

CHEM-BIO INFORMATICS AND COMPARATIVE QSAR

Fall 2002

BioByte Corp.

201 West 4th Street, Suite 204

Claremont, CA 91711

Phone number: (909) 624-5992

FAX Number: (909) 624-1398

E-Mail: clogp@biobyte.com

Table of Contents

| | |
|---|----|
| Preface | 4 |
| I. Introduction | 5 |
| II. Structure of the System | 6 |
| III. Using the Databases | 12 |
| A. Help | 12 |
| B. String Searching | 12 |
| C. Physical Database: | 13 |
| 1. Main Menu | 14 |
| 2. Searching and Show | 14 |
| 3. Browsing..... | 22 |
| 4. Statistics..... | 23 |
| 5. Omitted Data Points..... | 24 |
| 6. SMILES | 25 |
| 7. MERLIN and SMARTS | 27 |
| 8. Loading from 'Database Search' to 'Workspace'..... | 28 |
| D. Biological Database | 29 |
| 1. Browsing | 30 |
| 2. Searching | 30 |
| 3. Comparing New QSAR..... | 35 |
| IV. Searching for New Lead Compounds | 39 |
| A. Locating QSAR that are Based on Highly Active Compounds | 39 |
| B. MERLIN Searching | 41 |
| V. Searching Combined Databases | 43 |
| VI. SMILES Tutorial | 47 |
| VII. Parameter Definition | 51 |
| VIII. Regression Analysis: Example 1 | 56 |
| A. Title Information (set B633) | 57 |
| B. Naming Parameters | 57 |
| C. Naming and Entering Substituents | 57 |
| D. Entering Structures via SMILES | 59 |
| E. Auto-Loading of Parameters | 59 |

| | |
|---|----|
| F. Permuting | 60 |
| G. Checking for Parameter Collinearity | 62 |
| H. Jackknifing | 62 |
| I. Plotting Data | 63 |
| J. Cross Validation | 63 |
| K. Editing | 64 |
| L. Regression Analysis: Example 2 | 65 |
| M. Substituent Selection in Molecular Design | 68 |
| VIII. UDRIVE and Masterfile | 70 |
| IX System Crashes | 70 |
| X. Caveats | 70 |
| XI References | 75 |

Preface

It has now been 40 years since the first QSAR were developed at Pomona College. Our first attempt to cover the subject from the chemical as well as the biological point of view was published in 1996. Although it is still by far the most extensive publication of equations, much has changed since that time. Then our database of biological QSAR was only 6,000 and QSAR from mechanistic organic chemistry was much less. We now have a much more sophisticated, searchable database of over 17,700 equations of which 8,900 are for biological reactions. In 1995, the calculated octanol/water partition coefficient Clog P was based on only 7,500 measured values with $r^2 = 0.956$ and standard deviation of 0.336. At present, the accuracy of this most important parameter is much improved: $n = 12,600$, $r^2 = 0.973$, $s = 0.299$. The Hammett type parameters are still the most extensively tested electronic parameters. Where they have been compared with quantum chemical parameters (about 130 examples), about half the time one is significantly more effective than the other.

What remains most surprising to us is that, in the vast majority of studies on the interaction of chemicals with biological entities, no attempt has been made to make any kind of QSAR. Worse, structural changes in a parent compound are obviously being made with no thought given to including substituents that have significant variation in hydrophobic, electronic and steric properties! Although the need for interaction between synthetic and mechanistic organic chemists is becoming ever greater, as is the close interaction between chemists and biologists, progress is very slow.

Two important features of our program are: it can be used to see what has been done in thousands of examples and it can be used to formulate new QSAR. The latter use is greatly facilitated by the autoloading of parameters. This includes the automatic calculation of log P, σ , CMR (substituent polarizability) and MgVol (molecular volume). We believe that the majority of researchers will want to begin by reviewing what has been done in their area of interest. This often provides important possibilities for comparative QSAR. In this way one can obtain lateral support for any newly defined QSAR.

I. Introduction

Understanding how chemicals react with themselves and how they interact with the chemicals that compose living systems or parts thereof (DNA, proteins, enzymes, organelles, cells, membranes) has become a major concern of many phases of science. It is central to drug development, toxicology, environmental science, biochemistry and molecular biology. The information explosion, which is occurring at a constantly increasing rate, is an enormous challenge to cope with, despite the various online approaches. These methods are excellent *if* one knows the specific questions one needs answers to. They are rather helpless when it comes to searching for structure-activity relationships. For instance, simple Hammett equations that were initiated by Hammett in the mid-1930s are sometimes listed in *Chemical Abstracts* as LFER (linear free energy relationships), occasionally under Hammett, and these days sometimes under QSAR. However, searching under these terms would yield only a small fraction of the published data because many authors now do not include any indication of a QSAR in their abstract. The last attempt at complete coverage was made by Jaffe in 1953 when he listed about 400 that were termed LFER! We now have 8,900 such equations that cover every type of chemical reaction.

The situation with biological reactions is far more complex and it is difficult to find particular mathematical relationships between structure and activity (QSAR), primarily because *very* few authors attempt to publish their results in such terms. Usually it's, "here are the chemicals; here are biological activities—you take it from there." It has not been easy to collect data for these two classes. We have obtained leads from chemical abstracts or from perusing certain key journals. Then, from the references in those articles, uncovered new articles and so on and on.

What we have been attempting to do for the last 40 years is to develop a dynamic, integrated, computerized system covering all chemical-chemical interactions and chemical-biological interactions. We enter any information into this system of chem.-bioinformatics that can be described in mathematical terms with reasonable statistics. Of course any particular QSAR may be of interest, but that is not our primary concern. Comparative QSAR is our main goal. This becomes the foundation for developing a science of how chemicals affect living systems. This will take decades to develop properly, but it is already very helpful. We are constantly finding useful, unexpected relationships. For example, a study at the EPA on the toxicity of phenols to rat embryos *in vitro* set us on a path that yielded a QSAR that rationalized the estrogenic and

carcinogenic activity (or the *lack* of it) in many simple phenols and complex phenols such as diethylstilbestrol, Bisphenol A, estradiol, estriol, equilin and equilenin.¹ More recently we have unexpectedly discovered a way to establish a wide variety of allosteric interactions.^{2, 2a} Now that our database of QSAR has grown to 17,800 equations of which 8,900 are bio QSAR, the greatly enlarged database offers real possibilities for finding entirely new relationships that are truly exciting.

It must be emphasized that developing a science of chemical-biological interactions is not possible unless one employs standard parameters whose meaning is clearly understood. Moreover, one must be reasonably certain that the dependent variables for a given QSAR are reasonably uniform. If the chemicals are causing a similar reaction but by different mechanisms the final QSAR may be more confusing than helpful. There are many ways of formulating a QSAR these days, but so far ours is the only one the parameters of which can be compared among large numbers of equations.^{3, 4, 7, 8, 9, 9a}

Lead generation has always been of paramount interest to those designing new chemicals of biological interest. Our approach to lead generation is covered in section IV. In the past, clues were obtained from studying naturally occurring biologically active chemicals. Today the hot subject in lead generation is combinatorial synthesis. However, QSAR can play a vital role in lead exploitation. Possibly the best published example is that of Koga's⁵ demonstration that the rather mediocre drug nalidixic acid could be transformed with the help of QSAR into what has become the fabulous quinolone carboxylates. Other examples of QSAR success in the development of products of industrial importance have also been discussed by Fujita.⁶ Of course, as soon as a new drug hits the market competing companies start to work to make improved modifications. In doing so, the total application of QSAR is important. Also after getting a lead compound via combinatorial synthesis, the next step is modification guided by QSAR to insure the best overall properties in the compound going to market. Data mining is a current buzzword. We are doing model mining. Behind every molecule stands a QSAR model that points to ways to make more, *or less*, active congeners.

II. Structure of the System

An overview of our system is outlined in Tables 1 – 3. Our major concern, and most difficult problem, has been to organize topics and data so that information related to a particular problem

can be easily located with extraneous material omitted. Since one is most often concerned with either the biological or physical data, the system is divided into two sections; either of which can be searched independently or jointly.

The input data in Table 1 contains the information that must be associated with every QSAR. The systems are different for the biological and physical data. For the former, a name (enzyme, cell, animal, etc.) must be entered, while in the case of the phys data, the name of the reaction solvent is entered. The output data (fields 15-20) are that obtained from the regression analysis.

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) logP, 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 shows how the biological data are categorized. There are six major classes and these are broken down into sub-classes. All of these can be searched independently or in any

combination. For instance, loading B2A with B6F would gather all QSAR on oxidoreductases and fish so that one could look at QSAR for chemicals reacting with fish and compare these with oxidoreductases. The numbers in parenthesis indicates that number of QSAR in each sub-class.

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

| BO | Unknown (9) | B4 | Single-Celled Organisms |
|-----------|--|-----------|----------------------------------|
| | | B4A | Algae (42) |
| B1 | Nonenzymatic Macromolecules (DNA, Fibrin, Hemoglobin, Soil, Albumin, etc.) (286) | B4B | Bacteria (814) |
| | | B4C | Cells in culture (962) |
| | | B4E | Erythrocytes (79) |
| | | B4F | Fungi, Molds (295) |
| | | B4P | Protozoa (116) |
| B2 | Enzymes | B4V | Viruses (200) |
| B2A | Oxidoreductases (812) | B4Y | Yeasts (54) |
| B2B | Transferases (283) | | |
| B2C | Hydrolases (928) | B5 | Organs/Tissues |
| B2D | Lyases (58) | B5C | Cancer (259) |
| B2E | Isomerases (23) | B5G | Gastro-intestinal tract (79) |
| B2F | Ligases (2) | B5H | Heart (91) |
| B2G | Receptors (1665) | B5I | Internal/soft organs (64) |
| | | B5N | Nerves, Brain, Muscles (365) |
| B3 | Organelles | B5S | Skin (55) |
| B3A | Mitochondria (91) | B5L | Liver (30) |
| B3B | Microsomes (99) | | |
| B3C | Chloroplasts (84) | B6 | Multi-Cellular Organisms |
| B3M | Membranes (118) | B6A | Animal (vertebrates) (698) |
| B3R | Ribosomes (0) | B6B | Insects (232) |
| B3S | Synaptosomes (23) | B6F | Fish (202) |
| | | B6H | Human (43) |
| | | B6I | Invertebrates (non-insect) (114) |
| | | B6P | Plants (125) |

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

Classification is the most difficult problem facing the study of biological QSAR. The major areas such as biochemistry, medicinal and pesticide chemistry, toxicology, etc. all have large numbers of sub-classes, *e.g.*, enzymology, anesthesiology, cancer, mutagenesis, metabolism, cardiology, psychobiology, bacteriology, plant physiology, urology, etc. It is apparent from Table 2 that, beyond the few keywords listed, we have not as yet attempted a very extensive

classification. We hope that this can be handled by string searching (see below).

The phys database has been somewhat easier to organize, however, even here we have been forced to create a miscellaneous class that now contains almost 529 QSAR. In compiling the phys database from mechanistic physical organic chemistry studies, we have concentrated on reactions in solution. Although there are a few examples based on gas phase reactions and spectra, we gave up on these, as they seemed to be of little value in understanding biological processes. We started out with a class for Brønsted reactions but gave up on this and now enter these under the appropriate chemical type (*e.g.*, nucleophilic substitution). Many papers report results at a variety of temperatures. Generally for lack of resources, we report one example at the temperature nearest to 25°. In examples where results have been obtained in various concentrations of mixed solvents representative examples are given. We have not attempted to standardize the dependent variable hence intercepts cannot be compared as they can for biological reactions, where C in $\log 1/C$ is the molar concentration producing a standard response.

Many data sets have been reported with little, or no thought, about collinearity problems (we are now correcting these early results) and parameterizing the data so that all types of steric, electronic or hydrophobic properties of substituents have been covered. Synthetic chemists are often unfamiliar with such properties and are more concerned with the immediate problems, difficulties in synthesis. Nevertheless, we have tried to do the best we could to include such data as we have often obtained clues from modest studies that have helped our general understanding and suggested new experiments.

Of course it is by no means clear what quality of correlation equation one can expect for a bio QSAR even when all possible care has been taken in the design of say 30 to 40 congeners. Even in treating something as simple as a cell culture (let alone mice or people), the problems are awesome. Nonetheless, the enormous drug industry and the agencies trying to classify the environmental risks of industrial chemicals constantly face these problems. The DNA codes for 50 to 100 thousand cellular proteins that account for the many enzymes and components of the variety of organelles and cellular membranes. All sorts of biochemical processes such as endocytosis and exocytosis are apt to be effected. The possibility for 'side reactions' is enormous. No doubt the *weakest* links (with respect to a given set of congeners) in this complex machinery will be perturbed. However the concentration of chemical necessary to reach a particular 'end

point' may vary greatly. For example, rather high concentrations are required to disrupt a cell membrane (10^{-3} molar). Other compounds are active at 10^{-9} molar or even lower (10^{-11}). The time allowed for the experiment is also crucial. These factors allows us to gain meaningful QSAR whose shape can be shown to be supported with other bio QSAR and most surprising, in many instances, with those from studies in mechanistic physical organic chemistry.^{9a} An important factor is that over the eons, nature has favored the evolution of highly specific receptors that exclude extraneous molecules. Allosteric effects may be involved.^{2, 2a}

Our current premise is that the major interaction forces to consider for a set of congeners acting on a biological system are hydrophobic, steric and electronic (including polarizability). Less important are hydrogen bonding and dipole moments. Hydrogen bonding can be important but, as yet there is no general way to parameterize it. For example, the orientation and distance between an OH on a substrate or inhibitor and the bonding site on the receptor is so critical that a general method for parameterization appears to be impossible. Indicator variables can be helpful.

