History of QSAR

- Quantitative structure-activity relationship
- 1865 Crum-Brown and Fraser postulated a relation between "physiological activity Φ" of alkaloides and their chemical structure C ($\Phi = f(C)$).
- 1893 Richet correlated the toxicity of organic compounds with their aqueous solubility (toxicity $=$ solubility^{-1}).
- $\sim$ 1900 Meyer and Overton discovered linear relationships between the narcotic potencies of organic compounds and their partitioning behaviour.
- 1935 Hammett defined a linear free-energy relationship between reaction rates k and equilibrium constants K for reactions involving meta- and para-substituted benzoic acid derivatives.

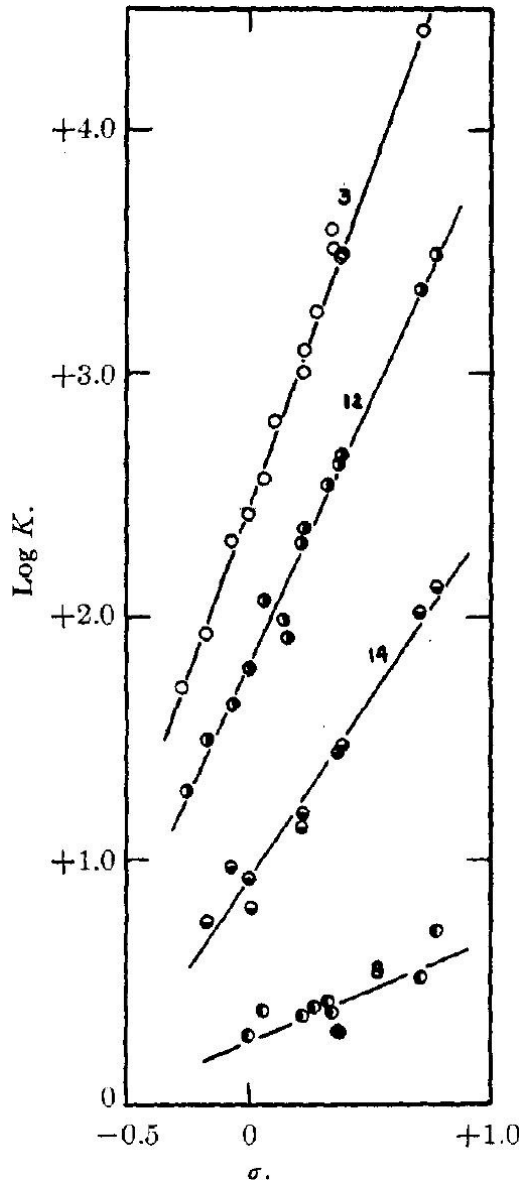
The Hammett Equation

$$\log \frac{K}{K_0} = \sigma\rho = \log \frac{k}{k_0}$$

σ is the **substituent constant**

ρ is the **reaction constant**, describing the susceptibility of the reaction to substituents, compared to the ionization of benzoic acid.

$\rho > 1$	more sensitive, neg. charge
$0 < \rho < 1$	less sensitive, neg. charge
$\rho = 0$	no sensitivity, no charge
$\rho < 0$	pos. charge



Hammett LP, (1937) The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives; J Am Chem Soc 59(1):96-103; DOI: 10.1021/ja01280a022

Equilibrium Constant

$$\alpha A + \beta B... \rightleftharpoons \rho R + \sigma S...$$

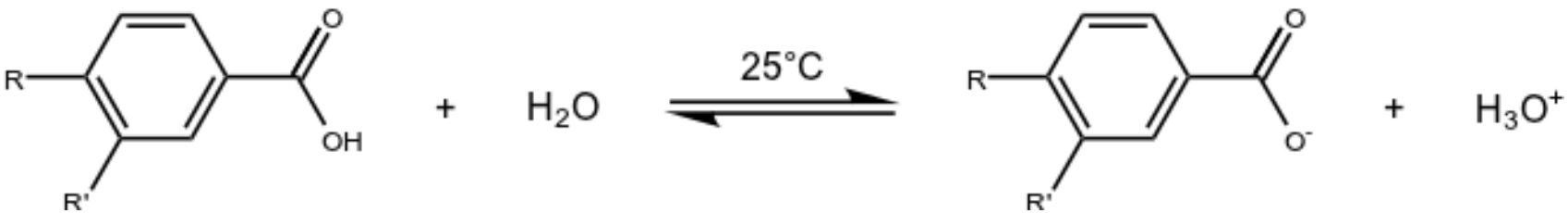
Thermodynamic equilibrium constant can be defined such that, at equilibrium

$$K = \frac{\{R\}^{\rho}\{S\}^{\sigma}...}{\{A\}^{\alpha}\{B\}^{\beta}...}$$

where $\{A\}$ is the activity of the chemical species A, usually concentrations $[A]$, rather than activities, defining a concentration quotient, K_c.

