

## C-QSAR

# A General Approach to the Organization of Quantitative Structure-Activity Relationships in Chemistry and Biology

Winter, 2000

*BioByte Corp.*

*201 West 4th St, Suite 204*

*Claremont, CA 91711*

*Phone number: (909) 624-5992*

*FAX Number: (909) 624-1398*

*E-Mail: [clogp@biobyte.com](mailto:clogp@biobyte.com)*

## **Preface to the 4<sup>th</sup> Edition of C-QSAR**

The C-QSAR database is growing nicely. Soon we will have three full-time researchers entering data instead of two and the pace will increase to the point where we should be able to keep up with the current rate of publication. We now have over 14,000 QSAR. The biggest change in this edition is in section VI: Searching For New Leads.

CH

## Table of Contents

I. Introduction.....	5
II. Organization of QSAR Database.....	15
III. Using the Databases.....	18
A. Help .....	18
B. String Searching.....	19
C. Physical Database.....	20
1. Main Menu.....	20
2. Searching and Show.....	20
3. Browsing.....	29
4. Statistics.....	29
5. Omitted Data Points.....	30
6. SMILES.....	31
7. Loading from 'Database Search' to 'Workspace'.....	32
D. Biological Database.....	33
1. Browsing.....	33
2. Searching.....	34
3. Comparing New QSAR .....	36
4. Log P <sub>o</sub> (Optimal Log P).....	38
IV. Searching the Parameter (THOR-Sigma) Database.....	40
A. Substituent Constants.....	40
B. Log P and MR .....	41
V. Regression Analysis: Example 1.....	42
A. Title Information.....	43
B. Naming Parameters.....	43
C. Naming and Entering Substituents .....	43
D. Entering Structures via SMILES.....	44

E. Auto-Loading of Parameters.....	45
F. Permuting.....	47
G. Jackknifing.....	50
H. Cross Validation.....	53
I. Editing.....	54
J. Regression Analysis: Example 2.....	54
K. Substituent Selection in Molecular Design.....	58
VI. Searching for New Leads.....	60
VII. SMILES Tutorial.....	66
VIII. Substructure Searching.....	70
IX. Searching Combined Databases.....	72
X. Caveats .....	76
XI. Appendix: Parameter Definition.....	81
XII. Notes on the Use of Substituent Parameters.....	85
XIII. QSAR and Combinatorial Synthesis.....	87
XII. References .....	89

## I. Introduction

The reader is cautioned that the subject embraced by this manual is not something to leaf through on a rainy Sunday afternoon. Over 60 years ago, one of the authors was told by his chemistry professor that organic chemistry was too complex for the human mind. It has become vastly more complex since then! Of course, it is quite simple compared to biology. We are taking some beginning steps to mechanistically merge the two subjects. Anyone who licenses the system will be given several days free instruction in Claremont plus assistance later via telephone.

Structure-activity relationships have been at the heart of chemistry since the work of Mendeleev in the 19th century. However, it was the work of Hammett in the mid 1930's that opened the way for QSAR in organic and eventually in biological chemistry. Since that time an enormous amount of work has been done. Our current system contains over 14,000 sets of data and the corresponding QSAR. Of these 6,300 are biological QSAR and 7,700 are from physical organic chemistry. These come from the many, many different sub-disciplines of physical organic and biological chemistries. An important purpose of this system is to facilitate comparative QSAR from as many points of view as possible.<sup>1-10</sup>

The purpose of our dynamic, integrated electronic system is much more than a compilation of QSAR from all areas of chemistry and chemical biology. We view it as laying the foundation for a science of chemical-biological interactions; that is, a science which undertakes to understand how any (or all) chemicals affect all living systems or parts thereof (see Table 1). Obviously, such a science will be decades in maturing, but we have reached the point where the subject can be taken seriously. In so far as possible we are relating the QSAR of classes in Table 2 to those of Table 3. That is, by *science* we mean mathematical relationships that can be compared with each other in many ways. Obviously qualitative knowledge has been accumulating for some time.

The field of QSAR has become so large that it is no longer possible to write a review of say, 14,000 equations plus the attendant chemical structures and data. To make matters worse, the results of quantitative chemical and biological studies appear in an incredible number of journals.

These studies are very poorly indexed by Chemical Abstracts. Without a continually updated electronic system, it is impossible to profit from past experience.

A truly major problem in advancing the science of QSAR is the split between those hoping to formulate QSAR (by whatever method) and those synthesizing variations in a lead compound. Unless *good* variation in the major properties (steric, electronic, hydrophobic) is achieved, it is of course impossible to establish the importance (or lack of it) for any given property. It is clear from the perusal of the *Journal of Medicinal Chemistry*, for example, that most papers give little or no thought to this important problem. Along with this is the matter of collinearity. If, in a set of derivatives, two parameters are almost parallel no decision can be made as to which is most important. The above problems need to be considered early on in a synthetic program. An important feature that has been recently added to our system is that of substituent selection (section V-K). Here it is shown how to select a set of substituents for the early phases of lead modification that will have all of the common parameters for all members of the set with the widest choice of substituent properties. Also at the start of the study collinearity can be minimized.

We started to seriously collect data about 1965. In 1970, David Elkins wrote our first program for searching the data, but it was cumbersome to use since structures were encoded in the Wiswesser Line Notation and the need to use IBM cards made it slow and inconvenient. Thus the current C-QSAR program is the outgrowth of over 35 years of research. It was designed and written by David Hoekman, and it employs the widely-used SMILES notation (invented by David Weininger) for structure entry. It also uses his Merlin searching program. Table 1 outlines the general organization of the system and tables 2 and 3 show the areas from which the QSAR come.

As of the present, QSAR development is booming with so many new ideas, often based on complex calculations, that it is virtually impossible for anyone to make critical evaluations of their relative merits. An important aspect of our data is that it provides many examples from all sorts of biological and mechanistic organic chemistry that can be employed for comparative purposes. Moreover, the QSAR are formed from traditional parameters whose intended meaning is easily understood.

C-QSAR does have merit for lead generation (section VI). The majority of Bio QSAR in our current system are from articles where the authors made no attempt to derive a QSAR of any sort. No doubt in many instances activities could be increased by means of QSAR. Possibly the best example of this approach is Koga's demonstration that the rather mediocre drug, nalidixic acid, could be transformed with the help of QSAR to what has become the fabulous quinolone carboxylates.<sup>11</sup> Other examples of QSAR success in the development of products of industrial importance, have been discussed by Fujita.<sup>12, 13</sup> Another way of searching for leads is to look for sets with large  $\log 1/C$  values.  $C$  is the molar concentration producing a standard biological effect. At the moment, we have 704 sets that contain compounds with  $\log 1/C > 8$  (*i.e.*,  $10^{-8}$  molar) and 1,327 with  $\log 1/C > 7$ .

At present there is nothing comparable to our large and growing database to guide the development of the science of chemical-biological interactions. Single isolated QSAR standing alone must be accepted with caution. It is only as they can be laterally related in a variety of ways to other QSAR that we can begin to place some confidence in their informational content. The subject will have to develop the way organic chemistry did. Someone discovers a new reaction and then many chemists working over the years define the limits of the reaction.

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*
11.	SUBSTITUENTS	labels of substituents
12.	SMILES	topological description of compounds
13.	DATA	table of parameter values
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

\* examined, even if not used in final equation.



Table 2  
Class Codes--Biological Database  
{Number of sets in parentheses}\*

<b>BO</b>	<b>Unknown</b>	<b>B4</b>	<b>Single-Celled Organisms</b>
		B4A	Algae (37)
<b>B1</b>	Nonenzymatic <b>Macromolecules</b> (DNA, Fibrin, Hemoglobin, Soil, Albumin, etc.) (220)	B4B	Bacteria (668)
		B4C	Cells in culture (568)
		B4E	Erythrocytes (80)
		B4F	Fungi, Molds (240)
		B4P	Protozoa (93)
<b>B2</b>	<b>Enzymes</b>	B4V	Viruses (145)
B2A	Oxidoreductases (639)	B4Y	Yeasts (47)
B2B	Transferases (155)		
B2C	Hydrolases (588)	<b>B5</b>	<b>Organs/Tissues</b>
B2D	Lyases (36)	B5C	Cancer (90)
B2E	Isomerases (11)	B5G	Gastro-intestinal tract (67)
B2F	Ligases (3)	B5H	Heart (83)
B2G	Receptors (884)	B5I	Internal/soft organs (62)
		B5N	Nerves, Brain, Muscles (326)
<b>B3</b>	<b>Organelles</b>	B5S	Skin (50)
B3A	Mitochondria (87)	B5L	Liver (17)
B3B	Microsomes (97)		
B3C	Chloroplasts (82)	<b>B6</b>	<b>Multi-Cellular Organisms</b>
B3M	Membranes (92)	B6A	Animal (vertebrates) (646)
B3R	Ribosomes (0)	B6B	Insects (168)
B3S	Synaptosomes (22)	B6F	Fish (150)
		B6H	Human (37)
		B6I	Invertebrates (non-insect) (101)
		B6P	Plants (118)

\* These numbers are constantly changing as new data are added daily.

Table 3  
Class Codes--Physical Database  
{Number of sets in parentheses}\*

<b>P1</b>	<b>Theoretical (28)</b>	<b>P7</b>	<b>Addition</b>
		P7D	Dimerization (10)
<b>P0</b>	<b>Unknown</b>	P7E	Electrophilic Addition (149)
		P7N	Nucleophilic Addition (215)
<b>P1</b>	<b>Ionization (1502)</b>	P7P	Polymerization (12)
P1P	Ionization Potential (29)		
P1X	Proton Exchange (67)	<b>P8</b>	<b>Elimination (147)</b>
		<b>P9</b>	<b>Rearrangement (182)</b>
<b>P2</b>	<b>Hydrolysis (773)</b>	<b>P10</b>	<b>Oxidation (420)</b>
		<b>P12</b>	<b>Radical Reactions (556)</b>
<b>P3</b>	<b>Solvolysis (613)</b>	<b>P13</b>	<b>Complex Formation (100)</b>
<b>P4</b>	<b>Spectra</b>	<b>P14</b>	<b>Partitioning (131)</b>
P4I	Ionization Spectra (61)	P14C	Chromatography (21)
P4E	ESR Spectra (2)		
P4M	Mass Spectra (12)	<b>P15</b>	<b>Pyrolysis (88)</b>
P4N	NMR Spectra (175)	<b>P16</b>	<b>H-Bonding (26)</b>
P4R	IR Spectra (9)	<b>P17</b>	<b>Electrochemical (226)</b>
P4U	UV Spectra (23)	<b>P18</b>	<b>Brønsted (119)</b>
		<b>P19</b>	<b>Esterification (232)</b>
<b>P5</b>	<b>Miscellaneous Reactions (418)</b>	<b>P20</b>	<b>Photochemical (39)</b>
		<b>P21</b>	<b>Hydrogenation (16)</b>
<b>P6</b>	<b>Substitution</b>	<b>P22</b>	<b>Isokinetic (3)</b>
P6E	Electrophilic Substitution (238)	<b>P23</b>	<b>Reduction (81)</b>
P6N	Nucleophilic Substitution (1047)		

\* These numbers are constantly changing as new data are added daily.

In compiling the physical database from mechanistic physical organic chemistry we have concentrated on chemical reactions in solution. Although there are some examples (275) based on spectra and gas phase reactions, no attempt was made to be complete in these areas. The same applies to the Brønsted reaction (108 examples).

Many papers report results from runs at a variety of temperatures. Generally we have reported only 1 example at the temperature nearest to 25°. 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. That is, intercepts 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 feel that we have 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, psychobiology, 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 do provide significant help to the researcher. A further complicating factor is that reports on these studies, that are 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 over 20,000 experimentally measured octanol/water log P values of which over 10,000 are unique and considered to be reliable), from which hydrophobic parameters are derived, have appeared in over 500 different journals. This does not include the many reports on various types of chromatography carried out for the same purpose. We believe the time has come to integrate these results. Since a *variety* of approaches are currently being studied for the formulation of QSAR, one might question whether this is the time to push this project. 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 and molecular dynamics.

This applies to many of the QSAR in our current system. Many sets have been poorly designed. The QSAR for these have low  $r^2$  values and outliers and sometimes have two few datapoints/variable. Nevertheless we have found such first attempts to be helpful in supporting each other and suggesting the next steps. 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 awesome. Nevertheless, the enormous drug industry constantly faces these problems. The DNA codes for 50 to 100 thousand proteins that account for the many enzymes and components of the variety of cellular membranes and those of the organelles. All sorts of biochemical processes are open to perturbation. Thus, it is not yet clear what quality (in terms of  $r^2$ ) one can expect with the complex Bio systems. However, a rational and statistically based analysis is far better than intuition.

Our current premise is that the major interaction forces to consider in a set of congeners acting on a biological system are electronic, steric and hydrophobic. Less important factors are hydrogen bonding and dipole moments. Hydrogen bonding can be important. However, as yet there is no general way to deal with it, in the way that one can use  $\text{P}_i$  ( ), 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 parameterization appears impossible. In this case, indicator variables can be helpful.

Before considering the details of how the total system operates, it is instructive to point out a few practical uses. For example, under P1 (table 3) there are 1,502 examples of QSAR based on the ionization of all sorts of acids and bases. Of these, 1,301 are in terms of pKa. By moving to the show mode (see below) one can by calling for system (1) and compound (3), scroll through all 1,300 sets for the ionization of acids and bases in less than one hour noting those set numbers of interest. These can then be loaded for detailed study. Of course, a similar approach can be used for

any of the much smaller classes in tables 2 and 3. One could load all of the summaries for ionization including the QSAR and page through them in a couple of hours. This approach provides mini reviews of massive amounts of research. By isolating all of those containing a term in either  $\sigma$ ,  $\sigma^-$ , or  $\sigma^+$  we obtain 6,471 equations restricted to aromatic systems. Eliminating those based on  $\sigma^-$  or  $\sigma^+$  we obtain 3,867 based on  $\sigma$ . There are now 1,113 different substituents that have measured values of both  $\rho_p$  and  $\rho_m$  (*i.e.*, 2,226 values). (There are 1,999  $\rho_p$  values, including calculated values). Thus we can calculate ionization constants for  $2,226 \times 3,867$  (8,607,942) monosubstituted acids or bases assuming that there are two positions free for substitution. This figure is high since in some instances there is more than one QSAR for a given acid (*e.g.*, benzoic) in different solvents. Fujita<sup>13</sup> has shown how the effect of ortho substituents on ionization can be taken into account so that these too can be considered as well as multiple substitution. The possibilities can be extended to acids of the type X-Ar-Y-Q where Q is an acidic or basic group. A method is now available<sup>3</sup> for calculating the attenuating effect on Q of many examples of the linking group, Y (*e.g.*,  $-\text{OCH}_2-$ ,  $-\text{CH}=\text{CH}-\text{CH}=\text{CH}-$ ,  $-\text{SO}_2\text{CH}=\text{CH}-$ , etc.).

The ionization of aromatic amines and phenols generally depends on  $\sigma^-$ . There are 375 substituents with measured  $\rho_p^-$  values<sup>14</sup> which along with the corresponding  $\rho_m$  provide many possibilities for calculating these ionization constants (it is generally found that  $\rho_m = \rho_m^- = \rho_m^+$ ). There are 879 Taft-type<sup>2</sup>  $\sigma^*$  for use with aliphatic reactions where the field/inductive effect is important. Thus for the 110 QSAR of aliphatic acids and amines many new ionization constants can be calculated.

The sterimol steric parameters exist for 1,078 different substituents.<sup>2</sup> With the huge variety of substituent parameters, and the ability of our current regression program to automatically load these and over 14,000 QSAR for study, we believe that comparative QSAR has been sufficiently organized so that it can do much to extend our understanding of chemical and biological reaction mechanisms. As will be discussed later, all sorts of evaluations of substituents and parameters can be made. For instance, we are finding the sterimol parameters B1<sup>15</sup> to often be more effective in the correlation of intramolecular steric effects than Taft's classic  $E_s$  parameter. At present in the

physical database there are 296 examples where B1 appears in QSAR and only 224 for  $E_s$  or  $E_s^C$ . We have not yet checked many of the early entries when only  $E_s$  was available. We suspect that sterimol parameters will do as well or better in many instances. Since sterimol parameters are calculated, rather than measured like  $E_s$ , many more can be readily obtained.

All outliers in the database are labeled. Searching the physical base we find 2,806 QSAR have one or more outliers. This figure is somewhat unreal. In entering the early work, ortho substituents were often entered, but not used in deriving the equation. Thanks to the work of Fujita and Nishioka<sup>13</sup> we can now cope with such steric problems. We can study the 2,806 outliers to see which substituents have caused problems.

The bank can be searched to find examples where a particular substituent has been used. For example in the physical base,  $CF_3$  appears in 921 equations,  $OCF_3$  in 31,  $SCF_3$  in 14 and  $SF_5$  in only 9. In Table 2 the numbers in parenthesis add up to more than the total for the database. The reason for this is that in some instances more than one label is attached to a single equation. A reaction might be labeled as P10 and P12 (radical and oxidation) or it might be tagged as B4B and B6A for drugs attacking *E. coli* in mice.

The major difference between the biological and physical database equations is the importance of hydrophobic terms, logP, and  $\pi$ , in the former. Out of 6,300 Bio-QSAR 4,367 contain such a term (69%). Considering the five major categories the following disposition is found:

Table 4

	QSAR with Hydrophobic terms	Total QSAR in class	%
Enzymes	1291	2400	54
Organelles	300	395	76
Cells	1386	1829	76
Organs	527	741	71
Whole organisms	1014	1218	83
Receptors	379	881	43

As might be expected hydrophobic terms show up least frequently in enzymes and receptors and of course, show up most often in QSAR for whole organisms.

There is much still to be learned about the role of hydrophobicity in the design of greater selectivity in bioactive compounds and less toxicity in industrial chemicals. The principle of minimal hydrophobicity in drug design is obviously important for bio availability,<sup>16,17</sup> and one must always be alert for the presence or absence of hydrophobic interactions. Sometimes compounds participating in electrophilic reactions with cells or whole organisms do not display a hydrophobic effect,<sup>18</sup> but the reason for this is not always clear at present.

The environmental toxicity of chemicals is constantly under review. Three widely studied test organisms: fathead minnow, tetrahymena and daphnia, are represented by 40, 53 and 47 sets respectively. Results with these toxicity studies can be compared with those from other organisms, cells, etc. For instance, pentachlorophenol occurs in 140 sets, atrazine in 10, and methoxychlor in 21.

The present report outlines the current state of our efforts to organize the QSAR literature to make it readily accessible for a virtually unlimited variety of comparative analyses. This will not be the last such effort and we hope that our ideas will be of value to others considering the next steps in comparative structure-activity analysis. Biological QSAR has become big business and large amounts of money are being invested in its development. We are only in the early stages of a field that computers have made intensely exciting.

## **II. Organization of QSAR Database**

The database has been organized in two major areas of chemistry: Biological (table 2) and physical-organic (table 3). Table 1 shows that the datasets in each of these areas is subdivided into input information and equation information (output). Class Codes in Tables 2 and 3 simplify searching. Normally the user is interested in either the physical or biological section. Splitting them saves searching time, but more important, it makes the evaluation of the results of the search easier. A major objective of our system is to obtain the desired data in a precise, quickly-scanable form.

Originally the 'systems' under study were entered and are accessible only by name. For example, there might be a QSAR study on mice with acute LD<sub>50</sub> as the reported activity. But if 'mice' were the system entry used in searching, one might miss comparisons where the QSAR was entered under the system 'mouse' or 'murine'. When there is ambivalence about names, enter both. Searching under mouse find 311 while searching with both mouse and mice finds 412. One might also want to expand the comparison to include certain vertebrates (such as rats or guinea pigs) but exclude others (such as fish). To make both entry and access more specific, yet keeping it simple, we have added a classification code as the second input data item in Table 1. It is apparent from table 2 that the code numbers increase with increasing complexity of the biological system. In all cases the 'class' is the broadest category and is narrowed by the 'system name'. Searching with **2 B6A** isolates all vertebrates (646), then string searching with **rat** finds 197 sets.

A major problem in searching either database for parameters of a desired type is knowing what label to use. As indicated below, we have developed standard labels for the most often used parameters, but any parameter can be used if it is defined in the note section. Of course finding sets containing such unusual parameters is difficult unless one knows the symbol. The following symbols define the major parameters. One can use the browsing mode to look for unusual labels.

Electronic Terms. S, S-, S+, S., SI, and S' are the English alphabetical representation of the Hammett type constants:  $\sigma$ ,  $\sigma^-$ ,  $\sigma^+$ ,  $\sigma^{\bullet}$ ,  $\sigma_I$  and  $\sigma^*$  (see ref. 1 for definition and many examples of their use) Note that for the radical parameter (sigma dot) a period is used, and the prime symbol is used instead of the asterisk to denote Taft's sigma star, since the asterisk has another programming role. S' is *not* sigma prime ( ' ).

At present the database contains a few QSAR based on molecular orbital calculations. The labels employed are: HOMO, LUMO and Q<sub>i</sub>. In some instances LUMO-HOMO has been used to define the HOMO-LUMO gap. E  $\frac{1}{2}$  is the potentiometric oxidation or reduction value, and U represents dipole moment. BDE stands for bond dissociation energy.

