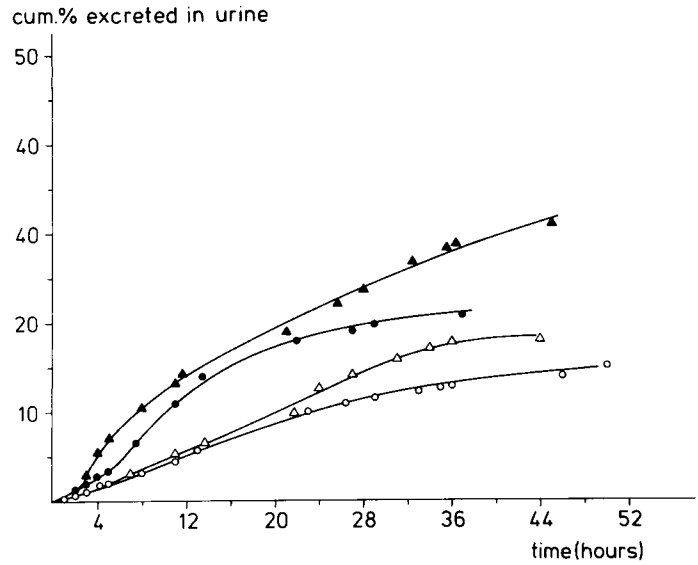


Fig. 3. Urine excretion after administration of sulfadimidine (○-), sodium sulfadimidine (-●-), sulfadiazine (-△-) and sodium sulfadiazine (-▲-).

recently reviewed (*86*). Absorption could be promoted in some cases, e.g., for ampicillin derivatives where the proper choice of the ester form could provide an optimal balance between the decrease in dissolution rate and the increase in absorption rate. Jonkman (*42*) has tried to improve the rectal absorption of thiazinamium by addition of ion-pair forming ions, but little effect has as yet been noticed.

The *surface properties*. First, the energy content of a solid determines directly its dissolution rate; this may therefore depend on the preparative method of the solid (*55*) and/or its polymorphic form (*28*). Examples are known for several compounds such as chloramphenicol (*1*) and sulfonamides (*43*), but whether this holds when drugs are given rectally is not known.

Second, the surface energy is important when suspended drug particles reach the interface between fatty vehicle and rectal fluid. The particle will then settle into an equilibrium position dictated by its surface properties. The smaller the contact angle formed by the underlying rectal fluid, the further the particle will immerse in that fluid, which might result in a faster dissolution process. This is important only for those drugs in which the approach to the interface is fast compared to the dissolution step (*18*). This contact angle is usually a more or less given fact, since only the vehicle can be varied. Addition of surface active agents may have an

appreciable influence. Not only will be the contact angle changed, but an influence on the agglomeration behavior of the suspension may also be expected. This results in an overall effect that is still unpredictable since, depending on the solid and the surface active agent, both the contact angle and the agglomeration will change simultaneously. An extreme situation was encountered when Span 85 was added to paracetamol suspensions in paraffin with the release to an aqueous phase being reduced to nearly zero even though the contact angle was still smaller than 180°, which would predict spontaneous contact with the qaueous phase.

The *particle size* of the drug in suspension suppositories has been widely considered important with respect to content uniformity. It is advisable to use particles no larger than ~ 150 μm, but no lower limit has been proposed; several authors state that smaller particles perform better than larger ones (*78*). This has been shown by Weiss Fogh for testosterone (*93*). In his experiments it is not clear, however, whether the drug is present in suspension or is, at least to some extent, dissolved in the vehicle used. Kraml *et al.* (*56*) have found that orally administered suspensions of griseofulvin in oil resulted in higher blood levels when micronized particles were used. For compounds with higher water solubility, the dissolution step is probably no longer rate-limiting for the release process. An analogy can be found in the work of Buckwalter and Dickison (with procain penicilline given intramuscularly) who found higher levels with larger particles (*15*). With NaCl as the model for a very soluble compound, it was shown that larger particles gave a more rapid release (*7*). This was confirmed for other water-soluble compounds by Rutten-Kingma (*80*) and by Schoonen *et al.* (*83*). An example is given in Fig. 4, confirming the *in*

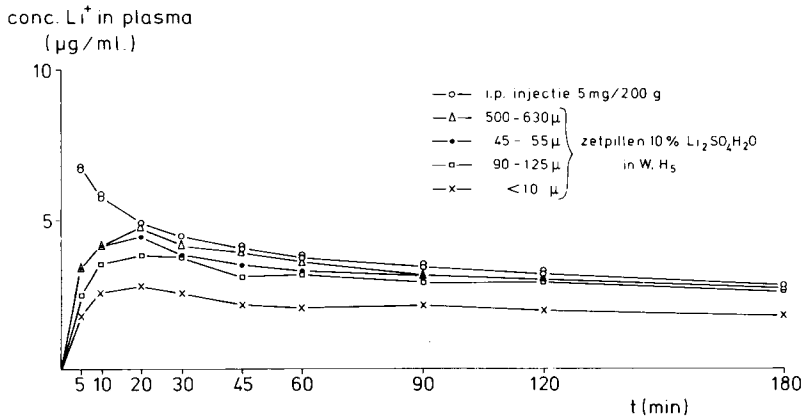


Fig. 4. Mean blood levels in 10 rats after administration of lithium sulphate IP and in suppositories using four different drug particle sizes. From (*80*).

vitro results. The effect is especially clear when particles of < 100 μm are compared with particles of 90–125 μm. Particles of 45–55 μm are somewhat slower than the 90–125 μm particles, whereas no difference is seen between particles of 500–630 μm and of 90–125 μm (see also ref. 18).

In man, the same effect was found upon administration of sodium pentobarbital, comparing particles of < 20 μm and 90–100 μm. In Table IV, the time of the maximum blood level is seen to be shorter for the larger particles.

TABLE IV
t_{max} (HOURS) AFTER RECTAL ADMINISTRATION OF SODIUM PENTOBARBITAL OF DIFFERENT PARTICLE SIZES IN WITEPSOL H5® [a]

Size (μm)	A	B	C	D	E
90–100	3	3	3	2	5
< 20	9	9	5	5	3

[a] From Rutten-Kingma (*80*).

This leads to the conclusion that it seems advisable to choose the particle size of readily water-soluble drugs to be between ~ 50 and ~ 150 μm, when given in fatty vehicles. From a technological point of view such a two-sided limitation is a serious drawback. It implies the need for a well-defined sieve fraction and the rejection and/or working up of deviating sizes. Furthermore, care has to be taken to avoid breaking of particles during further handling and preparation steps. Moreover, the larger particles present a potential problem as to content uniformity because of settling during the preparation; many synthetic vehicles cannot be cooled too rapidly, which would solve this problem, because of their tendency to cracking.

The size of a suppository is usually arbitrarily choosen at ~ 2 or 3 g. Thus the *concentration of suspended drug* may rise as high as 25%, for example with 500-mg aminophylline suppositories. More important still, in determining the rheological behavior, is the number of particles, i.e., introducing particle size again. For the theory, still mainly based on empiricism, we refer to the literature (*79*). Experimental evidence on the influence of drug concentration on the release is largely lacking, with regard to the viscosity determined by the particles themselves. In our laboratory we have found little influence when the concentration of lithium sulfate was varied between 5 and 20% in a fatty vehicle given to rats. This was true even for particles < 10 μm at the 20% concentration level, although the viscosity was raised by a factor of about three compared to the 5% series. These results suggest that the choice of supposi-

tory size and/or drug concentration need not be so limited. It is significant, however, that for aminophylline tested in man, Rassing (*77*), using 500 mg in ~ 2 g, usually found t_{max} values to be between 5 and 9 hours, while we (*11*) usually found such values to be between 2 and 5 hours—360 mg drug in 3 g cocoa butter was used in both investigations. It therefore seems better not to vary the drug concentration in suspension suppositories at random but, until further evidence becomes available, [e.g., Moolenaar *et al. Pharm. Weekbl. Sci. Ed.* **1**, 25–30 and 89–95 (1979)] to standardize this parameter. It might be of importance especially for high dose drugs, such as aminophylline or analgesics.

The last property in Table IV is the pK_a, which is not influenced very easily. This is also true of several other molecular properties, such as molecular weight and partitioning behavior. As molecular handling easily modifies the therapeutic properties, this type of handling is excluded as a tool in the design of delivery forms. An exception must be made for "handling" in the sense of prodrugs. The general aspects of pK_a in relation to absorption are discussed in Section II,B.

3. Factors Related to the Vehicle

For optimal release of the drug substance, it is essential that the suppository melts rapidly and spreads over the rectum wall. Thus, the following parameters, which are primarily determined by the *chemical composition,* have to be taken into account when judging the quality of a vehicle: (1) the melting range and the melting time at 37° C, (2) the rheological behavior of the melted or softened mass at 37° C, and (3) the properties of the interface with the rectal fluid. By Pharmacopoeial specifications, at best the melting range and the melting time are kept under control, and sometimes a determination of the hydroxyl value, indicating the presence of mono- and diglycerides, is prescribed.

A test for the rheological behavior of the vehicle or of the suppositories is not described in any Pharmacopoeia, notwithstanding its evident importance. Hennig (*33*) proposed a test that takes account, to some extent, of the rheology by measuring the melting time in a tube which is slightly compressed; Krówczynski (*59*) determined the drop point from a narrow tube and the time required to pass this capillary while applying a load on the mass.

The *rheological behavior* is influenced by the presence of suspended particles either of active or of auxiliary substances. Ober (*70*) demonstrated in 1958, for suspensions of procain penicillin, its importance for the release from depôt injections, and this may well be anticipated for suppositories. Moës (*60, 61*) made rheograms of pure vehicles and of suppositories containing 500 mg paracetamol and often found a plastic

behavior, implying a decreasing viscosity with increasing load. In some cases a slight thixotropy was also noticed. His results were obtained by melting the system to be measured within the rheometer cup instead of pouring the already melted mass into the cup. This technique has to be considered seriously since results may differ with the method used. This author relates the differences in amounts excreted in the urine to the static yield value of the different suppositories and comes up with a value of ~ 30 N/m^2, which should not be exceeded. As the pressure in the rectum exceeds this value by at least 100 times (*16a, 27, 60a*), the question is raised as to whether or not the viscosity at the pertinent shearing stress in the rectum is the determining factor, rather than the yield value.

Even when the melted mass behaves like a Newtonian liquid, which is the case for completely melted fats and oily deaggregated suspensions of low concentration, the viscosity is of paramount influence on the release. This was demonstrated *in vitro* by measuring the release rate of, for example, sodium chloride from masses with different viscosities obtained by a variation of the temperature (Fig. 5).

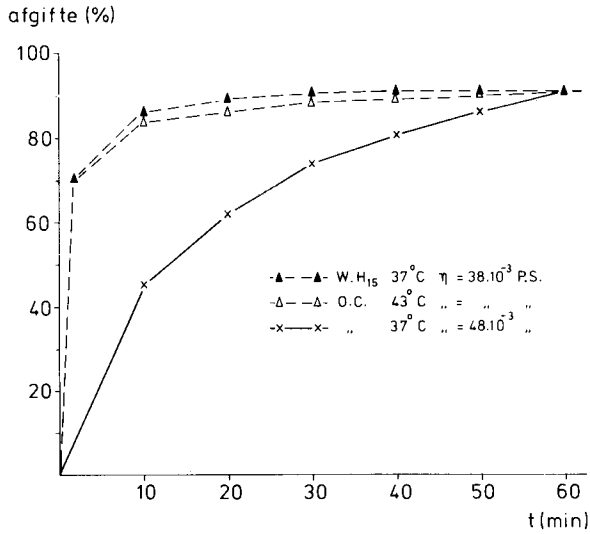


Fig. 5. *In vitro* release of sodium chloride from Witepsol (-▲-) and cocoa butter (-Δ- and -x-) at varying temperature. From (*80*).

From isoviscous masses, the release rates show no differences. At higher viscosity, e.g., for cocoa butter at 37°C, the release is much slower. These values were obtained by melting the masses beforehand; otherwise, the results may be misleading since the melting times are completely

different and do not necessarily correlate with differences in final viscosity. Therefore, without pertinent details, the explanation for some controversial results reported in the literature is difficult. Voigt c.s. (*25, 91a*) found a decreasing release when the viscosity of the mass was higher because of the use of specific additions, in particular when the drug substances to be tested were readily soluble in water. The addition of bentonite did influence the viscosity, but not the release rate, and since this mixture is thixotropic, an interpretation is impossible. Schoonen *et al.* (*83*) increased the viscosity by the addition of fumed SiO_2, and release was almost inhibited for water-soluble salts whereas the free acids (salicylic acid, phenobarbital) were still released, although very slowly. According to the authors, this is by virtue of the higher solubility in fat, so that the release takes place at least partially by dissolution in the base, followed by diffusion.

Baichwal and Lohit (*3*) reported that the release diminished proportionally with the logarithm of the viscosity. However, in this study viscosity was modified by addition of beeswax, nut oil etc., and thus inevitably the melting range was changed, with the direction depending on the concentration of these additives. Others (*68*) could find no difference in the release of sodium salicylate from microenemas when oils of different viscosity were used or when aqueous solutions, which were thickened with methylcellulose in varying concentrations, were administered.

In conclusion, more detailed studies are required to establish the influence of viscosity on the release. But it can be safely stated that for a fatty suppository base, a low viscosity, at 37°C and under the load present physiologically, is to be prefered.

Another complication in the interpretation of the viscosity studies is that the *interfacial tension* between the melted vehicle and the rectal fluid might be changed simultaneously by the additives. Some authors have tried to collect data on the spreading behavior of suppositories in relation to their hydrophilic–lipophilic balance (HLB) (*22*). The results are not consistent, and much depends on the correct choice of the subphase on which spreading is to occur. As said earlier, no data enabling such a choice are available. Neither HLB values nor equilibrium spreading pressures have offered a key to spreading behavior, and will probably be overruled anyway by the interrelation of the viscosity and the pressure exerted by the rectum wall.

The spreading of suppositories directly determines the area through which release from the delivery form, and thus also absorption, can take place, and is thus of obvious importance. Several attempts have been made to obtain data on *in vivo* spreading, e.g., Hennig (*32*) and Quevauviller (*76*) who have used radiopaque materials suspended in a suppository.

From their results they concluded that, soon after melting, the suppository is distributed throughout the rectum and perhaps also the first part of the colon. It was also concluded (*32*) that cocoa butter spreads further than the semisynthetic vehicle, Witepsol H15. Moreover, $BaSO_4$ was found back in the lower parts of the rectum in contrast to diiodone, which was distributed more evenly. The question remained, however, as to what happened with the vehicle itself and whether a suspended drug would remain homogeneously in the vehicle during spreading.

In our laboratory, we tried to answer this question by using two techniques. Qualitative and semiquantitative information was obtained from whole body coupes of a rat, showing fluorescence where the vehicle appeared (Fig. 6), and/or radioactivity when $Ba^{14}CO_3$ had been used. The

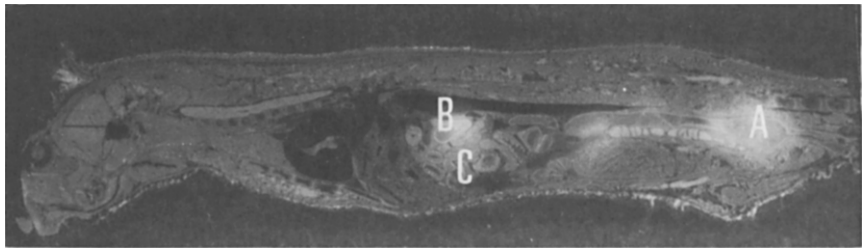


Fig. 6. Whole body coupe of a rat. Bright spots indicate the presence of suppository vehicle. A = insertion spot, B = colon close to small intestine, C = small intestine. From (*80*).

results are summarized in Fig. 7. Suspended colored compounds, such as methylene blue, could be detected visually. Quantitative information was obtained by subdividing the complete intestinal tract and analyzing its contents by combustion and subsequent liquid scintillation counting of the labeled tracers that were incorporated in the suppository. The same type of experiment was performed using lithium salts in suspension suppositories. As variables, particle size and particle concentration were also investigated. Rutten-Kingma, in her thesis (*80*), has thus formulated the following conclusions:

1. Cocoa butter spread somewhat less than Witepsol H5, which is consistent with the viscosities at 37°C (which are 48.10^{-3} and 33.10^{-3} Pa s, respectively) but not in agreement with Hennig (*32*); the latter did not follow the vehicle separately.
2. Suspended particles did not influence the spreading of the vehicle.
3. The spreading was completed within 15 minutes.
4. Drugs having densities of ~ 1.2–1.5 g/cm^3 were dragged along with

the vehicle completely; the same was true for lithium sulfate (density 2.2 g/cm^3), but not for barium carbonate (density 4.3 g/cm^3).

5. Little difference was observed between particle sizes of $< 30\ \mu$m and $\sim 100\ \mu$m and also between concentrations of 5 and 20%.

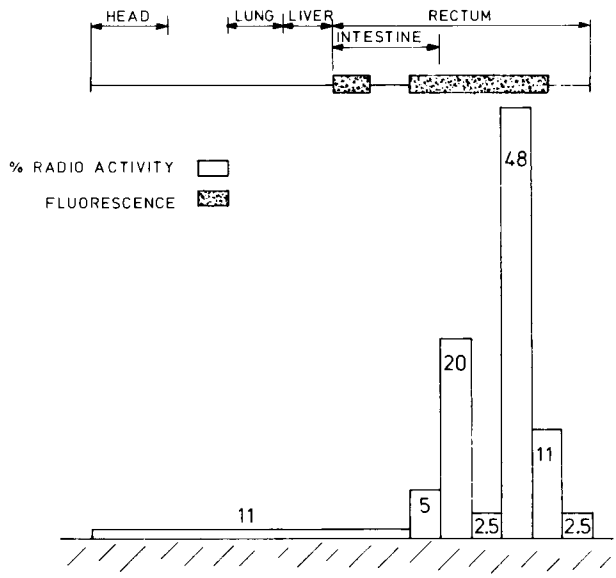


Fig. 7. Schematic representation of the distribution of vehicle (fluorescence) and $Ba^{14}CO_3$ (columns) through rectum/colon of a rat, 30 minutes after administration.

This suggests that the spreading is finished before much release from the vehicle has taken place. No definitive conclusions are possible until more results become available, especially since the results mentioned were obtained in rats.

The question remains, however, as to whether spreading is always an advantage. From the point of view of the enlarged area available for release and absorption, it certainly is; but from the point of view of the first pass effect, it need not always be so.

Many attempts have been made to improve the performance of suppositories by varying the composition of the vehicle. For water-soluble PEG vehicles, the influence of molecular weight has not always been consistent. Prednisolone release from higher-molecular-weight vehicles was decreased (*50*); salicylate release was reduced as well (*75*). Kakemi *et al.* (*44*) found an inverse relationship between the dielectric constant and the release rate of sulfonamides. Pagay *et al.* (*72*) could relate the

release of paracetamol from different mixtures of PEGs to solubility in the vehicle.

Fatty vehicles have been modified many times by the addition of surface active agents, but with varying success and often with conflicting results (*46, 60, 74, 77a, 91*). At present therefore, it seems impossible to draw useful generally valid conclusions in this area.

III. Vaginal Delivery Forms

In medical practice, vaginal administration of drugs is limited almost exclusively to the treatment of local disturbances, such as *fluor albus* and other infections. Thus amoebicides, sulfonamides, antibiotics, and disinfectants are frequently prescribed, sometimes combined with local anesthetics and astringents. Estrogens are also given to restore the vaginal mucosa. In contraception, spermicidal compounds have been applied. And recently, prostaglandins as abortion-inducing compounds have also been administered by this or the rectal route (*16, 89*). It remains doubtful that there is much use for this route with respect to drugs that act systemically.

In contrast to other sites of topical drug administration, the vaginal wall is very well suited to absorption, since it contains a dense network of blood vessels. For most topically intended treatments this is a serious disadvantage, since it may result in a loss of drug from the target site and in its appearance in the general circulation where it may be toxic systemically.

The design of a delivery form starts with the already difficult choice of the form itself, i.e., powders, creams, gels, solutions, capsules, tablets, and suppositories. This is greatly complicated by the fact that it is almost impossible, as it is for all topical delivery forms, to establish bioavailability at the site of action. It is therefore not even certain yet whether there even exists a biopharmaceutical problem. It may also well be that in many cases overdosing, undesirable per se, has taken place to obtain an effect. Underdosing, or insufficient *in situ* availability, is also undesirable, especially with antibiotics.

The information available today is grossly qualitative in nature and permits little more than reasoning as to a likely answer to these question (*78*).

With regard to the choice of delivery form, tablets, suppositories, creams, and in some cases solutions seem to be the best choices.

Tablets should not disintegrate since this would only result in loss of fragments with the mostly present vaginal secretion. Some loss seems

unavoidable, however. This delivery form needs an auxiliary application device in most cases. The design should then take place along the same lines as for buccal tablets, i.e., ensuring regular release rapid enough to permit the build up of an active concentration, but slow enough to avoid rapid loss from the target.

Suppositories are most often prepared with a glycerin–gelatin vehicle, since polyethyleneglycols provoke irritation. Fatty vehicles are not in much use either. The factors in design of the delivery form could well be the same as those mentioned in Table II (Section II,E,1) for rectal suppositories. This delivery form can be administered rather easily.

Creams (*77b*) are applied with the aid of an auxiliary device and have the same advantage as suppositories, i.e., the vehicle helps retain the drug at the site of action.

Solutions, used traditionally by irrigation, could well have a function when applied drenched in tampons. Dosing of the drug and/or adsorption to the carrier material should be kept under close control in that case, however.

An especially interesting point arises in the choice of the drug substance itself, which must not only be delivered to the target tissue but remain there long enough to be effective. This will require sufficient solubility and dissolution rates, whereas no absorption should occur. For this reason, the recent work on vaginal perfusion is of particular interest (*37, 39, 96*). The authors have shown that vaginal membrane transport obeys first-order kinetics for the compounds studied, i.e., aliphatic alcohols and alkanoic acids. The membrane behaved as an aqueous diffusion layer in series with two parallel routes, i.e., a lipoidal pathway and an aqueous pore pathway. This is not different from other absorption sites studied earlier, which means that un-ionized species will be primarily absorbed.

Drugs that are to remain unabsorbed should perhaps be ones that remain ionized at the physiological pH, such as quaternary ammonium compounds. Simultaneously, these authors concluded that the pH at the surface of the absorbing membrane follows the luminal pH very closely, and that negligible amounts of titratable species were secreted. This is in contrast with certain findings for other absorption sites (*20, 24*), but it could perhaps open the possibility of maintaining drug molecules in ionized form and thus limit their absorption. It remains to be seen in that case whether or not this affects the efficacy of the molecule in attacking, for example, the bacterial wall. It should be noticed, however, that at times the buffer species themselves influenced the rate of membrane transport, either promoting it (citrate and tromethamine) or decreasing it (borax). Others buffers left it unaffected (phosphate and phtalate).

As these results refer to rabbits, more data will be required on other

species, including ones that have a cyclus (*71, 52a*), to permit extrapolation to the human situation.