Substituent Constant σ

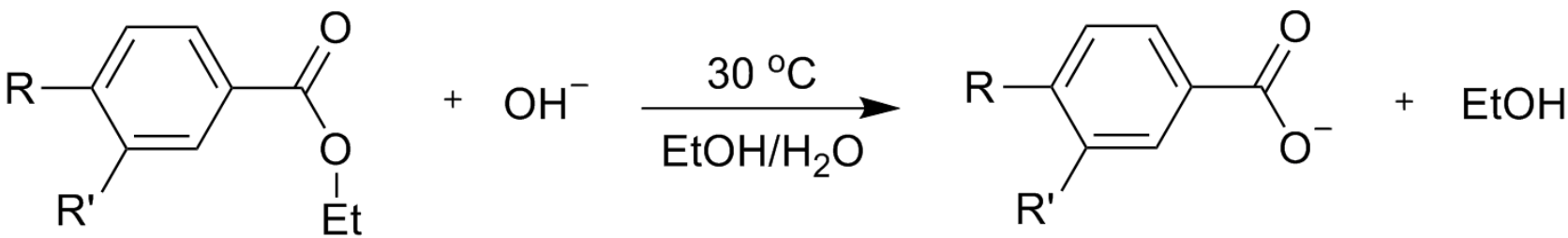
Reference: Set $\rho = \sigma = 1$ for the following reaction (ionization of benzoic acid) with R and R' being hydrogens.
Let K_0 be the equilibrium constant of this reference reaction.
Let k_0 be the reaction rate of this reference reaction.



Repeat with replaced R or R' in order to get K, and therefore the substituent constant σ.