Table 3: Class Codes--Physical Database

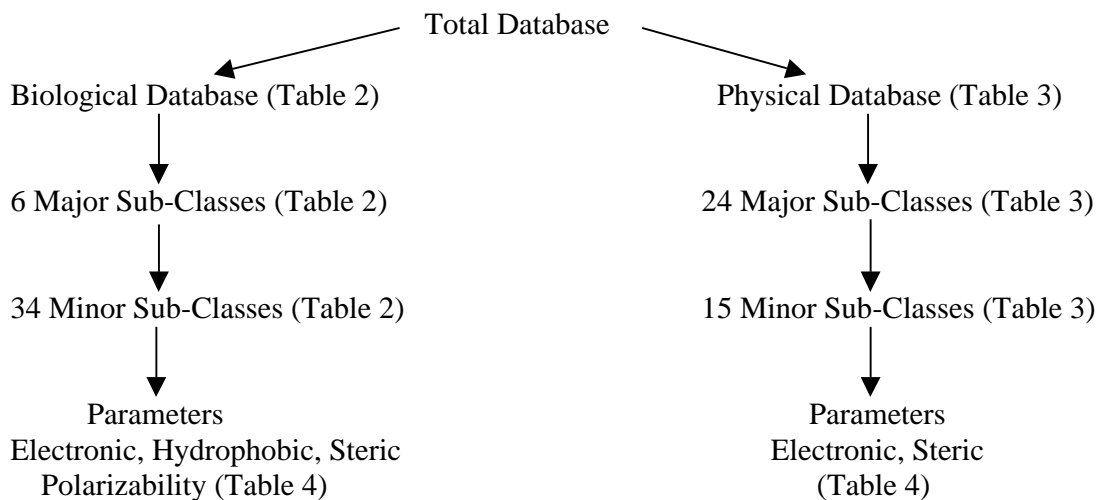
{Number of sets in parentheses}*

| | | | |
|-----------|--------------------------------------|------------|--------------------------------|
| PT | Theoretical (47) | P7 | Addition |
| | | P7D | Dimerization (12) |
| PO | Unknown | P7E | Electrophilic Addition (151) |
| | | P7N | Nucleophilic Addition (253) |
| P1 | Ionization (1732) | P7P | Polymerization (10) |
| P1P | Ionization Potential (39) | | |
| P1X | Proton Exchange (74) | P8 | Elimination (173) |
| | | P9 | Rearrangement (220) |
| P2 | Hydrolysis (872) | P10 | Oxidation (557) |
| | | P12 | Radical Reactions (613) |
| P3 | Solvolysis (646) | P13 | Complex Formation (105) |
| | | | |
| P4 | Spectra | P14 | Partitioning (130) |
| P4I | Ionization Spectra (60) | P14C | Chromatography (21) |
| P4E | ESR Spectra (6) | | |
| P4M | Mass Spectra (12) | P15 | Pyrolysis (90) |
| P4N | NMR Spectra (194) | P16 | H-Bonding (35) |
| P4R | IR Spectra (9) | P17 | Electrochemical (301) |
| P4U | UV Spectra (23) | P18 | Brønsted (121) |
| | | P19 | Esterification (238) |
| P5 | Miscellaneous Reactions (529) | P20 | Photochemical (48) |
| | | P21 | Hydrogenation (16) |
| P6 | Substitution | P22 | Isokinetic (2) |
| P6E | Electrophilic Substitution (264) | P23 | Reduction (91) |
| P6N | Nucleophilic Substitution (1192) | | |

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

III. Using the Databases

The primary idea to keep in mind is the structure of the system as outlined in Scheme I.



Scheme 1 outlines a biodynamic system that is like an electronic set of two books. One can read one book or the other or merge the two and page through them. The difference is that since paper is not involved, the books undergo continuous updating. Our present goal is about 100 new QSAR/month. Still, this is not enough to keep abreast of the voluminous flow of new literature.

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. Additional help can be obtained by calling Mike Medlin at BioByte Corporation [(909) 624-5992].

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 contexts in which it can appear.

| | | |
|------------------------|---------------|--|
| E.coli in mouse | a word | (both leading and trailing blanks) " <u>in</u> " |
| influenza | start of word | (leading blank, but no trailing blank) " <u>in</u> " |
| brain | end of word | (trailing blank, but no leading blank) " <u>in</u> " |
| pyridine, guinea | 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.

Any word or parameter can be negated by prefacing it with **NOT** (discussed later). This causes the result to be the reverse (logical complement) of what it would otherwise be.

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 4). Occasionally, because of a bad entry, the system will crash, enter **unlock** followed by **QSAR** and then **data physical** to restore the system.

Table 4: 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 151,840 data points (compounds). The physical base contains 88,280. 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

The search and show feature is accomplished in two steps: first one enters the **search** mode from the main menu (Table 4), which is obtained by entering **data** and then **sea**, and press **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 5.

Table 5: 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 | | VMS |
| | | max/min | | |
| | 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 database 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 3,968 sets, which is far too many for ionization alone. As entered the command locates, via string searching, 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 1,732. 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 8,900 sets and the search requested is 'Action...BROMIN'. To perform the search, enter search, then press return. 180 hits are found. To peruse the 'catch', enter show and then 4 (action) 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 then enter bl to blank out the last command then 2 P7 (Table 3) followed by 4 bromin and then press return, which shows the status check indicating that the search will be made on the hits made previously. Enter search, which finds 55 sets of bromine addition reactions. If you want to see bromine in radical reactions, you need to return to the original 180 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 55 hits. One might want to view these sets showing 'system', 'action' and 'terms', and would enter show 1 4 15. To return from any Menu to the Main Menu, enter q. To illustrate another way to search the bio database, enter data bio, press return then sea and enter 1 HIV to find 204 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: 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 3,867 hits. Adding to the search 1 not % eliminates mixed solvents and would reduce this to 1,633. Now searching with 2 " P1 " finds 617 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 8,900 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 725 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 797 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 (1,487). 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⁺, S⁻ and S' (computer-compatible letters for the Greek symbols σ , σ^+ , σ^- and σ') 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⁺,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⁺, S⁻, etc.), the following steps should be followed. Entering **15 " S** finds 7,979 sets including both positional and 'special' sigmas. To remove the latter, enter **15 not S- S+ S' S** and 4,209 remain. Notice that the 'not' command is not necessary in searching for the 'special' sigmas because of string searching. For example, **15 " S-** finds all QSAR based on sigma-minus parameters (1,145), 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:

| | |
|--------------------------|------------|
| 15 " S | 7,979 hits |
| 15 not S- S+ S' S | 4,822 hits |
| 15 not MR | 4,787 hits |
| 15 not B1 B5 | 4,556 hits |
| 15 not ES | 4,402 hits |
| 15 not logP PI | 4,278 hits |
| 15 not **2 bilin | 4,235 hits |

The symbol ****2** represents squared terms such as S² or PI² and bilin represents bilinear equations.^{14 (pg. 195)} Note that the NOTS must be listed as a separate entry. If one wants QSAR with only (S), a direct approach is to enter **15 " S," " S "**. This includes substituents σ , X etc., but excludes σ^- , σ^+ , σ' , σ .

Moving to **show** and entering **15** we see that except for a very few examples, all QSAR are

based on the single term S.

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 63 eq. parabolic in +
15 bilin locates 27 sets bilinear in any parameter

Range Searching

Searching the physical database with **15 " S "** finds 4,030 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. (1,229 examples with a slope between -0.8 and 1). 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 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 -0.769 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

In Section 3 of Table 1, entries such as phenols, X-C₆H₄-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 613 sets. For a quick overview of the compounds studied, enter **show** and then **3**. Scrolling through these examples takes less than ten minutes 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 incomplete, answer. A more systematic method 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 202 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 157 NMR studies. Searching with **2 P4N** locates 194 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 141 sets. These can be narrowed by checking certain years; *e.g.*, **5 (1998) (1999) (2000)**. Five QSAR came from recent papers published in these years. Entering **5 Bordwell, Cheng** isolates 35 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)**; 150 hits. **5 (1970) (1971) (1972)**; 946 hits. **5 (1998) (1999) (2000)**; 410 hits. It is necessary to put parentheses around the year to distinguish it from page numbers. 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 1,775. 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: 7162, 5983, and 4081. Thus out of 8,900 sets only about half have been published outside of these three journals. Entering **show** 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 with 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 1,192 hits for nucleophilic substitutions, and then **16 2.6 < S- < 2.97** which narrows the sets of interest to 12.

The first command isolates all nucleophilic substitution reactions. The second finds those with ρ in the range 2.6 to 2.97 (12 QSAR). Going to **Sh** and listing with /sort = **16 1 3 4 15 16 18** followed by **S-** yields the following representative examples.

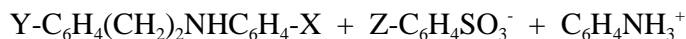
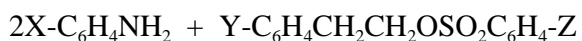
| | System | Compound | Reagent | ρ |
|----|--------------|---|---|--------|
| 1. | 20° Aqueous | X-C ₆ H ₄ OCO(CH ₂) ₃ NMe ₂ | Internal | 2.60 |
| 2. | 20° Methanol | 2-p-nitrophenoxy-3-NO ₂ -5-S-thiophenes | piperidine | 2.61 |
| 3. | 20° Aqueous | X-C ₆ H ₄ OCO(CH ₂) ₄ NMe ₂ | Internal | 2.65 |
|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

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** 219 hits
2. **15 " S," " S "** 147 hits
3. **15 not ES I MR D F B1 B5 **2 bilin** 32 hits
4. **18 n>70** 2 hits

The first search isolates all QSAR containing 3 terms. The second ensures that a term 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)



$$\text{Log } k_2 = -1.32(\pm 0.05) ,x - 0.13(\pm 0.01) ^+,y + 1.08(\pm 0.03) ,z - 3.93(\pm 0.01) \quad (1)$$

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

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 2,743 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 862 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

Information on the kinds of substituents that have been used in QSAR can be found in two ways. Common names can be employed or using SMILES (for those not familiar with the SMILES notation, see section VI). Searching on **11 CH3** locates 1,376 QSAR, **11 Me** finds 7,131 and **11 methyl** uncovers 86. Thus, out of 8,900 QSAR, 8,593 have a parent compound containing a methyl substituent. More complex structures can be isolated. **11 2-NH2 O-NH2** finds 82 QSAR with an amino group ortho to the functional group. **11 3,4,5-Cl 3,4,5-Cl3** locates

9 QSAR; **11 CF3** uncovers 1,167; **11 SF5**, 9; **11 NHSO2Me**, 6.

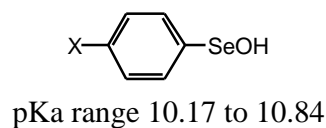
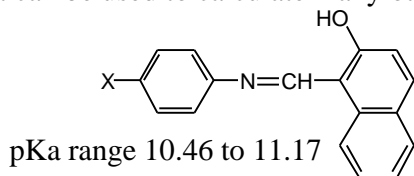
The SMILES approach can locate more complex 'substituents'. From the search menu, enter **12** and then enter **thiophene** and its SMILES is presented and one is asked if editing is desired. Answer **Yes**. Now place an asterisk on positions of interest *e.g.*, (*)clcccs1 yields *
Going to the search mode and searching finds 38 QSAR.

Our system contains the names and SMILES (53,647) for drugs and common chemicals. Again using **12** and entering **aspirin**, one obtains the SMILES and the 2-D structure. One can enter an asterisk in parenthesis at any point for a substituent.

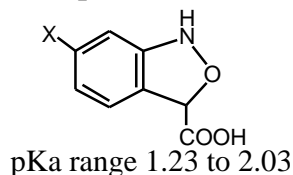
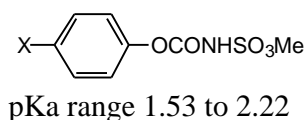
One of the most important classes of data in the phys databank is that of pKa values and ionization constants. These have been entered as the authors reported them without trying to place them on a common scale. The following search illustrates the potential where one wants to find solutes whose aqueous pKa's fall between 10 and 12.

1. **15 pK** 1,736 hits
2. **15 not logK** 1,646 hits
3. **1 aqueous** 1,172 hits
4. **1 not %** 586 hits
5. **14 10<pka<12** 9 hits

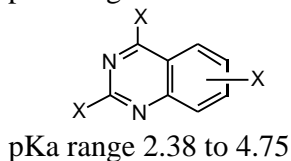
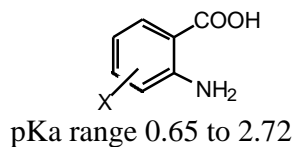
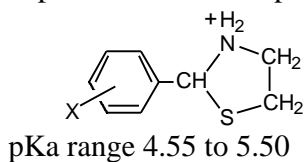
The potential for calculating new pKa values is great. We have 1997 _P, 1232 _M, 398 ⁺, 375 ⁻ and 887 * values. We have established that _M can be used with ⁺ and ⁻. The 565 pKa QSAR for the aqueous system can be expanded with these substituent constants. The first step isolates all pKa's, some of which have a different modifier than a. Step 2 then eliminates examples where pK is used as a dependent variable. Steps 3 and 4 isolate studies based in pure water and 5 illustrates how one can isolate compounds having pK's of values within a given range. The following examples illustrate the possibilities. Remember that each of the 9 hits is a QSAR that can be used to calculate many other values.



Searching for stronger acids we can change step 5 to **14 2<pka<3** which isolates 15 examples among which are:



By changing step 5 to **14** $0 < \text{pKa} < 6$ followed by **18** $n > 10$ sequesters 57 sets having eleven or more datapoints. Three examples are:



Of course one could search for any particular class of compound via the SMILES or the Merlin approach.

