Chem-Bioinformatics: Comparative QSAR at the Interface between Chemistry and Biology

Corwin Hansch,*,† David Hoekman,‡ A. Leo,† David Weininger,§ and Cynthia D. Selassie†

Department of Chemistry, Pomona College, Claremont, California 91711, David Hoekman Consulting Incorporated, 107 NW 82nd Street, Seattle, Washington 98117, and Daylight Chemical Information Systems Incorporated, 441 Greg Avenue, Santa Fe, New Mexico 87501

Received July 16, 2001

Contents

I. Introduction

This is a review of an approach to organizing data on chemical−chemical and chemical−biological reactions in numerical mechanistic terms such that numerous comparisons can easily be made and delineated. Ideas on how to mine these databases for very specific information are illustrated. In the development of our computerized system, a major point of interest has been to be able to make comparisons of quantitative structure−activity relationships (QSAR) between simple chemical reactions and reactions drawn from biological systems. Many instances have been noted where such comparisons are of definite value in understanding the more complex and sophisticated biological processes.

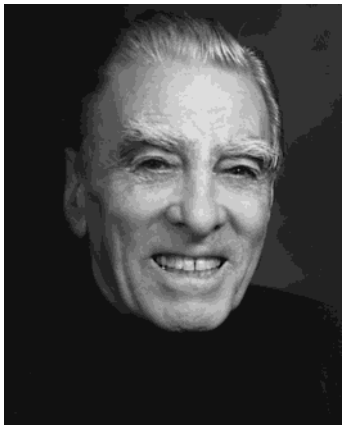
The glut in scientific information, which is growing at an exponential rate in conventional publications and on the world wide web, seriously taxes our ability to organize it *or* make proper use of it. In chemistry alone, *Chemical Abstracts* publishes almost 2000 abstracts/day (1949). A 3 month vacation would set you behind 175 410 abstracts! Thus, it is not surprising that researchers tend to work in narrowly defined compartments. Reviews tend to cover various focused interests, but what is lacking is more integration and cohesion. This problem is exacerbated at the interface between chemistry and biology. The advent of high-speed computing and enormous storage capacity allows us to organize what has been done in addition to generating new data. We have been trying to make a very small dent in the problem via the quantitative structure−activity relationships (QSAR) paradigm since its advent in 1962.[1]

In addition to the innumerable publications on the subject, there are now 12 500 web sites on QSAR. It is impossible to peruse 12 500 pages and collect what might be useful. The ability to keep track of what is happening in the field of QSAR is a daunting task. There are now numerous other approaches to QSAR. Many software companies market programs for SAR and QSAR. It is no surprise that most universities have started departments of information science and are struggling with their development. The flood of information in science has occurred with relatively little input from the continents of South America, Africa, and much of Asia. What will happen when these areas begin to produce like the United States and Europe? Newspaper reports indicate that there are about 1000 biotech companies in Europe and a comparable number in the United States. The needs of these companies as well as those of the large pharmaceutical enterprises, plus the constantly increasing interest of the major countries in environmental toxicology, greatly stimulates computerized attempts to understand the interactions between organic chemicals and every conceivable aspect of life from genes, enzymes, cells, membranes, plants, insects, animals to humans.

It has been a struggle to understand how to commence the development of a science of chemical−biological interactions. By science is meant mathematical descriptors using a relatively small number of well-tested parameters[2−9] and molecular graphics[10−12] to make the connections. A start on this problem has been made by creating a database of over 17 000 QSAR of which 8500 pertain to biological systems and 8600 are from mechanistic organic chemistry. This has not been an easy task, even for the development of simple QSAR from mechanistic organic chemistry, since there is no simplified method to collect such data! This illustrates the crux of the problem facing information science. *Chemical Abstracts* lists such equations under the heading of LFER (linear free energy relationships), Hammett,

* To whom correspondence should be addressed.
† Pomona College.
‡ Hoekman Consulting Incorporated.
§ Daylight Chemical Information Systems Incorporated.

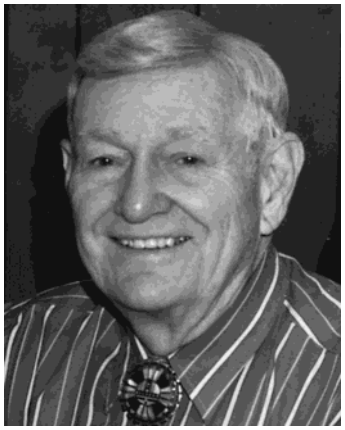
10.1021/cr0102009 CCC: $39.75
Published on Web 02/07/2002



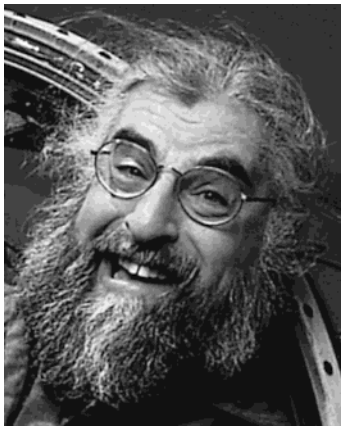
Corwin Hansch received his undergraduate education at the University of Illinois and his Ph.D. degree in Organic Chemistry from New York University in 1944. After working with the DuPont Company, first on the Manhattan Project and then in Wilmington, DE, he joined the Pomona College faculty in 1946. He has remained at Pomona except for two sabbaticals: one at the Federal Institute of Technology in Zurich with Professor Prelog and the other at the University of Munich with Professor Huisgen. The Pomona group published the first paper on the QSAR approach relating chemical structure with biological activity in 1962. Since then, QSAR has received widespread attention. Dr. Hansch is an honorary fellow of the Royal Society of Chemistry and recently received the ACS Award for Computers in Chemical and Pharmaceutical Research for 1999.

David Hoekman studied physics and biology at Pomona College, graduating in 1985 with his B.S. degree in Biology. He spent a year working on ecological wood anatomy at Rancho Santa Ana Botanic Garden and then did a further year of study in the Botany Department at University of California, Berkeley. In 1987 he joined Corwin Hansch's group as a scientific programmer, responsible for the design and implementation of a QSAR database and analysis package, and eventually served as Head of Computer Operations. Since 1996 he has worked as an independent consultant on a variety of database applications.

and sometimes correlation analysis. However, in many instances, the authors do not use these terms and no direct reference is possible. The only way to make progress is to check the references in each paper that is found and check the references in those papers and so on. The chemistry articles are easily entered into the system since, in most cases, the authors have formulated an appropriate equation. However, in the early work (1935−1965), before the advent of easy to use computers (the IBM 360 appeared in 1965), researchers made few attempts to explore more than one-variable equations. Regression analysis was unknown to chemists. Much of this work has been recast using steric and electronic parameters in a dual-parameter approach.



Albert Leo was born in 1925 in Winfield, IL, and educated in Southern California. He received his B.S. degree in Chemistry from Pomona College and his M.S. and Ph.D. degrees in Physical Organic Chemistry from the University of Chicago. His doctoral thesis, under Professor Frank Westheimer, was on reaction mechanisms based on rates of breaking carbon–deuterium bonds. After a number of years in industrial research and development, he returned to Pomona College to initiate and direct the Medicinal Chemistry Project under Professor Corwin Hansch. At present he is President and Research Director of the Biobyte Corporation, a vendor of computer software and databases for drug and pesticide design.



Dave Weininger is a self-actualized person who has spent most of his 50 years pursuing an obsession with chemical information and closely related subjects such as music, flying, and astronomy. He is currently President of Daylight Chemical Information Systems, Incorporated, which produces tools used for doing chemistry as an information science including chemical databases, high-performance search engines, chemical languages, and an object-oriented chemistry toolkit. Dr. Weininger was trained at the University of Rochester in Fine Arts, the University of Bristol in Chemistry, and the University of Wisconsin in Water Chemistry. His research experience includes four years at the USEPA's National Water Quality laboratory in Duluth, MN, and five years at Pomona College in Claremont, CA. He plays a small banjo, flies medium-sized aircraft, operates an astronomical observatory, and heads Daylight's research office in Santa Fe, NM.

Dealing with the biological QSAR was, and still is, a complex and difficult problem. Even today only a very small percent of researchers attempt any kind of a QSAR. In the last 20 years, SAR workers are slowly beginning to use a wide variety of approaches[13−17] to formulate equations or 3-D models to understand these interactions. Many of these approaches (as well as 2-D QSAR) have given the impression that various chemicals can be sequestered together to yield a QSAR with a good r^2. This means that at times the independent variable may not characterize a uniform mechanism of action/reaction. Such an approach can be grossly misleading. As yet,

Cynthia Selassie is a Professor of Chemistry at Pomona College, Claremont. She obtained her M.A. degree in Chemistry from Duke University and her Ph.D. degree in Pharmaceutical Chemistry from the University of Southern California, under the aegis of Professor Eric Lien. In 1980, she joined Professor Corwin Hansch as a postdoctoral Reserach Associate. In 1990, she joined the faculty at Pomona College as an Associate Professor of Chemistry. Her research interests include development of the QSAR paradigm, its coherence with molecular modeling, as well as its applications to drug design, multidrug resistance, and toxicity of phenols.

none of these new approaches have been shown to be capable of doing comparative QSAR. Until one can make such comparisons, one does not have the beginnings of a foundation for developing a science of chemical−biological interactions.

As with QSAR for mechanistic chemistry, locating satisfactory data for developing biological QSAR is a tenuous process. Each new QSAR generally has to be formulated from scratch. This process entails the rigorous perusal of certain sections of chemical abstracts and a few journals, followed by an investigation of interesting references. In some instances, emphasis has been placed on certain topics such as radical reactions,[6] potential HIV drugs,[7] compounds binding to the estrogen receptor,[8] QSAR lacking hydrophobic terms,[9a] and allosteric interactions.[154,159] Success stories using QSAR have been reported.[9b]

The design of 'search engines' is influenced greatly by how the data is entered and where ones interests lie. Our current system was started almost 30 years ago[18] when bioinformatics was not in vogue. Computers were in their infancy, and this too influenced design. The main problem with search engine design is careful organization so that a focused search does not warrant visual inspection to obtain relevant information. We admit that our present system needs improvement in this regard. Nevertheless, we believe that our experience will be of considerable help to others in developing more sophisticated approaches to the study of the chemistry of living systems and their components. Our data will be of help in the evolution of QSAR informatics systems.

II. Structure of the Database

An overview of our system is outlined in Tables 1−4. From the beginning, a major concern has been the arrangement of the structure so that one could sequester all the information related to a particular problem, leaving out extraneous material. Hence, since one is most often working on either the biological or physical data, our databank is divided into two sections. The two areas have been subdivided as shown in Tables 2 and 3 and Scheme 1. However, these subsections can be searched separately or in combination. There is one important difference in the two sections under the field 'SYSTEM'. In Table 1, the appropriate solvent has been entered as System for the organic reactions. Sequestering our system into a variety of classes means that all QSAR on one or more subjects can be analyzed singly or together. For instance, one could select B2A and B6B and garner equations for enzymes and insects for comparison. This might seem strange, but one can go further and next select out of this mixture of sets

Table 1. Organization of Sets

field	title	description
	input data	
1	SYSTEM	biological or physical system
2	CLASS	Pomona classification of system (Tables 2 and 3)
3	COMPOUND	parent compound (if any)
4	ACTION	measured action or activity
5	REFERENCE	journal reference or other source of data set
6	SOURCE	person who entered data set
7	CHECK	person who checked data set
8	NOTE	additional information about data set
9	DATE	date on which set was saved into database
10	PARAMETERS	list of parameters[a]
11	SUBSTITUENTS	labels of substituents
12	SMILES	topological description of compounds
13	DATA**	table of parameter values[b]
14	PRM MAX/MIN	maximum and minimum of each parameter
	output data (equation)	
15	TERMS IN EQN	parameters in regression equation
16	EQUATION	regression coefficients for each parameter
17	IDEAL	ideal (or optimal) log *P*, and confidence limits
18	STATISTICS	n, df, r, s, etc.
19	RESIDUALS	deviations between y-predicted and observed
20	PREDICTED	predicted values of dependent parameter

[a] Examined, even if not used in final equation. [b] Note: in SEARCH MENU (mode), this field is for MERLIN substructure searching.

Table 2. Class Codes−Biological Database (Number of Sets in Parentheses)[a]

B0	**unknown**
B1	nonenzymatic **Macromolecules** (DNA, fibrin, hemoglobin, soil, albumin, etc.) (237)
B2	**Enzymes**
B2A	oxidoreductases (676)
B2B	transferases (160)
B2C	hydrolases (668)
B2D	lyases (37)
B2E	isomerases (12)
B2F	ligases (3)
B2G	receptors (1065)
B3	**Organelles**
B3A	mitochondria (88)
B3B	microsomes (97)
B3C	chloroplasts (83)
B3M	membranes (98)
B3R	ribosomes (0)
B3S	synaptosomes (22)
B4	**Single-Celled Organisms**
B4A	algae (37)
B4B	bacteria (691)
B4C	cells in culture (702)
B4E	erythrocytes (82)
B4F	fungi, molds (251)
B4P	protozoa (104)
B4V	viruses (165)
B4Y	yeasts (47)
B5	**Organs/Tissues**
B5C	cancer (110)
B5G	gastrointestinal tract (77)
B5H	heart (86)
B5I	internal/soft organs (66)
B5N	nerves, brain, muscles (337)
B5S	skin (53)
B5L	liver (20)
B6	**Multicellular Organisms**
B6A	animal (vertebrates) (675)
B6B	insects (197)
B6F	fish (187)
B6H	human (42)
B6I	invertebrates (noninsect) (101)
B6P	plants (126)

[a] In some biological examples the numbers in parentheses may be smaller than indicated. This results from assigning more than one reference number to a particular study, e.g., for a study of compounds curing mice of a bacterial infection under class we might enter B4B and B6A.

Table 3. Class Codes−Physical Database (Number of Sets in Parentheses)

PT	**Theoretical** (30)
PO	**Unknown**
P1	**Ionization** (1618)
P1P	ionization potential (33)
P1X	proton exchange (72)
P2	**Hydrolysis** (791)
P3	**Solvolysis** (624)
P4	**Spectra**
P4I	ionization spectra (61)
P4E	ESR spectra (2)
P4M	Mass spectra (12)
P4N	NMR spectra (176)
P4R	IR spectra (9)
P4U	UV spectra (23)
P5	**Miscellaneous Reactions** (446)
P6	**Substitution**
P6E	electrophilic substitution (247)
P6N	nucleophilic substitution (1137)
P7	**Addition**
P7D	dimerization (10)
P7E	electrophilic addition (150)
P7N	nucleophilic addition (218)
P7P	polymerization (12)
P8	**Elimination** (153)
P9	**Rearrangement** (193)
P10	**Oxidation** (513)
P12	**Radical Reactions** (571)
P13	**Complex Formation** (104)
P14	**Partitioning** (132)
P14C	chromatography (22)
P15	**Pyrolysis** (90)
P16	**H-Bonding** (28)
P17	**Electrochemical** (242)
P18	**Brønsted** (121)
P19	**Esterification** (238)
P20	**Photochemical** (39)
P21	**Hydrogenation** (16)
P22	**Isokinetic** (3)
P23	**Reduction** (82)

Scheme 1

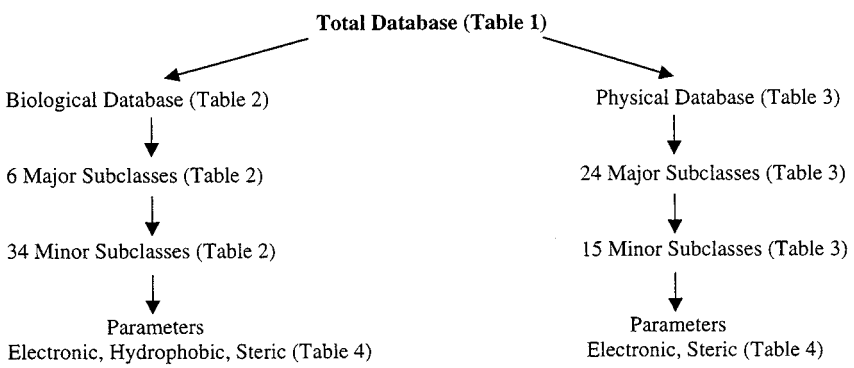


Table 4. [a]

1	PI	pi	ref
2	MR−SUB	substituent refractivity	76, 77
3	F	field effect (from S−L)	22
4	R	resonance effect (from S−L)	3
5	R+	resonance plus	3
6	R−	resonance minus	3
7	ES	E(s) from Taft	74
8	L-STM	length sterimol	75
9	B1−STM	width sterimol	75
10	B5−STM	width sterimol	75
11	S−P	sigma para	3
12	S−P+	sigma para plus	3
13	S−P−	sigma para minus	3
14	S−M	sigma meta	3
15	S−M+	sigma meta plus	3
16	S−M−	sigma meta minus	3
17	S−INDUC	sigma inductive	3
18	S−STAR	sigma star from Taft	3
19	ER−P	electronic radical, para	6
20	ER-M	electronic radical, meta	6
21	S.DOT-P	sigma dot, para	6
22	S.DOT-M	sigma dot, meta	6
23	S.-DOT-P	sigma dot, para (JJ)	6
24	S.-DOT-M	sigma dot, meta (JJ)	6
25	S.P-C	sigma para (C)	6
26	S.M=C	sigma meta (C)	6

[a] To those not familiar with terms from physical organic chemistry a glossary has been compiled. Muller, P. *Pure Appl. Chem*. **1994**, *66*, 1077.

those that contain certain features such as a term in σ^- or that lack hydrophobic terms. Or one might want to consider QSAR based on 20 or more data points with $r^2 > 0.90$, etc.

In compiling the physical database from mechanistic physical organic chemistry studies, we have concentrated on chemical reactions in solution. Although there are some examples (295) based on spectra and gas-phase reactions, no attempt was made to be complete in these areas. The same applies to the Brønsted reaction (121 examples). Reactions that constitute a Brønsted type are now entered without comment.

Many papers report results from kinetic runs at a variety of temperatures. Generally we have reported only one example at the temperature nearest to 25 °C. In cases where a reaction has been run in various mixtures of solvents (e.g., ethanol and water), we have reported representative examples. For lack of time, we have *not* attempted to standardize the dependent variables as we have in biological reactions. We have simply used the log of rate or equilibrium constants. For this reason, intercepts in the physical equations cannot be compared. Publication of Hammett-type equations has occurred at such a rapid rate and in such diverse areas that it was impossible to organize the results before modern interactive computing. Finally, after considerable effort, we acquired a large percentage of the data and devised the means to view it from many perspectives.

Biological QSAR has been in an even more confused state. The major areas−biochemistry, medicinal, and pesticide chemistry and the various toxicologies−all have a large number of subspecialties e.g., enzymology, anesthesiology, cancer, mutagenesis, metabolism, cardiology, pyschobiology, bacteriology, plant physiology, urology, etc. It is apparent from Table 2 that, beyond the few key words listed, we have *not* as yet attempted to include them in a systematic way. Yet they can provide significant help to the researcher. A further complicating factor is that reports on these studies, which are now appearing at an ever increasing rate, are published in hundreds of extremely diverse and sometimes obscure journals and hence are difficult to find. Our database shows that partition coefficients (at the moment we have almost 30 000 experimentally measured octanol/water log *P* and log *D* values of which over 12 000 are unique for the neutral species and considered to be reliable), from which hydrophobic parameters are derived, have appeared in over 600 different journals. Sources of biological data are even more diverse. We believe the time has come to integrate these results into a useful format. Since a *variety* of approaches are currently being studied for the formulation of QSAR, one might question whether this is the time to pursue such an approach. However, the experimental data reported and organized will be of value for decades to come regardless of how the methodologies evolve. In fact, our system will provide the testing ground for the various new approaches stemming from quantum chemistry, molecular dynamics, and modeling.