Hydrophobic Terms. Log P, Mlog P and Clog P parameters are based on octanol/water partition coefficients for the neutral form of the compound. Mlog P (M= measured) values can be



automatically loaded from our preferred list of over 10,000 experimental values while Clog P values are calculated from structure by a fragmental method<sup>19</sup> and are now calculated and automatically loaded for regression analysis. Recently, Leo has modified the method of calculating Clog P so that one no longer faces the missing fragment problem. In early work log P values were loaded by hand and may contain values calculated manually. Log P' represents something other than the standard octanol/water value for a neutral compound. It may come from another system (*e.g.*, chloroform/water) or it may be from buffered octanol/water when ionizable compounds are under consideration. It should be defined in the notes. Hydrophobic parameters have sometimes been obtained by chromatography, and these are termed RM. PI represents  $\pi$ , the hydrophobic constant for aromatic substituents. For automatic loading the PI value for the normal benzene system is chosen--not the value as seen on nitrobenzene, pyridine or aniline.

Steric Terms. ES is the classical Taft parameter<sup>20</sup> and ESC is the form of  $E_s$  corrected by Hancock for hyperconjugation.<sup>20</sup> B1, B5 and L are the sterimol parameters proposed by Verloop, Hoogenstraaten and Tipker<sup>15</sup>. MR represents the molar refractivity of a substituent while CMR is the molar refractivity which is automatically calculated for the whole molecule. Both are scaled by 0.1 and can be automatically loaded. MV represents molar volume that may have been obtained by a variety of means. MGVOL is the McGowan<sup>21</sup> molar volume for the whole molecule that can be automatically calculated and loaded for regression analysis. Values for most of these can be found in the Parameter database (THOR Sigma).

Very often substituent effects (and hence parameters) are position dependent. This is indicated by a suffix following a comma: S,M or PI,4 or B1-3; where M(eta),4 and 3 represent the ring positions where the substituent is attached. Multiple positions on a parent structure given in Field 3 in the input data are indicated by, for example, PI,X,Y or PI-SUM.

Ad Hoc Parameters. All sorts of special parameters have been used by different authors. Usually these are defined in the notes, but sometimes the original paper must be consulted. Most prominent are the indicator variables (1 or 0) used for the definition of many different kinds of structural variations. When they are not defined in the notes, their meaning is easily deduced from

an inspection of the columns of parameters. Recently solvatochromic parameters,  $\alpha$  and  $\beta$ , have been defined as measures of H-bond donating and accepting capability respectively. These are labeled 'alpha' or 'beta'.

Dependent Variables. In the physical database the dependent variable is most often a rate or equilibrium constant represented by K or KREL. The latter symbol generally means relative to H. In the biological database the most common dependent variable is  $\log 1/C$  where C is the molar concentration of chemical (moles/liter or moles/kg) used to produce a standard response. When the time is fixed this can be shown to be a rate constant. Sometimes RBR is used where RBR represents a relative biological response. When  $\log 1/C$  is used intercepts of the QSAR can often be compared with others. Of course this is not true for RBR. At present  $\log 1/C$  is used in 3,947 sets. For enzymic reactions KM, KCAT and VMAX represent  $K_m$ ,  $k_{cat}$ ,  $k_{cat}/K_m$  and  $V_{max}$ .

Integrated with the QSAR database is a model-building regression program (Section IV) into which any particular existing data set can be loaded and re-studied with new or additional parameters and structures.

Several other databases and calculation algorithms complete the suite of software in the C-QSAR™ 'package'. The MASTERFILE database contains over 45,000 log P values measured in over 300 solvent systems, with octanol/water dominating. It also lists pKa's and activity types for about 11,000 structures.

### III. Using the Databases

Note: In the following sections the commands to be typed in are in boldface and underlined, and a break in the underlining indicates a space is required, but case is not important.

**A. Help:** The user is encouraged to use the 'Help' function at just about any point in this exercise. Entering ? ( or **help**) delivers a brief help message, and choosing any of the functions listed preceded by ?? will give detailed help. For example, while in the regression mode, ? will return a list of functions, and if one were interested in 'eigenvalues', entering **?? eigenvalue** will

tell how to calculate them for a set of parameters. **?? Pred** shows the various way results from a calculation can be displayed.

**B. String Searching:** Grammar plays a very important role in this system, and the 'grammar' involved, while simple, must always be kept in mind or else spurious results will be obtained. The grammar for string searching can be illustrated with the word 'in'. It can stand alone (as a preposition) or be part of another word. The following examples show four distinct contents in which it can appear.

E.coli <b>in</b> mouse	a word	(both leading and trailing blanks) " <u>in</u> "
<b>in</b> fluenza	start of word	(leading blank, but no trailing blank) " <u>in</u> "
bra <b>in</b>	end of word	(trailing blank, but no leading blank) <u>in</u> "
pyrid <b>in</b> e, gu <b>in</b> ea	inside a word	(neither leading nor trailing blanks) <u>in</u>

To match only strings occurring at the start or end of a word the string must be 'extended' to include a leading or trailing blank. This is done by starting/ending the query with a quote and blank. A quote and blank before the set of searching letters restricts matches to the beginning of the letter string, while a blank and quote after the string of letters restricts matches to the end of words. To match only whole words, leading and trailing blanks must be present. Some further examples may make this clearer.

**"HEM"** or **HEM** matches **HEMOGLOBIN**, but not **CHEMOTHERAPY**.

**"ASE "** matches **LYASE**, but not **L. CASEI**.

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 **HEMOGLOBIN**, but not **CHEMOTHERAPY**

**ASE "** matches **LYASE**, but not **L.CASEI**

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**

In combining 'strings' it is important to note that a space denotes 'or' and expands the search, while a comma denotes 'and' and restricts it. Examples will follow which will demonstrate the power of string searching both databases. While learning to use the system, one should inspect the results from searches to be sure that one has found only the desired information.

### C. Physical Database:

After logging on, enter **QSAR** at the system prompt (usually the \$ sign), and follow it by **data physical**. A prompt then asks for a password. For read only access press RETURN and the main menu is displayed (table 5).

Table 5  
Main Menu

1	Summary	DIR
2	Show	HELP
3	Search	PRINT
4	Browse	QUIT
5	MedChem	READ
6	Manager	REG
7	Save Database	VMS
		WRITE

**1. Main Menu.** Entering **1** gives a summary of the number of sets, compounds and SMILES (see section VI). This table applies to both databases. The biological database has 113,750 data points (compounds). The physical base contains 76,111. Calling for 'help' will instruct you how to exit from any mode and call another as well as reinforce the explanations in this Manual.

**2. Searching and Show** is accomplished in two steps: first one enters the **search** mode from the main menu (3), then one enters the command and presses return to start the search. After the search is completed, use the **show** mode to display the hits. The Search Menu is shown in table 6.

Table 6  
Search Menu

SEARCH	0 Equations	8 Note	15 Terms in eq.	DIR
BACKUP	1 System	9 Date	16 Coefs. in eq.	HELP
BLANK	2 Class	10 Parameters	17 Ideal/logB	PRINT
NOT	3 Compound	11 Substituents	18 Statistics	QUIT
LIST	4 Action	12 SMILES	19 Residuals	READ
	5 Reference	13 MERLIN	20 Predicted	REG
	6 Source	14 Prm max/min		VMS
	7 Check			WRITE

The Show Menu differs from the Search Menu only in the left hand column which illustrates the sorting option.

As noted above, the grammar of string searching must be carefully followed or else one can get more than the expected result. For instance, one might search the physical base with the command **2 P1** where **2** refers to the 'class' as listed in table 1 and **1** is thought to refer to the code for 'Ionization' as listed in table 3. However with this entry one would recover 3,414 sets, which is far too many for ionization alone. As entered, the command locates all sets with classes P1, P10, P12...etc. and finds 'oxidation', radical reactions' etc. The correct entry has leading and trailing spaces **2 " P1 "** and finds only 1,493. Entering **SHOW** takes one to the show mode where the results can be inspected.

The value of string searching can be illustrated with a search of the action field while in the 'Search Menu'. Entering **4 bromin** returns the Search Menu with a status check showing the search will be performed on the entire physical database of 7,700 sets and the search requested is 'Action....BROMIN'. To perform the search, enter **search**, and 167 hits are found. To peruse the 'catch', enter **sh** and then **4** and one finds bromination, brominolysis, photobromination, dehydrobromination, etc. If you were interested only in equations where bromine is involved in an addition reaction, which is classed as P7 in table 3, enter **search** to return to that Menu, and enter **2**

**P7** and then press return, which shows the status check indicating that the search will be made on the 155 hits made previously. Enter search, which finds 54 sets of bromine addition reactions. If you want to see bromine in radical reactions, you need to return to the original 155 hits. This can be done by blanking out the last search with the entry **bl 2**. Then enter **2 P12** and after approving the status check, enter search. This returns the Menu showing 54 hits. One might want to view these sets showing 'system', 'action' and 'terms', and would enter **sh 1 4 15**. To return from any Menu to the Main Menu, enter **q**. To illustrate another way to search the bio database from the search mode enter **1 HIV** to find 128 QSAR associated with the AIDS virus.

**Important Note:** In combined commands used in the 'show' mode, the grammar differs somewhat from that used in string searching. A space specifies 'and' (not 'or'), and a comma specifies 'through'; *e.g.*, **1 4** means 'one and four' but **1,4** means 'one through four'. This will become more apparent in further exercises.

**More complex Searching Examples:** (The novice can skip to section III to become familiar with other basic operations before returning to this section, if desired.)

Category 1 in Table 1 is SYSTEM, and for the physical database it refers to the solvent in which the reaction was run. In searching this field the **NOT** command is very helpful. Often mixed solvents were used as the reaction medium and the % sign is always used in their identification. The importance of this feature can be illustrated if we wish to search for reactions run in aqueous solution. ('aqueous' is always used, not 'water'). Searching with **1 aqueous** would make 3,441 hits. Adding to the search **1 not %** eliminates mixed solvents and would reduce this to 1,398. Searching with **2 " P1 "** finds 533 QSAR for ionization in water.

Remember that results from all previous searches can be removed by entering **blank**. Alternatively, one or the other can be removed by **bl 1** or **bl 2**. To find which of the 7,700 sets are based on mixtures of ethanol and water enter, **1 aqueous, " ethanol "**. Note that several commands can be entered on one line by the use of commas which denote 'and', but the status report shows them on two lines. This search finds 698 sets based on ethanol-water mixtures. To see if certain steric parameters are significant in the QSAR of these sets, enter **15 MR ES B1**

**B5**. There are 49 hits. Entering **show** and then **15** displays the variables used in each set, with the dependent variable first (15 refers to terms in the equation, Table 1).

Any number of codes can be used simultaneously. Entering **2 P3 P2 "** in the search mode, collects all examples of hydrolysis and solvolysis (1,383). The quotes on P2 are required by string searching grammar so that P20, P21, etc. would not be included.

Collecting studies based on the various Hammett-Taft sigma constants requires some thought. Besides S, S<sup>+</sup>, S<sup>-</sup> and S' (computer-compatible letters for the Greek symbols  $\sigma$ ,  $\sigma^+$ ,  $\sigma^-$  and  $\sigma'$ ) various positional suffixes have been attached to sigma. Most common are S,X and S,Y where the sigma parameter may be of a different type at the two positions; *e.g.*, S,X and S<sup>+</sup>,Y. Also, it must be remembered that in string searching one must use a leading blank as in **" S** or sets with ES will be included. If you wish to find all occurrences of 'ordinary' sigma (S), including those at specified positions but excluding the 'special' sigmas (S<sup>+</sup>, S<sup>-</sup>, etc.), the following steps should be followed. Entering **15 " S** finds 7,111 sets including both positional and 'special' sigmas. To remove the latter, enter **15 not S- S+ SI S' S.** and 3,859 remain. Notice that the 'not' command is not necessary in searching for the 'special' sigmas. For example, **15 " S-** finds all QSAR based on sigma-minus parameters (980), including those at specified positions.

In the above examples we have focused on S alone, but the QSAR collected might contain other terms. These could be eliminated as follows:

- |    |           |              |                  |            |
|----|-----------|--------------|------------------|------------|
| 1. | <b>15</b> | <b>" S "</b> |                  | 3,618 hits |
| 2. | <b>15</b> | <b>not</b>   | <b>MR</b>        | 3,612 hits |
| 3. | <b>15</b> | <b>not</b>   | <b>B1 B5</b>     | 3,490 hits |
| 4. | <b>15</b> | <b>not</b>   | <b>ES</b>        | 3,435 hits |
| 5. | <b>15</b> | <b>not</b>   | <b>logP PI</b>   | 3,423 hits |
| 6. | <b>15</b> | <b>not</b>   | <b>**2 bilin</b> | 3,385 hits |

The symbol **\*\*2** represents squared terms such as S<sup>2</sup> or PI<sup>2</sup> and bilin represents bilinear equations. Note that the NOTS must be listed as a separate entry.

Moving to **show** and entering **15** we see that except for a very few examples, all QSAR are based on the single term S. The search could be changed by the command **blank 2 4** which removes these two commands leaving:

1. 15 " S "
2. 15 not B1 B5
3. 15 not logP PI
4. 15 not \*\*2 bilin 3,445 hits

Non linear QSAR. Over the years nonlinear QSAR have been found, and a variety of approaches have been devised to deal with them. Most of these involve using more than one variable. Because of the lack of general agreement on how to correlate such data we have normally used a single parameter and either a parabolic equation ( $a + b^2$ ) or a bilinear equation. These can be searched for as follows:

15 S+\*\*2 finds 114 eq. parabolic in +  
15 bilin locates 31 sets bilinear in any parameter

Range Searching. As noted above, searching the physical database with 15 " S " finds 3,618 sets. It does not include electronic parameters such as S, X, etc. Normally, this is too large a list to scroll through. If one wanted to see only those equations with a modest electronic dependence, one could enter 16 -.8 < " S " < 1. This locates those equations where the coefficient with sigma ( ) lies between -0.8 and +1. (1,139 examples). In the Search Menu, 16 is the code for the coefficient which, in the present case, is that for S. We might now check these sets to see how many examples of radical reactions have a term in S, with such slopes, by entering 2 P12 (The Search Menu shows '2' as representing 'class', and table 3 shows 'P12' representing Radical Reactions.). We find 73. This more modest list of equations can now be viewed, logically ordered by increasing sigma coefficient. First enter show. The Show Menu lists all the items that can be displayed with each equation. At first we may only want to see the essentials, and so we enter /sort = 16 1 3 4 15, 18. (As noted previously, in the grammar in the Show Mode a space means 'and' and a comma means 'through'.) The program then asks which coefficient to order the display on, and one enters " S ". The information specified by 1 3 4 15, 18 of the Show Menu is displayed and the sets are ordered in terms of increasing values of the coefficient, , from -0.798 to + 0.982. Since 16 delivers the equations with coefficients and terms, it might seem redundant to specify 15 also, but it is so often useful to spot the significant parameters at a glance.



Compound. Searching for specific compounds by their 'names' cannot be made effective. In Section 3 of Table 1, entries such as phenols, X-C<sub>6</sub>H<sub>4</sub>-COOH or benzodiazepines have been used. Compounds can more effectively be found by using the SMILES notation or by substructure searching using MERLIN as shown below. However, searching category (3) for generalized structures can be instructive, if preceded by a search by Class. For example, one might want to see what types of chemicals have been studied in radical reactions. In the Search mode enter **2 P12** followed by search. This locates 556 sets. For a quick overview of the compounds studied, enter show and then **3**. Scrolling through these examples takes less than ten minutes (imagine how long it would take to locate these in the library) and set numbers for those of interest can be noted. These can be examined in detail by switching to the regression mode (reg) which will be covered in Section III-7. Not surprisingly, it is seen that toluenes are a favorite subject for radical study. Returning to search and entering **3 C6H4CH3 C6H4Me Toluene** (note string searching is being called for) finds 73 cases where various derivatives of toluene have been the subject of investigation. This gives a quick, but probably incomplete, answer. A more systematic method which yields 111 hits is discussed later.

Action. (4 in Search Menu) Since uniform nomenclature for entry of this type of data has not been developed as yet, one can not be sure that a search will find all that one is interested in. Since searching is so rapid, a viable strategy is to start with a broad search and narrow it as you see what it finds. For example, enter **4 chemical** which sequesters 183 examples where this word was employed. Perusing the hits in the show mode (4), we find words such as photochemical, electrochemical and chemical shift. Repeating the search entering **4 chemical, shift** uncovers 141 NMR studies. Searching with **2 P4N** locates 175 NMR studies which includes those based on coupling and splitting constants in addition to chemical shifts. We have made very little effort to include QSAR on spectra or the Brønsted reaction.

References. (5 in Search Menu) Searching the reference category can often be of interest. One can often recall the name of a person who made a certain study, but cannot remember where or when. For instance, searching **5 Bordwell** locates 103 sets. These can be narrowed by checking

certain years; *e.g.*, 5 (1997) (1998). Eight QSAR came from recent papers published in these two years. Entering bl 2 to begin again with 103 hits and entering 5 Cheng isolates 19 QSAR by Bordwell and Cheng.

It might be of interest to see the trend in publications of QSAR in physical organic chemistry. The following searches could be made: 5 (1950) (1951) (1952); 149 hits. 5 (1970) (1971) (1972); 898 hits. 5 (1990) (1991) (1992); 352 hits. It is necessary to put parentheses around the year to distinguish it from page numbers. There are 114 QSAR from 1998. Remember that these are equations not individual papers. Several QSAR may come from one article. Although remaining strong, interest in the Hammett type equations in the 90's seems to be subsiding. Publications in particular journals can be checked. 5 J.Am.Chem.Soc. finds 1,619. Note that one should not leave spaces between the words in the title, but case is not important. From another point of view we can see where most publications occur by entering the following three searches in sequence:

5 not J.Chem.Soc.  
5 not J.Org.Chem.  
5 not J.Am.Chem.Soc.

The hits decline as: 6279, 5268, and 3652. Thus out of 7,700 sets only about half have been published outside of these three journals. Entering sh followed by 5 one can scan the list of the less 'popular' journals where publications have appeared. Note that no space has been left between the abbreviations.

Searching for Similar QSAR. One of the most important uses of the C-QSAR program comes *after* a new QSAR has been derived. The database can then be searched for comparable equations that may help validate the new one. For example, suppose that you have formulated a new QSAR for an aromatic nucleophilic substitution using the (S-) parameter and the resulting coefficient  $\rho = 2.7$  Table 3 gives the Class for this type of reaction as P6N, and so you can search for similar equations by entering:

2 P6N which delivers 1047 hits for nucleophilic substitutions, and then:

16 2.6 < S- < 2.97 which narrows the sets of interest to 9.

The first command isolates all nucleophilic substitution reactions. The second finds those with  $r$  in the range 2.6 to 2.97. The following are representative examples. Going to **Sh** and listing with /sort = **16 1 3 4 15 16 18** followed by **S-** yields:

	System	Compound	Reagent	-
1.	20° Aqueous	X-C <sub>6</sub> H <sub>4</sub> OCO(CH <sub>2</sub> ) <sub>3</sub> NMe <sub>2</sub>	Internal	2.60
2.	20° Methanol	2-p-nitrophenoxy-3-NO <sub>2</sub> -5-S-thiophenes	piperidine	2.61
3.	20° Aqueous	X-C <sub>6</sub> H <sub>4</sub> OCO(CH <sub>2</sub> ) <sub>4</sub> NMe <sub>2</sub>	Internal	2.65
4.	75° Benzene	6-X-2-NO <sub>2</sub> -C <sub>6</sub> H <sub>3</sub> Cl	piperidine	2.72
5.	20° Benzene	2-phenoxy-3-NO <sub>2</sub> -5-X-Thiophenes	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub> NH <sub>2</sub>	2.75
6.	50° Methanol	4-X-1-I-2-nitroiodobenzene	N <sub>3</sub> <sup>-</sup>	2.90
7.	20° Methanol	2-Br-3-NO <sub>2</sub> -5-X-thiophenes	piperidine	2.96

Two examples omitted.

It is sometimes desirable in comparative QSAR to focus on equations with a specific number of terms with a limited number of variables. This can be illustrated with a search for all QSAR containing only 3 terms, all of which are variations of .

1. 18 2<terms<4 196 hits
2. 15 " S 186 hits
3. 15 not ES I MR D F B1 B5 \*\*2 bilin 22 hits

The first search isolates all QSAR containing 3 terms. The second ensures that occurs at least once and the third that all QSAR which contain a term other than or nonlinear terms are eliminated. These generally are reactions involving two molecules in both of which substituents have been varied. It should be noted that until one obtains experience it is good practice to inspect search results at each step to see if ones expectations are being met. A good example of such a three term equation is the following: (set #4405)



$$\text{Log } k_2 = -1.32(\pm 0.05) \text{ , } x - 0.13(\pm 0.01) \text{ , } y + 1.08(\pm 0.03) \text{ , } z - 3.93(\pm 0.01)$$

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

Another way of looking at QSAR with a definite number of terms is the following from the Bio QSAR database:

1.	18	3<T<5	361 hits
2.	18	n > 20	274 hits
3.	18	r>0.90	213 hits

This finds all QSAR that have four terms, a minimum of 5 data points/variable and a fair quality of correlation.