From the search mode enter **12** and in the projected line enter the SMILES for the structure of interest or if there is a common name enter that and the SMILES will be generated. Entering benzoic acid, its 2-D structure is displayed. Moving to **search**, we find 253 QSAR. To examine the solvent systems used move to **show** and enter **1**, that will depict the wide variety of solvents that have been studied. Searching the bio section with pKa, we find 84 examples where pKa is the independent variable.

Now doing a MERLIN search, enter **13** and then the SMILES or compound name. This similarity search finds all examples where one or more H of the selected compounds has been substituted by another element. For example, **benzoic acid** finds 1,600 QSAR.

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 (none in Table 1). Checking the Phys data illustrates the useful searches. The command **18 n>20** in the search mode finds only 524 out of 8,900 QSAR that are based on 21 or more data points. **18 n>10** hits 2,166 sets and **18 n>4** locates 7,667 based on 5 or more compounds. 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 wants a minimum of 5 data points/variable with well spread parameter values. We have kept sets with 4 too, since they are better than nothing.

The quality of the correlations can be evaluated by means of the correlation coefficient r . Entering **18 r >.90** shows that 8,764 sets having correlation coefficients greater than 0.90. **18 r >.95** yields 8,144 QSAR and **18 r >.99** finds 3,731. Quality can also be checked with respect to the standard deviation S . **18 S<.10** hits 4,618 QSAR with standard deviations less than 0.1 while **18 S<.20** makes 7,017 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 1,217 examples where NO calculated value in the set deviates by more than ± 0.05 from the experimental value. This is a very stringent standard. Relaxing the standard to ± 0.2 finds 4,332 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. Keeping them in the equation can distort QSAR meaning.

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 (613). Now entering **18 omit>0** isolates all QSAR with one or more starred data points (256). Moving to **show** and entering **11** lists all substituents for each of these data sets along with the starred substituent. Now go to **sea** enter **bl** and add **2 P6N** followed by **18 omit>0**. This isolates 411 QSAR for nucleophilic substitution with one or more starred points.

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 529 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 can be determined.

The following 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 | | | Example 2 | | | Example 3 | | |
|-----------|-----------------------------|------------|-----------|-----------------------------|------------|-----------|-------------------|------------|
| 15 | S+ | 1,928 hits | 2 | P6N | 1,192 hits | 15 | S+ | 1,928 hits |
| 18 | n > 10 | 398 hits | 18 | n > 10 | 330 hits | 18 | n > 15 | 179 hits |
| 18 | r > .98 | 198 hits | 18 | r > .98 | 330 hits | 18 | r > .91 | 2 hits |
| 19 | -.1 < dev < .1 | 20 hits | 19 | omit < 1 | 330 hits | | | |
| | | | 19 | -.1 < dev < .1 | 69 hits | | | |

6. SMILES

Almost all structures in the databases are entered in the SMILES notation (section VI) 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 (from **regression**, enter **data** and then **sea**) in the Phys bank 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 340 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. Many of these sets can be removed by **3 not misc**. For more discussion of SMILES, see section VI and ref. 16 and 17.

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 21 examples. Moving to **show** and entering **15** it is seen that 7 QSAR are based on + , 12 on - and 2 on ^ . Entering **4** (action) shows that 9 examples involve bromination, 7 iodination, 4 on chlorination, 2 on nitrosation.

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.

From **search**, enter **12** and then **strychnine**. Two data sets are found; move to **show** and enter **5** from these the set numbers are found to be 881 and 886. Enter **reg** and then **load /d 886**. Entering **summary** lists key items in the data set.

| | | | | | |
|---------------|--|-------------|----|-----------|----|
| Dataset name: | PHYS_886 | | | | |
| Substituents: | 42 | Parameters: | 4 | SMILES: | 42 |
| Active: | 32 | Starred: | 10 | Inactive: | 0 |
| System | LOG P OCTANOL VS LOG P DIETHYL ETHER | | | | |
| Class | P14 ; Partitioning, binding | | | | |
| Compound | H-ACCEPTOR SOLUTES | | | | |
| Action | EQUATION B | | | | |
| Reference | LEO,A. HANSCH,C. J.ORG.CHEM. 36,1539(1971) R1309 | | | | |
| Source | LEO POMONA | | | | |
| Check | UNKNOWN | | | | |

Date 2001 September 24
Parameters YPRED DEV LP-ETHER LP-OCT

Next enter **eq /run** which shows the stored equation.

qsar> eq /run

LogP-ETHER = 1.142(0.128)LogP-Octanol - 1.070(0.122)

N = 32 R = 0.957 Q2 = 0.900 SS1 = 37.809 DEV+ = 20
DF = 30 R2 = 0.916 S = 0.326 SS2 = 3.187 DEV- = 12

seepparameters:

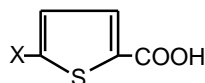
| | | | |
|-------|-----|----------|--------|
| 1 | 2 | 3 | 4 |
| YPRED | DEV | LP-ETHER | LP-OUT |

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**.

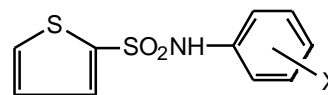
The following steps isolate all compounds containing a thiophene moiety:

1. **15 pKa** 1,592 hits
2. **13 thiophene** 29 hits **13** calls for a MERLIN search

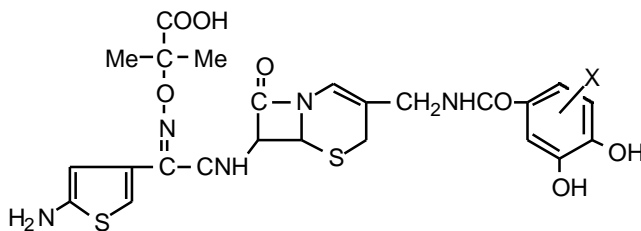
Examples of the 29 hits are:



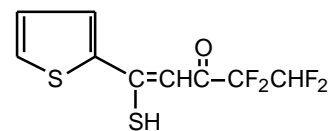
wide range of pKa
in various studies



pKa range 7.1 to 10.27



pKa range 5.5 to 8.5



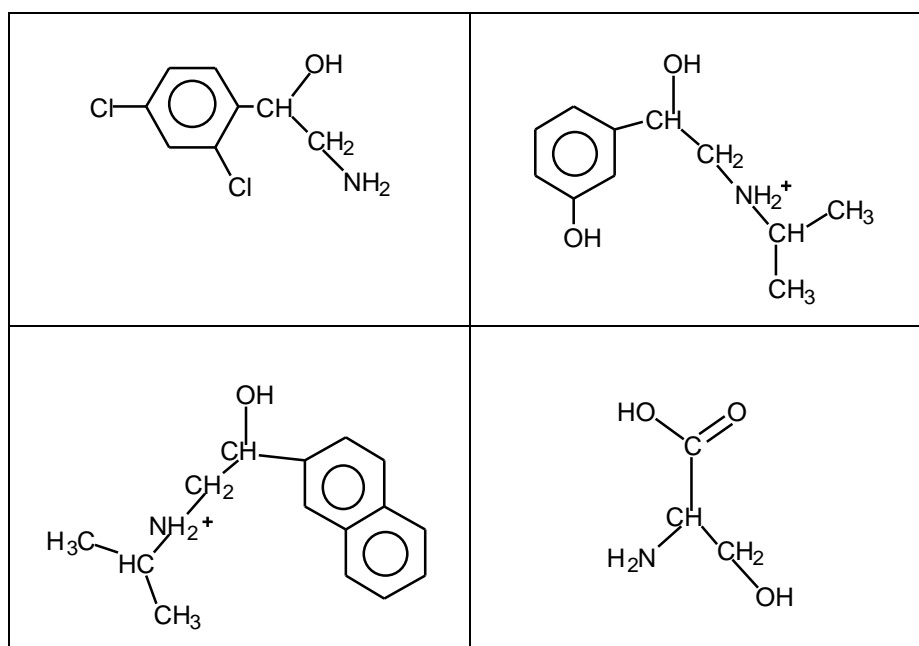
pKa 5.25

The range of structures for which pKa values have been determined is remarkable. The above examples are from studies in aqueous solution. In the last example, there were a variety of

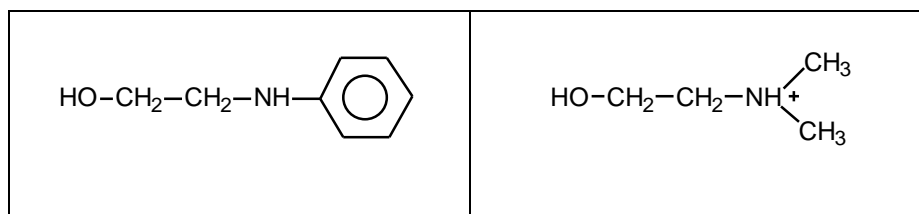
substituents where thiophene is attached, but only one example with thiophene. Other examples of the use of MERLIN will be discussed.

7. MERLIN and SMARTS

A good example of using MERLIN for searching would be finding the sets which contain the ethanolamine moiety. The SMILES entry is simple: NCCO, but if we wish the amine to be primary and the carbons to be 'hydrocarbon' (not carbonyl) the SMILES should be followed by this SMARTS: [N&H2][C&H][O&H]. SMARTS is just a scheme for putting inside brackets any restrictions on the bonding environment on any of the atoms in the SMILES. This search gives 169 hits when both bio and phys databases are searched. Some examples are:



If we allow the amine nitrogen to be secondary [N&H], the number of hits rises to 2,181. Some examples are:



Returning to the Phys database (**reg** then **data phys** then **sea**) we need to consider the subsection on miscellaneous reactions. Sequestering this group of QSAR by **2 P5** yields 529

QSAR. Of course, one could page through these by going to **show** and entering **1 3 4 15 16 18** to see the unusual reactions. However, if one has a particular reaction in mind, the whole database can be string searched as follows:

| 4 | <u>reaction name</u> | (# of hits) | |
|----------|-----------------------------|--------------------|----------------------------|
| | <u>Diels</u> | 44 | Diels-Alder reaction |
| | <u>Friedel</u> | 3 | Friedel Craft reaction |
| | <u>Cyclization</u> | 38 | |
| | <u>Mercuration</u> | 12 | |
| | <u>Salt</u> | 25 | salt formation |
| | <u>Alkyl</u> | 32 | Alkylation |
| | <u>Decomp</u> | 111 | Decomposition |
| | <u>Wolf</u> | 6 | Wolf-Kishner reduction |
| | <u>Dipole</u> | 15 | Dipole moments |
| | <u>Decarboxyl</u> | 27 | Decarboxylation |
| | <u>Racemi</u> | 2 | Racemization reactions |
| | <u>Meerwein</u> | 1 | Meerwein-Pondorf reduction |
| | <u>Bromi</u> | 268 | Reactions with bromine |
| | <u>Hydration</u> | 53 | Hydration reactions |

8. 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** (password = press **return**) followed by **reg**. The screen prompt becomes **qsar>**. Entering **load /d 464** (set number) transfers the set to the workspace. Any data set can be viewed in the following ways by entering the indicated commands:

Sum Lists key items of data set

System Human

Class B6H; Human

Compound X-2-amino-4-nitro benzenes

Action Sweet taste

Reference Blanksma, J.J.; Hoegen, D. *Rec. Trav. Chim. Pays-Bas* 65, 333, 1946. See Hansch & Deutsch, *Nature* 211, 75, 1966.

Source Hansch

| | | | | | | |
|--------------|--|---------|-------|-------|------|-------|
| Check | L. Zhang | | | | | |
| Seep | log RBR + Pi S ER PE Clog P | | | | | |
| Seeeq | log RBR = -0.66(±0.28) + 1.32(±.24) Clog P - 0.07(±0.48) (1) | | | | | |
| | $n = 9$ $r^2 = 0.973$ $s = 0.132$ $q^2 = 0.936$ | | | | | |
| Pred | | log RBR | Ypred | Dev | S+ | ClogP |
| | H | 1.60 | 1.592 | .008 | 0 | 1.258 |
| | F | 1.60 | 1.833 | -.233 | -.07 | 1.405 |
| | OMe | 2.52 | 2.426 | .094 | -.78 | 1.498 |
| | OC ₂ H ₅ | 3.15 | 3.127 | .023 | -.78 | 2.027 |
| | OC ₃ H ₇ | 3.70 | 3.827 | -.127 | -.78 | 2.556 |
| | Cl | 2.60 | 2.575 | .025 | -.11 | 2.055 |
| | Br | 2.90 | 2.813 | .087 | .15 | 2.255 |
| | I | 3.10 | 3.104 | -.004 | .13 | 2.465 |
| | Me | 2.52 | 2.392 | .128 | -.31 | 1.707 |

In this research, 'RBR' represents the relative potency of the compounds in several test humans. 'Source' indicates who entered the data, 'check' indicates the name of the person checking the results. 'Parameters' shows all parameters considered in the study. '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** , 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 5, enter **depict 5**, to view all structures up to 9, enter **depict 9**. To see those between 3 and 8, enter **depict 3,8**.

D. Biological Database

The biological database is more difficult to master than the Phys section. To get some feeling for this, think of the complexity of organic chemistry. One would not expect to quickly read through a textbook of organic reactions and then start to develop a new synthesis for quinine. To study the interface between chemistry and biology, one needs all of the chemistry and biology that one can master over a number of years. We believe that our system will greatly expedite the process.

Our database of 8,900 Bio QSAR can now be used as a chem-bioinformatics system 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. An important aspect of the bio database is that it contains many complex structures of drugs for which SMILES and synonyms have been entered even though they may not appear in a QSAR. There are 151,400 SMILES and 50,290 common names which, upon entry, yield the SMILES.

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 | cat papillary muscles |

Entering **1 sea** lists all of the 8,900 systems that have been studied. Paging through a few thousand gives an idea of what all has been studied. The more sophisticated work has high set numbers.

