Mathematical models in quantitative pharmacology: further development of Stephenson theory and QSAR

Optimization problems of ligand - receptor interactions

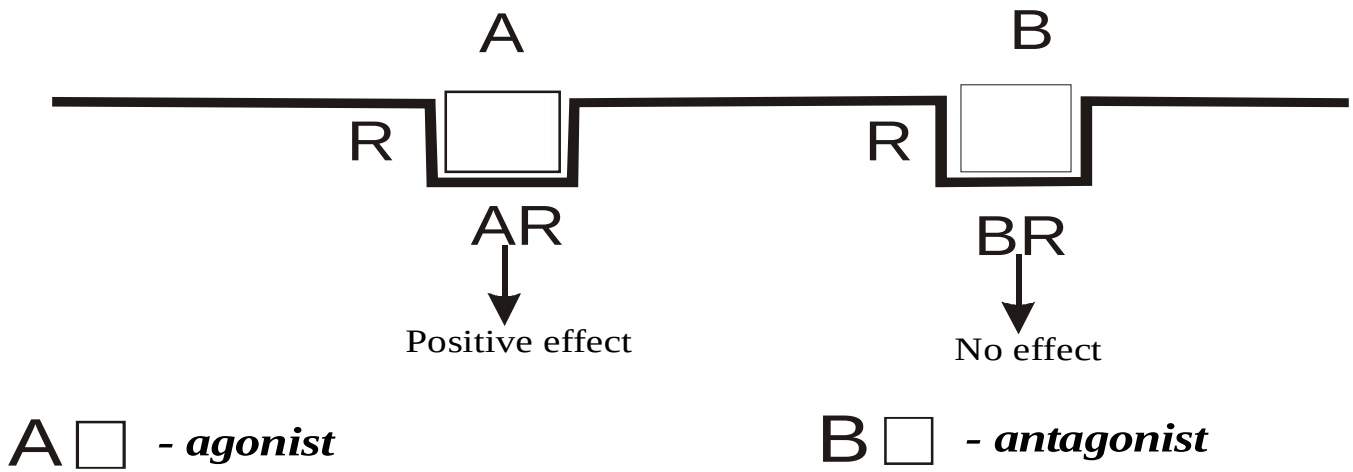
Peter Milanov

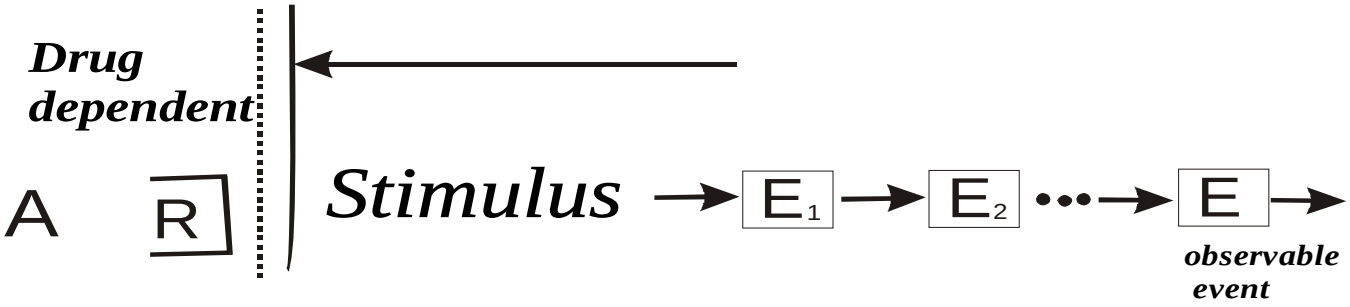
Special experimental data (biolog. experiment)

What kind of biological response is measured?

- inhibition or exhibition of muscles;
- electrical potential;
- other responses.

Drug agonist and antagonist





Quantitative pharmacology

1. Problem

Experimental data (ED)

A_1	A_2	….	A_n
E_1	E_2	…..	E_n

How to solve the problem of the best fitting?

- class of fitting functions;
- criteria of best fitting;
- methods solving these optimization problems.

Law of mass action

R – total number of receptors

A – total number of molecules

X – number of *AR* molecules

$$A + R \underset{k_2}{\overset{k_1}{\rightleftharpoons}} AR$$

$$V_{assoc} = k_1 (A - X)(R - X)$$

$$V_{dissoc} = k_2 X$$

The rate of formation

$$\frac{dX}{dt} = k_1(A - X)(R - X) - k_2 X$$

Ordinary differential equation-Riccati equation

At equilibrium (steady state)

$$k_1(A - X)(R - X) = k_2 X$$

$X << A$

$$X = \frac{A.R}{A + k_A}$$

$k_A = \frac{k_2}{k_1}$ - dissociation constant

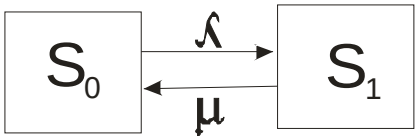
$1/k_A$ - affinity

Steady State

- Receptor R – System *S* with two states:

S_o free $\qquad\qquad$ S_1 occupied

Mass Service System



$$p_0 = \frac{m}{l+m} \qquad\qquad p_1 = \frac{l}{l+m}$$

Idea -What to do?

$$\sum_{j=1}^{R} x_j(i) = X(A)$$

$$\sum_{i=1}^{n} x_j(i) = r(A)$$

$$nX(A) = r(A)R$$

$$l = f(k_A, A)$$

$$m = g(k_A, A)$$

$$p_0 = \frac{g}{f+g}$$

$$p_1 = \frac{f}{f+g}$$

Absolute traffic capacity

$$\frac{fg}{f+g}$$

Ion Channels

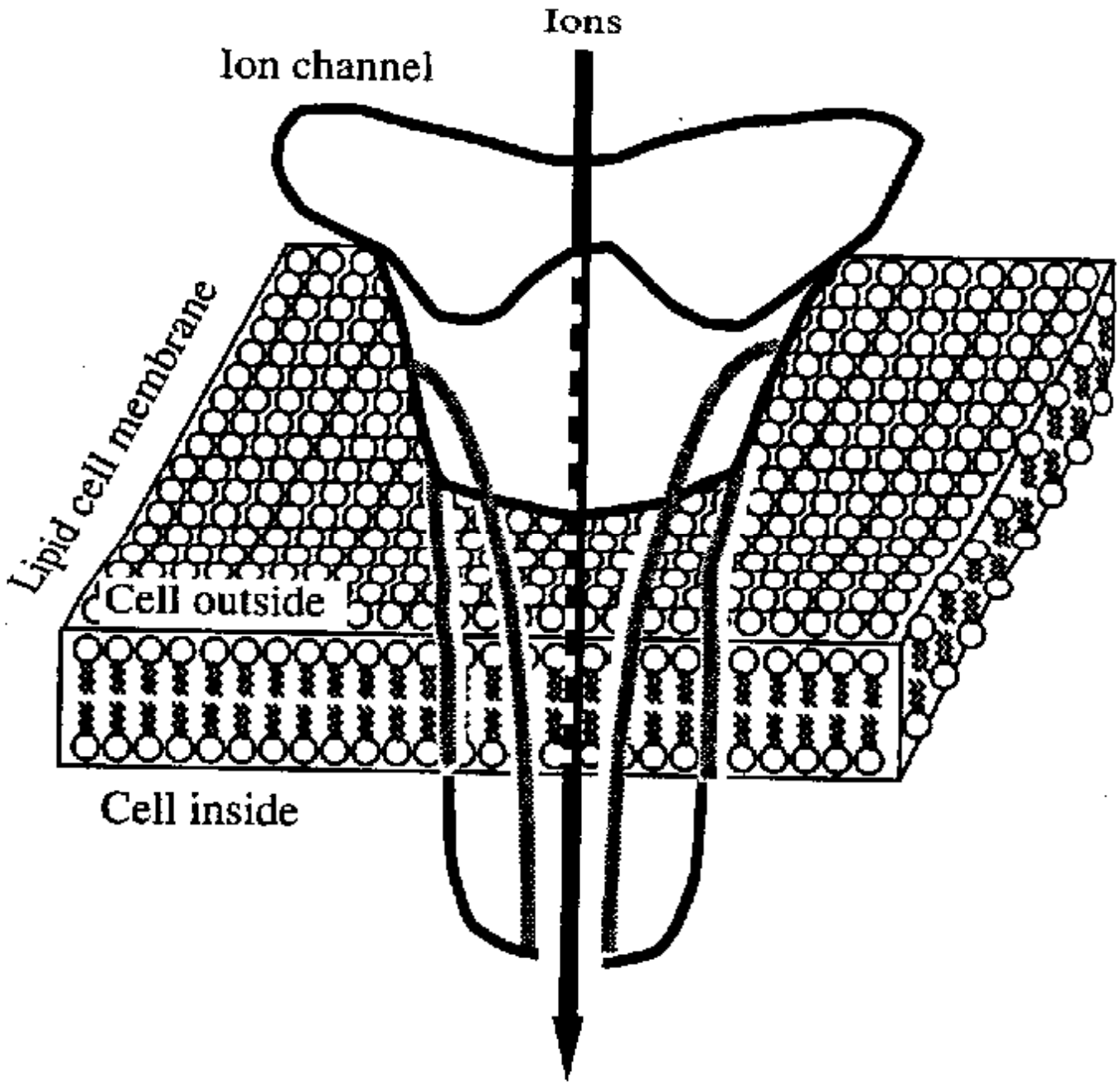
Ion channels are proteins that span the lipid bilayer

Bilayer forms the cell membrane

Ions, such as sodium, potassium, and chlorine, cannot cross the lipid bilayer

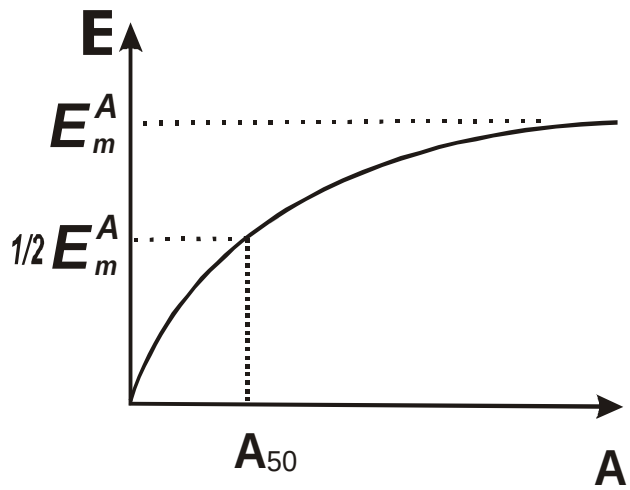
When the channel is in an open conformation state, ions can pass through the inside of the channel protein and thus enter or exit the cell

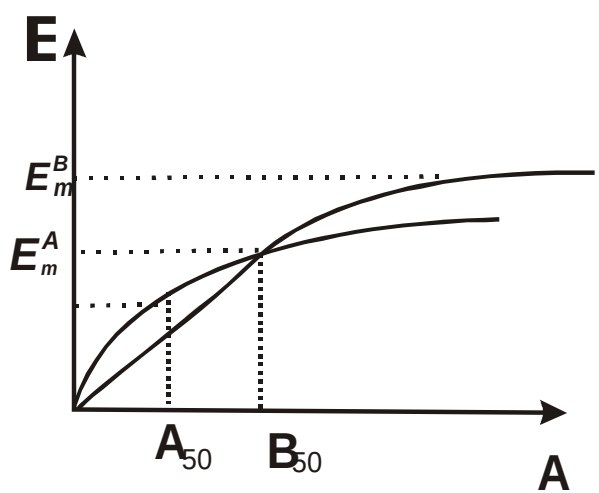
The life time of ion channels could not depend on the nature of the agonist. Another large group of receptors is whose effects are transduced by G-proteins