Continuing with the bio database:

Source. This compartment contains the name of the person who actually entered the data, not the paper from which it came. Entering **6 Gao** uncovers 1,445 data sets entered by Hua Gao.

Check. The name of the person who checked the entry is contained here. If it has not been checked 'unknown' is entered. Entering **7 not unknown** finds 2562 have at present been checked by someone other than the person deriving the QSAR.

Note, data and parameters. There is little or no value in searching these fields.

Substituents. This field can be of value for finding QSAR that contain a particular substituent if it has a generally accepted form. Searching the bio data shows the following. The methyl group might be entered as CH<sub>3</sub>, Me, or methyl. Searching **11 CH3** locates 949 sets; **11 Me**, 4,590 sets and **11 methyl**, 497 sets. All of these could be sequestered by one command; **11 CH3 Me methyl** . Searching substituents can be of value in locating unusual functions. **11 SO2C6H5** makes 44 hits. The location and type of aromatic substituents can be studied as follows **11 " 2-NH2 " "O-NH2 "** isolates 81 examples. Note that string searching is involved so that quote space notation is necessary. **11 " 3,4,5-Cl " " 3,4,5-Cl3 "** finds 26 examples. **" 2-NO2,4-NH2 "** locates 4 examples. **11 " 2,6-Cl " " 2,6-Cl2 "** uncovers 83 sets in which both ortho positions are substituted.

**3. Browsing.** From the Main Menu, entering **4** returns the following table.

Table 6  
Browse Menu

1. System	7. Check
2. Class	8. Note
3. Compound	
4. Action	10. Parameters
5. Reference	11. Substituents
6. Source	12. Smiles

By entering any of these code numbers all of the examples in the system will be listed exactly as they have been entered without (as yet) any kind of organization. It still can be useful for the beginner to become familiar with the terms that have been used, especially in the biological database. Ordinarily only categories 1, 3, 4 and possibly 5 are of interest to browse.

**4. Statistics.** With each QSAR the following statistics are given:

N	Number of datapoints
DF	Degrees of freedom
R	Correlation coefficient
R2	Squared correlation coefficient
S	Standard Deviation
SS1	Sum of squares about the mean of the dependent variable
SS2	Sum of squares from the deviations from the regression line
D+	Number of positive deviations from the QSAR
D-	Number of negative deviations from the QSAR
Omit	Number of datapoints omitted in deriving the QSAR
Q2	Indication of the quality of fit of the data

Any of these fields can be searched, but only a few are normally of value. They are accessed by 18 of the Search Menu (same in table 1). Checking the Phys data illustrates the useful searches. The command **18 n>20** in the search mode finds only 447 out of 7,700 QSAR that are based on 21 or more data points. **18 n>10** hits 1,882 sets and **18 n>4** locates 6,651 based on 5 or more compounds. **18 n<4** locates 123 based on 3 data points. Of course, three data points is not enough to properly define a QSAR, however, if it is very sharp with a good spread in the data points it is better than nothing. There are no equations based on 2 data points. Physical organic chemists have mostly been interested in establishing a value of  $r$  for a reaction with a minimum of

effort and hence often have not studied as many derivatives as those interested in Bio QSAR do. Normally one needs a minimum of 5 data points/variable with well spread parameter values.

The quality of the correlations can be evaluated by means of the correlation coefficient  $r$ . Entering **18 r >.90** shows that 7,552 sets have correlation coefficients greater than 0.90. **18 r ≥.95** yields 6,856 QSAR and **18 r >.99** finds 3,209. Quality can also be checked with respect to the standard deviation  $S$ . **18 S <.10** hits 3,973 QSAR with standard deviations less than 0.1 while **18 S < .20** makes 6,129 hits. Quality can also be analyzed in terms of the deviation of individual data points (Residuals, **19**). A search with **19 -.05 < dev < .05** hits only 1,075 examples where no calculated value in the set deviates by more than  $\pm 0.05$  from the experimental value. This is a very stringent standard. Relaxing the standard to  $\pm 0.2$  finds 3,842 examples.

**5. Omitted Data Points.** In developing a QSAR the question is often raised as to when, if ever, one should withhold data points. We believe that there are several good reasons for doing so. Outliers may be pointing to a failure in the mathematical model; or they might result from an error in the method of calculation of the dependent variable; they can be the result of an experimental error or they can result from a side reaction. Whatever their source, it is extremely important that omitted points not be forgotten.

In our system, to omit a point it must be marked by an asterisk (starred) which is held with the data point so that it will not be forgotten. Moreover it becomes possible to make generalized studies of outliers. For example, entering from the Phys search mode **2 P12** collects all examples of radical reactions (556). Now entering **18 omit > 0** isolates all QSAR with one or more starred data points (214). Moving to **show** and entering **11** lists all substituents for each data set.

Using the above approach any set of data can be rapidly surveyed to obtain an overview of the QSAR. For instance entering from the search mode **2 P5** isolates all miscellaneous reactions, at present 418 examples. Some of these can be re-classified. Moving to **show** and entering **1 3 4 11 15 16 18** one can look over all that has been done without loading each individual set to examine the results. One can scroll through the output in half an hour. Note that beside each substituent, is the residual, that is, the difference between the observed and experimental value for the current

stored equation. In this manner, the poorly behaved substituents for any type of reaction could be determined.

Example 1 shows how exemplars for any type of parameter could be selected and example 2 illustrates how good examples of a particular type of reaction (in this case nucleophilic substitution) can be found.

Example 1	
<u>15</u> " S+	1,653 hits
<u>18</u> n > 10	335 hits
<u>18</u> r > .98	161 hits
<u>19</u> -.1 < dev < .1	17 hits

Example 2	
<u>2</u> P6N	1,047 hits
<u>18</u> n > 10	276 hits
<u>18</u> r > .98	175 hits
<u>19</u> omit < 1	175 hits
<u>19</u> -.1 < dev < .1	52 hits

These sets could be used to test for calculation of electronic effects of substituents by newly developed methods.

Of course one might want to find poorly fit data to test methods to improve it.

<u>15</u> S+	1653 hits
<u>18</u> n > 15	142 hits
<u>18</u> r < .90	2 hits

**6. SMILES.** Almost all structures in the databases are entered in the SMILES notation (section VII) so that any compound can be located via its SMILES or a common name. For example, to find QSAR which contain phenol, go to the search mode in the Phys bank (enter **data phys** and then **12**). This returns a panel where one can enter either the SMILES (**clccccc1O**) or **phenol**. Pressing **return** shows the structure of phenol. If it is correct press **n** if not press **y** for editing. Pressing **n** returns the program to the searching mode. Entering **sea** finds 285 sets that contain phenol. Normally this locates a set of phenols; however, sometimes phenol may be one of a miscellaneous set of compounds. Going to **show** and entering **3** one sees a variety of examples other than all-phenol sets. For more discussion of SMILES, see section III-C, 6 and ref. 25.

To find more specific information on the phenol sets isolated now enter, for example, **2 P6E** to collect examples where phenol is involved in electrophilic substitution. **Searching** finds 12 examples. Moving to **show** and entering **15** it is seen that 6 QSAR are based on + and 6 on . Entering **4** (action) shows that 7 examples involve bromination, 4 iodination and 1 on nitrosation.

Substructure searching is another approach to isolating specific types of compounds (Section VIII).

Obtaining the SMILES structure by entering the name has serious shortcomings. For instance, entering p-chlorophenol yields the SMILES, but entering 4-chlorophenol fails. Using the name is very helpful with complex drugs or natural products where the SMILES takes time to write and where there is a standard name. Entering **Quinine** to get the SMILES and then searching, one set of data is found in the physical database that contains quinine.

**7. Loading from 'Database Search' to 'Workspace'.** After a data set of interest has been found by means of the 'search' and 'show' modes, it can be examined in greater detail if transferred via its set number to the regression mode. (Modifying an old set or preparing a new one for actual regression analysis will be described later.) This is done by entering **regression** from either the Bio or Phys database, and for this and the following exercises, transfer to the biological database with the entry **data bio** followed by **reg**. The screen prompt becomes **qsar>**. Entering **load /d 1890** (set number) transfers the set to the workspace. Any data set can be viewed in the following ways by entering the indicated commands:

<b>Summary</b>	Lists key items of dataset.
System:	Canine gastric mucosa H <sup>+</sup> , K <sup>+</sup> -ATPase
Class:	B2C Hydrolases
Compound:	4-X-phenyl-2-NR-Guaindinothiazoles
Action:	I50
Reference:	Ojha, T.N. et.al. Ind. J. Biochem. Biophys. 30, 239 (1993)
Source:	Hansch
Check:	L. Zhang
Note:	I=1 for NCH <sub>2</sub> C <sub>6</sub> H <sub>5</sub> , I-2=1 for 3,4-di-OH
Parameters:	Y Pred Dev log 1/C I S <sup>+</sup> I - 2
Eq/run:	log 1/C = 0.62(±.18)I - 0.15(±.15) + 1.25(±.29)I-2 + 4.49(±.11) n = 23, r <sup>2</sup> = 0.901, s = 0.185, q <sup>2</sup> = 0.863
<b>Seedata</b>	Displays table of parameters and their values together with predicted values and their deviations from the QSAR.
<b>Seeequation</b>	Displays the last derived equation.
<b>eq/run</b>	Shows stored equation for comparison with any newly derived equation.
<b>Seeparameters</b>	Lists parameter labels with their numbers. For example:

1	2	3	4	5	6
Ypred	Dev	log1/C	I	S+	I-2



'Ypred' is the predicted value from the latest stored equation. 'Dev' is the difference between this figure and the observed value (item 3). The dependent variable is usually entered in position 3.

To see the SMILES generated structures for the compounds of a data set enter **depict 1**, which depicts sequentially all of the compound structures as one presses **return** after each panel of 4 structures. The process can be stopped at any point by entering **q**. This can be very important in dealing with a large data set, say > 100 compounds. To view any particular structure enter **depict #** (number of compound in set). To check all structures following compound 16 enter **depict 16**. To view all structures up to 16 enter **depict ,16**. To see those between 16 and 20 enter **depict 16,20**.

#### **D. Biological Database**

Our database of 6,300 Bio QSAR can now be used as the starting point for developing a science of chemical-biological interactions. We believe that as it grows, it will help to minimize redundant research and provide new leads via the study of familiar chemicals.

To access the Bio database from the \$ prompt, enter **QSAR** then **data bio** and on request for password press **return**. As in case of the physical database the Main Menu is displayed.

**1. Browsing.** The inexperienced user can use the 'Browse' feature in the biological database to even better advantage than in the physical database to become acquainted with its contents. In the Main Menu, entering **4** displays the Browse Menu. As an example, in browsing the 'system' you might find 'cat' as an item of interest. The items in 'system' have not been alphabetized, and there might be several studies of cats that may be of interest. One can search by entering **1 " cat "**, which returns the display:

cat	muscle tibialis cat	aortic ring of cat
gut cat	cat intestine	phosphodiesterase III from cat heart
liver extract cat	muscle tibialis anterior cat	

At this point it might be advisable to review the grammar used in 'string searching' as given in Section III.B.

**2. Searching.** String searching the 'system' field (1) in this database can be of more value than it was for the physical database. Field 1 does not have standardized information, but use of 'Class' (2) can increase its utility. For instance, entering **2 B4B** yields 669 sets relating to bacteria. Moving to the **show** mode, one can quickly page through these sets by entering **1** to survey just what kind of microorganisms have been studied. Or entering **3** gives the types of compounds that can also be quickly scanned. Such a survey of all of the work on bacteria can be made in less than 5 minutes. It would take days in the library. Returning to **search** and entering **1 aureus** makes 107 hits all of which are for *S. aureus*, except 3 for *M. aureus* and 2 for *P. aureus*. Now entering **15 not logP PI RM** eliminates all QSAR with hydrophobic terms, leaving only 8 examples. These are of special interest since hydrophobicity is so often a factor in Bio QSAR. Entering **bl 2** removes the last command. Now entering **15 " S** locates 35 QSAR with various sigma terms. Going to **show** and entering **/sort=16 1 3 4 15,18** and **" S** on the prompt displays the QSAR in order of increasing value of .

Using a combination of 'system' and 'class' in searching can be effective in other ways. Searching with **1 " rat "** results in 797 hits but includes studies on, for example, 'oxidase monamine liver rat'. If only whole animal studies are of interest, this can be narrowed by entering **2 B6A**, which reduces the hits to 144. Reversing the order in this last search gives the same result. Sometimes the purpose of a search is to look for similarities that cross 'class boundaries' and the search is broadened rather than narrowed. For example, one might want to look for similarities in equations dealing with the enzyme oxidoreductase, the organelle microsome, and whole animals. This search would combine: **B2A B3B** and **B6A** to yield 1,357. Now we could isolate QSAR containing a + term for comparative analysis by **15 S+** to find 102 cases. Next move to **show** and enter **/sort=16 1 3 4 15 16 18** and follow this with **S+** on the prompt.

Searching on biological activity presents problems, because there are innumerable ways in which biological activities have been defined. We have not yet attempted to systematically define searchable terms. Still, searching this field can be helpful. For example, using **4 metab** finds 24 examples where the authors referred to their work as metabolism. Of course, there are hundreds of

examples of such studies that have been referred to in other ways: dealkylation, hydroxylation, hydrolysis, etc. Some examples are:

<b>4</b>	<b><u>bind</u></b>	781 hits for binding of chemicals to various bio materials
<b>4</b>	<b><u>uncoup</u></b>	42 hits for uncoupling of oxidative phosphorylation
<b>4</b>	<b><u>Biocon.</u></b>	21 hits for bioconcentration in various ways
<b>4</b>	<b><u>Glutath</u></b>	17 hits in which glutathione was employed
<b>1</b>	<b><u>Glutath</u></b>	28 hits including glutathione transferase
<b>4</b>	<b><u>Inflam</u></b>	13 hits on anti-inflammatory agents
<b>1</b>	<b><u>Cycloo</u></b>	29 hits on cyclooxygenase 1 and 2
<b>1</b>	<b><u>HIV</u></b>	128 hits on human immunodeficiency virus
<b>4</b>	<b><u>Mutag</u></b>	34 hits on mutagenesis
<b>4</b>	<b><u>LD</u></b>	227 hits mostly on LD or MLD plus a few outliers

Combined searches of the 'parameter' category can be fruitful. The following set of commands can be used to look for patterns in electronic and hydrophobic effects in various biological systems:

1.	<b><u>15</u></b>	<b><u>logP</u></b>		3,600 hits		
2.	<b><u>15</u></b>	<b><u>not</u></b>	<b><u>logP'</u></b>	<b><u>**2</u></b>	<b><u>bilin</u></b>	2,443 hits
3.	<b><u>15</u></b>	<b><u>" S</u></b>				2,048 hits
4.	<b><u>15</u></b>	<b><u>not S+ S- SI S' S. ES MR</u></b>				1,623 hits

The second command removes sets based on log P from systems other than octanol or instances where P' is the distribution coefficient for sets of partially ionized compounds. It also removes QSAR nonlinear in log P which makes for easier first time comparisons. Now moving to **show**, two types of comparisons can be made. These are mostly QSAR linear in log P. First, ordering on (S) we find a group of antitumor agents of the aniline mustard type (X-C<sub>6</sub>H<sub>4</sub>N(CH<sub>2</sub>CH<sub>2</sub>Y)<sub>2</sub> with  $\rho$  of about -1.5 to -2. and a small generally negative slope (h) with log P. The crucial factor for reactivity of these agents with DNA is the electron density on N<sup>6</sup>. Set 1780 for the enzymatic acylation of X-C<sub>6</sub>H<sub>4</sub>-NH<sub>2</sub> which is also dependent on the electron density on N has a  $\rho$  of -2. At  $\rho \sim 0.3$  to 0.5 we find examples of uncoupling of oxidative phosphorylation in mitochondria. Other examples of uncoupling occur at  $\rho \sim 2$ . The dependence of

biological activity on and comparison with examples from physical organic chemistry has been recently reviewed.<sup>3</sup> Ordering on log P yields a more complex picture.

Another example might be to check the bio system for possible oxidation of anilines that would likely be associated with  $\text{S}^+$  as follows:

- |    |           |                       |          |
|----|-----------|-----------------------|----------|
| 1. | <u>12</u> | <u>aniline</u>        | 137 hits |
| 2. | <u>15</u> | <u>S+</u>             | 14 hits  |
| 3. | <u>16</u> | <u>-10&lt;S+&lt;0</u> | 11 hits  |
| 4. | <u>2</u>  | <u>B2A</u>            | 7 hits   |

Most of the eleven examples involve oxidoreductases with, possibly, radical mechanisms.<sup>8</sup>

**3. Comparing New QSAR.** One of the most important uses of the searching capabilities of C-QSAR, one which will become of increasing value as the database grows, is that of finding QSAR that might be similar to one from current research. Range searching is useful in saving time inspecting the findings; however, it must be used with care since one does not normally know just how close two coefficients or intercepts (const) must be before one might consider the underlying mechanisms to be the same. To start with a simple example, assume we have just formulated a new equation linear in log P with no other terms except the intercept (const). If the slope of our equation were 0.8 and the intercept 1.7 we might select a comparative group of QSAR as follows:

- |    |           |                            |            |
|----|-----------|----------------------------|------------|
| 1. | <u>15</u> | <u>logP</u>                | 3,600 hits |
| 2. | <u>15</u> | <u>" log1/C _ "</u>        | 2,341 hits |
| 3. | <u>15</u> | <u>not logP' **2 bilin</u> | 573 hits   |
| 4. | <u>16</u> | <u>.7&lt;logP&lt;.9</u>    | 348 hits   |
| 5. | <u>16</u> | <u>1.5&lt;const&lt;2</u>   | 55 hits    |

Command 2 assures us that all sets have the same dependent variable where C is the molar concentration of chemical producing a standard end point. Under this condition we can compare intercepts if the slopes are essentially the same.

One might expect to find a rather uniform set of compounds and actions for the 52 QSAR. But, moving to show and perusing the catch with 1 3 4 reveals a variety of chemicals engaged in a variety of actions. We would term these nonspecific because of the low value of the intercept. An

intercept of 2 says that when  $P = 1$  ( $\log P = 0$ ) a  $10^{-2}$  molar concentration of chemical produces the standard response. For comparison isopropanol has a  $\log P$  of 0.05. Returning to **search** and **blanking** command 5, then replacing it with **16 4<const<10** and **searching** finds only 37 examples with intercepts above 4. These are for rather complex chemicals (compared to alcohols) and the activity would not be considered nonspecific. Studies with quaternary alkylammonium salts are exceptions. These charged molecules appear to disorganize membranes by a rather nonspecific mechanism. Their very low  $\log P$  and high potency results in the high intercept.

Searching for similarities in more complex QSAR is illustrated as follows:

- |    |           |                                 |            |
|----|-----------|---------------------------------|------------|
| 1. | <b>15</b> | <b>" log1/C "</b>               | 3,895 hits |
| 2. | <b>15</b> | <b>PI</b>                       | 485 hits   |
| 3. | <b>15</b> | <b>" S</b>                      | 140 hits   |
| 4. | <b>15</b> | <b>not S+ S- SI S' S. ES MR</b> | 58 hits    |

Quotes have been placed on the first example since there are instances where C is not molar and these have been noted by  $\log 1/C'$ . Quotes have not been placed on PI since the parameter is often position dependent and one might want to consider this information. Adding the stipulation that a term in must be present reduces the catch to a small number of QSAR that can easily be viewed in **show**, without ordering the QSAR on the slopes of PI or , by the entry **1,3,4 16**.

Molar refractivity has often been used as a measure of substituent bulk. It is presumed that the steric effect of substituent bulk should reduce activity, but sometimes bulk has a positive effect. Care must be taken in developing equations to see that MR is reasonably orthogonal with respect to hydrophobic parameters. The following search illustrates how such examples can be isolated for study. The whole biological database can be searched with the following commands:

- |    |           |                           |         |
|----|-----------|---------------------------|---------|
| 1. | <b>16</b> | <b>.15&lt;MR&lt;.36</b>   | 95 hits |
| 2. | <b>15</b> | <b>not **2 bilin</b>      | 71 hits |
| 3. | <b>15</b> | <b>" S</b>                | 19 hits |
| 4. | <b>15</b> | <b>not S+ S- S' SI S.</b> | 12 hits |

Command 1 sequesters QSAR with modest positive MR terms (note that MR has been scaled by 0.1 in C-QSAR). The second instruction eliminates the more complex QSAR and finally we

isolate those equations that also contain terms in  $\log P$ . From the **show** mode, entering **1 3 4 11**, we find that our catch consists mostly of a variety of hydrolases operating on aromatic substrates having substituents (mostly) in the para position that appear to be promoting reactivity by helping to properly position the substrates on the enzymes.