Many data sets have been poorly designed or suffer from a total lack of design. The QSAR for these sets have low r^2 values, too many outliers, and sometimes too few datapoints per variable. Nevertheless, we have found such preliminary attempts to be helpful in supporting other work and suggesting new options. Hence, we retained some QSAR that are rather weak. When one attempts to rationalize in numerical terms the results from treating even something as simple as a cell culture (let alone mice) with say 30 or 40 'congeners', the problems are mindboggling. Nevertheless, the pharmaceutical industry constantly faces these challenges. Human DNA codes for 50−100 thousand proteins that account for the many enzymes and components of various cellular membranes and organelles. Most biochemical processes are subject to perturbation. Hence, it is not yet clear what quality (in terms of r^2) one ought to expect with complex biosystems. However, a rational and statistically based analysis is vastly better than mere intuition.

Our main premise is that the major interaction forces to consider in a set of congeners acting on a biological system are electronic, steric, and hydrophobic in nature. Other important factors include hydrogen bonding, polarizability, and dipole moments. Hydrogen bonding can be important, but as yet there is no general way to deal with it in the way that one can use Pi (π), for example, to account for the hydrophobicity of a substituent. The orientation and distance between an OH on the substrate or inhibitor and the bonding site on the receptor is so critical that a general method for parametrization appears impossible. In this case, indicator variables can be helpful.

Graphically, our system can be viewed as in Scheme 1. Scheme 1 outlines a biodynamic system that is like an electronic set of two books. One can

read one book or the other or peruse the chapters as outlined in Tables 2 and 3 where the headings are listed (e.g., enzymes) or one can look at the paragraphs such as oxidoreductases. The difference is that since paper is not involved, the books can undergo continuous edition. New additions to the database occur at a rate of about 80 new QSAR/month, and yet this is not enough to keep abreast of the voluminous literature. Our singular approach is to bring understanding to chemical−biological dynamics via a mechanism-based analysis.

The combined database can be searched, or more commonly, the biological or physical bases can be searched independently. Then any of the major or minor subclasses can be sequestered for study. By means of item 15 in Table 1, QSAR can be isolated according to the parameters which form their basis.

Two general types of searching are string searching and searching via 2-D molecular structure. The objective of this scheme is to focus the output as narrowly as possible to limit the amount of data that must be examined. The complexity of the search engine is the result of the enormous variety of chemical−chemical and chemical−biological reactions.

III. Searching the Database

Our search engine operates in three broadly different ways. The first, string searching, is based on words. The second searches on 2-D molecular formulas using the SMILES notation. However, the SMILES search can be approached in two ways. One can identify every QSAR that contains a *specific* molecule, or else one can use a MERLIN search that finds all *derivatives* of a given structure. A third method searches on parameters, one or more at a time.

String searching can be utilized in several contexts, as illustrated with the simple string **in** (from this point on direct commands will be entered in bold letters and underlined) that can be involved in the following ways.

1	*E. coli* **in** mouse	as a stand alone word	(both leading and trailing blanks) **" in "**
2	**in**fluenza	as a start of a word	(leading blank, but no trailing blank) **" in**
3	bra**in**	as an end of word	(trailing blank, but no leading blank) **in "**
4	pyrid**in**e, gu**in**ea	inside a word	(neither leading nor trailing blanks) **in**

Searching on **in** with quotes separated by blanks would find every instance in the database where it is a stand-alone word. In the second example with a leading quote−blank every word in the system starting with **in** is found. In 3, searching with a trailing blank−quote locates all words ending with **in**. In example 4 with **in** alone, every possible form of **in** is located (2700 hits in the physical bank). String searching can be helpful when one is not sure how to spell a name or exactly how the subject of interest is classified.

A few other examples may be helpful. **"HEM"** or **HEM** matches **HEM**OGLOBIN but not C**HEM**OTHERAPY. **"ASE "** matches LY**ASE** but not L.C**ASE**I.

If you 'quote' a string but do not include either a leading or trailing blank, the query is no different than if you had not included the quotes at all. It is not required that quotes be matched up before and after a word. The two examples above could be stated. **" HEM** matches **HEM**OGLOBIN but not C**HEM**OTHERAPY. **ASE "** matches LY**ASE** but not L.C**ASE**I.

Any character search can be negated by prefacing it with **NOT**. This causes the result to be the reverse (logical complement) of what it would otherwise be. **NOT CAT** *does not* match **CAT**, **CATCH**, **CATTAIL**. **NOT ASE "** *does not* match **LYASE**, but does match **L.CASEI**. Note that we have underlined the commands to clarify each entry.

Another feature in our search system is illustrated by the use of the comma to signify 'and'. Entering **mouse** (space) **E. coli** would pull together all datasets where mouse *or E. coli* occurs. This would, in general, be pointless. Entering the two as **mouse**,**E. coli** first finds all sets based on mouse and then separates those that also have *E. coli* (i.e., *E. coli* interacting with mice).

An alternative means for searching is based on the SMILES language invented by David Weininger[19−21] and incorporated into our developing system while he was a member of the Pomona College MedChem Project. SMILES coupled with DEPICT was a truly outstanding advancement, since it constituted an unambiguous language for naming organic chemicals and displaying them in 2-D. SMILES allows one to use a line notation to enter two-dimensional structures into the computer, each in a unique format. We have now entered the SMILES for many compounds with unambiguous names such as benzoic acid or quinine so that input of a name results in the generation of the related SMILES for searches.

Two means are present for doing such searching. For example, one can enter **phenol** and find every data set that contains phenol. In so doing, we find 307 QSAR in the physical database that contain phenol. Many of these are mixtures of phenols and other compounds that researchers have used to formulate a single equation. Using the command **3 not miscellaneous** (see Table 1) reduces the number to 255. Unfortunately not all sets of mixtures were labeled as such, so further refinements are in order.

A searching program, also using the SMILES notation, is called MERLIN and was also invented by D. Weininger. Entering the SMILES for phenol into MERLIN using the command **13** in the search mode finds all derivatives of phenol in which substitution occurs at any or all of its six hydrogen atoms. This will find, for example, anisole and pentachlorophenol, among many other structures. This locates 4355 QSAR. The biological database contains the common names as well as the official names of over 10 000 drugs, currently on the market, discontinued, or interesting but not yet on the market. This means that one can do a MERLIN search on any one of these compounds to uncover QSAR on similar chemicals. The common names of many simple compounds are also stored, and their SMILES can also be generated by entering the name. Using command **13 p-ami-**

Table 5.

		hits			hits
SMILES	mescaline	5	**SMILES**	testosterone	19
MERLIN	mescaline	22	**MERLIN**	testosterone	39
SMILES	epinephrine	12	**SMILES**	phenoxyacetic acid	7
MERLIN	epinephrine	19	**MERLIN**	phenoxyacetic acid	67
SMILES	naproxen	8	**SMILES**	isoniazid	4
MERLIN	naproxen	9	**MERLIN**	isoniazid	10
SMILES	methotrexate	13	**SMILES**	adamantane	0
MERLIN	methotrexate	15	**MERLIN**	adamantane	70
SMILES	hexobarbital	21	**SMILES**	glucose	5
MERLIN	hexobarbital	21	**MERLIN**	glucose	38
MERLIN	[Pt]	7	**SMILES**	cortisone	13
MERLIN	[Se]	19	**MERLIN**	cortisone	13

nobenzoic acid finds 265 QSAR that contain this compound or a derivative of it where any H atom has been replaced by some other element. The biological database can be searched with SMILES using **12** from the search mode name or MERLIN using **13**. Some examples follow.In the examples of Pt and Se, only a MERLIN-type search is possible since no QSAR have been reported for the bare metals. This yields all compounds that contain such an element. It is interesting that adamantane itself has never been tested, but after the discovery of the antiviral activity of aminoadamantane, there was a wild flurry of testing derivatives of adamantane or using it as a substituent. In the case of cortisone, it was surprising to find no 'similar' compounds. The large number of hits with phenoxyacetic acid is due to the great interest in these chemicals as weed killers. In fact, QSAR was developed out of interest in this class of chemicals.[1]

IV. Parameters

The choice of parameters is of the utmost importance in the construction of a bioinformatics system where the ultimate objective is comparative QSAR. Table 4 lists some of the parameters that at present can be automatically loaded for QSAR calculations. S stands for Hammett sigma σ; -P and -M stand for para and meta values, respectively. In the broader sense para values are used for aromatic substituents conjugated with the reaction center and meta values for nonconjugated aromatic systems. The Hammett-type parameters (σ, σ^+, σ^-, σ^* (s-star), and σ_I (s-inductive) have received over half a century of study and testing on simple organic reaction mechanisms. Their use in formulating biological QSAR has been discussed, and a listing of published values has been made.[3] The field/inductive (F) and resonance parameters (R) have also been reviewed.[22] Molecular orbital parameters continue to be explored for use in both biological and physical QSAR since there are many instances where Hammett constants cannot be used.[23−68] Searching the biological database with **15 HOMO LUMO** finds 59 such QSAR. Some representative examples are in refs 24−68. Searching with **10 HOMO LUMO** finds every instance where HOMO or LUMO was tested (i.e., 120). This figure less 59 shows that in 61 of the examples, the molecular orbital parameters were tested but found to be not as sound as Hammett constants. However, this statistic must be considered with caution since not all calculations were made with some of the more rigorous computational programs now available.

Parameters 19−26 in Table 4 are of special interest to us as they have been specifically designed to correlate radical reactions.[6] The study of radical reactions is particularly fascinating. In living systems the effect of free radicals can be either useful or detrimental. That is, they can be carcinogenic, estrogenic, or valuable antioxidants, as in the case of flavonoids.[80] E_R was designed by Yamamoto and Otsu,[81] S. Dot by Dust and Aronald,[82] S.-Dot 22 by Jiang and Ji,[83] and S.C. by Creary et al.[84] There is a good correlation between E_R and Creary's parameter, but we have generally used E_R because of a better selection of substituents. However, one must always check σ^+. In general, we have found σ^+ to be most useful in correlating radical reactions, but there are instances where E_R or the other radical parameters are necessary. As yet it is not clear why there is poor correlation between σ^+ and the specially designed radical parameters, but it seems likely that the nature of the reaction transition states must be the critical factor.

The crucial parameter for the initial success of the biological QSAR paradigm[1] was the numerical accounting for hydrophobic interactions. Despite the great complexity of studies of all types of chemicals reacting with various kinds of biological systems (from DNA to whole animals), the octanol/water partition coefficient used in log terms provides surprising insights. It must be remembered that a compound entering a cell has a very large number of possible hydrophobic interactions besides those with a crucial receptor of interest. Most interesting are examples where no hydrophobic term appears even in whole animal studies.[9] The hydrophobic parameter for substituents (Pi)[2] can be of great assistance in delineating local hydrophobic interactions at the receptor level.[2] However, this parameter can be greatly affected by strong electron-attracting elements in close proximity. We have recently modified our system to calculate Pi values taking into account neighboring electronic effects.

Partition coefficients are rarely measured these days since this is a rather costly and time-intensive process. The use of data from the literature to formulate QSAR means that the compounds are not usually available for the measurement of their partition coefficients. In our set of 8500 QSAR, 4614

contain log *P* terms and 784 have Pi terms; hence, it is very important to have the best possible means for their calculations. There are now a wide variety of methods for the calculation of log *P*.[71] The most extensively supported method is that of Leo.[71,72] The quality of his method is illustrated by eq 1.[73]

$$\log P = 0.96(\pm 0.003)\text{Clog } P + 0.08(\pm 0.008)$$
$$n = 12{,}107,\ r^2 = 0.973,\ s = 0.299 \qquad (1)$$

This expression shows the relationship between 12 107 experimental and calculated (Clog *P*) values. Leo's program using SMILES or names as input calculates values on modern desktop machines at a rate of about 100/s. Our program calculates and automatically loads the parameters log *P* and Pi for regression analysis.

Steric parameters are the third cornerstone for QSAR formulation. The classic Es constant of Taft has been reviewed,[74] its use illustrated[2] and experimental values listed.[3] Es was designed for modeling intramolecular steric effects,[74] but sometimes it is helpful for intermolecular interactions. The calculated sterimol parameters of Verloop and Tipker[75] are generally much more useful and can be easily computed. Values for over a thousand different substituents have been published.[3] Originally five parameters were suggested as descriptors of a substituent, but then it was determined that three were just as effective: B1, B5, and L. B1 is essentially a measure of the size of the first atom in the substituent, and B5 is an attempt to define the effective volume, while L is a measure of the substituent length. Despite the simple nature of these terms, we have found them to be valuable in QSAR formulation. There are 907 examples where B1 has been used, 728 for B5, and 104 for L in the biological database.

Molar refractivity (MR) is a parameter first proposed for biological SAR by Pauling and Pressman[76] and then further developed by Agin et al.[77] It is defined as follows

$$\text{MR} = (n^2 - 1/n^2 + 2)\left(\frac{\text{MW}}{d}\right)$$

In this expression *n* is the refractive index, MW is the molecular weight, and *d* represents density. If refractive index does not vary greatly, MR is heavily dependent on molecular volume. Despite this strong association, it has been found to be superior to calculated molecular volume in QSAR formulations. 2553 QSAR are based on CMR for the whole molecule or MR for substituents, while there are only 422 based on molecular volume. The refractive index does incorporate a term for polarizability, which is directionally dependent on the position of the force causing the electrons to move.[78] Some of the limitations of this parameter have been discussed.[2] Despite these shortcomings, we have found many instances where it gives results superior to molecular volume. A recent most interesting discovery is that it can be used to delineate allosteric effects in enzymes and receptors.[79]

Some useful general searches of the literature can be illustrated by command 5 in Table 1 on references. To get some idea of the source of the original physical data, the following command can be used.

1	**5**	**J.Am.Chem.Soc.**	1750 hits
2	**5**	**J.Chem.Soc.**	1541 hits
3	**5**	**Indian**	339 hits
4	**5**	**Zh.Org.Khim**	363 hits
5	**5**	**Organic Reactivity**	366 hits
6	**5**	**J.Org.Chem.**	1111 hits

To determine the major contributors in the field of mechanistic organic chemistry, the combined databases can be searched in the following manner.

5	**Bowden,K.**	134	QSAR	**5**	**Taft,R.W.**	56	QSAR
5	**Bordwell,F.G.**	128	QSAR	**5**	**Grob,C.A.**	48	QSAR
5	**Lee,I.**	164	QSAR	**5**	**Kabachnik,M.I.**	44	QSAR
5	**Brown,H.C.**	89	QSAR	**5**	**Exner,O.**	51	QSAR
5	**Tsuno,Y.**	160	QSAR	**5**	**Jencks,W.P.**	95	QSAR

V. Mechanistic Organic Chemistry

Work with the Hammett equations and its extensions illustrates what is happening in all areas of science. The first and *last* attempt to list all such equations was made by Jaffe in 1953.[85] This was the most cited paper in *Chemical Reviews* in the period 1945−1995.[86] The second most cited paper in this period was that by Leo et al. on partition coefficients and their uses.[87] These two seminal works cover two of the three cornerstones of QSAR (the third being steric). There are a number of books that have been written on the Hammett equations and their use of which two are most useful.[88,89]

A good place to start with informatics is to use the search mode for Hammett parameters in the study of the ionization of organic compounds. Searching our physical database with **2 " P1 "** (where 2 represents field (Table 1) and P1 the subset in Table 3), we find 1618 QSAR. Note that quotation marks enclose leading and trailing blanks on P1, otherwise we would have found, via string searching, information on P12, P13, etc. Next, moving to the **show** mode, we can review any or all of the information in Table 1. In general, one would not want to page through all of the possibilities, but it could be done in less than an hour. A quicker review would entail a search on **1** and **3** of Table 1 to see the type of compound and solvent covered by each QSAR. The set number is shown so that all of the information in Tables 1 and 3 and the 2-D structures of all compounds can be viewed by entering the set number.

Usually one would want to review QSAR in a single solvent system. Searching with **1 aqueous** finds 1165 sets. This includes many examples where mixed solvents were used. In such examples, a percent is always present, e.g., aqueous 50% ethanol. Hence, entering **not %** reduces the hits to 588 sets based on water alone. Most studies have been published in terms of pK_a or ionization constants used as the dependent variable. The former can be isolated by searching the 588 by the command **15 pK$_a$**, which yields 491 sets. The search for any particular solvent can be illustrated by searching the 1618 with the

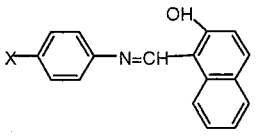
command **1 Ethanol Not %**, which finds only 80 examples in ethanol (sometimes 95% ethanol).

Again, returning to the 1618 sets, one can look for work by a particular author by using the command **5 authors name**. For instance, using **Jencks**, locates 13 studies by the noted biochemist W. P. Jencks. Other aspects of reference can be searched. One might want to look for recent studies on ionization that might cover more complex chemicals. Entering **5** (**1990**) (**1991**) (**1992**) (**1993**) (**1994**) (**1995**) and then searching on **2 " P1 "** uncovers 117 of the 1618 examples. These can then be perused in the **show** mode. Perusing the catch by compound one uses **3** (Table 1) in the **show** mode and finds an unusual study on capsaicin analogues. It must be noted that some examples are present where the same compound is listed in a series of several sets (e.g., phenylformamidines). In such instances it is usually found that the same set of compounds has been studied in several different solvents or solvent mixtures.

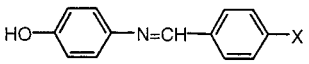
In some cases pK_a has been employed as an independent variable. These can be separated by searching with the entry **15 pK$_a$, logk**. First all sets with pK_a are isolated, and then those containing a log k term are pulled in. This yields 88 examples where the ionization constant log k is the dependent variable (left side of equation) and pK_a is the independent variable. It can be of interest to search for compounds having aqueous pK_a values within a certain range. This can be done using the physical database as follows.

1	**15**	**pK$_a$**	1515
2	**15**	**not logK**	1433
3	**1**	**aqueous**	1057
4	**1**	**not %**	505
5	**14**	**10<pK$_a$<12**	10

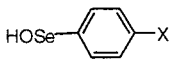
Command 5 isolates any sets having a compound with a pK_a value between 10 and 12. The following examples are illustrative of our catch.



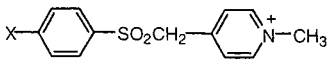
pKa range 10.46 to 11.17



pKa range 10.23 to 10.60

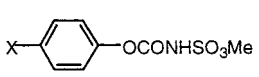


pKa range 10.17 to 10.84

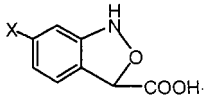


pKa range 10.01 to 11.68

Now in a search for stronger acids we can change step 5 to **14 2<pK$_a$<3**, which gets 15 hits, among which are

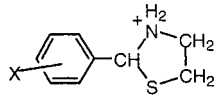


pKa range 1.53 to 2.22

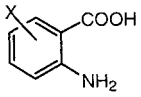


pKa range 1.27 to 2.03

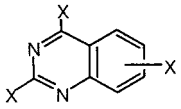
Note that in each example a QSAR is available from which hundreds of other pK_a values can be calculated. Another approach is to search over a wider range and ask for a relatively large group of congeners. By changing step 5 to **14 0<pK$_a$<6** and then **n >10** snags 57 hits on sets having 11 or more data points, and 4 of interest might be



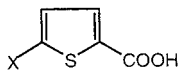
pKa range 4.55 to 5.50



pKa range 0.65 to 2.72

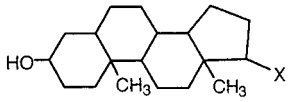


pKa range 2.38 to 4.75

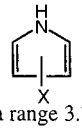


pKa range 2.78 to 3.78

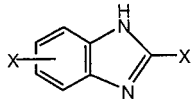
There are 88 QSAR in the biological database where pK_a is the independent variable.