Ions
Ion channel
Lipid cell membrane
Cell outside
Cell inside

Dose – Response relations

A - Dose **E -effect**

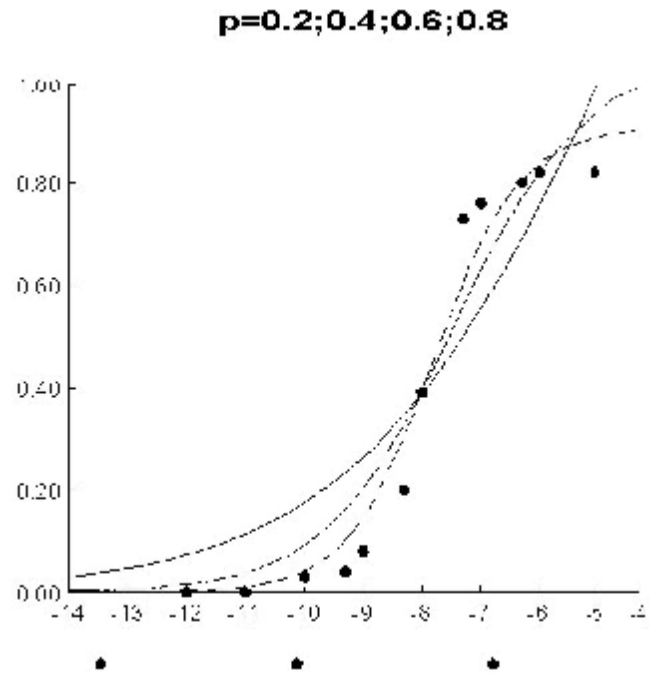




efficacy

potency

Experimental data
$(A_1,E_1), (A_2,E_2),\ldots,(A_n,E_n)$



$$E = \frac{L_1 A^p}{A^p + L_2}$$

Classical theory. Theory of Stephnenson axiom.

E. J. Ariëns 1954
Extension of Clark theory

$E = aX$ - directly proportional

$$E_{\max} = aR \Rightarrow E = \frac{E_{\max} A}{A + k_A}$$

R. P. Stephenson 1956
Modification of Ariëns theory

1. $E_{\max}$ can be produced by an agonist drug without total occupancy.

2. D –R complex provides a stimulus S to the tissue

$$S = e_A \frac{X}{R} = e_A \frac{A}{A + K_A}$$

e_A – Stephenson efficacy

3. The effect E is an unknown function *f(S): E = f(S)*

Katz interaction scheme

$$A + R \xrightarrow{k_A} AR \xrightarrow{k_{AR}} AR^*$$

- A year later, after Stephenson's work, another paper (Katz 1957)was published, where Katz was also seeking to explain partial agonism.
- His approach was entirely different from Stephenson's.
- He wrote down a simple explicit reaction scheme, which is an approximation to the real mechanism.

R. F. Furchgott 1964 – Nobel price

Method for of estimation of k_A

named “method of irreversible antagonist”

term “intrinsic efficacy” ε_A: $e_A = e_A.R$

D. Mackay 1966

It is theoretically impossible to estimate absolute value of e_A .

J.W. Black P. Leff 1983

Operational models of pharmacological agonism

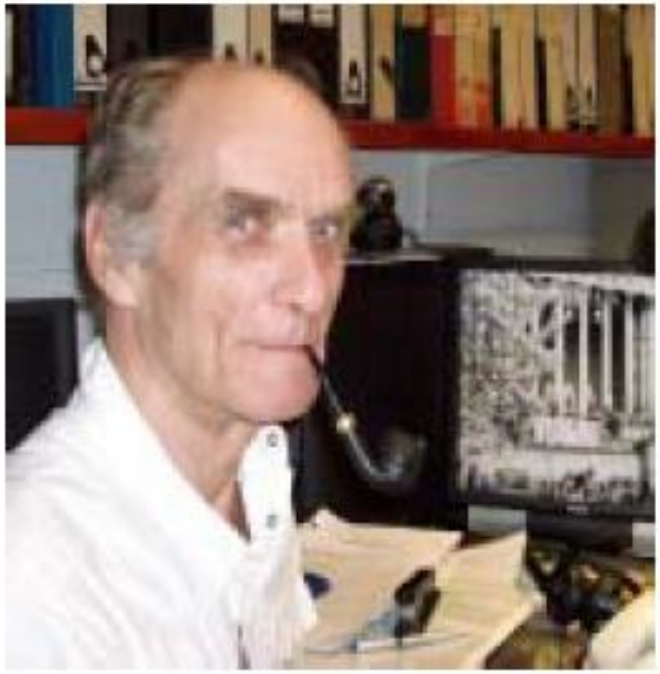
$$E = \frac{E_m X}{K_E + X}$$

K_E - value of X that elicits $1/2E_m$

$t = \frac{R}{K_E}$ - “operational efficacy”

History of the Problem

The quantitative analysis of drug–receptor interactions: a short history



Professor David Colquhoun

Prof. David Colquhoun
Dept. of Pharmacology
University College London

Key players

Archibald Vivian Hill (1886–1977, Cambridge and UCL). Hill (1909) discovered the Langmuir binding equation [9 years before Langmuir (1918)], and applied it to his studies on nicotine and curare.

Alfred Joseph Clark (1885–1941, UCL and Edinburgh). Clark made the first serious attempts after Hill to apply physical laws to receptors. His book and reviews were very influential, although his analysis of competitive antagonism failed to identify the advantages of the dose-ratio approach.

John Henry Gaddum (1900–1965, UCL and Edinburgh). Gaddum was the first to write the equation for competitive binding at receptors (in 1937, a Physiological Society abstract). But it referred to binding not response, and so was not usable until Schild's work. In fact, these equations date back to 1914, and appeared in Haldane's book *Enzymes*, published in 1930 [68].

Heinz Otto Schild (1906–1984, UCL). Schild showed, in 1949 and the 1950s, how to obtain the real equilibrium constant for an antagonist from measurements of responses, and so crude measurements such as IC_{50} values were no longer needed. This was enormously important because it was the first usable way of obtaining real physical information about receptors.

Jeffries Wyman (1901–1995) (UCL, Harvard and Rome). The seminal article of Wyman and Allen (1951) [35] described how selective affinity for an active state was linked to conformation change. This was written in the context of haemoglobin (and enzymes). If it had been read by pharmacologists at the time it might have saved us a lot of argument and misunderstanding.

Robert Stephenson (1925–2004, Edinburgh). Stephenson's influential 1956 paper proposed clearly that to understand an agonist it was important to distinguish between its ability to bind and its ability to activate once bound. He made a brave attempt to provide a general theory for agonists, based on the sort of null methods that Schild had exploited so successfully for antagonists. Sadly this proved over-ambitious (it is a pity that he was not aware of Wyman's work).

Bernard Katz (1911–2003, UCL). In 1957, del Castillo and Katz, characteristically, proposed not a general theory but a very simple physical mechanism, in an attempt to explain the supposed partial agonist action of decamethonium. This mechanism was sufficient to illustrate beautifully the nature of the affinity–efficacy (or binding–gating) problem. It provided a counter example that showed that the Stephenson approach was wrong (although Wyman's work had actually already shown that in a much more general way).

Alan Geoffrey Hawkes (1938–present, UCL, Durham and Swansea). Hawkes is responsible for much of the general theory underlying the interpretation of single-channel recordings. His work, in conjunction with the development by Neher and Sakmann of the patch-clamp method (1976), enabled the first separate measurements of affinity and efficacy (for the nicotinic acetylcholine receptor [52,72]).

Scheme of the THM

$$A \rightarrow AR \rightarrow S \rightarrow E$$

Theoretical Hyperbolic Model of drug-receptor interaction: affinity and efficacy of partial agonist

Basic Assumptions of the model

a) Interaction D-R bimolecular

$$X = \frac{A.R}{k_A + A}$$ law of mass action

b) Stimulus S

$$S = e_A \frac{X}{R} = \frac{e_A A}{A + k_A} = e_A X$$ Stephenson, Furchgott

c) D –R data is fitted by a hyperbolic function

$$E^A = \frac{L_1 A}{A + L_2}$$ (R. B. Barlow – 1999 – over 70%)

d).E_m^T exists (depends only the tissue; $\exists$ drug A producing E_m^T)

A- Full agonist

e) S – R relation – drug independent property

f) Equal stimuli lead to equal effect.

Consequences of axioms of THM

There exist constants C_1 and C_2 (depended only of T) such that

$$E^A = \frac{C_1 S}{S + C_2}$$

Explicit formulas for affinity and efficacy

$$k_A = \frac{C_1 L_2}{C_1 - L_1} \qquad e_A = \frac{C_2 L_1}{C_1 - L_1}$$

dissociation constant efficacy

Pharmacological interpretation of the parameters and their calculation

$$L_1 \approx E_m^A$$

max effect of A

$$L_2 \approx A_{50}$$

$0.5E_m^A$

D – R data

$$C_1 \approx E_m^T$$

max effect of A on the tissue T full agonist

$$C_2$$

elementary measure (unit) of stimulus elicits of $0.5E_m^T$

$$l_A = \frac{E_m^A}{E_m^T} \quad (l_A < 1)$$

$$k_A = \frac{A_{50}}{1 - l_A}$$

$$e_A / c_2 = \frac{l_A}{1 - l_A}$$ (Mackay)

Analysis of the model

"amplifier" $$m_A = \frac{l_A}{1 - l_A}$$

"intrinsic stimulus" $$c_2 = \frac{C_2}{R}$$

Stimulus $$S = c_2 m_A X$$

Biological effect $$E^A = \frac{E_m^T m_A X}{m_A X + R}$$

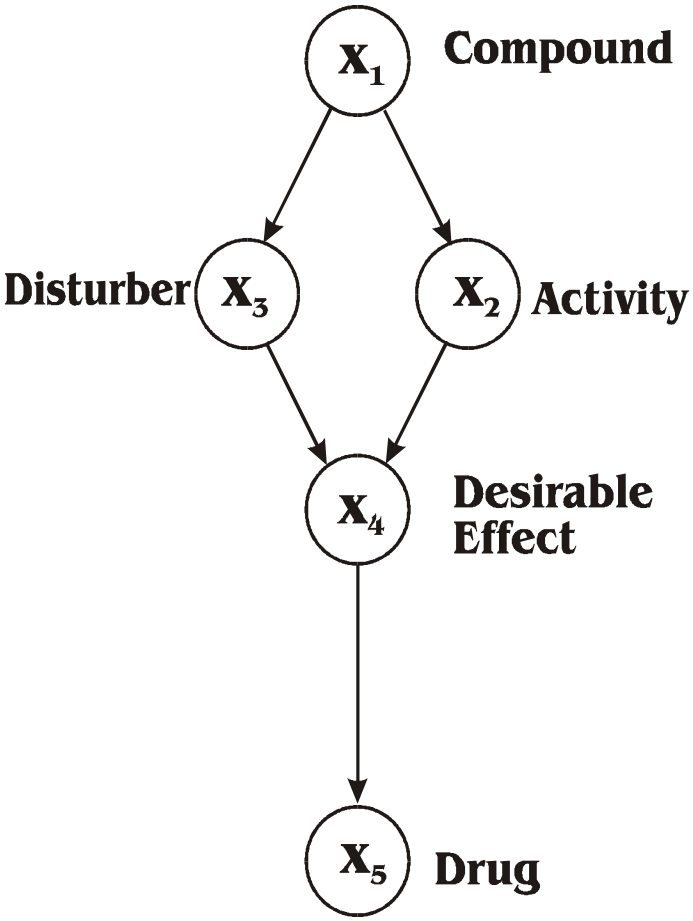
$$k_A = (m_A - 1) EC_{50}^A \qquad EC_{50}^A = \left(\frac{m_A + 1}{m_A - 1} \right) [A_{50}]$$