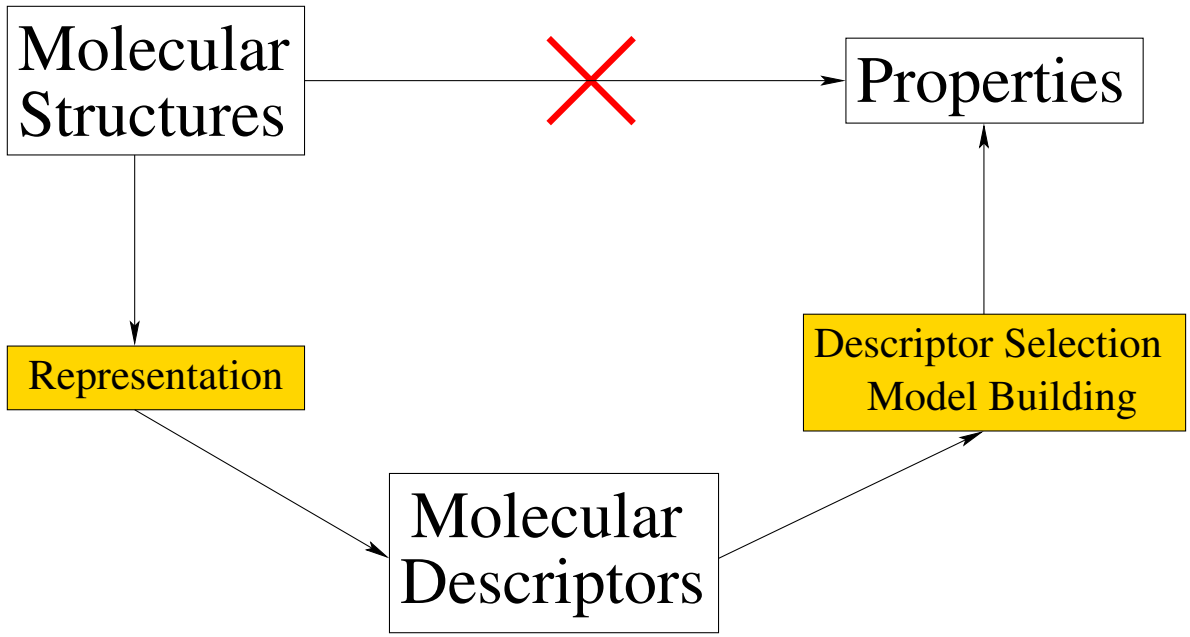
Reaction Constant ρ

With knowledge of the substituent constant σ, it is possible to determine the reaction constants for a range of reactions:
Example: Alkaline hydrolisis of ethyl benozoate in water/ethanol mixture:



(k_0 of the reference reaction combined with the knowledge of many σ-values leads here to the reaction constant $\rho = 2.498$)

The general QSAR Problem



1. Molecular structure determines properties.
2. Direct property prediction usually not possible.
3. Use of indirect method to tackle problem.
    - Encode molecular information in numerical descriptor.
    - Build a mathematical model (descriptor selection).
    - Validate the mathematical model.
    - Predict properties using the mathematical model.

QSAR: Basic Assumption

1. Similar structures exhibit similar biological activity.
2. Change in activity always comes from a change in structure.
3. The biological activity is caused by the same molecular mechanism.
4. Individual substituents or groups contribute additively to the observed biological activity.
5. ...

What is a Molecular Descriptor?

Derivation of knowledge from **real-world data** makes it necessary to **define properties** that differentiate certain objects from others.

Property	Elephant	Mouse
genus	mammal	mammal
extremity	4	4
eyes	2	2
food	herbivore	herbivore
similarity = 100%		
weight	6000kg	0.1kg
body length	7m	0.15m
ratio caudal/body length	0.2	1
similarity = 0%		

The way an object is described depends on the context of interest.

A molecular descriptor is an abstract, in most cases numerical, property of molecular structure, derived by some algorithm, describing a specific aspect of a chemical compound.

Fragment Descriptors (FD)

A FD is a selected subgraph of a molecular graph.

1. Any molecular graph invariant can be uniquely represented as linear combination of FDs. [Baskin, 1995]
2. Any symmetric similarity measure can be uniquely expressed in terms of FDs.
3. Any regression (or classification structure-property model) can be represented as a liear equation involving FDs.

FDs are easy to

- calculate
- store
- interprete

Baskin II et al., (1995) On the Basis of Invariants of Labeled Molecular Graphs;
J Chem Inf Comput Sci 35(3):527-531; DOI: 10.1021/ci00025a021;

Skvortsova MI et al., (1998) Molecular Similarity. 1. Analytical Description of the Set of Graph Similarity Measures;
J Chem Inf Comput Sci, 38(5):785-790; DOI: 10.1021/ci970037b;

J. Chem. Inf. Comput. Sci. **1995,** *35,* 527−531

On the Basis of Invariants of Labeled Molecular Graphs

Igor I. Baskin and Mariya I. Skvortsova

Institute of Organic Chemistry, Leninsky pr. 47, Moscow 117913, Russia

Ivan V. Stankevich*

Institute of Organoelement Compounds, Vavilov str. 28, Moscow 117813, Russia

Nikolai S. Zefirov

Department of Chemistry, Moscow State University, Moscow 119899, Russia

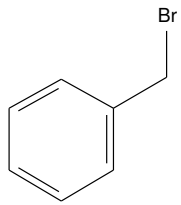
Received October 28, 1994[⊗]

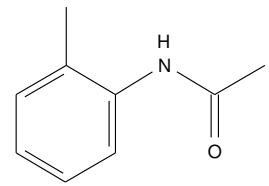
It is proved that any molecular graph invariant (that is any topological index) can be uniquely represented as (1) a linear combination of occurrence numbers of some substructures (fragments), both connected and disconnected, or (2) a polynomial on occurrence numbers of connected substructures of corresponding molecular graph. Besides, any (0,1)-valued molecular graph invariant can be uniquely represented as a linear combination (in the terms of logic operations) of some basic (0,1)-valued invariants indicating the presence of some substructures in the chemical structure. Thus, the occurrence numbers of substructures in a structure (or numbers indicating the presence or absence of substructures in a structure for the case of (0,1)-valued invariants) are shown to constitute the basis of invariants of labeled molecular graphs. A possibility to use these results for the mathematical justification of substructures-based methods in the "structure−property" problem is also discussed.

