Translated from *Uspekhi Khimii*, 402-421 (1983) U.D.C. 542.973

Prediction and Analysis of Heterogeneous Catalysis Mechanisms by Pattern Recognition Methods with a Computer

I.I.Ioffe, A.M.Dobrotvorskii, and V.A.Belozerskikh

The basic aspects of the prediction of catalytic activity in one-component and multi-component heterogeneous catalysts and of the analysis of the mechanism of heterogeneous catalytic reactions by using pattern recognition theory with a computer are discussed. The present state of development is reviewed, and examples of the use of this method in choosing catalysts for different types of reaction are given. 53 references.

CONTENTS

I. Introduction 228
II. Fundamentals of pattern recognition methods 229
III. Prediction of one-component catalysts or promoters 231
IV. Prediction of multi-component catalysts 232
V. Analysis of catalytic action mechanisms (the inverse recognition problem) 235
VI. Prospects of future developments 237

I. INTRODUCTION

Catalyst prediction should be clearly distinguished from the search for entirely new types of catalyst, or for catalysts for new and unexplored classes of reactions†. Catalyst prediction denotes the *a priori* search for suitable catalysts for known types of reaction on the basis of existing information on a range of potential catalysts and of a general understanding of the mechanism of their action. As well as having the right chemical composition, an effective industrial catalyst should have the required permeability (or porosity) parameters, including strength and geometric size. In this review we shall be concerned exclusively with the problem of predicting the chemical composition of the catalyst, though the pattern recognition algorithms are equally applicable to the prediction of the physical parameters of the catalysts.

By using the existing level of scientific understanding of catalytic reactions, and the vast fund of empirical information (extending over more than a century), we can often select suitable catalysts for various reactions, including new reactions, if they fall within one of the main classes of chemical transformation (e.g. hydrogenation, oxidation, condensation, etc.). However, this type of prediction offers no information on the productivity or selectivity of the catalyst, parameters of the utmost importance in industrial applications. We should note that in large-scale industrial production even a small improvement in performance could be very important economically.

Modern ideas on catalysis, such as the relationship between catalytic action and the acid–base properties of the surface[1,2], the energetics of the interatomic bonds of the lattice[3,4], the stabilisation energy of the crystal field[5], the geometric parameters of the lattice, etc. can be used to construct an acceptably rational scheme for the selection of catalysts for stated reactions[6–8]. However, in spite of its attractiveness and of its usefulness in research this scheme cannot predict quantitatively the main practical characteristics (activity, selectivity, etc.) of the proposed catalysts.

† Throughout this review the term catalyst will be used to denote heterogeneous catalysts.

In the present state of knowledge we can distinguish three possible paths for catalyst prediction: 1) the theoretical approach, based on the theory of the structure of chemical compounds and on quantum chemistry, 2) the semi-empirical approach, based on correlation methods, and 3) the mathematical (heuristic) approach, which relies on the methods and algorithms of the mathematical pattern recognition theory.

In spite of the important advances recently made in the quantum chemistry of heterogeneous catalysis[9–11] the first approach is still far from the state of development which would justify quantitative predictions of the output parameters of the catalysts. Nevertheless, as we shall show below, theoretical (and especially quantum-chemical) considerations play an important role in computer methods of catalyst prediction.

The second approach is equivalent, in practice, to using one of the many variants of the principle of linear free energy relationships (LFER), such as Balandin's principle of energy correspondence, or the Brønsted, Hammett–Taft, or Voevodskii–Semenov equation. Numerical calculations by these equations are possible only for empirical coefficients, and are only valid for catalysts of a single type used in one particular reaction of (in the case of a group of homologues) for one catalyst in one reaction. These methods are unsuitable for predicting multi-component catalysts and they do not reveal extrema in the quantities being correlated (except the Balandin method, which is not accurate enough, however). Nevertheless investigations in this direction are continuing. Empirical correlation equations have been obtained for the hydrogenolysis of ethane and of alcohols[12,13] and for the dehydrochlorination of chloro-ethanes on sulphates[14]. These problems are discussed in depth in Ref.9.