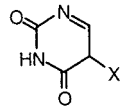
pKa range 0.86 to 10.36



pKa range 3.7 to 9.3



pKa range 0 to 9.95



pKa range 5.3 to 10.0

Data mining, the buzzword these days, is used to search huge sets of chemicals for various types of structures or properties. Our approach can be termed model mining, because behind every hit stands a QSAR that predicts the activity of many untested compounds.

There are two mechanisms for searching using the SMILES descriptor. Using the 1618 sets on ionization and the command **12** asks for the entry of a SMILES. Entering **quinoline** the program supplies the SMILES and searching yields seven sets in which the QSAR is based on quinolines and one set of miscellaneous chemicals that contain quinoline. A general similarity search using MERLIN finds every example in which the quinoline moiety is present or a derivative in which one or more H atom has been substituted. Searching on **13** and **quinoline** finds all such sets (20 examples) such as styrlquinolines, acridines, quinolones, and phenanthrolines. This type of search can yield a huge number of examples. Searching on CH_3CH_2OH uncovers 4398 sets. This number can be reduced by searching as follows.

2	**B4**	1133 hits	cells
15	**S′**	27 hits	QSAR that contain a σ^* terms
15	**Es**	22 hits	QSAR that contain an Es term

The third way of model mining is to search via parameters. Again starting with the 1618 sets and using the command **15 not logK** eliminates QSAR based on ionization constants and isolates 1528 examples where pK_a is the dependent variable. In checking for examples where through resonance is

important, we can use the command **15 S+ S−**, which then isolates 635 QSAR based on σ^+ or σ^-. Or we might be interested in electronic effects in aliphatic systems. Searching with **15 S′ SI** locates 199 possibilities.

Another way to mine the database that can be of interest is to find instances where certain substituents have been studied. Searching with **11 Me CH3 methyl** finds 7788 out of 8400 studies including a methyl group. Using **11 CF3** finds 1036 instances, **11 SO2CF3** uncovers only 34 examples, and using **11 SF5** locates 9 examples. More complex multisubstitution can also be uncovered, e.g., **11 2-NH2 o-NH2** finds 65 examples.

Even in the formulation of the relatively simple QSAR for organic reactions one finds it necessary to omit data points. In our system this is done by marking them with an asterisk (starred points). Such points are held in place and always shown when a listing of results is asked for so that they cannot be forgotten. These can be isolated and evaluated. For example, **2 P12** collects all QSAR on radical reactions (596). **18 omit>0** separates all QSAR with one or more data points starred (240). Moving from search to show and entering **11** lists all substituents for each example to see which ones are poorly fit as well as those that are well fit. The ρ-methoxy and nitro groups are often outliers.

So far we have only considered the subject of ionization that is by far the simplest of the examples in Table 3. The same search strategy can be applied to the other classes. A well-studied subject for physical organic chemists has involved nucleophilic substitution reactions. The search **2 P6N** locates 1146 examples. Remember **2** is from Table 1 and P6N is from Table 3. To check recent activity in this field we can use **5 (1995) (1996) (1997) (1998) (1999)**. This garners 107 hits showing that there is still considerable interest in this area. Similarly searching using **13 pyridine** on the 1146 examples yields 93 hits. This of course finds many examples with pyridine as the nucleophile, but in addition we uncover more complex structures such as quinolines, acridines, and pyridinium ions. One can peruse the 1146 hits with commands **3** and **4** to find interesting examples for comparative studies that can be similarly searched. Using **13 NH2NH2** uncovers 33 examples for a wide variety of derivatives such as X−C_6H_4NNO, X−C_6H_4-CONHNH_2. There is so much variation in the reagents and substrates that one would need to page through the 1146 examples to understand all that has been done. This review of the literature could be accomplished in less than an hour, which is much less time than that devoted to many narrow library searches.

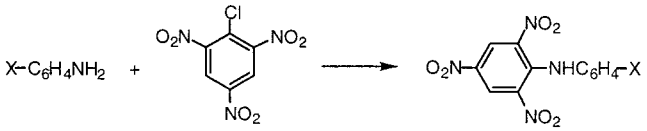
In dealing with over a thousand hits, another level of organization can be attained by organizing the output in terms of the coefficients with any given parameter as follows.

1	**2 P6N**	1,146 hits
2	**15 S−**	221 hits

Moving to the show mode and entering

3	**/sort=16 1 3 4 15 16 18**
4	**sort S−**

Command 3 says sort on slope coefficient (Table 1) and give information covered by some of the items of 1−18 in Table 1. On entering step 3 the program asks for the parameter to be sorted on (enter **S−**). The program then lists QSAR in terms of the coefficients with σ^- going from −6.9 to +8.5. The most negative slope (Hammett's rho value) is for the classical S_NAr reaction.



The most positive slope is associated with

$$X\text{-}C_6H_4Cl + C_6H_5SS^- \rightarrow X\text{-}C_6H_4SSC_6H_5$$

Rho values can also be examined by isolating datasets by using narrower ranges e.g., all negative or all positive coefficients or those coefficients with an intermediate range such as −0.5 to +0.5.

The same approach might be applied to radical reactions. Searching on **2 P12** finds 596 examples. Focusing this set with **15 S+** finds 310 correlated by σ^+, while searching with σ^- yields 63. The quality of 8500 QSAR can be examined in a variety of ways by means of the statistics search 18 (Table 1) as follows.

1	**18**	**2<terms<4**	204 hits−isolates all QSAR having 3 terms
2	**18**	**n>75**	5 hits−isolates QSAR based on more than 75 datapoints
3	**18**	**r>.99**	2 hits−selects QSAR with *r* greater than 0.99

Until rather recently, practitioners of physical organic chemistry rarely used more than two terms to rationalize their results, but faster and more efficient computers have changed the scene. As seen from the above example, the database contains 204 QSAR with three terms. Step 2 shows that some of these are based on large data sets containing a substantial number of data points with high-quality data. The following is an example of the result that we have derived from published data.[90]

$$X\text{−}C_6H_4\text{−}NH_2 + Y\text{−}C_6H_4(CH_2)_2OSO_2C_6H_4\text{−}Z \rightarrow Y\text{−}C_6H_4(CH_2)_2NHC_6H_4\text{−}X + Z\text{−}C_6H_4SO_3^-$$

$$\log k_2 = -1.32(\pm 0.05)\sigma_X - 0.13(\pm 0.02)\sigma^+_Y + 1.08(\pm 0.03)\sigma_Z - 3.93(\pm 0.01)$$

$$n = 80,\ r^2 = 0.992,\ s = 0.042,\ q^2 = 0.991 \quad (2)$$

The subsections of Table 3 are of the type that a physical organic chemist would be comfortable using. Searching by common reaction names can often be very helpful; for example, searching under action **4** isolates the following number of hits.

search command	hits	
DIELS	42	Diels−Alder reactions
Friedel	3	Friedel Craft reactions
Cyclization	32	
Mercuration	11	
Salt	23	salt formation
Alkyl	31	alkylation
Decomp	109	decomposition reactions
Wolf	6	Wolf−Kishner reductions
Dipole	14	dipole moments
Decarboxyl	27	decarboxylation reactions
Racemi	2	racemization reactions
Meerwein	1	Meerwein−Pondorf reduction
Bromi	179	reactions with bromine
Hydration	40	hydration reactions

Many of these QSAR come from P5 Table 3 for miscellaneous reactions.

VI. Chemical–Biological Interactions

Most of the general approaches to model mining that we have considered in mechanistic organic chemistry can be used with chemical−biological interactions. However, organizing biological QSAR is a vastly more difficult problem. The same major preliminary search mechanisms are available (string, SMILES, MERLIN, and parameter). Before or after factoring, as shown in Table 2, can be utilized to further focus the output. The major difficulty is that there is no simple way to categorize the system names or the types of actions. For example **2 B2A** isolates 716 sets and QSAR on oxidoreductases of all types. There is no uniform way to break this into smaller groups. By moving to **show** one can scan the names in less than 10 min and then sequester the ones of interest. The following are a few examples.

system name	number of hits
Cytochrome P450 P-450	63
Dehydrogenase	129
Microsome	19
Hydroxylase	25
Mitochondria	43
Monoamine	57
Dihydrofolate	95
Liver	170
Lipoxygenase	30
Peroxidase	36
Xanthine	18
Cyclooxygenase	51

These 12 examples illustrate some of the possibilities. Searching with cytochrome P450 or P-450 yields 63 examples. Sometimes P450 or P-450 have been used to characterize the system. There are many QSAR on dihydrofolate reductase, an area our laboratory has been working in for many years.

Comparing new QSAR from the biological database we have possibilities available that are not present with the physical database where we have not attempted to standardize the dependent variables. In the biological QSAR log 1/*C* is in molar terms except in a few cases marked by log 1/*C*′. The following approach is illustrative.

1	**15 " log1/C "**	4807 hits
2	**15 not **2 bilin**	3546 hits
3	**15 " logP " " ClogP "**	1738 hits
4	**15 not "S**	1435 hits
5	**15 not ES B1 B5 MR Pi PKA**	1127 hits
6	**16 0.6<logP<1**	481 hits
7	**16 0<const<0.5**	52 hits

The first step ensures that 1/*C* values are standard. The second eliminates all QSAR with nonlinear terms, and the third ensures that we have only octanol/water log *P* values. Searches 4 and 5 eliminate parameters other than log *P*. Step 6 selects only those QSAR where the coefficient with log *P* is between 0.6 and 1.0, and 7 eliminates QSAR whose intercept is outside of 0 and 0.5. The very weak activity (intercept 0−0.5) of the 52 QSAR in terms of slopes of compounds and biological activity is shown in the following examples: I_{50} of synaptosomes, guinea pig cerebral cortex by ROH; I_{50} of chloroplasts by $X-C_6H_4NHCOCH(CH_3)_2$; Inhibition of cholinesterase from electric eel by FCH_2COOR; Inhibition of microorganisms in pharmaceutical cream by $4\text{-}OH-C_6H_4CO_2R$; Hemolysis of red cells from Rabbits by ROH; 75% blockage cockroach nerve action by ROH; Inhibition of valinomycin induced potassium uptake by liver mitochondria by $X-C_6H_4\text{-}CH_2CH_2N(C_2H_5)_2$; I_{50} of Chinese hamster lung fibroblast cells by halobenzenes.

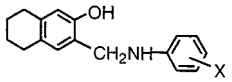
Note that of the above example, a number pertain to simple alcohols. Scores of such studies have been reported, and an extensive review of this work has been published.[91] For the most part, these constitute examples of nonspecific types of toxicity.

Now considering toxicity 100 times greater, we replace command 7 above with **16 2 <const<2.5** and obtain 48 hits for chemicals 100 times as potent. Examples are as follows: I_{50} of Algae by $X-C_6H_5$ and $X-C_6H_4OH$; I_{50} of bluegill fish by chlorophenols; I_{50} of acetylcholinesterase by physostigmine analogues; Uncoupling of phosphorylation in isolated thylakoids by $X-C_6H_4NHCONH_2$.

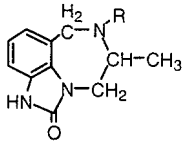
Many of these examples are based on phenols. We see that moving the OH from an alkyl to an aryl carbon increases the potency by 100-fold.

Now increasing 1000-fold over our first search by **16 3.0<const<3.5**, we uncover 16 examples among which are the following.

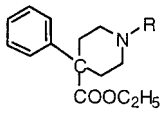
I_{50} of Human Polymorphonuclear Leukocytes by



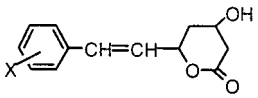
I_{50} to inhibit HIV-1-induced cytopathicity to MT-4 cells by



I_{50} of binding of [H^3] Naloxone rat brain opiate receptors by



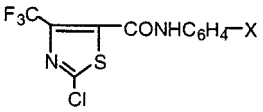
I_{50} of HMG-CoA reductase by



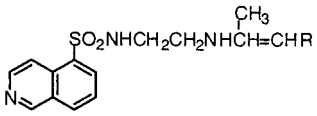
One must bear in mind that the value of the intercept will depend in part on the sensitivity and specificity of the test system and the toxicity of the chemicals.

Moving up another factor of 10 with **16 4.0<const<4.5** isolates 29 examples. The following again illustrates the wide range of chemicals and test systems in the database.

Inhibition of mitochondria succinate dehydrogenase by[92]



Concentration of needed for a 5-fold increase in



vinblastine accumulation in P388 cancer cells.[93]

I_{50} of $X-C_6H_4CONHOH$ to 5-lipoxygenase of red cells.[94a]

One can search for more complex QSAR as follows.

15	**logP**	4414 hits
15	**not **2 bilin PI**	3089 hits
15	**S+**	90 hits
16	**.6<logP<1**	17 hits
16	**−2<S+<O**	9 hits

The following are selected from the nine hits.

I_{50} sheep vesicle prostaglandin cyclooxygenase by phenols[94b]

$$\log 1/C = -1.71(\pm 0.25)\sigma^+ + 0.69(\pm 0.12)\text{Clog } P + 1.80(\pm 0.32)$$

$$n = 25,\ r^2 = 0.933,\ s = 0.186,\ q^2 = 0.910 \quad (3)$$

Acetyltransferase transfer of the acyl group from p-nitrophenylacetate to $X-C_6H_4NH_2$[95]

$$\log V_{max}/K_m = -1.25(\pm 0.46)\sigma^+ + 0.89(\pm 0.46)\log P + 0.65(\pm 0.31)Es_3 + 1.3(\pm 0.74)$$

$$n = 10,\ r^2 = 0.907,\ s = 0.243,\ q^2 = 0.787 \quad (4)$$

The positive Es term means that meta substituents are inhibitory since values of Es are negative.

I_{50} prostaglandin synthase by phenols[94a]

$$\log 1/C = -1.08(\pm 0.40)\sigma^+ + 0.74(\pm 0.33)\text{Clog } P + 1.23(\pm 0.70)$$

$$n = 7,\ r^2 = 0.939,\ s = 0.132,\ q^2 = 0.974 \quad (5)$$

The action classification presents the same difficulty. For example, isolating cell studies with **2 B4** we obtain 2078 QSAR for all kinds of cells. To get some idea of what has been studied, enter **show** followed by **1 4**. Now we can page through the 2078 sets in 30 min to get some idea of what has been done. Returning to search and using **2 B4**, we can search on the following terms.

system	name	hits
1	**hepatocyte**	5
1	**coli**	101
1	**HIV**	150
1	**caco**	8
1	**Aureus**	119
1	**Fungi**	68
1	**red Erythrocyte**	85
1	**Niger**	25
1	**Typhimurium**	39
1	**Diphtheria**	6

One needs to inspect the sequestered data since there can be some misleading information. In the search for coli, one set is for *E. coli* topoisomerase. In the case of the aureus search, one obtains mostly data on *S. aureus* but a few examples are for *M. aureus*. In the instance of the fungi search, checking the output we find three examples on wood destroying fungi. It would be suggested that this would be better entered under plants, but few would think to look for it there!

Next searching on action (4), we find the following examples.

system	name	hits
4	**Pen Perm** (cell penetration)	25
4	**Hemolysis**	51
4	**Narcosis**	13
4	**I50**	172
4	**Kill**	128
4	**Inh**	1058
4	**Mutagenesis**	23
4	**Luminescence**	16
4	**Cytolysis**	2
4	**oxidative**, **phosphorylation**	6

Hydrophobicity is important in 62% of the examples. What is even more remarkable is it's absence in so many examples.[9] Next moving to a subsection of cells **B4C**, we can scan 710 sets for work with cancer cells.

system	name	hits	type
1	**Chinese CHO**	40	Chinese hamster ovary
1	**Tumor**	20	Misc. tumor cells
1	**Ascites**	7	
1	**Leukemia**	41	
1	**Hela**	14	
1	**ovarian**	123	human cells
1	**colon**	24	human
1	**Myeloma**	4	
1	**Prostate**	5	human

Considering multicellular organisms (B6), we can illustrate subsection searching as follows on the 1350

QSAR in this class.

system	name	hits	type
1	mouse mice	289	
1	" cat "	17	
1	Dog	17	
1	Frog	7	
1	Rabbit	53	
1	Tadpole	30	
1	Guinea pig	278	
1	Not Guinea	22	isolates pig and pig parts
1	Fly	60	variety of flies
1	cockroach	29	including nerve chords
1	Goldfish	12	

Again we find that judicious thought must be used in entering the appropriate search commands. One always needs to inspect ones hits to be sure that unwanted data is not isolated. Further refinements of the search strategy are needed to minimize complexity yet increase recoverability and accuracy. In the case of cat, if we do not use quotes we obtain data on catfish. Searching on **1 guinea pig** isolates 17 examples on guinea pigs. In the case of a cockroach search, inspection of the results will disclose examples on both whole insect studies and isolated receptors (inhibition of nerve chord of cockroaches).

Using parameters as the searching tool can be helpful in getting lateral support for a newly developed QSAR. The following three examples illustrate esoteric kinds of studies that have been reported.

1	15 S−	282
2	2 B6	39
3	16 0<S−<3	32

The first step isolates all examples in the biological database having a σ^- term. The second narrows the focus to multicellular organisms; the third isolates all those having a positive coefficient with σ^- in the range of 0−3. Some examples are as follows.

Concentration of $X-C_6H_4-NH_2$ inhibiting root elongation of cabbage seeds[96]

$$\log 1/C = 0.44(\pm 0.12)\sigma^- + 0.69(\pm 0.10)\text{Clog } P + 2.10(\pm 0.18)$$

$$n = 7,\ r^2 = 0.991,\ s = 0.052,\ q^2 = 0.965 \quad (6)$$

Catalytic activity in generating NO from nitroglycerin by $X-C_6H_4SH$[97]

$$\log k = 1.18(\pm 0.68)\sigma^+ + 0.80(\pm 0.75)\text{I} - 9.18(\pm 0.35)$$

$$n = 8,\ r^2 = 0.941,\ s = 0.265,\ q^2 = 0.840\ \text{I} = 1 \text{ for X} = \text{COOH} \quad (7)$$

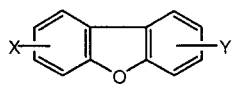
I_{50} of growth of pollen tubes in tobacco plants by $X-C_6H_4-NO_2$[98]

$$\log 1/C = 0.85(\pm 0.23)\sigma^- + 2.85(\pm 0.43)$$

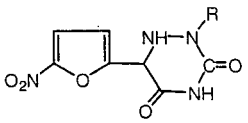
$$n = 8,\ r^2 = 0.932,\ s = 0.160,\ q^2 = 0.869$$

outliers: 2,3,6-tri-NO_2, 2,4,6-tri-NO_2 (8)

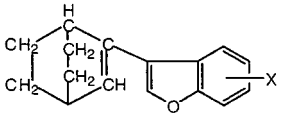
Turning now to a MERLIN search, we can use the furan nucleus to illustrate a structural approach to model mining. It must be noted that the furan unit may be present as a side chain attachment in only one or two members of the set. The hits should be inspected by first screening the 222 sets uncovered by the MERLIN search and then going to the show mode and scanning **3** and **4** for activity and compound name. One can then take the set number of interest and display the 2-D structures. Some representative examples follow.