Lipophilicity as an FD example

Lipophilicity is a measure for the partitioning of a compound between a lipidic and an aqueous phase, usually presented by the **partition coefficient** P.

$$P = \frac{[C]_{\text{lip}}}{[C]_{\text{aq}}} \qquad \log P = \sum_{i=1}^{n} a_i \cdot f_i + \sum_{j=1}^{m} b_j \cdot F_j$$





Bromide fragment	0.480	NH-amide fragment	-1.510
1 aliphatic isolating C	0.195	2 aliphatic isolating C	0.390
6 aromatic isolating C	0.780	6 aromatic isolating C	0.780
7 H on isolating C	1.589	10 H on isolating C	2.270
1 chain bond	-0.120	1 chain bond	-0.120
		1 benzyl bond	-0.150
		ortho substituted	-0.760
Total	2.924	Total	0.900

$\log P < 0$ means hydrophilic and $\log P > 0$ lipophilic.

Partition coefficient and log P (simplified)

The **partition coefficient** is a ratio of concentrations of a compound between the two solutions. The logarithm of the ratio of the concentrations of the solute in the solvents is called log P: The log P value is also known as a measure of lipophilicity.

$$\log \ P_{\mathrm{oct/wat}} = \log \left(\frac{[\mathrm{solute}]_{\mathrm{octanol}}}{[\mathrm{solute}]_{\mathrm{water}}} \right)$$

1D Descriptores

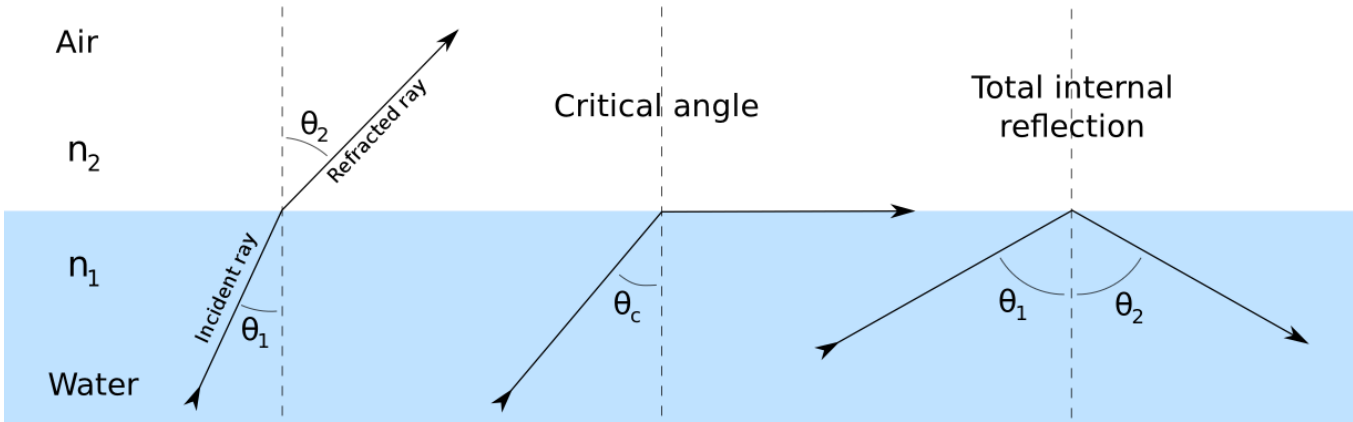
The Molar Refractivity index as a measure for polarizability.

$$MR = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{MW}{\rho}$$

n is the refractivity index. Since molecular weight (MW) devided by density (ρ) this descriptor equals molecular volume.

MR be easily calculated by atomic FD contribution methods.

FD contribuiton methods are often called **increment systems**.



