Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 1991, **1(4)**, 122–124

Mendeleev Communications

Molecular Screening of High-Energy Nitrocompounds

Tatyana S. Pivina,[*a] Vladimir A. Shlyapochnikov,[a] Marina S. Molchanova,[a] Gennadiy Kh. Agranov[b] and Vladimir L. Rukin[b]

[a] *N. D. Zelinsky Institute of Organic Chemistry, Academy of Sciences of the USSR, 117913 Moscow, USSR. Fax: 095 135 5328*
[b] *Lensoviet Technology Institute, 198013 Leningrad, USSR*

An effective method for the molecular screening of compounds within a particular range of formation enthalpies has been developed; this approach is identical to the classification problem and is handled by the method of logical correlations, the inclusion of a substance in a given range of enthalpies being determined by the presence of informative local features in the structural topology of the compound; the high predictive ability of the method is demonstrated by the screening of high-energy nitrocompounds.

The design of molecular systems with predetermined physical and chemical properties is an important and complex problem in chemistry. An efficient solution depends to a large extent on the knowledge of 'structure–property' cause-and-effect relationships. At present, these relationships are still formalized rather poorly. One of the methods which allows the most adequate formalization of empirical chemical knowledge and subsequent design of substances with the desired properties is symbolic mathematics, *i.e.* the logical-structural approach[1] of pattern recognition theory,[2] applied to molecular graphs.[3]

Application of the logical-structural approach is connected with two notions: an informative molecular descriptor and a logical correlation. An informative molecular descriptor is a feature of the compound's structural topology which has a significant correlation with the given property. A logical correlation is a set of rules determining which substructures must be present or absent *in the structural formula of a compound in* order for the value of the required property to lie within a certain range.

Our approach can be briefly described as follows. A given training set is divided into two classes of compounds in accordance with the value of the specified property. For compounds of the '+' class, this value must lie within the desired range, for those of the '−' class, this is not true. All the compounds from the training set (from both class '+' and class '−') are then decomposed into simple structural fragments (such as, for example, atoms in a certain valence state, functional groups, aromatic and heteroaromatic rings, or substructures of other types). All the various simple fragments, obtained from the compounds after this decomposition (we shall call them 'fragments of dimension 1') are united in a single set. On the basis of this set, a logical matrix $\boldsymbol{T}$ is constructed, with rows corresponding to the compounds from the training set and columns corresponding to the obtained fragments. The values of its elements indicate the presence ($\boldsymbol{T}_{pq} = 1$) or absence ($\boldsymbol{T}_{pq} = 0$) of the q-th fragment of dimension 1 in the structure of the p-th compound from the training set.

On the basis of the constructed matrix, the computer carries out a search for those fragments of dimension 1 whose presence or absence in the structure of the given compound correlates closest with the presence of this compound in the '+' or '−' class (such fragments may be referred to as informative molecular descriptors). When such descriptors are found, we can formulate the simplest rules ('precursors' of more complex rules) which must be fulfilled for compounds of class '+' and not fulfilled for elements of class '−'. These rules may be presented in the form $Q_k^{(i)}\Pi_k^{(i)}$, where k is the number of the fragment which is considered in the rule; the index i corresponds to the dimension (at the described stage, $i = 1$); $\Pi_k^{(i)}$ is the predicate for the presence of fragment k in the given structure, $k = 1$ if the fragment is present in the structure and $k = 0$ if it is not; $Q_k^{(i)}$ is either a 'blank' (if the rule requires the presence of the k-th fragment in the structure in order to consider the compound as belonging to the '+' class) or a logical negation (the absence of the fragment is required).