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 **data** then **sea** followed by **2 B4B** yields 814 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. Entering **1 3** yields both. Such a survey of all of the work on bacteria can be made in less than 30 minutes. It would take days in the library or on line. Returning to **search** and entering **bl** then (blanking out aureus and entering) **1 aureus**

makes 153 hits most of which are for *S. aureus* and *A. aureus*. Now entering **15 not logP PI RM** eliminates all QSAR with hydrophobic terms, leaving 23 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 10 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 (coefficient with).

Combining 'system' and 'class' in searching can also be effective in other ways. Searching with **1 " rat "** results in 1,000 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 221. 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 enzymes oxidoreductase, the organelle microsome, and whole animals. This search **2** followed by **B2A B3B** and **B6A** to yield 1,594. Now we could isolate QSAR containing a $+$ term for comparative analysis by **15 S+** to find 128 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, moving to **sea** and using **4 metab** finds 27 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:

| | | |
|----------|-----------------------|--|
| 4 | <u>bind</u> | 1,321 hits for binding of chemicals to various bio systems |
| 4 | <u>uncoup</u> | 43 hits for uncoupling of oxidative phosphorylation |
| 4 | <u>Biocon.</u> | 37 hits for bioconcentration in various ways |
| 4 | <u>Glutath</u> | 17 hits in which glutathione was employed |
| 1 | <u>Glutath</u> | 30 hits including glutathione transferase |
| 4 | <u>Inflam</u> | 18 hits on anti-inflammatory agents |
| 1 | <u>Cycloo</u> | 61 hits on cyclooxygenase 1 and 2 |
| 1 | <u>HIV</u> | 218 hits on human immunodeficiency virus |
| 4 | <u>Mutag</u> | 44 hits on mutagenesis |

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. **15** **logP** 5,038 hits
2. **15** **not logP' **2 bilin** 3,476 hits
3. **15** **" S," " S "** 298 hits

Note that log P also isolates sets based on ClogP via string searching. The second command removes a few 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 **2, which indicates a squared term and bilin a bilinear QSAR which makes for easier first time comparisons. Step 3 isolates QSAR based on simple terms (*i.e.*, not +, -, * or E_s). 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-C₆H₄N(CH₂CH₂Y)₂ with σ of about -2 to -1.5 and a small generally negative slope with log P. The crucial factor for reactivity of these agents with DNA is the electron density on N. Set 1780 for the enzymatic acylation of X-C₆H₄-NH₂ which is dependent on the electron density on NH₂ has a σ of -2.1. At σ ~ 0.3 to 0.5 we find examples of uncoupling of oxidative phosphorylation in mitochondria. The dependence of biological activity on σ and comparison with examples from physical organic chemistry has been reviewed.^{7, 9a} 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 σ^+ as follows:

1. **12** **aniline** 143 hits
2. **15** **S+** 16 hits
3. **16** **-10<S+<0** 12 hits
4. **2** **B2A** 8 hits

Most of the 16 examples involve oxidoreductases (B2A) with, possibly, radical mechanisms.⁸

Comparing a new QSAR with those in hand is more easily done with the Bio data where we have often been able to standardize the dependent variable log 1/C. Except when marked as log 1/C, C refers to molar concentration needed to produce a standard reaction in a standard time.

The following is illustrative.

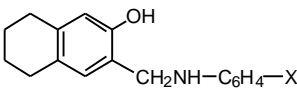
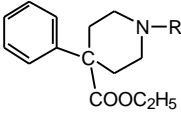
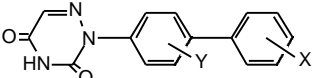
- | | | |
|----|---|------------|
| 1. | 15 " <u>log1/C</u> " | 5,618 hits |
| 2. | 15 <u>not **2 bilin</u> | 4,195 hits |
| 3. | 15 <u>not " S</u> | 3,213 hits |
| 4. | 15 <u>not Es B1 B5 MR Pi Pka I D</u> | 1,571 hits |
| 5. | 16 <u>.7<logP<.9</u> | 347 hits |
| 6. | 16 <u>0<const<0.5</u> | 48 hits |
| 7. | 3 <u>not misc</u> | 37 hits |

The first step insures that all dependent variables are uniform (*i.e.*, eliminates log 1/C'). The second eliminates all nonlinear QSAR. Step 3 removes all equations with any kind of term. Step 4 removes QSAR having any of the indicated independent variables. 5 selects only those with slopes between 0.7 and 0.9 and 6 insures the constant term lies between 0.0 and 0.5. The last step removes QSAR based on sets of miscellaneous compounds.

Some examples are:

- | | |
|--|----------|
| I ₅₀ chloroplasts by X-C ₆ H ₄ NHCOCH(CH ₃) ₂ | set 223 |
| Inh. cholinesterase from electric eel by FCH ₂ COOR | set 493 |
| Inh. potassium uptake of liver mitochondria by X-C ₆ H ₄ COOCH ₂ N(C ₂ H ₅) ₂ | set 880 |
| I ₅₀ of Na ⁺ -k ⁺ in mouse brain synaptosomes by CH ₃ COOR | set 3251 |

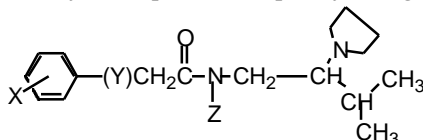
Now looking for examples 1000 times more potent we can change step 6 to **3<const<3.5** and isolate 176 examples such as:

- | | | |
|---|--|----------|
| I ₅₀ Human polymorphonuclear leukocytes by |  | set 5160 |
| I ₅₀ binding of [H ₃]-Naloxone to rat brain opiate receptors by |  | set 4919 |
| Increase in nerve membrane potential by 20mV in mollusk buccal ganglion by salicylates. | | set 411 |
| MIC of eimera tenella in leghorn cockerels by |  | set 5073 |

The above examples of model mining illustrate how simple examples can be found. One could of course focus a search by placing all sorts of limitations on the search engine. One might study cells and whole animals by entering **2 B4** and **2 B6A**. This garners 3,101 QSAR. Now the following steps might be of interest:

1. **15** " log1/C " 2,613 sets
2. **15** not **2 bilin 1,862 sets
3. **15** logP 1,098 sets
4. **15** " S," " S " 95 sets – isolates only QSAR containing a term
5. **16** .7<logP<1 8 sets – log P terms in the range 0.7 to 1.0 ant.
6. **16** .8<" S," " S "<1.5 4 sets – a term in the range 0.8 to 1.5

*I*₅₀ displacement of [3H] bre mazocine from opioid receptor from guinea pig brain by set 5068



$$\log 1/C = 0.84(\pm 0.28)\text{Clog P} + 1.30(\pm 0.45) X - 1.13(\pm 0.53)\text{MR} + 3.90(\pm 1.0) \quad (2)$$

$$n = 14, \quad r^2 = 0.912, \quad s = 0.229, \quad q^2 = 0.785 \quad 1 \text{ outlier}$$

*I*₅₀ antinociceptive activity of the above compounds in mice set 5069

$$\log 1/C = 0.73(\pm 0.26)\text{Clog P} + 0.88(\pm 0.46) X + 2.92(\pm 1.01) \quad (3)$$

$$n = 11, \quad r^2 = 0.889, \quad s = 0.212, \quad q^2 = 0.809 \quad 1 \text{ outlier}$$

The same sets of rather complex chemicals yield rather similar QSAR for two different systems. The largest difference being in the intercepts and the negative MR term. The guinea pig system is about ten times more sensitive. Small changes in the linker unit (Y) do not appear to be especially important. Another search for more complex QSAR the activity of which might be dependent on radicals is the following:

$$\mathbf{15} \quad \mathbf{logP} \quad 5,038 \text{ sets}$$

$$\mathbf{15} \quad \mathbf{not \text{**}2 \text{ bilin}} \quad 3,566 \text{ sets}$$

$$\mathbf{15} \quad \mathbf{S+} \quad 103 \text{ sets}$$

$$\mathbf{16} \quad \mathbf{.6<logP<1} \quad 18 \text{ sets}$$

$$\mathbf{16} \quad \mathbf{-2<S+<0} \quad 10 \text{ sets}$$

Two examples of interest are:

*I*₅₀ sheep vesicle prostaglandin cyclooxygenase by phenols set 1593

$$\log 1/C = -1.71(\pm 0.25) + + 0.69(\pm 0.12)\text{Clog P} + 1.80(\pm 0.32) \quad (4)$$

$$n = 25, \quad r^2 = 0.933, \quad s = 0.186, \quad q^2 = 0.911$$

We have often found + to correlate radical reactions.⁸

Acetyltransferase of acyl group from p-nitrophenyl acetate to X—C₆H₄NH₂ set 1431

$$\log V_{\max}/K_m = -1.25(\pm 0.46) + + 0.89(\pm 0.46)\log P + 0.65(\pm 0.31)\text{Es}_3 + 1.3(\pm 0.74) \quad (5)$$

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

In equation 4, δ^+ suggests a reaction whereby anilines are converted to phenoxyl radicals. In eq. 5 increase in the electron density on NH_2 makes it a more potent reagent for displacing the $^-\text{O}-\text{C}_6\text{H}_4\text{NO}_2$ moiety.

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. | 15 | <u>logP</u> | 5,038 hits |
| 2. | 15 | <u>" log1/C "</u> | 3,299 hits |
| 3. | 15 | <u>not logP' **2 bilin</u> | 2,256 hits |
| 4. | 16 | <u>.7<logP<.9</u> | 444 hits |
| 5. | 16 | <u>1.5<const<2</u> | 73 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 these conditions 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 73 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 51 examples with intercepts above 4. These are for rather complex chemicals (compared to alcohols) and the activity would not be considered nonspecific.

The major difference between the parameters of the physical and biological QSAR is the importance of the hydrophobic terms $\log P$ and P_i . Out of 8,900 bio QSAR, 4,378 (49%) contain such terms.

Considering the five major categories the following disposition is found.

| | QSAR with hydrophobic terms (log P or PI) | Total QSAR in class | Percentage |
|-----------------|---|---------------------|------------|
| Enzymes | 1999 | 3983 | 51% |
| Receptors | 733 | 1665 | 44% |
| Organelles | 319 | 433 | 70% |
| Cells | 1755 | 2497 | 73.7% |
| Organs | 672 | 1008 | 66.7% |
| Whole Organisms | 1135 | 1418 | 80% |

As might be expected, enzymes have rather hydrophilic terms and whole organisms have the most hydrophobic. In the last few years we have found a great increase in studies with more or less purified receptors. We now have 1,665 such QSARs of which only 44% have hydrophobic terms. Electronic terms (including HOMO and LUMO) appear in 2,067 and the steric parameters Es, B1, B5 and L appear in 2,167. Clearly hydrophobic terms are most important. Of course our means for defining steric effects is limited. We have recently reviewed QSAR lacking positive hydrophobic terms.⁹

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 design¹³ is obviously important for bioavailability, 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,⁹ but the reason for this is not always clear at present. Compounds that appear to react with DNA seem to fall in this class.

Searching for studies of a particular drug or similar chemicals is illustrated in the following table. The database contains over 52,500 common chemical names as well as official names of drugs currently on the market or discontinued, or interesting, but not yet on the market. Table 7 shows the results of both SMILES or MERLIN searches on a variety of compounds. For SMILES based search enter **12** from the **search** mode and for MERLIN enter **13**.

Table 7

| | | hits | | | hits |
|--------|-------------|------|--------|--------------------|------|
| SMILES | Mescaline | 5 | SMILES | Testosterone | 21 |
| MERLIN | Mescaline | 22 | MERLIN | Testosterone | 41 |
| SMILES | Epinephrine | 14 | SMILES | Phenoxyacetic Acid | 7 |

| | | | | | |
|--------|--------------|-------|--------|--------------------|----|
| MERLIN | Epinephrine | 21 | MERLIN | Phenoxyacetic Acid | 69 |
| SMILES | Naproxen | 9 | SMILES | Isoniazid | 4 |
| MERLIN | Naproxen | 10 | MERLIN | Isoniazid | 15 |
| SMILES | Methotrexate | 13 | SMILES | Adamantane | 0 |
| MERLIN | Methotrexate | 15 | MERLIN | Adamantane | 81 |
| MERLIN | [Pt] | 6 | SMILES | Glucose | 5 |
| MERLIN | [Se] | 19 | MERLIN | Glucose | 39 |
| MERLIN | [Fe] | 5 | SMILES | Cortisone | 13 |
| MERLIN | [P] | 265 | MERLIN | Cortisone | 13 |
| MERLIN | [S] | 2,587 | | | |

In the examples of Pt, Fe, Se, P and S only a MERLIN type search is possible since no QSAR have been reported for the bare elements. This yields all compounds that contain such an element. It is interesting that adamantane itself has never been tested, but after the discovery of the antiviral activity of aminoadamantane there was a wild flurry of testing derivatives of adamantane or using it as a substituent. In the case of cortisone it was surprising to find no 'similar' compounds. The large number of hits with phenoxyacetic acid is due to the great interest in these chemicals as weed killers. In fact QSAR was developed out of interest in this class of chemicals.¹⁰

The same approach can be used with CMR or any other parameter. There are 1,292 QSAR with CMR terms of which 190 have CMR² terms. Of the 190, 55 have positive CMR² terms and 135 have a corresponding negative CMR term. That is, the shape of curve is that of an inverted parabola. The search used to uncover this information is as follows:

1. **15 CMR**2** 190 hits
2. **16 0<CMR**2<10** 148 hits
3. **15 CMR** 1292 hits
4. **16 -10<CMR<0** 734 hits

We have taken the result of step 4 to imply that an allosteric reaction is occurring between ligand and receptor. That is, at first interaction between ligand and receptor decreases as CMR increases, but then at an inversion point as CMR continues to increase activity turns around and increases.² Clearly a change in mechanism has occurred. We assume this to be the result of change in shape of the receptor.³¹ We have a few examples where such inverted parabolas are found with Clog P or molar volume, but most occur with CMR. The data must be plotted to be sure that the inversion point is well defined.