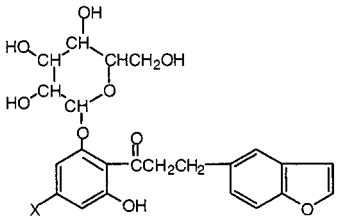
Displacement of tetrachloro dioxin from protein



Inhibition of *E. coli* growth



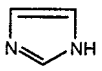
Binding to muscarinic receptor in guinea pig cerebral cortex



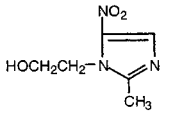
Increase in urinary glucose excretion in rats

Keep in mind that behind each structure there is a QSAR that can be loaded for suggestions to make more active congeners or avoid making less potent or toxic derivatives.

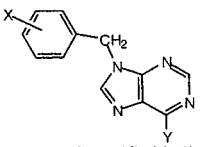
Similarly searching on produces 550 hits. Reducing



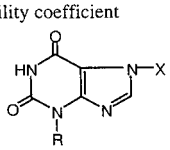
this by **2 B5** (organs and tissues) yields 106 examples. Perusing this in the show mode with **1 3 4** we can view the system, compound, and action where we note a large number of examples related to the brain. Searching with **1 brain cerebral** isolates 29 QSAR of which the following are examples.



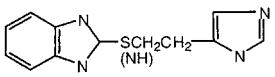
One example of radio labelled heterocycles. Rat brain capillary permeability coefficient



I_{50} of specific binding of [H^3] Diazepam to rat brain benzodiazepine receptor

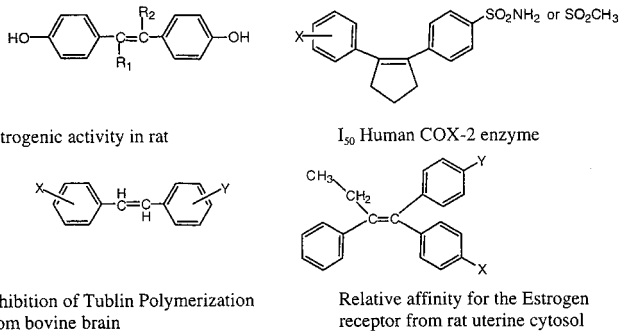


K_i affinity for Al-receptor of cerebral cortex of guinea pig

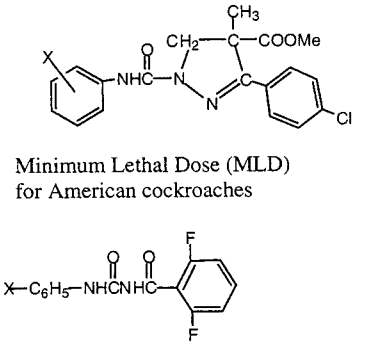


Inhibition of H^3-N-2-methylhistamine binding to rat cortex

Another example of the huge number of possibilities is similarity searching on $C_6H_5CH{=}CHC_6H_5$ that gets 79 hits of which the following are interesting.



To explore the area of insecticides, use command **2 B6B** to get 196 sets. Then **13 urea** isolates 17 sets from which the following two examples were selected.



Optimal Hydrophobicity. Up to this point we have avoided consideration of QSAR with nonlinear terms. These often may be of primary interest. They appear in two forms: parabolic (e.g., $a(\log P) - b(\log P)^2$) or the bilinear model in which activity normally increases linearly up to an optimum and then descends linearly or levels off. These are obtained via nonlinear regression analysis. Neither set of terms is an ideal solution. The parabola forces data into a symmetrical relationship, and it is often apparent that the relationships are not perfectly symmetrical. The most unsatisfactory aspect of the parabola in terms of comparative QSAR is that the slopes are not comparable with linear QSAR. In principle, the bilinear form is ideal in that the initial (upward) slopes can be compared with linear QSAR. Moreover, it is often found that an increase in hydrophobicity increases activity only up to a certain point which then levels off. This is especially true for enzymes where hydrophobic space may be limited. A serious problem with the bilinear terms is that unless there is a good spread in values of the dependent variable, the slopes have completely unrealistic values. Generally, this is easy to spot for someone who has had experience in the QSAR field. For instance, it is known that slopes of log P and π in simple linear equations rarely exceed ± 1.2.[4] Despite the unrealistic slopes, the estimates of the optimum value are usually good when they can be compared with that obtained via the parabolic QSAR.

To search the database for compounds having log P_O, use the following commands:

1	**15 logP**	4123 hits
2	**15 logP**2 bilin(logP) bilin(ClogP)**	1026 hits
3	**17 1.5<logP<2.5**	101 hits

In step 2, log **2 represents log P^2. Command 3 narrows the catch to log P_O values between 1.5 and 2.5. To inspect the results, we move to **sh**ow and enter **17**. For parabolic equations, log P_O is displayed with its confidence limits, when it is possible to calculate them.

One of the advantages of the parabolic model is that an estimate of log P_O can be obtained without having datapoints on the down side of the curve, which is necessary to derive the bilinear model. Further information on these QSAR can be obtained using the usual codes. **1 3 4 17** displays system, compound, action, and log P_O. It is instructive to compare log P_O for QSAR on cells with that on whole animals. Entering **2 B4** finds 2063 QSAR on all types of cells. Then **15 logP**2 bilin(logP) bilin(ClogP)** isolates 295 cases where log P_O is established. Moving to **sh**ow and entering **3 17** and surveying the results, we find that charged compounds (quaternary ammonium and guanidinium analogues) have distinctly lower log P_O . When these and those without good confidence limits as well as partially ionized acids and bases are omitted, the remaining sets have an average log P_O of about 4.3. Repeating the process for vertebrates using **2 B6A** locates 179 examples with an average log P_O of about 2.8. This is significantly lower than the value for cells. We believe the difference is due to entrapment of hydrophobic chemicals in the fatty sites in animals (compared to cells) and also to P-450 metabolism (there is evidence that hydrophobic compounds induce P-450, ref 2, p 313). log P_O can be a measure of optimum bioavailability. We have found that log P_O of about 2 is ideal for CNS penetration by neutral compounds.[99] This figure could be shifted up or down depending on the nature of the receptor and any special metabolic liability. It is our belief that it is prudent to make drugs as hydrophilic as possible commensurate with efficacy.[99] Of course, ascertaining exactly what efficacy is in humans is by no means simple. Short-term use is one problem, but long-term use is quite another. This is especially true today when a person may be dependent on one or more drugs for a decade or even longer. The trend to do the screening of potential drugs on cells, rather than animals, makes selection for animal studies difficult. We believe that QSAR will gradually increase our ability to anticipate toxic molecular configurations.[27]

VII. Model Mining for Active Lead Compounds

A major challenge in the development of new bioactive compounds is that of finding a promising lead molecule. Sometimes luck plays an important role. The drug Viagra for erectile dysfunction was stumbled upon during the development of a heart drug. Thalidomide, a drug that caused terrible birth

defects in children in pregnant women, now shows real promise in the treatment of leprosy. Cisplatin, which emerged from a study on the effects of an electric field on growing bacteria, is one of the most successful albeit toxic anticancer agents. Nalidixic acid, a mediocre antibiotic, was converted into the first of the fabulous quinolone carboxylates (floxins) with the help of QSAR.[100] "Me too" drugs are the bane of every drug company. Once a potential drug starts showing promise in the FDA phase I−III trials, all efforts increase in attempts to find more effective variations. On the other hand, scores of drugs have been found by random screening of extracts from plants and simple organisms.

Once a lead compound has been selected, there are two options. One can proceed with combinatorial synthesis or the use of classical QSAR to optimize activity and minimize toxicity. With the combinatorial approach, at some point QSAR and/or structure-based design will be necessary to maximize activity and avoid toxicity and vice versa with the initial QSAR approach.

With our present system there are two approaches for looking for new leads. One can look for highly active compounds by the search

14 log1/C>n or **14 log@max>n**

The first finds all sets in which every compound that has a log 1/*C* of *n* or greater. The second finds those sets in which at least one compound has a log 1/*C* of *n* or greater. The possibilities in our present biological database are as follows.

1	**14 log1/C>9**	8 hits
2	**14 log1/C@max>9**	317 hits
3	**14 log1/C>8**	29 hits
4	**14 log1/C@max>8**	890 hits
5	**14 log1/C>7**	163 hits
6	**14 log1/C@max>7**	1670 hits
7	**14 log1/C>6**	563 hits
8	**14 log1/C@max>6**	2392 hits

One can select any of the above four mining levels or lower ones. Once a level of activity is set, then that output can be further refined using the parameters of Tables 1−3. For instance, after selecting the level of 8 (item 4), the following operations might be used to narrow ones focus.

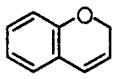
15 S+	isolates 59 sets having terms in σ^+
15 " S," " S "	finds 923 sets having a term in σ
1 HIV	finds 118 sets pertaining to HIV
2 B4	sequesters 800 sets of various cells
2 B6	picks up 346 sets in multicellular organisms
15 logP2 bilin(logP) bilin(ClogP)**	180 sets

It is of interest to consider the group of 317 QSAR of search 2 above to inspect the distribution according to system.

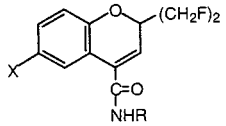
B1	macromolecules	1 hit
B2	enzymes	161 hits
B3	organelles	4 hits
B4	single cell organisms	95 hits
B5	organs/tissues	53 hits
B6	multicellular organisms	45 hits
total		357 hits

The difference between 317 and 357 is that, as mentioned above, some sets are given two labels.

Next, we further illustrate similarity searching via MERLIN by scanning the data (317 sets) obtained by **14** log1/C@max>9 using **13**

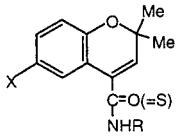


to obtain two examples of interest.



EC_{50} rat potassium channel openers

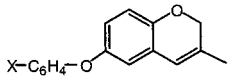
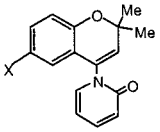
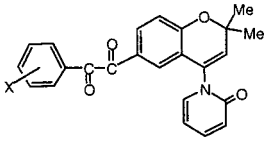
highest log 1/C = 9.4



EC_{50} rat potassium channel openers

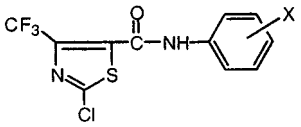
highest log 1/C = 9.6

Searching the whole database of 8500 sets we find the above two plus seven others, three of which contain other structures. The following illustrates the methodology.

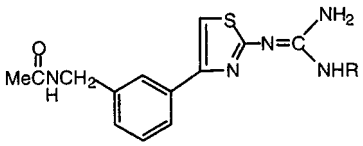
Parent structure	Activity	Highest log 1/C
	Inhibition of Human Platelet 5-lipoxygenase	7.6
	Rat aorta EC_{50} relaxant of potassium channel activator	8.0
same as above	Rate trachea relaxation of potassium channel activation	7.7
	EC_{50} relaxant of rat aorta	7.1

Similarity searching on

and then **14 log1/C@max>9** obtains four hits: in two of which the heterocyclic unit is only a substituent. The other two sets are of interest



Inhibition of Mitochrondrial succinate dehydrogenase



Inhibition (MIC) of Helicobacteria Pylori

Note the large numbers of searches that are possible. Any subsets from Table 2 can be used at any of the eight levels of searching suggested above; in addition, the subsets could be narrowed by the use of any of the parameters. The data sets selected have a QSAR that has information suggesting the next moves to avoid or to cultivate in designing better molecules.

VIII. On the Use of the Combined Databases

The reason for having a database of QSAR from mechanistic organic chemistry is 2-fold. This project was started over 30 years ago in part because of our curiosity about the large number of 'Hammett' equations that were constantly appearing in scientific literature. A more compelling reason slowly became apparent. Familiarity with QSAR from physical organic chemistry can provide an excellent basis for understanding and supporting the enormously more complex QSAR from biomedicine.[4,5,7,9]

From the very beginning of our work in the early 1960s, we have worried about formulating meaningless QSAR. In the early days we did bolster our spirits by finding similar QSAR for comparative support. For instance, an extensive review of the QSAR of simple alcohols showed general agreement in a number of ways.[91] Most encouraging were the early studies using molecular graphics[2,10−12] and QSAR to analyze ligand binding to a variety of enzymes whose crystallographic structures had been established.

A worrisome factor is the occurrence of outliers. Sometimes these are easy to understand when the structural changes in a parent molecule are very different from the other members of a set. Also, our parameters are not perfect, and this too may be hard to fathom. Finally, we have found that experimental errors are easy to make but difficult to establish. Another serious problem is that of collinearity caused by poor selection of substituents or other structural changes. Hence, it is very important to find support for a new QSAR by all reasonable means. Similar studies from the same or similar systems are the best way. At present, when possible, we like to make comparisons with studies from mechanistic organic chemistry. There are a variety of ways to do so.

For example, we might search the double database via functional groups as follows

1	**12 nitrobenzene**	155 hits
2	**3 not misc**	155 hits
3	**15 not logP**	86 hits
4	**15 S−**	23 hits

A quick 30-s scan of the data after step 4 finds a number of QSAR of interest containing the parameter σ^-.

Many environmental studies of *mixed* sets of chemicals have been made and correlated with log *P*, but the above results suggest that often log *P* does not enter the picture in variety of toxicology studies. Note that polynitro compounds behave according to a different mechanism, see eq 13. Care must be taken before sequestering chemicals together for a correlation analysis until it is established that we are dealing with a homogeneous reaction mechanism.

The following are representative examples of the activity of the NO_2 function.

Reduction of 4-X−$C_6H_4NO_2$ by $CH_3\dot{C}HOH$ in N_2O-saturated solution[106b]

$$\log k = 0.85(\pm 0.15)\sigma^- + 8.26(\pm 0.11)$$

$$n = 13,\ r^2 = 0.932,\ s = 0.125,\ q^2 = 0.915$$

outlier: X = H (9)

Reduction of X−$C_6H_4NO_2$ by pyrimidine-saturated N_2O[106c]

$$\log k = 1.05(\pm 0.13)\sigma^- + 0.06(\pm 0.09)$$

$$n = 13,\ r^2 = 0.965,\ s = 0.120,\ q^2 = 0.944 \quad (10)$$

Reduction of X−$C_6H_4NO_2$ by xanthine oxidase[106d]

$$\log k = 0.98(\pm 0.16)\sigma^- - 0.35(\pm 0.23)B5_2 + 2.13(\pm 0.27)$$

$$n = 26,\ r^2 = 0.884,\ s = 0.201,\ q^2 = 0.865$$

outliers: 4-SO_3^-, 4-SO_2NH_2, 4-CHO (11)

Acute toxicity of X−$C_6H_4NO_2$ to fathead minnows[106e]

$$\log 1/C = 1.44(\pm 0.31)\sigma^- + 3.85(\pm 0.22)$$

$$n = 12,\ r^2 = 0.914,\ s = 0.242,\ q^2 = 0.866$$

outliers: 3,4-di-Cl, 4-Br (12)

I_{50} of X−$C_6H_4NO_2$ to Daphnia Magna[106f]

$$\log 1/C = 0.98(\pm 0.22)\sigma^- + 2.62(\pm 0.41)$$

Equations 9 and 10 suggest that a radical reaction

$$n = 10,\ r^2 = 0.927,\ s = 0.186,\ q^2 = 0.888$$

outliers: 4-Br; 3-NO_2, 4-CH_3 (13)

is involved in the reduction of the nitro group. The biological QSAR eqs 11−13 are also correlated by σ^- with similar ρ values.

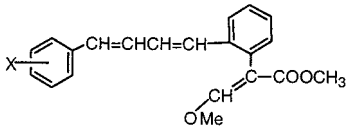
Another area of interest is the toxicity of olefins. Searching the combined databases using $CH_2{=}CHCH{=}CH_2$ followed by **15** **S+** gets 15 hits, two of which are of interest.

Addition of :CCl_2 to trans-X−C_6H_4CH=CHCH=CH_2[101]

$$\log k_{rel} = -0.42(\pm 0.03)\sigma^+ - 0.01(\pm 0.02)$$

$$n = 9,\ r^2 = 0.994,\ s = 0.025,\ q^2 = 0.991 \quad (14)$$

I_{50} to P. falciparum NF54[102] *by*

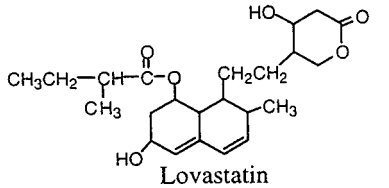


$$\log 1/C = -1.19(\pm 0.55)\sigma^+ + 1.43(\pm 0.32)B5_2 + 0.41(\pm 0.25)L_4 + 6.21(\pm 0.70)$$

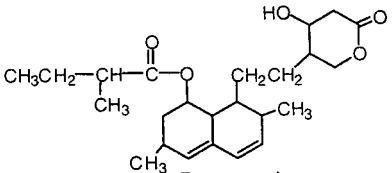
$$n = 12,\ r^2 = 0.948,\ s = 0.222,\ q^2 = 0.896$$

outlier: 2,4-di-CH_3 (15)

A reason for our interest in the above two equations is the fact that butadiene has long been known to be carcinogenic. The fact that σ^+ correlates the electronic effect with a negative coefficient ρ suggests a radical reaction.[6] Also of interest is the reported[103] carcinogenicity in rodents of the two widely used cholesterol-lowering statins that contain a butadiene unit.

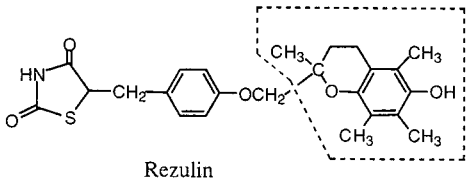


Lovastatin



Pravastatin

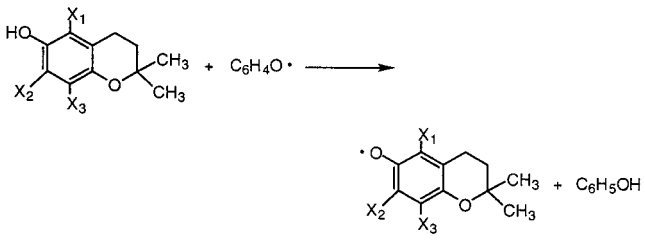
QSAR can alert one to toxicity features that can then be checked experimentally. Another example of toxicity that might have been anticipated today, but was not in the past, is the drug rezulin. Rezulin was withdrawn from the market when it was found to cause serious liver damage.



Rezulin ClogP 5.58

The encircled portion of the above structure is identical to that in vitamin E. However, vitamin E has a long hydrophobic carbon chain that gives it a calculated Clog *P* of 12. In fact, this is so high that it cannot be measured. This chain evolved over time for a reason. It would anchor the vitamin into a large hydrophobic region (e.g., the cell membrane) with its polar phenolic moiety near the surface to scavenge radicals. The more hydrophilic rezulin is freer to wander about and form a reactive radical intermediate via interaction with ROS (reactive oxygen species produced by cells burning oxygen).