References

1. A. J. Aguiar, *Drug Inform. Bull.* **3,** 17–26 (1968).
2. S. Agurell, A. Berlin, H. Ferngren, and B. Hellström, *Epilepsia* **16,** 277–283 (1975).
3. M. R. Baichwal and T. V. Lohit, *J. Pharm. Pharmacol.* **22,** 427–432 (1970).
4. G. M. Basart-Westgeest, C. J. de Blaey, J. J. Kingma, and R. F. Lammens, *Pharm. Weekbl.* **109,** 738–740 (1974).
5. E. Bechgaard, *Acta Pharmacol. Toxicol.* **33,** 123–128 and 129–137 (1973).
6. K. B. M. Bevernage, and J. Polderman, *Pharm. Weekbl.* **108,** 429–438 (1973).
7. K. B. M. Bevernage, E. L. Noach, and J. Polderman, *Abstracts F.I.P.—Congress, Stockholm,* p. 184, 1973.
8. V. P. Bhavnagri, and P. Speiser, *Pharm. Acta Helv.* **51,** 10–18 (1976).
9. W. Bitterman, J. Spencer, K. A. Huizenga, and R. G. Shorter, *Gastroenterology* **53,** 288–290 (1967).
10. C. J. de Blaey, and J. J. Rutten-Kingma, *Pharm. Acta Helv.* **51,** 186–192 (1976) and **52,** 11–14 (1977).
11. C. J. de Blaey, and A. G. de Boer, *Pharm. Weekbl.* **111,** 1216–1221 (1976).
12. A. G. de Boer, J. M. Gubbens-Stibbe, G. H. P. de Koning, and D. D. Breimer, *Abstracts F.I.P.-Congress, The Hague,* p. 86, 1977, and A. G. de Boer, Ph.D. Thesis, Leiden, 1979.
13. D. D. Breimer, H. L. M. Cox, and J. M. van Rossum, *Pharm. Weekbl.* **108,** 1101–1110 (1973).
14. K. Bucher, *Helv. Physiol. Acta* **6,** 521–525 (1948).
15. F. H. Buckwalter, and H. L. Dickison, *J. Am. Pharm. Assoc.* **47,** 661–666 (1958).
16. M. Bygdeman, A. Ganguli, K. Kinoshita, V. Lundstrom, K. Green *et al., Contraception* **15,** 129–141 (1977).

16a. A. M. Connell, *Gut* **2,** 175–186 (1961).

17. H. L. M. Cox, and D. D. Breimer, *Abstracts F.I.P.-Congress, Stockholm,* p. 186, 1973 (see also ref. 10).
18. D. J. A. Crommelin, and C. J. de Blaey, *Abstracts F.I.P.-Congress, The Hague,* p. 58, 1977, and *Int. J. Pharm.,* submitted.
19. D. J. A. Crommelin, C. Mol, and C. J. de Blaey, *Abstracts F.I.P.-Congress, The Hague,* p. 66, 1977.
20. D. J. A. Crommelin, J. Modderkolk, and C. J. de Blaey, *Int. J. Pharm.,* submitted.
21. W. G. Crouthamel, G. H. Tan, L. W. Dittert, and J. T. Doluisio, *J. Pharm. Sci.* **60,** 1160–1163 (1971).
22. H. von Czetsch-Lindenwald, M. M. Motawi, and M. M. Abdel Khalek, (1960). *Sci. Pharm.* **28,** 270–280 (1960).
23. F. M. Delfs, and J. Kuhne, *Arzneim. Forsch.* **13,** 304–306 (1963).
24. K. Desai, *Ph.D. Thesis,* Univ. of Michigan, Ann Arbor, 1975.
25. G. Falk, and R. Voigt, *Pharm. Zentralhalle* **105,** 573–581 (1966).
26. J. H. Fincher, D. N. Entekin, and C. W. Hartman, *J. Pharm. Sci.* **55,** 23–28 (1966).
27. B. Frenckner, *Gut* **16,** 638–644 (1975).
28. W. E. Hamlin, E. Nelson, A. M. Ballard, and J. G. Wagner, *J. Pharm. Sci.* **51,** 432–435 (1962).
29. J. Halls, *Proc. Roy. Soc. Med.* **58,** 859–860 (1964).
30. H. Hammerl, *Wiener Ztschr. Inn. Med.* **43,** 15–18 (1962).

31. W. Hennig, and G. Severin, *Fortschr. Geb. Röntgenstr. Nuclearmed.* **86,** 251–253 (1957).
32. W. Hennig, "Über die rektale Resorption von Medikamenten." Juris Verlag, Zürich, 1959.
33. W. Hennig, *Pharm. Ind.* **33,** 537 (1971).
34. T. Higuchi, *J. Soc. Cosm. Chem.* **11,** 85–97 (1960).
35. C. A. M. Hogben, D. J. Tocco, B. B. Brodie, and L. S. Schanker, *J. Pharmacol. Exptl. Therap.* **125,** 275–282 (1959).
36. L. C. Hoskins, and N. Zamcheck, *Ann. N.Y. Acad. Sci.* **106,** 767–774 (1963).
37. S. Hwang, E. Owada, T. Yotsuanagi, L. Suhardja, N. F. H. Ho, G. L. Flynn, and W. I. Higuchi, *J. Pharm. Sci.* **65,** 1574–1578 (1976).
38. S. Hwang, E. Owada, L. Suhardja, N. F. H. Ho, G. L. Flynn, and W. I. Higuchi, *J. Pharm. Sci.* **66,** 778–781 (1977).
39. S. Hwang, E. Owada, L. Suhardja, N. F. H. Ho, G. L. Flynn, and W. I. Higuchi, *J. Pharm. Sci.* **66,** 781–785 (1977).
40. M. Höbel, and M. Talebian, *Arzneim.-Forsch.* **10,** 8–12 (1960).
41. M. J. Jaycock, *In* "Dispersion of Powders in Liquids" (G. D. Parfitt, ed.), pp. 44–85, Appl. Sci. Publ. Ltd., Barking, Essex, England, 1973.
42. J. H. G. Jonkman, "Thiazinamium Methylsulphate, Bioanalysis and Pharmacokinetics," Ph.D. Thesis, Groningen, 1977.
43. H. Junginger, *D.A.Z.* **116,** 1880–1883 (1976).
44. K. Kakemi, T. Arita, and S. Muranishi, *Chem. Pharm. Bull.* **13,** 969–975 (1965).
45. K. Kakemi, T. Arita, R. Hori, and R. Konishi, *Chem. Pharm. Bull.* **15,** 1883–1887 (1967).
46. K. Kakemi, H. Sezaki, S. Muranishi, and H. Matsui, *Chem. Pharm. Bull.* **15,** 172–178 (1967).
47. K. Kakemi, T. Arita, R. Hori, R. Konishi, and K. Nishimura, *Chem. Pharm. Bull.* **17,** 248–254 (1969).
48. K. Kakemi, T. Arita, R. Hori, R. Konishi, K. Nishimura, and H. Matsui, *Chem. Pharm. Bull.* **17,** 255–261 (1969).
49. S. A. Kaplan, *In* "Current Concepts in the Pharmaceutical Sciences: Dosage Form Design and Bioavailability" (J. Swarbrick, ed.), pp. 1–30, Lea and Febiger, Philadelphia, 1973.
50. I. W. Kellaway, and C. Marriott, *J. Pharm. Sci.* **64,** 1162–1166 (1975).
51. H. P. M. Kerckhoffs, and T. Huizinga, *Pharm. Weekbl.* **102,** 1183–1200 (1967).
52. H. Kilian, "Participation of the Lymphatic System in the Absorption of Pharmaca from the Gut." Ph.D. Thesis, Nijmegen, 1973.
52a. K. T. Kirton, T. J. Roseman, and A. D. Forbes, *Contraception* **8,** 561 (1973).
53. F. U. Knudsen, *Acta Paedr. Scand.* **66,** 563–567 (1977).
54. F. Köchel, *D.A.Z.* **94,** 821–824 (1954).
55. S. S. Kornblum, and J. O. Hirschom, *J. Pharm. Sci.* **59,** 606–609 (1970).
56. M. Kraml., J. Duboc, and D. Beall, *Can. J. Biochem. Physiol.* **44,** 1449–1451 (1962).
57. V. E. Krogerus, and M. Tolvi, *Acta Pharm. Suec.* **2,** 327–344 (1965).
58. V. E. Krogerus, R. Lahti, J. Tervola, and L. Utriainen, *Farm. Aikak.* **80,** 98–105 (1971).
59. L. Krowczynski, *Diss. Pharm.* p. 259 (1959) (see also *10*).
60. G. Mezey, *Pharmazie* **26,** 166–169 (1971).
60a. J. J. Misiewicz, *Gut* **16,** 311–314 (1975).
61. A. Moës, *J. Pharm. Belg.* **29,** 113–125 and 319–322 (1974).
62. A. Moës, and F. Jaminet, *Pharm. Acta Helv.* **51,** 119–125 (1976).
63. E. Nelson, *J. Am. Pharm. Assoc. Sci. Ed.* **47,** 297–299 (1958).

64. F. H. Netter, *In* "The Ciba Collection of Medical Illustrations" (E. Oppenheimer, ed.), Vol. 3, Part II, Ciba, Summit, N.J., 1969.
65. F. Neuwald, J. Kuhne, and K. Soehring, *Pharm. Ztg.* **106,** 1544–1547 (1961).
66. F. Neuwald, and F. Kunze, *Arzneim.-Forsch.* **14,** 1029–1037 and 1162–1167 (1964).
67. F. Neuwald, and P. Ackad, *Am. J. Hosp. Pharm.* **23,** 347–351 (1966).
68. F. Neuwald, and P. Ackad, *Galenica Acta* **19,** 179–189 (1966).
69. H. Nogami, M. Hanano, and J. Watanabe, *Chem. Pharm. Bull.* **10,** 1161–1167 (1962).
70. S. S. Ober, H. C. Vincent, D. F. Simon, and K. J. Frederick, *J. Am. Pharm. Assoc. Sci. Ed.* **47,** 667–676 (1958).
71. E. Owada, C. R. Behl, S. S. Hwang, L. Suhardja, G. L. Flynn, N. F. H. Ho, and W. I. Higuchi, *J. Pharm. Sci.* **66,** 216–219 (1977).
72. S. N. Pagay, R. I. Poust, and J. L. Colaizzi, *J. Pharm. Sci.* **63,** 44–47 (1974).
73. G. D. Parfitt, *In* "Dispersion of Powders in Liquids" (G. D. Parfitt, ed.) pp. 1–43, Appl. Sci. Publ. Ltd., Barking, Essex, England, 1973.
74. J. M. Plaxco, C. B. Free, and C. R. Rowland, *J. Pharm. Sci.* **56,** 809–814 (1967).
75. H. W. Puffer, and W. J. Crowell, *J. Pharm. Sci.* **62,** 242–245 (1973).
76. A. Quevauviller, and Y. Jund, *Ann. Pharm. Franc.* **9,** 593–608 (1951).
77. M. Rassing, "Laegemiddelformulering og rektal absorption," Ph.D. Thesis, Kopenhagen, 1976.
77a. S. Riegelman, and W. Crowell, *J. Pharm. Sci.* **47,** 123–133 (1958).
77b. L. A. Rigg, H. Hermann, and S. S. C. Yen, *Med. Intell.* **298,** 195–196 (1978).
78. W. A. Ritschel, "Angewandte Biopharmazie." Wiss. Verlagsgesellschaft, Stuttgart, 1973.
79. R. Roscoe, *In* "Flow Properties of Disperse Systems" (J. J. Hermans, ed.), North-Holland Publ., Amsterdam, 1953.
80. J. J. Rutten-Kingma, "Biofarmaceutisch onderzoek van suspensiezetpillen," Ph.D. Thesis, Leiden, 1973, also *Int. J. Pharm.* **2,** 39–53, 179–186, and 187–194 (1979).
81. U. Sammelius, and A. Aström, *Acta Pharmacol. Toxicol.* **14,** 240–250 (1958).
82. T. W. Schwarz, and A. Bichsel, *Pharm. Acta Helv.* **38,** 861–870 (1963).
83. A. J. M. Schoonen, F. Moolenaar, E. Haverschmidt, and T. Huizinga, *Pharm. Weekbl.* **111,** 585–589 (1976).
84. A. J. M. Schoonen, G. W. de Vries-Nijboer, and T. Huizinga, *Abstracts F.I.P.-Congress, The Hague,* p. 54, 1977, also *J. Pharm. Sci* **68,** 163–168 (1979).
85. N. Senior, *In* "Advances in Pharmaceutical Sciences" (H. S. Bean *et al.*, eds.), Vol. IV, pp. 363–437, Academic Press, New York, 1974.
86. A. A. Sinkula, and S. H. Yalkowski, *J. Pharm. Sci.* **64,** 181–210 (1975).
87. V. F. Smolen, *J. Pharm. Sci.* **62,** 77–79 (1973).
88. K. H. Soergel, and F. Ingelfinger, *Gastroenterology* **47,** 610–616 (1964).
89. P. G. Stubbefield, *Contraception* **15,** 175–181 (1977).
90. W. H. Thomas, and R. McCormack, *J. Pharm. Pharmacol.* **23,** 490–494 (1971).
91. L. Turakka, and V. E. Krogerus, *Farm. Aikak.* **83,** 59–70 (1974).
91a. R. Voigt and G. Falk, Pharmazie **23,** 709–714 (1968).
92. J. Wellauer, M. del Buono, A. M. Ory, and K. Steiger-Trippi, *Pharm. Acta Helv.* **35,** 619–631 (1960).
93. O. Weiss-Fogh, *Arch. Pharm. Chemi* **68,** 885 (1961).
94. A. Widman, *Pharm. Ind.* **22,** 348–352 (1960).
95. C. Wolf, and R. Voigt, *Pharmazie* **23,** 345–353 (1968).
96. T. Yotsuanagi, A. Molokhia, S. Hwang, N. F. H. Ho, G. L. Flynn, and W. I. Higuschi, *J. Pharm. Sci.* **64,** 71–76 (1975).

Chapter 6 Interactive Graphics in Medicinal Chemistry

Edgar F. Meyer, Jr.

ISBN 0-12-060309-8

I. Introduction

The convergence of information regarding structure–function relationships promises to open new frontiers in drug design. The interaction of biological molecules is well understood in chemical terms: The making or breaking of a chemical bond may be an object of wonder but it is not magic. Yet, without the help of facile models, the results of structural studies (geometric parameters) are beyond the perceptions of most scientists. This chapter describes the use of interactive computer graphics as an emerging tool for modeling the complex stereochemical interactions of biological molecules.

As the emphasis here is on numeric techniques in general and the manipulation and display of crystallographic data in particular, the question occasionally heard among chemists "Why do I need a computer?" applies. Indeed, three recent syntheses based upon known or hypothesized receptor geometries (cf. Section VI) were made without the aid of computer graphics. Rather, this chapter is intended to highlight existing and emerging methods that may serve the needs of a variety of users and, quite generally, that may address structure–activity questions of a growing number of molecular systems, thus bringing a large body of knowledge within the scope of the medicinal chemist.

Even as a variety of quantitative, as well as chemical information, techniques converge with the aid of the computer, so the computer can be taught to "speak," and understand a graphical language more suitable to our perceptions. Specifically, the common principles that recur among the ca. 50 macromolecules studied thus far by X-ray crystallography may serve as a useful guide in seeking to visualize the receptor site geometry ("lock") and its interaction with a variety of substrates ("keys"). Then, in the near future, as the three-dimensional (3-D) architecture of additional protein receptor site–substrate complexes is elucidated, modeling techniques will be available to permit the medicinal chemist to develop heuristics for studying the interaction of analogs, some of which may have been synthesized "ad hoc" in the computer.

To see these molecules in 3-D, the chemist needs help (*78*). Mechanical models have numerous limitations. Current computer technology offers the powers of computer graphics and may be joined with computational

chemistry and chemical information techniques to extend an emerging field of structural chemistry.

No extensive reference will be made to the use of other numeric methods (e.g., pattern recognition, molecular orbital calculations, multiple regression analysis) currently receiving attention, other than to observe that they are representative of yet other tools available—the emphasis here is one of adding the sense of "sight" to a variety of "blind" tools.

Finally, it is the sharpened perceptions of the investigator that will derive from the methods the insights needed to suggest candidate analogs for subsequent synthesis and/or testing, not the methods themselves.

II. Applications of Interactive Graphics

When a flexible software system is linked to an available graphics display, several modeling operations may be greatly facilitated.

1. One may evaluate intermediate and final results of force field calculations (*21*) of conformations of pharmacophores, especially in the environment of known or putative receptor sites. Although the choice of force field constants and refinement methods are current limitations to this method, encouraging results have been obtained (*47*) for both small and large molecules.
2. Simply by viewing a model of a molecule in 3-D, spatial relationships and stereochemical properties become apparent: "A picture *is* worth a thousand words." It is easy to predict that as chemists begin studying complex biological macromolecules, diagrams heretofore bound to the blackboard or printed page will prove inadequate; a new, three-dimensional language natural to our perceptions is needed. Perhaps these efforts and illustrations may be among the first words in this new language, which would be only as extension to the century-old chemical language of van't Hoff and LeBel.
3. We may wish to study the molecular geometry of macromolecular receptor sites, either as rigid models or with interactively guided internal rotations. For example, the active site models of a homologous series like the serine proteinases or the dehydrogenases can be superimposed for steric comparisons, an operation impossible with mechanical models. Likewise, a series of homologous pharmacophores can be superimposed for structural comparisons. The pattern recognition capabilities of the trained eye could leap over the rigid logic of computer algorithms and give the intuitive powers of

the chemist full play. Direct read-out of molecular geometry is made possible in the interactive mode; questions of binding geometry may be answered as they are considered.

4. Finally, an appropriate use of molecular graphics is the controlled fitting of: structure to structure (*75*); model to electron density (*4, 54, 66, 68*); molecules into a crystalline lattice (with symmetry supplied) (*32*); a molecule within a clathrate solvent cage (*35a*); substrate, activated complex, or inhibitor to an active site of an enzyme; a drug to a receptor; and conformationally induced biological information transfer from macromolecule to macromolecule. Although a good modeling system would faithfully reproduce the biological interactions dynamically, the simple modeling of rigid entities is surely the first step in the conceptualization of these processes. Interactive manipulation, with internal degrees of freedom, allows subgroup flexibility to mimic more complex interactions. The trained eye of the investigator receives the 3-D output from the computer, and various graphical input devices let the user respond in kind so that the loop is closed for immediate response. In this sense, interactive graphics is fun; it can also be aesthetic.

III. Hardware and Software

Two trends in computer hardware have become manifest: Large computers are becoming larger and small computers are becoming larger. The promised availability of large, coordinated computer facilities, like the National Resource Center for Computational Chemistry (*79*) in the U.S., should help rationalize the developmental efforts for exceedingly complex modeling programs [MNDO (*18*), molecular dynamics calculations (*47*), large data base searching (*76*), etc.]. Currently, the proliferation of minicomputers of several colors makes it more likely that a chemist will interact with a laboratory computer (which, however, may then be linked to a maxi-computer in a computer network). Even desk calculators are becoming more "intelligent," so that the chemist will have, or even now has, a remarkable amount of computational power at his fingertips. In retrospect, the typical laboratory computer now has the power of the university computer of ten years ago, at a fraction of the initial investment in hardware (electronics, peripherals, etc.). The increasing demands of ever more complex tasks, however, have required the development of programs (software) that are both "user-friendly" and capable of faithful modeling of chemical systems. The challenge to the structural chemist then is to develop programming tools to use this advanced hardware.

A. COMPUTER LANGUAGES

The "lingua franca" of computational chemistry is FORTRAN, a language adequate for algebraic problems but all too lacking in flexibility for logical operations and especially for string processing (where a string may be a series of numbers and/or alphabetic characters, words, chemical descriptors, etc.). More expressive languages like ALGOL (*62*), PL/1 (*80*), PASCAL (*77*), or SAIL (*70*) are not as widely used so that a program in a specialized language may not be understood by a neighboring computer. Some help is afforded by a recent preprocessor, RATFOR (*40*) or *rat*ional *FOR*TRAN, which permits more coherent grouping of instructions, compressed coding, and added expressiveness, especially of logical statements which are the heart of chemical information systems, including graphics.

B. HARDWARE

1. *A Minimal Configuration*

By current standards, the minimum hardware configurations needed for computational chemistry on a laboratory computer are the following: mini-computer with 24 to 32K words of memory; disk operating system that supports FORTRAN; teletype (or alphanumeric CRT) input; and some type of hard-copy output, e.g., printer–plotter. Useful peripherals might also include: additional core and disk memory; industry compatible magnetic tape; and communications interface to a computer network via telephone.

2. *Graphics Hardware*

To the basic system may be added a computer graphics display (*23, 24, 37, 56*). As a working definition, an interactive graphics display should be able to draw a picture before the user has forgotten the questions put to the computer. Thus, plotters are excluded. However, even plotted output is more expressive than printed output in suggesting structural relationships, especially if stereo output [e.g., from programs like ORTEP (*36*), or PLUTO (*55*)] is used.

Until recently a bimodal distribution of graphics equipment has existed. Simple, yet useful, storage displays such as the Tektronix series or rapid vector display terminals (e.g., Hewlett-Packard) are most frequently encountered at the low-cost end of the spectrum. At the other end are powerful displays with "intelligent" picture-drawing capacity (Fig. 1) that allow rapid translation and/or rotation of all or part of the picture, perspective, hidden-line calculations, and even representation of surfaces.

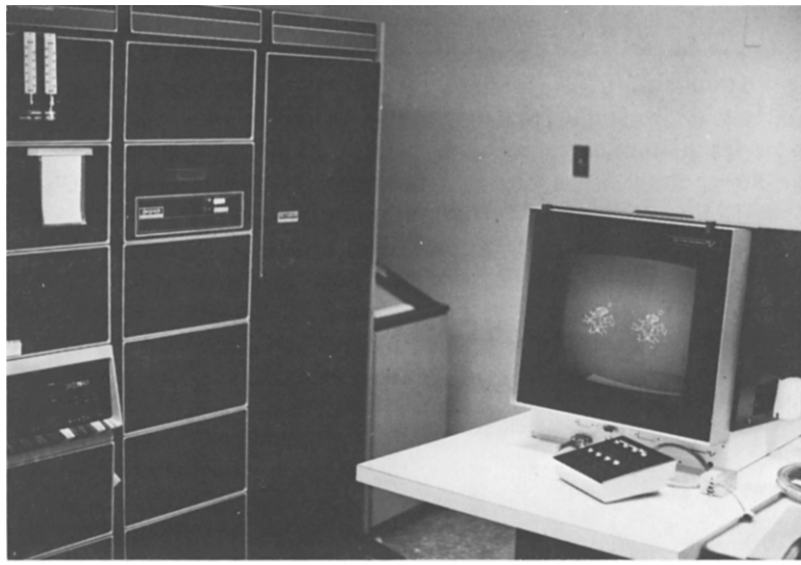


Fig. 1. View of the interactive graphics display in the Biographics Laboratory, Texas A&M University. Primary control of the display is through the 10 control dials.

Typical vendors are Evans and Southerland or Vector General; costs of display hardware may range from $75,000 to $100,000.

But, more than an output device, it is highly essential that the user be able to interact with the display process and control the sequence of events. Although a variety of interactive devices exist (dials, tablets, light pens, joy sticks, etc.), they are all essentially 2-D devices, and for complex 3-D modeling "most pointers are disappointing." Typically, one learns to make do with available devices.

C. CAVEAT EMPTOR

Several caveats may be worth considering in seeking an overview of computer systems as an aid in molecular modeling leading to drug design possibilities.

1. The hardware purchase is only the first expense.
2. Annual maintenance costs typically run ca. 10% of the hardware costs.
3. Software and systems development costs can equal or double hardware costs.
4. A series of programs does not necessarily constitute a usable sys-

tem. Integration and flexibility are highly important for a user-friendly system, especially if it is to be provided as a service or tool for computer-shy chemists.

5. The time required to develop and/or adapt requisite software can exceed several man-years for ambitious systems.
6. Finally, tasks and expectations should be clearly delineated in advance of hardware acquisition and software development. The broader aspects of software management problems have been covered, with perceptive wit, by Brooks (*10*).