Pauling and Pressman¹¹ and Agin, *et al.*,¹² first used molar refractivity in attempts to rationalize biological processes, although the concept of bond polarizability had been used for many years to help understand simple chemical reactions. We use the following definition:

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

Where n is the refractive index, MW is the molecular weight and d is density. Many years ago, A. Leo developed an additive constitutive approach using interatomic bonds to calculate CMR from structure. Our program automatically calculates CMR and loads it for regression analysis. Our MR values are scaled by 0.1 to make them somewhat more equiscalar with log P. Although CMR is rather collinear with molar volume it is often definitely superior for correlation analysis.² Fortunately, during the first part of the 20th century, the Russian chemists routinely published index of refraction and density of new chemicals from which Leo devised the means for calculating CMR.

To search the database for compounds having log P_o (optimum log P), use the following commands:

- | | |
|--|------------|
| 1. <u>15 logP</u> | 5,038 hits |
| 2. <u>15 logP**2 bilin(logP) bilin(ClogP)</u> | 1,303 hits |
| 3. <u>17 1.5<logP<2.5</u> | 127 hits |

Command 3 narrows the catch to log P_o values between 1.5 and 2.5. To inspect the results, move to **show** and enter **17**. For parabolic equations, log P_o 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 10-G).

One of the advantages of the parabolic model is that an estimate of log P_o 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_o. It is instructive to compare log P_o for QSAR on cells with that on whole animals. Entering **2 B4** finds 2,497 QSAR on all types of cells. Then **15 logP** 2 bilin(logP) bilin(ClogP)** isolates QSAR where log P_o is found (434). 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_o. When these and those without confidence limits and partially ionized acids and bases are omitted, the remaining sets have an average optimum log P_o of about 4.3. Repeating the process for vertebrates by entering **2 B6A** finds 187 sets where the mean log P_o 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 may not be desirable.¹³

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 – 3 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.¹³ 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.

These results have some bearing on Lipinski's rule that to maximize an oral drug's probability of surviving development, the following properties are desirable: molecular weight < 500; number of hydrogen donors > 5; number of hydrogen bond acceptors < 10; Clog P < 5. His proposal has been criticized.^{13a} Our results indicate that in cell culture tests one would tend to obtain high activity drug candidates having a $\log P$ greater than 4, but this would not be ideal for animals or humans.

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

IV. Searching for New Lead Compounds

A. Locating QSAR that are Based on Highly Active Compounds

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 the fabulous quinolone carboxylates, QSAR played an important role in this process.⁵ 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. Viagra,

designed as a heart drug, turned out to be great for erectile dysfunction. 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 6 hits. All members of the three sets have $\log 1/C$ values > 9 .
2. **14 " log1/C "@max>9** finds 431 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 47 and 1,157 sets, respectively, for the two searching modes. At $\log 1/C$ of 7, we find 249 and 2,168 sets. Searching the entire bank with **15 " log1/C "** 5,611 sets are characterized by molar $\log 1/C$ terms.

These approaches can be generalized in the following ways:

3. **14 " log1/C " <=2** 965 hits
4. **14 2<" log1/C " <=4** 222 hits
5. **14 logP>6** 1,570 hits
6. **14 logP@max>6** 1,875 hits
7. **14 6<logP<8** 3 hits
8. **14 -1<S+<0** 11 hits

As illustrated in the above examples one can find ranges for any parameter. Example 7 finds only 3 QSAR based on $\log P$ values in the range 6 to 8. Example 8 uncovers all QSAR having $\log P$ values between -1 and 0 .

As the database increases, it becomes more arduous to get exactly what one wants without sorting out extraneous material. Another example of focusing is the following:

1. **2 B4 B3** 2,929 hits (cells and organelles)
2. **14 " log1/C " @max>7** 805 hits
3. **5 (1990) (1991) (1992) (1993) (1994) (1995)** 219 hits
4. **13 Imidazole** 11 hits

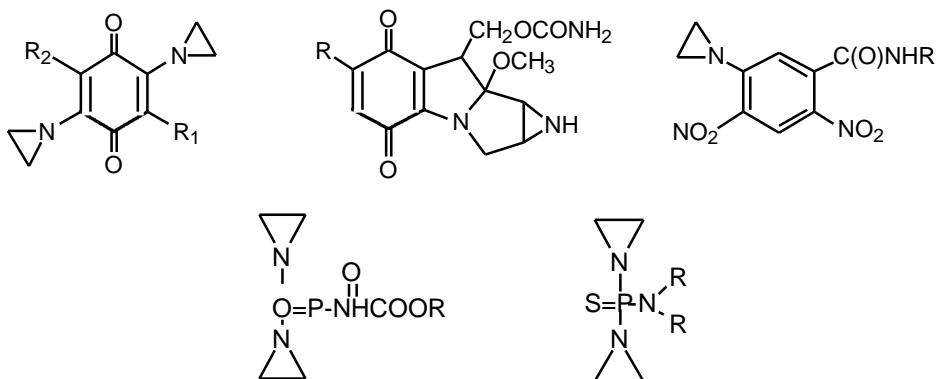
This searches all QSAR from organelles and cells published in 1990 to 1995 that contain an imidazole moiety.

B. MERLIN 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 has been 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, or **depict** , enters all structures and the set number, where these structures occur. The depiction process can be stopped at any point by entering **quit**.

Carrying out this process for **aziridine** in the bio database (entered as C1NC1 or by name) finds 323 examples. Entering **depict** , displays the 2-D structures and the set numbers in which they occur. Note that after the first the program says that's a lot of hits - - continue? Press **Y** to continue until the final number is obtained. In some instance, they occur in so many sets that the numbers are hard to read. The process can be stopped by **q** and a particular compound can be inspected by **depict #**.

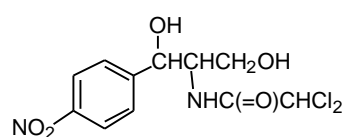
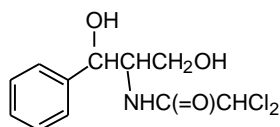


Note that even when aziridine is fused into a second ring system, it is also captured. Carrying out the same operation from the Phys mode finds 59 examples and of course searching from **data double** 319 hits on aziridines are made.

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

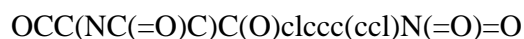
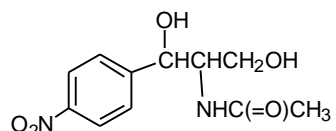
merlin and **c1ccccc1C#N** to find derivatives of benzonitrile. Choosing **n** starts the search and, on the prompt asking whether or not to continue, entering **y**, sequesters 692 derivatives of benzonitrile. Enter **depict 1**, displays them 4 at a time. It would be rather laborious to look at all 678 structures, but not impossible. The same search with **bio** yields 937. If we limit the search by asking for derivatives of 4-chlorobenzonitrile (SMILES N#Cc1ccc(Cl)cc1), only 12 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

Entering **depict**, we find 24 examples with substitution on phenyl ring or the C of CHCl2. If we wish instead to study only the variations on the methyl of the acetamido group, we can enter the SMILES for :



Now 20 examples are found and are viewed by entering **depict 1**. All are 4-NO2-phenyl analogs with variations in the side chain. Searching on chloramphenicol makes only 2 hits: chloramphenicol and the trichloroacetamido analog. There are no examples with substituents in the 2, 3, 5 or 6 positions.

We can search the double database (see Section V) from the search mode by entering **13** and **[Te]**. So doing finds 10 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. Substituent searching is different from a MERLIN search using the **13** entry (Phys base) where entering **13 C#N** locates 1,841 data sets each of which may have one or more compounds with a CN group.

V. 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 our primary reason for combining the two is to search the Phys database for mechanistic support for Bio QSAR. Of course, physical organic chemists would be very interested in the phys section.

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 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,^{8, 14, 15} 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, $\text{RCOOC}_2\text{H}_5(\text{CH}_3)$. 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 623 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 590 examples. To narrow the 'catch' enter **16 .7<ES<.9** and **sea** which reduces the number to 43. Going to **show** and entering **/sort=16 1 3 4 15 16 18** on request for a label to sort on, enter **ES**. The following representative examples illustrate comparative analysis.

1. Benzene set # 1384 phys (10018 double)
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. Aqueous 50% Ethanol set # 286 phys (8832 double)
 Trans C₆H₅CB = CACOOH
 Ionization 25 deg
 $\text{pKa} = 0.73(\pm 0.35)\text{Es} - 2 - 5.61(\pm 1.2) - A + 4.61(\pm 0.30)$
 $n = 9, \quad r^2 = 0.960, \quad s = 0.235$
3. Isopropanol set # 5245 phys (13791 double)
 X-COCH₃
 Reduction with sodium borohydride 50 deg
 $\log K = 0.75(\pm 0.17)\text{Es} + 2.19(\pm 1.0) * + 0.99(\pm 0.34)$
 $n = 7, \quad r^2 = 0.982, \quad s = 0.139$
4. Muscle Rectus Abdominis Frog set # 969 bio
 X-CH₂COOCH₂CH₂N(CH₃)₃⁺
 Contraction
 $\log 1/C = 0.76(\pm 0.17)\text{Es} + 1.19(\pm 0.37) + 4.65(\pm 0.28)$
 $n = 6, \quad r^2 = 0.984, \quad s = 0.076$
5. Acetonitrile set # 3424 bio (double 11970)
 X-pyridines
 Reaction with CH₃I 25 deg
 $\log K_2 = 0.79(\pm 0.24)\text{Es} - 2 - 2.21(\pm 1.9) + 0.08(\pm 0.30)$
 $n = 10, \quad r = 0.921, \quad s = 0.178$
6. Aqueous 50% Dioxane set # 6632 (double 15170)
 2-X-C₆H₄CONH₂
 Acid hydrolysis 90 deg
 $\log k = 0.80(\pm 0.09)\text{Es} - 2 + 0.08(\pm 0.12)$
 $n = 13, \quad r^2 = 0.972 \quad s = 0.086$
7. *S. aureus* set # 2028 bio
 BrCH₂CONHR
 I₁₀₀
 $\log 1/C = 0.83(\pm 0.30)\text{Es} + 1.59(\pm 0.30)\log P - 0.26(\pm 0.05)(\log P)^2 + 2.06(\pm 1.80) * + 3.20(\pm 0.40)$
 $n = 15, \quad r^2 = 0.953, \quad s = 0.208$

8. Aqueous set # 7063 (double 15609)
 $\text{RCH}_2\text{CONH}_2$
 Acid hydrolysis 75°
 $\log k = 0.84(\pm 0.12)\text{Es} + 1.62(\pm 0.20)$
 $n = 9, \quad r^2 = 0.974, \quad s = 0.088$
9. Chloroplast, Pea set #2453 bio
 $\text{X-C}_6\text{H}_4\text{NHCON(Me)CONHMe}$
 I_{50} of photosystem II
 $\log 1/C = 0.88(\pm 0.18)\text{Es} - 2 + 5.69(\pm 1.6)\log P - 5.63(\pm 1.7) \log(1 + 10^{\log P}) -$
 $0.55 (\pm 0.34) + 2.62(\pm 0.39)\text{F} - 2 - 2.51(\pm 2.3)$
 $n = 37, \quad r^2 = 0.937, \quad s = 0.203$
10. *M. Tuberculosis* set # 2553 bio
 2-X-4-CONHNH₂ pyridines
 MIC
 $\log 1/C = 0.89(\pm 0.29)\text{Es} - 2 - 3.70(\pm 1.1)\text{F} + 5.78(\pm 0.50)$
 $n = 17, \quad r^2 = 0.835 \quad s = 0.368$
11. Aqueous 60% Dioxane set # 3863 (double 12409)
 RCOOMe
 Alkaline hydrolysis
 $\log k_2 = 0.89(\pm 0.15)\text{Es} + 1.90(\pm 0.29)$
 $n = 7, \quad r^2 = 0.978, \quad s = 0.085$

The 11 equations are for widely different reactions. Seven represent simple chemical reactions and five 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. 8 and 11. Several correlate substituents ortho to the functional group (1, 5, 6, 10) 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. The molecules of equations 4 and 7 are closely related to the system defining Es that suggests a possible reaction involving nucleophilic attack on the carbonyl group. 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 better for modeling steric effects.

Another example involves radical reactions a subject of special interest to us.^{1,8} Searching the double database as follows is instructive.

1. **12** clccccclN 397 hits
2. **15 S±** 44 hits
3. **15 not **2 bilin** 43 hits
4. **3 not misc** 42 hits
5. **sh**
6. **/sort=16 1 3 4 6 15 16 18**
7. Sort on S+

The following examples are instructive:

Oxidation of X-C₆H₄NH₂ by MnO₂ in aqueous solution set # 6791 (double)

$$\log k = -3.80(\pm 1.28) + 1.49(\pm 0.65)$$

$$n = 6, \quad r^2 = 0.944, \quad s = 0.567, \quad q^2 = 0.876$$

Oxidation of X-C₆H₄NH₂ by vanadium V in 50% aqueous acetic acid set # 2393 (double 13904)

$$\log k_2 = -3.31(\pm 0.79) + 0.58(\pm 0.41)$$

$$n = 7, \quad r^2 = 0.958, \quad s = 0.263, \quad q^2 = 0.916$$

Hydrogen abstraction by (C₆H₅)₂NNN[•] of X-C₆H₄NH₂ in CCl₄ set #4703 (double 13249)

$$\log k_3 = -2.83(\pm 0.63) + 4.69(\pm 0.17)$$

$$n = 6, \quad r^2 = 0.949, \quad s = 0.138, \quad q^2 = 0.932$$

The above QSAR establish a radical mechanism for the oxidation of anilines with which the following biological QSAR can be compared.