Unlike the LFER methods, the pattern recognition methods do not need to be confined to catalyst (or reaction) analogues: though the mathematical recognition methods when applied to structure–property problems are effectively methods of detecting multi-dimensional or multi-level correlations they differ from the classical correlation analysis methods in their ability to reveal the inter-relationship of ten or more properties used as indicators of the quality of the samples by taking the combination of criteria into account. This feature of the pattern recognition methods has already led

to their successful application in chemical problems such as the determination of molecular structure from spectral data, the *a priori* calculation of the properties of inorganic compounds, and the prediction of biological activity in organic compounds[15–17].

The most important aspect of the pattern recognition methods as applied to catalysis problems is their ability to exploit the very large fund of experimental information on catalysts in standard handbooks[18] or in original publications as a predictive basis. The published data differ widely in precision and reliability: they include misleading information introduced for patent reasons, and they are only valid in statistical, computer-aided information processing methods such as the pattern recognition methods.

The use of these methods for the prediction of catalyst behaviour is made easier by the level of reliability required, which can be very high in some applications but not in catalyst prediction. Thus in radar, medicine, and geological prospecting the probability of the prediction must be 99.5% or more. In the case of catalyst selection an economically effective forecast is achieved simply by short-circuiting a laborious empirical search, and increasing the probability of finding the optimum catalyst. Since in any case the prognosis must be tested experimentally, there is no danger of making a "catastrophic" mistake. As compared to catalyst selection by the conventional semi-empirical methods, a quantitative prediction with a probability of ~80–85% would be economically effective.

The potential value of pattern recognition methods in the search for optimum catalysts is confirmed by calculations of their effectiveness[19,20]. In recent years pattern recognition methods and related methods of mathematical statistics have been used by many workers. Thus in Refs.21–24, discussed below in detail, pattern recognition algorithms are used to identify catalysts for the hydrogenation and synthesis of thiophen. The factor analysis method has been used[25,26] to predict catalysts for electrochemical processes. Some results obtained by the present writers since the publication of their monograph[9] and only partially reported[27–29] will also be discussed in this review.

Essentially all recognition methods consist in establishing step-by-step relationships between the parameters of the solid responsible for its catalytic properties and its other physicochemical properties. These parameters include (depending on technological and cost considerations) either the productivity of the catalyst in terms of the desired product (i.e. its activity) or the selectivity of the reaction (or process) when catalysed by this material. The latter quantity is defined as the yield of desired product expressed as a percentage of the theoretically possible yield (or of the potential yield in petroleum chemistry), and it corresponds to the selectivity of the anion of the catalyst itself. In special cases other selection criteria are possible; for example, the operational life of the catalyst, its thermal stability, etc.

Since the recognition algorithms allow the ranking of the properties with respect to their effect on the useful parameters of the catalyst we can also solve the inverse recognition problem, i.e. identify the most "influential" physicochemical properties of the catalysts and reactants from the point of view of ensuring a fast and selective reaction. The results obtained give a basis for proposing the mechanism of the catalytic action. The main value of this method of analysing the mechanism of catalysis lies (as in the case of catalyst prediction) in the ability to exploit the extensive but not always very accurate experimental information, and to obtain reliable results by virtue of the statistical treatment implicit in the pattern recognition algorithms.

II. FUNDAMENTALS OF PATTERN RECOGNITION METHODS

Pattern recognition theory is a branch of cybernetics. The mathematical foundations of the methods used in pattern recognition are fully discussed in textbooks[30–32]. In this Section we shall merely give a short description of the methods which have been used in chemistry to predict the properties of substances.

The solution of the recognition problem is based on empirical input data, called the "training sample". In our case the "training" consists in "tuning" the specially developed algorithms in a way which enables the computer to determine the property of interest to the experimenter from the structure of the materials. The term structure denotes the set of primary properties of the material which can be used as criteria. The training process includes separating the materials into classes according to the value of the property being predicted. The assignment to one of these classes of a material not included in the training sample is the solution of the pattern recognition problem.

Existing recognition algorithms are either "deterministic" or "statistical". The former include the algorithms based on geometric considerations, and the perceptron algorithms, developed as a model of the perception of patterns by human beings[33]. The latter class includes the algorithms of the stochastic approximation and those based on the principles of statistical decision making. A special group, only recently recognised, consists of algorithms which have been called syntactic[34]. So far only algorithms from the deterministic group have been used in chemical problems. The desirability of using other types of algorithm to solve these problems will be considered below.