**4. Log P<sub>o</sub> (Optimal Log P).** 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.*,  $\log P - b(\log P)^2$ ) or the bilinear model in which activity normally increases linearly to an optimum and then begins to decrease linearly or level off. Neither set of terms is an ideal solution. The parabola is 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 and then becomes flat. 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 considerable experience in the QSAR field. For instance, it is known that slopes of  $\log P$  and  $\log P^2$  in simple linear equations rarely exceed  $\pm 1.2$ .<sup>3</sup> 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. Searching with **15 logP\*\*2 PI\*\*2** finds 430 examples. This is a small fraction of 6,300 and shows that most researchers have not covered hydrophobic space with their QSAR. This overlooks **logP'\*\*2** and **PI-X\*\*2** cases. Replacing the command with **15 \*\*2** locates 713 QSAR with any parabolic terms (including MR, etc.), while the command **15 bilin** finds 864. Many of the QSAR entered as parabolic equations in the early stages of building the database have not yet been examined to see if they can be improved as bilinear equations. In QSAR where the dependence on hydrophobicity is non linear (parabolic or bilinear) the optimum value for  $\log P$  or  $\log P^2$  is indicated by the subscript o ( $\log P_o$ ). This information is very

important in the design or rationalization of activity of bioactive compounds, and it is instructive to see how this parameter varies with the biological system. The hydrophobic effect is complex and it is determined by at least 3 major factors: first the ease of penetration into simple cells or the 'random walk' inside more complex organisms; second, the interaction with the final receptor; and third, metabolism of the compound, especially by the P450 set of enzymes. For example, while increasing log P may increase the potency of a compound acting on an isolated enzyme, it also may experience metabolic degradation, leaving less material available to reach the active site in the whole organism.

To search the database for compounds having log P<sub>o</sub>, use the following commands:

- |    |                  |  |            |
|----|------------------|--|------------|
| 1. | <b><u>15</u></b> | <b><u>logP</u></b>                             | 3,600 hits |
| 2. | <b><u>15</u></b> | <b><u>logP**2 bilin(logP) bilin(ClogP)</u></b> | 965 hits   |
| 3. | <b><u>17</u></b> | <b><u>1.5&lt;logP&lt;2.5</u></b>               | 88 hits    |

Command 3 narrows the catch to log P<sub>o</sub> values between 1.5 and 2.5. To inspect the results, move to **show** and enter **17**. For parabolic equations, log P<sub>o</sub> is displayed with its confidence limits, when it is possible to calculate them. For the bilinear, log B is given as well as the optimal log P. The confidence limits are found by jackknifing (Section V-G).

One of the advantages of the parabolic model is that an estimate of log P<sub>o</sub> can be obtained without having data points 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<sub>o</sub>. It is instructive to compare log P<sub>o</sub> for QSAR on cells with that on whole animals. Entering **2 B4** finds 1,832 QSAR on all types of cells. Then **14 logP\*\* bilin(logP) bilin(ClogP)** isolates QSAR where log P<sub>o</sub> is found (271). Moving to **show** and entering **3 17** and surveying the results we find that charged compounds (quaternary ammonium and guanidinium analogs) have distinctly lower log P<sub>o</sub>. When these and those without confidence limits and partially ionized acids and bases are omitted, the remaining sets have an average optimum log P<sub>o</sub> of about 4.3. Repeating the process for vertebrates by entering **2 B6A** finds 174 sets where the mean log P<sub>o</sub> value is about 2.8. This is significantly less than that for

cells. We believe that the difference is largely due to the random walk process and metabolism. That is,  $\log P_o$  is an important measure of Bio availability. However,  $\log P_o$  of about 2 is often, but not always, ideal for penetration of the CNS, which may or not be desirable.<sup>16, 17</sup>

This was one of the earliest generalizations of the QSAR paradigm, stemming from the discovery that 16 examples of CNS agents (barbituates, t-alcohols carbamates) acting on a variety of animals had a mean  $\log P_o$  value of 1.98. This figure is close to the above finding of 2.8 for a variety of biological end points. That is,  $\log P$  of about 2 seems to be about ideal for general Bio availability. This figure could be shifted up or down depending on the nature of the receptor and any special metabolic liability. It does not hold for charged or partially ionized compounds. However, it is our belief that one wants to make drugs as hydrophilic as possible commensurate with efficacy.<sup>16, 17</sup> Of course, ascertaining exactly what efficacy is in humans is by no means simple. Short term use of a drug is one thing, but long term use is another. Increased hydrophobicity allows drugs to penetrate into hydrophobic compartments and disrupt a wider variety of processes.

The trend to do screening of potential drugs on cells, rather than animals, makes the selection of rather few compounds for trial in animals difficult. One may be misled to select compounds that are too hydrophobic for ideal results in animals.

#### **IV. Searching the Parameter (THOR-Sigma) Database**

##### **A. Substituent Constants**

The 'preferred' substituent constants (*i.e.*, the most reliable  $\sigma$ ,  $\sigma^+$ , MR, etc.) can be automatically loaded into regression format as a QSAR is being developed, as will be explained in a later section. However, one often wants to check parameter values of one kind or another for a given substituent, especially to see how other methods of determination might have led to values different from those chosen as 'preferred'.

To look up substituent parameter values one accesses the Parameter (THOR-Sigma) database from the \$ prompt by entering **thor**. Respond to the query for database name by entering **sigma**.



For read-only password, press return, and the prompt becomes THOR:sigma>. If, for example, one wanted to scan the parameters for the trifluoroacetamido substituent, -NHCOCF<sub>3</sub>, one enters **find smi \*NC(=O)C(F)(F)F**, where \* indicates the attachment bond (See Section VII for SMILES tutorial.) This returns the statement: 'found record 20 data items'. These can be displayed, under the headings: 'Item' 'Value' 'Label' 'Reference' 'Sel' 'Footnotes', with the command 'type table values' entered as: **t t v**. An asterisk(\*) in the 'Sel' column indicates a preferred value. These are the ones that will be automatically loaded in constructing a new dataset as described in a following section. To save space the Reference and Footnote entries are truncated, but usually they give sufficient information for initial selection. To see the full entry on any item use the 'type item number' command as: **t i 7**, which in this example returns the following:

SMILES	*N(C=O)C(F)(F)F
TIMESTAMP	1990 Jul 13 08:52:52
VALUE	0.30
LABEL	S.META
SEL	*
REFERENCE	Exner,O. & Lakomy,J., Coll.Czech.Chem.Comm., (1970) 35,1371
FOOTNOTE	Apparent pKa of substituted acids (50% Aq.-EtOH or Water @ 25 Deg.)
2	80% Me-Cellosolve and 50% EtOH

To return to the VMS prompt \$ enter **quit**.

**B. Log P and MR:** In the THOR Sigma search described in the previous section, it can be seen that the hydrophobic substituent parameter, PI, and the substituent molar refractivity parameter, MR are listed as Item 5 and 9, respectively. Often it is preferable to have these parameters apply to the entire molecule rather than just the substituent. The calculations and any measurements are both conveniently available through UDRIVE.

From the regression mode, enter **udrive**. The panel that returns on the screen lists a number of helpful suggestions, but the default in the command line is 'S', a prompt for a SMILES entry. (Section VII). Pressing return puts the cursor on the SMILES entry line and if, for example, aspirin is desired, a proper entry, if one elected to begin at the carboxyl group, would be:

**OC(=O)c1ccccc1OC(=O)C**

This delivers a panel with the 2-D structure displayed by entering **DEPICT**. Entry via name is often easier, and to do this the 'help' screen directs one to enter: **F** followed by **2**. The upper line then calls for "name" whereupon **aspirin** can be entered. Entry via 'name' is even more desirable for very complex structures, such as strychnine. Occasionally entry via CAS number is the preferred route, when SMILES is long and the correct nomenclature doubtful, as might be the case with PCB analogs. In the panel's lower section is a summary of: (1) Preferred Measured log P (if available); (2) Estimated Log P; (3) a statement of Error Level; and (4) the CMR (calculated molar refractivity). In the right hand panel is a summary of MASTERFILE data that can be accessed. Entering **T** (in caps) at the command line returns a prompt showing that log P data is going to be shown by default. (One could enter **pka** or **activity** here if desired). The table displays 25 log P values for aspirin measured in 9 solvent systems and at a variety of temperatures and pHs. Entering **t** returns the entire THOR page with all the data. Entering **quit** at any time or **return** at the 'nomore' prompt returns to the main panel, which now has lost the depiction. (One gets it back by entering **p**). Calculation details can be viewed at moderate length by entering **c**, or in 'verbose' form by entering **C**. Details of the Molar Refractivity calculation are seen by entering **M**.

## V. Regression Analysis: Example 1

For several reasons it is best to begin the regression program while in the proper area, either **database bio** or **database phys**. Searching for similar equations can then be carried out directly, and if the developed equation(s) is worthwhile, it is much easier to save it in the proper space.

First enter **regression** followed by **clear** to be sure your workspace is empty. Next assign a name for storage and retrieval; *e.g.*, **name Smith-1**. Next enter the title information as in the following example:

## A. Title Information

T/system **mouse embryo fibroblast cells**

T/compound **X-C6H4-NH2**

T/action **I50 Growth**

T/reference **Harada,A,Hanazawa,M,Saito,J,Hashimoto,K.**

Environ. Toxicol. Chem. **11,973(1992)**

T/source **Name of person entering data**

T/class **B4C**

Now check by entering summary

## B.Naming Parameters

The next step in data entry is to name the parameters which one plans to use. Automatic loading will be demonstrated in this example, and so only the dependent variable need be entered. Enter getp. The program asks for a label for parameter 3. (As noted previously, parameter 1 is reserved for predicted values and 2 is used for deviation). In the present instance log1/C is entered. The prompt is then for parameter 4, but since automatic loading is to be used, just enter end at this point. If parameters other than those in THOR-sigma are being used, such as M.O. parameters or pKa, they would be entered at this point. Often data is not in logarithmic form. They can be entered as such and then converted to log P form by using gettran.

## C. Naming and Entering Substituents

Now enter newsub. This prompts one to enter each substituent label. (In data sets of miscellaneous structures, the whole name, such as ethanol would be entered at this point.) In the present example entering the first label, H, then pressing return, returns a prompt to enter a value for parameter 3:

Label for substituent 1: H

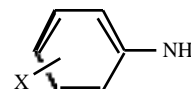
parameter value 3: 2.73

Next it prompts for substituent label 2, etc.

Label for substituent 2: 2-NO2

parameter value 3: 3.28

It is important to note that there must be *no spaces* within the label. *i.e.*, 2,4-di-CL not 2 4 diCl. After all of the labels and parameter 3 values are entered, enter **end**.



No.	Substituent	log 1/C
1.	H	2.73
2.	2-NO <sub>2</sub>	3.28
3.	3-NO <sub>2</sub>	3.44
4.	4-NO <sub>2</sub>	3.49
5.	2-NH <sub>2</sub>	4.85
6.	3-NH <sub>2</sub>	3.92
7.	4-NH <sub>2</sub>	4.68
8.	2-Me	3.83
9.	3-Me	3.72
10.	4-Me	3.70
11.	2-OMe	4.13
12.	3-OMe	3.52
13.	4-OMe	4.55
14.	2-Cl	3.68
15.	3-Cl	3.41
16.	4-Cl	3.89
17.	2-OH	4.82
18.	3-OH	4.08
19.	4-OH	4.70
20.	4-C <sub>2</sub> H <sub>5</sub>	3.64
21.	4-C <sub>3</sub> H <sub>7</sub>	3.46

It is advisable to save data entry frequently by entering **save**. To view Title information, enter **summary**. To view parameter data entry, enter **seedata**. If entry errors need to be corrected, one can enter **editsub** or **editdat**. Additional details on editing are found in Section I which follows.

Often some of the variation of activity with structure is sensed as one enters the data, and hopefully, some idea of the possibly significant parameters can be obtained. In the present example we have well-known electron releasing substituents ortho and para to the amino group which are seen to increase toxicity. Note that the two most hydrophobic analogs, 20 and 21, are not especially potent.

#### D. Entering Structures via SMILES (Section VII)

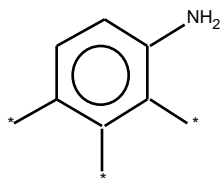
If the data set is not based on a parent structure, then the SMILES for each structure must be entered one at a time and auto-loading of parameters is not possible. In that case, entering **getsmi**

provides a panel with a prompt for structure 1. After the SMILES is entered, 'return' displays the 2-D structure. If the structure needs editing, enter y, if not, enter n and the prompt for the second SMILES appears. Since there are a large number of SMILES stored in the database together with name(s), the name can be entered at this point and the SMILES will be picked up from the database. The name of the compound must be entered exactly as at least one of the stored synonyms, but the program helps one search for 'near-misses'. Common names have been used, *i.e.*, acetic acid not ethanoic acid; use p-chlorophenol not 4-chlorophenol. This can be a very time saving procedure for complex structures such as strychnine. When all of the SMILES have been added enter end or quit.

### E. Auto-Loading of Parameters

With the present set of anilines, automatic loading is to be used, and the parent structure is entered via getsmi /parent. A panel is displayed into which one enters the SMILES with an \* for each substituent position. In the present example substituents are at the 2, 3 and 4 positions, and a proper SMILES for the parent is: **Nc1c(\*)c(\*)c(\*)cc1** or **c1(\*)c(\*)c(\*)ccc1N**.

Note that the asterisks are placed in parentheses which is how SMILES denotes branching from the main pathway. Pressing return should then display:



If the structure is not correct, enter y for editing. Deletions and additions can be made to take place just left of the cursor. When the structure is correct, enter n and the prompt returns to qsar. Next enter getsmi and the panel returns for entry of the first compound. Enter \*H \*H \*H and 'return' to see the unsubstituted aniline. If editing is not needed, then in the panel asking for the second structure enter \*N(=O)=O \*H \*H for 2-nitroaniline. 3,4-dichloroaniline would be entered as \*H \*Cl \*Cl. (Note: If a substituent contains a ring, [*e.g.*, OCH2C6H4-4-Me =

\*OCc2ccc(C)cc2] denote it with a higher number for ring closure than any used in the parent.)

After entering all structures enter **end** (or **quit**) which returns to the qsar prompt.

It is good practice at this point to enter **depict** , which will display all structures for a final check to be sure that the name assigned matches the structure displayed.

Occasionally, it is not possible to enter a SMILES. The structure may not be known for a member of the set. Entering \*\* is necessary to account for the presence of the compound in the set.

For automatic loading of parameters enter **f**etch which then shows: Nlc(\*)c(\*)c(\*)ccl where 1, 2 and 3 refer to the ortho, meta and para positions of aniline.

Enter **1** to load parameters for ortho substituents. A list of labels for all possible substituents is displayed. Since it is not known what kind of electronic effect to expect, we will begin with 'normal' S, S+ and S-, accepting the approximation that, in terms of resonance effects, ortho is much like para. Therefore we enter **15 16 17** to load  $\sigma^+$  and  $\sigma^-$  for ortho substituents. The program returns a report of what it has done. Next enter **f**etch followed by **2**. Now enter **18** for  $\sigma_m$  for substituents in the meta position, with the good assumption that for the meta position  $\sigma_m = \sigma_m^+ = \sigma_m^-$ . For para substituents use **f** and enter **3** and then **15 16 17**. Now view the data matrix with **seed**ata. It is obvious that for each of the parameter types we really need the summation of all the positions, ortho, meta and para, and this is done by the command **get**transformation. With this program, parameters can be transformed in many ways as indicated by the panel after entering **get**trans. For example we might want to test a cross product term as  $\sigma^+ \cdot \sigma^-$  or an exponential term such as  $(\log P)^3$ . Following the instructions in this panel all sorts of transformation can be accomplished. In the present instance we want to combine positional sigma constants. First enter the new label **S** and press **return**. A panel for the transformation appears into which is typed from the upper left column: **S-P-1 + S-M-2 + S-P-3** a space must be present on each side of the +. Enter **seed**ata and the new parameters that cover all 3 positions are shown. (Note that it will be parameter 11 and on following 'screen'.) Next the process can be repeated using **get**t for the other two types of sigmas: **S-P+-1 + S-M-2 + S-P+-3**. and **S-P--1 + S-M-2 + S-P--3**. Now the no longer needed precursors can be removed. Enter **seed** and note

the parameter numbers of **S-P-1**, **S-M-2**, **S-P-3**, **S-P+1** etc. The use of the delete command solves the problem: **del /para** followed by the parameter numbers to be deleted (**4 5 6 7 8 9 10**). Parameters cannot be deleted if there is a stored equation. **del /eq** removes it and the parameters can now be deleted.

Hydrophobic parameters must be considered from two points of view: global (log P) and local position dependent ( ). There are two options for automatic loading of the global log P once the SMILES have been entered. Entering **add mlogp** loads all experimentally determined values from a carefully evaluated list of over 10,000. Of course this is relatively *very small* compared to all possible organic compounds, but sometimes for common chemicals (such as phenols or alcohols) all desired values can be obtained. Generally the command **add clogp** is used and the values are calculated.<sup>19</sup> It is extremely important to remember that both add MlogP and add ClogP supply log P for the **NEUTRAL** form of the compound. Also it must be remembered that PI ( ) values are from benzene. These will not be very accurate when the substituents are near a group with a strong electronic effect (*e.g.*, a NO<sub>2</sub> group or a N or O atom in a heterocycle). It can be of value to add both MlogP and ClogP for comparison to see how well ClogP works for a particular class of compound.

Although ClogP is calculated for the neutral form of the solute, these values are often satisfactory for a set of partially ionized compounds if there is not a great range in the degree of ionization. Of course even if one obtains a reasonable QSAR, the intercept would differ from that where the apparent log P (log D) was used instead and log P<sub>o</sub> would need correction. Also using a sigma constant can correct for the degree of ionization.

Now entering in order, **f**, then **1**, followed by **7 11 12 13**, parameters to check on possible steric effects of ortho substituents. The data set should now be saved by entering **save**. It can be reloaded by entering **load Smith-1**.

## **F. Permuting**

At this point the permutation routine is valuable for uncovering the most important variables. Enter **seep** to view parameters:

1      2      3      4      5      6      7      8      9      10      11  
 YPRED   DEV   LOG1/C   S   S+   S-   MLOGP   ES-1   L-STM-1   B1-STM-1   B5-STM-1

All of the steric parameters are labeled 1 when in fact, they are for ortho substituents. These should be edited to ES,2, L-2, B1-2 and B5-2. One must use the dash and not a comma for the sterimol parameters (not B1,2) a comma is used with ES; otherwise string searching finds S-. To edit enter **editpara 8**. The label is displayed and the 1 is deleted and replaced by **2**. The same process can be used for parameters 9, 10 and 11.

Entering **3 perm 4 5 6 7 8 9 10** derives equations for all possible linear combinations of the eight parameters and then displays the results in tabular form . One could also explore the parabolic and bilinear forms of log P by using P7 or B7 (case is important).

#### 1 TERM REGRESSIONS

S.D.	MLOGP	S+	S	S-	ES-1	L-STM-1	B1-STM-1	B5-STM-1	CONST
1 .344		-.73							3.73
2 .417			-.95						3.88
3 .447	-.44								3.91
4 .459				-.65					4.37
5 .571								.17	3.65
6 .578						0.9			3.66
7 .579					0.7				3.90
8 .580							.11		3.76

#### 2 TERM REGRESSIONS

S.D.	MLOGP	S+	S	S-	ES-1	L-STM-1	B1-STM-1	B5-STM-1	CONST
1 .298		-1.83	1.72						3.51
2 .316		-1.03		.74					3.58
3 .331	-.18	-.61							3.95
4 .352		-.73					.14		3.57
5 .352		-.74			-.06				3.71
6 .353		-.73						.04	3.68
7 .353		-.73				.03			3.67
8 .394	-.24		-.71						4.14
9 .411	-.28			.48					4.12
10 .412			-2.12	.95					3.83



### 3 TERM REGRESSIONS

S.D.	MLOGP	S+	S	S-	ES-1	L-2	B1-2	B5-2	CONST
1 .271	-.200	-1.753	1.803						3.751
2 .300	-.175	-1.174		.731					3.803
3 .300		-1.979	1.930			-.102			3.730
4 .301		-1.978	1.916					-.092	3.609
5 .305		-1.877	1.795		.042				3.518
6 .306		-1.869	1.772				-.076		3.594
7 .306		-1.857	1.853	-.084					3.513
8 .321		-1.397		.847				-.084	3.672
9 .322		-1.377		.831		-.080			3.756
10 .322		-1.376		.855	.077				3.587

The clue to the relative value of the various parameters is the standard deviation (S.D.). From this it is seen from the 1 term regressions, that  $S^+$  is the single most important parameter. The coefficient with this term is -0.73 and the constant for the equation is 3.73. Now considering the 2 term regression, S is the next most important parameter. However, note that the coefficient with  $S^+$  has changed greatly ( $S^+$  and  $S$  are not independent of each other). The most solid parameter is  $S^+$  which occurs in 7 of the 10 equations. The other parameters show no consistent pattern. Considering the 3 term regressions, MlogP is the next most important parameter. All of the coefficients with the steric parameters are very small and there is no pattern in their entry into the equations. Looking at the set of 4 term regressions (not shown) we see that there has been no reduction in the size of the standard deviation, hence no improvement in the correlation.