Oxidation of X-C₆H₄-NH₂ by horseradish peroxidase I set #2393

$$\log k_2 = -3.17(\pm 0.46) + 0.34(\pm 0.08)\text{Clog P} + 5.14(\pm 0.21)$$

$$n = 9, \quad r^2 = 0.990, \quad s = 0.227, \quad q^2 = 0.973$$

Oxidation of X-C₆H₄-NH₂ by horseradish peroxidase II set # 1682

$$\log k = -3.00(\pm 0.51) + 0.25(\pm 0.10)\text{Clog P} + 4.52(\pm 0.24)$$

$$n = 8, \quad r^2 = 0.988, \quad s = 0.240, \quad q^2 = 0.972$$

Oxidation of X-C₆H₄NH₂ by cytochrome C peroxidase set # 3072

$$\log k_2 = -2.86(\pm 0.43) + 1.09(\pm 0.21)$$

$$n = 7, \quad r^2 = 0.984, \quad s = 0.193, \quad q^2 = 0.962$$

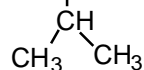
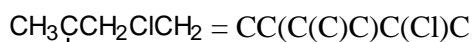
The toxic action of radicals is an important factor to understand in drug research and environmental toxicology. To help in this process, we have over 600 phys QSAR on radicals.

VI. SMILES Tutorial



Inventing the SMILES language for linear entry of complex structures of organic chemicals into computers was a truly major achievement by David Weininger.^{16,17} **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** (Section VIII) 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 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. Entering **sea** completes the search. Now entering **show**, one can peruse the resulting catch of QSAR. For example, entering **indole** from the bio database after the **12** command and then **searching** makes 11 hits.


The following examples illustrate the four basic rules.

1. Use ordinary atomic symbols. These represent aliphatic structures except for Cl and Br. B, C, N, O, P, S, F, Cl, Br, I
2. Hydrogens are not normally entered except on compounds such as pyrrole, indole, carbazole, in these instances [nH] is used. The small n, c, o, p, s represent aromatic atoms.
3. Other atoms and any charges are placed in brackets, *e.g.*, [Si]; [N+]. Also use H to fill the valences of non-ordinaries, *e.g.*, cl[SiH2]cl
4. Use upper case letters for aliphatic atoms and lower for aromatic. Except for the halogens. These are entered as Br, Cl. The second letter is entered in lower case, [Sn] for aliphatic tin, [Se] for aromatic or aliphatic selenium.
5. For nonlinear structures branches are enclosed in parentheses. Neopentane is written as C(C)(C)(C)(C).



Other examples:

- | | |
|--|---|
| a. $\text{CH}_3\text{CH}_2\text{OH}$ | CCO |
| b. $\text{CH}_3\text{CH}_2\text{NHCH}_3$ | CCNC |
| c.  | clccncl (the two "l"s serve to connect the linear structure to the cyclic aromatic ring). |
| d.  | C1CCCCC1 |

- e.  clc[nH]ccl
- f. Pb[Cl]4 Pb(Cl)(Cl)(Cl)(Cl)
- g. Deutero chloroform [2H]C(Cl)(Cl)(Cl) (the isotopic mass number precedes the atom. For tritium, use [3H]).

6. Bonds are not specified except as follows:

= for double bond

for triple bond

a period (.) separates disconnected structures, *e.g.*, ions, complexes or two or more reactants.

Examples:

CH3C CN CC#N

CH3COO- Na+ CC(=O)[O-].[Na+]

HC#CNO2 C#CN(=O)=O

 clccsclS(=O)(=O)N

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

-NH3+ [N+](H)(H)(H)

Reaction between 4-X—C₆H₄C(=O)Cl and 4-Y-C₆H₄NH₂=clcc(*)ccclC(=O)Cl.lccl(*)ccclN

The (*) indicates positions where various forms of X and Y can be attached. A period separates the two molecules.

7. A branched group is placed in parentheses

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

 clccc[nH+]cl

C6H5N#NBF4- clcccccl[N+]#N.[B-](F)(F)(F)(F)

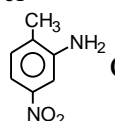
8. To make a ring pathway linear one bond must be 'broken' for each ring.

a. One numbers the atoms on each side of the break with the same number.

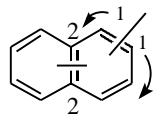
b. Any number can be reused after that ring is closed.

Examples:

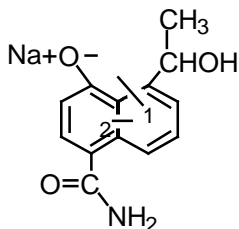
Benzene clcccccl

2-amino, 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; O=N(C)(C)C.



c1ccc2ccccc2c1 or clc2ccccc2ccc1



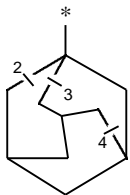
CC(O)clcccc2c(C(=O)N)ccc(c12)[O-]. [Na+]

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:

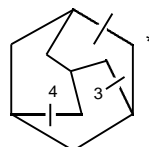
getsmi / parent **c1c(*)cccc1**

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**

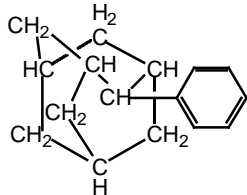


1-Adamantyl

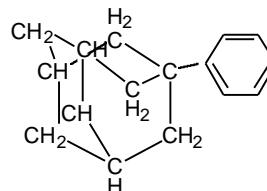


2-Adamantyl

As Shown by DEPICT:

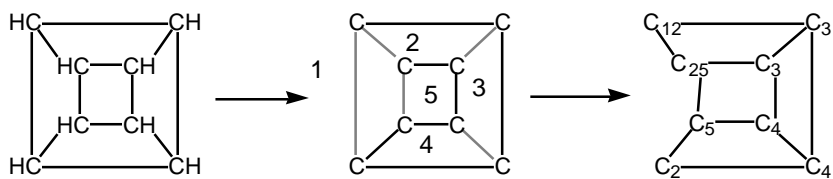


2-Adamantyl



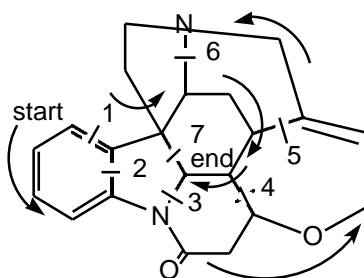
1-Adamantyl

An atom can be involved in more than one bond as shown above. A more complex example is that of cubane.



SMILES for cubane = C12C3C4C1C5C4C3C25

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 pyrrolide 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 clccc2N3C(=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, but if one can handle the strychnine example, there are few in all of chemistry that will be more formidable.

VII. Parameter Definition

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

| | | | | | |
|----|---------|-----------------------------|----|----------|------------------------------|
| 0 | CPI | Calculated Pi Value | 22 | S-O- | sigma ortho minus |
| 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 © |
| 20 | S-M- | sigma meta minus | 42 | S.M-C | sigma meta © |
| 21 | S-O | sigma ortho | 43 | CMR-SUB | calc. MR for Sub |

Any of the 43 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, 3, 4, 5, 6, 7, 11, 12, 13, 15, 16, 17, 18, 23, 32, 43) are discussed and their use illustrated in ref. 14.

A brief definition and references follows.

1. PI (). Hydrophobic parameter for substituents defined by partitioning of X-C₆H₅ between octanol and water. (P)¹⁴
$$P_x = \log P_{X-C_6H_5} - \log P_{C_6H_6}$$
2. MR-SUB. Molar refractivity of a substituent defined analogously to .
$$MR = \left(n^2 - \frac{1}{n^2 + 2} \right) \frac{MW}{d}$$

where n = refractive index, MW = molecular weight and d = density.
MR values are scaled by 0.1. MR is highly collinear with substituent volume.¹⁴
3. F. Swain-Lupton inductive/field effect parameter for aromatic systems.¹⁵
4. R. Corresponding Swain-Lupton resonance parameter.¹⁵
5. R+. Taft resonance parameter for substituent delocalization of a + charge.¹⁵
6. R-. Taft resonance parameter for substituent delocalization of a - charge.¹⁵
7. ES. Classic steric parameter for substituents defined by Taft from the hydrolysis of X-CH₂COOCH₃(C₂H₅).¹⁸
8. ES-HYBO. An Es type parameter obtained from the hydroboration of substituted ethylenes.
9. ES-V. Charton's steric parameter.¹⁹ Using the command **parameter /nolimit** and entering **7** and **9** it is found that there are 117 substituents for which both Es parameters are available. Entering **3** **reg 4** yields a correlation between Es and Es-V with r² = 0.942. This can be improved to r² = 0.964 by jackknifing to remove three badly correlated points: C(CH₂CH₃)₃, CHBr₂, CH(Me)CH₂CMe₃.
10. Austel's²⁰ 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: r² = 0.793.
11. L-STM. Verloop sterimol parameter for substituent length.^{14, 21}
12. B1-STM. Sterimol parameter for the width of the first atom of the substituent.^{14, 21}
13. B5-STM. An estimate of the overall width of the substituent.^{14, 21}
14. O-STER. There are 30 substituents having this parameter for the effect of adjacent substituents inhibiting the reaction of pyridines with CH₃I. There is little correlation between O-STER and Es or B1.²⁹
15. S-P (). Normal Hammett constant for para substituents. It is based on the ionization constants of benzoic acids.¹⁴
16. S-P+ (⁺). Brown parameter where substituents delocalize a + charge or radical via resonance.^{14, 34}

17. S-P- (ρ^-). Hammett constant where substituents delocalize a negative charge via resonance. It is derived from the ionization constants of phenols.¹⁴
18. S-M (ρ). Hammett constant for meta substituents (non conjugated substituents).¹⁴
19. S-M+(ρ^+). Brown parameter for meta substituents. There is little difference between ρ_m and ρ_m^+ .¹⁴
20. S-M- (ρ^-). Hammett constant for meta substituents (non conjugated substituents).¹⁴
21. S-O- (ρ). Hammett constant for ortho substituents. It correlates poorly with ρ_p , for 51 substituents $r^2 = 0.303$.³⁵
22. S-O (ρ). The parameter for ortho substituents.
23. S-INDUC (ρ). for the field/inductive effect. Originally defined from 4-X-bicyclo [2.2.2] octane-1-carboxylic acids.¹⁴
24. S-AN-RS . Resonance parameter (ρ^-) obtained from anilines. There are 25 substituents for which both ρ^- and S-AN. The correlation between the two is poor: $r^2 = 0.718$.
25. S-RES+. Resonance parameter for delocalization of + charge.¹⁵
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.¹⁴
27. S-PARNO. A set of normalized ρ_p values.³⁶
28. S-ORTH+. ρ^+ for ortho substituents.³⁶
29. S-PHOSP. ρ for substituents attached to phosphorus.³⁷
30. S-L. ρ_L for field/inductive effect.¹⁴
31. S-TWST. Effect on resonance by twisting substituent 90° out of plane.
32. S-STAR. Classic ρ^* defined by Taft.¹⁴
33. S-IND.P. Field/inductive parameter for substituents attached to phosphorus.³⁷
34. S-RES.P. Resonance parameter for substituents attached to phosphorus.³⁷
- 35,36. ER-P, ER-M. Radical parameters defined by Yamamoto and Otsu.³⁸
- 37,38. S.DOT-P, S. DOT-M. Radical parameters (ρ^\cdot) defined by Dust and Arnold.³⁹
- 39,40. S.-DOT-P, S.-DOT-M. Radical parameters (ρ^\cdot) defined by Jiang.⁴⁰
- 41,42. S.P-C, S.M-C. Radical parameters (ρ^\cdot) defined by Creary.⁴¹
43. Calculated MR for substituents.

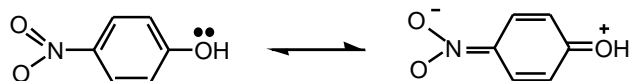
These parameters were collected over the past 35 years. Many are obsolete and have been kept

for historical reasons. The following are those that we find useful today.

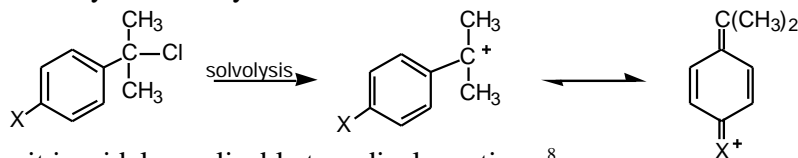
1. ρ . Calculated ρ value that takes into account electronic effects of the molecule on the substituent of interest.
2. MR-Sub. Calculated values for substituents.
3. F. Field/inductive effect of substituents on aromatic rings.
4. Es. Steric parameters for intramolecular interactions.
5. L-STM. Sterimol parameter for substituent length.
6. B1-STM. Sterimol parameter for size of first atom in substituent.
7. B4-STM. An attempt to account for substituent bulk.
8. S-P. Classic σ constant for aromatic substituents.
9. S-P+. Classic σ^+ when there is direct resonance between substituent and the reaction center involving delocalization of a + charge.
10. S-P-. Through-resonance involving delocalization of a - charge.
11. S-inductive. (σ_{ind}) for aliphatic reactions.
12. S-star. (σ^*) for aliphatic reactions. Sometimes σ^* is best, sometimes σ_{z} . They are rather collinear.
13. ER-P / ER-M. Used for aromatic radical reactions.
14. S.P-C / S.M-C. Used for aromatic radical reactions.

The electronic parameters for aromatic substituents (ρ , σ^- , σ^+) have been developed and extensively studied over the last 50 years. Simple ρ is obtained from the ionization constants of benzoic acids where little through resonance is present.

The σ^- parameter holds when direct through resonance is involved in the delocalization of a pair of electrons.