Partial agonists haven't a receptor reserve

Quantitative Structural – Activity Relationship (QSAR)

Dose –effect (response)

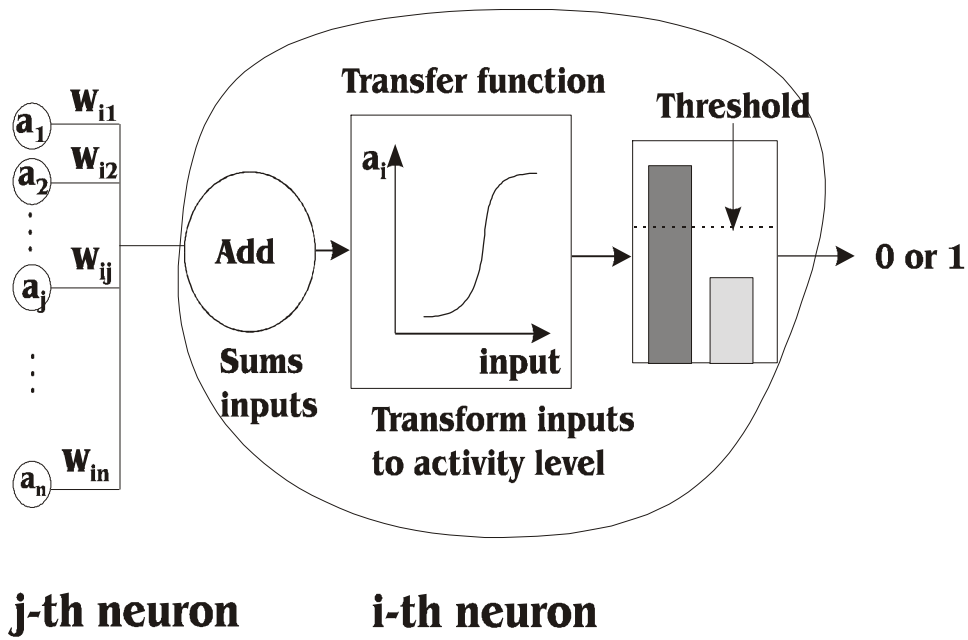
Structure of the drug – effect (response)

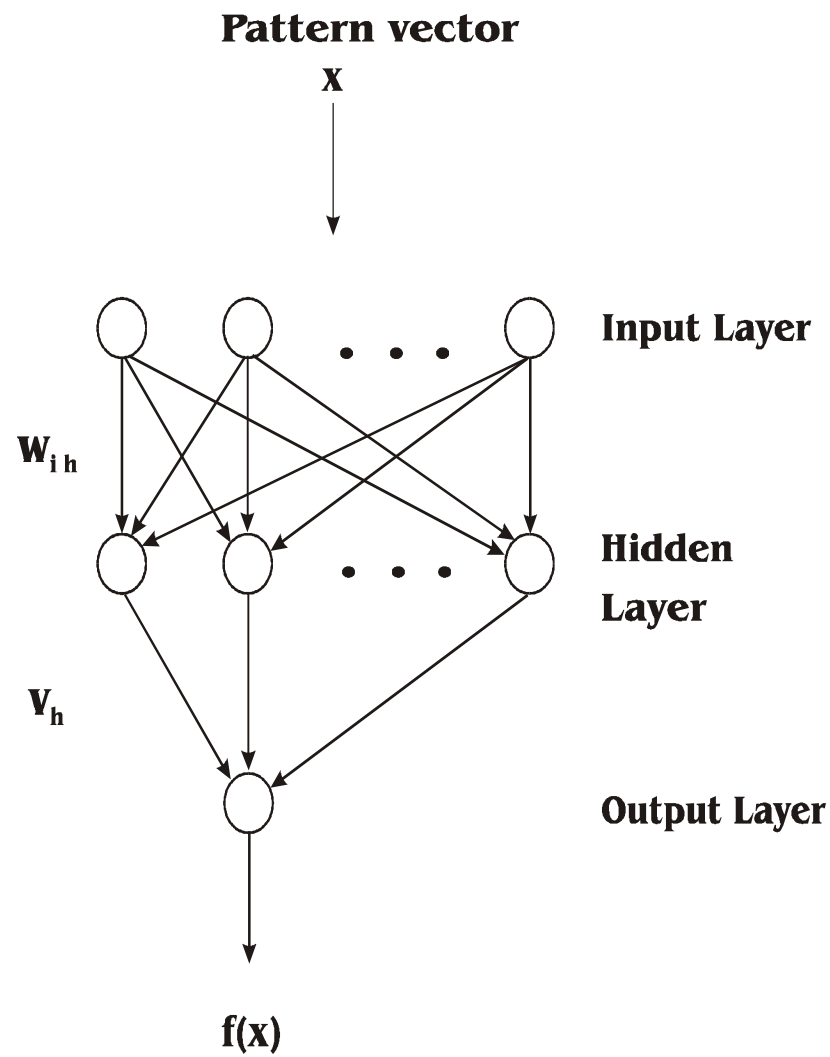


Problem: investigation of structure – receptor relations

What kind of mathematical tools have been used?

Artificial neural network (ANN)





Activity, affinity and efficacy

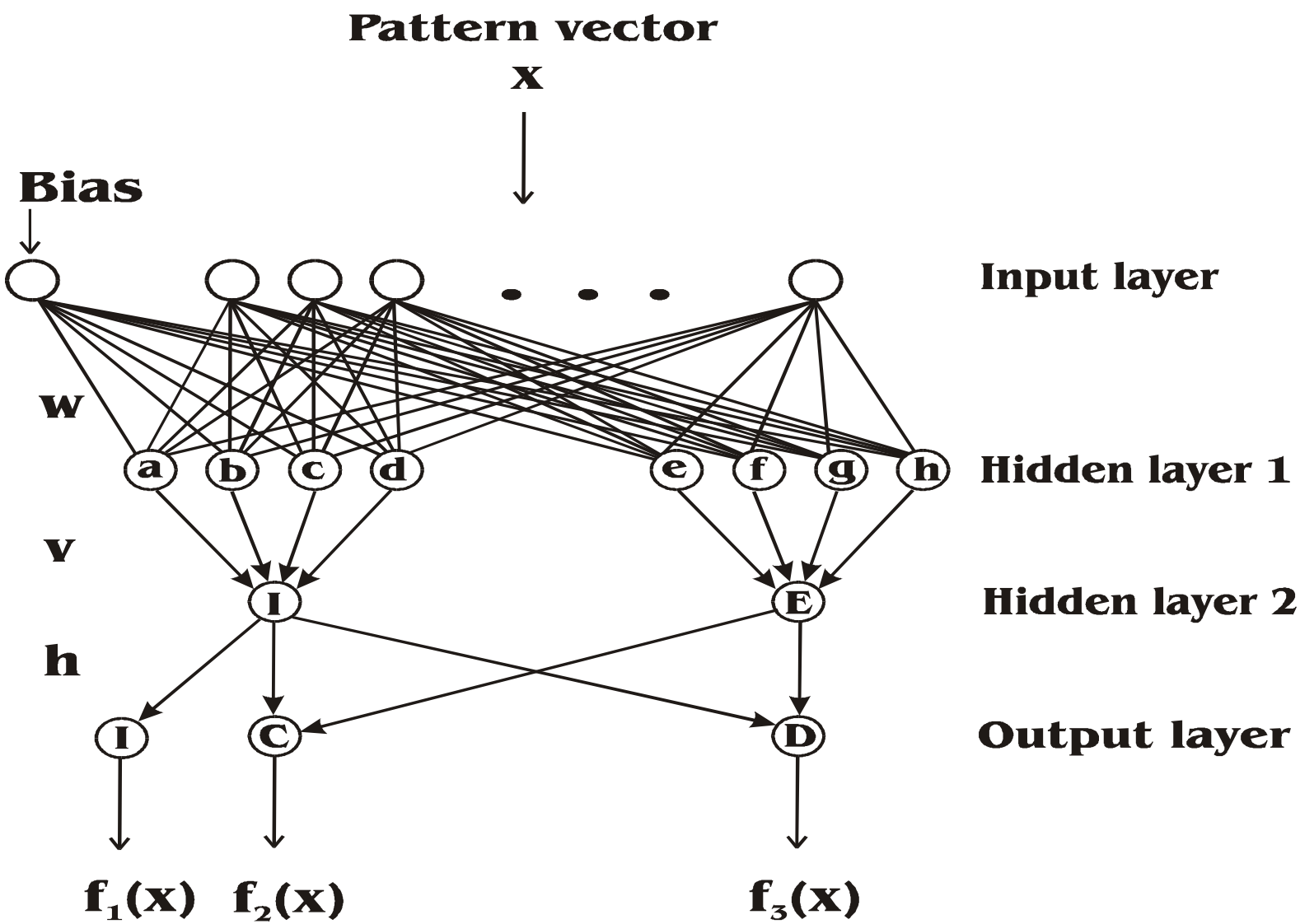
$$K_A = \frac{EC_{50}^{A} E_m^{T}}{E_m^{T} - E_m^{A}} \quad (1) \qquad\qquad e_A = \frac{E_m^{A}}{E_m^{T} - E_m^{A}} \quad (2)$$

For Output layer we can use formula (1) as transfer function of unit C and formula (2) as transfer function for unit D.

Training of ANN – using of database NCBI, KEGG and ExPaSy

After training of this neural network, we expect to predict the following three characteristics for the compounds with novel structure: EC_{50} - a measure for their activity and K_A and ε_A - parameters which allow to compare their selectivity. The commonly used architecture for modeling of QSAR in the pertinent literature is a three layered feed forwarded network with sigmoidal hidden-unit activity and a single linear output neuron. This architecture does not allow to predict efficacy and selectivity of the compounds.

Network architecture in modeling selectivity and efficacy of enkephalin analogues.



Since the goal of the present neural network modeling concerns not only activity (potency) of the enkephalin analogues, but their selectivity and efficacy too, we suggest the following network architecture: a four-layered feed-forward network with sigmoidal hidden-unit activity of Hidden layer 1, linear units activity for neurons from Hidden layer 2 and Output neuron I. The sigmoidal transfer function for Hidden layer 1 activity is:

$$f(x) = (1 - \exp^{-}(\sum_{i=1}^{n} w_{ij} x_i - v_j))^{-1}$$

where x is a n-dimensional input vector, coding the structure of the enkephalins; w, v and h are the weight matrixs of the Hidden layer 1, Hidden layer 2 and Output layer respectively. The threshold ν_j, which is the weight of the bias neuron, is the EC_{50} value of the compounds and concerns a, b, c and d units. For the next e - h neurons form Hidden layer 1, the threshold ν_j is the peptides:

The linear activity function in the Hidden layer 2 for neuron I is*:*

$$EC_{50}^{A}(x) = \sum_{j=1}^{4} n_j (1 - \exp^{-}(\sum_{i=1}^{n} w_{ij} x_i - EC_{50}))^{-1}$$