However, *in the overwhelming majority of cases*, only a small number of compounds from the training set can be correctly classified using a single descriptor. Therefore, the computer forms specific 'subrules'. Within their framework, the presence or absence of several fragments is considered simultaneously (the number of descriptors included into a subrule can be varied within a certain range, specified by the user of the program). This is shown by eqn. (1),

$$P_j^{(i)} = Q_{j_1}^{(i)}\Pi_{j_1}^{(i)}Q_{j_2}^{(i)}\Pi_{j_2}^{(i)} \ldots Q_{j_m}^{(i)}\Pi_{j_m}^{(i)} \ldots Q_{j_n}^{(i)}\Pi_{j_n}^{(i)} \qquad (1)$$

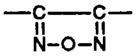
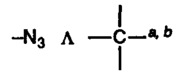
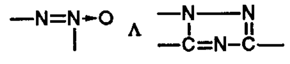
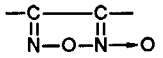
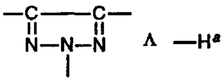
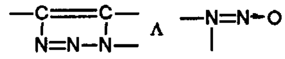
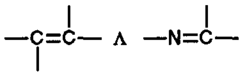
where j is the number of the subrule, the index (i) represents the highest possible dimension of fragments in the given subrule (at the stage we are considering now, $i = 1$); $j_1, j_2, \ldots, j_m, \ldots, j_n$ are the numbers of the fragments used as descriptors in this subrule.

Examples of the subrules found, which include fragments of dimension 1, can be seen in Nos. 1–7 of Table 1. For instance, subrule 7 states that a compound belongs to a class of substances with high enthalpies of formation (the '+' class) if its structure contains a 1,2,3-triazole ring and an azoxy group (*i.e.* $Q_{2_1}^{(1)}$ and $Q_{2_2}^{(1)}$ are 'blanks'); subrule 2 requires the presence of an azido group (*i.e* $Q_{2_1}^{(1)}$ is a 'blank') and the absence of alphatic carbon atoms (*i.e.* $Q_{2_2}^{(1)}$ is a logical 'not', a logical negation). The obtained subrules, which include fragments of dimension 1, are united into the so-called partial rule with the help of a logical 'or' (disjunction) [eqn. (2)],

$$R^{(i)} = P_1^{(i)}\Lambda\, P_2^{(i)}\Lambda \ldots \Lambda\, P_j^{(i)}\Lambda \ldots \Lambda\, P_l^{(i)} \qquad (2)$$

where L is a logical 'or', l is the number of subrules P_j in the partial rule. Yet, as a rule, only a small number of molecules from the training set can be correctly classified with the help of the obtained partial rule $R^{(1)}$. This is due to the fact that the fragments included in this rule can still be present (or absent) both in some compounds from the '+' class and in some compounds from the '−' class.

Table 1 Subrules of the decisive rule determing the inclusion of a compound into a class of substances with ($\Delta_f H$ > 100 kcal mol^{-1})

No.	Type of rule
1	
2	
3	
4	
5	$-N{=}N-\ \Lambda\ -H$[a]
6	
7	
8	

[a] A logical 'not'. [b] An aliphatic carbon atom, *i.e.* not included in a macrofragment such as a furazan or a furoxan ring, or other aromatic cycles.

Then, for molecules which remain *non-classified after the* partial rule $R^{(1)}$ is applied (and only for these molecules), the computer generates fragments of dimension 2. These fragments may be considered as all the possible pairwise combinations of those fragments of dimension 1 which are present in the non-classified molecules (for example, the ethylene fragment in the 8th line of Table 1 is a combination of two ethylene-type carbon atoms, and, since the atoms are fragments of dimension 1, the dimension of the ethylene fragment is 2). On the basis of the newly generated fragments, the logical matrix is extended. It becomes possible to formulate new subrules for the compounds which were not classified at the previous stage. These subrules $P_j^{(2)}$may include descriptors both of dimension 1 and of dimension 2. From these subrules, the partial rule $R^{(2)}$ is formed with the help of a logical 'or'.