To test radical scavenging ability, Mukai et al.[104] examined the following reaction; the data from this study was used to derive the following QSAR (σ^+ is selected with respect to OH).



$$\log k = -1.08(\pm 0.32)\sigma^+ + 0.37(\pm 0.28)B1_3 + 2.35(\pm 0.39)$$

$$n = 10,\ r^2 = 0.908,\ s = 0.095,\ q^2 = 0.790 \quad (16)$$

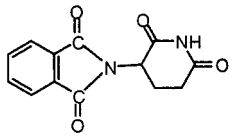
In this system, $C_6H_4O^\bullet$ is a model for the ROS. $B1_3$ accounts for the steric effect of X_3 and shows that substituents in this position have a positive effect on the reaction. We assume this may inhibit solvation by the solvent ethanol that would tend to localize electrons on the ether oxygen, thus inhibiting hydrogen abstraction. For example, 4-methoxyphenol is carcinogenic but phenol is not. An equation similar to eq 16 has been formulated for the toxicity of simple phenols, having electronic releasing substituents, to fast growing leukemia cells.[27]

$$\log 1/C = -1.35(\pm 0.15)\sigma^+ + 0.18(\pm 0.04)\log P + 3.31(\pm 0.11)$$

$$n = 51,\ r^2 = 0.895,\ s = 0.227,\ q^2 = 0.882 \quad (17)$$

Phenols with electron- attracting substituents do not fit this QSAR, and their toxicity is correlated by log *P* alone. Thus, as our database grows, it will provide more information understandable in mechanistic terms to help in the design of better drugs and to aid in the understanding of ligand−receptor interactions at the molecular level. There are numerous examples, especially with potential anticancer drugs, where studies of QSAR from mechanistic organic chemistry can be compared with chemical−biological interactions to clarify reaction mechanisms.[4,9]

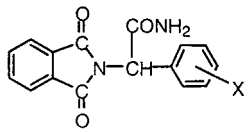
A compound that has recently attracted renewed interest is thalidomide, a teratogenic drug that is now being investigated in the treatment of leprosy and cancer.



Phthalidomide

A MERLIN search on the combined physical−biological database using phthalimide finds four datasets. One has a substituent pattern too complex for consideration and hence a very weak correlation. Equations 18, 19, and 21 are of potential interest. We also find that there are no physical QSAR based on this phthalimide structural feature.

Inhibition of necrosis factor[106g] *by*

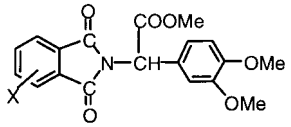


$$\log 1/C = -0.25(\pm 0.26)\sigma^+ + 0.69(\pm 0.24)\mathrm{CMR} - 1.70(\pm 2.1)$$

$$n = 9,\ r^2 = 0.938,\ s = 0.153,\ q^2 = 0.886$$

outlier: 3,4-di-OC_3H_7 (18)

Inhibition of necrosis factor[106g] *by*



$$\log 1/C = -0.97(\pm 0.15)\sigma^+ + 5.14(\pm 0.12)$$

$$n = 7,\ r^2 = 0.983,\ s = 0.124,\ q^2 = 0.967$$

outlier: 2-OH (19)

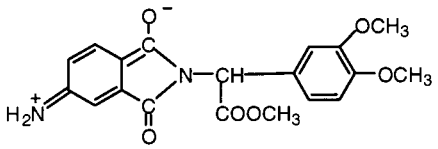
QSAR 18 has a σ^+ term of borderline value, while its activity is mainly a function of size as delineated by CMR. The range in log 1/*C* for eq 18 is 3.6−5.3, while the range in eq 19 is 4.4−6.4. Not only are congeners of QSAR eq 19 more potent, the correlation of QSAR eq 19 is much sharper. A SMILES search with benzamide yields a number of studies on hydrolysis, three of which have very similar ρ^+ terms, of which the following is an example.

Hydrolysis[105] *of* $X-C_6H_4CONH_2$ *in 40% aqueous ethanol at 65 °C*

$$\log k = -0.28(\pm 0.08)\sigma^+ - 5.10(\pm 0.03)$$

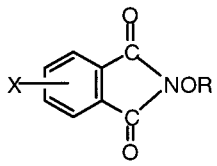
$$n = 4,\ r^2 = 0.996,\ s = 0.014,\ q^2 = 0.902 \quad (20)$$

Through resonance implied by eq 20 would suggest the following resonance form to be important in the case of compounds from QSAR eq 19.



Reaction with an electrophilic binding site or agent is implied.

An interesting comparison comes from the study of Chan et.al.[106h] for the I_{50} toxicity of a similar amide to L1210 leukemia cells.



R = H or SO_2Y where Y = alkyl or $-C_6H_4-Z$

$$\log 1/C = -0.93(\pm 0.24)\sigma^+ - 3.48(\pm 2.30)R_m - 1.30(\pm 0.82)I + 4.10(\pm 0.80)$$

$$n = 15,\ r^2 = 0.936,\ s = 0.293,\ q^2 = 0.893$$

outlier: X = H, Y = SO_2CH_3 (21)

R_m is a measure of hydrophobicity derived from chromatography. Its negative coefficient is evidence of a polar receptor. $I = 1$ for two examples where R = H. These are unique structures where the OH has a very deleterious effect on activity. It is of interest that σ^+ has the same ρ value as in QSAR eq 19. Thus, eqs 19 and 21 might be clues as to why thalidomide is effective against leprosy or cancer. At this point it would be of interest to study the reactions of thalidomide and phthalimide in more detail via classical LFER.

The above examples are only illustrative and reflective of the type of datasets that are incorporated in these database. Many more such comparisons are possible, and as the database expands, it will become much more fruitful to search via MERLIN for novel comparisons. Again using similarity searching on $C_6H_5CH{=}CH_2$, we find a number of reactions of styrenes and styrene derivatives with radicals from mechanistic organic chemistry.

Reaction of $X-C_6H_4CH{=}CH_2$ *with 4-*$Cl-C_6H_4S^\bullet$ [106i] *in cyclohexane*

$$\log k = -0.58(\pm 0.15)\sigma^+ + 7.73(\pm 0.06)$$

$$n = 7,\ r^2 = 0.949,\ s = 0.055,\ q^2 = 0.924 \quad (22)$$

Reaction of $X-C_6H_4CH{=}CH_2$ *with* $C_6H_5S^\bullet$ [106i] *in cyclohexane*

$$\log k = -0.33(\pm 0.08)\sigma^+ + 7.45(\pm 0.03)$$

$$n = 6,\ r^2 = 0.970,\ s = 0.026,\ q^2 = 0.842$$

outlier: 4-Br (23)

Reaction of $X-C_6H_4CH{=}CH_2$ *with* $(CH_3)_3COO^\bullet$ [107] *in benzene*

$$\log k_{rel} = -0.31(\pm 0.22)\sigma^+ + 0.04(\pm 0.09)$$

$$n = 5,\ r^2 = 0.862,\ s = 0.063,\ q^2 = 0.645 \quad (24)$$

Reaction of $X-C_6H_4CH=CH_2$ with $Cl_3C^{\bullet}$[108] *in benzene*

$$\log k_{rel} = -0.49(\pm 0.13)\sigma^+ + 0.05(\pm 0.04)$$

$$n = 8,\ r^2 = 0.937,\ s = 0.044,\ q^2 = 0.882$$

outliers: 4-CN, 4-NO_2 (25)

Reaction of $X-C_6H_4C(CH_2CH_3)=CH_2$ with $:CCl_2$[109]

$$\log k_{rel} = -0.37(\pm 0.13)\sigma^+ - 0.03(\pm 0.05)$$

$$n = 5,\ r^2 = 0.964,\ s = 0.032,\ q^2 = 0.786 \quad (26)$$

Reaction of $X-C_6H_4CH=CHC_6H_5$ with $^{\bullet}SCH_2COOH$ heat[110]

$$\log k_{rel} = -0.40(\pm 0.18)\sigma^+ - 0.01(\pm 0.07)$$

$$n = 5,\ r^2 = 0.944,\ s = 0.041,\ q^2 = 0.828$$

outlier: 3,4-di-OMe (27)

There are fewer examples from biological systems for comparison.

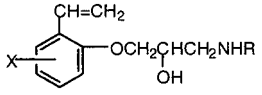
Elevation of serum alanine transaminase in mice due to hepatic toxicity by $X-C_6H_4CH=CH_2$[111]

$$\log 1/C = -0.46(\pm 0.26)\sigma^+ + 3.22(\pm 0.18)$$

$$n = 6,\ r^2 = 0.862,\ s = 0.118,\ q^2 = 0.738$$

outlier: H (28)

β-Adrenoceptor blocking activity of complex styrenes in right atria of guinea pigs[112]

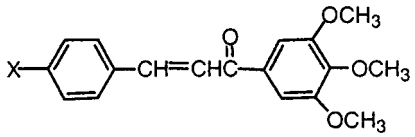


$$pA_2 = -0.98(\pm 0.22)\sigma_X^+ + 0.53(\pm 0.32)B1_{X,5} - 1.36(\pm 0.31)B1_R + 7.41(\pm 0.43)$$

$$n = 21,\ r^2 = 0.894,\ s = 0.184,\ q^2 = 0.839$$

outliers: R = CMe_3, X = H; R = $CHMe_2$, X = 3,5-di-Cl; R = CMe_3, X = 3,5-di-Cl; R = CMe_3, X = 3-Me, 5-Cl; R = CMe_3, X = 3-Me (29)

Toxicity to HeLa cells compared to colchicine of[113]

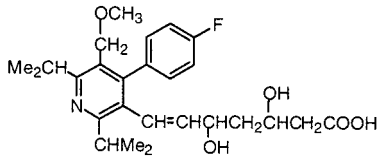


$$\log 1/C = -1.51(\pm 0.32)\sigma^+ - 0.62(\pm 0.26)B5_4 + 4.36(\pm 0.69)$$

$$n = 12,\ r^2 = 0.931,\ s = 0.321,\ q^2 = 0.880$$

outliers: 4-NH_2, 4-Br, 6-CF_3, 4-NHC_4H_9 (30)

It is clear that σ^+ is the parameter of choice, suggesting a radical mechanism in these pharmacological actions. Hence, it would be an exercise in futility to try to develop a drug in which an aromatic $CH=CH_2$ is conjugated to an electron-rich moiety. As in the butadiene cases, all of these examples are correlated with σ^+ having negative ρ values. Another drug with liver toxicity recently withdrawn from the market is baycol.



Here we find a styrene-like moiety that may well be the cause of toxicity.

Another functional group that has received attention from chemists and biologists is the sulfonamido entity. Equation 31 shows the substituent effect on ionization of $X-C_6H_4SO_2NH_2$.[114]

$$pK_a = -0.87(\pm 0.07)\sigma + 10.0(\pm 0.04)$$

$$n = 13,\ r^2 = 0.985,\ s = 0.058,\ q^2 = 0.977 \quad (31)$$

Thus, the ρ value for ionization would be 0.87.

The following biological examples can be compared with the effect of substituents on the acidity of the sulfonamide function. One can then determine if the ionization of sulfonamides impacts their biological activity.

Inhibition of lyase, carbonic anhydrase by $X-C_6H_4-SO_2NH_2$[115]

$$\log 1/C = 0.90(\pm 0.23)\sigma + 0.23(\pm 0.17)\text{Clog } P + 5.36(\pm 0.15)$$

$$n = 16,\ r^2 = 0.930,\ s = 0.176,\ q^2 = 0.884$$

outlier: 2-Me, 2-Cl, 2-NO_2 (32)

Naturiuretic action in rats of $X-C_6H_4SO_2NH_2$[116]

$$\log 1/C = 0.77(\pm 0.22)\sigma - 0.16(\pm 0.16)\text{Clog } P + 0.30(\pm 0.13)$$

$$n = 14,\ r^2 = 0.849,\ s = 0.151,\ q^2 = 0.734$$

outliers: 3-NO_2, 4-Cl; 4-NO_2, 3-CF_3 (33)

Despite the extra term in eq 33 and the fact that the action is occurring in rats, the agreement with eq 31 in terms of ρ is good. The following is another example in whole animals.

ED_{50} against electroshock seizures in mice by $X-C_6H_4SO_2N(Y)_2$[117]

$$\log 1/C = 0.91(\pm 0.25)\sigma_X + 0.47(\pm 0.16)\text{Clog } P - 0.58\log(\beta \cdot 10^{\text{Clog } P} + 1) + 3.03(\pm 0.12)$$

$$n = 16,\ r^2 = 0.913,\ s = 0.100,\ q^2 = 0.836,\ \beta = -1.31$$

outliers: X = 4-Br, Y = OCH_3, H; X = 4-Br, Y = CH_3, CH_3 (34)

Despite the complexity of QSAR eq 34, the ρ value is in good agreement with eqs 31−33.

We have been interested in studying the use of the sterimol parameter B1 for the correlation of steric effects emanating from the ortho position. From the combined datasets we can make the following search.

15 B1	1220 hits
12 phenol	33 hits

This finds 33 QSAR based on phenols. Inspecting the results of a mechanistic organic chemical reaction for comparison with a biological QSAR can be done by viewing the results with machine sorting on the coefficient associated with B1.

Bond dissociation energy (BDE) of phenols in kcal/mol[118]

$$\text{BDE} = -2.16(\pm 0.54)B1_2 + 3.91(\pm 0.80)\sigma^+ + 88.9(\pm 0.97)$$

$$n = 14,\ r^2 = 0.955,\ s = 0.584,\ q^2 = 0.926$$

outlier: H (35)

Sulfation of phenols by human liver sulfotransferase[119]

$$\log V_{max}/K_m = -1.91(\pm 0.60)B1_2 - 0.93(\pm 0.26)B5_4 + 0.71(\pm 0.51)\sigma^- + 0.05(\pm 1.1)$$

$$n = 17,\ r^2 = 0.870,\ s = 0.422,\ q^2 = 0.670$$

outliers: 3-NH_2, 4-NH_2, 3-CH_3, 3-C_2H_5 (36)

Although eq 36 is not a very good correlation since four data points had to be omitted, the comparison of the two steric effects would seem to make sense in that the removal of hydrogen in each example is critical. The electronic effects in the two sets are quite different, reflecting a homolytic bond dissociation reaction in QSAR eq 35 (removal of $^{\bullet}H$) and a heterolytic reaction in QSAR eq 36 (removal of a proton) where one normally finds σ^- to be the parameter of choice for phenols. Steric effects in QSAR eqs 35 and 36 are independent of electronic effects.

Running a similarity search on the double database with σ^- turns up many interesting QSAR for comparison. Searching with **15 S−** finds 1362 QSAR with σ^- terms. Next, using 16 in Table 1 **16 .7<S−<2** isolates 329 QSAR with ρ between 0.7 and 2. Now using the sort procedure all QSAR are listed in order of increasing slopes on σ^-. One of the first equations that appears is QSAR eq 36 above for the enzymatic sulfation of phenols. Another example of enzymatic sulfation is that of $X-C_6H_4CH{=}NOH$.[120]

Sulfation by arylsulfotransferase

$$\log V_{max}/K_m = 0.75(\pm 0.25)\sigma^- + 0.56(\pm 0.40)\text{Clog } P + 6.21(\pm 0.86)$$

$$n = 5,\ r^2 = 0.990,\ s = 0.072,\ q^2 = 0.897 \quad (37)$$

This makes sense in that the removal of hydrogen in each example is critical. The electronic effects in the two sets are equivalent. This resembles phenols H-bonding in 1,2-dichloroethane with pyridine[121]

$$\log k = 0.73(\pm 0.13)\sigma^- - 0.67(\pm 0.13)B1_2 + 1.95(\pm 0.20)$$

$$n = 17,\ r^2 = 0.941,\ s = 0.099,\ q^2 = 0.896$$

outlier: 2,4,5,-tri-Cl (38)

Now a search for examples with σ^- in the range 2−3 finds 104 examples of which the following are illustrative.

Ionization of phenols in aqueous solution[122]

$$\log K = 2.01(\pm 0.15)\sigma^- + 1.94(\pm 0.34)F_2 - 9.86(\pm 0.08)$$

$$n = 23,\ r^2 = 0.979,\ s = 0.146,\ q^2 = 0.966$$

outliers: 4-F, 2-$C(Me)_3$, 2-NO_2 (39)

In this expression, F_2 is the field/inductive parameter for ortho substituents. Fujita and co-workers[123] established that this parameter adequately accounts for the importance of the electronic effect of ortho substituents beyond that accounted for σ, constants used for ortho substituents. Our analyses substantiate this finding.

A comparable biological example involves the uncoupling of phosphorylation of mitochondria from ascaris muscle[124]

$$\log 1/C = 2.04(\pm 0.21)\sigma^- + 0.93(\pm 0.20)\text{Clog } P + 0.47(\pm 0.48)$$

$$n = 21,\ r^2 = 0.967,\ s = 0.393,\ q^2 = 0.955$$

outliers: 2-I, 4-CN, 6-NO_2; 2,6-di-I, 4-NO_2; 4-COMe (40)

Next we consider the parameters σ_I and σ^* that have developed from two different systems to model field/inductive effects of substituents. Searching the double database we find 260 QSAR based on the former and 816 based on the latter. These two parameters, as one might expect, are highly collinear. We have 362 substituents with both values that show a mutual correlation of $r^2 = 0.911$.

Searching the double database with σ^* **16** **2**<**S′**<**4** (S′ = σ^*) finds 79 examples.

Alkaline hydrolysis at 35 °C in 15% aqueous ethanol of $RCOOC_2H_5$[125]

$$\log k = 2.25(\pm 0.89)\sigma^* + 1.04(\pm 0.18)\text{Es} - 0.42(\pm 0.35)$$

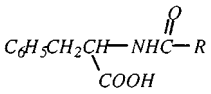
$$n = 9,\ r^2 = 0.988,\ s = 0.150,\ q^2 = 0.980 \quad (41)$$

Alkaline hydrolysis at 65 °C in 20% aqueous methanol of $RCOOC_2H_5$[126]

$$\log k = 2.51(\pm 0.42)\sigma^* + 0.91(\pm 0.08)\text{Es} - 0.22(\pm 0.34)$$

$$n = 13,\ r^2 = 0.989,\ s = 0.196,\ q^2 = 0.964 \quad (42)$$

Rate of hydrolysis of by carboxypeptidase[127]



$$\log k = 1.98(\pm 0.80)\sigma^* - 3.50(\pm 1.80)\text{B1} + 6.10(\pm 2.3)$$

$$n = 8,\ r^2 = 0.897,\ s = 0.416,\ q^2 = 0.801$$

outlier: $CHCl_2$ (43)

Rate of hydrolysis of 4-NO_2*-*C_6H_4*COOR by chymotrypsin*[128]

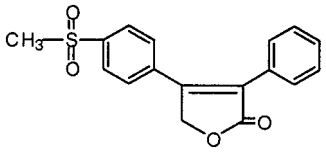
$$\log k_3 = 2.09(\pm 0.34)\sigma^* + 1.21(\pm 0.27)\text{Es} + 0.34(\pm 0.10)\text{MR} - 0.95(\pm 0.71)I - 1.91(\pm 0.29)$$

$$n = 36,\ r^2 = 0.950,\ s = 0.320,\ q^2 = 0.933$$

outliers: 3-indolyl, $(CH_2)_3NHCOCH_3$; C_6H_4-4-NO_2 (44)

In these four different examples we find rather close agreement with the σ^* terms and in three of the four cases agreement with Es terms. The common point of reaction is with the carbonyl group that is influenced by R. The positive Es coefficient implies a negative steric effect since Es values are negative. QSAR eq 44 is most interesting because of the small MR term and the indicator variable *I* that is assigned the value of 1 for instances where R = $-C_6H_4-X$. In eight such examples the $-C_6H_4-X$ moiety is assigned the value of 1 for R = $X-C_6H_4-$. Despite the complexity of QSAR eq 44, the electronic and steric effects shine through clearly and fall in line with the much simpler eqs 41−43. This is the kind of lateral support that one surely needs in formulating biological QSAR.

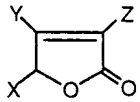
Similarity searching using the double database is of interest in examining the hydrofuranone function since it occurs in the highly successful drug Vioxx.



Vioxx

Similarity searching on 2-hydrofuranone yields 33 QSAR. Reducing this to sets that contain electronic terms yields eight QSAR.