Although several reviews have discussed high-performance computer graphics techniques (*23, 37*), one may hope that the marketplace begins to provide more options in medium price–medium performance graphics. Promising candidates are the color raster displays in the $20,000–$25,000 range. Previous work (*48, 50*) has shown this type of display to be capable of modeling complex molecules in an interactive mode. Likewise, the MMS-X (*5*) development is an effort to design hardware specifically for molecular modeling (Fig. 2).

D. MOLECULAR MODELING

1. *Introduction*

In considering the utility of computer graphics techniques and modeling of 3-D images, it is important to realize the distortions frequently encountered in schematic, 2-D diagrams. At the beginning of a survey of the structures of neurohumoral transmitters (biogenic monoamines), the authors give several examples of the confusion that can arise, for example, when diagrams of noradrenaline and serotonin are compared (*14*) (see Fig. 3). Yet, textbook diagrams and sketches on blackboard or paper are the usual, international symbols for depicting molecular conformations and configurations. As the molecules become more complex, a 2-D projection becomes strained to convey the 3-D information available. By comparing the several possible graphical output modes, the user will gain a feeling for the usefulness and inherent limitations of each.

2. *Representations*

Stick or ball-and-stick models are used to represent the bonds and atoms of the molecules; lines represent chemical (usually covalent) bonds and line junctions represent atoms. Such a representation is reasonably natural for both the chemist and computer (*7, 8*) but does lack an expression for structural features such as lone-pair electrons or hard-sphere bulk of individual atoms. Circles may be used to represent atom size or type,

Fig. 2. The controls of the MMS-X molecular graphics system developed by the Computer Systems Laboratory of Washington University, St. Louis and now in use at Purdue University (courtesy of M. Rossmann, Purdue University).

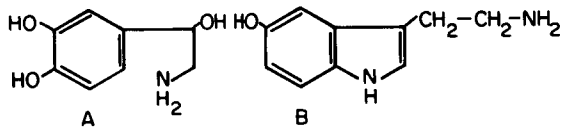


Fig. 3. See text. Reprinted with permission from Carlström (*14*); copyright by Cambridge Univ. Press.

but most commonly the user learns to add atomic bulk onto the important feature of "fuzziness" required to reflect coordinate imprecision.

Stick models are easily generated by vector graphics systems. Simple lines may be rendered relatively transparent on occasions when a view of the "inside" is partially eclipsed by the "outside" of a receptor site. Finally, they are a simple means of modeling the complex reality of rap-

idly bouncing atoms leashed together by springy bonds to form a unique composite. Reality must mirrored in the mind's eye of the beholder.

3. *Stereography*

By supplying an appropriately programmed computer with the Cartesian coordinates of the atoms together with a connectivity table such as the Chemical Abstracts FROM and RING lists (*49, 53*), one may generate a two-dimensional projection of a three-dimensional molecular structure on a graphics display. A second image may be generated by projecting the model on a plane rotated 6° about the vertical axis, and translating the image to the right to achieve a side-by-side stereo representation, such as in Fig. 4. Small, pocket spectacles* facilitate the viewing of such images.

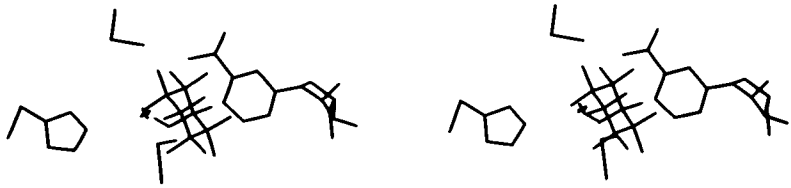


Fig. 4. Stereo view of a model of the coordination sphere of Zn in alcohol dehydrogenase. Amino acid side chains (His-67, Cys-46, and Cys-174) form the base of the trigonal pyramid. The fourth ligand to the Zn atom (marked by a cross) is the O-atom of the substrate, cyclohexanol. It is at the apex of the pyramid pointing towards the viewer. The substrate molecule is also positioned with respect to the NAD^+ cofactor so that the H-atom to be removed from the substrate is directed towards the C-4 atom of NAD^+. Only the nicotinamide ribose part of NAD^+ is shown.

Another common technique, the Ortony box (*58*), involves rotation and reflection to create a second image. One image is reflected, the other projected through a half-silvered mirror with crossed polarizing filters and viewed with polarizing glasses. A third technique (*57*) uses separate colors (i.e., colored images on a color television monitor) to generate two rotated, superimposed images, which are then viewed through narrow-band optical filters (e.g., Wratten filters no. 29 and 61 are favored for television monitor viewing). In all cases, normal, stereooptical vision is achieved by presenting a separate image to each eye. A fourth, quite effective method of generating a 3-D image (the kinetic depth effect) involves continuous rotation or rocking back and forth of the image. Depth perception may be enhanced by (hardware) attenuation of intensity as a

* Obtainable from Hubbard Scientific Co., Northbrook, Illinois. Press-on prisms may be obtained (*5*) from optometrists or MENTOR, Codman & Shurtleff Inc., Randolph, Massachusetts 02368.

function of distance of individual parts of the model from the viewer. Even as one should occasionally check the first three methods to see that the correct optical enantiomorph is created, the kinetic depth effect derived from a rotating image suffers from an optical illusion (the Sinsteden effect) (*29*) that causes the image to appear to reverse direction of rotation with a resulting change in chirality.

When the graphics system is able to create realistic 3-D models, the time needed to draw a single view of a large macromolecule may exceed the flicker perception rate (ca. 1/30 second) of the eye. The image may be too complex for simple viewing or the receptor site may be buried in a crevice of a macromolecule or otherwise visually obscured; hidden line algorithms are not the answer because one frequently wishes to look through parts of the model to keep an holistic view. More expensive computer systems let you zoom in on the region of interest. For more modest systems, programs have been written (e.g., SEARCH) (*51*) that let one anticipate regions of interest and selectively gather all atoms (amino acids, etc.) within a specified distance from the center of interest.

Interactive graphics permits user manipulation of all or part of the molecule. Typical pointing techniques employ a flashing bond that moves throughout the molecule under user control. When the data base has been correctly designed, translations and rotations forward or backward of the flashing bond are provided (e.g., torsional rotation about the C_α–C_β bond of amino acid side groups). 3-D pointers (like a child's toy, the Etch-a-Sketch®) are occasionally used (cf. Fig. 5) to indicate a point of interest for calculations of molecular geometry.

When the display is composed of several files (pictures), subgroup manipulations may be applied independently. This is a useful technique for fitting a substrate into a receptor site. Likewise, intensity levels may be varied for different picture elements (e.g., side chains displayed or removed while the protein backbone is continuously displayed) (cf. Figs. 7–10).

4. *Program Examples*

A computer program, FIT (*15, 42, 54*), was written for the PDP11/40 and 3-D Vector General display in this laboratory (cf. Fig. 1) for the 3-D display of electron-density maps and the fitting of amino acids leading to the construction of a model of a macromolecule based upon crystallographic measurements (*43*). Because a general data structure was chosen, we have also been able to use this program to depict the architecture of the active site of a variety of proteins. The interactive techniques discussed here have permitted us to fit several substrate models interac-

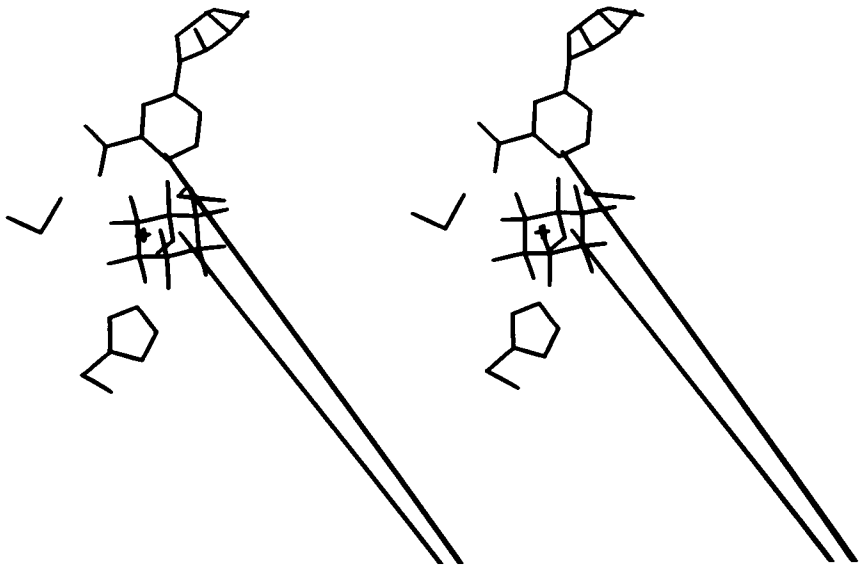


Fig. 5. Stereo view of the model of the key participants in the active site of ADH. The atom Zn is indicated by a +. Only the nicotinamide–ribonucleoside part of NAD^+ is shown. One pointer is directed to the C-4 atom of NAD^+ which receives an H-atom from the substrate, cyclohexanol. The second 3-D pointer indicates the C-1–hydroxyl bond to be oxidized in the ternary complex.

tively. The data base structure chosen is sufficiently general so that non-molecular 3-D data sets (*43*) may be treated in a straightforward manner.

A second program, PIXMOV (*52*), currently permits the user to control the intensities of a maximum of 10 picture files. Moreover, other global options include scaling, orientation and rate of rotation, stereopair generation, and 3-D pointer activation. All these commands are given to the computer through the rational implementation of 10 control dials. Program PIXMOV may be instructed to record as well as execute the commands under operator control. Then, the recorded sequence can be played back. Finally the play-back operation is joined with a photographic option to permit proper exposure of a grey scale range of ca. four f-stop values onto motion picture film. A 16 mm Arriflex® camera is electronically coupled to the computer for this purpose.

Within the modest capacity of this computer, program ENERG (*60*) takes advantage of the digital nature of the model to optimize the non-bonded interactions, hydrogen bonding capabilities, and electrostatic interactions of the visually fit substrate with the entire protein molecule as an off-line calculation.

IV. Data Bases and Information-Retrieval Techniques

Computers are frequently used to record and retrieve bibliographic information (*76*) (which is usually one-dimensional). To this may be added data bases with spectroscopic, chemical, or physiological data. Current sub-structure search systems are essentially two-dimensional but, nevertheless, offer considerable power beyond simple searches based upon chemical names. Emphasis here is placed on storage–retrieval systems that reflect the dimensional complexity of substrate–receptor site interactions, based upon the published results of 3-D X-ray crystallographic determinations of the coordinates of individual atoms (*7*). Thus, we should prefer to search for candidate molecules on the basis of molecular architecture as an enzyme would, with full cognizance of spatial properties.

A. THE CRYSTAL DATA CENTRE

At present the crystallographic method has provided definitive studies of some 15,000 organic (carbon-containing) compounds, as registered by the Cambridge Crystal Data Centre (*38*). This number would pale in comparison to the some four million entries contained in the Chemical Abstracts file, were it not for the high degree of internal consistency in molecular architecture; one need only think back to the size of the data base available to Linus Pauling in the 1930s when his revolutionary "The Nature of the Chemical Bond" (*59*) was conceived.

Searching Techniques

Several techniques exist for searching the literature of structural chemistry. As Gutenberg still reigns, the most widely available method is a manual search of the assembled data in book form (*39*). Computer search methods have been designed by Feldmann, Motherwell, and Villarreal for searching the holdings of the Crystal Data Centre files.

For users with access to a large DECSYSTEM 10 or with a telephone link to the TYMESHARE network, the X-ray System (*76*) provides keyword and substructure input capabilities and bibliographic output in an interactive mode. Several noninteractive graphics systems are supported by the OMNIGRAPH (*65*) routine, which has been incorporated into the X-ray system.

A batch-stream system that offers superior flexibility of query input at the expense of relatively slower file searching speeds has been written by Dr. Motherwell (*55*). As the Cambridge Data Files have been subdivided into bibliographic, structure, and coordinate files, hits obtained from the

search of one file may be linked to entries in another file by means of a unique, six-character descriptor assigned to each molecule.

An ambitious retrieval program, CRYST, written by Villarreal (*72*) for a laboratory mini-computer, is adequate when a specific class of compounds (e.g., amino acids, or steroids, or carbohydrates) is to be searched. However, limited disk storage capacity makes global searches currently impossible. In a few years, the bulk memory capacity of bubble memory devices may make this system much more attractive in a laboratory setting.

B. THE PROTEIN DATA BANK

The Protein Data Bank (*7, 81*) is a computerized depository of macromolecular data, bibliographic but especially structural in nature. A relatively straightforward method of accessing these data is provided by Feldmann's AMSOM microfiche atlas (*25*), which contains bibliographic, geometric, and 3-D representations of all and parts of each entry, for a modest cost. Not all questions have been anticipated and no interactive capacity is possible, but an otherwise overwhelming data base has been brought within easy reach.

1. *SEARCH*

Computer-aided queries of the macromolecular file have thus far been limited to individual entries because of the size of the molecules involved (an average-sized protein has perhaps 1500 atoms). Program SEARCH (*51*) was first written for a large computer (CDC6600) but now runs on our mini-computer. Typical searches (e.g., extraction of a region about a certain amino acid or the active site region of a protein) require several minutes; a complete extraction of the largest macromolecule (ca. 3000 atoms) with no structural constraints other than covalent bonding requires ca. 20 minutes on our PDP 11/40. Symmetry operations may be included to study more complex interactions (cf. Sections VI,A and D,5, esp. Figs. 6 and 10.

2. *Limitations: Contents and Resolution*

For structure–activity studies two critical limitations exist with the current macromolecular data base: contents and resolution. For some types of drug receptor systems, the isolation–purification–crystallization process leading up to a crystallographic determination is largely a matter of emphasis and time. This limitation may be overcome, in part, by the collaboration of the biochemist and the crystallographer. Membrane-bound proteins are not usually amenable to crystallization, however.

The question of resolution has witnessed a happy improvement over the

last several years, thanks to a variety of refinement methods (*33, 41, 82*). In studying small and large molecule interactions, questions of chemical bonding, stereochemistry, and molecular geometry must ultimately be resolved. Yet, most macromolecules whose studies are now published or under investigation do not permit the visualization of individually resolved atoms. When questions of atom shifts of 0.1 Å are essential to the study of substrate–receptor bonding, a resolution of better than 1.5 Å and a precision (of bond lengths) of better than 0.01 Å are required, assuming problems with thermal and statistical disorder can be minimized. The current norm is ca. 2.0 Å resolution. Several experimental techniques, especially low-temperature studies (*61*), plus constrained refinement techniques (*41*), promise to ameliorate our view of biological macromolecules. Indeed, a recent refinement of the room-temperature data of the zymogen, trypsinogen, in the laboratory of Prof. Robert Huber has produced a precision in bond lengths of 0.02 Å. Low-temperature (− 60°C) studies by Dr. Tej Pal Singh (*82*) suggest that the desired precision (0.01 Å) may be reached or exceeded, opening the way for careful receptor + substrate binding studies.

V. Properties of Molecules and Crystals

A. In Motion

The typical biomolecule is awash with solvent or buried in a nonpolar matrix of lipids. Individual atoms of small molecules typically undergo root-mean-square displacements of 0.03 Å with a frequency of 10^{14} oscillations/sec at room temperature (*34*). Summed together, chemical bonds tend to keep the molecule intact but a large degree of flexibility affords a range of energetic parameters on the characteristic Born-Oppenheimer energy surface (*21*). Especially for macromolecules, computer simulations have suggested the degree of floppiness that may be exhibited by biomacromolecules over a short span of time: Molecular dynamics calculation (*47*) of bovine pancreatic trypsin inhibitor simulated atomic motion over 8.8 ps. The time-averaged structural changes obtained are near to, but not identical with, the X-ray determination: Root-mean-square deviations were 1.2 Å for α-carbons and 1.7 Å for all atoms.

The X-ray crystallographic study (*8*) averages the position of molecules over, for example, 10^{12} molecules in all the translationally "identical" unit cells that comprise a crystal. This averaging takes place not over picoseconds but over minutes or days. Thus, until recently, the resulting picture of the average position of the individual atoms had a limited precision. Yet these solid-state structural results have been consistent with our

understanding of molecular interactions (e.g., charge transfer and electrostatic interactions, van der Waals contacts, acid–base: electron–proton transfer interactions) from organic chemistry and have been confirmed in particular cases in solution studies (*45*).

B. CONTACTS

1. *Small Molecules*

An especially compelling story has emerged from the crystallographic investigation of methadone (*12, 13*) which possesses an intramolecular N to C=O interaction. Comparisons with related compounds "have shown that for decreasing N to C distance, the carbonyl C atom is increasingly displaced from the plane of its three ligands toward N and the C—O distance increases" as the incipient N—C bond begins to form, thus giving a geometric evaluation of a nucleophilic addition reaction. A perusal of the structural literature revealed that the contacts of individual atoms in separate molecules were found to range continuously from normal, covalent contacts at the short end to van der Waals contacts at the long end, with concomitant geometrical changes within the molecule as the contact decreases. Thus, not only does an experimental basis for the geometry of individual molecules exist, but a synthesis of these studies reveals the nature of steric modifications forced upon an atom by its environment.

2. *Macromolecules*

A high-resolution crystallographic determination of the trypsin–trypsin inhibitor complex by Huber and colleagues (*35*) shows the close interactions of the catalytic site with the atoms of a scissile peptide bond. A nucleophilic γ-oxygen (from Ser-195) is found to be 2.6 Å from the peptide carbonyl carbon (of the scissile bond), which is displaced from the plane of the atoms related to the peptide bond into a position intermediate between trigonal and tetrahedral conformations. A "good" substrate would permit the trigonal carbonyl C-atom to assume a fully tetrahedral conformation with a concomitant bond rupture leading to proteolysis.

VI. Molecular Interactions

Although the purpose of this chapter is to show the powers and potentials of interactive computer graphic modeling, several drug design efforts have produced significant results without the aid of interactive graphics.

A. INDUCING OXYGEN RELEASE IN HEMOGLOBIN

Beddell and co-workers (*27*) used the results of the crystallographic investigation of hemoglobin for the *ab initio* design of effective model oxygen releasing compounds. Arthur Arnone (*2*) had shown that 2,3-diphosphoglycerate (2,3 DPG) binds in a cavity between the β-subunits of the hemoglobin tetramer and promotes the release of bound oxygen. Beddell's group recognized that the hemoglobin–2,3 DPG interaction had many similarities to a drug receptor system. Several substituted bibenzyl compounds were synthesized, tested, and shown to mimic the role of 2,3 DPG.

Although crystallographic evidence for the specific bonding geometry of analogs is not yet available, a comparison with bonding of bis-aryl-hydroxysulfonic acid to horse and human hemoglobins produced the expected correlation of structure to function (O_2 partial pressure) (cf. Fig. 6).

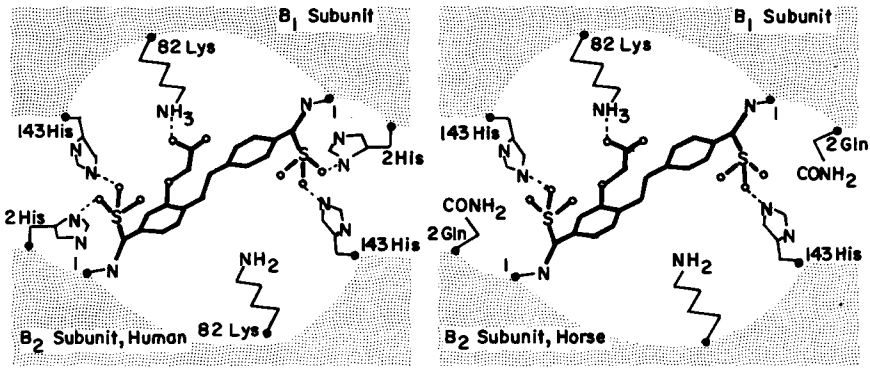


Fig. 6. Schematic diagram of the 2,3 DPG receptor site. The hydroxysulfonate-substituted analog of 2,3 DPG was predicted to interact favorably with the four His residues of human haemoglobin but with only two of the horse hemoglobin. Observations are compatible with these predictions. Reprinted with permission from Goodford (*27*); copyright by MacMillan.

In the region near the diad axis where 2,3 DPG had been located, human hemoglobin has a His as the second amino acid in the β-subunit; in horse haemoglobin the second amino acid is Gln. Gln does not have an ionizable proton to give in association with the sulfite moiety whereas His does; thus, stronger binding to human hemoglobin was predicted. Oxygen-binding studies correlate with the model and give results similar to the binding of 2,3 DPG. Likewise, support for the proposed mode of binding of these analogs is provided by an NMR study (*11*).

B. INHIBITION OF THE ANGIOTENSIN-INHIBITING ENZYME

By deductive logic, Ondetti and coworkers (*17*) speculated on the structural similarity of the receptor–active sites of the angiotensin-converting enzyme (ACE, structure unknown) with carboxypeptidase A (CPA). Whereas CPA cleaves a single amino acid from the end of a peptide chain, ACE is known to cleave a dipeptide. ACE was hypothesized to have a Zn^{2+} atom in the active site, like CPA. The structure of the active site cavity of CPA was therefore taken as a model of the active site of ACE. With the role of the labile dipeptide included, spatial accomodation of ionic, hydrogen, and covalent bonding was postulated; the simple compound, succinyl-L-proline, was shown to be active. Substitution of a sulfur atom for oxygen in the terminal carboxyl group gave greatly enhanced activity. The hypertensive properties of these novel compounds have been demonstrated *in vivo*. This work represents a logical leap between the "known" active site conformation of a (monopeptide) carboxypeptidase and the "unknown" (dipeptide) carboxypeptidase.

C. VASOPRESSIN ANALOG

Using NMR studies of the "biologically active" conformation of vasopressin (*73*), Smith and Walter (*64*) have used as a working hypothesis an approximately disk-shaped (attributed to β-structure) molecule, one surface of which is "featureless and hydrophobic"; the opposite surface has charged residues (Gln, Asp, COOH-terminus). One part of the rim of the disk was taken to be involved with receptor binding. Guided by this working model, amino acid substitutions and chemical modifications were proposed to enhance the antidiuretic properties of the analog while reducing its pressor activity. Synthesis and testing confirmed both the predicted properties of the analog and the method: "The systematic design of vasopressin analogs now rests on a rational foundation."

In all three of these cases, completely new compounds were derived from modeling studies. In the first two cases, the active series were totally unrelated to known pharmacophores.

D. ALCOHOL DEHYDROGENASE

Because of our interest in developing computer graphics techniques to facilitate the study of molecular interactions, work was begun in this laboratory on modeling an enzyme–substrate system extracted from the Protein Data Bank. Brändén and his group (*20*) have studied the enzyme,

horse liver alcohol dehydrogenase (ADH), at a resolution of 2.4 Å (Figs. 4–5 and 7–10). ADH accepts a variety of substrates with alcohols, or ketones and aldehydes, as functional groups, and, with cofactors NAD^+ or NADH, it oxidizes or reduces the substrate. Thus, the enzymic reaction is reversible under varying conditions.