We can now check for collinearity of the parameters by entering **corr 4 5 6 7 8 9 10 11** to obtain the correlation matrix values of r between the various parameters.

	MLOGP	S+	S	S-	ES-1	L-2	B1-2	B5-2
MLOGP		.540	.537	.490	-.218	.164	.200	.078
S+	21		.967	.921	-.168	-.085	.019	-.177
S	21	21		.975	-.243	.013	.096	-.075
S-	21	21	21		-.337	.068	.151	-.001
ES-1	21	21	21	21		-.758	-.884	-.751
L-2	21	21	21	21	21		.885	.956
B1-2	21	21	21	21	21	21		.807
B5-2	21	21	21	21	21	21	21	

The upper portion of the matrix gives the correlation coefficient  $r$  for the correlation between each of the various parameters, while the lower portion indicates how many data points the comparisons are based on. Note that the correlation between the three parameters is very high. To avoid this problem care must be taken in the design of the congener set before starting the project (Section V-K) or at least to correct it as the project proceeds. At this point it would be a mistake to use two terms. Using  $\log 1/C$  and the next best parameter MlogP, enter **3 reg 4 5** to obtain:

$$\log 1/C = -0.18 (\pm 0.24) \text{MlogP} - 0.61 (\pm 0.30) + 3.95 (\pm 0.33)$$

$$n = 21, \quad r^2 = 0.693, \quad q^2 = 0.615, \quad s = 0.331$$

This is not a very good equation; it accounts for only 69.3% of the variance in the data ( $100 \times r^2$ ).

### G. Jackknifing

At this point it would be good to know if there are any outliers among the data points. These might be due to the inadequacy of the model, an error in entering data (if it was done manually) or an experimental error in determining the activity of one of the analogs. Another major cause of outliers that is very difficult to understand are side reactions yielding a different kind of biological response. In some instances, it has been possible to resolve those problems.<sup>9, 23</sup> To carry out 'jackknifing' one enters: **3 j 4 5**. In this operation 21 regression equations are derived dropping in each instance a different data point. The best equation is the one with the highest  $r^2$  after the dropping of one point. From this operation we obtain:

<u>Omitted</u>	<u>R<sup>2</sup></u>	<u>S</u>
none	0.693	0.331
1 H	0.849	0.211
16 4-Cl	0.713	0.329
18 3-OH	0.702	0.334
6 3-NH <sub>2</sub>	0.701	0.336
4 4-NO <sub>2</sub>	0.699	0.337

Under omitted, none means all 21 data points were included and with the two variable equation,  $r^2 = 0.693$  as shown above is found. Dropping the parent compound, aniline, yields a QSAR with  $r^2 = 0.849$  and S.D. = 0.211.

To delete data point 1 use the command **star /add 1**. This places an asterisk on 1 and it is not used in deriving equations. Any number of points can be withheld in this fashion. **star /a 1 5,10 18** would withhold points 1, 5 to 10 and 18. **star /delete** can be used to remove all asterisks in the reverse manner. Or **star /d 5** could restore data point 5 (remember that case is *not* important in entering commands).

Starring 1 yields the following QSAR:

$$\log 1/C = -0.22 (\pm.15) M\log P - 0.55 (\pm.19) + + 4.07 (\pm.22)$$

$$n = 20, \quad r^2 = 0.849, \quad q^2 = 0.783, \quad s = 0.211$$

We can explore the possibility that a three variable equation might be worthwhile. Enter **3 reg 4 5 7** which yields:

$$\log 1/C = -0.23 (\pm.12) M\log P - 1.36 (\pm.51) + + 1.26 (\pm.77) + 3.91 (\pm.20)$$

$$n = 20, \quad r^2 = 0.914, \quad q^2 = 0.868, \quad s = 0.211$$

Although this equation has a significantly better  $r^2$ , we would not recommend accepting it until some kind of support for such a model can be found from physical organic chemistry. The object of the game is not to get the highest  $r^2$ , but to get understanding.

One might suspect from the opposite signs of the two terms that  $\log 1/C$  depends parabolically on the electronic effect of substituents. This can be checked by performing the regression **3 reg 4 p5** which yields a correlation with  $r^2 = 0.886$  or **3 reg 4 p6** which gives  $r^2 = 0.781$ . Placing a P before any parameter number derives an equation with this term in parabolic form while placing a b (*e.g.*, b5) yields a result with the parameter in bilinear form (ref 1, page 195).

Further insight can be gained by plotting the data. Any two parameters can be plotted against each other by the command **a graph b** where a and b are the parameter numbers. Entering **3 graph 5** it is seen that the 2-, 3- and 4- nitro points fall off the straight line. Deleting these points by **star /a 2 3 4** and then comparing the results shows that **3 reg 4 5** gives just as good an  $r^2$  as **3 reg 4 P5**. More work is needed with strong electron withdrawing substituents, such as CN, to see if the problem is electronic or if the  $\text{NO}_2$  might be undergoing reduction in the cells. It is

readily reduced to the toxic nitrogen group. We have also found in possible radical reactions of phenols that strong electron withdrawing substituents block the reaction.<sup>22</sup>

At this point we need to look for support from other QSAR. Enter **database Bio** or **Phys** and then **sea** to get the search program. To find other comparable Bio QSAR enter **15 S+** and **sea**. Narrowing the search to equations having coefficients near ours by searching on **16  $-.75 < S+ < .45$**  reduces the list to 59 having values of (slope) bracketing that found in our 2-variable equation. Enter **show** and **15** to inspect the results. So doing we find a number of QSAR having parabolic or bilinear terms in S+. These can be deleted by returning to **search** and entering **not S+\*\*2 bilin (S+)** leaving 54 with linear S+ terms. Moving to **show** and entering **/sort=16 1 3 4 15 16 18** and then S+ on the prompt gives a list ordered on increasing values of .

Equations of interest are found for developmental effects of phenols to rat embryos in culture, phenol toxicity to *P. aeruginosa*, phenol toxicity to *P. phosphoreum*, phenol complexing with NAD and phenols inhibiting Listeria. These QSAR contain only a term in  $S^+$  for which the values are: -0.56, -0.58 and -0.65. Thus we find the developmental (toxic) effect of the phenols to depend on  $S^+$  in a parallel manner to the toxicity of the anilines to mouse embryo cells. A fourth QSAR is found for phenols, the endpoint measured as a 10g weight loss in rats. Weight loss in pregnant rats shows parabolic dependence on logP and contains a  $S^+$  term with  $\beta = -0.61$ . Adding  $\log P + (\log P)^2$  to the rat embryo equations also makes a slight improvement, but with only 11 data points, one hesitates to rely on such an equation. More data (20) were available for the weight loss correlation. Of course, the above results do not *prove* that anilines and phenols operate by similar mechanisms, but they are suggestive.

We can now turn to the physical database for further insight. Enter **reg** to go to qsar and then **data phys** when asked for the password press **return**. This returns the main menu. Enter **sea** for the search mode. We now repeat the routine used for the Bio search. Entering **15 S+** locates 1,656 sets. **15 not S+ \*\*2 bilin (S+)** removes the non linear equations, leaving 1,579 examples. By searching on **16  $-.75 < S+ < -.45$**  this is reduced to 136 sets. These can be perused in the **sh** mode by the entry **/sort=16 1 3 4 15, 18**. However, since we suspect a radical

reaction the 136 hits can be further reduced by **2 P12** which yields 50 cases. Viewing these we find one of special interest: radical abstraction of •H from phenols (X-C<sub>6</sub>H<sub>4</sub>OH + (CH<sub>3</sub>)<sub>3</sub>CO• → X-C<sub>6</sub>H<sub>4</sub>O• + (CH<sub>3</sub>)<sub>3</sub>(OH)) with a term of -0.71. In fact, there are many examples of H abstraction by radicals which are correlated by  $\rho$  of which the following are representative for C-H bonds.

Compound	Reagent	$\rho$
X-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub> CH=CH <sub>2</sub>	Br•	-0.75
X-C <sub>6</sub> H <sub>4</sub> CH <sub>3</sub>	Cl•	-0.73
X-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	Br•	-0.72
X-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	Br•	-0.62

## H. Cross Validation

The value of cross validation in avoiding poor quality or meaningless QSAR has been discussed by Cramer et al.<sup>24</sup> We have included two ways of viewing the problem. The program automatically calculates a cross validated  $r^2$  (Q<sup>2</sup>) and prints this out with each equation. One wants to see the normal  $r^2$  and Q<sup>2</sup> in 'reasonable' agreement, but exactly what amounts to reasonable is a matter of judgment. If they are not close (say  $r^2 = .90$ , Q<sup>2</sup> = .7) then jackknifing to remove one or more points may be worthwhile.

Q<sup>2</sup> is basically a kind of warning that the QSAR may not be as good as implied by  $r^2$ . In some fashion the data are not well fit. If this is due to an outlier or two, jackknifing can be used to find the deviant point(s). More complex problems, such as nonlinearities in the model may be hard to find.

In another approach one can set one's own standard. For example, in the above case of the anilines using the command **3 cross 4 5** randomly drops 10% of the data points in 10 trials and displays the QSAR. In these examples  $r^2$  ranges from 0.807 to 0.896.

One can also set an arbitrary percentage drop: **3 cross /omit=30% 4 5** which yields 6 QSAR with  $r^2$  from 0.803 to 0.896. In this example, 30% of the data points were randomly dropped. Any percentage can be selected by inserting the proper number along with %.

In the above examples it is found that the coefficients with  $M\log P$  and  $\dots$  vary but little from our first QSAR with only one point omitted.

### I. Editing

There is always the need to change data values or correct errors, and the most general way to do this is by entering **editset**. This displays all the data and puts one into a general edit mode allowing: (1) movement of the cursor by the arrow keys, (2) the use of the delete key to erase the character to the left of the cursor, and (3) the insertion of new characters at that spot. Of course a completely new variable could also be added by first assigning it a symbol and then entering the values. When finished, one must exit this edit mode, using **control Z**. When making minor changes a quicker mode is available. Entering **editdata** prompts for the substituent number. Entering that number then prompts for the parameter number, whence the current value is displayed. When this is edited, a box for the new parameter value is displayed. After editing is complete, the command **seedata** enables one to check the results. Other editing choices are shown after entry of **edit**.

After experimenting with a wide variety of parameters including squared and bilinear ones which have been added to the data matrix, it is convenient to use the delete command to clean up the set before saving it. However parameters cannot be deleted if an equation has been saved. To check for saved equations, enter **eq /run**. All equations can be deleted by **del /eq #**. Now entering **seep**parameter the parameter numbers are displayed. The command **del /para 4 8 16** would remove these parameters. Sometimes it may be necessary to delete a data point and the information associated with it including the SMILES. The command **del /sub #** removes the substituent and its SMILES with that number.

### J. Regression Analysis: Example 2

We now consider a data set (Bio #1557) based on a study of benzanilides ( $X-C_6H_4CONHC_6H_4-R$ ) inhibiting mitochondria by G. A. White. The set can be loaded for study when in the Bio mode by entering **load /d 1557**. Entering **summary** gives a view of the set including the parameters that were studied in the derivation of the equation. Entering **eq /run**

displays the stored equation. Entering **predict** shows the parameters used to derive the equation, substituents and calculated values. In the summary under compound it is indicated that substitution has been made on both rings; one labeled X and the other R. **Depict** 2 pictures the 2-D structures which can be paged through by pressing **return**. A particular structure or set of structures can be viewed by **depict #** or **depict 1 3,10**; for example.

Describing how to go about formulating a QSAR is rather like giving directions for solving a jigsaw puzzle. There are many ways to go about it and they depend on clues one gets from inspecting the scene. In the case of the QSAR one does not usually have all of the pieces to unambiguously picture the roles of steric, electronic and hydrophobic properties of all positions on a parent molecule.

A good first move is to check for a role for hydrophobic effects since 50 to 80% of the Bio QSAR depend on log P or on    at some position on the parent. In the present case entering **3 graph 4** one sees not only a strong nonlinear relationship, but evidence for a bilinear effect. By starring various points one can check up on the behavior of subsets of the data. For example we might compare each of the two rings independently. First enter **star /d** to free all data points. Then entering **3 reg B4** yields the bilinear relationship:

$$\log 1/C = 0.99 (\pm.34) \text{ ClogP} - 1.73 (\pm.74) \text{ bilin} (\text{ClogP}) + 1.41 (\pm 1.2)$$

$$n = 34, \quad r^2 = 0.542, \quad s = 0.682 \quad \text{optimum} = 5.62, \quad = -5.49$$

Activity initially increases with increase in log P with slope (h) of 0.99 and then falls off linearly with slope of -0.74 (0.99 - 1.73). The optimum (log P<sub>o</sub>) listed above is actually rapidly calculated from the parabolic equation. To get the exact value enter **3 j B4**. The optimum shown here is 5.19.

The above bilinear term actually represents:  $-1.73(\pm.74)\log( \cdot 10^{\text{ClogP}} + 1)$  where  $= -5.49$ . Generally researchers use the form  $\log( \cdot P + 1)$ . Since many of our parameters are on a log basis ( , , E<sub>s</sub>, MlogP) or are calculated as such ClogP, it is simpler to use  $\log( \cdot 10^x + 1)$  where X is in logarithmic form.

Note that the Dev + and Dev - which represent the number of points below and above the regression line are not badly out of balance. Also the 95% confidence limits on the coefficients are not unreasonable. Although the correlation is not high in terms of  $r^2$ , it seems promising. Next we need to see the effect of adding the steric and electronic terms. Inspecting the substituents we see only one example (#15) that contains a strong electron withdrawing via resonance substituent. Hence it is not surprising to see the QSAR does not contain a  $\sigma$  term. It is of interest to see that there is no electronic term for X-substituents. It is quite possible that the person entering this set or any other of the sets in the database may have overlooked something. The best way to consider this point is to study X substituents alone. Enter **star /a 8**, which stars all substituents above 7. Now enter **3 perm 4 5 6 7 14 15 16**. The most important variable is Es - X<sub>2</sub> followed by Clog P. A very small lowering of the standard deviation occurs on the addition of B1-X, but a 3 parameter equation is not justified by 7 data points. The two variable equation is:

$$\log 1/C = 0.73 (\pm.47) \text{ ClogP} - 1.24(\pm.32) \text{ Es-X}_2 + 0.97 (\pm 1.4)$$

$$n = 7, \quad r^2 = 0.973, \quad s = 0.113 \quad q^2 = 0.923$$

At this point recall that all substituent values of Es are negative, that is, the Es term implies that bulky substituents are good! Note also that all of the log P values for this subset are much below log P<sub>0</sub> of 5, so that linear dependence on ClogP is expected. Deriving single parameter equations in ClogP and ES-X<sub>2</sub> one finds that the latter is more important. The coefficients are in reasonable agreement with the stored QSAR.

Since there is significant variation in the properties of X, the fact that useful electronic terms are not found might be due to a collinearity problem. Enter **corr 4 5 6 7 14 15 16** which provides the correlation matrix of r values between the various substituents. The correlation (r) between ES-X<sub>2</sub> and the electronic parameters is not high. Entering **3 reg 4 14** shows that, indeed, electronic effects do not appear to be significant. The substituents studied in this position are rather small. Our results suggest that using larger hydrophobic substituents would yield more potent compounds as long as we do not exceed log P<sub>0</sub> of 5 for the whole molecule. It would be interesting to study the isopropyl and t-butyl groups.



Caution must always be used when working with the bilinear terms. Unless there is a good range in the values of the dependent variable unreasonable coefficients may be found for these terms. Sometimes the numbers are so large the problem is obvious. In the case of log P or a large amount of experience teaches that the coefficient (h) is rarely outside of the range of  $\pm 1.2$ .<sup>3</sup> Large negative slopes (more negative than -1.2) are sometimes due to steric effects of large substituents when log P and steric terms are collinear. The big advantage of the bilinear model is that one can compare the slopes, especially the initial slope, with other QSAR which are only linear in log P, for example.

Now to consider the other ring enter **star /d** followed by **star /a 1,8 32**, this provides a set in which all substituents on the X-ring are constant having only a 2-methyl group. Entering **3 perm B4 8 13 17 18** finds only the parameters of the stored equation to be significant and entering **3 reg B4 5 8 13** yields a very similar equation to that stored, but with a lower  $r^2$ . Destarring all data points and then removing two by the jackknifing procedure produces the stored equation.

Considering this equation what should the next move be? Obviously the negative sign with B1-R2 shows that ortho substituents in this ring are bad. Surprisingly no para substituents were tested. The negative coefficient with (S,R) shows that electron releasing substituents are beneficial. One of the best would be 4-NH<sub>2</sub> ( = -0.66) or 4-OH( = -0.37).

If further testing showed that the 2-isopropyl or 2-t-butyl groups on the X-ring increased potency, then one of these could be combined with the 4-NH<sub>2</sub> in the R ring. In addition, it would be necessary to add alkyl or alkoxy groups in the 3-position of the R ring to obtain a log P value near 5 for maximum potency.

To gain experience, the beginner can enter any dataset by **load /d #** where # is the set number. Unfortunately parent SMILES have only been provided for the more recent entries. We plan to correct this deficiency (above #3500 in physical databank and above #3000 in the Bio databank). Once the set has been loaded one can explore it as we have done with the above example.

## K. Substituent Selection in Molecular Design

A critical problem in undertaking a program to modify a parent compound in a structure-activity study is the selection of suitable substituents.<sup>1</sup> Not only does one need convenient access to the widest possible variety of substituents, but it is time saving to be sure if a particular substituent is selected that it has a range of known substituent constants ( $\rho$ ,  $\rho^+$ ,  $\rho^-$ ,  $\sigma$ , MR, etc.). At the start of such a process it is generally not possible to anticipate which parameters will be significant in the final QSAR. To utilize this feature of the C-QSAR program there are two commands that can be entered from the regression mode: **parameter**, **parameter /nolimit**. Entering **parameter** produces a table of labels (same as the **fetch** command). Any group can be selected. For example, entering **15 16 17** ( $\rho$ ,  $\rho^+$ ,  $\rho^-$ ) the prompt asks for the minimum and maximum to set a range for each parameter. We might enter **0 1 0 1 0 1** to define the range for each of the three parameters. After the data have been collected, enter **seed** and we find that only 21 substituents fall in this range.

In selecting a set of substituents for combinatorial synthesis it can be very important to limit the range of  $\rho$  or possibly other parameters. Thus for  $\rho$ , one might set limits such as -1 to 3 for  $\rho$  and let values for the others fall where they may by simply pressing return to pass the options by. If one had pressed return at each prompt so that no limit had been set and then entered **seed**, 79 substituents are found. Notice that when entering more than one parameter, the data are ordered on the first parameter called for. One can stop the listing at any point by entering **q**.

Entering **parameter /nolimit** asks for no limits. This saves time in specifying limits when many parameters are requested. Entering **seed** shows the same 79 substituents. One can stop the listing at any point by entering **q**.

Entering **parameter /nolimit** and then selecting **1 2 12 15 18** followed by **seed** produces a set of 256 substituents ordered on increasing values of  $\rho$  (first parameter selected) all of which have  $\rho$ , MR, B1,  $\rho$  and  $\rho_m$  values. Since this operation is in the regression mode these substituents are loaded for regression. Enter **seep** and we find the parameter labels. Entering **3**

**reg 4** we find  $r^2$  for the correlation between and MR ( $r^2 = 0.127$ ). Entering **corr 3 4 5 6 7** we obtain the following correlation matrix for the set of 5 parameters in terms of r.

	P1	MR	B1	S-P	S-M
P1		.357	.161	-.225	-.275
MR			.292	-.052	-.106
L				.343	.250
S-P					.918

The above approach may be sufficient, but especially in Combinatorial Synthesis one wants to be sure that data space is well explored and collinearity minimized. To deal with this problem enter **parameter /E /P** (E denotes euclidian space and P stands for pick). From the displayed table select parameters of interest and all substituents having all of these parameters will be sequestered. In the present instance enter **1 11 12 13 15 18** for , B1, B5, L,  $p$ . These are entered in regression mode. Enter **seep** this shows euclid P1, B1, B5, L, Sp with their numbers. Entering **Corr 4 5 6 7 8** yields the correlation matrix. Entering **seed** lists the substituents in order of increasing distance from H. The collinearity between any two substituents can be checked. **seep** provides the parameter numbers and **4 reg 8** shows that the collinearity between and  $p$  is almost 0 for 294 substituents. Next, enter **star /A** which stars all substituents. Now use **seed** to peruse substituents picking the numbers of those of interest in terms of their ease of synthesis and distance in euclidean space. Then enter **star 1d** followed by the number of the selected substituents. Now enter **corr 4 5 6 7 8** or any combination of these to explore the collinearity among the selected substituents.

Another way to select the most effective substituents to break collinearity enter **parameter /nolimit**. Next, select **15** and **16** from the table of parameters. Since you are now in regression mode, **3 reg 4** yields the correlation equation for two terms and also informs one as to the number of substituents having both and + (199). Next, use **gett** and enter **S+-S** for the name of the new variable. This then is defined on the prompt by entering **S-P+ S-P**. This becomes new parameter 5. The command **sort /null /abs /des** and then **5** on the prompt orders the

substituents in decreasing size of difference. Of course one could treat any number of parameters in this fashion.