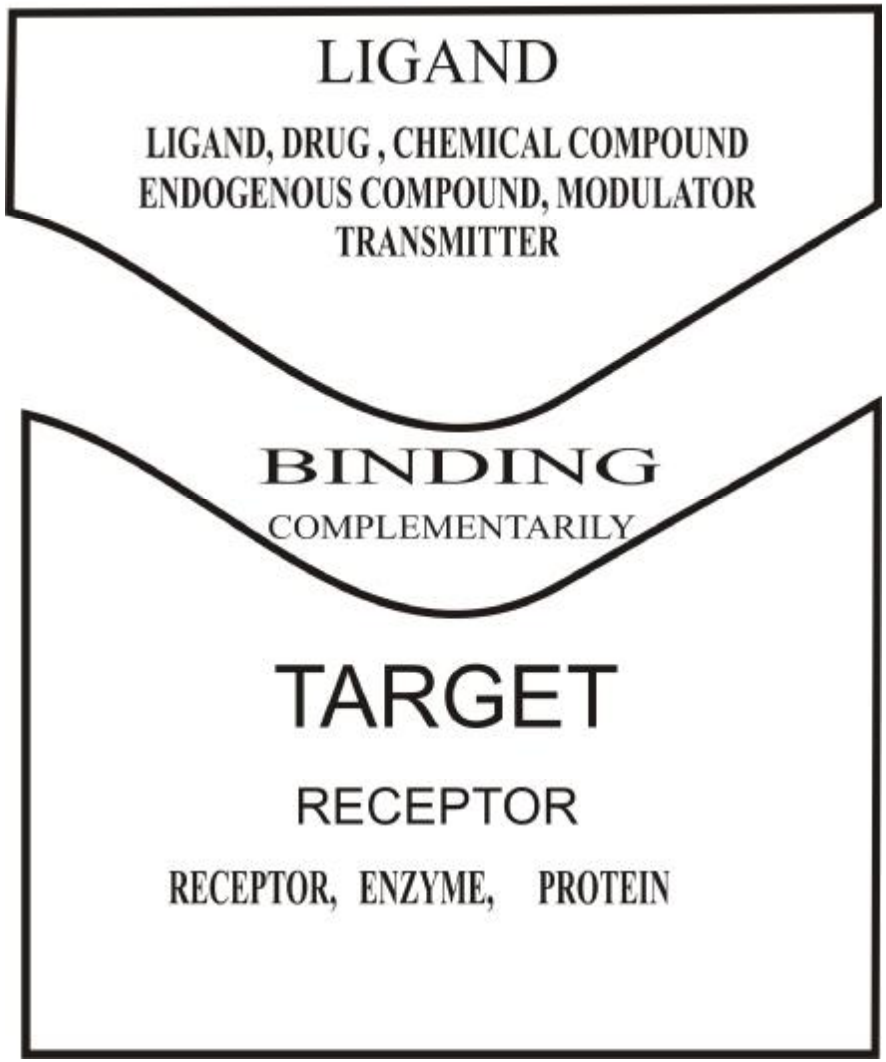
For neuron E it is:

$$E_m^{A}(x) = \sum_{j=5}^{8} n_j (1 - \exp^{-}(\sum_{i=1}^{n} w_{ij} x_i - E_m^{T}))^{-1}$$

Models of similarity of chemical compounds

- QSAR Models- ANN
- Models of Protein Threading Problem

COMPLEMENTARILY



Chemical spaces and molecular similarity

- Similar Property Principle – Molecules having similar structures and properties should also exhibit similar activity. (Often but not always true)

- Thus, molecules that are located closely together in chemical reference space are often considered to be functionally related.

LARGE MOLECULAR SIMILARITY

The training phase proceeds as follows:

1. Extract a random set of training patterns $\{\mathbf{p}_i, i = 1,2,\ldots,k; \mathbf{p}_i \in P\}$ from the data set P.
2. Map the patterns $\mathbf{p}_i$ onto $\Re^m$ using a conventional nonlinear mapping algorithm ($\mathbf{p}_i \rightarrow \mathbf{y}_i$, $i = 1,2,\ldots,k$, $\mathbf{y}_i \in \Re^m$).
3. Select a set of reference patterns $\{\mathbf{r}_i, i = 1, 2,\ldots,l; \mathbf{r}_i \in P\}$ from the data set P.
4. Compute the similarity $\{s_{ij}, i = 1,2,\ldots,k; j = 1,2,\ldots,l: s_{ij} = sim(\mathbf{p}_i, \mathbf{r}_j)\}$ of each pattern in the training set, $\mathbf{p}_i$, to each of the reference patterns, $\mathbf{r}_j$, identified in step 3. Denote $T = \{(\mathbf{s}_i, \mathbf{y}_i), i = 1,2,\ldots,k\}$ as the training set.
5. Train a neural network, *net*, to recognize the mapping $\mathbf{s}_i \rightarrow \mathbf{y}_i$ using the input/output pairs in the training set T. Export the network *net* and its associated parameters.

Molecular descriptors and chemical spaces

TABLE 1.2. Different types of molecular descriptors

TABLE 1.2. Different types of molecular descriptors

Descriptor category	Examples
Physical properties	Molecular weight logP(o/w)
Atom and bond counts	Number of nitrogen atoms Number of aromatic atoms Number of rotatable bonds
Pharmacophore features	Number of hydrogen bond acceptors Sum of van der Waal surface areas of basic atoms
Charge descriptors	Total positive partial charge Dipole moment from partial charges
Connectivity and shape descriptors	Kier and Hall molecular shape indices
Surface area and volume	Solvent-accessible surface area

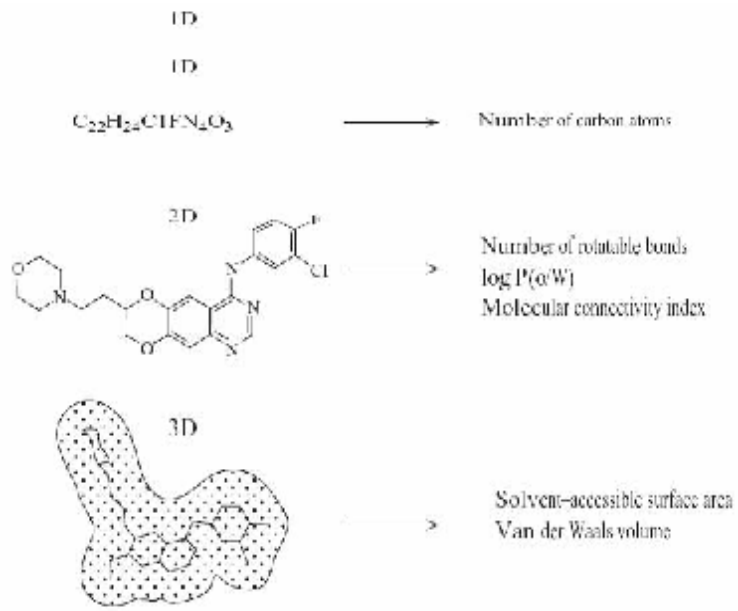


Figure 1.3. Examples of descriptors classified according to dimensionality (adapted from Bajorath 2002)

- There are no generally preferred descriptor spaces.
- Require to generate reference spaces for specific application on a case by case

Aim was the definition of a set of substructures that cover a large diversity of organic molecules. The strategy applied for the creation of substructures was as follows: (i) estriction to most common elements; (ii) systematic generation of substructures by using an isomer generator; (iii) selection of substructures by chemical experiences; (iv) elimination of very exotic substructures. Finally, a set of 1365 substructures was obtained, divided into eight groups as shown in Table I.

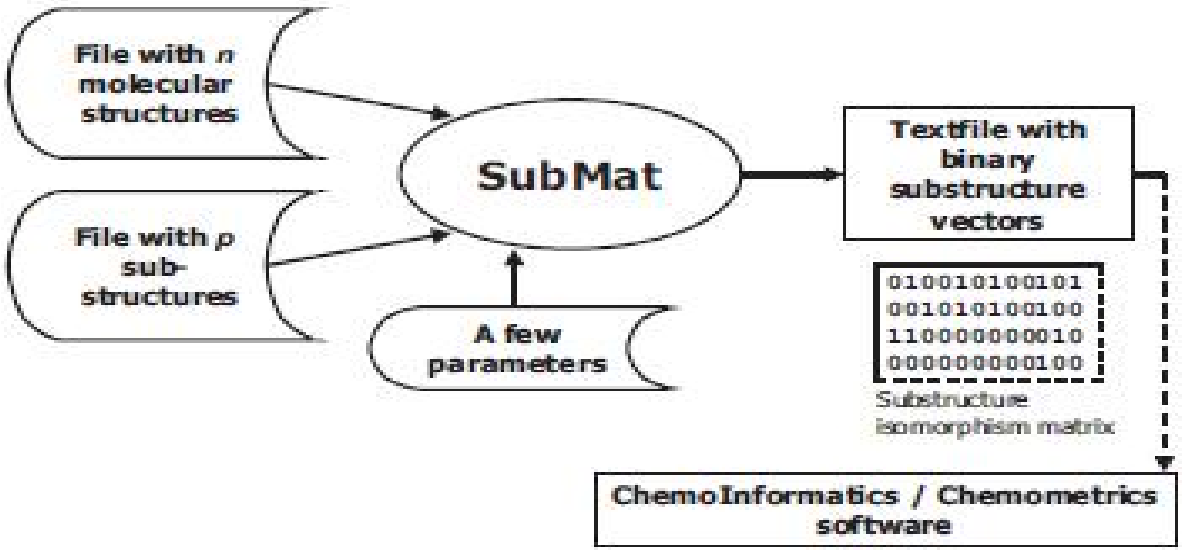


Figure 7. Software SubMat for the generation of binary substructure descriptors from a file with molecular structures and a file with substructures (both in Molfile format). The result file in text format can be easily imported into other software.

1. Compute the similarity $\{s_i, i = 1,2,\ldots,l: s_i = sim(\mathbf{p},\mathbf{r}_i)\}$ of the new pattern, $\mathbf{p}$, to each of the reference patterns, $\mathbf{r}_i$, identified in training step 3.
2. Map $\mathbf{s} \rightarrow \mathbf{y}$, $\mathbf{s} \in \Re^l$, $\mathbf{y} \in \Re^m$ using the neural network *net* derived during training step 5. Store the coordinates $\mathbf{y}$.

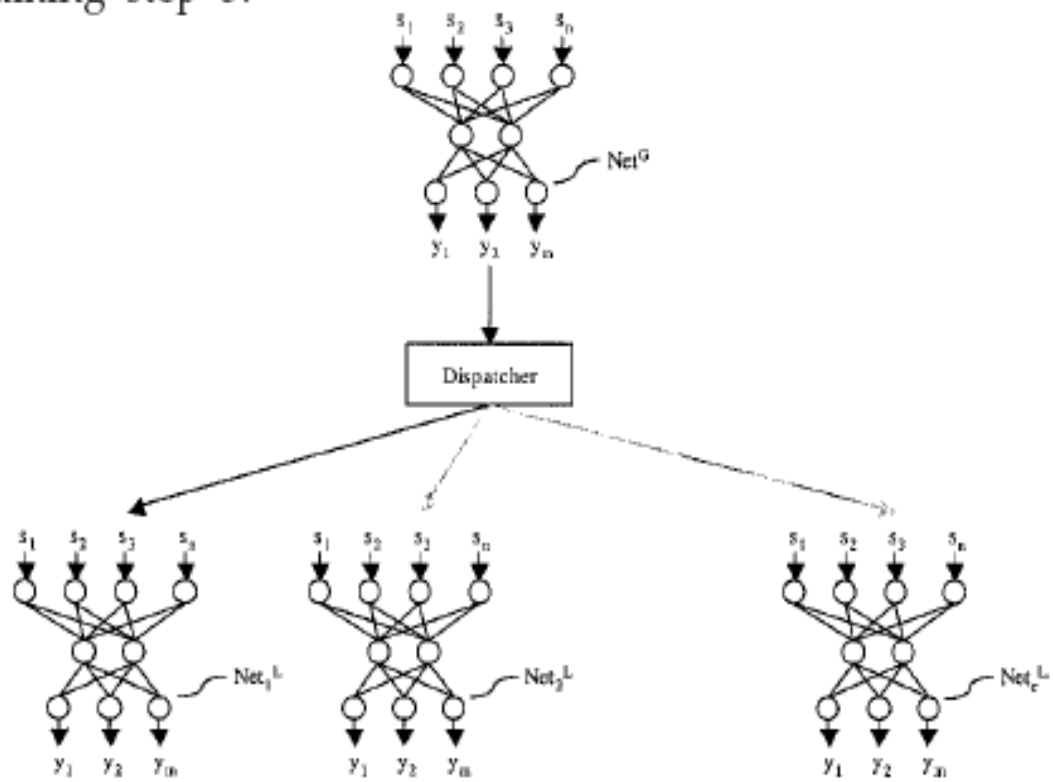


FIGURE 2. Tandem nonlinear mapping network architecture.