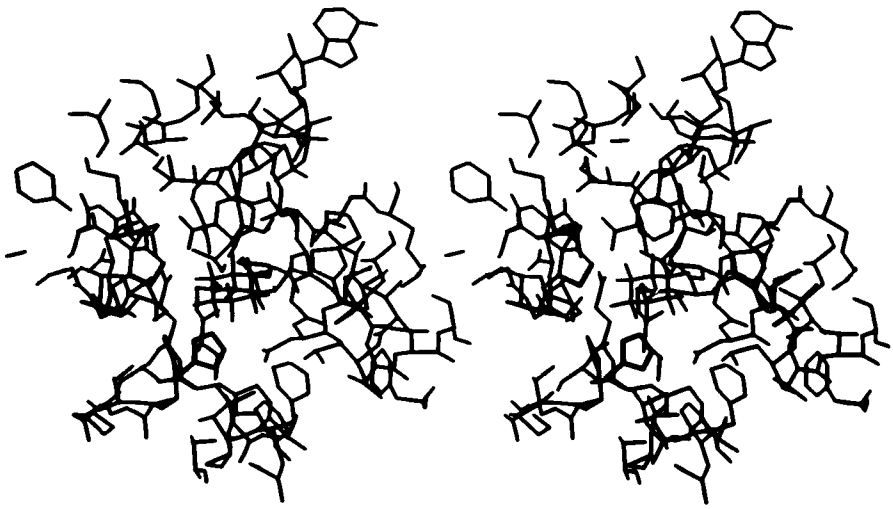


Fig. 7. Stereo view of the active site model of ADH with the identical orientation as Figs. 6 and 8. The central Zn atom (marked by a cross) is seen by looking through the cyclohexanol substrate. All amino acid residues within 12 Å of Cys-46 are drawn as extracted by program SEARCH. NAD^+ is seen from center to top right. As in Fig. 8, this view shows the substrate in the active site along the entrance cavity.

1. *Enzyme Kinetics*

Extensive studies of the *in vitro* kinetics of this enzyme as a function of the stereochemically substituted (H atom, methyl, ethyl, isopropyl, or *tert.*-butyl substituents) substrate, cyclohexanone, have been carried out by H. Dutler (*19*), who then correlated his results with a brass-rod model of the active site geometry.

2. *Computer Graphics Modelings*

Gerry White and I have undertaken the task of applying interactive modeling programs available on our laboratory graphics system to duplicate Dutler's comparison in order to show the flexible utility of these graphical methods in studying the mode of interaction of small molecules in receptor and active site regions of macromolecules (*52*). Although Dut-

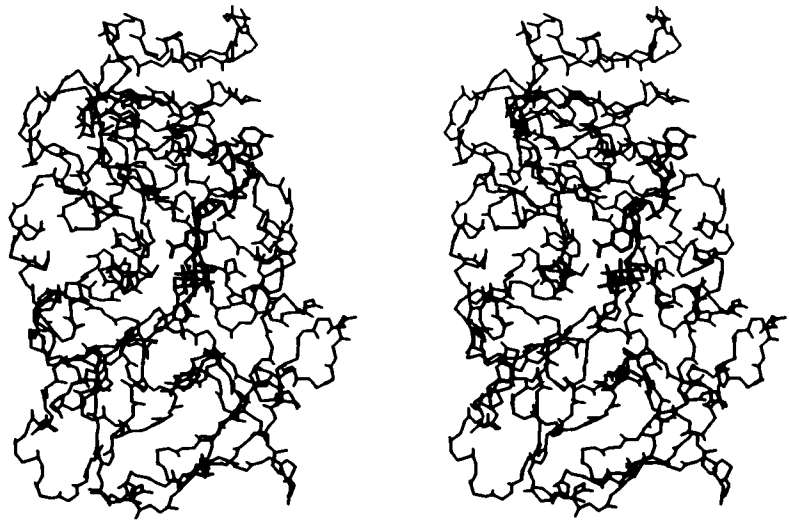


Fig. 8. A rotated, expanded, and simplified stereo view of the model of the ternary complex, ADH + NAD^+ + cyclohexanol, as seen along the entrance cavity to the active site. The enzyme is drawn as the amino acid backbone skeleton. Secondary structural features are seen: α-helices to the right, left and above, β-pleated sheet regions are above and below the active site region. Although three-dimensional, these still pictures are unable to convey the freedom available with program FIT to position the substrate and ask geometric questions interactively.

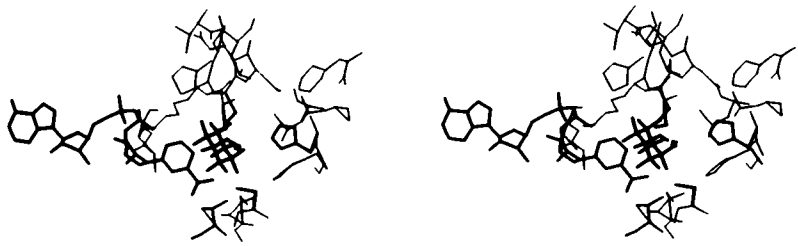


Fig. 9. Amino acids within 8 Å of Cys-46 are shown in this stereo model, together with NAD^+ and cyclohexanol. Catalytically active groups are drawn more intensely.

ler chose to use ADH to reduce cyclohexanones with NADH for experimental reasons, our modeling efforts were based on the related oxidation reaction with NAD^+ and cyclohexanols. In either case the C—O bond must be juxtaposed in the active site cavity between the nicotinamide moiety of NAD^+ (NADH) and the central Zn atom, and thus fixed in space. Although the six-carbon ring structure would be free to rotate in

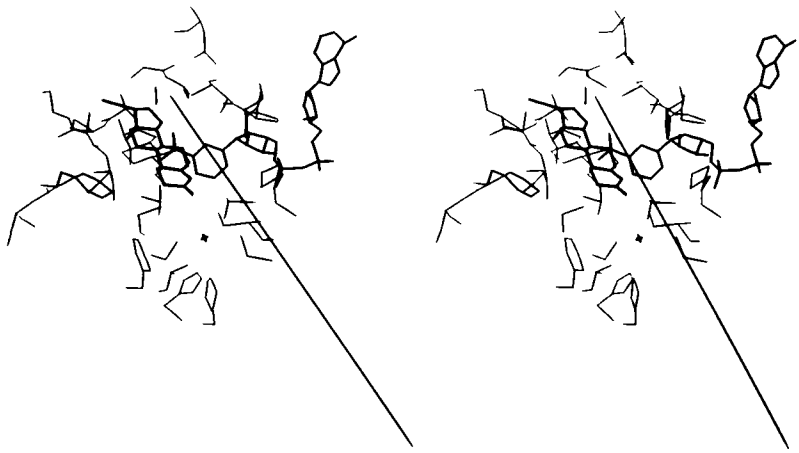


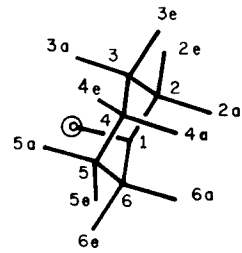
Fig. 10. Stereo view of two superimposed models of steroids in the entrance cavity of ADH. The A-ring and C-3 keto groups have been superimposed, the C-3 oxygen atom has been placed into binding proximity of the Zn atom. The pointer is directed at the fragment of two amino acids (Met-306, Leu-309) in the entrance cavity which belong to the diad-related ADH in the dimeric form of the enzyme.

space about the C—O vector, the hydrogen to be removed from the substrate C-1 atom must lie between the C-4 atom of the nicotinamide group and C-1 for the reaction to occur with favorable kinetics (cf. Figs. 4, 5, 8). Thus the ring is likewise fixed in space. Bulky substituents on the cyclohexanol (one) ring could possibly interact unfavorably with NAD^+, the Zn atom, or the protein active site region.

3. *Fitting the Cofactor*

The X-ray study used adenosine diphosphoribose (ADP ribose) as an inhibitor to ADH. The first task in the interactive model construction was to optimize the position of NAD^+ as provided by Saenger (*63*). The adenosine ring of NAD^+ was caused to superimpose with the ADP ribose by rotation and translation operations. A series of twists about allowed torsional angles were applied to bring the nicotinamide group into the active site cavity. Next, a model of cyclohexanol was translated and rotated until the C—O bond was fixed (Figs. 4 and 5). The oxygen atom assumed a position of regular tetrahedral coordination with the Zn atom, the C-1 to H bond was positioned to point at the C-4 atom of nicotinamide at an approximate approach angle of ca. 60°, corresponding to an optimum reaction coordinate geometry.

4. *Observations*



There are 10 stereochemically unique positions on the cyclohexanol ring (i.e., 5 axial and 5 equatorial). It was immediately clear that bulk was not a critical factor at the C-4 "tail" region, para to the hydroxyl group (cf. Fig. 7). Likewise, the axial (meta) positions of C-3 and C-5 atoms offered little possibility of steric crowding with the active site region as they point away from NAD^+. However, 1,3- (or 1,5-) diaxial interactions are known to induce a conformational transition in the cyclohexane ring so that bulky co-axial groups are shifted to equatorial positions in order to avoid each other. Yet, an equatorial C—O bond would cause NAD^+ to have to approach from a sterically crowded axial direction. Our modeling of 1,3- or 1,5-diaxially substituted cyclohexanol predicted that these would be poor substrates. With the C—O bond axial, 3e and 5e bonds are uncomfortably close to active site amino acid residues, and the oxygen atom would be crowded by bulky (nonhydrogen) substituents at the 3a and 5a positions, inducing an equatorial-to-axial twist in the ring. Dutler finds that the 5e position shows essentially no steric impediment to the reaction. The contradiction is resolved if, in the course of reduction and ejection from the active site, the cyclohexanol product undergoes an axial-to-equatorial twist.

5. *Modeling Structure–Activity Relationships*

As with a few but growing number of examples, questions of structure–function relationships may be applied to the ADH system. Briefly, although the geometric parameters of the catalytic site (i.e., Zn^{2+}, C-4 of $NADH^+$) are relatively fixed, the entrance cavity expands outward in size so that, in addition to the common ethanol substrate, a variety of larger substrates may be oxidized or reduced. In terms of quite small molecules, the question of why isobutanol is a poor substrate while cyclohexanol is a good substrate can be answered by looking for the presence or absence of stabilizing (van der Waals) interactions. Dutler has studied these relationships as well as those of more complicated systems (*19*).

However, alcohol dehydrogenase is not specific for small aliphatic alcohols (aldehydes, ketones) only; it is also a redox enzyme for larger substrates such as steroids. Indeed the A-ring of the steroid skeleton may superimpose partially or completely onto cyclohexanol, depending upon whether the C-3 hydroxyl group of the steroid is α or β to the plane of the skeleton. Theorell and co-workers (*16*) have studied the reaction kinetics of a variety of steroid substrates. They found that steroids with the C-3 hydroxyl group in the α-position were inactive while the β-position permitted oxidation to occur. These relationships have been discussed by Brändén (*9*).

We have begun a study of the interaction of steroids with ADH using Theorell's work as a guide. Although steroids mediate a number of highly specific physiological responses, this is the only available structure of a macromolecule known to bind and react with steroids, which, because of the generous dimensions of the entrance cavity, may be oxidized or reduced like the simpler alcohols or ketones.

Figure 10 illustrates the case of the models of two steroid molecules with planar A-rings and keto oxygen atoms at C-3. Program FIT permits the operator of the graphics system to adjust the rotation and translation of a foreground (steroid) model to assume any number of trial locations. Interactions with groups near the catalytic site and in the entrance cavity may be observed in this process. Especially interesting are the two amino acids, Met-306 and Leu-309, being pointed at in Fig. 10. Because these two amino acids come from the neighboring molecule in the dimeric form of the enzyme, steric interactions in their vicinity could be transmitted through the polypeptide backbone to the catalytic site of the diad-related enzyme. Thus, symmetry may be involved in the structure–function relationship of ADH. Physiologically, ADH is active as a dimer, the individual enzymes related by a diad (two-fold) axis in the solid state. Because the diad-related enzyme forms part of the entrance cavity, bulky interactions would cause steric changes in the symmetrically related molecule as well. The methods of information propagation in proteins (e.g., in allosterism) are not well understood in detail due to the current absence of definitive, high resolution studies. Although steroids react highly specifically with a variety of macromolecules, this is one of the first cases where a physiologically active receptor–substrate complex involving steroids may be modeled for structure–activity studies.

6. *Summary*

Using existing computer programs, we have (a.) extracted the active site region of an enzyme in the Protein Data Bank; (b.) "synthesized" a series of substrate models (substituted cyclohexanols); (c.) used small-

molecule coordinates to generate the NAD^+ cofactor and position it in the active site region; (d.) positioned a model of a substrate molecule into bonding position; (e.) observed which bulky substituents would be clear and which would be likely to interfere with the narrow confines of the enzyme active site; (f.) obtained a good correlation of interactive modeling results with the enzyme kinetics measurements of Dutler; and (g.) used the bond geometry of the receptor + substrate complex as a pattern for the binding of steroid substrates in ways heretofore inaccessable to the medicinal chemist.

Two sweeping assumptions have been made: 1. The molecules are essentially rigid and 2. the unrefined atomic position of the protein at 2.4 Å resolution and the adjusted cofactor orientation correspond sufficiently to dynamically possible conformations. Nevertheless, our modeling efforts support and extend those of Dutler and both correlate with the related enzyme kinetic measurements.

E. CARBONIC ANHYDRASE

Other attempts at using interactive graphics have been recently reported. Chignell and coworkers (*74*) sought to correlate the fit of spin-labeled sulfonamides with electron spin resonance measurements of carbonic anhydrases from a variety of mammalian species. Unfortunate difficulties in visualizing the models in 3-D caused the interactive feature to be reduced to the one-dimensional parameter of distance of the spin-label from the active site zinc atom. This study suggests a method for probing the depth of penetration of a spin-label within the active site crevice; it also suggests that the appropriate spin-labeled substrate is able to differentiate between isozymes. Finally, it suggests the need of a full panoply of visualization techniques and sharpened perceptions in order to take advantage of available information, a curious limitation in some current graphics systems which remain quasi-2-D.

F. THE OPIATE RECEPTOR MODEL

Gorin and Marshall (*28*) have proposed a stereochemical model of the opiate receptor using crystallographic data and their graphics system. They then calculated the related conformation of the minimal peptide of enkephalin (Tyr-Gly-Gly-Phe) that was consistent with the postulated receptor model. Using the MMS-X computer graphics display (Fig. 2), they are able to superimpose the ring center and substituent atoms to show the structural analogy of a set of molecules (cf. Fig. 11). Likewise, a type of "cage" contour was used to sculpt the averaged cavity for a series of

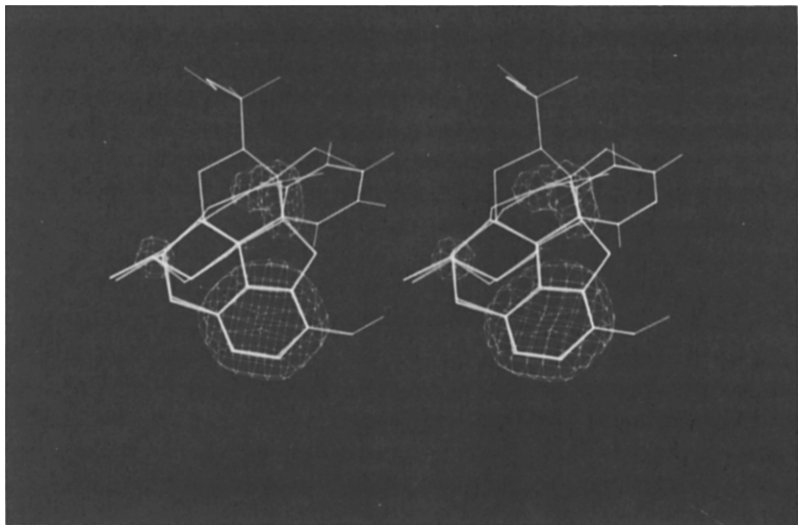


Fig. 11. A stereo view of three potent opiates fitting a postulated target pharmacophore based on the crystal structure of morphine (courtesy of Barry, Gorin, and Marshall, Washington University, St. Louis).

analogs. Candidate analogs could then be tested visually for goodness-of-fit.

G. DIHYDROFOLATEREDUCTASE

A combined crystallographic and modeling effort is currently being made to study the interaction of amethopterin (methotrexate) and similar drugs with the enzyme, dihydrofolatereductase (*46*). In a 2.5 Å resolution study, the pteridine ring has been located in a 15 Å-deep cavity and found to have interactions with 13 to 15 amino acid residues. Most attractions involve hydrophobic and hydrogen-bonding interactions. Asp-26, located nearby in a hydrophobic pocket, is taken to be the proton source in the reduction reaction. Further refinement or higher resolution studies may be expected to yield full geometric details of the drug–receptor interactions and suggest correlations with functional parameters, leading to the modifications of existing drugs or to the design of novel drugs.

H. CURRENT LIMITATIONS

The limitations of current modeling methods are as follows:

1. As discussed above, the resolution of the diffraction study limits the precision (and accuracy) of individual atomic positions.

2. The macromolecule and substrate systems discussed here have been treated as rigid molecules. Subgroup manipulation permits partial flexibility (e.g., torsion angle rotation about single bonds). Thus, the computer graphics model has some similarities to commonly used brass or plastic atomic models without the disadvantages of mechanical instability.
3. The current approach has not taken into account excited state conformational changes, deformations from electron orbital overlap, etc.

A likely next level of elegance assumes a high-resolution macromolecular structure. Then force field calculations could permit the refinement of protein and substrate positional parameters, thus permitting internal degrees of freedom (*33, 47*). Continued development of molecular dynamics programs and energy minimization routines should permit the modeling procedure to be initiated visually and then extended computationally with many parameters allowed to vary simultaneously.

The several successful, noninteractive cases discussed at the beginning of Section VI illustrate the usefulness of even approximate modeling techniques. The main purpose of this presentation is to show how the powers of 3-D computer graphics can be applied to creating simpler models of complex macromolecules and how manipulative procedures can be applied to modeling molecular interactions. Because of the digital nature of the data, quantitative (e.g., molecular geometry) calculations are easily applied, and finally, the results may be recorded both photographically and magnetically for later use.

VII. A Digital Approach to Drug Design

A. EARLY EFFORTS

Early modeling attempts used structural similarities to help postulate receptor site geometries without the assistance of computer graphics: Thus, two-dimensional silhouettes of a variety of substrates were used to demonstrate a high degree of correlation between molecular shape to ant alarm pheromone activity using a "scanning computer" (*1*).

Three dimensional structure–activity correlations have been confirmed for a variety of compounds, for example, muscarinic stimulants (*3*), sympathomimetic amines (*6*), psychoactive indolealkylamines (*22*), and neuroleptic drugs (*67*). A model for the active site of glutamine synthetase was postulated from the 2-D model of the extended form of L- and D-glutamic acid (*26*).

B. A SCENARIO IN 3-D

1. *Method*

I would propose the following scenario for a computer-aided approach to drug design: First, the investigator would choose the desired physiological–chemical properties, together with known molecules that exhibit these properties with respect to a unique receptor site. Steric parameters (*71*) could be established for the starting set of compounds. Stereochemical similarities (spatial proximities of charged or polar regions, H-bonding donors/acceptors, nucleo- or electrophilic atoms, aromatic or hydrophobic groups, heteroatoms, etc.) would be sought by visual inspection and empirical weights assigned. A computer program (not yet written) would sum the weighted steric parameters and sculpt a postulated receptor geometry. The composite picture could be input to a 3-D substructure search of structural data files. Candidates for synthesis could be generated by fixing atoms at points of high correlation, linked by a rigid or flexible network of bridging atoms.

Alternatively, if the 3-D structure of a receptor site is available, models of candidate molecules could be constructed interactively to possess the required steric correlation and the desired internal rigidity leading to a targeted synthesis of promising candidates. A candidate molecule could be proposed for computer-aided synthesis. Spectroscopic or X-ray diffraction experiments could then monitor the postulated structural agreement. The possibilities of pattern recognition techniques have recently been reviewed by Gund (*30*). The Merck Molecular Modelling System (*31*) has been applied to patterns of antileukemic and analgesic pharmacophores. The role of computers in drug information processing has likewise been recently reviewed (*44*).

2. *Step-Wise Approach*

Three levels of elegance may be envisioned for the modeling of molecular interactions: (1) rigid substrate–rigid receptor, (2) flexible substrate–rigid receptor, and (3) flexible substrate–flexible receptor.

The first level is the most straightforward and is already producing results. Molecular dynamics or molecular orbital calculations of excited states (i.e., strained geometries) could aid in the identification of likely conformers. Most current interactive modeling techniques permit internal (e.g., torsional angle) rotations, so the second level is already partially obtainable.

As the resolution and precision of the macromolecular data base continues to improve, the stage is set for full-scale molecular dynamics calculations of both protein and substrate. This is a separate area which will be increasingly discussed as it emerges.

C. CRYSTALLOGRAPHIC CONFIRMATION

1. *Modeling*

Interactive computer graphics modeling programs (e.g., FIT) facilitate the construction of the macromolelule model into the 3-D electron-density map which is derived from the crystallographic experiment. As this model emerges, it may be subjected to further refinement until it is ready for model substrate binding studies. Here, a 3-D view of the "lock" would be available for fitting known and postulated "keys." Potential substrates may be "synthesized" within the computer for testing before they have ever been synthesized in the laboratory.

2. *Testing*

Candidate molecules may be tested *in vitro* or *in vivo;* binding to isolated receptor molecules may be followed by ESR or NMR; and finally, conformational changes may be calculated as above, or, so long as the complex of the protein and trial substrate remain *isomorphous* with the previous crystallographic study, difference Fourier and refinement techniques will provide crystallographic evidence of specific binding stereochemistry, especially when measurements can be made at low temperatures.

3. *Stop-Action and Refinement*

Currently developing low-temperature techniques (*61, 83*) may permit the substrate–receptor complex to be frozen in place as well as allow longer crystal life and improved precision in the measurement of diffraction intensities, leading to improved resolution and precision in atomic positions. Several current refinement techniques could be applied to complete the structural investigation. In addition, these latter results should add considerably to the understanding of enzymology.

D. PATTERN SEARCHING IN 3-D AND 4-D

As a definitive data base is extended, creative 3-D pattern-searching techniques would be applied for the interactive (or even automated) modeling of potential pharmacophores. Other structure–activity techniques could be introduced at appropriate steps in the process to maintain the overall requirements of a good drug molecule, which would have been designed as a "magic bullet" aimed at a well-defined target.

Finally, pattern-searching techniques could be extended to include receptor molecules responsive to a variety of drugs but not amenable to crystallographic techniques (e.g., membrane-bound receptors, transport proteins).

Because complex molecular interactions involve 3-D relationships, any attempt to model or study these interactions in a space with fewer dimensions will be severely handicapped. Indeed, efforts at 4-D modeling might offer unifying principles even as the several 3-D polygons (cube, rectangular, and hexagonal parrallelepiped) are all "shadows" of the same 4-D tesserac (*69*).

VIII. Summary

This chapter was originally envisioned as a somewhat speculative article on the potential application of computer techniques, especially computer graphics, to medicinal chemistry. However, recent events have confirmed many of the originally tentative aspects. A merging of chemical and structural sciences with available computer technology has produced a method with real potential for improving the acuity of the medicinal chemist to target his synthetic efforts.

Traditionally, the synthetic chemist was not usually exposed to computers. Yet, in considering the educational and information retrieval aspects alone, one may easily predict that current computer technology will have a profound impact on synthetic chemistry. Part of this article attends to the potentials—and the pitfalls—of laboratory computers. The main intent, however, is to review the most recent events that signal a markedly new approach to drug design and the broader study of molecule–molecule interactions by matching the highly developed visual skills of chemists with current interactive display hardware by means of well-designed computer software systems. Several such systems have now been developed and the direction of these research efforts may easily be predicted to turn to modeling well-defined cases. Two essential limitations exist for this method: (1) availability of structural information and (2) insufficient resolution of atomic positional parameters. A number of laboratories are actively seeking to minimize both limitations. The time is ripe for a dialog between the medicinal biochemist and the crystallographer.