Let us examine more closely the geometrical treatment of the recognition problem. In this treatment the set of criteria of the objects to be recognised is treated as a hyper-space with one coordinate axis for each criterion. In such a space each object is represented by a point whose coordinates give the numerical values of the criteria. Coordinates corresponding to qualitative criteria can also be used: this is done by numbering the possible realisations of the qualitative criterion and equating the numbers to the values of the coordinate on the axis for that criterion. Since the limits and the ranges of the characteristics used as criteria are always known it is convenient to adopt a system of dimensionless units such that all possible objects lie within a unit cube of the criterion space.

In property forecasting it is a natural assumption (and one consistent with our chemical intuition) that substances whose Euclidean distance apart in the criterion space is small will have similar properties. This hypothesis underlies the application to the chemical problem of algorithms based on geometrical considerations. The geometric treatment includes algorithms which have been used successfully in catalyst prediction problems (potential functions, and binary or segmented linear classifiers).

The simplest of these algorithms is the binary linear classifier. Its action consists in finding a hyper-plane which divides the objects into two classes: the "dichotomy" problems. It is convenient to make this plane pass through the origin of the coordinate system. This can be done by introducing an additional $(N + 1)$-th coordinate (where N is the number of criteria of interest) whose value is assumed to be the same for all the objects (Fig.1). In this space any dividing plane is uniquely defined by the unit vector of the normal **W**, called the weight vector. The vector **W** is found by an iterative process, during which the positions of the objects in the training sample relative to an arbitrarily chosen initial plane are tested sequentially. If an object

(radius vector $\mathbf{R}_k$) is found to be incorrectly classified, the position of the plane is correlated accordingly, and its new vector is expressed as the linear combination

$$\mathbf{W}' = c_1\mathbf{W} + c_2\mathbf{R}_k$$

The coefficients c_1 and c_2 are determined from the normalising condition of the weight vector and from the expression

$$\mathbf{W}' \cdot \mathbf{R}_k = -\mathbf{W} \cdot \mathbf{R}_k$$

If this condition is satisfied the kth object is on the correct side of the new plane and at the same distance from it as from the initial plane. It can be shown[35] that if a dividing plane exists, i.e. if the classes are linearly separable, this iterative process always converges.

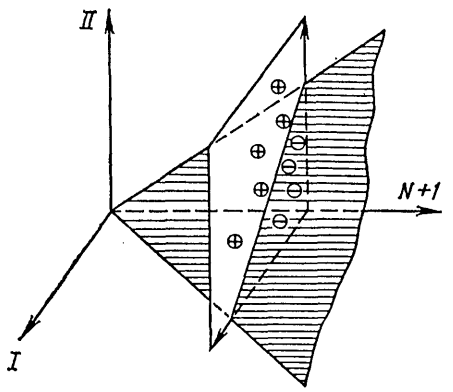


Figure 1. Dividing plane in criterion space, with an additional coordinate.

Unfortunately, linear separability into classes is seldom found. The boundaries of the classes usually have a more complex structure, and a simple linear classifier is inadequate. This weakness of the algorithm can be partly overcome by replacing the dividing plane by a plate of finite thickness and treating the objects lying within its thickness as unclassifiable. The thickness of the dividing plate is called the threshold, and the algorithm is called a linear classifier with non-zero threshold[15].

Even more flexible is the segmented linear classifier: a generalised form of the linear classifier in which the dividing surface is formed by a set of planes. In this algorithm a correction is applied to the position of the planes when the object is incorrectly classified with respect to the majority of the planes[35].

Linear classifiers have been widely used, and in spite of their simplicity they have been found to be effective in a wide range of recognition problems (including chemical ones)[15]. Many more complex binary classifiers, in which the dividing surfaces are not linear, can be reduced[35] to algorithms of this type. With any of the above algorithms the objects can be classified into three or more groups by sequential dichotomies, i.e. by repeating their application to the already classified populations.