TABLE I. Substructure groups and number of substructures per group

Group number	Group definition	No. of substructures
1	Elements (single atom substructures)	46
2	Two-atom substructures	78
3	Single, not aromatic rings	404
4	Condensed, not aromatic rings	130
5	Aromatic rings	97
6	Other rings	39
7	Trees (chains and branches)	418
8	Functional groups	153
	Total	1365

TABLE II. Examples for two-atom substructures

Subgroup	Element		Bond Type[a]				
	Atom 1	Atom 2	s	d	t	a	n
C and another	C	C	+	+	+	+	+
	C	N	+	+	+	+	+
	C	O	+	+			+
	C	S	+	+			+
	C	A	+	+	+	+	+
	C	F	+				
	C	Cl	+				
	C	Br	+				
	C	I	+				
N and another	N	N	+	+	+	+	+
	N	O	+	+			+
	N	S	+	+			+
	N	A	+	+	+	+	+
	N	Q	+	+	+	+	+

[a]Bond types are s, single; d, double; t, triple; a, aromatic; n, not defined. A plus (+) indicates that this substructure is used.

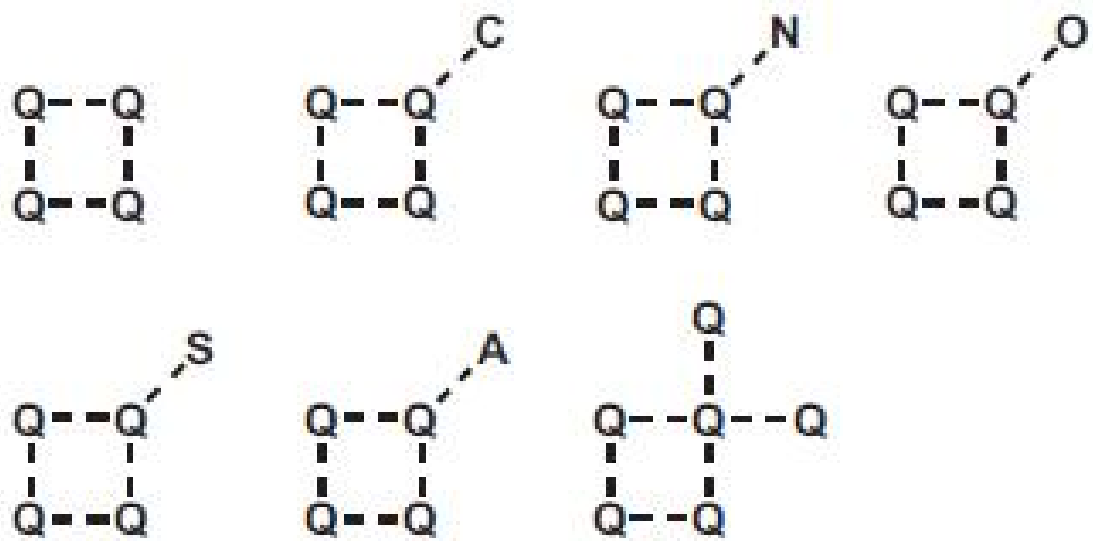


Figure 1. Examples for substructures in group 3 (single, not aromatic rings). Four-membered rings made only by Q-atoms and the used substitutions are shown. All bonds have not-defined type. Such ring substructures have been defined for ring sizes 3 to 8.

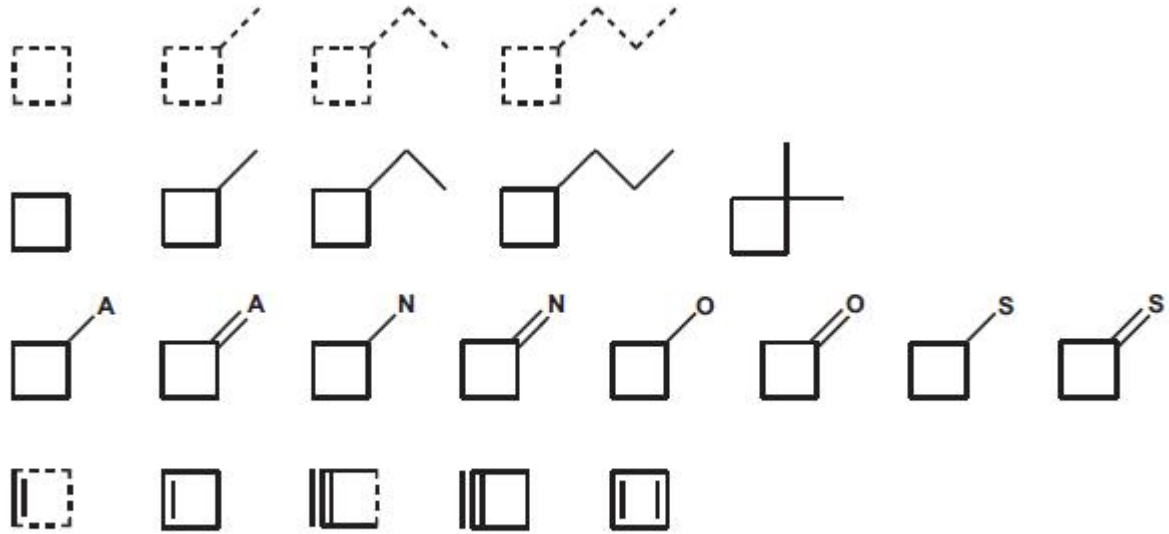
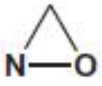
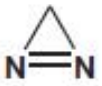


Figure 2. Examples for substructures in group 3 (single, not aromatic rings). Rings made only by C-atoms and used substitutions as well as unsaturations are shown for 4-membered rings. Such ring substructures have been defined for ring sizes 3 to 8. A dotted lined denotes a not defined bond type.

$$c + n + o = r \text{ for } r = 3, 4, 5 \tag{1}$$

$$h = h_{max}, h_{max} - 2, h_{max} - 4, \ldots \text{ with } h > 0 \tag{2}$$

$$h_{max} = 2c + n \tag{3}$$

						
BCF	1,603	3	1,023	12	912	75,463
IR	0	0	4	0	0	93
MS	15	0	13	0	3	1,232

						
BCF	35	16,425	24	625	54	0
IR	0	2	0	0	0	0
MS	0	229	0	11	0	0

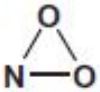
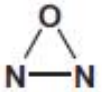
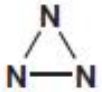
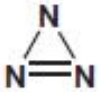
					
BCF	4	0	31	185	41
IR	0	0	0	0	0
MS	0	0	0	4	0

Figure 3. Exhaustive set of 3-membered hetero cyclic rings made from elements C, N, and O, containing at least one hetero atom and having at least one free valence. Number of occurrences in the Beilstein Crossfire Database (BCF; 4 million compounds), in an infrared spectral database (IR; 13,484 compounds), and in a mass spectral database (MS; 106,955 compounds) are given.

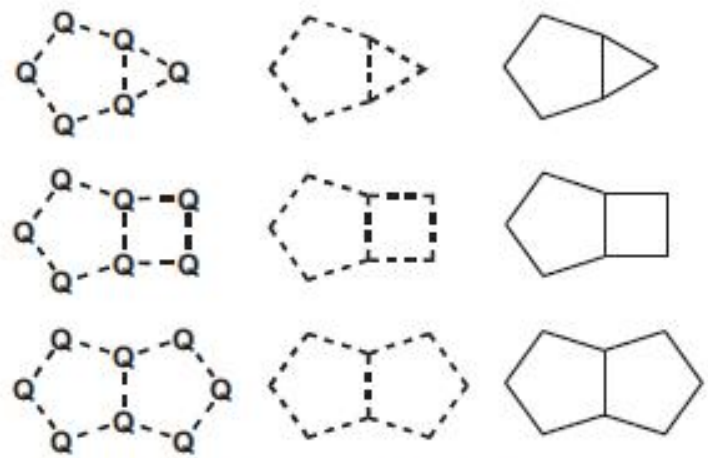


Figure 4. Selected substructures with condensed rings (group 4) obtained by the combination of a 5-membered ring with a 3-, 4- or 5-membered ring.

TABLE III. Number of tree substructures (isomers) with three to six C-atoms and one or two double bond equivalents (DBE)

C-atoms	Number of isomers		
	DBE = 1 one double bond	DBE = 2 two double bonds	DBE = 2 one triple bond
3	1	1	1
4	3	2	2
5	5	6	3
6	13	16	7
sum	22	25	13

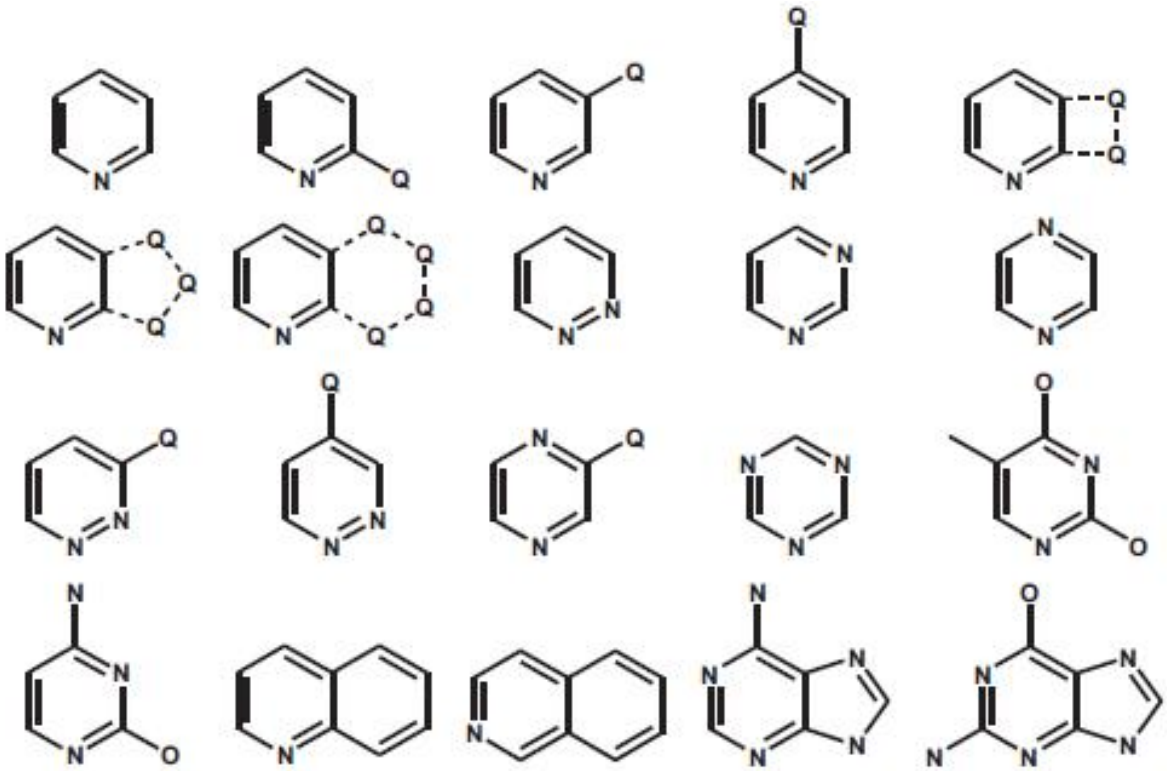
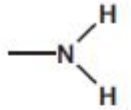
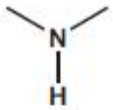
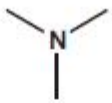
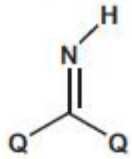
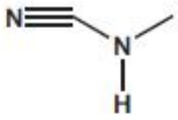
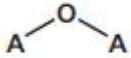
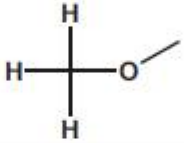
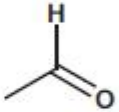
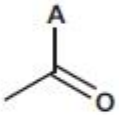


Figure 5. Substructures used containing a 6-membered N-aromatic ring (group 5).