The default approach can be had by the command **parameter /nolimit /E**. This selects automatically , MR, L, B1, B5,  $\rho$ ,  $\rho^+$ ,  $\rho^-$ . **seed** lists only 61 substituents having all of these parameters. The reason for this is that there are relatively few values for  $\rho^+$  and  $\rho^-$ . However, these electronic parameters are very valuable in providing mechanistic insight via QSAR<sup>3, 8, 9, 47</sup> and should be studied separately. There is a high degree of collinearity among them. In a way this is helpful as it means that  $\rho$  can detect the presence of electronic effects that later can be explored with the other two parameters.

## VI. Searching for New Leads

The most difficult and important aspect of medicinal chemistry is finding new lead compounds for drug development. Strange to say, they are all around us, and yet almost impossible to recognize. A good example is the conversion of nalidixic acid, a mediocre antibiotic, into a quinolone carboxylate. QSAR played an important role in this process.<sup>11</sup> The old estrogenic sedative thalidomide now looks promising for treatment of leprosy and multiple myeloma. Improved "me too" drugs are always coming into the market. The great excitement at present is combinatorial synthesis. Of course with our present C-QSAR system, one can select any subject and study it for possible clues to increase potency or decrease toxicity. Since the vast majority of the bio QSAR in our system were developed by us from data published by others in which no attempt to formulate a QSAR was made, it is highly unlikely that the most active congeners were discovered. A selective approach is to find QSAR that cover highly active compounds. The dependent variable  $\log 1/C$  has been entered in molar terms whenever possible, hence  $\log 1/C = 9$  means activity at  $10^{-9}$  molar concentration.

A selective approach is to consider QSAR that cover highly active compounds. Two search modes are possible:


1. **14 log1/C>9** makes 2 hits. In these instances, all members of the set have log 1/C values > 9.
2. **14 log1/C@max>9** finds 243 data sets. In this case, any set having a congener with log 1/C greater than 9 is selected. Lowering the standard to 8, we find 17 and 708 sets respectively for the two searching modes. At log 1/C of 7, we find 118 and 1,334 sets. Searching the entire bank with **15 " log1/C "** 3,906 sets are characterized by molar log 1/C terms.

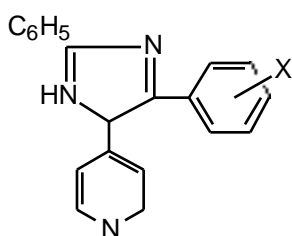
These approaches can be generalized in the following ways:

- |                                      |            |
|--------------------------------------|------------|
| 3. <b><u>14 log1/C&lt;2</u></b>      | 866 hits   |
| 4. <b><u>14 2&lt;log1/C&lt;4</u></b> | 212 hits   |
| 5. <b><u>14 logP&gt;6</u></b>        | 24 hits    |
| 6. <b><u>14 logP@max&gt;6</u></b>    | 986 hits   |
| 7. <b><u>15 not logP**2</u></b>      | 5,940 hits |
| <b><u>15 not bilin(logP)</u></b>     | 5,760 hits |
| <b><u>15 not bilin(ClogP)</u></b>    | 5,414 hits |
| <b><u>15 not logP'</u></b>           | 5,232 hits |
| 8. <b><u>14 logP@max&gt;6</u></b>    | 986 hits   |
| <b><u>14 6&lt;logP&lt;8</u></b>      | 1 hit      |

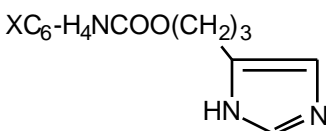
The above table lists the number of different examples and the range of some of the more common classes of substituents that provide all sorts of opportunities for modification and testing of a known QSAR.

As the database increases, it becomes more arduous to find exactly what one wants to study. There are several ways to make searching more selective. Besides setting a high level of activity one can limit the search to biological areas of interest. For instance, we might search on organelles and cells by entering **2 B3 B4**. This makes 2,232 hits, about 1/3 of the database. Next, selecting the cutoff at log 1/C = 7 by **14 log1/C@max>7** reduces the catch to 584 sets. Going to **show**, we see that these are listed in order of increasing size of set number. To peruse recent work, search

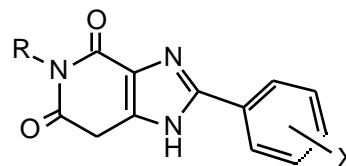
with **5** (1990) (1991) (1992) (1993) (1994) (1995). This reduces the number to 211. A quick check of the catch finds many sets for HIV. To eliminate these enter **1 not HIV**. This cuts the number to 158. Once a compound of interest has been found, the MERLIN search can be employed. Going to **sea** and entering **13** and then on the prompt entering the SMILES for  and then searching, locates 172 sets that contain one or more compounds having this moiety. Moving to **sh** and entering **1 3 4** the main features of activity can be scanned in a few minutes. Some sample hits are as follows:



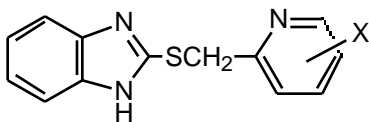
$I_{50}$  C-RAF  
Protein Kinase



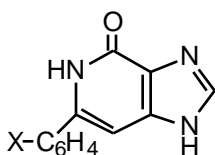
Antagonistic activity  
vs. synaptosomes  
Rat cerebral cortex



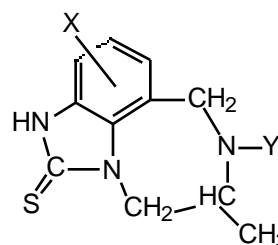
Displacement of  
cyclohexyladenosine  
from rat cortex receptor



Mic vs. *Helicobacteria Pylori*

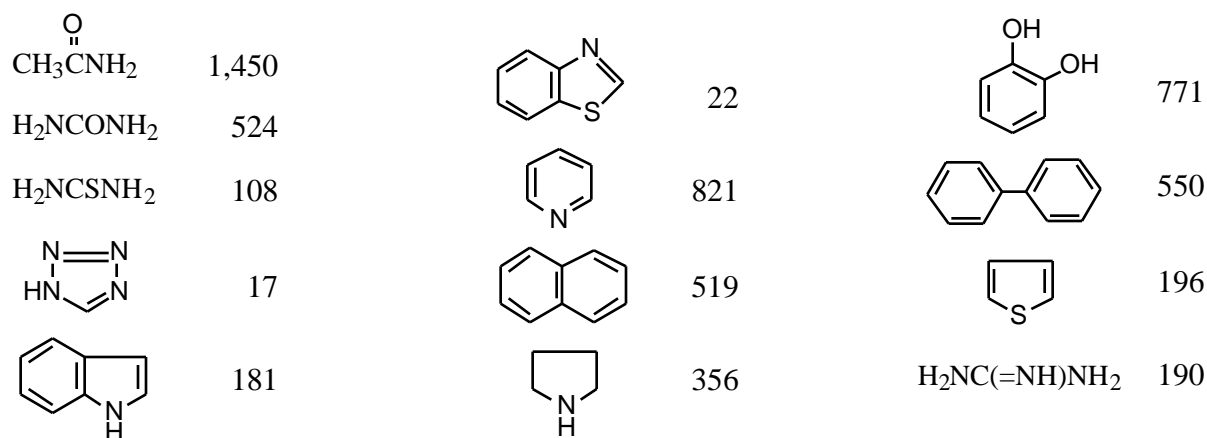


$I_{50}$  Thymidine Kinase



$I_{50}$  to protect MT-4 cells  
from HIV

Searching on the following structures finds the indicated number of hits for sets that contain one or more compounds having the indicated feature.




Searching can be made much more specific as illustrated in the following examples. After sequestering 821 pyridines, we can narrow the search to more specific points of interest. Using **2** **B2** finds 356 QSAR on enzymes and receptors. Finally, **15 S+** narrows this to 6 that contain <sup>+</sup> terms.

**Table 7**

	range	number of substituents
	-5.96 to 10.3	1088
0.1MR	0.09 to 12.3	1536
<sub>p</sub>	-1.58 to 2.42	1999
<sub>m</sub>	-0.53 to 2.18	1234
<sub>p</sub> <sup>+</sup>	-7.2 to 1.88	397
<sub>m</sub> <sup>+</sup>	-1.15 to 3.20	75
<sub>p</sub> <sup>-</sup>	-0.82 to 5.40	375
<sub>m</sub> <sup>-</sup>	-0.33 to 2.25	92
*	-5.40 to 7.6	887
<sub>i</sub>	-0.96 to 1.53	998
<sub>p</sub> <sup>•</sup>	-0.08 to 0.90	26
<sub>m</sub> <sup>•</sup>	-0.12 to 0.03	13
E <sub>s</sub>	-8.68 to 0	249
B1	1 to 4.65	1078
B5	1 to 12.17	1077
L	2.06 to 17.2	1083

Because of the relatively few examples of  $m^+$  and  $m^-$  we have normally made the assumption that  $m = m^+ = m^-$ . We can now obtain some idea of how valid this is. Using **parameter /nolimit** and entering **18 19** via **seed**, finds 45 instances where values for both  $m$  and  $m^+$  substituents are known. Enter **seep** to get parameter number. Next **3 reg 4** shows that  $r^2 = 0.796$ . Using **3 j 4** and starring the bad points we can eliminate three outliers:  $O^-$ ,  $N^+(Me)_3$ , 3-thienyl. Correlation of the remaining 42 finds  $r^2 = 0.915$ ;  $s = 0.113$ . Note that by using the regression 'ploy' one can quickly ascertain the number of substituents that have values for any pair of parameters (in this case  $n = 45$ ). If one had used **seed** and retrieved several hundred, it would require scrolling through to find matching pairs.

Repeating the process for  $m$  and  $m^-$  it is found that 76 substituents have both parameters and the correlation between them has  $r^2 = 0.930$ . Omitting  and  $NH_3^+$  the remaining 75 examples correlate with  $r^2 = 0.962$ ,  $s = 0.084$ .

At present there are 366 substituents for which both  $\sigma^*$  and  $\sigma_I$  are known. Removing the four most poorly correlated substituents ( $-(CH_2)_3P(=O)(OEt)_2$ ,  $SO_2CH=CH_2$ , 9-carbazolyl and  $B(OH)_2$ ) we find  $r^2 = 0.911$ ,  $s = 0.073$ .

The most useful  $\sigma^*$  or  $E_R$  values for radical reactions are  $\sigma^*$  for Creary and  $E_R$ , simply because more of these are available. Values are also given for  $\sigma^*$  determined by Arnold and by Jiang.

Sterimol parameters (B1, B5 and L) are known for a number of substituents that have cis and trans forms. At present, for technical reasons, we can not enter values for both, hence we have only given values for the trans forms. The cis values can be found in ref. 2.

We have greatly enlarged the number of  $\rho_p$  and  $m$  values in our system by means of the following correlation equations obtained from experimental values for  $-C_6H_4-4-X$  and  $-C_6H_4-3-X$ .

$$(-C_6H_4-4-X) = 0.310(\sigma_p) + 0.026(\sigma) + 0.009$$

$$n = 35, \quad r^2 = 0.956, \quad s = 0.026$$

$$(-C_6H_4-3-X) = 0.238(\sigma_m) + 0.012(\sigma) + 0.006$$

$$n = 23, \quad r^2 = 0.968, \quad s = 0.012$$



Inserting any  $\rho_p$  or  $\rho_m$  value into these expressions yields a value for the substituted phenyl moiety. To study the development of any QSAR in the system that has the parent SMILES entry one loads it to **reg** by set number. If the set does not have parent SMILES (check by entering **f**etch) then all SMILES must be deleted and the parent entered. Next, check for the relative importance of the various parameters. Naturally if the parameters are nonlinear then one has the most to gain by maximizing them. The equation should be printed out along with the numbered parameters (**seep**). Delete the current parameters after deleting the stored equation (**del /eq**). Now enter **news**ub and receive the prompt for the first new substituent. Enter the symbol and on the prompt for the parameter value, enter a period (.) (this is to be calculated). After all new substituents of interest have been entered, enter **end**. Now the SMILES for the new derivatives must be added via **getsmi**. This provides the prompt for adding the first new substituent to the parent SMILES. When all the new substituents have been added, enter **end** or **quit**. At this point, it is good practice to enter **depict** to check all entries to see that no errors in entry have occurred. Now one must enter a new set of parameters for those that were deleted via the **f**etch command. (Section V-E).

Of course, one is searching for the limits of known substituent constants to maximize activity. At the same time, care must be taken to minimize collinearity. The approach in Section V-K. can be most helpful. For illustration, consider the example of a QSAR with the following independent variables:  $\log P$ ,  $\log P^2$ ,  $\rho_p$ ,  $\rho_m$ , B1. Of course,  $\log P$  values can be calculated from the SMILES, but in selecting substituents, the need is to select those that have  $\rho$  values which, when added to parent  $\log P$ , do not yield a  $\log P$  much beyond  $\log P_0$  or at least the upper confidence limit on  $\log P_0$ . For parameter selection assume  $\rho_p$  is most important to consider (we already know the limit on  $\log P$ ). We need to consider both  $\rho_p$  and  $\rho_m$  so that the two can be summed to get the overall  $\rho$ . After entering **parameter** we see the list to select from. Selection will be ordered on the first parameter entered that would be  $\rho_p$ . Select **15 18 1 12**. On the prompt for limits assume we have tested  $\rho_p$  up to 0.5. Hence, enter .5 and 3. Any large number can be selected for the upper limit simply to insure that all substituents with large values will be covered. We can use the same limits

for  $\beta_m$ . For the decision will have to be made after one calculates what the ideal value of  $\beta_m$  should be. For the present, we might set  $\beta_p$  of -.4 and 3 and then 1 and 3 for B1. One must set some kind of real limit for each parameter. For example, setting 0.5 and 4 for B1 the search fails, as the smallest value for B1 is 1 for H. Entering **seed** we find a catch of 19 substituents. The most interesting might be  $\text{SO}_2\text{CF}_3$  with  $\beta_p$  of 0.96. That is, if B1 has a positive coefficient in the QSAR and its  $\beta_p$  of 0.55 is not too high. Of course, two substituents *e.g.*, 3,5-di- $\text{SO}_2\text{CF}_3$  might be used if the synthetic chemists did not object too strongly!

To check for collinearity, enter **corr 3 4 5 6**. From the correlation matrix, it is apparent that only  $\beta_m$  and  $\beta_p$  are highly collinear. In the present case, this makes little difference since our interest is in getting the highest sum of  $\beta_p$  and  $\beta_m$  possible. This could be very important if one were trying to establish the importance of  $\beta^-$  or  $\beta^+$  for mechanistic perspective.

To further be sure that we are not facing a collinearity problem at a lower level in selecting  $\text{SO}_2\text{CF}_3$  we can use the regression program since parameter has placed us in that mode. The following regressions can be run: **3 reg 5 ; 3 reg 6 ; 5 reg 6**. In each case, using **pred** allows us to see the correlations between individual substituents. For more complicated problems, **pred /des /abs** can be used to find the most noncollinear examples from an inspection of the deviations.




## VII. SMILES Tutorial

Inventing the SMILES language for linear entry of complex structures of organic chemicals into computers was a major achievement by David Weininger.<sup>25a,b</sup> **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem has four basic rules that cover 98% of all structure notation. These can be learned in a few minutes, but of course it will take some practice to become proficient. One can practice by entering **udrive** from either regression mode. It is a bit more convenient to enter from the **search** mode by entering **12**. In the displayed panel one can either enter the common name **ascorbic acid** or the SMILES to see the 2-D structure. If the common name was used entering **n** goes to the

search mode and pictures the SMILES. Entering sea completes the search. The following examples illustrate the four basic rules.

1. Use ordinary atomic symbols.
  - a. B, C, N, O, P, S, F, Cl, Br, I.
  - b. Suppress hydrogen except on pyrrole nitrogen and derivatives such as indole carbazole etc. In these instances [nH] is used. The small n represents aromatic nitrogen.
  - c. Other atoms and any charges are placed in brackets; *e.g.*, [Si]; [N<sup>+</sup>]. Also use H to fill the valences of non-ordinaries; *e.g.*, dichlorosilane: Cl[SiH2]Cl.
  - d. Use upper case letters for aliphatic atoms and lower for aromatic, but the second letter of any symbol is always lower case; *e.g.*, [Sn] for aliphatic tin; [se] for aromatic selenium.
  - e. For non-linear structures, branches are enclosed in parentheses.

Examples:

- |  |   |
|--|---|
| a. CH <sub>3</sub> CH <sub>2</sub> OH  | CCO   |
| b. CH <sub>3</sub> CH <sub>2</sub> NHCH <sub>3</sub>                                   | CCNC  |
| c.   | n1ccccc1 1 is to connect n and c to form aromatic ring      |
| d.  | C1CCCCC1  |
| e.  | c1c[nH]cc1  |
| f. Pb[Cl] <sub>4</sub>   | [Pb](Cl)(Cl)(Cl)Cl  |
| g. DCCl <sub>3</sub> (deutero chloroform)  | [2H]C(Cl)(Cl)Cl (Isotopic mass number <u>precedes</u> atom) |

2. Bonds need not be specified except:
  - a. Equal sign (=) for double bond.
  - b. Pound sign (#) for triple bond.
  - c. A period separates disconnected structures; *e.g.*, ions, complexes, or reactions between two compounds. An asterisk represents the bond of a substituent to its parent.

Examples:

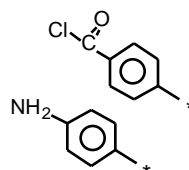
- |  |                   |
|--|-------------------|
| CH <sub>3</sub> C≡N  | CC#N              |
| CH <sub>3</sub> CH <sub>2</sub> COO <sup>-</sup> Na <sup>+</sup> | CCC(=O)[O-].[Na+] |
| HC≡CNO <sub>2</sub>  | C#CN(=O)=O        |
| -SO <sub>2</sub> NH <sub>2</sub> substituent                     | *S(=O)(=O)N       |

Azido substituent  $*N=[N+]=[N-]$

Ammonium substituent  $*[N+H3]$

Reaction between 4X-Benzoyl Chloride + 4X-Aniline

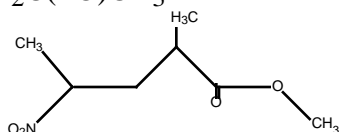
clcc(\*)cc1C(=O)Cl.clcc(\*)cc1N (see p. 27).



3. A branched group is placed in parentheses.
  - a. Branches can be nested if desired.
  - b. Can follow any path desired; will be made unique for storage.

Examples:

CF3CH2C(=O)CH3 FC(F)(F)CC(=O)C or C(F)(F)(F)CC(C)=O



C6H5N+ NBF4-

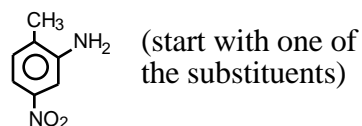
CC(N(=O)=O)CC(C)C(=O)OC  
clcccc1[N+]#N.[B-](F)(F)(F)F

4. To make a ring pathway linear one bond must be 'broken' for each ring.
  - a. One numbers the atoms on each side of the break with the same number.
  - b. Any number can be reused after that ring is closed.

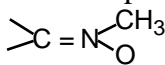
Examples:

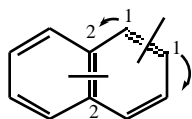
Benzene clcccc1

2-amine, 4-nitrotoluene Cc1C(N)cc(N(=O)=O)cc1

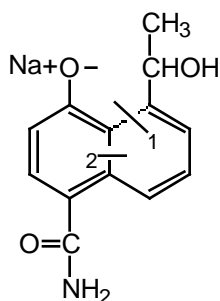


Note that N is either trivalent or pentavalent (NO2 is written as N(=O)=O); (CH3)3N as

CN(C)C; (CH3)3N-O  as C=N(=O)C.

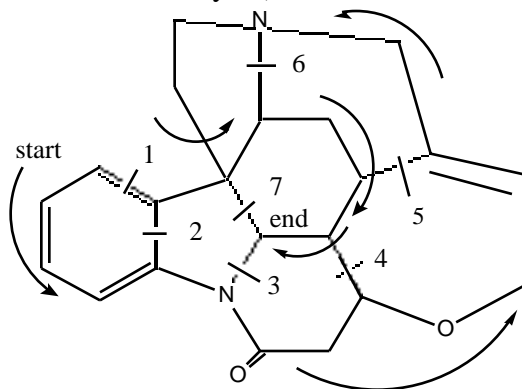


c1ccc2ccccc12 or clc2ccccc2ccc1



CC(O)clcccc2c(C(=O)N)ccc(c12)[O-].[Na+]

In some routines (*e.g.*, **getsmi**) the SMILES for many complex structures can be retrieved by name, if present in MASTERFILE. For example, it is much faster to recover a SMILES for strychnine by name, **strychnine**, than by entering SMILES atom by atom. However, the latter task is not too difficult if one chooses a path which first follows the periphery of the structure and then leads into the center. Beginning arbitrarily at the top of the benzene ring as shown, the lower case 'c' is followed by the number '1' showing that it will later be joined to the carbon fused to the pyrrolidine ring (that 'c' will also be followed by '1').