Mutagenicity in the Ames test[129] *with S. typhimurium TA100 of*

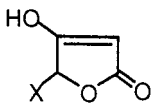


$$\log k = -14.5(\pm 1.91)E_{\text{LUMO}} - 13.5(\pm 1.90)$$

$$n = 20,\ r^2 = 0.937,\ s = 1.14,\ q^2 = 0.921 \quad (45)$$

This is a very unusual equation since there was considerable variation in X, Y, and Z; nevertheless, an excellent QSAR based on only one parameter (the energy level of the lowest unoccupied molecular orbital) is found. The QSAR would suggest care needs to be exercised in incorporating this unit into commercial products. There are three other similar equations for mutagenesis.

Only one equation from the physical database is found−that for the ionization[130] of



$$pK_a = -3.96(\pm 1.1)\sigma_I + 3.91(\pm 0.35)$$

$$n = 10,\ r^2 = 0.904,\ s = 0.343,\ q^2 = 0.861$$

outlier: CO_2Me (46)

It is hard to say whether there is any relation between these two QSAR. Of course, one would expect electron withdrawal to promote ionization. However, QSAR eq 45 shows that electron-releasing substituents promote activity. Vioxx does not contain such groups.

Now scanning the double database as follows allows us to compare two different subsections.

1	**15 S+**	2110 hits
2	**15 not **2 bilin**	2003 hits
3	**2 B2A P12**	379 hits
4	**16 −3<S+<−.9**	138 hits
5	**12 Phenol**	23 hits

The first step yields a tremendous amount of information. Step 2 eliminates QSAR with nonlinear terms; while step 3 sequesters oxidoreductase enzymes from the biological database and radical reactions from the physical database. Step 4 narrows the search to datasets with ρ in the range from -3 to -0.9, and finally in step 5, we limit the study to phenols as substrates.

The following examples display some of the results.

Oxidation by Horseradish peroxidase I[131]

$$\log k_2 = -2.68(\pm 0.78)\sigma^+ + 1.31(\pm 0.71)\pi_4 + 6.36(\pm 0.30)$$

$$n = 12,\ r^2 = 0.872,\ s = 0.397,\ q^2 = 0.741$$

outliers: 3-OH, 3,4-di-Me (47)

Hydrogen abstraction with $(C_6H_5)_2NN^{\bullet}$-2,4,6-tri-$NO_2C_6H_2$[132]

$$\log k = -2.68(\pm 0.37)\sigma^+ - 1.21(\pm 0.32)\mathrm{B1}_2 + 3.19(\pm 0.45)$$

$$n = 18,\ r^2 = 0.941,\ s = 0.291,\ q^2 = 0.901$$

outliers: H, 3-OMe, 2,3,4,5,6-penta-Cl_5 (48)

Oxidation by Mn III[133]

$$\log k = -2.60(\pm 0.69)\sigma^+ - 6.48(\pm 0.19)$$

$$n = 7,\ r^2 = 0.951,\ s = 0.190,\ q^2 = 0.921$$

outlier: 4-COMe (49)

Oxidation by fungal laccases[134]

$$\log k_{cat}/K_m = -2.28(\pm 0.55)\sigma^+ + 1.52(\pm 1.48)\mathrm{B1} - 0.82(\pm 0.63)I + 3.05(\pm 1.9)$$

$$n = 18,\ r^2 = 0.912,\ s = 0.349,\ q^2 = 0.855$$

outliers: 2-OMe-4-CH_2COO^- (50)

I_{50} of prostaglandin cyclooxygenase, sheep vesicle[135]

$$\log 1/C = -1.71(\pm 0.25)\sigma^+ + 0.69(\pm 0.12)\mathrm{Clog}\,P + 1.80(\pm 0.32)$$

$$n = 25,\ r^2 = 0.933,\ s = 0.186,\ q^2 = 0.910$$

outliers: 2,3,5,6-tetra-Me (51)

Oxidation with peroxydisulfate in aqueous solution[136]

$$\log k = -1.56(\pm 0.17)\sigma^+ + 0.20(\pm 0.07)$$

$$n = 34,\ r^2 = 0.919,\ s = 0.177,\ q^2 = 0.909$$

outlier: 2-COOH, 4-CMe_3 (52)

In QSAR eq 47, π_4 accounts for the specific hydrophobicity of para substituents. There is no overall hydrophobic effect. There is considerable evidence that a radical reaction underlies all of these equations, as we have found σ^+ to be a general parameter for radical reactions.[6,27] QSAR eq 48, a well-established radical reaction, reveals a similar ρ but with a negative steric effect for ortho substituents. Nevertheless ρ is in close agreement with eq 47.

These results are also supported by QSAR eq 53 for the cytotoxic action of simple and complex phenols (Bisphenol A, Diethylstilbestrol, Estradiol, Estriol, Equilin, Equilenin) against L1210 leukemia cells.[27]

$$\log 1/C = -1.35(\pm 0.15)\sigma^+ + 0.18(\pm 0.04)\log P + 3.31(\pm 0.11)$$

$$n = 51,\ r^2 = 0.895,\ s = 0.227,\ q^2 = 0.882 \quad (53)$$

Actually, a better correlation is obtained using calculated homolytic bond dissociation energies (BDE) in place of σ^+ ($r^2 = 0.925$). This points more directly to a radical reaction, in this cellular system.

Equation 52 is another type of radical reaction that has a similar ρ. Equation 50 is more complicated, having a positive B1 term for ortho substituents and an indicator variable that is assigned the value of 1 for 2,6-disubstituted compounds. Although it is based on a mixture of laccases, ρ is qualitatively similar to the other examples. Equation 51 has a lower ρ value similar to that of the peroxydisulfate oxidation. Other factors being equal, we have found that low ρ values suggest action by a stronger radical or a more labile H.[6]

Up to this point we have considered mostly electronic parameters for aromatic systems in making comparisons between biological and physical QSAR. Two parameters that provide easy to see connections are Es and B1. The former was developed by Taft from the hydrolysis of X-CH_2COOR

$$\sigma^* = 1/2.48[\log(k_X/k_H)_B - \log(k_X/k_H)_A]$$

In this expression σ^* represents the field inductive effect of X, k_X is the rate constant for the hydrolysis of X-CH_2COOR, and k_H is that for the hydrolysis of CH_3COOR. B denotes hydrolysis in basic solution, while A denotes hydrolysis in acid solution. Es = $\log(k_X/k_H)_A$. The above equation is based on the assumption that there is little or no electronic effect in acid hydrolysis. It is hard to be sure that the two terms are completely independent, but the evidence over the years in hundreds of examples indicates that the separation is reasonable. The Verloop-calculated values of B1 pertain to the first atom of the substituent, while Es is related to the whole substituent. B1

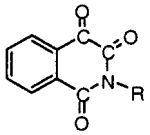
is free of electronic effects. These parameters have been discussed and illustrated,[2,75] and compilations of them have been published.[3]

The early entries into our system were primarily Es based. However, around 1990 it was discovered that B1 was often superior to Es. Also, the large number of available B1 values and their ability to be calculated made them a viable option in terms of structure−activity analysis. Comparisons of recent work entered into our system since 1995 revealed the following.

1	**5** (**1995**) (**1996**) (**1997**) (**1998**) (**1999**) (**2000**)	3187 hits
2	**15** **Es**	60 hits
3	**16** **.4<Es<2**	15 hits

In these 60 examples Es is found to be the superior parameter. In biological systems this may account for an intermolecular steric effect, while in chemical systems it is often indicative of an intramolecular effect. The following examples constitute interesting comparisons.

Relative toxicity to weeds of[137]

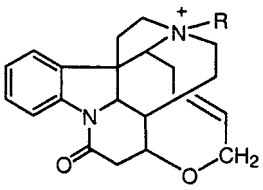


$$\log k_{rel} = 0.46(\pm 0.27)Es - 1.36(\pm 0.50)\sigma^* + 1.23(\pm 0.44)$$

$$n = 7,\ r^2 = 0.936,\ s = 0.092,\ q^2 = 0.633$$

outlier: $CHMe_2$ (54)

Affinity of derivatives of strychnine for muscarinic receptor of type 1[138]

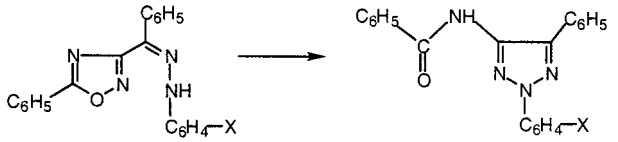


$$\log 1/C = 0.50(\pm 0.09)Es + 0.22(\pm 0.09)B5 + 4.65(\pm 0.20)$$

$$n = 10,\ r^2 = 0.858,\ s = 0.105,\ q^2 = 0.742 \quad (55)$$

Es values are lacking for R=$CH_2C{\equiv}CH$, $CH_2C_6H_4$-3-NO_2, $CH_2C_6H_4$-4-NO_2. Recall that Es values are negative, so that the positive coefficient with Es indicates a deleterious effect (steric hindrance). There is a very small positive effect from B5 that suggests that the width of a substituent enhances receptor affinity. This parameter works better than CMR or molar volume.

Rearrangement in aqueous dioxane at pH 3.8 at 313 K[139]



$$\log k_X/k_H = 0.53(\pm 0.07)Es_2 - 1.37(\pm 0.10)\sigma - 0.98(\pm 0.24)F_2 - 0.04(\pm 0.04)$$

$$n = 20,\ r^2 = 0.994,\ s = 0.065,\ q^2 = 0.988 \quad (56)$$

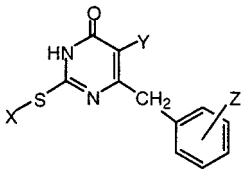
Ionization of ph $Es_{2,6}$ enols in DMSO[140]

$$pK_a = 0.57(\pm 0.15)Es_{2,6} - 6.43(\pm 0.64)\sigma^- + 17.9(\pm 0.53)$$

$$n = 15,\ r^2 = 0.975,\ s = 0.641,\ q^2 = 0.950$$

outliers: 2,4,6-tri-C_6H_5; 2,6-di-CMe_3, 4-OCOMe (57)

Inhibition of reverse transcriptase in MT-4 cells by DABO derivatives[141a]



X = Alkyl groups, Y = H or CH_3
Z = various substituents

$$\log 1/C = 0.59(\pm 0.34)Es_{Z-2,6} + 1.35(\pm 0.61)\sigma + 3.25(\pm 1.00)\text{Clog } P - 0.44(\pm 0.12)\text{Clog } P^2 - 0.50(\pm 0.25)L_{Z-4} + 2.36(\pm 0.62)F_{Z-2,6} + 0.54(\pm 2.2)$$

$$n = 41,\ r^2 = 0.869,\ s = 0.277,\ q^2 = 0.801$$

outliers: X = Me, Y = H, Z = 2,6-di-Cl; X = $CHMe_2$, Y = H, Z = 2,6-di-Cl; X = Me, Y = Me, Z = 2,6-di-Cl; X=$CHMeC_2H_5$, Y = Me, Z = 2,6-di-Cl (58)

The above five QSAR have similar Es coefficients. In addition, there are a few redundant QSAR and a few with coefficients above 1.

The biological QSAR (eqs 54, 55, 58) all have coefficients between 0.45 and 0.59 for a wide range in activities. In QSAR eq 56, the slope is close to that of eq 58. However, eq 57 is based on pK_a values, and so one needs to multiply by −1 to place the results on a log K basis which would give the Es term a negative coefficient, meaning that ortho substituents promote loss of a proton. The effect is additive since Es values are assigned to each of the two ortho positions.

In an earlier comparative study of Es, where the whole double database was considered not just the

recent years, we found 13 examples with the Es coefficient in the range 0.67−0.83. However, these had not been checked to see if B1 could replace any of the Es terms. In any case, the results show that Taft's parameter can be profitably used to deal with steric effects in biological systems that are similar to those found in physical organic chemistry. Es was designed to account for the steric effect of the whole substituent, while B1 is primarily for the first atom. In some instances we have found that B1 plus B5 can more than adequately replace Es.

Searching the double database with Es we find 579 QSAR with this term. However, checking using Esc, we find that 29 of these examples are based on Esc, a form of Es that was designed to correct for substituent hyperconjugation (see ref 141b). Searching with B1 we find 1203 examples where B1 is superior to Es. Actually it is anticipated that this disparity will increase when the data is reexamined in order to establish the superior parameter.

Focusing on more recent work we can do the following search using the double database.

5 (**1999**) (**2000**) (**2001**)	1330 hits
15 **S+**	76 hits

Scanning the 76 sets, a study on the inhibition (I_{50}) of endothelial cell nitrous oxide synthetase by substituted 2-aminopyridines attracts our attention.[142] *I* is an indicator variable that accounts for substitution in position 5.

Inhibition of nitrous oxide synthetase by 2-amino-X-pyridines[142]

$$\log 1/C = -2.48(\pm 0.76)\sigma^+ - 0.84(\pm 0.30)\text{Clog } P - 0.73(\pm 0.50)I + 6.70(\pm 0.51)$$

$$n = 17,\ r^2 = 0.853,\ s = 0.394,\ q^2 = 0.747$$

outliers: H, 6-Me (59)

This can be compared with QSAR eq 60.

Complex formation between X-pyridines and H_9 tetraphenylporphin[143]

$$\log k = -1.36(\pm 0.19)\sigma^+ + 1.20(\pm 0.13)$$

$$n = 5,\ r^2 = 0.994,\ s = 0.090,\ q^2 = 0.971 \quad (60)$$

The correlation between these two QSAR may be fortuitous, but it could be a lead of interest. While our main interest is in comparative QSAR analysis, searching for new leads is a prime interest of many.

IX. QSAR Based on Data from Humans

The most interesting subject for the development of comparative QSAR is that of humans. Although there is little such work, there are some interesting examples. Searching with **2** **B6H**, we find 42 sets of which we have selected the following examples.

Sweet taste of X-2-amino-4-nitrobenzenes[144]

$$\log \text{RBR} = -0.66(\pm 0.28)\sigma^+ + 1.32(\pm 0.24)\text{Clog } P - 0.07(\pm 0.48)$$

$$n = 9,\ r^2 = 0.973,\ s = 0.132,\ q^2 = 0.936 \quad (61)$$

RBR stands for relative biological response. Although response is strongly dependent on Clog *P*, σ^+ accounts for 17% of the variance in the data. In another report, Iwamura[145] collected data from the literature as well as that used in QSAR eq 61 to derive and report QSAR eq 62, where *L* and *W* represents substituent width and length while *A* denotes taste potency.

$$\log A = 0.52(\pm 0.14)L - 1.37(\pm 1.08)W_1 + 3.71(\pm 3.49)$$

$$n = 20,\ r^2 = 0.810,\ s = 0.32 \quad (62)$$

A reexamination of his data results in the development of the following equation

$$\log k = -0.51(\pm 0.28)\sigma^+ + 1.19(\pm 0.24)\text{Clog } P + 0.25(\pm 0.46)$$

$$n = 18,\ r^2 = 0.894,\ s = 0.239,\ q^2 = 0.844 \quad (63)$$

Outliers 3-NO_2, 6-OC_4H_9; 3-NO_2, 6-OCH=CH_2 had to be omitted for lack of a σ^+ value.

The σ^+ term is close to that of eq 61. The above two equations can be compared with QSAR eq 64 for the oxidation of aniline with chloramine-T in ethanol/water.[149a]

$$\log k_2 = -1.41(\pm 0.49)\sigma^+ + 0.72(\pm 0.12)$$

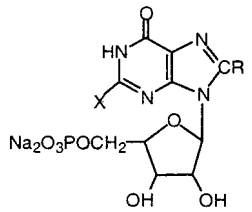
$$n = 6,\ r^2 = 0.941,\ s = 0.107,\ q^2 = 0.870$$

outlier: 2-Cl (64)

The similarity of the σ^+ terms in the three examples makes one wonder if oxidation could possibly be involved in taste. QSAR eq 63 is based on a more complex set of data in that in a number of examples the 4-NO_2 group has been replaced with 4-CN.

Equations 61 and 62 illustrate an important point that we have been concerned with. Although Iwamura was well aware of our work in eq 61, his model only focused on the length and width of substituents and neglected hydrophobic and electronic parameters. The discrepancy in eqs 61 and 62 provides compelling evidence for the importance of lateral validation in the generation of an appropriate QSAR.

Another interesting study on taste came from Mizuta et al., on the flavor-enhancing activity of ribonucleotides.[149b]



$$\log 1/C = 0.51(\pm 0.14)\mathrm{CMR} + 0.71(\pm 0.83)$$

$$n = 12,\ r^2 = 0.873,\ s = 0.102,\ q^2 = 0.824$$

outliers: $SCH_2C_6H_5$; SCH_3; C_6H_5 (65)

Although this equation touts the overall dependence on mostly size and polarizability, it does not clearly delineate which molecular features of this complex compound are crucial for the biological activity.

Turning now to another type of activity, 61 QSAR in the databank focus on studies of cytochrome P450.

Demethylation of $X-C_6H_4N(CH_3)_2$ by isolated P450[146]

$$\log k_{cat}/K_m = 0.53(\pm 0.20)\log P + 3.47(\pm 0.53)$$

$$n = 8,\ r^2 = 0.878,\ s = 0.093,\ q^2 = 0.823$$

outlier: 4-CHO (66)

Microsomal demethylation of miscellaneous compounds[147]

$$\log 1/K_m = 0.70(\pm 0.14)\log P + 2.86(\pm 0.29)$$

$$n = 13,\ r^2 = 0.915,\ s = 0.260,\ q^2 = 0.884$$

outlier: Ephedrine (67)

Dealkylation of $C_6H_5CH(Me)NR_2$ by one person[148]

$$\log k = 0.61(\pm 0.16)\log P - 3.09(\pm 0.51)$$

$$n = 12,\ r^2 = 0.874,\ s = 0.221,\ q^2 = 0.762$$

outliers: *sec*-butyl, benzyl (68)

These examples indicate that dealkylation in the isolated enzyme, in the organelle, and in humans is a very similar process. This is the ideal to strive for in building up a science of chemical−biological interactions.

Another interesting study with humans is that of nonrenal and renal clearance of β-adrenoreceptor antagonists: bufuralol, tolamolol, propanolol, alprenolol, oxprenolol, acebutolol, timolol, metoprolol, prindolol, atenolol, and nadolol.

Non enal clearance of miscellaneous alcohols acting as β-adrenoreceptor antagonists[150]

$$\log K = 1.94(\pm 0.61)\mathrm{Clog}\ P - 2.00(\pm 0.80)\log(\beta \cdot 10^{\mathrm{Clog}\ P} + 1) + 1.29(\pm 0.30)$$

$$n = 10,\ r^2 = 0.950,\ s = 0.168,\ q^2 = 0.918$$

outlier: oxprenolol

$$\mathrm{Clog}\ P_O\text{:}\quad 2.6\ (\pm 1.5),\ \log \beta = -0.813 \qquad (69)$$

Using a parabolic model instead of the bilinear model, one obtains a better defined optimum Clog *P* of 2.5 (2.1−3.2).