	1	2	3	4	5
					
IR	9.52	14.65	13.36	0.76	0.08
MS	5.07	14.81	14.84	0.34	0.01

	6	7	8	9	10
					
IR	0.49	0.88	14.33	3.70	37.88
MS	0.92	4.68	16.37	1.66	36.12

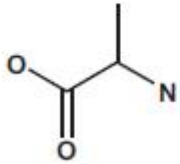
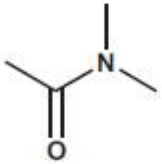
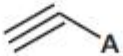
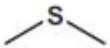
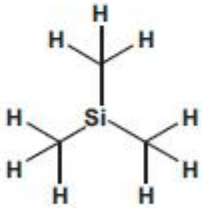
	11	12	13	14	15
					
IR	1.10	4.78	0.05	9.03	0.39
MS	2.52	4.58	0.16	7.39	4.80

Figure 6. Examples of substructures from group 8 (functional groups). The percent of compounds containing the substructure is given for two spectral databases, one with 13,484 infrared spectra, the other with 106,955 mass spectra; see Figure 9.

Discrete-Valued Feature Vectors

The components of discrete feature vectors may indicate the presence or absence of a feature, the number of occurrences of a feature, or a finite set of binned values such as would be found in an ordered, categorical variable. Each component of an *n-component binary feature vector, also called bit vectors or molecular fingerprints,*

$$\mathbf{v}_{\mathrm{A}} = \left(v_{\mathrm{A}}(x_1), v_{\mathrm{A}}(x_2), \ldots, v_{\mathrm{A}}(x_k), \ldots, v_{\mathrm{A}}(x_n)\right)$$

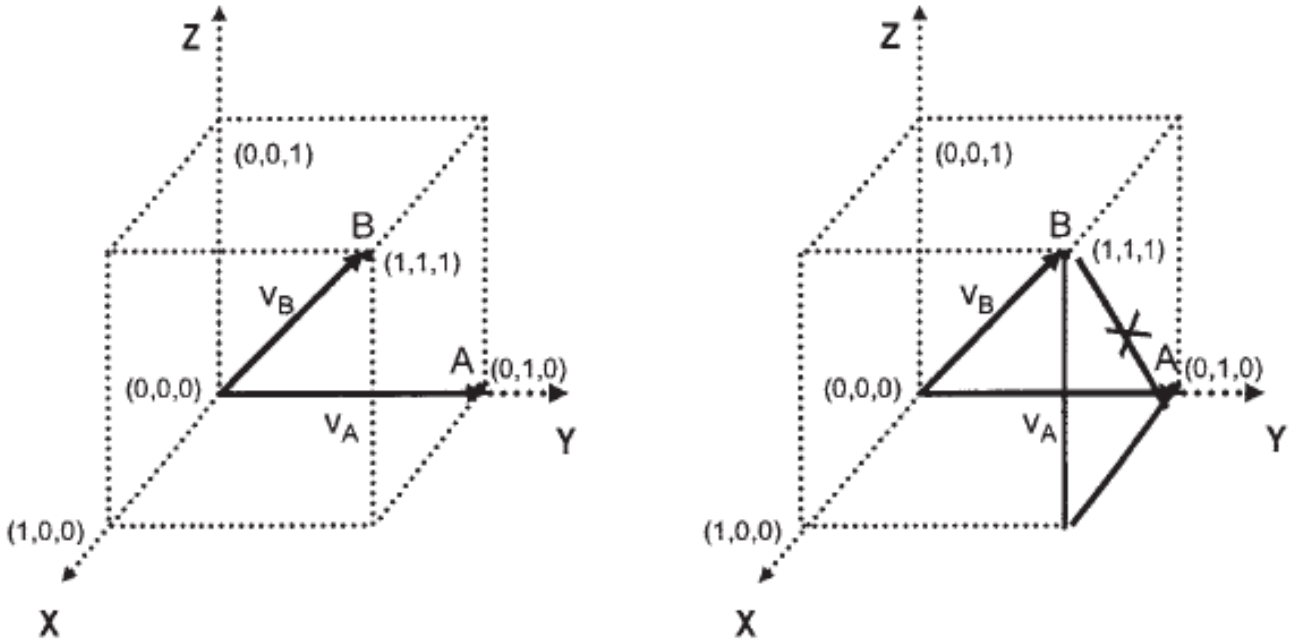
indicates the presence or absence of a given feature, x_k, *that is*

$$v_{\mathrm{A}}(x_k) = \begin{cases} 1 & \text{Feature present} \\ 0 & \text{Feature absent} \end{cases} .$$

A wide variety of features have been used in bit vectors, including molecular fragments, 3-D “potential pharmacophores,” atom pairs, 2-D pharmacophores, topological torsions, and variety of topological indices.
Binary feature vectors are completely equivalent to sets. Care must be exercised when using them to ensure that appropriate mathematical operations are carried out. The number of components in a bit vector is usually quite large, normally *n >> 100. In some cases n* can be orders of magnitude larger, sometimes exceeding a million components.

Bit vectors of this size are not handled directly because many of the components are zero, and methods such as hashing are used to reduce the
size of the stored information.
Bit vectors live in an n-dimensional, discrete hypercubic space, where each vertex of the hypercube corresponds to a set. Figure 2 provides an example of sets with three elements. Distances between two bit vectors, v_A and v_B, measured
in this space correspond to Hamming distances, which are based on the
city-block l1 metric

$$d_{\mathrm{Ham}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) = \left|\mathbf{v}_{\mathrm{A}} - \mathbf{v}_{\mathrm{B}}\right| = \sum_{k=1}^{n} \left| v_{\mathrm{A}}(x_k) - v_{\mathrm{B}}(x_k) \right| \quad .$$



$$d_{\mathrm{Ham}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) = \left|\mathbf{v}_{\mathrm{A}} - \mathbf{v}_{\mathrm{B}}\right| = \sum_{k=1}^{n}\left|v_{\mathrm{A}}(x_k) - v_{\mathrm{B}}(x_k)\right| = [1-0]+[1-1]+[1-0] = 2$$

Fig. 2. Distance between two binary-valued feature vectors v_A and v_B is not given by the Euclidean distance but the Hamming distance between the two.

The most widely used similarity measure by far is the Tanimoto similarity coefficient S_{Tan}, which is given in set-theoretic language as

$$S_{Tan}(A,B) = \frac{|A \cap B|}{|A \cup B|} \ .$$

$$S_{Tan}(A,B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k \max[A(x_k), B(x_k)]} \ .$$

The Tanimoto similarity coefficient is symmetric,

$$S_{Tan}(A,B) = S_{Tan}(B,A) \ ,$$

as are most of the similarity coefficients in use today, and is bounded by zero and unity,

$$0 \le S_{Tan}(A,B) \le 1 \ .$$

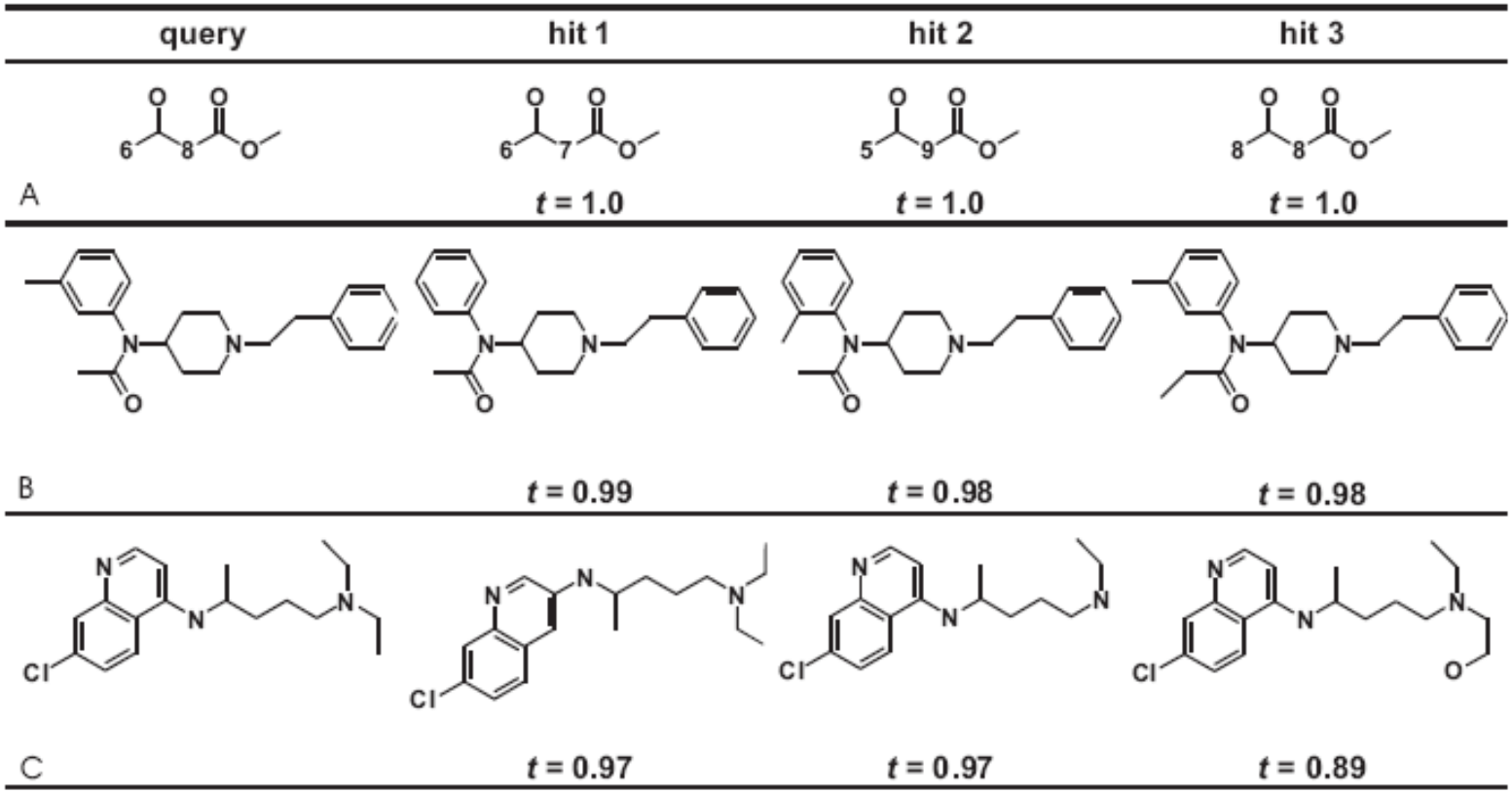


Figure 8. Examples for structure similarity searches. The query structures have been searched in a spectral database containing 106,955 compounds. *t*, Tanimoto index (structure similarity) between query and hit. Numbers within a structure denote a chain of C-atoms of the given length. Query structures are (A) 10-hydroxypalmitic acid methylester; (B) fentanyl; (C) resochine.