If the training set contains compounds which cannot be classified with the help of the partial rules $R^{(1)}$ and $R^{(2)}$, then various fragments of dimension 3 are generated for these compounds (by pairwise combination of fragments with dimension 2 and dimension 1). In the way described above, the subrules $P_j^{(3)}$ and the partial rule $R^{(3)}$ for the fragments of dimension less than or equal to 3 are formulated, and so on.

Thus, the decisive rule R is a logical sum of partial rules $R^{(i)}$ [eqn. (3)],

$$R = R^{(1)}\Lambda\, R^{(2)}\Lambda \ldots \Lambda\, R^{(3)}\Lambda \ldots \Lambda\, R^{(k)} \qquad (3)$$

where k is the dimension of the fragments which ensures that the number of non-classified compounds in the training set is decreased to a certain given limit. From the formal-logical point of view, all the rules obtained in this way are logical correlations.

It should be noted that our algorithm makes it possible to carry out the necessary search among fragments of smaller dimension, while their dimension is increased only for those molecular graphs which have not yet been classified. Thus, we obtain the solution best protected from artefacts and randomness.

In order to evaluate the predictive ability of the logical correlations found at different stages of the process, the obtained rules are tested on compounds from a given control set.

This approach was realized as a set of computer programs.[5] In order to demonstrate its work, we shall use computer screening of high-energy nitrocompounds as an example.

The search for 'structure–enthalpy of formation' correlations was based on a training set composed of nitrocompounds of different chemical classes: nitro- and polynitroalkanes, nitroamines (cyclic and acyclic), aromatic and heteroaromatic compounds and others. The whole data base (more than 400 compounds) was divided into two sets: class '$-$' (310 compounds with enthalpies < 100 kcal mol^{-1}) and class '$+$' (95 compounds with $\Delta_f H^0 > 100$ kcal mol^{-1}).† The decisive rule R (a logical correlation) was used to classify molecular structures from the test set (it consisted of 28 compounds from class '$+$' and 28 compounds from class '$-$'. The predictive ability, calculated as the ratio of the number of correctly classified structures to the total number of structures in the test set, was 96.43%. Table 1 below shows the subrules which form the partial rules and the decisive rule of 'structure–property ($\Delta_f H^0$)' logical correlations for the classification of test compounds.

Comparison of the fragments on which the decisive rules are based leads us to the conclusion that there exists a structural conditionality of such a property as the enthalpy of formation. Thus, with the help of the logical-structural approach, it became possible to reveal the key fragments and their relationship and to create a non-contradictory basis of rules. Using this basis, it is possible to predict high-energy substances and to organize a constructive search based on computer design of high-enthalpy compounds. There is every reason to believe that this approach will also be effective for searching 'structure–property' correlations in the case of other physical and chemical characteristics and of substances for various applications.

Received in USSR, 25th February 1991
Received in UK, 12th June 1991; Com. 1/01162F

References

1 V. E. Golender and A. B. Rosenblit, *Logiko-combinatornye metody v konstruirovanii lekarstv* (Logical and Combinatorial Methods in Drug Design), Zinatne, Riga, USSR, 1983 (in Russian).

2 J. Tu and R. Gonzales, *Printsipy rasposnavaniya obrazov* (The Principles of Pattern Recognition), Mir, Moscow, 1978 (in Russian).

3 D. Smith, C. Reese and J. Stewart, *Iskusstvennyi intellekt: primenenie v khimii* (Artificial Intelligence: Application to Chemistry), eds. T. Pierce and B. Honey, Mir, Moscow, 1988 (translated into Russian).

4 T. A. Pivina, G. H. Agranov, V. L. Rukin and V. A. Shlyapochnikov, *Izv. Akad. Nauk SSSR, Ser. Khim.*, 1990, **6**, 1349.

5 G. H. Agranov, V. V. Sotnikov, V. L. Rukin and S. V. Fedorov, *Program Complex for the Search of Structural Laws in Molecular Systems*, Proceedings of the Conference 'Molecular Graphs in Chemical Studies', Odessa, USSR, 1987.

† 1 cal = 4.184 J.