The parameter σ^+ (S+) is used when delocalization of a positive charge occurs. It was developed from solvolysis of cumyl chloride as follows:³⁴



Surprisingly, it is widely applicable to radical reactions.⁸

The field/inductive effect F is useful in some aromatic cases. It is especially valuable for substituents ortho to reaction centers. Such substituents should be tested by the combination F,

Es, B1, -p. That is, using $F + -p$ can account for the extra electronic effect.²²

For substituents on aliphatic molecules two sets of parameters have been developed σ^* (S') and σ_1 (SI) although they are highly collinear sometimes one works somewhat better than the other. There are many more values of σ^* .

In the case of radical reactions σ^+ (S+) is the most widely useful parameter, however, E_R or S.-Dot. sometimes are better than σ^+ .⁸ We still do not understand the reasons behind this.

There are two classes of hydrophobic parameters: $\log P$ and π (Pi). They are based on partitioning between octanol / water. The expense of measuring $\log P$ has inspired a large effort by many laboratories to develop methods of calculation. The following result comes from Leo's work at BioByte.

$$\log P = 0.96 \text{ Clog P} + 0.07$$

$$n = 12,443, \quad r^2 = 0.973, \quad s = 0.299$$

There are 228 outliers out of our present set carefully evaluated experimental values that are not included in the above equation.

We have a large number of experimental values from X-C₆H₅: $\sigma = \log P \text{ X-C}_6\text{H}_5 - \log P \text{ C}_6\text{H}_6$. These values have a serious limitation when there is a strong electronic interaction between the substituent and other parts of the molecule. For such cases we subtract calculated Clog P from the calculated value for the derivative (CPI). The steric parameters Es and the sterimols are interesting to compare. Searching the double data base of bio and phys of 17,800 QSAR, we find the following:

| | | | |
|-------|------------|-------|----------|
| 15 Es | 623 hits | 15 L | 63 hits |
| 15 B1 | 1,373 hits | 15 MR | 109 hits |
| 15 B5 | 917 hits | | |

We have not attempted to survey all of our earlier formulated QSAR to see where B1 can replace Es. Verloop's sterimol parameters are an important extension of traditional QSAR in the direction of 3-D QSAR. MR is a measure of the bulk of a substituent and it has not found extensive use. Now that we can calculate MR for a substituent, we may find more use for it.

The first problem in designing a new set of congeners from a parent molecule is to proceed carefully by studying one position at a time and choosing a set of substituents that shows good variation in electronic, hydrophobic and steric properties. Now one needs to know the values that are available for the set selected.

The problem can be approached from either the bio or phys systems. Entering **parameter /nolimit** in the regression mode displays the table of parameter values. Enter **2 13 16 17 18**. This isolates substituents all of which have values for Pi, B1, p , p^+ , p^- . Entering **sum** shows that there are 61 such substituents. Entering **seed**data displays the results ordered on increasing values of the first substituent (Pi) selected. The data set obtained can be worked with in the same fashion as a standard data set. Next, first entering **corr** followed by **3 4 5 6 7** displays the following correlation matrix:

Correlation matrix: R**2 and N

| | PI | B1 | S-P | S-P+ | S-P- |
|------|----|------|------|------|------|
| PI | . | .014 | .023 | .004 | .039 |
| B1 | 61 | . | .147 | .122 | .079 |
| S-P | 61 | 61 | . | .852 | .860 |
| S-P+ | 61 | 61 | 61 | . | .732 |
| S-P- | 61 | 61 | 61 | 61 | . |

The only significant collinearity is that among the electronic parameters. Although this might seem serious, we have found that with some care in the selection of substituents, the most significant parameter can be determined. We have established this fact via comparative QSAR.⁸

Another way to obtain a wider view is to select only S () for which we have many more values. This would be sufficient to establish whether or not an electronic effect is present. Often it is not. Entering **parameter /nolimit** and then **2 13 14 16** followed by **sum** finds 294 substituents ordered on . Enter **seed**. If S was entered first, they would have been ordered on .

If interest is only in a single substituent, the following results can be obtained:

2 Pi yields 1088 values. Enter **seed** to find values from -5.96 to 3.03

13 B1 yields 1080 values from 1 to 4.65

16 p yields 1997 values from -1.58 to 2.42

17 p^+ yields 398 values from -7.17 to 1.88

In each instance, one can enter **seed** to view the individual values.

VIII. 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 useful, it is much easier to save it in this area.

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 (set B633)

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** (get parameters). 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 by hand. Often data is not in logarithmic form. They can be entered as such and then converted to log or other form by using **gettran**.

C. Naming and Entering Substituents

Now enter **newsu**. 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 di-Cl. After all of the labels and parameter 3 values are entered, enter **end**. Entering **seed** yields the following table:

| No. | Substituent | log 1/C |
|-------|---------------------------------|---------|
| 1. | H | 2.73 |
| 2. | 2-NO ₂ | 3.28 |
| 3. | 3-NO ₂ | 3.44 |
| 4. | 4-NO ₂ | 3.49 |
| 5. | 2-NH ₂ | 4.85 |
| 6. | 3-NH ₂ | 3.92 |
| 7. | 4-NH ₂ | 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 ₂ H ₅ | 3.64 |
| 21. | 4-C ₃ H ₇ | 3.46 |
| Enter | End | |

It is advisable to save data entry frequently by entering **save**. To view Title information, enter **summary**. To view parameter data entry, enter **seeddata**. If entry errors need to be corrected, one can enter **editsub** for editing substituents or **editdat** for editing data. 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.

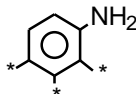
D. Entering Structures via SMILES

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(*)ccl** or **cl(*)c(*)c(*)ccclN**.

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 | 2-nitroaniline |
| *H *N(=O)=O *H | 3-nitroaniline |
| *H *H *N(=O)=O | 4-nitroaniline |
| *N *H *H | 2-aminoaniline (Note that) H on NH ₂ or OH on CH ₃ is not entered, these are supplied by the system |
| *C *H *H | 2-methylaniline |
| *H *OC *H | 3-methoxyaniline |

2 TERM REGRESSIONS

| | S.D. | MLOGP | S | S+ | S- | F-1 | ES-1 | B1-1 | B5-1 | CONST |
|----|------|-------|--------|--------|-------|------|-------|------|------|-------|
| 1 | .298 | | 1.716 | -1.833 | | | | | | 3.514 |
| 2 | .316 | | | -1.303 | .743 | | | | | 3.583 |
| 3 | .331 | -.180 | | -.609 | | | | | | 3.955 |
| 4 | .349 | | | -.744 | | .305 | | | | 3.703 |
| 5 | .352 | | | -.733 | | | | .143 | | 3.567 |
| 6 | .352 | | | -.742 | | | -.056 | | | 3.712 |
| 7 | .353 | | | -.726 | | | | | .037 | 3.683 |
| 8 | .394 | -.241 | -.708 | | | | | | | 4.143 |
| 9 | .411 | -.284 | | | -.484 | | | | | 4.241 |
| 10 | .412 | | -2.117 | | .949 | | | | | 3.832 |

3 TERM REGRESSIONS

| | S.D. | MLOGP | S | S+ | S- | F-1 | ES-1 | B1-1 | B5-1 | CONST |
|----|------|-------|-------|--------|-------|-------|------|-------|-------|-------|
| 1 | .271 | -.20 | 1.803 | -1.753 | | | | | | 3.751 |
| 2 | .30 | -.175 | 1.916 | -1.174 | .731 | | | | | 3.803 |
| 3 | .301 | | 1.873 | -1.978 | | | | | -.092 | 3.609 |
| 4 | .304 | | 1.795 | -1.927 | | -.203 | | | | 3.513 |
| 5 | .305 | | 1.772 | -1.877 | | | .042 | | | 3.518 |
| 6 | .306 | | 1.853 | -1.869 | | | | -.076 | | 3.594 |
| 7 | .306 | | | -1.857 | -.084 | | | | | 3.513 |
| 8 | .321 | | | -1.397 | .847 | | | | -.084 | 3.672 |
| 9 | .322 | | | -1.376 | .855 | | .077 | | | 3.587 |
| 10 | .324 | | | -1.357 | .823 | -.175 | | | | 3.583 |

If we run **5 reg 6** this shows that S and S^+ are highly collinear ($r^2 = 0.935$). There seems to be no special role for ortho substituents. Mlog P has a weak role. We can check for the overall collinearity as follows

corr 4 5 6 7 8 9 10 11

Correlation matrix: R**2 and N

| | MLOGP | S | S+ | S- | F-1 | ES-1 | B1-1 | B5-1 |
|-------|-------|------|------|------|------|-------|------|-------|
| MLOGP | . | .289 | .292 | .236 | .047 | -.218 | .200 | .078 |
| S | 21 | . | .935 | .943 | .241 | .059 | .09 | -.075 |
| S+ | 21 | 21 | . | .915 | .017 | .028 | .019 | -.077 |
| S- | 21 | 21 | 21 | . | .099 | .114 | .144 | .000 |
| F-1 | 21 | 21 | 21 | 21 | . | .713 | .845 | .756 |
| ES-1 | 21 | 21 | 21 | 21 | 21 | . | .782 | -.751 |
| B1-1 | 21 | 21 | 21 | 21 | 21 | 21 | . | .807 |
| B5-1 | 21 | 21 | 21 | 21 | 21 | 21 | 21 | . |

G. Checking for Parameter Collinearity

We can check for overall collinearity problems as follows: **corr 4 6 5 10 11 15 13 14**

| | MLOGP | S | S+ | S- | F-2 | ES-1 | B1-1 | B5-1 |
|-------|-------|------|------|------|------|-------|------|-------|
| MLOGP | | .289 | .292 | .240 | .047 | -.218 | .200 | .078 |
| S | 21 | | .935 | .952 | .241 | .059 | .096 | -.075 |
| S+ | 21 | 21 | | .921 | .017 | .028 | .019 | -.177 |
| S- | 21 | 21 | 21 | | .101 | .114 | .151 | -.001 |
| F-2 | 21 | 21 | 21 | 21 | | .713 | .845 | .756 |
| ES-1 | 21 | 21 | 21 | 21 | 21 | | .782 | -.751 |
| B1-1 | 21 | 21 | 21 | 21 | 21 | 21 | | .807 |
| B5-1 | 21 | 21 | 21 | 21 | 21 | 21 | 21 | |

There is high collinearity among σ , σ^+ and σ^- and considerable collinearity with the steric parameters. The upper part of the matrix gives the correlation among the 21 different substituents.

Next exploring the two best terms: we obtain the following QSAR:

$$\log 1/C = -0.18(\pm 0.24) \text{Mlog P} - 0.61(\pm 0.30) \sigma^+ + 3.96(\pm 0.33) \sigma^-$$

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

This is obviously not a good equation.

H. Jackknifing

In picking the current data set for study we have not selected a good, easy to correlate example. It is better to study how to do the best with a non-ideal set of data. At this point, we can look for outliers. This could be due to poor experimental data. A major problem is a lack of uniform reaction mechanisms of the various 'congeners'.⁴² Poor selection of substituents so that bad collinearity is present from the start is so often the problem. It is rarely considered in the design of the project. We can now resort to jackknifing to obtain some perspective. **3 j 4 11** derives all possible regression equations by dropping a different data point in each instance.

| Omitted | r^2 | s |
|---------|-------|-------|
| none | 0.693 | 0.331 |
| H | 0.849 | 0.211 |
| 4-Cl | 0.713 | 0.329 |
| 3-OH | 0.702 | 0.334 |

Dropping parent compound aniline (H) yields r^2 of 0.849. To delete this compound use the command **star /add 1**. This places an asterisk on data point 1 and it is not then used in deriving

future equations. Any number of points can be withheld in this fashion. **star /a 1 5,10 18** would place asterisks on compound 1, 5 to 10, and 18. Asterisks can be removed by entering the command **star /d 5** to restore data point 5 for study.

Starring 1 yields the following QSAR:

$$\log 1/C = -0.22(\pm.15) \text{ Mlog P} - 0.55 (\pm.19) + + 4.07 (\pm.21)$$

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

Jackknifing a second point points to 3-NH₂ as being poorly fit. **star /a 6** withholds this point.

The resulting QSAR is:

$$\log 1/C = -0.22 (\pm.16) \text{ Mlog P} - 0.55 (\pm.20) + + 4.07 (\pm.22)$$

$$n = 19, \quad r^2 = 0.876, \quad s = 0.197, \quad q^2 = 0.833$$

Checking the reliability of Mlog P we can **add ClogP**. Substituting this in the above QSAR yields an equations with $r^2 = 0.853$. Removing a third point (3-NH₂) offers improvement.

$$\log 1/C = -0.30 (\pm.17) \text{ Mlog P} - 0.50 (\pm.20) + + 4.20 (\pm.25)$$

$$n = 18, \quad r^2 = 0.879, \quad s = 0.201, \quad q^2 = 0.833$$

Note that q^2 is now approaching r^2 in value. This implies that dropping another point would have little effect on the correlation. A problem with the jackknifing procedure is that it forces the data to fit the model equation one starts with. Now repeating the perm procedure with Mlog P, S+, F, Es, B1 and B5, we cannot find any value in parameters other than Mlog P and +.

I. Plotting Data

Sometimes insight can be gained by plotting the data. Any two parameters can be plotted against each other by the command **a graph b**, *i.e.*, **3 graph 6** gives some idea of the fit to the most important variables. In the present instance this is of little help. Of more help is the command **predict Absolute values Descending order**. Viewing the deviations does not reveal any pattern.

A point of interest is the use of amino groups as substituents when the study is concerned with anilines. It is surprising that two of the three amino groups are well fit. This is not what we normally find.