Tversky - *Asymmetric Similarity Indices:*

$$S_{Tve}(A,B) = \frac{|A \cap B|}{\alpha|A-B| + \beta|B-A| + |A \cap B|},$$

where α,β ≥ 0 . This generalizes the typical symmetric Tanimoto similarity measure given**, which obtains when α = β = 1. For all other values** of α and β $S_{Tve}(A,B)$ is asymmetric, that is, $S_{Tve}(A,B) \neq S_{Tve}(B,A)$. Only the two extreme forms will, however, be considered here, namely, those when α = 1 and β = 0 and α = 0 and β = 1. Their set-theoretic forms are given by

$$S^*_{Tve}(A,B) = \frac{|A \cap B|}{|A-B| + |A \cap B|} = \frac{|A \cap B|}{|A|}$$ Fraction of A similar to B

$$S^*_{Tve}(B,A) = \frac{|A \cap B|}{|B-A| + |A \cap B|} = \frac{|A \cap B|}{|B|}$$ Fraction of B similar A

$$S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) = \frac{\sum_{k} \min[v_{\mathrm{A}}(x_k), v_{\mathrm{B}}(x_k)]}{\sum_{k} v_{\mathrm{A}}(x_k)}$$

$$S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) = \frac{\sum_{k} \min[v_{\mathrm{A}}(x_k), v_{\mathrm{B}}(x_k)]}{\sum_{k} v_{\mathrm{B}}(x_k)}$$

$$0 \le S_{\mathrm{Tan}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) \le S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}), S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) \le 1$$

$$S^{*}_{Tve}(A,B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k A(x_k)} \quad .$$

As was the case for the symmetric similarity coefficient

$$0 \le S^{*}_{Tve}(A,B), S^{*}_{Tve}(B,A) \le 1 \quad ,$$

although generally **$S_{*Tve}(A,B) \neq S_{*Tve}(B,A)$.**

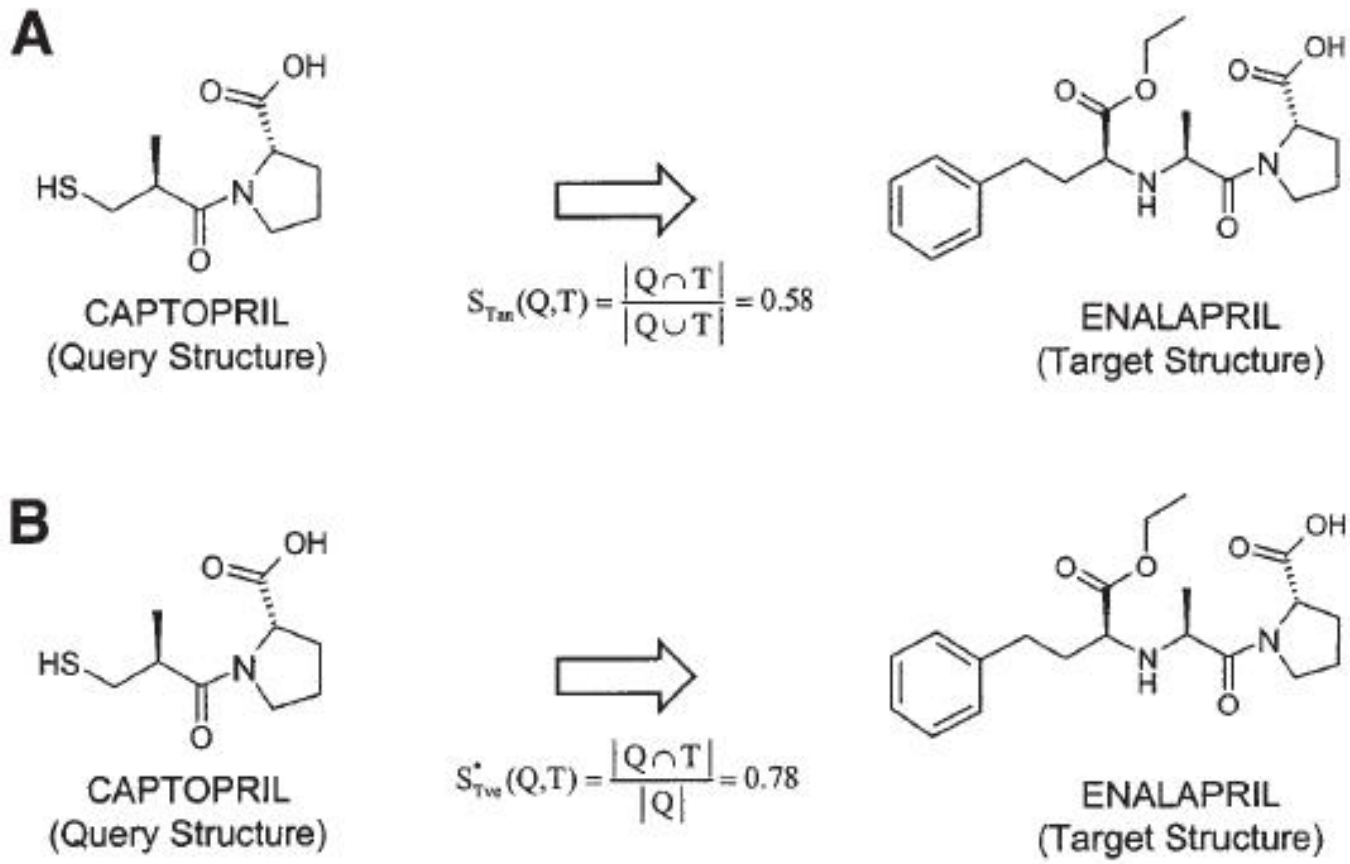


Fig. 3. Asymmetric similarity searching might provide some benefits not afforded by symmetric similarity searching. (**A**) Database searching using ISIS keys and symmetric similarity searching, S_{Tan}, will not yield enalapril as a "database hit" because the similarity value is too low, 0.58. (**B**) Whereas database searching using asymmetric similarity searching, S^*_{Tve}, could yield enalapril as a "database hit" because the asymmetric similarity value is 0.78.

Petke similarity indexes:

$$S_{Pet_{max}}(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)}$$

and

$$S_{Pet_{min}}(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} .$$

$$0 \le S_{Pet_{max}}(A,B) \le S_{Tan}(A,B) \le S_{Pet_{min}}(A,B) \le 1 .$$

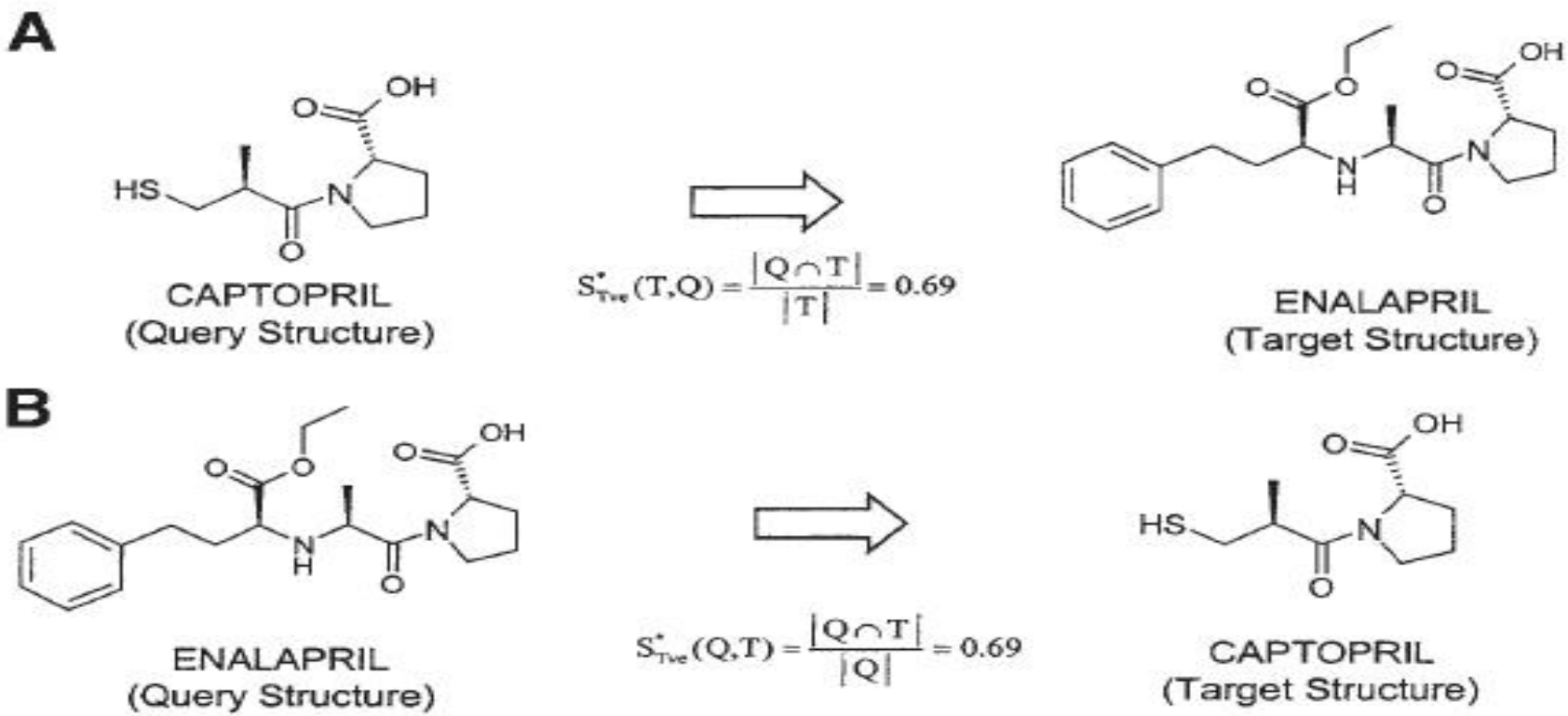


Fig. 4. (A) The other asymmetric Tversky similarity index, S^*_{Tve}, has a value of 0.69. Exchanging the roles of the query and target molecules (Q⇔T) gives (B), which shows that smaller target molecules are more likely to be retrieved from a large query structure using the asymmetric Tversky similarity index than the Tanimoto similarity index.

Chemical Graphs

Chemical graphs are ubiquitous in chemistry. A chemical graph, G*k, can be*
defined as an ordered triple of sets