SMILES entry for strychnine

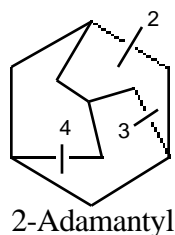
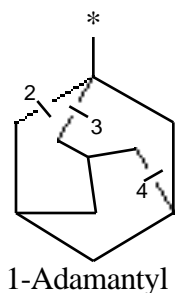
As the path follows the periphery, bonds to the interior are 'broken' and numbered in succession, until the path enters the interior at the quaternary carbon and the seventh and last 'ring break' is made. (Note that the highest number should equal the number of rings present.) The SMILES for this path is: clcccc2N3C(=O)CC4OCC=C5CN6CCC7(c12) C6CC5C4C37. Before entering the string, it is a good practice to check to see that each numeral appears as a pair. Note that before the path enters the central region, it branches to make connection to the beginning carbon (at c12). Note also that these numbers refer to 'one' and 'two' and not 'twelve'. Numbering the bond breaks is sufficient to "keep score", and it is more legible in small diagrams than numbering both sides of the break as was done in the previous example. It takes a few hours of practice to become comfortable in entering complex structures via SMILES, but if one can handle the strychnine example, there are few in all of chemistry that will be more formidable.

When entering structures in a QSAR using the **getsmi /parent** routine, and both the parent and substituent contain a ring, remember to begin the substituent ring with a number higher than used in the parent. The following example with the adamantyl as a substituent on, for example, benzene will illustrate this:

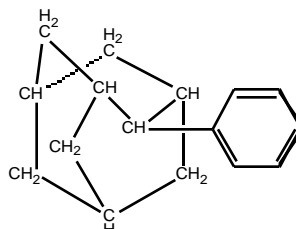
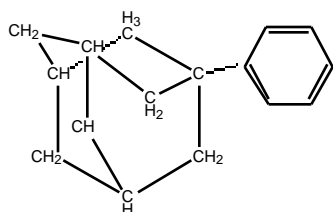
**getsmi / parent    c1c(\*)cccc1**

For the 1-adamantyl substituent the getsmi entry would be: **\*C23CC4CC(C2)CC(C3)C4**

For the 2-adamantyl substituent it would be: **\*C2C3CC4CC2CC(C3)C4**



As Shown by DEPICT:



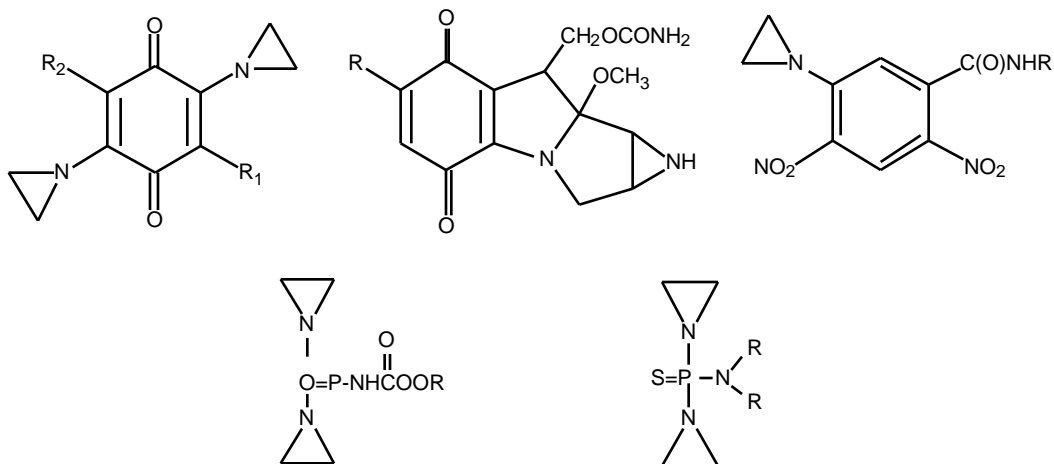
## VIII. Substructure Searching

A more general method of searching for compounds of interest is that of substructure searching. When a desired substructure is provided, the program finds all compounds in the database where a hydrogen atom of that substructure can be replaced with another group.

**Merlin** is entered from the Bio (or Phys) regression mode. A panel is presented in which the SMILES can be entered or, in many cases, the compound name. If the structure is correct, choose **n** and the program sequesters all derivatives in which a hydrogen has been replaced. Entering

**depict** , displays the 2-D structures. Entering **depict 1,10** would show the first 10, etc. or the depiction process can be stopped at any point by entering **q**.

Carrying out this process for **aziridine** in the bio database (entered as C1NC1 or by name) finds 260 examples of which the following are representative:

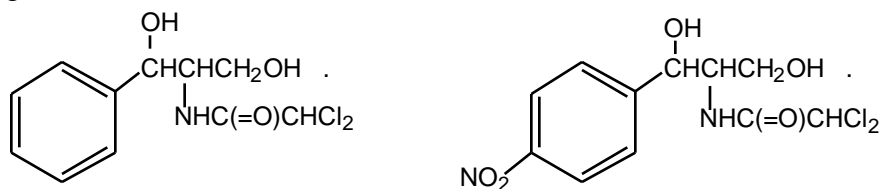


Note that even when aziridine is fused into a second ring system, it is also captured. Carrying out the same operation from the Phys mode finds 48 examples and of course searching from **data double** 308 hits on aziridines are made.

One of the shortcomings of the present system of simplified substructure searching is that it often finds too many examples for consideration. For example, in the Phys reg mode enter **merlin** and **c1ccccc1C#N** to find derivatives of benzonitrile. Choosing **n** starts the search and, on the prompt asking whether or not to continue, entering **y**, sequesters 600 derivatives of benzonitrile. Enter **depict** , displays them 4 at a time. It would be rather laborious to look at all 600 structures, but not impossible. If we limit the search by asking for derivatives of 4-chlorobenzonitrile (SMILES N#Cc1ccc(Cl)cc1), only 4 examples are found: 4-chlorobenzonitrile and 3 derivatives which contain 1 other substituent in addition to 4-Cl.

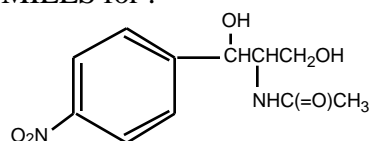
Searching for derivatives of chloramphenicol in the **bio** mode using **merlin** is illustrative. One can enter the SMILES for the analog without the nitro group, but a simpler method is to enter

**chloramphenicol** by name, choose y for editing, and then delete the **N(=O)=O** at the end of the SMILES string.



Chloramphenicol

In the resulting search we find 20 examples with substitution on phenyl ring or the C of  $\text{CHCl}_2$ . If we wish instead to study only the variations on the methyl of the acetamido group, we can enter the SMILES for :



Clcc(N(=O)=O)ccc1 C(O)CO)N(=O)C

Now 3 examples are found and are viewed by entering **depict 1**. All are 4- $\text{NO}_2$ - phenyl analogs with variations in the side chain. Searching on chloramphenicol makes only 2 hits: chloramphenicol and the trichloroacetamido analog.

We can search the double database (see Section VIII) from the search mode by entering **[Te]**. So doing finds 55 examples of compounds containing Te.

After some practice one learns how to phrase questions to limit the number of hits. However, this is difficult with aliphatic compounds. Searching with CCO (ethanol) makes more hits than can be handled. In such cases one must resort to guessing which simple structures will be contained in the dataset for QSAR of interest and resort to the simple SMILES (Section VIII).

## IX. Searching Combined Databases

There are countless ways to search the combined Phys and Bio databases. By the proper choices the searching can be limited to either half, but the primary reason for combining the two is search the Phys database for mechanistic support for Bio QSAR.



To obtain the combination enter from either regression mode **data double** then press return and enter **sea**. Care must be taken in the use of set numbers in this mode. Set numbers for Bio data are the same in data double and data Bio and hence can be used to retrieve sets for study in either mode. However, Phys data sets have been assigned new larger numbers that must be used in **data double** to extract sets for study. To see the normal Phys dataset number enter **6** in show. The number shown here is the normal Phys label which is used in the Phys mode to retrieve sets.

A large number of instructive comparisons of the electronic role of substituents have previously been published<sup>6</sup>, and thus steric factors will be used for illustration here. Taft's steric parameter,  $E_s$ , was defined using the rates of acid hydrolysis of the esters,  $\text{RCOOC}_2\text{H}_5$ . Hence, for this reaction, the coefficient of  $E_s$  is defined as 1.0. The values of  $E_s$  for all substituents larger than hydrogen are negative, and so a positive coefficient for  $E_s$  indicates the reaction is hindered by larger substituents.

Entering **15 ES** makes 508 hits on the combined database. Since we have not placed quotes on ES, it is good practice to see what has been found. Going to **show** and perusing under **15** we see that a variety of labels other than  $E_s$  are found; *i.e.*, KES, ESC, PRES, EST, ESTD. These can be removed by the not command (**15 not KES ESC** etc.). Doing so leaves 478 examples. To narrow the 'catch' enter **16 .7<ES<.9** and **sea** which reduces the number to 40. Going to **show** and entering **/sort=16 1 3 4 16 18**. The following representative examples illustrate comparative analysis.

1. Benzene

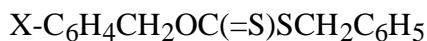
Benzoic acids

Ionization 25 deg

$$\log 1/K = 0.71(\pm.22)E_s - 2 + 1.82(\pm.19) - 1.70(\pm.70)F - 2 + 5.37(\pm.07)$$

$$n = 29, \quad r^2 = .956, \quad s = 0.160$$

2. Sea Urchin Eggs



$\text{LD}_{50}$

$$\log 1/C = 0.72(\pm.18)Es-B + 1.23(\pm.59)\log P - 2.24(\pm.10)\log(\bullet 10^{\log P} + 1) + 2.88(\pm.25)$$

$$n = 28, \quad r^2 = 0.846, \quad s = 0.354$$

3. Aqueous 50% Ethanol



Ionization 25 deg

$$pK_a = 0.73(\pm.35)Es-2 - 5.61(\pm 1.2) - A + 4.61(\pm.30)$$

$$n = 9, \quad r^2 = 0.960, \quad s = 0.235$$

4. Isopropanol

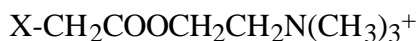


Reduction with sodium borohydride 50 deg

$$\log K = 0.75(\pm.17)Es + 2.19(\pm 1.0) * + 0.99(\pm.34)$$

$$n = 7, \quad r^2 = 0.982, \quad s = 0.139$$

5. Muscle Rectus Abdominis Frog

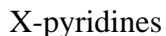


Contraction

$$\log 1/C = 0.76(\pm.17)Es + 1.19(\pm.37) + 4.65(\pm.28)$$

$$n = 6, \quad r^2 = 0.984, \quad s = 0.076$$

6. Acetonitrile



Reaction with CH<sub>3</sub>I 25 deg

$$\log K_2 = 0.79(\pm.24)Es-2 - 2.21(\pm 1.9) + 0.08(\pm.30)$$

$$n = 10, \quad r = 0.921, \quad s = 0.178$$

7. Aqueous 50% Dioxane



Acid hydrolysis 90 deg

$$\log k = 0.80(\pm.09)Es-2 + 0.08(\pm.12)$$

$$n = 13, \quad r^2 = 0.906 \quad s = 0.086$$

8. *S. aureus*



I<sub>100</sub>

$$\log 1/C = 0.83(\pm.30)Es + 1.59(\pm.30)\log P - 0.26(\pm.05)(\log P)^2 + 3.20(\pm.40)$$

$$n = 15, \quad r^2 = 0.953, \quad s = 0.208$$

9. Mouse  $\begin{array}{c} \text{CH}_3 \\ | \\ 3\text{-X-C}_6\text{H}_4\text{CH}_2\text{CHNH}_2 \end{array}$  (amphetamines)

Stimulation of locomotor activity

$$\log k = 0.84(\pm.12)\text{Es} + 0.35(\pm.12) + 2.50(\pm.30)$$

$$n = 9, \quad r^2 = 0.891, \quad s = 0.181$$

10. Aqueous



Acid hydrolysis 75 deg

$$\log k = 0.84(\pm.12)\text{Es} + 1.62(\pm.20)$$

$$n = 9, \quad r^2 = 0.974, \quad s = 0.088$$

11. Chloroplast, Pea



$I_{50}$  of photosystem II

$$\log 1/C = 0.88(\pm.29)\text{Es} - 2 + 5.69(\pm.16)\log P - 5.63(\pm1.7) \log(10^{\log P} + 1) -$$

$$0.55 (\pm.34) + 2.62(\pm.39)\text{F} - 2 - 2.51(\pm2.3)$$

$$n = 37, \quad r^2 = 0.937, \quad s = 0.203$$

M. Tuberculosis

2-X-4CONHNH<sub>2</sub> pyridines

MIC

$$\log 1/C = 0.89(\pm.22)\text{Es} - 2 - 3.70(\pm1.1)\text{F} + 5.78(\pm.50)$$

$$n = 17, \quad r^2 = 0.835 \quad s = 0.368$$

12. Aqueous 60% Dioxane



Alkaline hydrolysis

$$\log k_2 = 0.89(\pm.15)\text{Es} + 1.90(\pm.29)$$

$$n = 7, \quad r^2 = 0.978, \quad s = 0.085$$

The 13 equations are for widely different reactions. Seven represent simple chemical reactions and six are from biological systems. The parameter Es (with slope of 1) is based on the acid hydrolysis of  $\text{RCOOC}_2\text{H}_5(\text{CH}_3)$  that have a slope close to eq. 10 and 13. Several correlate substituents ortho to the functional group (1, 6, 7, 12) which brings out a similar steric effect. The examples with biological systems may involve intra- and/or intermolecular steric effects. Bear in

mind that values for Es are all negative except H so that a positive coefficient indicates a deleterious effect and vice versa. Equations 5 and 9 are closely related to the system defining Es which suggests a possible reaction involving nucleophilic attack on the carbonyl group. In the example of eq. 9 meta substituents cannot affect the side chain, hence the effect must be intermolecular. These equations illustrate one way to search for lateral support for a new QSAR. Es and B1 are similar, but we find that B1 is often the better for modeling steric effects.

Although we are only in the very early stages of Comparative QSAR, the use of a common set of classical parameters from physical organic chemistry is promising. Of course, until much more work has been done, one cannot always be sure exactly what such comparisons are telling us. Nevertheless, these results can be used as points of comparison by those researchers working with the more ambitious approaches to QSAR, such as CoMFA, neural networks and genetic algorithms.

## X. Caveats

It must not be forgotten that the present databanks were created over a period of about 30 years by many individuals. The biological database was developed by the chance discovery of data sets that could be correlated with the tools available at that time. As time went by, improvements in QSAR formulation were made, and more complex datasets could be entered. Most of the QSAR have *not* been published. Values of the dependent variables have been checked, but the values of the independent variables may have been improved (especially log P) since the original entry, but as of the present have not been checked.

To check a QSAR for possible errors the best way to start is to use the stored QSAR (**seeeq**) with the command **pred /abs /des**. This lists the results with the most poorly fit points first in increasing order of goodness of fit. The data can also be examined by the commands **pred /des /abs /unstar** or **pred /des /abs /star**. This is handy for viewing large sets that may contain a number of starred datapoints. The parameters values can be checked and one can consider other ways of parameterizing the data. It is easy to try other parameters with automatic loading now

available. About 55% of the physical QSAR have been entered using the parent SMILES procedure, but only about 45% of the biological have been so entered. We plan to remedy this deficiency so that one can rapidly check new ideas with the latest parameters.

Recently we have made a concerted effort to obtain a more complete set of QSAR from the area of physical organic chemistry. This was done by going through the indices of a number of the major journals cited in Chemical Abstracts. In this way we obtained about 4,000 QSAR. By checking references to other work in these papers, we found over 3,000 more equations. Still we have neglected certain areas such as spectra, dipole moments and Brønsted type correlations. Also, we have no doubt missed many examples not published in English or German. We have noticed that those publishing in the "standard journals" (*i.e.*, J. Am. Chem. Soc., J. Chem. Soc., and J. Org. Chem.) often do not reference papers published in "foreign" language journals. We hope to rectify these omissions in time, if we can find collaborators in other countries to help us with this 'language problem'.

We have retained sets that are obviously not very good correlations for various reasons. Many authors have not studied more than a few compounds. This generally means that variation in the substituent properties is poor. In other cases, a number of data points have had to be omitted and sometimes correlation coefficients are low. We have felt that such examples (especially the biological data) might be of help to others studying similar reactions, a poor example being better than none at all. This is where comparative QSAR plays a most helpful role. When a QSAR for a similar set of chemicals acting on a similar system can be found, a weak QSAR can be supported or refuted. Such information is more helpful than conventional statistics.

In developing QSAR we have tended to under parameterize rather than over parameterize equations. There are a variety of ways in which dual parameter equations have been developed in physical organic chemistry ( $I + R$ ,  $+ \bullet$ ;  $F + R$ , etc.). We have rarely resorted to such refinements, since a major goal has been to develop QSAR for comparative purposes. We feel that the under parameterized equations may be more easily compared.

The task of delineating the role of hydrophobic effects in QSAR is difficult. In whole organisms there is the problem of the random walk of the chemical from the site of entry to the site of action. There is good reason to expect this to be related nonlinearly to log P. At the receptor, one would expect more specific hydrophobic interactions of certain parts of the ligands and these would not necessarily parallel those of the random walk process. Hydrophobic effects at the receptor, would best be modeled by  $\sigma$  of certain substituents. In rare examples <sup>26</sup> it is possible to find clear roles for both  $\sigma$  and log P in the same equation, but usually one accepts some kind of "average" hydrophobic term. To really understand the problem studies would have to be made at the isolated receptor, on cell culture and in the whole organism. Only a very few such studies have been made.<sup>6</sup>

Another factor confounding the hydrophobic effect in whole organisms is that of P450 metabolism. In general hydrophobic compounds are more rapidly metabolized. Thus metabolism can have much to do with setting Log P<sub>o</sub>.

In building a database of QSAR for comparative analysis, having the means for the calculation of log P is of paramount importance. One simply cannot begin to measure log P for all of the SAR studies being published and very few authors, give the problem much attention. At present, the calculation of Clog P has reached a fairly reliable status as shown by the following equation:

$$M\log P = 0.96C\log P + 0.08$$

$$n = 10,000, \quad r^2 = 0.979, \quad s = 0.278$$

The ten thousand measured values have been carefully selected over a period of many years.

A short coming in all of the current methods of calculating log P is that while the relative values may be reasonable, the absolute values may be off the mark. QSAR obtained with these will have false log P<sub>o</sub> and intercepts. The only way to be sure of avoiding this uncertainty is to measure at least one log P for the set (ideally the parent compound) and use this to adjust the calculated values. When using the automatic method for obtaining log P (**add ClogP**) it is a good idea to also use **add mlogP** to see if any measured values are available for comparison. Even if only one to two

values exist this does help in getting a better estimate of  $\log P_o$ .  $\log P_o$  is of paramount importance in understanding bioavailability in whole organism research.

Parameterizing for local hydrophobic effects, can be difficult. It is not unusual to find examples where meta substituents show a hydrophobic effect and para substituents do not and this problem can be complicated by ring flipping.<sup>27, 28</sup> Ideally the substituent parameter,  $\sigma$ , can be used to check for these possibilities. However, our current automatic loading uses only  $\log P$  values measured for X-C<sub>6</sub>H<sub>5</sub> or values calculated from simple benzene derivatives. If there is a strong electron withdrawing group nearby, it will affect the  $\log P$  value of the substituent being parameterized, especially if lone pair electrons are on the attachment atom of the substituent of interest. For example, the following are measured  $\log P$ s: benzene = 2.13; aniline = 0.90; nitrobenzene = 1.85; and 4-nitroaniline = 1.39. With benzene as the parent solute,  $\log P$  values for the amino and nitro substituent are:  $-\text{NH}_2 = 0.90 - 2.13 = -1.23$ ; and  $-\text{NO}_2 = 1.85 - 2.13 = -0.28$ . When these two substituents are on the same phenyl ring, electronic interaction raises their apparent  $\log P$  values as seen in these calculations:

(1) For **nitrobenzene** as the parent solute system:  $-\text{NH}_2 = 1.39 - 1.85 = -0.46$

(2) For **aniline** as the parent solute system:  $-\text{NO}_2 = 1.39 - 0.90 = +0.49$

Each  $\log P$  value has been raised by +0.77 compared to the value using benzene as a parent.

This increase in  $\log P$  values must be considered in heteroaromatic rings if the heteroatom is electronegative; *e.g.*, for aminopyridines. One can get around this problem for relatively simple systems by taking  $\log P$  as the difference in the CLOGP values for the substituted compound and the parent compound. The ClogP program takes into account electronic interactions between substituents.

Even though  $\log P$  from the benzene system is not ideal for more complex problems it generally is good enough to spot local hydrophobic effects. Depending on the importance of the problem it might then be worthwhile to measure  $\log P$  for a variety of substituents to assess the seriousness of the problem.