Radically new structural information is now becoming available to the medicinal chemist. The lock and key mechanism of receptor + substrate interactions is no longer a nebulous concept. Interactive computer graphics modeling techniques now make it possible to visualize the interactions of molecules, chemical groups, and even individual atoms. Thus, facile modeling methods are now unlocking the gate to a new avenue to drug design.

ACKNOWLEDGMENT

The author acknowledges Stanley M. Swanson for software modifications and Gerry White and Greg Cole for help with the interactive fitting. Peter Gund and Eli Schefter read an early draft of the manuscript and offered useful criticisms. Garland Marshall and David Barry provided examples of the MMS-X system. Hans Dutler provided biochemical information, but especially, the critical views of a structural biochemist.

References

1. J. E. Amoore, G. Palmieri, E. Wanke, and M. S. Blum, *Science* **165,** 1266–1269 (1969).
2. A. Arnone, *Nature (London)* **237,** 146–147 (1972).
3. R. W. Baker, C. H. Chothia, P. Pauling, and T. J. Petcher, *Nature (London)* **230,** 439–445 (1971).
4. C. D. Barry, H. E. Bosshard, R. A. Ellis, and G. R. Marshall, *In* "Computers in Life Science Research" (W. Siler and D. A. B. Lindberg, eds.), Plenum Press, New York, 1974; cf. also *Fed. Proc.* **33:12,** 2368–2372 (1974).
5. C. D. Barry, Tech. Memo. 228, Computer Systems Laboratory, Washington University, St. Louis, Missouri. cf. also *MMS-X News* **1,** 1 (1978).
6. R. Bergin, "The Molecular Structure of Some Sympathomimetic Amines and Related Substances," Ph.D. Thesis, Stockholm, (1971).
7. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535–542 (1977).
8. T. L. Blundell and L. N. Johnson, "Protein Crystallography." Academic Press, New York, 1976.
9. C. I. Brändén, "Functional Significance of the Structure of Liver Alcohol Dehydrogenase" *In* "Pyridine Nucleotide-Dependent Dehydrogenases" pp. 325–338, H. Sund, ed. W. de Gruyter, Berlin, 1977.
10. F. P. Brooks, Jr., "The Mythical Man-Month." Addison-Wesley, Reading, Mass., 1975.
11. F. F. Brown and P. J. Goodford, *Br. J. Pharmacol.* **60,** 337–341 (1977).
12. H. B. Bürgi, J. D. Dunitz, and E. Shefter, *Nature (London) New Biol.* **244,** 186–187, (1973).
13. H. B. Bürgi, J. D. Dunitz, and E. Shefter, *Acta Cryst.* **1330,** 1517–1527 (1974).
14. D. Carlström, R. Bergin, and G. Falkenberg, *Quart. Rev. Biophys.* **6,** (No. 3), 257–310 (1973).
15. D. M. Collins, F. A. Cotton, E. E. Hazen, Jr., E. F. Meyer, Jr., and C. N. Morimoto, *Science* **190,** 1047–1053 (1975).
16. T. Cronholm, C. Larsén, J. Sjörall, H. Theorell, and Å. Åkeson, *Acta Chem. Scand. B29,* 571–576 (1975).
17. D. W. Cushman, H. S. Cheung, E. F. Sabo, and M. A. Ondetti, *Biochemistry* **16,** 5484–5490 (1977).
18. M. J. S. Dewar *In* "Further Perspectives in Organic Chemistry." Ciba Foundation Symposium 53, 1978.
19. H. Dutler, *In* Structure–Activity Relationships in Chemoreception" (G. Benz, ed.), pp. 65–73, Information Retrieval, London, 1976.

19a. "Substrate Orientation in the Active Site of Liver Alcohol Dehydrogenase", *In* "Pyridine Nucleotide-Dependent Dehydrogenases", p. 339–350, H. Sund, Ed., Walter deGruyter, Berlin (1977).

20. H. Eklund, B. Nordström, E. Zeppezauer, G. Söderlund, I. Ohlsson, T. Boiwe, B. O. Söderlund, O. Tapia, C. I. Brändén, and Å. Åkeson, *J. Mol. Biol.* **102,** 27–60 (1976).
21. O. Ermer, *In* "Structure and Bonding" (J. D. Dunitz *et al.*, eds.), pp. 161–211, Springer-Verlag, Berlin and New York, 1976.
22. G. Falkenberg, "The Molecular Structure of Some Psychoactive Indolealkylamines and Related Substances," Doctoral Thesis, Stockholm, 1972.
23. R. J. Feldmann, *Annu. Rev. Biophys. Bioeng.* **5,** 477–510 (1976).
24. R. J. Feldmann, C. R. T. Bacon, and J. S. Cohen, *Nature (London)* **244,** 113–115 (1974).
25. R. J. Feldmann, AMSOM: Atlas of Macromolecular Structure on Microfiche, Tracor Sitco, Inc., Rockville, Maryland, 1977.
26. J. D. Gass and A. Meister, *Biochemistry* **9,** 1380–1390 (1970).
27. P. J. Goodford, *In* "Drug Action at the Molecular Level," (G. C. K. Roberts, ed.), pp. 109–126. MacMillan, 1977.
28. F. A. Gorin and G. R. Marshall, *Proc. Natl. Acad. Sci. U.S.A.* **74:11,** 5179–5183 (1977).
29. R. L. Gregory, *Sci. Amer.*, Nov., 1968.
30. P. Gund, J. D. Andose, and J. B. Rhodes, Proc. III Internat. Conf. on Computers in Chem. Res., Educ., and Technol., Plenum, New York, in press.
31. F. E. Hahn and P. Gund, *In* "Drug Receptor Interaction in Antimicrobial Chemotherapy" (J. Drew and F. E. Hahn, eds.), pp. 245–266. Springer-Verlag, Berlin and New York, 1978.
32. B. S. Hass, T. V. Willoughby, C. N. Morimoto, D. L. Cullen, and E. F. Meyer, Jr., *Acta. Cryst.* **B31,** 1225–1229 (1975).
33. J. Hermans and J. E. McQueen, *Acta. Cryst.* **A30,** 730–739 (1974).
34. G. Herzberg, "Molecular Spectra and Molecular Structure," Vols. 1 and 2. D. Van Nostrand Co., Princeton, N.J., 1945, 1950.
35. R. Huber and W. Bode, *Acc. Chem. Res.* **113,** 114–121 (1978).
35a. G. A. Jeffrey and R. K. McMullan, *Prog. in Org. Chem.* **8,** 43–108 (1967).
36. K. Johnson, ORTEP, Report ORNL-3749 Revised (1965), Oak Ridge Natl. Lab., Oak Ridge, Tennessee, 1965.
37. L. Katz and C. Levinthal, *Annu. Rev. Biophys. Bioeng.* **1,** 465–504 (1972).
38. O. Kennard, D. G. Watson, F. H. Allen, N. W. Isaacs, W. D. S. Motherwell, R. C. Pettersen, and W. G. Town, "Molecular Structure and Dimension," Vol. A1, Oosthoeck, Utrecht, 1972.
39. O. Kennard, D. G. Watson, F. H. Allen, and S. M. Weeds, "Molecular Structure and Dimensions," (Bibliography) Vols. 1–9, Bohn, Scheltema, and Holkema, Utrecht. 1970–1979.
40. B. W. Kernigan and P. J. Pflauger, "Software Tools." Addison-Wesley, Reading, Massachusetts, 1976.
41. J. Konnert, *Acta. Cryst.* **A32,** 614 (1976).
42. M. J. Legg, "Protein Crystallography: New Approaches to X-ray Data Collection, Direct Space Refinement and Model Building," Ph.D. Dissertation, Texas A&M University, College Station, Texas, 1978.
43. R. G. Legg, "Geographic Variation in the Cactus Mouse, *Peromycus eremicus* (Rodenta: Muridae)," Ph.D. Dissertation, Texas A&M University, College Station, Texas, 1978.
44. P. J. Lewi, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. VI, pp. 209–278. Academic Press, New York, 1976.
45. M. W. Makimer and A. L. Fink, *Ann. Rev. Biophys. Bioeng.* **6,** 301–343 (1977).
46. D. A. Matthews, R. A. Alden, J. T. Bolin, D. J. Filman, S. T. Freer, R. Hamlin, W. G. J.

Hol, R. L. Kisliuk, E. J. Postore, L. T. Plante, N. Xuong, and J. Kraut, *J. Biol. Chem.* **253,** 6946–6954 (1978).
47. J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature (London)* **267,** 585–590 (1977). (cf. also the commentary by B. Robson, pp. 577–578.)
48. E. F. Meyer, Jr., *J. Appl. Cryst.* **3,** 392–395 (1970).
49. E. F. Meyer, Jr., *J. Chem. Doc.* **10,** 85–86 (1970).
50. E. F. Meyer, Jr., *Nature (London)* **232,** 255–257 (1971).
51. E. F. Meyer, Jr., *Biopolymers* **13,** 419–422 (1974).
52. E. F. Meyer, G. M. Cole, S. M. Swanson, and W. G. White, "Modelling Receptor and Substrate Interactions," Symposium, "Computer Assisted Drug Design," *Am. Chem. Soc.* (1979).
53. H. L. Morgan, *J. Chem. Doc.* **5,** 107–113 (1965).
54. C. N. Morimoto and E. F. Meyer, Jr., *In* "Crystallographic Computing Techniques" (F. R. Amned, ed.), pp. 488–496. Munksgaard, Copenhagen, 1976.
55. S. Motherwell and P. Murray-Rust; *Acta Cryst.* **B34,** 2518–2546 (1978).
56. W. M. Newman and R. F. Sproull, "Principles of Interactive Computer Graphics." McGraw-Hill, New York, 1973.
57. D. Ophir, R. J. Shephard, and R. J. Spinrad, *Commun. ACM* **12,** 309–310 (1969).
58. A. A. Ortony, *Comput. J.* **14,** 140–144, (1970).
59. L. Pauling, "The Nature of the Chemical Bond," 3rd ed., Cornell University Press, Ithaca, New York, 1960.
60. G. Pepe, E. F. Meyer, Jr., in preparation.
61. G. Pesko, *J. Mol. Biol.* **96,** 381–392 (1975).
62. H. Rütishauser, "Handbook for Automatic Computing," Vol. 1, Part a, Springer-Verlag, New York, 1967.
63. W. Saenger, B. S. Reddy, K. Mühlegger, and G. Weimann, *Nature (London)* **267,** 225–229 (1977).
64. C. W. Smith and R. Walter, *Science* **199,** 297–299, (1978).
65. R. F. Sproull, OMNIGRAPH, CSL74-4, Xerox, Palo Alto Res. Center, Palo Alto, CA., 1973.
66. S. M. Swanson, *J. Mol. Biol.,* **129,** 637–642 (1979).
67. J. P. Tollenaere, H. Moereels, M. H. J. Koch, *Eur. of Med. Chem.* **12,** 199–211 (1977).
68. D. Tsernoglou, G. A. Petsko, J. E. McQueen, Jr., and J. Hermans, *Science* **197,** 1378–1381 (1977).
69. M. Van Bavel, private communication.
70. K. A. Van Lehn, SAIL Users Manual, Note 204, Stanford Artificial Intelligence Project, Palo Alto, California, 1973.
71. A. Verloop, W. Hoogenstraaten, and J. Tipker, *In* "Drug Design" (E. J. Ariëns, ed.), Vol. VII. Academic Press, New York, 1976.
72. V. Villarreal, Jr., E. F. Meyer, Jr., R. W. Elliott, and C. Morimoto, *J. Chem. Info. and Comp. Sci.* **15,** 220–225 (1975).
73. R. Walter, C. W. Smith, P. K. Metha, S. Boojurern, J. A. L. Arruda, and N. A. Kurtzman, *In* "Disturbances in Body Fluid Osmolality" (T. E. Andreoli, et al. eds.). Am. Physiol. Soc., Bethesda, Maryland, 1977.
74. V. T. Wee, R. J. Feldmann, R. T. Tanis, and C. F. Chignell, *Mol. Pharmacol.* **12,** 832–843 (1976).
75. T. V. Willoughby, C. N. Morimoto, R. A. Sparks, and E. F. Meyer, Jr., *J. Appl. Cryst.* **7,** 430–434 (1974).

76. W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, "Computer Presentation and Manipulation of Chemical Information." Wiley, New York, 1974.
77. N. Wirth, "The Programming Language PASCAL," *Data Inform.* 1, 35–63 (1971).
78. J. A. Yankeelov, Jr. and J. R. Coggins, *In* "Structure and Function of Proteins at the Three-Dimensional Level," Vol. XXXVI, Cold Spring Harbor Symposia, New York, 11724, 1972.
79. "Needs and Opportunities for the National Resource for Computation in Chemistry (NRCC)," Natl. Acad. Sci., Washington, D.C., 1976.
80. IBM System/360, PL/1 Reference Manual, Form C-88-8201.
81. *Acta. Cryst.* **B29,** 1746 (1973).
82. Cf. collected papers on structure refinement, *In* "Crystallographic Computing Techniques," (F. R. Ahmed, ed.), pp. 291–335. Munksgaard, Copenhagen, 1976.
83. T. P. Singh, W. Bode, and R. Huber, *Acta Cryst., in press.*

Chapter 7 Logico-Structural Approach to Computer-Assisted Drug Design

V. E. Golender and A. B. Rozenblit

ISBN 0-12-060309-8

I. Introduction

A. Current Trends in Computer-Assisted Drug Design

A scientist searching for new drugs must avoid many pitfalls. The process of drug research and development involves many steps, and requires both strenuous effort and considerable expenditure. Each stage of the process (e.g., pharmacological screening, clinical study, development of technological procedures) sees the majority of the tested compounds discarded. Only a tiny portion of the newly synthesized compounds are useful in the long run. Hence, the urgent need to design drugs via a rational basis which will reduce trial and error to a minimum (*1*). The rapid developments in life sciences and the quantity of theoretical and empirical knowledge on the biological effects of chemicals available today provide a sound foundation for such an approach.

However, our present knowledge of the mechanisms of drug action do not always allow us to make theoretical predictions of the biological activity of a compound from its structure. Hence, an important role is retained by empirical structure–activity relationships.

Establishment of structure–activity relationships (SAR) is often the goal for which chemists and pharmacologists are striving. But owing to the complexity of chemical structures, the large number of compounds studied, and the complex pattern of structure–activity relationships, the scientist is unable, in many cases, to identify by mere "visual analysis" the correlations for which he is looking. Furthermore, many of the correlations appear statistically insignificant on closer examination.

The search for empirical correlations can be improved and facilitated by applying various computer-assisted methods. It is not our aim to provide a survey of the numerous publications in the field, since many such reviews exist (2–4).

It should be pointed out, however, that several limitations of the classical Hansch (*5*) and Free-Wilson (*6*) methods have been recognized recently (*2, 4*). As a result, there has been observed a growing interest in the application of pattern recognition methods to the prediction of the biological activity of chemical compounds. The applications of cluster analysis (*7–9*), perceptron (*10*), various linear (*11–15*) and nonlinear (*16, 17*) discriminant functions, nearest-neighbor methods (*13, 14*), factor analysis (*18–20*), etc. have been reported. The structures of compounds in the majority of cases are described by a set of structural fragments.

Most of these studies have earned serious attention, demonstrating applicability of the appropriate methods. However, the main handicap inherent in the majority of algorithms is an inability to identify meaningful

features, that is, structural fragments responsible for the expression of a particular activity. Therefore, the investigator is forced to regard the computer as an "electronic fortune-teller" and to take the results for granted (or reject them) without a proper basis for judgment. This may result, in some cases, in misleading or trivial conclusions as to SAR, despite the high recognition rate observed on a restricted sample (*21–24*).

In relation to the aforesaid, there arises the problem of developing an approach that would provide for selection of the meaningful features responsible for biological activity, based on scrupulous analysis of experimental data and with the investigator participating at all stages.

In our view, the logico-structural approach (LSA), the subject of the present chapter, meets these requirements. It is based on a simulation of the logical processes of a drug designer engaged in establishing correlations between particular structural features of compounds and their activity.

It is interesting to note in this context that although traces of intrinsic logic inherent in the design of new medicinal compounds can be found in almost every publication in this field, it is most clearly revealed by Austel and Kutter (*26*). This logic is essentially identical with the principal laws of Mill's inductive logic (*27*).

The basic inductive methods, such as the Methods of Agreement, Difference and Concomitant Variations, are the principles underlying the LSA developed by us for analysis of SAR.

B. EFFICIENCY CRITERIA OF THE ACTIVITY PREDICTION SYSTEM

The LSA is based on the analysis of information obtained for earlier compounds studied; and, like other QSAR procedures, it is only a statistical method, and cannot guarantee faultless classification of a new substance as either active or inactive. Therefore, it is desirable to assess the efficiency of the statistical prediction system employed.

Suppose the input data of a prediction system (Fig. 1) is composed of L compounds, L_1 of which are endowed with activity. The system will identify with the active class M compounds, of which only M_1 are virtually active. Hence, two kinds of errors are possible.

Errors of the first kind imply false classification of an effective compound within the inactive class of substances. These compounds will be deleted from screening and, therefore, lost.

Errors of the second kind result when inactive compounds fall within the active class. These compounds are bound to undergo further screening, overloading it with useless substances.

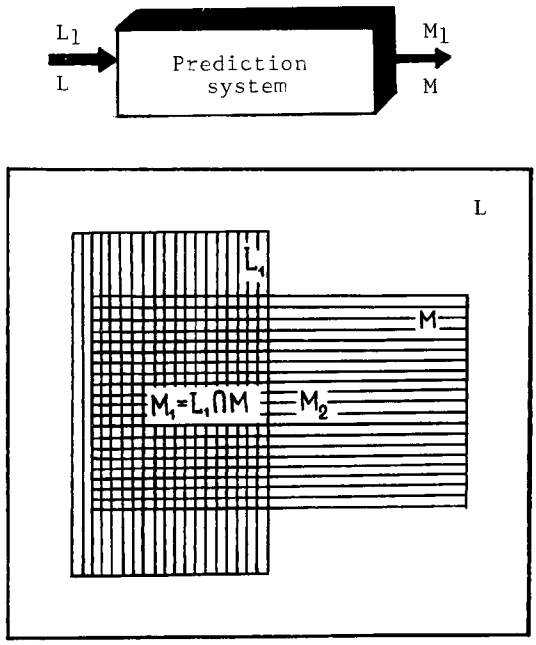


Fig. 1. Characteristics of activity prediction system. L, complete set of compounds; L_1, subset of active compounds; M_1, subset of compounds classified by the prediction system as active; L_1/L, the portion of active compounds at the input of prediction system; M_1/M, the portion of active compounds at the output of prediction system.

Let P_1 denote probability of errors of the first kind and P_2 denote probability of errors of the second kind.

As a criterion of economic efficiency we have chosen the ratio K_e demonstrating reduction in costs to obtain a single effective compound by applying the prediction system (*4, 16*). Making certain assumptions, it can be estimated as:

$$K_e = LM_1/L_1M \tag{1}$$

or

$$K_e = (1 - P_1)/P_2$$

The latter formula shows that the efficiency of the prediction system grows with decreasing P_1 and P_2 values. It might seem possible to develop a relatively efficient system even when P_1 remains rather high by appropriately lowering the P_2 value. For example, $P_1 = 0.4$, $P_2 = 0.3$, $K_e = 2$; that is to say, the costs have been reduced by half. However, in our considerations we have to assume that the "stock" of possible chemical compounds at our disposal is practically unlimited. In fact, for the solu-

tion of any particular problem it often appears limited. Therefore, caution must be exercised not to discard active compounds. For this purpose, it is desirable to use as a second efficiency criterion, the probability of correct recognition of effective compounds

$$P_A = 1 - P_1 = M_1/L_1 \tag{2}$$

II. General Principles of Logico-Structural Approach

The logico-structural approach is based on simulating some aspects of the intelligence of the drug designer engaged in the analysis of structure and activity data. It involves selection and further utilization of biological activity features, that is, structural fragments which discriminate as to compounds with different biological activity.

A. The Basic Assumptions and Steps of the Logico-Structural Approach

The LSA is based on the following assumptions:

1. The chemical structure of compounds lends itself to formal description S.
2. The compounds can be grouped, according to their activity, into N classes $A_k, k = 1, \ldots, N$. In the general case, the classes of activity may form intersections when certain compounds belong to several classes of activity.
3. The discriminating biological activity features are defined as certain fragments of structure description, $\varphi(S)$, whose occurrence is indicative of a relatively high probability P that the compound in question is likely to exhibit this particular activity.

Realization of the LSA involves the following steps: (1) structure description of chemical compounds, (2) separation of compounds into classes according to the activity criteria defined by the investigator, (3) selection and statistical estimation of features in each class, and (4) human analysis of computer-selected features and their application to the purposeful synthesis of compounds possessing desirable properties, as well prediction of activity for unexplored substances. These steps deserve a more detailed examination.

1. *Chemical Structure Descriptions*

Structure description is the central step in SAR analysis, since only representation of the structure adequate to the particular task will provide a successful solution.

At the present time, there are several methods of describing a structure for LSA that can be classified into 2 groups according to the data structure: (1) vector or linear representation or (2) matrix or graph representation.

Vector representation of structure makes use of a set of discrete structure parameters such as (1) various structure fragments (included herein are fragments used in different line notations) and (2) physicochemical parameters, substituent constants, and values describing electron and space structure of compounds. Continuous characteristics converted into discrete form can also be employed.

Matrix representation of the structure is performed by using special matrices which give description of (a) molecular topology and (b) molecular topography.

Although matrix representation offers a more detailed description of the chemical structure of compounds, its application requires more complex algorithm manipulating of chemical information and results in higher computational costs. Therefore, from the practical point of view, it appears reasonable to use simpler description methods at the initial stages of SAR analysis.

A more detailed account of the present approaches to structure representation will be described in the next sections.

2. *Selection and Estimation of Biological Activity Features*

The particular form of biological activity features is largely dependent on the language used for structure representation. However, the structures of algorithms employed for selection and estimation of features may appear similar for different representations. For instance, the feature selection algorithms can be based on the following scheme.

1. Potential features of each class are selected by extracting common fragments shared by each pair of compounds. Identification of common fragments φ^r is carried out by the intersection of the structure description of the appropriate compounds

$$\varphi^r = S^m \cap S^n$$

where S^m and S^n are structure representation of the mth and nth compounds respectively.

The $\cap$ denotes common (in the case of vector representation) or generalized (when the structure is represented by matrix or graph) intersection.

2. Occurrence numbers are calculated for potential features of compounds in all classes l_k, $k = 1, \ldots, N$.

3. Statistical estimation of selected intersections on the basis of l_k, $k = 1, \ldots, N$ values is performed and statistically significant features for

each activity class are selected. One of the principal criterion of feature estimation used is the probability that the new compound with the class A_k feature will belong to this class. The mean value of this probability in the case of two activity classes, $N = 2$, can be computed from the expression

$$P(A_k/\varphi^r) = (l_k + 1)/(l_1 + l_2 + 2) \qquad k = 1,2$$

representing Bayes' estimation of probability. The potential feature φ^r is defined as a discriminating feature of class A_k, if

$$P(A_k/\varphi^r) > \lambda_k$$

where λ_k equals a certain threshold value.

When considerable variation in prior probability is obtained among the classes, it is advisable in the place of the probability $P(A_k/\varphi^r)$ to use a relative value, i.e., prognostic utility of the feature

$$U_k^r = P(A_k/\varphi^r)/P(A_k) \tag{4}$$

which is expressed as a ratio of the probability that a compound would exhibit kth activity in the presence of feature φ^r to the prior probability of obtaining a compound with the above activity.