Renal clearance of β-adrenoreceptor antagonists[150]

$$\log K = -0.42(\pm 0.12)\mathrm{Clog}\ P + 2.35(\pm 0.24)$$

$$n = 10,\ r^2 = 0.888,\ s = 0.185,\ q^2 = 0.793$$

outliers: acebutolol, pindolol (70)

It is clear that the two processes have different hydrophobic requirements for clearance. A most unusual QSAR is obtained by assessing human kill by miscellaneous drugs.[151]

LD_{100} for humans

$$\log 1/C = 1.17(\pm 0.34)\log P + 1.70(\pm 0.70)$$

$$n = 12,\ r^2 = 0.869,\ s = 0.498,\ q^2 = 0.825 \qquad (71)$$

The data for this QSAR comes from England, where the practice in cases of suicide or accidental overdose of drugs is that the individuals blood is analyzed to determine the concentration of drug. In QSAR eq 71, the concentrations from cases of poisoning were averaged to obtain a single value for each compound. As one might expect, the standard deviation is high. The data pertains to the following chemicals: ethanol, ether, paraldehyde, chlormethiazole, chloroform, phenobarbital, secobarbital, (maprofiline outlier) dothiepin, amitriptyline, propoxyphene, and chlorpromazine. For partially ionized compounds, log *D* was employed, where *D* is the distribution coefficient at ca. pH 7.

The shape of QSAR eq 71 is similar to what has been termed nonspecific toxicity in our earlier discussion. Hundreds of such QSAR are known for all sorts of biological systems. In the early days of biological SAR, it was often assumed that nerve damage was the critical factor in such toxicity. It is now clear that many biological processes show results similar to QSAR eq 71, in which nerves are not involved. Cell membranes may also be inplicated. In any case, it is the most sensitive site in the cell or organism that determines the shape of the QSAR.

The hallucinogenic activity of X−$C_6H_4CH_2CH(R)$-NH_2 in humans[152a,152b]

$$\log \text{RBR} = 1.17(\pm 0.25)\log P - 3.28(\pm 1.0)\log(\beta \cdot 10^{\log P} +) - 0.18(\pm 0.15)\sigma^+ - 1.49(\pm 0.49)$$

$$n = 24,\ r^2 = 0.850,\ s = 0.232,\ q^2 = 0.801,\ \log P_O = 3.24,\ \log \beta = -3.49$$

outliers: X = 2,5-di-OMe, 4-Me, R = Me; X = 2,5-di-OMe, 4-Br, R = Me; X = 2,3-di-OMe-4,5-OCH_2O, R = Me (72)

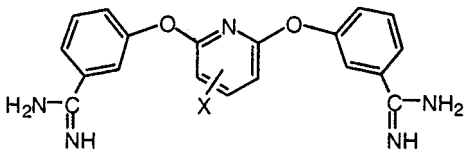
The end point was comparing the potency relative to mescaline in the human subjects. This work was conducted at the University of Chile; it would have been illegal in the United States!

X. Allosteric Interactions

Having such a large collection of information based on QSAR has enabled us to constantly uncover new relationships. We were recently surprised to discover instances where correlation with CMR (or sometimes Clog *P*) gave inverted parabolic QSAR. That is, activity first decreased and then at a certain point turned upward and increased. Obviously a change in mechanism has occurred. This is in stark contrast to many hundreds of examples where biological activity increases to a maximum and then levels or falls off. The inverted curve suggests a change in the configuration of the receptor structure. We have classified this as an allosteric change. The term comes from allostery, a Greek word for another shape.

The following examples illustrate our finding based on CMR.[154]

Inhibition of bovine trypsin by



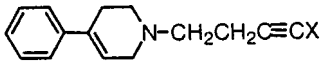
$$\log 1/k_i = -3.02(\pm 1.2)\text{CMR} + 0.14(\pm 0.05)\text{CMR}^2 + 0.46(\pm 0.25)\text{B1}_4 + 21.7(\pm 0.70)$$

$$n = 22,\ r^2 = 0.837,\ s = 0.131,\ q^2 = 0.772$$

outlier: 3-NHCO-gly-NH_2

inversion point: 10.8 (10.2−11.1) (73)

Inhibition of dopamine D_2 receptor from rat striatal membrane by[153,154]



$$\log 1/k_i = -14.2(\pm 8.3)\text{CMR} + 0.72(\pm 0.41)\text{CMR}^2 - 0.47(\pm 0.19)\text{Clog } P + 78.5(\pm 41.7)$$

$$n = 14,\ r^2 = 0.837,\ s = 0.186,\ q^2 = 0.665$$

outlier: 4-HO−C_6H_4-; 2-pyridinyl

inversion point: 9.85 (9.43−10.0) (74)

Inhibition of angiogenesis in mixed mouse lymphocyte cell cultures[155] *by analogues of TNP-470 and ovalicin*[159]

$$\log 1/C = -3.98(\pm 1.46)\text{Clog } P + 0.95(\pm 0.39)\text{Clog } P^2 + 0.92(\pm 0.72)\text{I} + 10.5(\pm 1.5)$$

$$n = 11,\ r^2 = 0.941,\ s = 0.375,\ q^2 = 0.812$$

inversion point: 2.09 (1.92−2.35) (75)

I = 1 for congeners having two epoxide units.

There has been great interest in allosteric interactions since Monod et. al. first introduced the idea.[156,157] Recently, Changeux and Edelstein reviewed the subject.[158]

Note that in the above two examples CMR has an initial negative slope, but at the value of CMR = 10.8 of eq 72 and 9.85 in eq 73, the slope becomes positive. Care must be taken to see that the inversion point is solidly established. A plot of the data and confidence limits on the point of inversion are necessary, otherwise one may have an L-shaped result where the activity first falls and then more or less levels off. As discussed above, CMR does contain a molecular volume component. However, we have observed in 11 published examples that CMR cannot be replaced by a molecular volume term. Thus, it appears that polarizability does play a role in these inverted parabolic relationships.

The first clear understanding of allosteric interactions was elucidated by Monod et al.[156] from the interaction of ligands with hemoglobin. The above three examples are of course for quite different systems. We have recently found evidence on hemoglobin that is related directly to Monod's study.

Rate constants for the binding of isonitriles ($R-N^+{=}C-$) to the alpha subunit of human hemoglobin[159]

$$\log k = -0.77(\pm 0.44)\text{Clog } P + 0.35(\pm 0.23)\text{Clog } P^2 - 1.72(\pm 0.44)\text{B1} + 4.76(\pm 0.78)$$

$$n = 12,\ r^2 = 0.949,\ s = 0.188,\ q^2 = 0.833$$

outlier: R = $CH_2CH(CH_3)_2$

inversion point: 1.11 (0.9−1.7) (76)

The above QSAR shows that hydrophobic properties of the ligand can also induce an allosteric interaction. Many such examples based on CMR or Clog *P* have been uncovered, and a review on the subject is now in progress.[159]

A more interesting example is the following:

Binding of X−$C_6H_4NO_2$ to hemoglobin in Wistar rats[160]

$$\log \mathrm{HBI} = 3.62(\pm 1.4)\sigma^+ - 11.1(\pm 5.64)\mathrm{Clog}\,P + 1.97(\pm 1.0)\mathrm{Clog}\,P^2 + 1.51(\pm 1.0)\mathrm{B1}_4 + 14.2(\pm 7.9)$$

$$n = 14,\ r^2 = 0.874,\ s = 0.507,\ q^2 = 0.743$$

outlier: 2,4-di-F

inversion point: 2.82 (2.61−3.1) (77)

In the above expression, HBI is the hemoglobin binding index (i.e., mmol of compound/mol of HB/mmol of compound/kg of body weight). Although the above equation is not as sharp as one would like and the ratio of data points to variables is rather low, the inversion point is well defined. The sterimol parameter $B1_4$ brings out the presence of a positive steric effect of 4-substituents, and the electronic term σ^+ suggests that the nitro group is reduced to a radical which then binds to hemoglobin and, no doubt, to other targets too.

Normally σ^- is determined to be the best parameter for this process with regard to the nitro moiety; however, in this instance it yields a slightly poorer result ($r^2 = 0.854$). The two parameters σ^- and σ^+ are in the present instance highly collinear ($r^2 = 0.964$). These preliminary results suggest that QSAR can be used to uncover allosteric interactions with hemoglobin, enzymes, or in cells and animals.

A possibility that needs to be considered in such studies is that if the structure of the receptor or enzyme is undergoing a large change, would the points of contact on the down side and the up side change in ways so the electronic properties of the system would be incongruent. At present, a method of searching our system is to isolate all QSAR that have −CMR and +CMR^2 terms or the same for Clog *P*. This can be done in less than a minute.

XI. Conclusions

The above review outlines one informatics approach to developing some understanding about the interface between chemical−chemical and chemical−biological interactions. Certainly it will not be the last effort. We believe that specialized efforts such as this will also be forced to evolve in other areas as the output of information continues to burgeon in all areas of science. *Chemical Abstracts* or online searching of the literature is too nonspecific to provide the necessary structure that is so important for understanding a particular subsection of science. Scheme 1 outlines the makeup of our current system.

The major design problem is to decide on how many levels of searching to provide and how to name these levels for defining and collecting data. Tables 2 and 3 outline our nomenclature that grew as specific needs emerged. The physical database has 23 major classes that seemed to do a fair job; however, we were forced to introduce a miscellaneous class that has slowly grown to almost 500 QSAR. Nevertheless, this area can be rapidly searched in terms of parameters or chemical structures using the SMILES or MERLIN options. At present, one can survey these rather quickly, but as the system continues to grow, new classes may be needed. Even as it stands, it is easy to use for someone having a little background in mechanistic organic chemistry.

The biological database presents the onerous problems. Under enzymes there are so many potential kinds of subclassifications for oxidoreductases, hydrolases, and receptors. Indeed, receptors, the fastest growing class, needs a separate subclassification that must soon be undertaken. No doubt, this will also be true for nucleic acids. At present, we can quickly isolate the 676 QSAR for oxidoreductases and then scan the names in a few minutes to find one of interest that can be downloaded for detailed analysis.

Cells present some ambiguity. At present, we are going to label bacteria as Gram positive or negative. Most cells are clearly named and can be located easily. The sets involving organs and tissues can be viewed for leads, but eventually more subsections will be needed. In the case of whole animals, mice present a minor problem as sometimes they are denoted in the singular form. Searching **2 B6a** and then **1 mouse mice** finds 289 QSAR which when searched individually yields 88 on mice and 201 on mouse.

The most serious drawback to general usage of the database is that of the researchers background. Even chemists have trouble with the meaning of the Hammett parameters, and of course, these are opaque to most biological researchers. Many chemists have limited backgrounds in molecular biology. One also needs experience in building models and some understanding of simple statistics. There are no simple solutions to the problem of understanding chemical−biological interactions.

Various approaches to QSAR tend to minimize the real complexity of mathematically delineating the significant structural features of a set of 'congeners' acting on just a single cell culture, not to mention a mouse or a man. The possibilities for side reaction/interaction are enormous. Recently, Wermuth and Clarence-Smith[161] reviewed some of the well-established multiple targets of known drugs. For example, the antipsychotics clozapine and olanzapine have been shown to bind to at least 14 different receptors. The hope of medicinal chemists is that testing modifications of old drugs can lead to more potent and more selective new drugs. We believe that our system of bioinformatics will be of help in such work. For example, the drug chloramphicol, an excellent antibiotic, had to be withdrawn from the market because of serious side reactions. It was assumed by many that it was the nitro group that was the source of the toxicity. We have shown instead that it is the benzylic moiety that is easily converted to a radical, a reaction well correlated by E_R.[162] This propensity for radical formation (and the basis for a solid mechanistic interpretation of chemical reactions)

could have been lowered by replacing the nitro with a substituent having a lower E_R value. However, the nitro group can also readily undergo a radical reduction. Today, the incorporation of a nitro group into a prospective drug target would be frowned upon. However, there would be little concern about using a hydroxymethyl group attached to an aromatic ring, which can be made more biologically susceptible to a radical reaction by conjugation with a substituent having a large E_R value. Recently we were informed by a researcher at a leading drug company that management has suggested that it is not a good idea to incorporate an aromatic OH function in a prospective drug. Again, we have shown that it is a matter of what the OH is conjugated with.[27] Electron-releasing groups increase the propensity for radical formation, but electron-attracting groups inhibit such a reaction.[27] This kind of information can be gathered from simple biological systems early on in a research project. Once a drug goes to market, it is very difficult to detect certain types of radical toxicity. Such toxicity could result in cancer after many years of use. As we have noted, computational chemistry for drug design has been making rapid strides in the last 10 years. It is not unusual for companies to have 50 or more computerized programs for such work. However, the problems are daunting. One can quickly learn to punch in the numbers, but careful evaluation of the output warrants extensive experience. It is of critical importance that we utilize the enormous amount of work that has laid the basis for a sound mechanistic interpretation of chemical reactions. Phenol is not mutagenic or carcinogenic, but 4-methoxyphenol is carcinogenic to rodents. Gradually an expert system of chemical−biological informatics will educate us about the complexity of drug interactions.

A word needs to be said about the Hammett parameters. They is the achievement of over 60 years of study by thousands of chemists. These results are invaluable in studying how chemicals react with each other, and the results can readily be compared with the enormous number of studies on many, many types of reactions. Quantum chemistry offers no such possibilities yet, although it may sometime in the distant future. In the final analysis, comparative QSAR, regardless of how it is attained, is the only guide in the evolution of our understanding of how chemicals affect living systems or their parts.

XII. Acknowledgments

The following individuals derived and loaded into our system the indicated number of QSAR over the past 40 years: Akamatsu, M. (101); Allister, D. (15); Arms, P. (3); Briggs, M. (252); Calef, D. F. (27); Clayton, D. F. (27); Coats, E. A. (5); Coubeils, J. L. (92); Debnath, A. K. (123); Dixon, J. (2); Dunn, W. J. (47); Dull, G. (1); Engle, R. (7); Fukunaga, J. Y. (2); Fujita, T. (6); Garg, R. (1993); Gao, H. (4190); Ghose, A. (1); Glave, W. R. (133); Good, P. (2); Grieco, C. (5); Hadjipavlou-Litina, D. (19); Hansch, C. (6213); Hatheway, G. J. (8); Hinshaw, M. (19); Hoekman, D. (1); Jon (2); Kapur, S. (3); Kiehs, K. (2); Kurup, A. (1661); Leo, A. (37); Li, R. (26); Lien, E. J. (80); Mekapati, S. B. (1331); McFarland, J. (1); Musallan, M. (1); Munson, R. (8); Nikaitani, D. (2); Panthananickal, A. (12); Li, P. (415); Portoghese, P. S. (1); Quin, F. (2); Recanatini, M. (20); Schaeffer, H. J. (56); Schmidt (3); Silipo, C. (9); Unger, S. (6); Van der Aa, E. (165); Verhaar, H. J. M. (8); Venger, B. H. (1); Verma, R. P. (11); Ware (2); Wilcox, A. (2); Win Yu (3); Yamakawa, M. (3); Ye, S. (13); Yoshimoto, M. (1); Zhang, L. (16).

A special mention must be made of Peng Li, who entered SMILES for several thousand QSAR that were derived before the advent of SMILES. Also, Litai Zhang and Michael Medlin did extensive checking of entered data.

Our computer program, including all of the data, can be obtained from BioByte Corporation: 201 West 4th Street, Suite 204, Claremont, California 91711. All of the QSAR can be inspected on our website: www.biobyte.com.

XIII. References

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178.
(2) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(3) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.
(4) Hansch, C.; Hoekman, D.; Gao, H. *Chem. Rev.* **1996**, *96*, 1045.
(5) Hansch, C. *Acc. Chem. Res.* **1993**, *26*, 147.
(6) Hansch, C.; Gao, H. *Chem. Rev.* **1997**, *97*, 2995.
(7) Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S.; Debnath, A. K.; Hansch, C. *Chem. Rev.* **1999**, *99*, 3525.
(8) Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. *Chem. Rev.* **1999**, *99*, 723.
(9) (a) Hansch, C.; Kurup, A.; Garg, R.; Gao, H. *Chem. Rev.* **2001**, *101*, 619. (b) Hansch, C. In *Classical and Three-Dimensional QSAR in Agrochemistry*; Hansch, C., Fujita, T., Eds.; ACS Symposium Series 606; American Chemical Society: Washington, DC, 1995; p 254.
(10) Hansch, C.; Li, R. L.; Blaney, J. M.; Langridge, R. *J. Med. Chem.* **1982**, *25*, 777.
(11) Hansch, C.; Klein, T. *Acc. Chem. Res.* **1986**, *19*, 392.
(12) Blaney, J. M.; Hansch, C. *Comprehensive Medicinal Chemistry*; Pergamon Press: Elmsford, NY, 1990; p 459.
(13) In *3-D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/Escom: Norwell, MA, 1998; Vols. 3 and 4.
(14) *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol. 11.
(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studios Press: 1986.
(16) Kier, L. B.; Hall, L. H. *Molecular Structure Descriptors*; Academic Press: New York, 1999.
(17) Cramer, R. D., III; Patterson, D. E.; Bunce *J. Am. Chem. Soc.* **1988**, *110*, 5959.
(18) (a) Elkins, D.; Leo, A.; Hansch, C. *J. Chem. Doc.* **1974**, *14*, 65. (b) Leo, A.; Elkins, D.; Hansch, C. *J. Chem. Doc.* **1974**, *14*, 61. (c) Hansch, C.; Leo, A.; Elkins, D. *J. Chem. Doc.* **1974**, *14*, 57.
(19) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31. (a) Selassie, C. D.; DeSoyza, T. V.; Rosario, M.; Gao, H.; Hansch, C. *Chem.-Biol. Interact.* **1998**, *113*, 175.
(20) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
(21) Weininger, D.; Weininger, J. L. *Comprehensive Medicinal Chemistry*; Pergamon Press: Elmsford, NY; Vol. 4, Chapter 17.3, p 59.
(22) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165.
(23) Debnath, A. K.; Hansch, C. *Environ. Mol. Mutagen.* **1992**, *20*, 140.
(24) Pritykin, L. M.; Selyutin, O. B. *Russ. J. Org. Chem.* **1969**, *34*, 1143.
(25) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.
(26) Zhang, L.; Gao, H.; Hansch, C.; Selassie, C. D. *J. Chem. Soc., Perkin Trans. 2* **1998**, 2553.
(27) Selassie, C. D.; Shusterman, A. J.; Kapur, S.; Verma, R. P.; Zhang, L.; Hansch, C. *J. Chem. Soc., Perkin Trans. 2* **1999**, 2729.
(28) Debnath, A. K.; de Compadre, R. L. L.; Shusterman, A. J.; Hansch, C. *Environ. Mol. Mutagen.* **1992**, *19*, 53.