Topological Incices: The Wiener Index

The Wiener index is defined as the sum of all topological distances ($N \times (N-1)/2$ many) in the H-depleted molecular graph.

$$W = \sum_{i=1}^{N} \sum_{j>i}^{N} d_{ij} \qquad \text{with } N := \text{number of atoms}$$

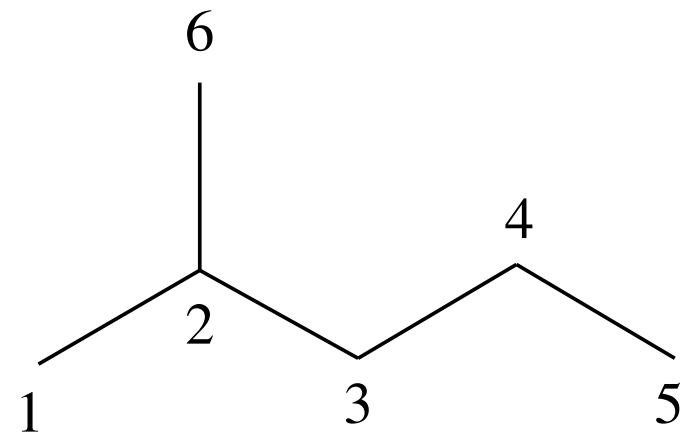
The Wiener index increases with the number of atoms, and for a constant number of atoms, reaches a maximum for linear structures and a minimum for the most branched and cyclic structures.

W is a convenient measure for the **compactness** of a molecule and has low degenerecy (many molecules have the same Wiener index).

Wiener H, (1947) Structural Determination of Paraffin Boiling Points; J Am Chem Soc, 69(1):17-20; DOI: 10.1021/ja01193a005

Topological Indices: Wiener Index (cont'd)

0	1	2	3	4	2
1	0	1	2	3	1
2	1	0	1	2	2
3	2	1	0	1	3
4	3	2	1	0	4
2	1	2	3	4	0



Distance matrix

$$
\begin{aligned}
W \quad = \quad & 1+2+3+4+2 \\
& +1+2+3+1 \\
& +1+2+2 \\
& +1+3 \\
& +4 \\
= \; & 32
\end{aligned}
$$

W is is the oldest topological index related to molecular branching. It correlates with the van der Waals surface area of the molecule.

Complexity Descriptors

Can be devided into **topological** and **chemical complexity** indices.
The Minoli Index is a measure of topological molecular complexity

$$MI = \left(\frac{|V| \times |E|}{|V| + |E|}\right) \cdot \sum_{l} P_l$$

Information-theoretic indices are derived from the Shannon formula

$$I = -\sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \log_2 \left(\frac{n_i}{n}\right)$$

An example is the Bonchev-Trinajstic Index

$$BT = n \log_2 n - \sum_{l} n_l \log_2 n_l \qquad \text{with } n = \sum_{l} n_l$$

n_l is then number of distances of length l.

Complexity Descriptors (cont'd)

The number of spanning trees (undirected graphs) $t(G)$ is a topological complexity measure

$$t(G) = |L_i| \qquad \text{with} \quad L(G) = V(G) - A(G)$$

$V(G)$ is the diagonale matrix of vertex degrees, $A(G)$ is the adjacency matrix and $L(G)$ the Laplacian matrix (L_i is the Laplacian Matrix with row i and column i deleted).

For directed graphs: The value is the number of "directed spanning trees" rooted at node i.

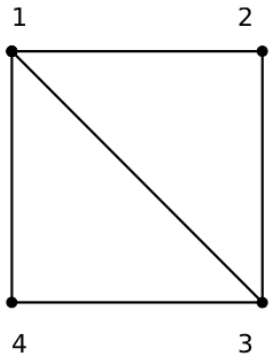
Complexity Descriptors (cont'd)

The number of spanning trees (undirected graphs) $t(G)$ is a topological complexity measure,

$$t(G) = \frac{1}{n} \prod_{i=1}^{n-1} \lambda_i$$

λ_i are the non-zero eigenvalues of the Laplace matrix (see Kirchhoff's Matrix Tree Theorem)