Other feature characteristics connected with the probability that a given feature is not just a "freak of chance" are also examined. A more detailed discussion of the problems faced in the course of statistical feature estimation is presented elsewhere (*4*).

4. Optimization of the threshold values of statistical criteria for feature estimation is performed in accordance with the results of the leave-one-out technique.* The threshold is chosen by means of K_e and P_A criteria optimization [see Eqs. (1) and (2)]. It must be pointed out that the problem of threshold choice enabling us to draw the line between "the good and the bad" is not thereby eliminated, but is only transferred to another (higher) criteria level.

5. Activity prediction and purposeful synthesis of new compounds. Discriminating features are used to predict biological activity of unstudied compounds. The prediction procedure followed in this case may have several modifications. For example, the compound in question can be classified according to its predominant features. If two activity classes A_1 and A_2 are present, the rule can be described as follows:

$$C^m \in A_1, \quad \text{if} \quad V_1 - V_2 > \eta$$

$$C^m \in A_2, \quad \text{if} \quad V_2 - V_1 > \eta$$

* Using the leave-one-out technique, each compound in turn is left out of account by the feature selection procedure and its activity predicted. The number of both correct and incorrect predictions is used for the estimation of the prediction system criteria.

The decision is rejected if $|V_1 - V_2| \leq \eta$, where C^m represents the predicted compound, V_1 and V_2 are the number of A_1 and A_2 class features, respectively, that are present in compound C^m; η denotes threshold, in the particular case $\eta = 0$. This rule is employed in the STRAC program package.

In the case of many overlapping classes, it is more convenient to assign the compound to the class whose features are present on it. This rule is used in the ORACLE program package.

Discriminating features provide a good basis for goal-directed synthesis of compounds with predetermined profiles of biological activity. By combining in a given structure various feature-fragments responsible for desirable biological effects, and by eliminating substructures eliciting unfavorable effects, the researcher provides new opportunities for molecular manipulations.

B. EXAMPLE OF COMPUTATIONAL PROCEDURE

The basic principles and steps of the LSA can be illustrated by this simple example.

Suppose the compounds under study, having different substituents R_1, R_2, R_3, and R_4 (see Table I) and belonging to the same series, have been

TABLE I
LIST OF STUDIED COMPOUNDS (ILLUSTRATIVE EXAMPLE)

No.	R_1	R_2	R_3	R_4	LD_{50}	ED_{50}	I
1	H	CH_3	Br	NH_2	5000	400	12.50
2	CH_3	NH_2	H	C_2H_5	2000	150	13.33
3	H	C_6H_5	Br	C_2H_5	3500	300	11.67
4	CH_3	NH_2	CH_3	NO_2	1800	100	18.00
5	H	NH_2	Cl	NO_2	2000	800	2.50
6	CH_3	C_6H_5	H	NH_2	1000	500	2.00
7	H	CH_3	CH_3	NO_2	70	40	1.75
8	CH_3	C_6H_5	Br	C_2H_5	200	70	2.86

subjected to the SAR analysis using the algorithms of the LSA. The principal steps of the analysis are as follows.

1. *Description of the Chemical Structure*

As the compounds in question are closely related (being members of the same series), it may be sufficient to apply vector represention using various types of substituents attached in the corresponding positions of the

molecule as structure parameters. Thus, compounds 1 and 2 (Table I) are described, respectively, by vector (H, CH_3, Br, NH_2) and vector (CH_3, NH_2, H, C_2H_5).

2. *Separation of the Compounds into Activity Classes*

The results of biological testing expressed as LD_{50} and ED_{50} values are summarized in Table I. It can be seen from the table that the compounds fall into two classes, as judged by the safety index values $I = LD_{50}/ED_{50}$. Compounds 1–4 ($11.7 \leq I \leq 18.0$) belong to the first class, whereas class 2 is composed of compounds 5–8 ($1.75 \leq I \leq 2.86$). The two classes were designated as class A_1 (active compounds) and class A_2 (inactive compounds), respectively.

3. *Selection and Estimation of Class-Specific Features*

Features of each given class are selected by intersecting compound vectors within this class. The feature extraction procedure is exemplified by the intersection of vectors of compounds 1 and 3:

$$\begin{array}{llll} (H, & CH_3, & Br, & NH_2) \\ (H, & C_6H_5, & Br, & C_2H_5) \\ \hline (H, & \varnothing, & Br, & \varnothing) \end{array}$$

The $\varnothing$ signifies that the corresponding vector component may assume any arbitrary value. The selected combination of structural parameters represents the conjunctive feature, which can also be expressed as (R_1 = H) & (R_3 = Br). Successive pair by pair comparison of all compounds results in the extraction of features summarized in Table II.

TABLE II
BIOLOGICAL ACTIVITY FEATURES

No.	Features	Numbers of occurencies		Statistical characteristics	
		l_1^r	l_2^r	$P(A_1/\varphi^r)$	$P(A_2/\varphi^r)$
1	(R_1 = H)&(R_3 = Br)	2	0	0.75	0.25
2	(R_4 = C_2H_5)	2	1	0.60	0.40
3	(R_1 = CH_3)&(R_2 = NH_2)	2	0	0.75	0.25
4	(R_1 = H)&(R_4 = NO_2)	0	2	0.25	0.75
5	(R_1 = CH_3)&(R_2 = C_6H_5)	0	2	0.25	0.75

Statistical estimation of these features requires that occurrence numbers be calculated for each feature present on the compounds of classes A_1 and A_2: l_1^r and l_2^r. Feature φ^1 is shared, for example, by two class 1 compounds, but is absent in the compounds of class 2, hence $l_1^r = 2$ and $l_2^r = 0$.

The values l_1^r and l_2^r are used in Eq. (3) to calculate probability parameters of the features, viz. $P(A_1/\varphi^1 = (2 + 1)/(2 + 2) = 0.75$ and $P(A_2/\varphi^1 = 1/(2 + 2) = 0.25$. Note that the occurrence frequency of the feature in question equals one in the class 1 compounds and zero in the compounds of class 2. However, occurrence of a feature in 2 active compounds is far from being sufficient to claim that its presence would help to discriminate, unequivocably, between active and inactive compounds. The probability (0.75 in this particular case) that the new compound possessing this feature would be active, as determined by Eq. (3), appears more practical.

Assuming that the threshold parameters, λ_1 and λ_2, are equal to 0.75, 1 and 3 appear to be the discriminating features of activity, whereas 4 and 5 are discriminating inactivity features, feature 2 being nondiscriminating.

4. *Activity Prediction for New Compounds*

The selected discriminating features can be used to predict biological activity of chemical compounds, as exemplified in Table III.

TABLE III
BIOLOGICAL ACTIVITY PREDICTION

No.	R_1	R_2	R_3	R_4	Prediction
9	CH_3	NH_2	Br	NH_2	Active (feature 3)
10	CH_3	C_6H_5	H	C_2H_5	Inactive (feature 5)
11	CH_3	CH_3	H	H	Prediction reject

It is evident from the table that compound 9 has been classified as active because activity feature 3 is present on it, whereas compound 10 was ranked within the inactive class because of the presence of inactivity feature 5. On the other hand, compound 11 cannot be included in either class since it lacks discriminating features.

Discriminating activity features can be a useful basis for the synthesis of new, potentially active compounds. For example, the substituents R_2 and R_4 are "inessential" parameters of the feature φ^1 and may assume various values, resulting in a set of new, potentially active compounds, such as (H, NH_2, Br, C_2H_5) and (H, C_6H_5, Br, NO_2).

C. IMPLEMENTATION OF LSA ALGORITHMS

This section outlines the general concepts of LSA. In practice, realization of this approach requires the construction of complex algorithms for logical and statistical analysis of experimental data on the structure–activity relationships in chemical compounds. The usefulness of these algorithms is measured by the degree of their simplicity and reliability in handling by a chemist without prior training in computer technology. To meet these requirements, several program packages have been devised for the elaboration of LSA algorithms. The packages follow, to a certain extent, the pattern of developments in LSA and are based on diverse feature-selection algorithms. The algorithms use various structure representations making their application to the solution of widely different problems possible. The STRAC package, for example, can be employed for lead optimization, whereas the ORACLE package is applicable to lead generation. A more detailed account of the principles and specific applications of LSA will be presented in the next section.

III. Activity Feature Selection within Series of Chemical Compounds

A. INTERACTIVE PROCEDURE OF FEATURE SELECTION AND STRUCTURE DESCRIPTION

In the initial approach to any particular SAR problem, it is difficult to select *a priori* the structural characteristics specifically responsible for the activity manifestation in a given group of compounds. Consequently, it appears logical to combine in a single procedure both structure description and feature selection. The set of structure parameters must be accumulated by the investigator through interaction with the computer in such a manner as to provide means for the establishment of required relationships. The procedure for such an approach is now examined.

1. *Description of the Procedure*

1. To begin with, a starting set of structure parameters is selected that includes the simplest and most easily definable structure characteristics of the compounds studied.

2. The next step involves selection of the logical features in individual activity classes. An algorithm described in Section II is used in this process.

In many cases, the starting set appears to be insufficient for classification purposes and, hence, the discriminating features describe only part of the compounds studied. For this reason the next steps of the procedure

are aimed at expanding the starting set and converting nondiscriminating features into discriminating ones, making use of additional parameters.

3. A nondiscriminating feature is selected and displayed or printed out with the compounds on which it is present.

4. The investigator performs analysis of the selected parts of compounds and extracts a parameter in accordance with the following rules:

(a) An additional parameter is chosen which permits conversion of a nondiscriminating feature into a discriminating one. If such a parameter can be found go to (c), otherwise go to (b). (b) A structural parameter is selected that is typical of the compounds of one class and is present in a number of compounds in some other class. In case of success go to (c), otherwise go to (d). (c) Enter into the computer, the denomination of the new structural parameter and indicate which of the compounds analyzed at this stage contain this parameter. (d) Enter into the computer, information on the impossibility of extracting an additional parameter.

5. Two new features are now generated in the computer: (a) conjunction of the initial nondiscriminating feature and the new parameter and (b) conjunction of the initial nondiscriminating feature and negation of this parameter. Statistical characteristics of these features are then determined. For nondiscriminating features steps 3–5 are performed. The discriminating features are stored. New discriminating features are generated in the course of the decision tree formation similar to the one described by Hunt *et al.* (*28*). Nevertheless, in distinction to the CLS (*28*) program, the decisions in each tree node are made by the investigator on the basis of a nonformal analysis.

6. Extraction of parameters is continued until the whole set of nondiscriminating features has been exhausted or until a set of discriminating features has been obtained which describes all compounds studied.

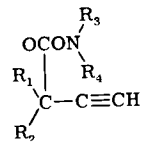
7. In the case in which the list of discriminating features has been exhausted and some of the compounds remain undescribed, these compounds are analyzed according to 4 and 5.

8. When the procedure of parameter extraction and feature generation in the interactive mode have been completed, the values of the selected parameters must be entered for all compounds studied and final feature selection must be performed. This implies the possibility of new informative feature formation.

One can note a similarity between this procedure, originally proposed by Golender and Rozenblit in 1974 (*25*) and the idea of interactive SAR analysis presented by Mathews (*23*) as an alternative to the common pattern recognition techniques (*13*). This similarity confirms the assertion that LSA algorithms simulate the intelligence of an experienced drug designer.

2. *An Example of Interactive SAR Analysis*

The relationship between structure and activity was studied in a series of acetylenic carbamates



considered as potential antitumor agents (*29*). The series under study (presented in Fig. 2) numbers 62 compounds, 44 of which show antitumor activity against tumor C 1498 and 18 being inactive.

ACETYLENIC CARBAMATES

COMPOUNDS OF 1 CLASS

N	NU	R1	R2	R3	R4
1	5	C6H5	C6H5	CH3	H
2	6	C6H5	C6H5	C2H5	H
3	7	C6H5	C6H5	CH2CH=CH2	H
4	8	C6H5	C6H5	CH2C-=CH	H
5	9	C6H5	C6H5	CYCLOPROPYL	H
6	12	C6H5	C6H5	CH2CH2C6H5	H
7	13	C6H5	C6H5	CYCLOPENTYL	H
8	14	C6H5	C6H5	CYCLOHEXYL	H
9	15	C6H5	C6H5	CYCLOHEPTYL	H
10	16	C6H5	C6H5	CYCLOOCTYL	H
11	17	C6H5	C6H5	(CH2)3N(CH3)2	H
12	18	C6H5	C6H5	4-CLC6H4	H
13	19	C6H5	C6H5	NH2	H
14	20	C6H5	C6H5	CH3	CH3
15	24	4-CLC6H4	C6H5	H	H
16	25	4-CLC6H4	C6H5	CH3	H
17	26	4-CLC6H4	C6H5	CH2CH=CH2	H
18	27	4-CLC6H4	C6H5	CYCLOHEXYL	H
19	28	4-CLC6H4	C6H5	CH3	CH3
20	31	3-CLC6H4	C6H5	CH3	H
21	32	3-CLC6H4	C6H5	CH3	CH3
22	34	2-CLC6H4	C6H5	CH3	CH3
23	35	4-CLC6H4	4-CLC6H4	H	H
24	36	4-CLC6H4	4-CLC6H4	CH3	H
25	37	4-CLC6H4	4-CLC6H4	CH3	CH3
26	45	4-BRC6H4	C6H5	H	H
27	46	4-BRC6H4	C6H5	CH3	H
28	47	4-BRC6H4	C6H5	CYCLOHEXYL	H
29	48	4-BRC6H4	C6H5	CH3	CH3
30	50	3-BRC6H4	C6H5	H	H
31	51	3-BRC6H4	C6H5	CH3	CH3
32	52	4-FC6H4	C6H5	H	H
33	53	4-FC6H4	C6H5	CH2C-=CH	H
34	54	4-FC6H4	C6H5	CYCLOHEXYL	H

Fig. 2. Data set for SAR analysis.

35	55	4-FC6H4	C6H5	CH3	CH3
36	57	4-FC6H4	4-FC6H4	H	H
37	58	4-FC6H4	4-FC6H4	CYCLOHEXYL	H
38	59	4-FC6H4	4-FC6H4	CH3	CH3
39	61	4-IC6H4	C6H5	H	H
40	66	4-C6H5C6H4	C6H5	H	H
41	67	4-C6H5C6H4	C6H5	CH3	CH3
42	68	2-C10H7	C6H5	H	H
43	69	2-C10H7	C6H5	CH3	CH3
44	71	1-C10H7	C6H5	H	H

COMPOUNDS OF 2 CLASS

N	NU	R1	R2	R3	R4
45	2	C6H5	CH3	H	H
46	3	C6H5	H	H	H
47	10	C6H5	C6H5	CH2CH2OH	H
48	11	C6H5	C6H5	CH2C6H5	H
49	30	3-CLC6H4	C6H5	H	H
50	33	2-CLC6H4	C6H5	H	H
51	38	3,4-CL2C6H3	C6H5	H	H
52	39	3,4-CL2C6H3	C6H5	CH3	CH3
53	41	2,4-CL2C6H3	C6H5	H	H
54	42	2,4-CL2C6H3	C6H5	CH3	CH3
55	43	3,4-CL2C6H3	4-CLC6H4	H	H
56	44	2,4-CL2C6H3	4-CLC6H4	H	H
57	62	4-CF3C6H4	C6H5	H	H
58	65	4-CH3C6H4	C6H5	H	H
59	72	2-C5H4N	C6H5	H	H
60	73	2-C5H4N	C6H5	CH3	H
61	75	2-C5H4N	C6H5	CH3	CH3
62	76	2-C4H2S	C6H5	H	H

Fig. 2. (*continued*)

A total of 33 substituents have been investigated in this series of compounds.

The types of substituents in positions 1, 2, 3, and 4 were selected as the starting set of parameters. Following the algorithm described in Section II, discriminating features were selected and are given below (the threshold values, λ_1 and λ_2, were chosen to equal 0.83 and 0.80, respectively).

No.	Feature	l_1^r	l_2^r	$P(A_1/\varphi^r)$	$P(A_2/\varphi^r)$
1.1	(R_3 = Cyclohexyl)&(R_4 = H)	5	0	0.86	0.14
1.2	(R_1 = 4-ClC_6H_4)&(R_2 = C_6H_5)	8	0	0.90	0.10
1.3	(R_1 = 4-BrC_6H_4)&(R_2 = C_6H_5)	4	0	0.83	0.17
1.4	R_1 = 4-FC_6H_4	7	0	0.89	0.11
1.5	R_1 = 3,4-$Cl_2C_6H_3$	0	3	0.20	0.80
1.6	R_1 = 2,4-$Cl_2C_6H_3$	0	3	0.20	0.80
1.7	(R_1 = 2-C_5H_4N)&(R_2 = C_6H_5)	0	3	0.20	0.80

These features describe approximately one-half of the studied compounds. A set of nondiscriminating features, presented in part below, was also selected.

No.	Feature	l_1^r	l_2^r	$P(A_1/\varphi^r)$	$P(A_2/\varphi^r)$
2.1	$(R_2 = C_6H_5)\&(R_3 = H)\&(R_4 = H)$	8	8	0.50	0.50
2.2	$(R_3 = CH_3)\&(R_4 = CH_3)$	11	3	0.75	0.25
2.3	$(R_2 = C_6H_5)\&(R_4 = CH_3)$	13	4	0.74	0.26
2.4	$(R_2 = C_6H_5)\&(R_4 = H)$	29	11	0.71	0.29

The following description is the procedure for the generation of new discriminating features based on the nondiscriminating feature 2.1.

1. The compounds possesing nondiscriminating feature $(R_2 = C_6H_5)\&(R_3 = H)\&(R_4 = H)$ are selected and printed. Since the types of substituents on positions 2, 3, and 4 are fixed, only the substituents in position 1 are printed out.

Class A_1 compounds		Class A_2 compounds	
No.	R_1	No.	R_1
24	$4\text{-}ClC_6H_4$	30	$3\text{-}ClC_6H_4$
45	$4\text{-}BrC_6H_4$	33	$2\text{-}ClC_6H_4$
50	$3\text{-}BrC_6H_4$	38	$3,4\text{-}Cl_2C_6H_3$
52	$4\text{-}FC_6H_4$	41	$2,4\text{-}Cl_2C_6H_3$
61	$4\text{-}IC_6H_4$	62	$4\text{-}CF_3C_6H_4$
66	$4\text{-}C_6H_5C_6H_4$	65	$4\text{-}CH_3CH_4$
68	$2\text{-}C_{10}H_7$	72	$2\text{-}C_5H_4N$
71	$1\text{-}C_{10}H_7$	76	$2\text{-}C_4H_2S$

The investigator subjects these compounds to analysis, selects a new structural parameter—*Halogen in* R_1, and indicates that it is present in compounds Nos. 24, 45, 50, 52, 61, 30, 33, 38, 41, and 62. New features are generated in the computer (1) $(R_2 = C_6H_5)\&(R_3 = H)\&(R_4 = H)\&$(Halogen in R_1) and (2) $(R_2 = C_6H_5)\&(R_3 = H)\&(R_4 = H)\&$(No Halogen in R_1).

2. Compounds possessing the nondescriminating feature (1) are selected and printed out.

Class A_1 compounds		Class A_2 compounds	
No.	R_1	No.	R_1
24	$4\text{-}ClC_6H_4$	30	$3\text{-}ClC_6H_4$
45	$4\text{-}BrC_6H_4$	33	$2\text{-}ClC_6H_4$
50	$3\text{-}BrC_6H_4$	38	$3,4\text{-}Cl_2C_6H_3$
52	$4\text{-}FC_6H_4$	41	$2,4\text{-}Cl_2C_6H_3$
61	$4\text{-}IC_6H_4$	62	$4\text{-}CF_3C_6H_4$

The analysis of these compounds shows that the representatives of class A_1 typically possess *4-substituted phenyl derivatives in* R_1. The investigator prints out the name of the new structural parameter and the numbers of compounds in which it is present: 24, 45, 52, 61, 38, 41, and 62. New features are generated in the computer—(3) (R_2 = C_6H_5)&(R_3 = H)&(R_4 = H)&(Halogen in R_1)&(4-substituted phenyl radicals in R_1) and (4) (R_2 = C_6H_5)&(R_3 = H)&(R_4 = H)&(Halogen in R_1)&(Not 4-substituted phenyl radicals in R_1).

3. Compounds possessing nondiscriminating feature (3) are selected and displayed.

Class A_1 compounds		Class A_2 compounds	
No.	R_1	No.	R_1
24	$4\text{-}ClC_6H_4$	38	$3,4\text{-}Cl_2C_6H_3$
45	$4\text{-}BrC_6H_4$	41	$2,4\text{-}Cl_2C_6H_3$
52	$4\text{-}FC_6H_4$	62	$4\text{-}CF_3CH_4$
61	$4\text{-}IC_6H_4$		

The investigator notes that class A_1 compounds possess one atom of halogen, whereas those in class A_2 contain several halogen atoms, and he enters the name of the structural parameter into the computer—*Monohalogen substituted phenyl radicals in* R_1, indicating that it is present in compounds 24, 45, 52, and 61.

The following features are generated in the computer (5) (R_2 = C_6H_5)&(R_3 = H)&(R_4 = H)&(Halogen in R_1)&(4-substituted phenyl radicals in R_1)&(monohalogen substituted phenyl radicals in R_1) and (6) (R_2 = C_6H_5)&(R_3 = H)&(R_4 = H)&(Halogen in R_1)&(4-substituted phenyl radicals in R_1)&(not monohalogen substituted phenyl radicals in R_1). It must be noted that both these features are discriminating.

Subsequently, the investigator analyses the compounds with nondiscriminating features (4) and (2), which results in the generation of new

structural parameters—*monochloride derivative in* R_1 and R_1 = *bulky substituent*.

Figure 3 shows a decision tree reflecting these steps in the feature-selection procedure. The names of the structural parameters extracted by the investigator are given in the nodes of the decision tree. A dialog based on other nondiscriminating features can be conducted in the same manner.

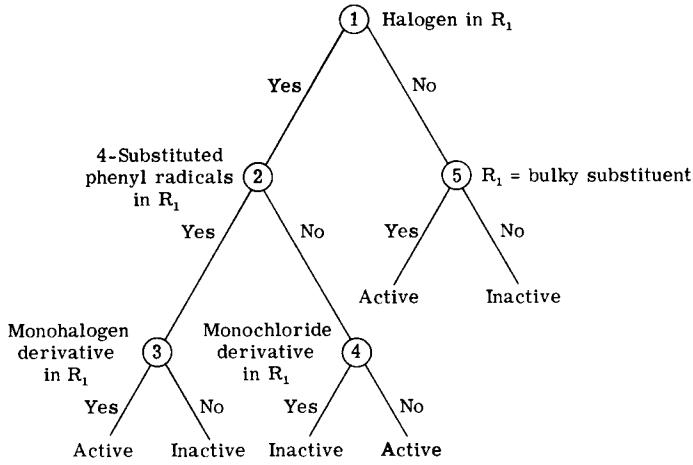


Fig. 3. Decision tree formed during interactive procedure.

New parameters, selected in the course of the dialog, are entered into the compound descriptions and the feature-selection procedure is repeated. This results in the selection of new discriminating features, for instance, R_1 = *4-monohalogen substituted phenyl,* with statistical characteristics $l_1 = 20$, $l_2 = 0$, $P(A_1/\varphi) = 0.95$, $P(A_2/\varphi) = 0.05$.

Interestingly, the authors of this series of compounds have not found this nor some other strong activity features (*25, 4*).