$$G_k = (V_k, E_k, L_k)$$

where V*k is a set (see the Appendix for notation) of n vertices ("atoms")*

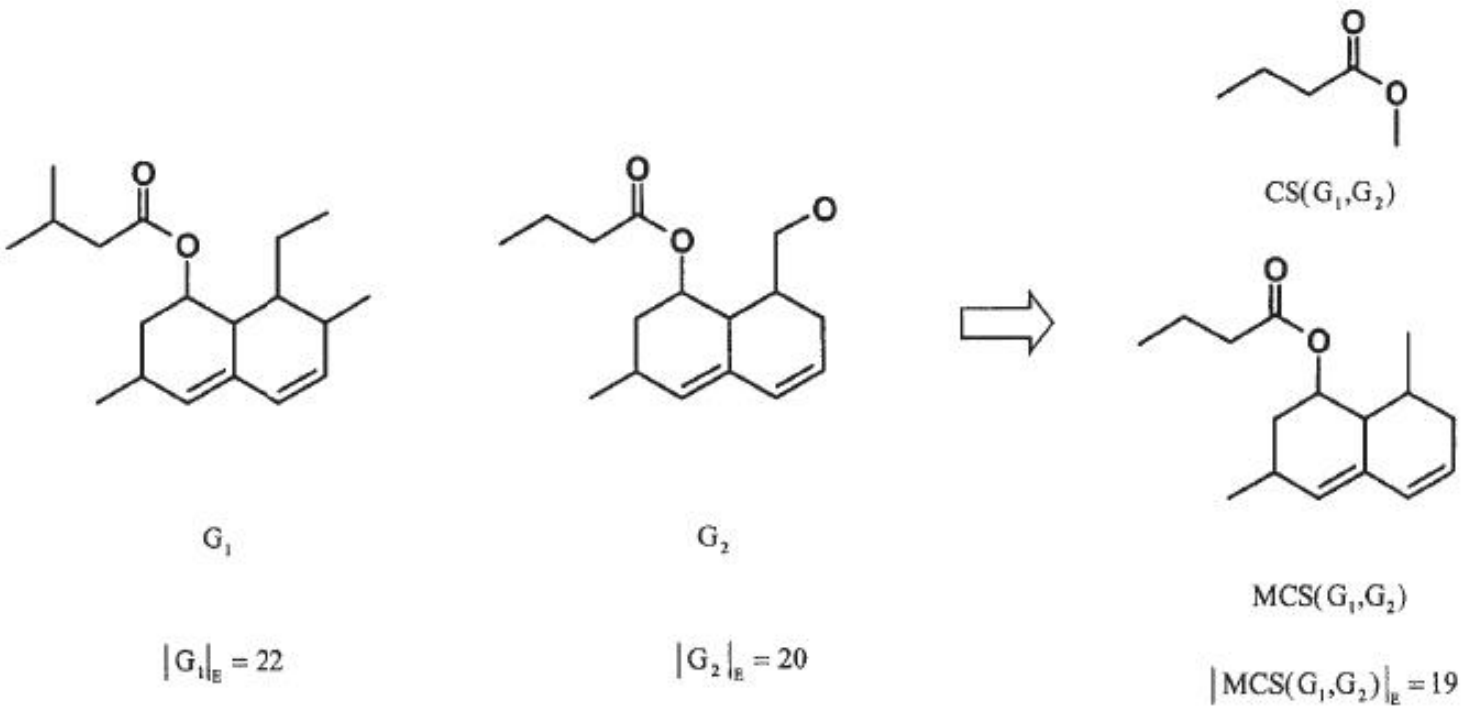
$$\begin{aligned} V_k &= \{V_k(x_1), V_k(x_2), \ldots, V_k(x_n)\} \\ &= \{v_{k,1}, v_{k,2}, \ldots, v_{k,n}\} \end{aligned},$$

$$\mathrm{E}_k = \{e_{k,1}, e_{k,2}, \ldots, e_{k,m}\} \ ,$$

where each edge corresponds to an unordered pair of vertices, that is $e_{k,i} = \{v_{k,p}, v_{k,q}\}$ *, and* L_k *is a set of r symbols*

$$\mathrm{L}_k = \{\ell_{k,1}, \ell_{k,2}, \ldots, \ell_{k,r}\}$$

that label each vertex (“atom”) and/or edge (“bond”). Typical atom labels include hydrogen (“H”), carbon (“C”), nitrogen (“N”), and oxygen (“O”); typical bond labels include single (“s”), double (“d”), triple (“t”), and aromatic (“ar”), but other possibilities exist. Whatever symbol set is chosen will depend to some degree on the nature of the problem being addressed. In most chemoinformatics applications *hydrogen-suppressed chemical graphs, which are* obtained by deleting all of the hydrogen atoms, are used. **Figure 1 depicts an** example of two hydrogen-suppressed chemical graphs, G_1 and G_2, which are clearly related to a chemist’s 2-D representation of a molecule.



$$S_{Tan}(G_i,G_j)=\frac{|G_i\cap G_j|_E}{|G_i\cup G_j|_E}=\frac{|MCS(G_i,G_j)|_E}{|G_i|_E+|G_j|_E-|MCS(G_i,G_j)|_E}=\frac{19}{22+20-19}=0.83$$

$$d(G_i,G_j)=|G_i|_E+|G_j|_E-2|MCS(G_i,G_j)|_E=22+20-2(19)=4$$

Fig. 1. An example of two hydrogen-suppressed graphs G_1, G_2 and a common substructure $CS(G_1,G_2)$ and the maximum common substructure $MCS(G_1,G_2)$ are shown above. The Tanimoto similarity index and the distance between the two chemical graphs are computed below.

$$G'_k \subseteq G_k \Rightarrow V'_k \subseteq V_k \text{ and } E'_k \subseteq E_k \quad ,$$

that is, the vertex and edge sets V'_k *and* E'_k *associated with the subgraph,* G'_k*,* are subsets of the corresponding vertex and edge sets V_k *and* E_k *of the graph,* G_k*. Many operations defined on sets can also be defined on graphs. One such* operation is the norm or cardinality of a graph,

$$|G_k| = |V_k| + |E_k|$$

which is a measure of the "size" of the graph. Another measure the *edge norm,* which is of interest in this work, is given by

$$|G_k|_E = |E_k| \ ,$$

where the subscript E explicitly denotes that the cardinality refers only to the edges ("bonds") of the graph. For the two chemical graphs depicted in **Fig. 1,** $|G_1|_E = 22$ and $|G_2|_E = 20$. Note that only the number of bonds and not their multiplicities (e.g., single, double) are considered here. However, many other possibilities exist, and their use will depend on the problem being addressed **.**

A key concept in the assessment of molecular similarity based on chemical graphs is that of a *maximum common substructure, MCS(G_i,G_j), of two chem-*

ical graphs, which derives from the concept of maximum common subgraph employed in mathematical graph theory. There are several possible forms of MCS . Here we will focus on what is usually called the maximum common edge substructure, which is closest to what chemists perceive as “chemically meaningful” substructures, but we will retain the simpler and more common nomenclature MCS. A common (edge) substructure (CS) of two chemical graphs is given by

$$\mathrm{CS}(\mathrm{G}_i,\mathrm{G}_j)_{k,\ell} = \mathrm{E}_i^k \cap \mathrm{E}_j^\ell = \mathrm{E}_i^k = \mathrm{E}_j^\ell$$

Where E_i^k and E_i^l *are subsets of their respective edge sets ,* $\mathrm{E}_i^k \subseteq \mathrm{E}_i$ *and* $\mathrm{E}_j^l \subseteq \mathrm{E}_j$, and are equivalent. Thus, the intersection (or union) of these two equivalent subsets is equal to the sets themselves. As there are numerous such common substructures, $\mathrm{CS}(G_i,G_j)_{k,l}$, *k,l = 1,2,3, . . . , determining the MCS between two* chemical graphs is equivalent to determining the edge intersection-set of maximum cardinality, that is

$$\mathrm{MCS}(\mathrm{G}_i,\mathrm{G}_j) = \mathrm{CS}(\mathrm{G}_i,\mathrm{G}_j)_{p,q} \text{ such that } \left|\mathrm{CS}(\mathrm{G}_i,\mathrm{G}_j)_{p,q}\right|_{\mathrm{E}} = \max_{k,\ell}\left|\mathrm{CS}(\mathrm{G}_i,\mathrm{G}_j)_{k,\ell}\right|_{\mathrm{E}}$$

Thus,

$$\mathrm{G}_i \cap \mathrm{G}_j \equiv \mathrm{MCS}(\mathrm{G}_i,\mathrm{G}_j) \ ,$$

Asymmetric similarity indices developed by Tversky

$$S_{Tve}(G_Q,G_T) = \frac{\left|G_Q \cap G_T\right|_E}{\left|G_Q\right|_E} = \frac{\left|MCS(G_Q,G_T)\right|_E}{\left|G_Q\right|_E} \quad ,$$

Two complementary compound sources are accessible for virtual screening, databases of known structures and de novo designs (including enumerated combinatorial libraries). Some major databases frequently employed for virtual screening experiments are listed in **Table 1**. In addition, several companies offer large libraries of both combinatorial and historical collections on a commercial basis. Usually the combinatorial collections contain 100k–500k structures, whereas commercially available historical collections rarely exceed 100k compounds. Most of the major pharmaceutical companies have compound collection in the 300k+ range.

Table 1.

Some major databases that are useful for virtual screening experiments (adapted from Eglen et al[33])

Database	No. of molecules	Description
ACD[a]	> 250,000	Available Chemicals Directory; catalogue of commercially available specialty and bulk chemicals from over 225 international suppliers
Beilstein[b]	> 7,000,000	Covers organic chemistry from 1779
CSD[c]	> 200,000	Cambridge Structural Database; experimentally determined three-dimensional structures of small molecules
CMC[a]	> 7,000	Comprehensive Medicinal Chemistry database; structures and activities of drugs having generic names (on the market)
MDDR[a]	> 85,000	MACCS-II (MDL) Drug Data Report; structures and activity data of compounds in the early stages of drug development
MedChem[d]	> 35,000	Medicinal Chemistry database; pharmaceutical compounds
SPRESI[d]	> 3,400,000	Substances and bibliographic data abstracted from the world's chemical literature
WDI[e]	> 50,000	World Drug Index; pharmaceutical compounds from all stages of development

[a]Molecular Design Limited, San Leandro, CA, U.S.A.

[b]Beilstein Informationssysteme GmbH, Frankfurt, Germany

[c]CSD Systems, Cambridge, UK.

[d]Daylight Chemical Information Systems Inc., Claremont, CA, U.S.A.

[e]Derwent Information, London, U.K.

Combinatorial libraries usually provide small amounts of uncharacterized compounds for screening. Once these samples are fully characterized—e.g., by HPLC and mass spectroscopy, the data are of interest for structure-activity purposes. In most companies, these compounds are also present with the “historical” collection of compounds, generally derived from classical medicinal chemistry programs, most of which have very well-defined chemical characteristics. Commercial compound collections can also be purchased that fall between these two extremes. Collectively, therefore, the information used to relate biological activity and chemical structure must clearly integrate all of these types of compounds.

Similarity Searching

Chemical similarity searching is a straightforward practical approach to identify candidate molecules by pair-wise comparison of compounds. In its simplest form, the result of a similarity search in a compound database is a ranked list, where high-ranking structures are considered to be more similar to the query in a certain sense than low-ranking molecules. If either the query structure(s) or the database structures or both structures reveal a certain (desired or undesired) property or activity, some conclusions may be drawn for the molecules under investigation. Structures are compared based on a similarity value that is calculated from their molecular descriptors. There are two assumptions inherent to this idea, representing the *hypothesis* “if molecule A is more similar to the query molecule *R* than molecule *B*, then molecule *A* might *more likely* show some biological activity that is comparable to the activity of *R*”:

- The molecular representation (descriptor) is assumed to appropriately cover those molecular attributes which are relevant for the underlying SAR/SPR /***Specific absorption rate, S*****ociety for Psychophysiological Research/**

- The similarity measure applied is assumed to accurately relate differences in molecular descriptions to differences in the quality function (*Principle of Strong Causality*).

In the past, the analysis of assay data was primarily performed by medicinal chemists, looking at the active compounds and then deciding which hits the efforts should be focused on. First, with the increase in the number of experimentally determined hits, this approach becomes increasingly ineffective and computational techniques are increasingly used to classify the hits and derive hypotheses. Second, one should keep in mind that it is basically impossible for a human being also to take into account the large number of inactive compounds. The development of pharmacophore hypothesis, for example, typically requires the incorporation of information on inactive compounds.

By similarity searching, sets of candidate structures can be rapidly compiled from databases or virtual chemical libraries. Practical experience shows that such hypotheses are often weak and there clearly is no cure-all recipe or generally valid hypothesis leading to success in chemical similarity searching. Nevertheless, similarity searching provides a useful concept. A practicable measure of success can be expressed by an enrichment factor, ef, giving the ratio of the fraction of active molecules in the selected subset compared to the fraction of actives in the total pool (database). This value may be regarded as an estimate of the enrichment obtained compared to a random selection of molecules, as given by Equation.

$$ef = \frac{fractionofactiveinsubset}{fractionofactiveinpool}$$

A large number of molecular descriptors has been developed over the past decades (**Definition**). The particular selection of a molecular representation defines a chemical space, and thus the ordering of molecules within this space. The choice of descriptors influences the distribution of structures.

The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." (according to Todeschini and Consonni)

Thank you!!!