(29) Cnubben, N. H. P.; Peelen, S.; Borst, J.-W.; Vervoort, J.; Veeger, C.; Rietjens, I. M. C. M. *Chem. Res. Toxicol.* **1994**, *7*, 590.
(30) You, Z.; Brezzell, M. D.; Das, S. K.; Espadas-Torre, M. C.; Hooberman, B. H.; Sinsheimer, J. E. *Mutat. Res.* **1993**, *319*, 19.
(31) Snyder, S. H.; Merril, C. R. *Proc. Nat. Acad. Sci. U.S.A.* **1965**, *54*, 258.
(32) Debnath, A. K.; Hansch, C. *Environ. Mol. Mutagen.* **1992**, *20*, 140.
(33) Zoete, V.; Bailly, F.; Maglia, F.; Rougee, M.; Bensasson, R. V. *Free Radical Biol. Med.* **1999**, *26*, 1261.
(34) Wald, R. W.; Feuer, G. *J. Med. Chem.* **1971**, *14*, 1081.
(35) Tuppurainen, K. *J. Mol. Struct. (THEOCHEM)* **1994**, *112*, 49.
(36) Kato, S.; Kawasaki, T.; Urata, T.; Mochizuki, J. *J. Antibiot.* **1993**, *46*, 1859.
(37) Xu, S.; Li, L.; Tan, Y.; Feng, J.; Wei, Z.; Wang, L. *Bull. Environ. Contam. Toxicol.* **2000**, *64*, 316.
(38) Sami, S. M.; Iyengar, B. S.; Tarnow, S. E.; Remers, W. A.; Bradner, W. T.; Schurig, J. E. *J. Med. Chem.* **1984**, *27*, 701.
(39) Shusterman, A. J.; Johnson, A. S.; Hansch, C. *Int. J. Quantum Chem.* **1989**, *36*, 19.
(40) Shusterman, A. J.; Debnath, A. K.; Hansch, C.; Horn, G. W.; Fronczek, F. R.; Greene, A. C.; Watkins, S. F. *Mol. Pharm.* **1989**, *36*, 939.
(41) Taskinen, J.; Vidgren, J.; Ovaska, M.; Baeckstroem, R.; Pippuri, A.; Nissinen, E. *Quant. Struct.-Act. Relat.* **1989**, *8*, 210.
(42) Schultz, T. W.; Sinks, G. D.; Hunter, R. S. *SAR QSAR Environ. Res.* **1995**, *3*, 27−36.
(43) Tyrakowska, B.; Cnubben, N. H. P.; Soffers, A. E. M. F.; Wobbes, T.; Rietjens, I. M. C. M. *Chem.-Biol. Interact.* **1996**, *100*, 187.
(44) Tuppurainen, K. *Chemosphere* **1999**, *38*, 3015.
(45) Cnubben, N. H. P.; Soffers, A. E. M. F.; Peters, M. A. W.; Vervoort, J.; Rietjens, I. M. C. M. *Toxicol. Appl. Pharmacol.* **1996**, *139*, 71.
(46) Tuppurainen, K.; Lotjonen, S. *Mutat. Res.* **1993**, *287*, 235.
(47) Tuppurainen, K.; Lotjonen, S.; Laatikainen, R.; Vartiainen, T. *Mutat. Res.* **1992**, *266*, 181.
(48) Crebelli, R.; Andreoli, C.; Carere, A.; Conti, G.; Conti, L.; Ramusino, C. M.; Benigni, R. *Mutat. Res.* **1992**, *266*, 117.
(49) Tuppurainen, K.; Lotjonen, S.; Laatikainen, R.; Vartiainen, T.; Maran, U.; Strandberg, M.; Tamm, T. *Mutat. Res.* **1991**, *247*, 97.
(50) Veith, G. D.; Mekenyan, O. G. *Quant. Struct.-Act. Relat.* **1993**, *12*, 349.
(51) Dimoglo, A. S.; Chumakov, Y. M.; Dobrova, B. N.; Saracoglu, M. *Arzneim.-Forsch./Drug Res.* **1997**, *47*, 415.
(52) Lewis, D. F. V.; Brantom, P. G.; Ioannides, C.; Walker, R.; Parke, D. V. *Drug Metab. Rev.* **1997**, *29*, 1055.
(53) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. *Environ. Toxicol. Chem.* **1998**, *17*, 15.
(54) Tollenaere, J. P. *Chim. Ther.* **1971**, *6*, 88.
(55) Anusevicius, Z.; Soffers, A. E. M. F.; Cenas, N.; Sarlaukas, J.; Segura-Aguilar, J.; Rietjens, I. M. C. M. *FEBS Lett.* **1998**, *427*, 325.
(56) deCompadre, R. L. L.; Debnath, A. K.; Shusterman, A. L.; Hansch, C. *Environ. Mol. Mutatgen.* **1990**, *15*, 44.
(57) Ridder, L.; Briganti, F.; Boersma, M. G.; Boeren, S.; Vis, E. H.; Scozzafava, A.; Veeger, C.; Rietjens, I. M. C. M. *Eur. J. Biochem.* **1998**, *257*, 92.
(58) Oikawa, S.; Tsuda, M.; Endou, K.; Abe, H.; Matsuoka, M.; Nakajima, Y. *Chem. Pharm. Bull.* **1985**, *33*, 2821.
(59) Klimesova, V.; Palat, K.; Waisser, K.; Klimes, J. *Int. J. Pharm.* **2000**, *207*, 1.
(60) Schultz, T. W.; Cronin, M. T. D. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 304.
(61) Yuan, X.; Lu, G.; Lang, L. P. *Bull. Environ. Contam. Toxicol.* **1997**, *58*, 123.
(62) Habicht, J.; Brune, K. *J. Pharm. Pharmacol.* **1983**, *35*, 718. Zoete, V.; Bailly, F.; Maglia, F.; Rougee, M.; Bensasson, R. V. *Free Radical Biol. Med.* **1999**, *26* 1261.
(63) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775.
(64) Van Haandel, M. J. H.; Claassens, M. M. J.; Van der Hout, N.; Boersma, M. G.; Vervoort, J.; Rietjens, I. M. C. M. *Biochim. Biophys. Acta* **1999**, *1435*, 22.
(65) Brown, D.; Woodcock, D. *Pestic. Sci.* **1975**, *6*, 371.
(66) Schmitt, H.; Altenburger, R.; Jastroff, B.; Schüürmann, G. *Chem. Res. Toxicol.* **2000**, *13*, 441.
(67) Sinha, S.; Bano, S.; Agrawal, V. K.; Khadikar, P. V. *Oxid. Commun.* **1999**, *22*, 479.
(68) Zhang, L.; Gao, H.; Hansch, C.; Selassie, C. D. *J. Chem. Soc., Perkin Trans. 2* **1998**, 2553.
(69) Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175.
(70) Hansch, C.; Unger, S. H.; Forsythe, A. B. *J. Med. Chem.* **1973**, *16*, 1217.
(71) Leo, A. *Chem. Rev.* **1993**, *93*, 1281.
(72) Leo, A.; Hansch, C. *Perspect. Drug Discovery Des.* **1999**, *17*, 1.
(73) Leo, A. Unpublished results.
(74) Unger, S.; Hansch, C. *Prog. Phys. Org. Chem.* **1976**, *12*, 91.
(75) Verloop, A.; Tipker, J. In *Drug Design and Toxicology*; Hadzi, D., Jorman-Blazic, B., Eds.; Elsevier: New York, 1987.
(76) Pauling, L.; Pressman, D. *J. Am. Chem. Soc.* **1945**, *67*, 103.
(77) Agin, D.; Herch, L.; Holtzman, D. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *67*, 103.
(78) Ingold, C. K. *Structure and Mechanism in Organic Chemistry*, 2nd ed; Cornell University Press: Ithaca, NY, 1969; p 293.
(79) Hansch, C.; Garg, R.; Kurup, A. *Bioorg. Med. Chem.* **2001**, *9*, 283.
(80) Rice-Evans, C. A.; Packer, L. *Flavanoids in Health and Disease*; Marcel Dekker: New York, 1998.
(81) Yamamoto, Y.; Otsu, T. *Chem. Ind.* **1967**, 787.
(82) Dust, J. M.; Aronald, D. R. *J. Am. Chem. Soc.* **1983**, *105*, 1221.
(83) Jiang, X.-K.; Ji, G. Z. *J. Org. Chem.* **1992**, *57*, 6051.
(84) Creary, X.; Mehrsheikh-Mohammadi, M. E.; McDonald, S. *J. Org. Chem.* **1987**, *52*, 3254.
(85) Jaffé, H. H. *Chem. Rev.* **1953**, *53*, 191.
(86) *Chem. Rev.* **2000**, *100*, 1.
(87) Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 525.
(88) *Advances in Linear Free Energy Relationships*; Chapman, N. B., Shorter, J., Eds.; Plenum Press: New York, 1972.
(89) *Correlation Analysis in Chemistry*; Chapman, N. B., Shorter, J., Eds.; Plenum Press: New York, 1978.
(90) Lee, I.; Choi, Y. H.; Lee, H. W.; Lee, B. C. *J. Chem. Soc., Perkin Trans. 2* **1988**, 1537.
(91) Hansch, C.; Kim, D.; Leo, A. J.; Novellino, E.; Silipo, C.; Vittoria, A. *Crit. Rev. Toxicol.* **1989**, *19*, 185.
(92) Phillips, W. E.; Rejda-Heath, J. M. *Pestic. Sci.* **1993**, *38*, 1.
(93) Nakamura, S.; Wakusawa, S.; Tajima, K.; Miyamoto, K.-I.; Hagiwara, M.; Hidaka, H. *J. Pharm. Pharmacol.* **1993**, *45*, 268.
(94) (a) Hsuanyu, Y.; Dunford, H. B. *J. Biol. Chem.* **1992**, *267*, 17649. (b) Dewhirst, F. E. *Prostaglandins* **1980**, *20*, 209.
(95) Riddle, B.; Jencks, W. P. *J. Biol. Chem.* **1971**, *246*, 3250.
(96) Feng, L.; Wang, L.-S.; Zhao, Y.-H.; Song, B. *Chemosphere* **1996**, *32*, 1575.
(97) Chong, S.; Fung, H.-L. *Biochem. Pharmacol.* **1991**, *42*, 1433.
(98) Shüürmann, G.; Somashekar, R. K.; Kristen, U. *Environ. Toxicol. Chem.* **1996**, *15*, 1702.
(99) Hansch, C.; Björkroth, J. P.; Leo, A. *J. Pharm. Sci.* **1987**, *76*, 663.
(100) Fujita, T. In *Drug Design: Fact or Fantasy?*; Jolles, G., Woolridge, K. R. H., Eds.; Academic Press: New York, 1984; p 18.
(101) D'Yakonov, I. A.; Kostikov, R. R.; Aksenov, V. S. *Organic Reactivity* **1970**, *7*, 248EE.
(102) Alzeer, J.; Chollet, J.; Heinze-Krauss, I.; Hubschwerlen, C.; Matile, H.; Ridley, R. G. *J. Med. Chem.* **2000**, *43*, 560.
(103) Newmann, T. B.; Hulley, S. B. *J. Am. Med. Assoc.* **1996**, *275*, 55.
(104) Mukai, K.; Yokoyama, S.; Fukuda, K.; Uemoto, Y. *Bull. Chem. Soc. Jpn.* **1987**, *60*, 2163.
(105) Meloche, I.; Laidler, K. J. *J. Am. Chem. Soc.* **1951**, *73*, 1712.
(106) (a) Mekapati, S. B.; Kurup, A.; Garg, R.; Hansch, C. Unpublished results from data taken from (b) Jagannadnam, V.; Steeken, S. *J. Am. Chem. Soc.* **1984**, *106*, 6542. (c) Jagannadhnam, V.; Steeken, S. *J. Phys. Chem.* **1988**, *92*, 111. (d) Tatsumi, K.; Yoshimura, H.; Kawazoe, Y. *Chem. Pharm. Bull.* **1978**, *26*, 1713. (e) Zhao, Y.-H.; Wang, L.-S.; Gao, H.; Zhang, Z. *Chemosphere* **1993**, *26*, 1971. (f) Zhao, Y.-H.; He, Y.-B.; Wang, L. S. *Toxicol. Environ. Chem.* **1995**, *51*, 191. (g) Muller, G. W.; Corral, L.; Shire, M. G.; Wang, H.; Moreira, A.; Kaplan, G.; Stirling, D. *J. Med. Chem.* **1996**, *39*, 3238. (h) Chan, C. L.; Lien, E. J.; Tokes, Z. A. *J. Med. Chem.* **1987**, *30*, 509. (i) Ito, O.; Matsuda, M. *J. Am. Chem. Soc.* **1982**, *104*, 1701.
(107) Sawaki, Y.; Ogata, Y. *J. Org. Chem.* **1984**, *49*, 3344.
(108) Sakurai, H.; Hayashi, S.; Hosomi, A. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 1945.
(109) Kostikov, R. R.; Molchanov, A. P.; Ogloblin, K. A. *Zh. Org. Khim.* **1973**, *9*, 2473EE.
(110) Cadogan, J. I. G.; Sadler, I. H. *J. Chem. Soc. (B)* **1966**, 1191.
(111) Yamamoto, K.; Kato, S.; Mizutani, T.; Irie, Y. *Res. Commun. Pharm. Toxicol.* **1996**, *1*, 211.
(112) Ogata, M.; Matsumoto, H.; Takahashi, K.; Shimizu, S.; Kida, S.; Ueda, M.; Kimoto, S.; Haruna, M. *J. Med. Chem.* **1984**, *27*, 1142.
(113) Edwards, M. L.; Stemerick, D. M.; Sunkara, P. S. *J. Med. Chem.* **1990**, *33*, 1948.
(114) Dauphin, G.; Kergomard, A. *Bull. Soc. Chim. Fr.* **1961**, 468.
(115) Lien, E. J.; Hussain, M.; Tong, G. L. *J. Pharm. Sci.* **1970**, *59*, 865.
(116) Kakeya, N.; Yata, N.; Kamada, A.; Aoki, M. *Chem. Pharm. Bull.* **1970**, *18*, 191.
(117) Keasling, H. H.; Schumann, E. L.; Veldkamp, W. *J. Med. Chem.* **1965**, *8*, 548.
(118) Lucarini, M.; Pedrielli, P.; Pedulli, G. F.; Cabiddu, S.; Fattuoni, C. *J. Org. Chem.* **1996**, *61*, 9259.
(119) Temellini, A.; Franchi, M.; Giuliani, L.; Pacifici, G. M. *Xenobiotica* **1991**, *21*, 171.
(120) Mangold, J. B.; McCann, D.; Spina, A. *Biochim. Biophys. Acta* **1993**, *217*, 1163.

(121) Pilyugin, V. S.; Vasin, S. V.; Maslova, T. A. *Zh. Ohshch. Khim.* **1981**, *51*, 1238 EE.
(122) Fujita, T.; Kamoshita, K.; Nishioka, T.; Nakajima, M. *Agr. Biol. Chem.* **1974**, *38*, 1521.
(123) Fujita, T.; Nishioka, T. *Prog. Phys. Org. Chem.* **1976**, *12*, 49.
(124) Tollenaere, J. P. *Comp. Biochem. Parasite Relat. Proc. Int. Symp. 2nd* **1976**, 629.
(125) Roberts, D. D. *J. Org. Chem.* **1964**, *29*, 2714.
(126) Bowden, K.; Chapman, N. B.; Shorter, J. *J. Chem. Soc.* **1964**, 3370.
(127) Fones, W. S.; Lee, M. *J. Biol. Chem.* **1953**, *201*, 847.
(128) Hansch, C.; Grieco, C.; Silipo, C.; Vittoria, A. *J. Med. Chem.* **1977**, *20*, 1420.
(129) Tuppurainen, K.; Lötjönen, S. *Mutat. Res.* **1993**, *287*, 235.
(130) Charton, M.; Charton, B. I. *J. Org. Chem.* **1969**, *34*, 1871.
(131) Job, D.; Dunford, H. B. *Eur. J. Biochem.* **1976**, *66*, 607.
(132) Hogg, J. S.; Lohmann, D. H.; Russell, K. E. *Can. J. Chem.* **1961**, *39*, 1588.
(133) Stone, A. T. *Environ. Sci. Technol.* **1987**, *21*, 979.
(134) Xu, F. *Biochemistry* **1996**, *35*, 7608.
(135) Dewhirst, F. E. *Prostaglandins* **1980**, *20*, 209.
(136) Behrman, E. J. *J. Am. Chem. Soc.* **1963**, *85*, 3478.
(137) Mitchell, G.; Clarke, E. D.; Ridley, S. M.; Greenhow, D. J.; Gillen, K. J.; Vohra, S. K.; Wardman, P. *Pestic. Sci.* **1995**, *44*, 49.
(138) Gharagozloo, P.; Lazareno, S.; Popham, A.; Birdsall, N. J. M. *J. Med. Chem.* **1999**, *42*, 438.
(139) Frenna, V.; Macaluso, G.; Consiglio, G.; Cosimelli, B.; Spinelli, D. *Tetrahedron* **1999**, *55*, 12885.
(140) Bordwell, F. G.; Zhang, X.-M. *J. Phys. Org. Chem.* **1995**, *8*, 529.
(141) (a) Mai, A.; Artico, M.; Sbardella, G.; Massa, S.; Novellino, E.; Greco, G.; Loi, A. G.; Tramontano, E.; Marongiu, M. E.; La Colla, P. *J. Med. Chem.* **1999**, *42*, 619. (b) Fujita, T.; Takayama, C.; Nakajima, M. *J. Org. Chem.* **1973**, *38*, 1623.
(142) Hagmann, W. K.; Caldwell, C. G.; Chen, P.; Durette, P. L.; Esser, C. K.; Lanza, T. J.; Kopka, I. E.; Guthikonda, R.; Shah, S. K.; MacCoss, M.; Chabin, R. M.; Fletcher, D.; Grant, S. K.; Green, B. G.; Humes, J. L.; Kelly, T. M.; Luell, S.; Meurer, R.; Moore, V.; Pacholok, S. G.; Pavia, T.; Williams, H. R.; Wong, K. K. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1975.
(143) Kirksey, C. H.; Hambright, P. *Inorg. Chem.* **1970**, *9*, 958.
(144) Deutsch, E. W.; Hansch, C. *Nature* **1966**, *211*, 75.
(145) (a) Iwamura, H. *J. Med. Chem.* **1980**, *23*, 308. (b) Radhakrishnamurti, P. S.; Rao, M. D. P. *Indian J. Chem.* **1976**, *14B*, 790.
(146) Macdonald, T. L.; Gutheim, W. G.; Martin, R. B.; Guengerich, F. P. *Biochemistry* **1989**, *28*, 2071.
(147) Martin, Y. C.; Hansch, C. *J. Med. Chem.* **1971**, *14*, 777.
(148) Donike, V. M.; Iffland, R.; Jaenicke, L. *Arzneim.-Forsch.* **1974**, *24*, 556.
(149) (a) Radhakrishnamurti, P. S.; Padhi, S. C. *Indian J. Chem.* **1978**, *16A*, 541. (b) Mizuta, E.; Toda, J.; Suzuki, N.; Sugibayashi, H.; Imai, K.-I.; Nishikawa, M. *Chem. Pharm. Bull.* **1972**, *20*, 1114.
(150) Hinderling, P. H.; Schmidlin, O.; Seydel, J. K. *J. Pharmacokinet. Biopharm.* **1984**, *12*, 263.
(151) King, L. A. *Human Toxicol.* **1985**, *4*, 273.
(152) (a) Shulgin, A. T.; Sargent, T.; Naranjo, C. *Nature* **1969**, *221*, 537. (b) Shulgin, A.; Shulgin, A. *PIHKAL*; Transform Press: Berkeley, LA, 1991.
(153) Glase, S. A.; Akunne, H. C.; Heffner, T. G.; Jaen, J. C.; Mackenzie, R. G.; Meltzer, L. T.; Pugsley, T. A.; Smith, S. J.; Wise, L. D. *J. Med. Chem.* **1996**, *39*, 3179.
(154) Hansch, C.; Garg, R.; Kurup, A. *Bioorg. Med. Chem.* **2001**, *9*, 283.
(155) Turk, B. E.; Su, Z.; Liu, J. O. *Bioorg. Med. Chem.* **1998**, *6*, 1163.
(156) Monod, J.; Wyman, J.; Changeux, J.-P. *J. Mol. Biol.* **1965**, *12*, 88.
(157) Koshland, D. E.; Nemethy, G.; Filmer, D. *Biochemistry* **1966**, *5*, 365.
(158) Changeux, J.-P.; Edelstein, S. J. *Neuron* **1998**, *21*, 959.
(159) Garg, R.; Kurup, A.; Mekapati, S. B.; Leo, A.; Hansch, C. Submitted for publication.
(160) Sabbioni, G. *Chem. Res. Toxicol.* **1994**, *7*, 267.
(161) Wermuth, C. G.; Clarence-Smith, K. *Pharm. News* **2000**, *7*, 53.
(162) Hansch, C.; Garg, R. *J. Chem. Soc., Perkin Trans 2* **2001**, 476.

CR0102009