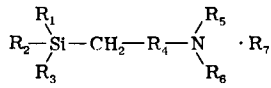
In addition to the aforesaid, the same series of compounds has been subjected to analysis by the Free-Wilson method (*30*). Application of that method in this case results in a regression equation with relatively high values of the explained variance. However, this is in fact a misleading approximation, since most substituents have been found only on one compound of the analyzed series (*30*). Contributions to activity from such substituents are statistically insignificant. Nevertheless, the interactive procedure allows, as was demonstrated, the identification of structural fragments shared by various substituents, which can be further used for the selection of statistically significant biological activity features in acetylenic carbamates.

B. THE STRAC SOFTWARE PACKAGE AND ITS PRACTICAL APPLICATIONS

The feature-selection procedure previously described can be implemented by virtue of the STRAC software package developed at the Institute of Organic Synthesis in Riga (USSR) (*4*). The STRAC package, coded in FORTRAN, was executed on a HP 2116 C mini-computer. A modification of the STRAC package has been developed at the G. Donegani Institute (Italy) to be applied in the Univac 1100 computer system (*31*).

The STRAC package is constructed so as to permit the user to edit and expand the experimental data files. Additional information can be fed into the computer during the course of operation in the interactive mode.

The STRAC package has been applied to the analysis of a new class of antimicrobial compounds, alkylaminosilanes, synthesized by Lukevics, Voronkov, and co-workers (*32*). The logico-structural analysis of the structure–antimicrobial potency relationship in this group of chemicals made use of the results obtained in four activity assays, T_1–T_4, carried out on 128 compounds (*33*). The compounds in question have the general formula:



During the SAR analysis, a compound was viewed as active unless its lowest concentration inhibitory to the growth of the test-microorganism exceeded 17 μg/ml, otherwise such a compound was ranked as inactive.

In accordance with the strategy of the LSA, initial feature selection was performed for the compounds described by the starting set of structural parameters (viz. the list of substituent types in positions 1–7). The resulting features are not sufficient for classification of the compounds according to their activity. Thus, for the activity threshold T_2 = 17 μg/ml, 15% errors and 22% classification rejections in the class of active compounds are observed. For the class of inactive compounds, the appropriate values are 11% and 20%, respectively. Two additional structure parameters were revealed in the course of the dialog procedure: (a) the total number of carbon atoms in the radicals attached to silicon (R_1–R_3), within 8–13 range and (b) the presence of an alkyl radical in position 6.

The results of the final feature selection accomplished for two assays are given in Table IV.

The decision rules, based on the selected activity features, are characterized by high stability, as evidenced by the experiments in which both feature selection and activity prediction were performed on two different

TABLE IV

FEATURES RESPONSIBLE FOR ANTIMICROBIAL ACTIVITY IN ORGANOSILICON AMINES $R_1R_2R_3SiCH_2R_4NR_5R_6 \cdot R_7$

Activity criterion	r	R_1	R_2,R_3	R_4	R_5	R_6	R_7	Σn_c R_1–R_3	l_1^r	l_2^r	$P(A_1/\varphi^r)$	$P(A_2/\varphi^r)$
$T_1^a \leq 17$	1.1	∅	∅	CH_2CH_2	∅	Alk	∅	8–13	17	20	0.46	0.54
20 active	1.2	∅	∅	∅	H	Alk	—	8–13	12	11	0.52	0.48
compounds	1.3	∅	∅	CH_2CH_2	H	Alk	∅	8–13	12	11	0.52	0.48
108 inactive	1.4	CH_3	∅	∅	∅	Alk	—	8–13	13	11	0.54	0.48
compounds	1.5	CH_3	∅	CH_2CH_2	∅	Alk	∅	8–13	13	13	0.50	0.50
$P(A_1) = 0.16$								not 8–13	2	69	0.13	0.87
$P(A_2) = 0.84$												
$T_2^b \leq 17$	2.1	∅	∅	CH_2CH_2	∅	∅	∅	8–13	27	24	0.53	0.47
27 active	2.2	∅	∅	CH_2CH_2	∅	Alk	∅	8–13	23	14	0.62	0.38
compounds	2.3	∅	∅	CH_2CH_2	∅	∅	—	8–13	26	19	0.57	0.43
101 active	2.4	∅	∅	CH_2CH_2	∅	Alk	—	8–13	22	11	0.66	0.34
compounds	2.5	∅	∅	CH_2CH_2	H	Alk	∅	8–13	15	8	0.64	0.36
$P(A_1) = 0.21$	2.6	∅	∅	CH_2CH_2	∅	∅	—	8–13	15	8	0.64	0.36
$P(A_2) = 0.79$	2.7	∅	∅	∅	∅	∅	∅	not 8–13	0	71	0.01	0.99

[a] Test-microorganism *Candida albicans* 67/846.

[b] Test-microorganism *Epidermophyton Kaufman-Wolf* 41.

groups of chemicals with 64 compounds in each group. The results of prediction, averaged for 10 random divisions of 128 compounds into two groups, are presented in Table V.

TABLE V
ACTIVITY PREDICTION RESULTS FOR 64 COMPOUNDS FROM ORGANOSILICON AMINES SERIES (ACTIVITY CRITERION $T_2 \leq 17$)[a]

Errors (%)			Prediction rejects (%)				
A_1	A_2	Mean	A_1	A_2	Mean	K_e	P_A
0 ± 0	23 ± 5	17 ± 5	5 ± 4	7 ± 4	6 ± 4	2.4 ± 0.2	1.0 ± 0

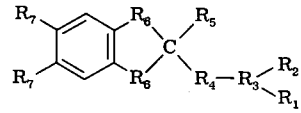
[a] 95% confidence interval.

This experiment demonstrates that by applying a prediction procedure, a representative group of compounds comprising only half of the given series need be investigated to reduce the expenditures necessary for screening new compounds by half, without loss of active compounds.

The application of the STRAC package to the SAR analysis of a series of organosilicon amines has been subjected to careful examination, resulting in the following conclusions (*34*): (1) the total sum of carbon atoms (ranging from 8 to 13) implies a high probability of active compounds; (2) active compounds are more likely to contain three methyl groups located between silicon and nitrogen; and (3) antimicrobial activity tends to decline on transition from saturated to ethylenic and acetylenic amines.

Bearing these limitations in mind, 12 organosilicon derivatives of pyrrolidine, piperidine, and perhydroazepine have been synthesized. Three of the compounds were characterized by the presence of inactive features; the remaining ones carried activity features. *In vitro* antimicrobial activity assays found that the three compounds in the first group were barely active, whereas seven out of the nine compounds in the second group exhibited high fungistatic activity.

The STRAC package has been also successfully applied to the analysis of SAR in a series of indandione-1,3 derivatives (*35*).



Analysis of 139 compounds in the series revealed essential structure parameters and identified a pharmacophoric pattern responsible for anticonvulsive effect (*4*).

A specially designed activity prediction experiment has been performed on a series of indandione-1,3 piperazine derivatives. Feature selection was carried out on 68 compounds studied and the results used to predict the activity and toxicity of 15 newly synthesized compounds that were already undergoing biological trial. Toxicity prediction was correct in 11 cases (73%). Activity was predicted for five compounds only, since the remaining compounds possessed none of the discriminating features. Prediction was correct in the case of four compounds, and wrong for one, however, the latter's activity was close to the threshold value.

IV. Activity Prediction Using a Data Base on Chemical Structure and Biological Activity

Successful application of statistical methods for lead generation is only feasible with a data base for the structure and activity of chemical compounds. Similar types of data bases have generally been created for information retrieval and applied only sporadically to the study of either structure–activity relationships or to biological activity prediction (*7, 12, 36*).

A structure and pharmacological activity data base specifically designed for SAR analysis was set up in 1976 at the Institute for Biological Screening of Chemical Compounds in Moscow and at the Institute of Organic Synthesis in Riga (*37*). The file, numbering about 5000 medicinal compounds, was formed by the data derived from Negver's Handbook (*38*), supplemented by 1000 newly synthesized, active compounds. The data base contains information on 55 major types of pharmaceutical activity. The chemical structures are coded by means of substructure superposition fragment notation (SSFN) developed by Avidon and co-workers (*39*).

A. STRUCTURE DESCRIPTION AND ACTIVITY PREDICTION ALGORITHM

The activity prediction algorithm, using a data bank, is based on the main principles of LSA, although the algorithm has some peculiarities because of the large amount of information stored in the data base and the characteristics of SSFN (substructure superposition fragment notation) employed for structure representation.

The basic principles of this notation, designed for structure representation of biologically active compounds, are now described.

1. *Substructure Superposition Fragment Notation (SSFN)*

The coding of chemical structures using SSFN is based on the hypothesis which postulates that the biological effects of compounds are determined by the capacity of certain active centers to bind to the receptor. In accordance with this concept, descriptor centers are selected in the molecule. They include potential active centers, like all heteroatoms and aromatic systems as integrated entities, as well as ≡CH, =CH_2, and —CH_3 groups. A linear descriptor serving to code any fragment that is started and ended by a descriptor center is the principal unit of structure notation. It consists of four parts, schematic representation of which is given by:

Code of descriptor center C_1	Number of carbons between the centers C_1 and C_2	Code of descriptor center C_2	Conjugation feature along the chain
2 digits	2 digits	2 digits	1 digit

As shown, each descriptor is coded by a 7-digit number. The conjugation feature equals 1 when conjugation is observed and 0 in the absence thereof. Descriptor centers and their notations are summarized in Table VI. The coding of a compound involves the listing of all chains starting with and ending in an active center, but not of those chains running across them. The sequence in which descriptors are cited is defined by several rules, the descriptors being ranked in the order of increasing 7-digit numbers.

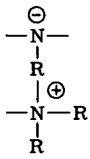
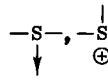
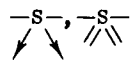
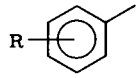
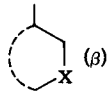
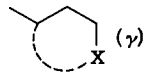
Apart from the linear descriptors, SSFN makes use of descriptor coding of the cyclic moieties contained in the compound. A cyclic descriptor consists of four parts as shown by:

Ring shape notation	Delimiter (comma)	Number of π-electrons in the cyclic conjugated system	Notation of heteroatom location
varying number of characters	1 character	2 numbers	varying number of characters

An example of the coding procedure is presented in Fig. 4.

Such notation has been easy in operation. Manual coding of a single

TABLE VI
SSFN DESCRIPTOR CENTERS

Group	Code	Group	Code	Group	Code
	00 01	—SH —SR	21 22	$-CH_3$ $=CH_2$ $\equiv CH$	41
$-NH_2$, —NHR	02		23	Md (metalloid)	42
	03		24	Met (metal)	43
$=N-H$	04	$=S$	25		
$=N-R$	05	Cl, Br, I	31		44
$\equiv N$	06	F	32	$-C\equiv X$ in aromatic systems	
	07		33		
—OH	11				
—OR	12		34		
$=O$	13				
	15		35		
	16		36		
			37		
			40		

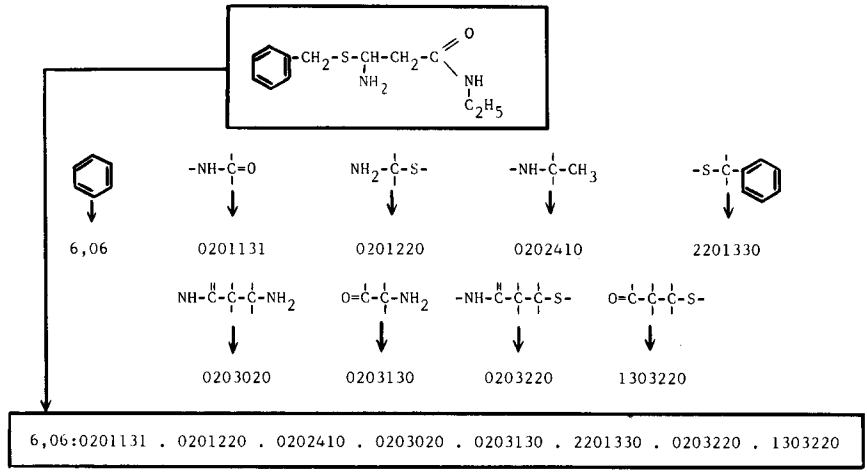


Fig. 4. Example of structure coding procedure using SSFN.

compound takes, on the average, 2–2.5 minutes. Efforts are now aimed at developing computer programs to handle this coding.

SSFN is not immune to the typical shortcomings of fragment notation. For instance, a certain loss of information concerning the relative location of descriptor centers in the molecule is observed upon transition from the structural formula to SSFN. In order to eliminate this drawback, the method has been improved by including the topology of descriptor centers in the molecular structure notation (see Section V).

2. *Feature Selection Algorithm*

Selection of activity features in compounds coded by SSFN is characterized by some peculiarities imposed by this notation. In principle, selection of features-conjunctions (descriptor combinations) can be accomplished by means of the basic procedure of pair-wise intersection of compound representations. However, the features chosen by this procedure are often of rare occurrence and, hence, are statistically insignificant. To select features of more frequent occurrence, it appears preferable to generate descriptor disjunctions.* At the same time, generation of these features requires substantial computational expenditure, especially when working with a data base. For this reason we have used, as the first approximation, an algorithm in which SSFN descriptors are implied to be

* $\bigcup_{i=1}^{m} D_i = D_1 \vee D_2 \vee, \ldots, D_m$, where $D_1, \ldots, D_m$ = descriptors, $\cup$ = the disjunction sign, $\vee$ = logical OR. We assume the descriptors to be chemically close.

the features. It must be pointed out that linear descriptors can in fact be regarded as conjunctive expressions of the following type.

The molecule contains (descriptor center 1)&(descriptor center 2)&(the length of the chain connecting center 1 and 2 equals L)&(conjugation is (not) observed along the chain).

Two statistical criteria for feature estimation have been applied in order to choose descriminating features in the substructures corresponding to SSFN descriptors.

A descriptor is considered to represent a particular activity feature if its presence on this activity is not a chance event. Furthermore, in stricter terms, this criterion can be formulated as follows. Let l_k^r compounds with A_k activity be contained among the l^r compounds with φ^r descriptor. The probability P_k^r of chance occurrence of less than l_k^r compounds in the A_k class can be derived from the expression:

$$P_k^r = \sum_{m=0}^{l_k^r - 1} \binom{l^r}{m} [P(A_k)^m (1 - p(A_k))^{l^r - m}]$$

where $P(A_k)$ is prior probability of the A_k class.

At comparatively high P_k^r values, the probability of chance occurrence of l_k^r or a greater number of compounds is remote and, therefore, it allows one to conclude that the presence of φ^r descriptor implies the presence of A_k activity. Thus, the following requirement must be met for the criterion of feature selection

$$P_k^r > \omega_k$$

where ω_k stands for the threshold value.

Selection of descriminating features makes use of still another criterion, the coefficient of prognostic feature utility [Eq. (4)]. A feature φ^r is classified as discriminating if

$$U_k^r > \lambda_k$$

where λ_k denotes a certain threshold value.

The threshold values ω_k and λ_k are chosen for all activities by optimization of K_e and P_A criteria [see Eqs. (1) and (2)], using the leave-one-out technique. In the activity prediction procedure, making use of a data base, the prediction system is considered acceptable if it provides reliable recognition of active compounds based on the condition:

$$P_{A_k} \geq 0.9 \qquad k = 1, \ldots, N$$

where N (number of activities) equals 55 for this particular case.

Among the threshold values meeting this requirement, one has to choose those with the greatest efficiency coefficient K_e.

3. *Activity Prediction Algorithm*

The activity prediction procedure involves calculation of P_k^r and U_k^r, $k = 1, \ldots, N$ parameters for each φ^r descriptor contained in the compound. A compound is assumed to possess A_k activity if these parameters exceed the threshold values, ω_k and λ_k.

The results of activity prediction are obtained in an intellectually manageable form, with all substructures corresponding to the descriminating activity features and their statistical characteristics being printed out. Thus, the drug designer is provided with a good basis for further meaningful feature evaluation.

Activity prediction is also accompanied by the following numerical criteria.

1. The confidence coefficient is a measure of probability that a given activity is possessed by the compound in question.

$$\delta_k = \max P(A_k/\varphi^r) \qquad k = 1, \ldots, N$$

where r denotes the numbers of all descriminating features responsible for A_k activity in the compound.

2. The prediction efficiency coefficient is given by:

$$e_k = \max U_k^r \qquad k = 1, \ldots, N$$

where δ_k and e_k provide the investigator with additional information which can be used for qualititative evaluation of the results of computer prediction. The activities characterized by equally high confidence and efficiency coefficient values are of the greatest prognostic value. When the confidence coefficient has a low value, the efficiency coefficient being high, it is advisable to undertake activity testing only in the case of special interest.

B. ORACLE SOFTWARE PACKAGE AND RESULTS OF ITS APPLICATION

The data base (*36*) storing data files on both known drugs (4962 compounds) and newly synthesized highly active substances (1002 compounds) provides a good foundation for the performance of a wide range of activity prediction experiments. Four such experiments, accomplished by virtue of the ORACLE package, are now presented.

Experiment 1 involves threshold optimization by the leave-one-out recognition technique, carried out on a file of drugs. The results of the exper-

iment are summarized in Table VII. Threshold selection permits correct recognition of active compounds in about 91% of cases. The efficiency coefficient of the prediction system equals 2.3. The selected activity features appear fairly reliable, judged by the results, thus permitting one to proceed to the next step of the analysis.

Experiment 2 predicts biological activity on a file of active compounds, making use of the optimal threshold values found in Experiment 1 (see Table VII). Leaving aside results with statistically insignificant small sample classes, low probability of correct recognition of active compounds was found on activities 25 (P_A = 0.67) and 38 (P_A = 0.65). Note that activity 25 numbers twice as many compounds in the prediction set as compared with the training one. For an explanation of high probability of prediction errors on activity 38, see the following discussion. The average probability of correct recognition estimated for all classes was 0.8 and the mean efficiency coefficient was 1.94.

One of the reasons for the prediction errors lies in the poor extrapolation of some features. Several descriptors recognized as inactivity features appear to determine activity on active compounds. For example, for descriptor 5.06S1 (tryptophan), the ratio of the number of compounds with activity 29 to the total number of compounds is only 4/54 for drugs and 26/61 for active compounds. The observed phenomenon may be due to (1) the lack of information in the data base on certain types of biological activity possessed in fact by some compounds or (2) the influence of some structural parameters not taken into account by the algorithm and the notation used. Occurrence of compounds with new substructures during the test may also result in prediction errors. For example, the 11 active compounds possessing activity 38 are the only ones that contain the cyclic descriptor 56,10M1M3. Nevertheless, this descriptor is not present on any of the drugs. It must be noted, however, that similar features are very rare.

Experiment 3 involves optimization of threshold values by the leave-one-out technique, performed on the complete set of compounds. Higher accuracy of the leave-one-out method is attained in this case owing to the larger sample size, as compared with Experiment 1. The mean correct prediction probability amounts to 0.91, the mean efficiency coefficient being 2.61.

Analysis of the results obtained in Experiments 1–3 reveals a surprising phenomenon. Large-size classes tend to be prone to a greater number of second-kind errors and to exhibit a lower prediction efficiency coefficient than small-size ones (see Table VIII). This effect may be due to several factors. In the first place, the large classes in many cases consist of a

TABLE VII

RESULTS OF ACTIVITY PREDICTION EXPERIMENTS 1 AND 2

		Experiment 1						Experiment 2			
k	Activity	L_k	ω_k	λ_k	P_{A_k}	K_{e_κ}	P_{2k}	L_k	P_{A_k}	K_{e_κ}	P_{2k}
1	α-Adrenergic blocking agents	38	0.700	3.0	0.92	1.8	0.48	4	1.00	2.3	0.43
2	β-Adrenergic blocking agents	50	1.000	2.5	0.90	3.6	0.24	24	0.92	4.7	0.17
3	Central adrenergic blocking agents	8	0.980	6.0	1.00	6.5	0.15	0			
4	α-Adrenomimetics	61	0.700	4.0	0.93	4.0	0.22	2	0.50	2.8	0.18
5	β-Adrenomimetics	57	0.700	5.0	0.93	4.2	0.31	3	0.67	4.2	0.16
6	Analeptics	61	0.800	2.5	0.95	2.1	0.44	1	1.00	2.6	0.39
7	Analgesic, narcotic	133	0.700	2.5	0.92	1.8	0.48	35	0.83	1.8	0.45
8	Analgesic, non-narcotic	228	0.700	2.5	0.90	2.0	0.43	34	0.76	1.7	0.44
9	Androgens	5	0.999	5.0	1.00	16	0.06	1	1.00	35	0.03
10	Anesthetics, local	158	0.700	2.0	0.91	1.6	0.57	1	1.00	2.1	0.47
11	Antidepressants	119	0.700	2.0	0.95	1.8	0.51	1	1.00	1.5	0.52
12	Anticoagulants	41	0.999	2.5	0.90	2.3	0.39	4	0.75	2.6	0.29
13	Antimetabolites and metabolites	76	0.999	6.0	0.92	3.0	0.30	12	0.92	2.7	0.33
14	Antiseptics	208	0.600	1.0	0.91	1.7	0.53	25	0.84	1.5	0.54
15	Ganglion blocking agents	67	0.600	1.5	0.90	4.9	0.17	4	0.75	8.4	0.09
16	Hestgens	15	0.999	6.0	1.00	5.2	0.19	2	1.00	8.4	0.12
17	Hypoglycemic, non-steroid	68	0.700	3.0	0.90	3.2	0.27	20	0.90	3.5	0.25
18	Hypocholesterolemic, non-steroid	53	0.700	2.0	0.91	2.1	0.43	7	1.00	3.0	0.33
19	Histamine antagonists	132	0.850	4.0	0.90	2.8	0.31	10	0.60	1.7	0.35
20	Glycocorticoids	34	0.999	6.0	1.00	4.0	0.20	0			
21	Cardiotonic	34	0.700	3.0	0.91	2.2	0.40	1	0.00	0.0	0.28
22	Curaric	55	0.700	2.0	0.96	3.0	0.31	18	0.89	4.3	0.20
23	Coronary dilatants	96	0.850	2.5	0.92	2.3	0.40	2	1.00	2.4	0.42
24	Mineralcorticoids	3	0.999	6.0	1.00	46	0.02	0			
25	MAO inhibitors	22	0.400	2.0	1.00	2.5	0.39	48	0.67	1.4	0.46

26	Neuroleptics	131	0.850	5.0	0.90	4.4	0.19	24	0.75	3.3	0.21
27	Antiallergic	76	0.700	2.0	0.91	1.8	0.50	7	0.29	0.8	0.37
28	Antiarrythmic	57	0.700	2.0	0.93	1.6	0.59	16	0.75	1.5	0.50
29	Antibacterial	396	0.980	2.5	0.90	1.5	0.58	151	0.82	1.3	0.58
30	Antiviral	78	0.700	4.0	0.90	3.0	0.29	33	0.70	2.2	0.30
31	Antiinflammatory, non-steroid	218	0.800	1.8	0.91	1.6	0.57	57	0.93	1.5	0.60
32	Anthelmintic	121	0.400	2.0	0.90	1.9	0.46	15	0.93	1.6	0.56
33	Antifungal	114	0.700	1.0	0.91	1.5	0.61	69	0.80	1.3	0.62
34	Antitussive	84	0.85	1.5	0.93	1.6	0.57	1	1.00	2.2	0.45
35	Antineoplastic	277	0.850	2.5	0.92	1.8	0.48	67	0.85	1.6	0.52
36	Antiprotozoal and antispirochetal	205	0.910	3.0	0.91	3.1	0.27	29	0.79	2.1	0.36
37	Anticonvulsant	110	0.500	1.0	0.90	1.3	0.69	40	1.00	1.5	0.66
38	Antituberculous and antileprotic	110	0.700	2.5	0.91	2.3	0.39	49	0.65	1.6	0.40
39	Psychostimulants	43	0.910	1.8	0.91	1.8	0.49	3	0.67	1.4	0.48
40	Psychomimetics	19	0.800	2.5	0.95	1.9	0.50	1	1.00	2.5	0.39
41	Saluretics	54	0.850	5.0	0.94	3.8	0.24	6	1.00	4.6	0.21
42	Sympatholytics	18	0.400	3.0	0.94	2.9	0.33	2	0.50	1.7	0.30
43	Hypnotic and sedative	241	0.700	1.0	0.92	1.4	0.64	61	0.80	1.3	0.62
44	Vasodilator	149	0.800	1.5	0.90	1.5	0.57	13	0.77	1.2	0.64
45	Vasoconstrictor	46	0.990	6.0	0.91	6.6	0.13	0			
46	Spasmolytics and bronchodilatants	322	0.700	1.5	0.95	1.4	0.67	17	0.88	1.5	0.57
47	Thyroid hormone antagonists	24	0.900	6.0	0.92	4.0	0.23	0			
48	Tranquilizers	111	0.700	1.5	0.94	1.4	0.66	11	0.91	1.5	0.59
49	Uricosuric	24	0.900	6.0	0.92	6.8	0.13	0			
50	Cholinolytic-M	167	0.950	6.0	0.92	3.8	0.22	11	0.55	3.6	0.15
51	Cholinolytic-N	160	0.800	2.5	0.91	2.6	0.33	14	0.71	3.4	0.21
52	Cholinomimetic-M	34	0.700	1.5	0.94	1.9	0.48	3	0.67	1.7	0.39
53	Cholinomimetic-N	27	0.400	3.5	1.00	4.0	0.25	9	0.67	3.2	0.20
54	Cholinesterase inhibitors	38	0.700	1.0	0.92	1.4	0.64	7	0.86	1.6	0.53
55	Estrogens	49	0.999	6.0	0.98	5.1	0.18	6	0.50	4.2	0.12

[a] L_k—number of compounds possessing activity; A_k, P_{2k}—probability of errors of the 2nd kind for the activity A_k.