One of the most difficult problems is that of outliers. These are generally found taking the 'best' QSAR that one can obtain in terms of  $r^2$  and then jackknifing to remove aberrant data points. Of course, the danger in this procedure is that one is forcing the data to fit the 'best' model. The procedure of marking outliers is important in that it helps one to get clues about the cause behind the problems. There are four major reasons for outliers: 1. The mathematical model may be incorrect. 2. Shortcomings in the parameter values—especially for steric parameters. 3. Experimental errors. 4. Finally and probably, most serious is that of side reactions. There are innumerable possibilities for members of a set of 'congeners' to react with the components of even a 'simple' cell that might affect the measured activity. Of course, this can be minimized by using relatively unreactive substituents. As we gain background from Comparative QSAR, we expect that it will be possible in some instances to identify very similar chemicals operating by different mechanisms. Two examples illustrate this point.<sup>22, 23</sup>

To our knowledge ours is the first attempt to develop a computerized database for storing and comparing QSAR for all kinds of chemical and biological reactions. As such, it no doubt has many shortcomings. However, it can be relatively easily modified in many ways and we welcome any suggestions users might care to offer. At sometime in the future we may want to use more complex equations and include 3-D graphics.

Finally, we hope that the C-QSAR program will induce as well as assist others in the next phase of QSAR, that of developing an organized science of structure-activity relationships for chemical-biological interactions such as that which has evolved for organic chemistry in the past 125 years.



## XI. Appendix: Parameter Definition

When in either the Bio or Phys regression mode loading a set and then using the `fetch` command displays the following table:

1	PI	pi	23	S-INDUC	sigma inductive
2	MR-SUB	substituent refractivity	24	S-AN-RS	sigma resonance, anilines
3	F	field effect (from S-L)	25	S-RES.+	sigma resonance plus
4	R	resonance effect (from S-L)	26	S-'	sigma prime
5	R+	resonance plus	27	S-PARNO	sigma para normalized
6	R-	resonance minus	28	S-ORTH+	sigma ortho plus
7	ES	E(s) from Taft	29	S-PHOSP	sigma phosphoric acid
8	ES- HYBO	E(s) from hydroboration	30	S-L	sigma localized [Charton]
9	ES-V	E(s) from Charton	31	S-OTWST	sigma orthogonal twist
10	ES-A	E(s) from Austel	32	S-STAR	sigma star from Taft's
11	L-STM	length sterimol	33	S-IND.P	sigma inductive [phosphorus]
12	B1-STM	width sterimol	34	S-RES.P	sigma resonance[phosphorus]
13	B5-STM	width sterimol	35	ER-P	electronic radical, para
14	O-STER	ortho quats with MeI	36	ER-M	electronic radical, meta
15	S-P	sigma para	37	S.DOT-P	sigma dot, para
16	S-P+	sigma para plus	38	S.DOT-M	sigma dot, meta
17	S-P-	sigma para minus	39	S.-DOT-P	sigma dot, para (JJ)
18	S-M	sigma meta	40	S.-DOT-M	sigma dot, meta (JJ)
19	S-M+	sigma meta plus	41	S.P-C	sigma para (C)
20	S-M-	sigma meta minus	42	S.M-C	sigma meta (C)
21	S-O	sigma ortho			
22	S-O-	sigma ortho minus			

Any of the 42 different parameters can be automatically loaded for regression analysis, perused for drug design or compared with each other for theoretical analysis. The more commonly used (1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 15, 16, 17, 18, 23, 32) are discussed and their use illustrated in ref. 4.

A brief definition and references follows.

1. PI. Hydrophobic parameter (  $\pi$  ) for substituents defined by partitioning of X-C<sub>6</sub>H<sub>5</sub> between octanol and water. (P)<sup>1, 2</sup>

$$\pi_x = \log P_{X-C_6H_5} - \log P_{C_6H_6}$$

2. MR-SUB. Molar refractivity of a substituent defined analogously to  $\pi$ .

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

where n = refractive index, MW = molecular weight and d = density.

MR values are scaled by 0.1. MR is highly collinear with substituent volume.<sup>1, 2</sup>

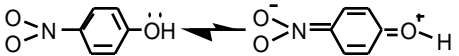
3. F. Swain-Lupton inductive/field effect parameter for aromatic systems.<sup>14</sup>
4. R. Corresponding Swain-Lupton resonance parameter.<sup>14</sup>
5. R+. Taft resonance parameter for substituent delocalization of a + charge.<sup>14</sup>
6. R-. Taft resonance parameter for substituent delocalization of a - charge.<sup>14</sup>
7. ES. Classic steric parameter for substituents defined by Taft from the hydrolysis of X-CH<sub>2</sub>COOCH<sub>3</sub>(C<sub>2</sub>H<sub>5</sub>).<sup>20</sup>
8. ES-HYBO. An Es type parameter obtained from the hydroboration of substituted ethylenes.
9. ES-V. Charton's steric parameter.<sup>29</sup> Using the command **parameter /nolimit** and entering **7** and **9** it is found that there are 117 for which both Es parameters are available. Entering **3** **reg 4** yields a correlation between Es and Es-V with  $r^2 = 0.942$ . This can be improved to  $r^2 = 0.964$  by jackknifing to remove three badly correlated points: C(CH<sub>2</sub>CH<sub>3</sub>)<sub>3</sub>, CHBr<sub>2</sub>, CH(Me)CH<sub>2</sub>CMe<sub>3</sub>.
10. Austel's <sup>30</sup> version of Es. A calculated value available for 1738 substituents. There are 198 substituents with both Es-A and ES, between which the correlation is weak:  $r^2 = 0.793$ .
11. L-STM. Verloop sterimol parameter for substituent length.<sup>1, 2, 15</sup>
12. B1-STM. Sterimol parameter for the width of the first atom of the substituent.<sup>1, 2</sup>

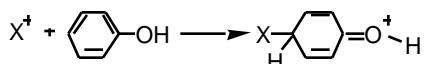
13. B5-STM. An estimate of the overall width of the substituent.<sup>1,2</sup>
14. O-STER. There are 30 substituents having this parameter for the effect of adjacent substituents inhibiting the reaction of pyridines with CH<sub>3</sub>I. There is little correlation between O-STER and Es or B1.<sup>31</sup>
15. S-P. Normal Hammett constant for para substituents. It is based on the ionization constants of benzoic acids.<sup>1,2</sup>
16. S-P+. Brown  $\rho^+$  parameter where substituents delocalize a + charge or radical via resonance.<sup>1,2</sup>
17. S-P-. Hammett constant ( $\rho^-$ ) where substituents delocalize a negative charge via resonance. It is derived from the ionization constants of phenols.<sup>1,2</sup>
18. S-M. Hammett constant for meta substituents (non conjugated substituents).<sup>1,2</sup>
19. S-M+. Brown  $\rho^+$  parameter for meta substituents.<sup>4</sup> There is little difference between  $\rho_m$  and  $\rho_m^+$ .<sup>1,2</sup>
20. S-M-. Hammett  $\rho^-$  for meta substituents (non conjugated substituents).<sup>1,2</sup>
21. S-O-. Hammett  $\rho^-$  for ortho substituents. It correlates poorly with  $\rho_p$ , for 51 substituents  $r^2 = 0.303$ .<sup>32</sup>
22. S-O. The  $\rho^-$  parameter for ortho substituents.<sup>33</sup>
23. S-INDUC.  $\rho_i$  for the field/inductive effect.<sup>1,2</sup> Originally defined from 4-X-bicyclo [2.2.2] octane-1-carboxylic acids.
24. S-AN-RS.<sup>34</sup> Resonance parameter ( $\rho^-$ ) obtained from anilines. There are 25 substituents for which both  $\rho^-$  and S-AN. The correlation between the two is poor:  $r^2 = 0.718$ .
25. S-RES+. Resonance parameter for delocalization of + charge.<sup>35</sup>

26. S-'. Field/inductive parameters from bicyclo [2.2.2] oct-ene-1-carboxylic acids, 4-X-dibenzobicyclo [2.2.2] octa-2, 4-diene-1-carboxylic acids and cubanedicarboxylic acids.<sup>36</sup>
27. S-PARNO. A set of normalized  $\rho$  values.<sup>37</sup>
28. S-ORTH+.  $\rho^+$  for ortho substituents.<sup>38</sup>
29. S-PHOSP.  $\rho$  for substituents attached to phosphorus.<sup>39</sup>
30. S-L.  $\rho_L$  for field/inductive effect.<sup>40</sup>
31. S-TWST. Effect on resonance by twisting substituent 90° out of plane.<sup>41</sup>
32. S-STAR. Classic  $\rho^*$  defined by Taft.<sup>1, 2</sup>
33. S-IND.P. Field/inductive parameter for substituents attached to phosphorus.<sup>42</sup>
34. S-RES.P. Resonance parameter for substituents attached to phosphorus.<sup>42</sup>
- 35,36. ER-P, ER-M. Radical parameters defined by Yamamoto and Otsu.<sup>1, 2, 43</sup>
- 37,38. S.DOT-P, S. DOT-M. Radical parameters ( $\rho^{\cdot}$ ) defined by Dust and Arnold.<sup>44</sup>
- 39,40. S.-DOT-P, S.-DOT-M. Radical parameters ( $\rho^{\cdot}$ ) defined by Jiang.<sup>45</sup>
- 41,42. S.P-C, S.M-C. Radical parameters ( $\rho^{\cdot}$ ) defined by Creary.<sup>46</sup>

## XII. Notes on the Use of Substituent Parameters

The electronic parameters for aromatic substituents,  $\rho$ ,  $\rho^+$ ,  $\rho^-$  have been developed and extensively tested over the last 40 years. The 'normal' electronic parameter  $\rho$  is used when there is not significant through resonance (*i.e.*, direct resonance between a substituent conjugated with the reaction center.) The parameter  $\rho^-$  holds when direct resonance is involved in the delocalization of

a negative charge: *i.e.*, . While  $\rho^+$  is used when delocalization of a positive charge is involved:

. Surprisingly,  $\rho^+$

serves very well to correlate radical reactions.<sup>9</sup> In fact, it is more generally useful to detect and correlate radical reactions than parameters formulated expressly for this purpose ( $\rho$  and  $E_R$ )<sup>9</sup>.

Those inexperienced in physical organic chemistry should test all three. In fact experienced researchers are sometimes surprised by unexpected results that shed light on the underlying reaction mechanism. A serious problem is the high collinearity among these parameters that must be minimized by careful substituent selection. For example, there are 199 substituents for  $\rho_p$  and

$\rho_p^+$  for which  $r^2 = 0.878$ . However, this collinearity can be greatly reduced by the proper selection of substituents. In the case of  $\rho_p$  and  $\rho_p^-$  where 196 values are available for both types of substituents correlation is also high  $r^2 = 0.829$ . Of course, meta substituents are not conjugated with a reaction center so that  $\rho_m$ ,  $\rho_m^+$ ,  $\rho_m^-$ .

For reactions with aliphatic systems where resonance is not involved two so-called field/inductive parameters are available  $F^*$  and  $F$ . Sometimes one gives a somewhat better answer than the other.  $F^*$  may contain a small component associated with steric effects. The correlation between these parameters is also high  $r^2 = 0.856$  for 366 substituents for which both values are known.  $F$  is a field/inductive parameter for aromatic systems. It is most commonly tested in

combination with one of the aromatic parameters and a steric parameter for substituents ortho to a reaction center.

The commonly used steric parameters are Es, B1, B5, L and MR. Es is the classic steric parameter defined from ester hydrolysis by Taft. It is somewhat comparable to B1, however for 147 substituents common to both series  $r^2$  is only 0.3. But the low collinearity is the result of including large bulky substituents in the comparison. The collinearity is much higher with the smaller commonly used groups. B1, B5 and L are calculated parameters. B1 is used for steric effects around the first atom in the substituents, B5 is bulk parameter and L is for the overall length. MR (molar refractivity) is essentially a measure of the overall bulk of a substituent. It is highly collinear with molar volume. Unless care is taken it may also be highly collinear with log P.

Of course, there are two major steric problems: intra- and inter- molecular and generally in biological systems one has no idea whether one or the other or both are involved. One of the great advantages of our system is that parameters can be very rapidly load for regression analysis and then using the **perm**utation command one can quickly sort the significant variables. All of the above parameters have been discussed in more detail and there uses illustrated.<sup>4</sup> A parameter that cannot be automatically loaded is the indicator variable. These are often of considerable help. They are added by using the **edit set** command. Then one must take care to add the name of the new variable. Then a 1 or 0 is added for the presence or absence of the feature under consideration. In adding values for substituents that cannot be autoloading care must be taken not to overrun the space allowed for each line. When this is done, a marker is shown on the far right of the screen. This must be removed by moving the cursor away from the last value and using delete until the marker disappears.

### XIII. QSAR and Combinatorial Synthesis

The almost infinite possible variation in the structure of organic chemicals presents the drug designer or the toxicologist with overwhelming possibilities. For example, we noted in 1979<sup>6</sup> that if one were to substitute all seven positions on the quinoline ring, with (at that time) the number of well characterized substituents (166), there would be approximately  $3.5 \times 10^{15}$  possible derivatives. Combinatorial synthesis has greatly magnified the problem as Brown and Martin have shown. They note that if one were to start with 360 commercially available precursors for  $R_1$  and 259 for  $R_2$  and if one were to select 100  $R_1$ 's and 100  $R_2$ 's for the production of 10,000 compounds by combination of the  $R_1$  and  $R_2$  precursors the number of possibilities would be  $2.5 \times 10^{164}$ . To provide perspective for the size of such a number, they note that the age of the universe in seconds is about  $10^{17}$  and its size is  $10^{108} \text{ e}^3$ .

Combinatorial synthesis is booming in tandem with the stock market! Although the approach outlined in this manual is not designed to expedite the random search for new leads it does have some features that may be of help. There are hundreds of QSAR that are based on trivial numbers of data by combinatorial standards. These QSAR provide points of departure for combinatorial studies as well as old fashioned synthesis as mentioned in VI.

In undertaking the expense of designing, synthesis and testing of a large number of compounds it is very important to avoid property redundancy, the collinearity problem and to be sure that electronic, steric and hydrophobic data space is well covered. For example, electronic terms are found in 1,509 out of 6,300 bio QSAR and hydrophobic terms are found in 4,097. In what we would call a *conservative* approach to multicongener synthesis, starting with a highly active parent molecule (active at  $10^{-6}$  to  $10^{-9}$  M) the interest would be in substituents whose properties are known. The only way to do so at present is via groups with known substituent constants. In section V. K, we have discussed this problem. In addition to substituent constants one could use calculated log P values as well. However we believe, despite their shortcomings, for substituents can often provide better insight. We have encountered hundreds of instances where

only the hydrophobicity of one part of the molecule is important in the QSAR. One must calculate or measure log P for the parent and then limit the selection of  $\log P$  values so that the log P range for the compounds to be prepared is reasonable. Rather few drugs have log P outside the range of -1 to 5.<sup>48</sup>



## XII. References

1. Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, 1995.
2. Hansch, C.; Leo, A. and Hoekman, D. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*. American Chemical Society, 1995.
3. Hansch, C.; Hoekman, D. and Gao, H. *Chem. Rev.* 1996, 96: 1045-1075.
4. Gao, H., Katzenellenbogen, J.A., Garg, R. and Hansch, C. *Chem. Rev.* 1999, 99: 723.
5. Hansch, C., Gao, H. and Hoekman, D. In, *Comparative QSAR* J. Devillers, Ed. Taylor and Francis, Washington D.C. 1997, p. 285-368.
6. Hansch, C. and Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley-Interscience, 1979, pp. 48.
7. Hansch, C. *Acc. Chem. Res.* 1993, 26: 147.
8. Hansch, C. and Gao, H. *Chem. Rev.* 1997, 97: 2995-3059.
9. Selassie, C.D.; Shusterman, A.J.; Kapur, S.; Verma, R.P.; Zhang, L. and Hansch, C. *J. Chem. Soc., Perkin Trans 2* 1999,
10. Garg, R.; Gupta, S.P.; Gao, H.; Babu, M.S.; Debnath, A.K. and Hansch, C. *Chem. Rev.* 1999, 99: .
11. Fujita, T. In, *Drug Design: Fact or Fantasy?*. G. Jolles and K.R.H. Wooldridge, Eds. Academic Press 1984, pp 17.
12. Fujita, T. in *Comprehensive Medicinal Chemistry, Vol. 4*. C. Ramsden, Ed., Pergamon, 1990.
13. Fujita, T. and Nishioka, T. *Prog. Phys. Org. Chem.* 1976, 12: 49.
14. Hansch, C.; Leo, A. and Taft, R.W. *Chem. Rev.* 1991, 91: 165.
15. Verloop, A., Hoogenstraaten, J. and Tipker, J., In, *Drug Design, Vol. VII*. E. J. Ariens, Ed., Academic Press, 1976, pp.165.
16. Hansch, C.; Bjorkroth, J.P. and Leo, A. *J. Pharm. Sci.* 1987, 76: 663.
17. Ref. 4, pp.392.
18. Debnath, A.K.; Shusterman, A.J.; deCompadre, R.L.L.; and Hansch, C. *Mutat. Res.* 1994, 305: 63.
19. Leo, A. *J. Chem. Rev.* 1993, 93: 1281.
20. Unger, S.H. and Hansch, C. *Prog. Phys. Org. Chem.* 1976, 12: 91.
21. Abraham, M.; and McGowan, J. *J. Chromatographia* 1987, 23: 243.
22. Selassie, C.D., DeSoyza, T.V., Rosario, M., Gao, H. and Hansch, C. *Chem. Biol. Inter.* 1998, 113: 175.
23. Smith, C.J., Hansch, C. and Morton, M.J. *Mutat. Res.* 1997, 379: 167.

24. Cramer, R.D.III; Bunce, J.D.; Patterson, D.E. and Frank, I.E. *QSAR* 1988, 7: 81.
- 25.a. Weininger, D. and Weininger, J.L. In, *Comprehensive Medicinal Chemistry* Ramsden, C.A., Ed. Pergamon Press, 1990, Vol. 4, pp. 59.
- 25.b. Weininger, D. *J. Chem. Inf. Comput. Sci.* 1988, 28: 31.
26. Kim, K.H., Hansch, C., Fukumaga, J.Y., Steller, E.E., Jow, P.Y.C., Craig, P.N. and Page, J. *J. Med. Chem.* 1979, 22: 473.
27. Hansch, C. and Klein, T. *Acc. Chem. Res.* 1986, 19: 392.
28. Selassie, C.D., Li, R.-L., Poe, M. and Hansch, C. *J. Med. Chem.* 1991, 34: 46.
29. Charton, M. *Prog. Phys. Chem.* 1971, 8: 235.
30. Austel, V., Kutter, E. and Kalbfleisch, W. *Arzneim.-Forsch.* 1979, 29: 585.
31. Berg, U., Gallo, R., Klatte, G. and Metzger, J. *J. Chem. Soc. Perk. 2* 1980, 1350.
32. Perrin, D., Dempsey, B. and Serjeant, E. *pKa Predictions for Organic Acids and Bases.* Chapman and Hall 1981, pp.107.
33. Tribble, M. and Traynham, J. *J. Am. Chem. Soc.* 1969, 91: 379.
34. Dayal, S., Ehrenson, S. and Taft, R.W. *J. Am. Chem. Soc.* 1972, 94: 9113.
35. Palm, V., Ed. Summary of Science and Technology Questions in Organic Chemistry: Tables of Rate and Equilibrium Constants of Heterocyclic Organic Reactions. Moscow, 1979.
36. Baker, F.W., Parish, R.C. and Stock, L.M. *J. Am. Chem. Soc.* 1967, 89: 5677.
37. Van Bekkum, H., Verkade, P.E. and Webster, B.M. *Rec. Trav. Chim.* 1959, 78: 815.
38. LeGuen, M.M.J. and Taylor, R. *J. Chem. Soc. Perk. 2* 1976: 557.
39. Mastryukova, T.A. and Kabachnik, M.I. *J. Org. Chem.* 1971, 36: 1201.
40. Charton, M. *Prog. Phys. Org. Chem.* 1981, 13: 119.
41. Grindley, T.B., Johnson, K.F., Katritzky, A.R., Keogh, H.J., Thirkettle, C., Brownlee, R.T.C., Munday, J.A. and Topsom, R.D. *J. Chem. Soc. Perk. 2* 1974: 276.
42. Kasukhim, L. and Gololobov, Yu, G. *Organic Reactivity* 1978, 15: 463EE.
43. Yamamoto, Y. and Otsu, T. *Chem. Ind.* 1967: 787.
44. Dust, J.M. and Arnold, D.R. *J. Am. Chem. Soc.* 1983, 105: 1221.
45. Jiang, X.-K. and Ji, G.Z. *J. Org. Chem.* 1992, 57: 6051.
46. Creary, X., Mehrsheikh-Mohammadi, M.E. and McDonald, S. *J. Org. Chem.* 1987, 52: 3254.
47. Gao, H., Denny, W.A., Garg, R. and Hansch, C. *Chem. Biol. Inter.* 1998, 116, 157.
48. Ghose, A.K., Viswanadhan, V. N. and Wendoloski, J.J. *J. Comb. Chem.* 1999, 1, 55.