Example

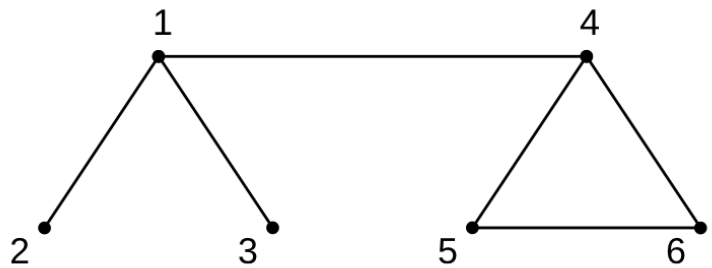


$$L(G) = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

$$\begin{aligned} \lambda_1 &= 4 \\ \lambda_2 &= 4 \\ \lambda_3 &= 2 \\ \lambda_4 &= 0 \end{aligned}$$

$$t(G) = \frac{1}{4}(4)(4)(2) = 8$$

Example



$$L(G) = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}$$

$\lambda_1 \approx 4.5646$
$\lambda_2 = 3$
$\lambda_3 = 1$
$\lambda_4 = 1$
$\lambda_5 \approx 0.4384$
$\lambda_6 = 0$

$$t(G) = \frac{1}{6}(4.5616)(3)(1)(1)(0.4384) = 3$$

Complexity Descriptors (cont'd)

The Zagreb Indices sum up vertex degrees over vertices and edges

$$M_1 = \sum_{\text{vertices}} (d_i)^2 \qquad M_2 = \sum_{\text{edges}} (d_i \cdot d_j)$$

Requirements for "good" molecular complexity descriptors:

1. Increase with the number of vertices and edges.
2. Reflect the degree of connectedness.
3. Be independent from the nature of the system.
4. Differentiate non-isomorphic systems.
5. Increase with the size of the graph, branching, cycling, and number of multiple edges.

(Even More) Popular Descriptos

1. Atom Environments
   - Hierarchical Orderd Spherical Description of Environment (HOSE) Codes.
   - Radial Distribution Function.
     Is a correlation based method and takes the distribution of atomic properties in the molecule into account.
   - Local Atom Environment Kernel.
     Takes local atomic similarity and spherical neighborhood into account.
2. Eigenvalue Decomposition
   - Characteristic Polynomial.
   - Burden Matrix and BCUT Descriptors.
     Encode atomic properties relevant to intermolecular interactions.
3. 3D Structure
   - Weighted Holistic Invariant Molecular (WHIM) descriptor.
     Capture information regarding size, shape, symmetry and atom distribution.

Chemical Space *versus* Reference Space

The high dimensional chemical space might often be too complex to carry out a meaningful analysis.

Observation:

- Correlation effects between descriptors distort reference space.

Solutions are:

1. Design low-dimensional reference space.
2. Remove "simple" correlated descriptors.
3. Reduce dimensionality by combining important descriptors (e.g. Principle Component Analysis (PCA)).

QSAR: Basic Equation

A QSAR equation relates biological activity to variations in computed (or measured) parameter values for a series of molecules:

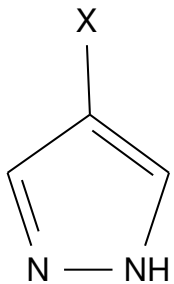
$$\text{Biological Activity} = C_0 + \sum_{i=1}^{n} C_i \cdot P_i$$

where the parameters $P_1 - P_n$ (molecular descriptors) are computed for each molecule and the coefficients $C_0 - C_n$ are calculated by fitting variations in the parameters and biological activity *via* a regression analysis.

Rule of thumb: The most parsimonious QSAR model, explaining a maximum of the variation with a minimum of variables is in general the preferable model.

QSAR: Example Inhibition of ADH

Inhibition of alcohol dehydrogenase (ADH) by 4-substituted pyrazoles.
The QSAR equation was derived by Hansch for rat liver ADH.



$$\log(1/K_{\text{inh}}) = 4.87 + 1.22 \cdot \log P - 1.80 \cdot \sigma_{\text{meta}}$$

1. The near unit value of $\log P$ suggests that the partitioning of X between enzyme and water parallels that between octanol and water, with a complete desolvatisation of the substituent.
2. The negative value of σ_{meta} is consistent with the fact, that the pyrazole binds to the catalytic zinc atom. Electron releasing substituents X increase the electron density on the nitrogen and hence tighten the binding to the metal atom.

Hansch C, Klein T, (1986) Molecular graphics and QSAR in the study of enzyme-ligand interactions. On the definition of bioreceptors; Acc Chem Res 19(12):392-400; DOI: 10.1021/ar00132a003
Hansch C et al., (1986) A quantitative structure-activity relationship and molecular graphics analysis of hydrophobic effects in the interactions of inhibitors with alcohol dehydrogenase; J Med Chem, 29(5):615-620; DOI: 10.1021/jm00155a005