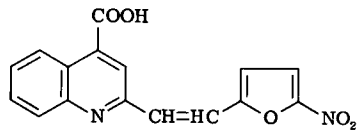
TABLE VIII
AVERAGED RESULTS OF ACTIVITY PREDICTION EXPERIMENTS USING A DATA BASE[a]

No. experiment	For all activities			For rare activities (less than 100 compounds)			For frequent activities (more than 100 compounds)		
	P_A	K_e	P_2	P_A	K_e	P_2	P_A	K_e	P_2
1	0.91	2.3	0.45	0.91	3.0	0.35	0.91	1.9	0.45
2	0.80	1.9	0.47	0.76	2.7	0.32	0.81	1.6	0.52
3	0.91	2.6	0.39	0.93	3.5	0.30	0.91	2.2	0.41

[a] Calculated as weighted averages.

number of isolated subclasses with different mechanisms of biological action. Attempts to find common features for such subclasses result in the selection of descriptors distributed among several classes. Reliability of prediction is therefore likely to be enhanced by a more thorough class division. Furthermore, it seems advisable to strengthen feature selection criteria somewhat in the case of large-size classes. One of the approaches

TABLE IX
ACTIVITY PREDICTION EXAMPLE



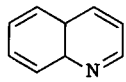
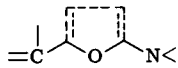
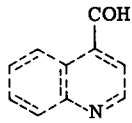
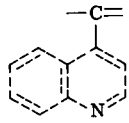
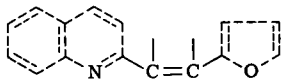
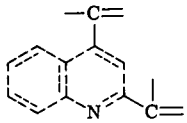
Structure fragment[a]	SSFN descriptor	Total occurence number	Predicted activities and corresponding occurence numbers (in parentheses)
	5,06Q1	120	29(67), 33(13)
	66,10M1	134	8(17), 14(16), 22(6), 25(4), 36(36), 49(8)
NO_2	0100131	281	29(101), 32(18), 33(25), 36(50), 52(4)
$-O-\overset{\|\|}{C}-N<$	0100151	141	29(79), 32(8), 33(14), 36(33)

TABLE IX—*Continued*

Structure fragment[a]	SSFN descriptor	Total occurence number	Predicted activities and corresponding occurence numbers (in parentheses)
	0164441	98	29(59), 33(14), 36(12)
COOH	1101131	689	18(16), 31(87), 35(61)
	1101371	42	8(7), 31(6), 38(22), 49(3)
	1301371	81	8(15), 25(5), 31(11), 38(27), 49(7)
	3502351	16	29(11), 32(3)
	4463440	0 (new descriptor)	—

[a] Heteroaromatic systems corresponding to the descriptor centers 35, 37, 44 are pictured by dotted lines and can be substituted in a descriptor by any heteroaromatic system.

PREDICTION

Activity	Confidence coefficient	Efficiency coefficient
8—analgesic, nonnarcotic, or antipyretic	0.19	4.2
14—antiseptic	0.12	3.1
18—hypocholesterolaemic agent, nonsteroid	0.02	2.3
22—curaric	0.04	3.7
25—MAO inhibitor	0.06	5.3
29—antibacterial	0.69	7.5
31—antiinflammatory, nonsteroid	0.14	3.1
32—anthelmintic	0.19	8.2
33—antifungal	0.14	4.7
36—antiprotozoal or antispirochetial	0.27	6.8
38—antituberculous or antileprotic	0.52	19.6
49—uricosuric	0.09	21.5
52—cholinomimetic-M	0.01	2.3

may be to attempt to reach a reasonable compromise between K_e and P_A. For example, by selecting the appropriate threshold values, it is possible to obtain $P_A = 0.87, K_e = 2.3$ instead of $P_A = 0.90, K_e = 1.8$ for activity 46 in Experiment 3 (356 compounds).

Results similar to those obtained in Experiment 3 were reported (*40*) in describing application of a modified method of substructural analysis (*12*). The prediction algorithm based on the logico-structural approach has some advantage by comparison with the substructural analysis. It is characterized by a logical, easily manageable prediction procedure providing facilities for meaningful evaluation of results.

Experiment 4 has been undertaken to predict the activity of 150 compounds in the furan series being studied at the Institute of Organic Synthesis in Riga (USSR). The optimal threshold values found in Experiment 3 were employed in the prediction procedure. For the purpose of illustration, we shall present the results of activity prediction (summarized in Table IX) obtained for 2-[2′-(5″-nitrofuryl-2)-vinyl]quinoline-4-carboxylic acid. Antibacterial activity substantiated by Miura *et al.* (*41*), was predicted with the highest confidence coefficient values, the nitrofuran ring and the conjugation of quinoline and nitrofuran cycles being the activity features. Antituberculous activity (confidence coefficient 0.52) has also been confirmed (*42*). This activity is determined by the quinoline 4-carboxylic acid. Antiprotozoal activity was determined by the quinoline ring (with higher probability) and several nitrofuran substructures (with lower probability). This activity was reported for the compound in question by Katsuaro *et al.* (*43*). Antifungal activity, also documented in the literature (*42*) (confidence level 0.14), is due to the nitrofuran moiety. It should be noted, however, that our recognition system failed to predict a slight antitumor activity reported in the literature (*44*).

The activity prediction system correctly recognized the presence of earlier established activities for the majority of compounds. And a number of unknown activities were predicted (with high confidence coefficient) for many substances. The data obtained were used for planning further biological experiments, the results of which will be published.

The results of the experiments just described show that the proposed approach is useful for the purposes of lead generation. They also suggest means for improving the method. Particularly useful for the refinement of the prediction system parameters are attempts to improve the language of the chemical structure description and the algorithms of activity feature selection. The possibilities for such refinements will be the subject of the next section.

V. Selection of Topological and Topographical Activity Features

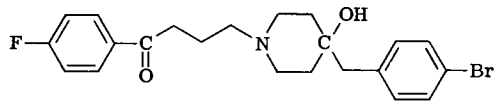
A. STRUCTURE DESCRIPTION

The main drawback of the SSFN resides in the disintegration of structure, since it reflects information on structural fragments chosen as descriptor centers but fails to provide information about their mutual disposition.

In the following method for the description of chemical compounds, advantages of the SSFN are extended by including information on molecular topology. The basic principles of this notation are that (1) potentially active (descriptor) centers in the molecule are identified and coded and that (2) compound descriptions contain information on the distances between all pairs of descriptor centers expressed as the number of bonds in the chain connecting each pair.

The structure of a compound is presented as an undirected graph, the nodes correspond to the appropriate descriptor centers, and the edges denote the shortest distances between adjacent descriptor centers. Descriptor centers are considered adjacent if they are directly connected by a chain of carbon atoms (even if they are part of any cycle). Such structure notation will be referred to as a descriptor center graph (DCG).

The notation is illustrated by DCG construction for the heterocyclic butyrophenon



I

The list of descriptor centers in this case cites fluorine, bromine, oxygen (double bonded), nitrogen (with three single bonds), an OH-group and two benzene rings. The shortest distances between descriptor centers have been estimated to contain six bonds between fluorine and oxygen (three of the bonds are part of the benzene cycle), five bonds between oxygen and nitrogen, etc. A descriptor center graph for the compound in question is depicted in Fig. 5.

Biological activity of chemical compounds is often determined by the total distance between the active centers, regardless of the atom types in the chain connecting these centers. Therefore, it is preferable to describe the structure as a complete graph, i.e., a graph in which each pair of nodes is connected by an edge. The edges denote the shortest distance between descriptor centers.

Such a graph is not very demonstrative and is, therefore, not pictured

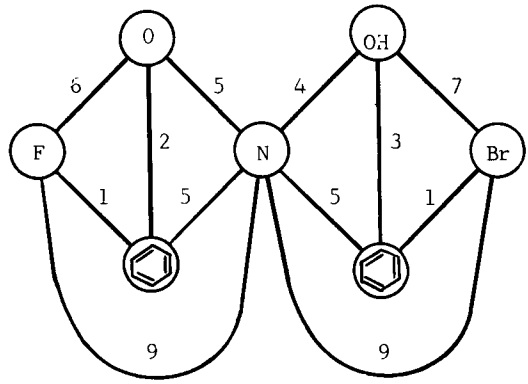


Fig. 5. Descriptor center graph of compound I.

here. Given here is an equivalent representation of the graph in the form of descriptor center matrix (DCM).

32	1	6	9	13	14	18
	33	2	5	9	10	14
		13	5	9	10	14
			03	4	5	9
				11	3	7
					33	1
						31

Diagonal elements of the matrix contain descriptor center codes (see Table VI) whereas the crossing of the Ith row with the Jth column represents the shortest distance between descriptor centers I and J. The matrix is symmetrical with respect to its main diagonal, hence only half of it is presented.

The DCM contains information concerning molecular topology only, which is not always unambiguously connected with the three-dimensional structure of the molecule. However, matrix representation of structure can also be used to describe topography of the molecule. This involves estimation of geometrical distances between charged centers in the molecule used as nondiagonal matrix elements and estimation of parameters of these centers (e.g., charge and volume) used as diagonal elements. Such structure notation will be referred to as a charged center matrix (CCM).

The distances between charged centers can be estimated on the basis of X-ray analysis and by conformational calculations. As with DCM, it is expedient to combine atoms of planar aromatic systems into a single

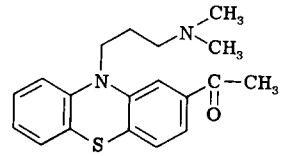
charged center and to estimate distances with respect to its center of gravity.

B. FEATURE SELECTION ALGORITHM

The basic principles of the logico-structural approach are also applicable to the chemical structures represented by descriptor center graphs (matrices). The general pattern of the feature selection procedure remains unaltered.

In accordance with the basic procedure of feature selection, generalized intersection of structure descriptions must be performed on the graphs (matrices). The generalized intersection of two graphs is defined as the maximum common subgraph shared by them, i.e., a subgraph wherein the addition of a single node or edge prevents it from being a subgraph of either or both intersecting graphs. The procedure of generalized intersection of graphs or matrices requires the use of special algorithms (*45*, *46*).

The intersection procedure is illustrated by the example in which the compounds under comparison are compound I, mentioned in Section V,A and compound II belonging to the phenothiazine series.



II

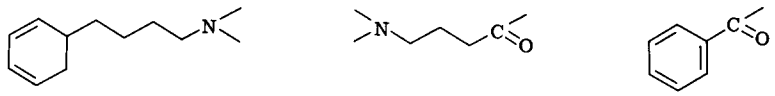
This compound can be described by the descriptor center matrix

$$\begin{vmatrix} 22 & 1 & 3 & 1 & 6 & 6 & 7 & 8 & 8 \\ & 33 & 1 & 2 & 2 & 2 & 5 & 6 & 6 \\ & & 03 & 1 & 5 & 5 & 4 & 5 & 5 \\ & & & 33 & 6 & 6 & 5 & 6 & 6 \\ & & & & 13 & 2 & 9 & 10 & 10 \\ & & & & & 41 & 9 & 10 & 10 \\ & & & & & & 03 & 1 & 1 \\ & & & & & & & 41 & 2 \\ & & & & & & & & 41 \end{vmatrix}$$

A special procedure used to compare the two matrices results in the following common submatrices found for the compounds in question:

$$\begin{vmatrix} 33 & 5 \\ & 03 \end{vmatrix} \quad \begin{vmatrix} 13 & 5 \\ & 03 \end{vmatrix} \quad \begin{vmatrix} 33 & 2 \\ & 13 \end{vmatrix}$$

The above matrices correspond to the following structural fragments:



The described example of feature selection on DCM has been used exclusively for the purpose of illustration. However, it is interesting to note that the first fragment appears to be a known feature responsible for neuroleptic activity (*47, 48*).

On the other hand, this fragment is not revealed by the SSFN. Using SSFN the two compounds can be written:

I. 6,06. 6,00N1; 0303110. 0304130. 0304330. 1102330. 1301331. 3100331. 3200331

II. 6,06. 66B6, 16N6S13; 0300331. 0301410. 0303030. 0362221. 0363440. 1301331. 1302410. 2200331. 2264441. 3302411

Comparison of the two notations leads to the identification of common descriptors. Such a common feature shared by the compounds is conjunction of the descriptors 6.06 and 1301331, corresponding to the last fragment.

The procedure of generalized matrix intersection can be applied to find topographical features using CCM. This can be achieved by the introduction of additional threshold values to determine "coincidence" of distances and charges.

The procedures used for the selection of topological and topographical features permit automated searching for pharmacophoric patterns (*49, 50*) that had usually been selected by mere "visual" comparison of structures. The present section outlines only the basic principles of automated pharmacophoric pattern identification. A more detailed treatment of different methods of structure representation, feature selection algorithms, and results of practical applications of these procedures will be given in forthcoming papers.

VI. Conclusion

The rational strategy of computer-assisted drug design does not imply complete replacement of the knowledge, intuition, and logic of the investigator by the computer intelligence, but it does involve creating conditions favorable to better utilization of both human and computer capacities. These have been the principal guidelines pursued in the elaboration of the logico-structural approach. The logic of the approach simulates the logic of the drug designer. Application of computers provides a more

thorough and unbiased analysis of a much larger number of compounds. At the same time, the investigator working in the interactive mode has at his disposal a means for influencing the course of SAR analysis and is able to feed additional information into the computer as well as to evaluate the results obtained.

The LSA applied to the analysis of SAR has resulted in the selection of biological activity features, i.e., fragments of structure description, responsible for any particular type of biological activity. These features are applicable to activity prediction in unstudied compounds, to purposeful synthesis of compounds with desirable biological action and, finally, to postulate hypotheses on the mechanism of action of chemical compounds.

The LSA has particularly good prospects, as evidenced by experiments, for activity feature selection using data bases which provide successful lead generation.

Further developments of the LSA are directed toward the use of topological and topographical structure representations in combination with quantum chemical and conformational calculations to bridge the gap between empirical and theoretical approaches to drug design. In the future, theoretical methods will, beyond a doubt, gain increasing recognition. However, the methods of logical and statistical analysis will always retain their importance, since they reflect a certain stage in the development of scientific knowledge from data accumulation through empirical laws toward theoretical generalizations.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support and collaboration of their colleagues who contributed to the implementation of the described approach. Our thanks go to L. S. Gitlina and V. V. Drboglav for carrying out a considerable amount of work on the software development for the LSA and to Prof. E. J. Lukevics, Dr. R. S. Sturkovich, Prof. M. J. Lidak, Drs. N. M. Sukhova, S. K. Germane and E. S. Lawrinowich who were the first to apply our methods in practice. Many years of fruitful cooperation with Drs. V. V. Avidon and V. S. Arolovich have resulted in the development of a data base on the structure and activity of chemical compounds and the topological notation. We are grateful to Prof. E. I. Ariëns who drew our attention to the work of Austel and Kutter (*26*) and who pointed out the similarity between the principles of drug design employed in that paper and those of the LSA.

The authors are indebted to Dr. S. A. Avertsev for the translation of the manuscript into English and to Ms. Solomatova for the preparation of the manuscript.

References

1. E. J. Ariens and A.-M. Simmonis, *In* "Topics in Current Chemistry," Vol. 52, p. 1. Springer-Verlag, Berlin, 1974.

2. G. Redl, B. D. Cramer III, and C. E. Berkoff, *Chem. Soc. Rev.* **3,** 273 (1974).
3. P. J. Lewi, *In* "Drug Design" (E. J. Ariens, ed), Vol. 7, p. 209. Academic Press, New York, 1976.
4. V. E. Golender and A. B. Rosenblit, "Computer Assisted Methods of Drug Design." Zinatne, Riga, 1978 (in Russian).
5. C. Hansch, *Il Farmaco* **23,** 293 (1968).
6. S. M. Free and J. W. Wilson, *J. Med. Chem.* **7,** 395 (1964).
7. P. J. Harrison, *Appl. Stat.* **17,** 226 (1968).
8. C. Hansch, S. Unger, and A. Forsythe, *J. Med. Chem.* **16,** 1217 (1973).
9. K. C. Chu, *Anal. Chem.* **46,** 1181 (1974).
10. S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, *Comput. Biomed. Res.* **6,** 411 (1973).
11. K. L. Ting, R. C. Lee, G. W. A. Milne, M. Shapiro, and A. M. Guarino, *Science,* **180,** 417 (1973).
12. R. D. Cramer III, G. Redl, and C. E. Berkoff, *J. Med. Chem.* **17,** 533 (1974).
13. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **96,** 916 (1974).
14. K. C. Chu, R. J. Feldman, N. B. Shapiro, G. F. Harard and R. I. Geran, *J. Med. Chem.* **18,** 539 (1975).
15. A. J. Stuper and P. C. Jurs, *J. Pharm. Sci.* **67,** 745 (1978).
16. S. A. Hiller, A. B. Glaz, V. E. Golender, L. A. Rastrigin, and A. B. Rosenblit, *Khim.-Pharm. Z.* **6,** No. 12, 18, (1972).
17. P. J. Brown, *The Statistician* **20,** 3 (1971).
18. A. Cammarata and G. K. Menon, *J. Med. Chem.* **19,** 739 (1976).
19. W. J. Dunn III, S. Wold, Y. C. Martin, *J. Med. Chem.* **21,** 922 (1978).
20. M. Weiner and P. H. Weiner, *J. Med. Chem.* **16,** 655 (1973).
21. J. T. Clerc, P. Naedeli, and J. Seibl, *Chimia* **27,** 639 (1973).
22. C. Perrin, *Science* **183,** 551 (1974).
23. R. J. Mathews, *J. Am. Chem. Soc.* **97,** 935 (1975).
24. S. H. Unger, *Cancer Chemother. Repts.,* Part 2 **4,** 45 (1974).
25. V. E. Golender and A. B. Rozenblit, *Avtomat. Telemekh.* **11,** 99 (1974).
26. V. Austel and E. Kutter, *In* "Arzneimittellentvicklung, GrundlagenStrategien-Perspektiven" (E. Kutter, ed.), p. 113. G. Thieme Verlag, Stuttgart, 1970.
27. J. S. Mill, "A System of Logic, Rationative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation" 8th ed. Harper, New York, 1900.
28. E. B. Hunt, J. Marin, and P. J. Stone, "Experiments in Induction." Academic Press, New York, 1966.
29. R. D. Dilard, G. Poore, D. R. Cassady, and N. R. Easton, *J. Med. Chem.* **10,** 199 (1967).
30. W. P. Purcell and J. M. Clayton, *J. Med. Chem.* **11,** 40 (1968).
31. E. Pescatori, STRAC Package. System Manual. Soc. Montedison, Inst. Ricerche "G. Donegani". Novara, 1976.
32. E. Ya. Lukevic and G. M. Voronkov, *Dokl. Akad. Nauk SSSR,* **216,** 103 (1974).
33. R. S. Sturkovich. Dissertation Abstract, Riga, 1976 (in Russian).
34. S. A. Hiller, V. E. Golender, A. B. Rosenblit, R. S. Sturkovich, and E. Ya. Lukevic, *Khim.-Farm. Zh.* **10,** No. 3, 29 (1976).
35. S. K. Germane, *In* "Epilepsy" Vol. 2, p. 410. Medicina, Moscow, 1964.
36. R. J. Marcus, E. E. Gloye, and E. T. Florance, *Comput. and Chem.* **1,** 235 (1977).
37. L. A. Piruzjan, V. V. Avidon, A. B. Rosenblit, V. S. Arolovich, V. E. Golender, S. P. Kozlova, E. M. Michailovsky, and E. G. Gavrischuk, *Khim.-Farm. Zh.* **11,** No. 5, 35 (1977).

38. M. Negver, "Organisch-chemische Arzneimittel und ihre Synonyma." Academie-Verlag, Berlin, 1971.
39. V. V. Avidon, *Khim.-Farm. Zh.* **8,** No. 8, 22 (1974).
40. V. V. Avidon and V. S. Arolovich, *Khim.-Farm. Zh.* **11,** No. 6 (1977).
41. K. Miura, M. Ikeda, T. Oohashi, Y. Igarashi, I. Okada, T. Shikimi, and S. Ischwi, *Yakugaku Zasshi,* **85,** 289 (1965). Chem. Abstr. **67,** 31148 (1967).
42. Switzerland Patent No. 396907, 1961.
43. F. Katsuoro, S. Masanao, T. Yoshiyuki, N. Isao, and N. Shinichi, *Nippon Kagaku Ryohogakukai Zasshi,* **15,** 535 (1967). Chem. Abstr. **68,** 47352 (1968).
44. K. Miura, M. Ikeda, T. Oohashi, I. Okada, and Y. Igarashi, *Yakugaku Zasshi,* **84,** 341 (1964). Chem. Abstr. **61,** 8772 (1964).
45. L. S. Gitlina, V. E. Golender, and A. B. Rosenblit, IV Int. Conf. on Comp. in Chem. Res. and Educ. Abstracts, Novosibirsk, 2–17 (1978).
46. M. M. Cone, R. Venkataraghavan, and F. W. McLaferty, *J. Am. Chem. Soc.* **98,** 7668 (1977).
47. P. A. Jansen, *In* "International Encyclopedia of Pharmacology and Therapeutics" (C. J. Cavalito, ed), Vol. 1, Sect. 5, p. 37. Pergamon Press, New York, 1973.
48. J. J. Kauffman, W. S. Koski, *In* "Drug Design" (E. J. Ariëns, ed.) Vol. 5, p. 251. Academic Press, New York, 1975.
49. L. B. Kier, *In* "Fundamental Concepts in Drug-Receptor Interactions" p. 15. Academic Press, New York, 1970.
50. A. Korolkovas, "Essentials of Molecular Pharmacology." Wiley, New York, 1973.

Index

D

E

M

N

O

T