J. Cross Validation

Cross validation is a means for avoiding poor quality or meaningless QSAR. We have included two ways of viewing the problem. The program automatically calculates a cross

validated r^2 (q^2) and prints this out with each equation. One wants to see the normal r^2 and q^2 in 'reasonable' agreement, but exactly what amounts to reasonable is a matter of judgment. If they are not close (say $r^2 = 0.90$, $q^2 = 0.7$) then jackknifing to remove one or more points may be worthwhile.

q^2 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 //full 4 5** randomly drops 10% of the data points in 20 trials using 17 data points and displays the QSAR. In these examples r^2 ranges from 0.849 to 0.900.

One can also set an arbitrary percentage drop: **3 cross//full /omit=30% 4 5** which yields 4 QSAR with r^2 from 0.792 to 0.891. 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 Mlog P and π vary but little from our first QSAR with only two points omitted.

K. 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.

L. Regression Analysis: Example 2

We now consider a data set (Bio #1557) based on a study of benzanilides (X-C₆H₄CONHC₆H₄-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** , 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 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 more than 50% of the Bio QSAR depend on log P or on σ 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:

$$\log 1/C = 1.02(\pm 0.36) \text{ ClogP} - 1.75(\pm 0.78) \text{ bilin (ClogP)} + 1.30(\pm 1.29)$$

$$n = 34, \quad r^2 = 0.520, \quad s = 0.699 \quad q^2 = 0.454$$

Activity initially increases with increase in log P with slope (h) of 1.02 and then falls off linearly

with slope of -1.75. The optimum ($\log P_o$) listed above is actually rapidly calculated from the parabolic equation. To get the exact value enter **3 j B4**. The optimum shown here is 5.27.

The above bilinear term actually represents: $-1.75 (\pm 0.78) \log (\cdot 10^{\text{ClogP}} + 1)$ where $\cdot = -5.43$. Generally researchers use the form $\log (\cdot P + 1)$. Since many of our parameters are on a log basis (, , E_s , $M\log P$) 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 $\text{Dev} +$ and $\text{Dev} -$ which represent the number of points below and above the regression line are not badly out of balance. 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. It is not surprising to see the QSAR does not contain a $-$ or $+$ 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 $E_s - X_2$ 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:

$$\log 1/C = 0.73 (\pm 0.47) \text{ClogP} - 1.24 (\pm 0.32) E_s - X_2 + 0.97 (\pm 1.4)$$

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

At this point recall that all substituent values of E_s are negative, that is, the E_s term implies that bulky substituents are good! Note also that all of the $\log P$ values for this subset are much below $\log P_o$ of 5, so that linear dependence on ClogP is expected. Deriving single parameter equations in ClogP and $E_s - X_2$ 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^2 values between the various substituents. The correlation (r^2) between $E_s - X_2$ and the electronic parameters is not high. Entering **3 reg 4 14** shows that, indeed, electronic effects do not appear to be significant (note confidence limits on coefficients). 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_o 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 ± 1.2 .⁷ 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, but the QSAR is very bad. Note r^2 and entering **3 reg B4 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-NH₂ ($\beta = -0.66$) or 4-OH ($\beta = -0.37$).

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-NH₂ 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 are correcting this deficiency (above #3500 in physical databank and above #3000 in the Bio databank). Once a set has been loaded, one can explore it as we have done with the above example.

M. 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.^(see ref. 14, Ch. 13) 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 (σ_p , σ_p^+ , σ_p^- , 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**, or **parameter /nolimit**. Entering **parameter** produces a table of labels (same as the **fetch** command). Any group can be selected. For example, entering **2 16** (σ_p , MR) the prompt asks for the minimum and maximum to set a range for each parameter. We might enter **-1 1 -1 1** to define the range for each of the two parameters. After the data have been collected, enter **seed** and we find that only 206 substituents fall in this range.

In selecting a set of substituents for combinatorial synthesis it can be very important to limit the range of σ_p or possibly other parameters. Thus for σ_p , one might set limits such as -1 to 3 for σ_p 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**, 876 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 **/ 2 3 13 16 19** followed by **seed** produces a set of 256 substituents ordered on increasing values of σ_p (first parameter selected) all of which have σ_p , MR, B1, σ_p and σ_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 σ_p 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^2 .

| | P1 | MR | B1 | S-P | S-M |
|-----|-----|------|------|------|------|
| P1 | . | .127 | .026 | 0.50 | .275 |
| MR | 256 | . | .085 | .052 | .011 |
| L | 256 | 256 | . | .118 | .062 |
| S-P | 256 | 256 | 256 | . | .843 |
| S-M | 256 | 256 | 256 | 256 | . |

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 **2 12 13 14 16** for ρ , B1, B5, L, ρ . 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 (294) 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 ρ and ρ 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 /d** followed by the number of the selected substituents of interest. 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 **16** and **17** 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 ρ and ρ^+ (199), where $r^2 = 0.878$. Next, use **gett** (get transformation) 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 ρ , MR, L, B1, B5, ρ , ρ^+ , ρ^- . **seed** lists only 61 substituents having all of these parameters. The reason for this is that there are relatively few values for ρ^+ and ρ^- . However,

these electronic parameters are very valuable in providing mechanistic insight via QSAR^{7, 8, 14} and should be studied separately. There is a high degree of collinearity among them. In a way this is helpful as it means that ρ can detect the presence of electronic effects that later can be explored with the other two parameters.

VIII. UDRIVE and Masterfile

UDRIVE allows interactive, unified access to the software modules which calculate log P (oct), and molar refractivity. In addition, it accesses the measured properties in Masterfile, such as log P values in octanol and other solvent systems, pKa, activity types, and also the calculated McGowan molar volumes, molformula and molecular weight.

Entry of **udrive** will return a panel which has the cursor on the default entry by SMILES, 'S'. Entry of the find command, **F(ind)**, returns a list of other access routes, such as name, CAS number, etc. Entry of **2** gives the prompt that a name is expected. For example, entering **AZT** at the name prompt results in a depiction (DEPICT) of the structure and calculated and measured log P values. Other data on the THOR page of Masterfile can be accessed with the entry of **d**. Note that other synonyms, such as zidovudine or retrovir, could have been entered to access this calculation and the stored data. Entry of **c** will retrieve the details of the log P calculation, such as the contribution of each fragment and their interactions. Masterfile also lists the Activity Type as 'Reverse Transcriptase Inhibitor' and 'Antiviral'.

IX System Crashes

Occasionally, the system crashes. First enter **unlock** and then **QSAR** followed by **data Bio** or **Data Phys**. Sometimes, if errors are made in the search and show process, the system does not display results. Enter **q** until one reaches the \$, then reload by entering **QSAR** followed by **data bio** or **data phys**.

X. Caveats

It must not be forgotten that the present databanks were created over a period of about 40 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. At present, we are re-studying all this early work to see what improvements can be made.

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 are remedying 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 4,600 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 ($\log I + R$, $\log k + R$, $\log F + R$, 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 $\log P$ of certain substituents. In rare examples it is possible to find clear roles for both $\log P$ and $\log P_0$ 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.⁷

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 $\log P_0$.

A shortcoming 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 $\log P_0$ 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 or two values exist this does help in getting a better estimate of $\log P_0$. $\log P_0$ 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.²⁴ Ideally the substituent parameter, $\log P$, can be used to check for these possibilities. However, our current automatic loading uses only $\log P$ values

measured for X-C₆H₅ or values calculated from simple benzene derivatives. If there is a strong electron withdrawing group nearby, it will affect the 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 Ps: benzene = 2.13; aniline = 0.90; nitrobenzene = 1.85; and 4-nitroaniline = 1.39. With benzene as the parent solute, values for the amino and nitro substituent are: $\sigma_{\text{NH}_2} = 0.90 - 2.13 = -1.23$; and $\sigma_{\text{NO}_2} = 1.85 - 2.13 = -0.28$. When these two substituents are on the same phenyl ring, electronic interaction raises their apparent values as seen in these calculations:

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

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

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

This increase in 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 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 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. We have now developed a system for the automatic calculation of values (C_{Pi}) that avoids the problem of electronic interactions.

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 q^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.^{25, 42} **Although outliers are intimidating, they can be very helpful. If a compound is more active than expected this can be a lead worth following up. If less active than predicted, it provides information on a direction to be avoided.**

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 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 100 years.

XI References

1. Selassie, C.D.; Shusterman, A.J.; Kapur, S.; Verma, R.P.; Zhang, L.; Hansch, C. *J. Chem. Soc. Perkin Trans 2* 1999, 2729.
2. Hansch, C.; Garg, R.; Kurup, A. *Bioorg. Med. Chem.* 2001, 9, 283.
- 2a. Garg, R.; Kurup, A.; Mekapati, S.B.; Hansch, C. *Bioorg. Med. Chem.* 2002, in press.
3. Kurup, A.; Garg, R.; Carini, D.J.; Hansch, C. *Chem. Rev.* 2001, 101, 2727.
4. Kurup, A.; Garg, R.; Hansch, C. *Chem. Rev.* 2001, 101, 2573.
5. Fujita, T. in *Drug Design: Fact or Fantasy?* Jolles, G.; Wooldridge, K.R.H., Eds. Academic Press, 1984, p. 17.
6. Hansch, C.; Fujita, T. in *Classical and Three-dimensional QSAR in Agrochemistry*, Hansch and Fujita, Eds., American Chemical Society, Washington, D.C. 1995.
7. Hansch, C.; Hoekman, D.; Gao, H. *Chem. Rev.* 1996, 96, 1045.
8. Hansch, C.; Gao, H. *Chem. Rev.* 1997, 97, 2995.
9. Hansch, C.; Kurup, A.; Garg, R.; Gao, H. *Chem. Rev.* 2001, 101, 619.
- 9a. Selassie, C.D.; Garg, R.; Kapur, S.; Kurup, A.; Verma, R.P.; Mekapati, S.B.; Hansch, C. *Chem. Rev.* 2002, 102, .
10. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. *Nature* 1962, 194, 178.
11. Pauling, L.; Pressman, D. *J. Am. Chem. Soc.* 1945, 67, 1003.
12. Agin, D.; Hersch, L.; Holtzman, D. *Proc. Natl Acad. Sci. U.S.A.* 1965, 53, 952.
13. Hansch, C.; Björkroth, J.P.; Leo, A. *J. Pharm. Sci.* 1987, 76, 663.
- 13a. Carr, R.; Hann, M. *Modern Drug Discovery* April 2002, 45.
14. Hansch, D.; Leo, A. *Exploring QSAR* American Chemical Society, 1995.
15. Hansch, C.; Leo, A.; Taft, R.W. *Chem. Rev.* 1991, 91, 165.
16. Weininger, D.; Weininger, J.L. in *Comprehensive Medicinal Chemistry*, Ramsden, C. Ed., Pergamon Press, 1990, Vol. 4, pp. 59.
17. Weininger, O. *J. Chem. Inf. Computer Sci.* 1988, 28, 31.
18. Unger, S.H.; Hansch, C. *Prog. Phys. Org. Chem.* 1976, 12, 91.
19. Charton, M. *Prog. Phys. Org. Chem.* 1971, 8, 235.
20. Austel, V.; Kutter, E.; Kalbfleisch, W. *Arzneim. Fösch.* 1979, 29, 585.
21. Verloop, A.; Hoogenstraaten W.; Tipker, J. in *Drug Design Vol. VII*, Ariens, E.J., Ed. Academic Press, 1976, pp. 165.

22. Fujita, T.; Nishioka, I. *Prog. Phys. Org. Chem.* 1976, 12, 49.
23. Cramer, R.D.; Bunce, J.D.; Patterson, D.E.; Frank, I.E. *QSAR* 1988, 7, 81.
24. Smith, R.N.; Hansch, C.; Kim, K.H.; Omiya, B.; Fukumura, G.; Selassie, C.D.; Jow, P.Y.C.; Balney, J.M.; Langridge, R. *Arch. Biochem. Biophys.* 1982, 215, 319.
25. Garg, R.; Kurup, A.; Hansch, C. *Crit. Rev. Toxicol.* 2001, 31, 223.
26. Charton, M. *Prog. Phys. Chem.* 1971, 8: 235.
27. Berg, U., Gallo, R., Klatter, G. and Metzger, J. *J. Chem. Soc. Perk. 2* 1980, 1350.
28. Tribble, M. and Traynham, J. *J. Am. Chem. Soc.* 1969, 91: 379.
29. Taft, R.W.; Grob, C.A. *J. Am. Chem. Soc.* 1974, 96, 1236.
30. Yamamoto, Y. and Otsu, T. *Chem. Ind.* 1967: 787.
31. Garg, R.; Kurup, A.; Hansch, C. *Bioorg. Med. Chem.* 2001, 9, 3161.
32. Hansch, C; Hoekman, D; Leo, A; Weininger, D; Selassie, C.D. *Chem. Rev.* 2002, 102, 783.
33. Creary, X., Mehrsheikh-Mohammadi, M.E. and McDonald, S. *J. Org. Chem.* 1987, 52: 3254.
34. Okamoto, Y.; Brown, H.C. *J. Org. Chem.* 1957, 22, 485.
35. Charton, M. *Prog. Phys. Org. Chem.* 1981, 13: 119.
36. LeGuen, M.M.J. and Taylor, R. *J. Chem. Soc. Perk. 2* 1976: 557.
37. Mastryukova, T.A. and Kabachnik, M.I. *J. Org. Chem.* 1971, 36: 1201.
38. Yamamoto, Y. and Otsu, T. *Chem. Ind.* 1967: 787.
39. Dust, J.M. and Arnold, D.R. *J. Am. Chem. Soc.* 1983, 105: 1221.
40. Jiang, X.-K. and Ji, G.Z. *J. Org. Chem.* 1992, 57: 6051.
41. Creary, X., Mehrsheikh-Mohammadi, M.E. and McDonald, S. *J. Org. Chem.* 1987, 52: 3254.
42. Mekapati, S.B.; Hansch, C. *J. Chem. Inf Comput. Sci.* 2002, 42, 956.