Another geometric approach used in pattern recognition theory is the potential functions method, which relies on the concept of a potential for the kth object in the criterion space, $u(|\mathbf{R}-\mathbf{R}_k|)$, and a potential for the α class of objects,

$$U_\alpha(\mathbf{R}) = \frac{1}{N_\alpha} \sum_{k \equiv \alpha} u(|\mathbf{R} - \mathbf{R}_k|)$$

where N_α is the number of objects in the class. The boundary between the two classes (α and β) is taken to be the surface on which the potentials are equal:

$$U_\alpha(\mathbf{R}) = U_\beta(\mathbf{R})$$

It is convenient to choose for $u(x)$ a continuous, monotonically decreasing function having a finitie value at $x = 0$, for example

$$u(R) = 1\ (a + bR^2),\ a > 0$$

As in the case of binary classifiers, in the potential functions method objects are most easily classified into a large number of groups by successive dichotomies. At each step the assignment to one of two classes is determined by the sign of the difference between the potentials, $U_\alpha - U_\beta$. A more detailed discussion of the application of the potential functions algorithm can be found in textbooks[9,36].

Without attempting to examine in depth the principles of the perceptron theory (see, for example, Refs.30 and 33) we can define the method as the classification of a set of objects in terms of the sets of values of the criteria prescribed for each class. In catalysts prediction and similar problems the algorithms and programs based on logic functions[37] are more useful than the classical perceptron. The advantage of this approach lies in the possibility of defining classes forming an interpenetrating system of the "water in a sponge" type. The principle has been used for catalyst prediction in the "Prognoz-73" program[9], which relies on the following algorithm. Ranges of values are defined for all the properties of the catalysts. The number of criteria is assumed to be equal to the total number of ranges, and each criterion can take the values 1 or 0 depending on whether the property lies within or outside the range. Thus each catalyst is associated with a Boolean (logic) vector. We can now construct the discriminant conjunctions, i.e. the Boolean vectors whose termwise logic product with any vector from the class of effective catalysts is equal to the same vector and whose product with the vectors from any other class is equal to the zero vector. By using the discriminant conjunctions any catalyst can be assigned to one of the classes.

The recognition algorithms described in this Section are far from exhausting the enormous range of possible approaches to the solution of problems in this rapidly expanding field of applied mathematics. They do, however, give a clear illustration of one of the most important ways of "recognising": the assignement of an unknown object to a defined class. This includes preliminary training (in other words, tuning) of the algorithm by using an existing sample of already classified objects. As a rule, recognition algorithms are not "ideally" tunable. We must therefore introduce a new concept: the classifying power α_{cl} of the algorithm, defined as the ratio of the number of correctly classified objects to the total number of objects in the training sample#. The next step is to apply the trained algorithm to the unknown objects. The ratio of the number of objects which subsequent experience shows to have been correctly classified to the total number of objects investigated is called the predictive power of the algorithm α_{pr}.

The parameter α_{cl} is determined by carrying out a prediction for each object in the training sample with the recognition algorithm trained by using the remaining (N − 1) objects.

The coefficients α_{cl} and α_{pr} calculated in this way should be treated as arbitrary quantities, depending on the size and composition of the sample. Adequate estimates of the true classifying power and predictive power are obtained if sufficiently representative (or generally valid) samples are used. An important characteristic of the general validity of the sample is the ratio of its volume to the number of criteria used in the recognition, denoted by the symbol λ. It is usually assumed that successful prediction requires a fairly large value of the ratio λ, which we shall term the relative sample volume ($\lambda > 2$, say). However, there is evidence that satisfactory solutions of applied problems can be obtained also with much smaller values of λ.[38]

We should stress that problems requiring detailed statistical analysis of the pattern recognition results are only seldom included in current publications, and will not be considered in this review.

III. PREDICTION OF ONE-COMPONENT CATALYSTS OR PROMOTERS

The problem most simply solved by pattern recognition theory is the prediction of one-component catalysts. In this case the choice of a new catalyst (metal, oxide, etc.) can be made by comparing existing experimental data on the activity and selectivity of the proposed catalysts and on the physical and chemical characteristics (used as selection criteria) of all the substances involved. Promoters, or any other modifier or additive for the chosen catalyst, can be selected in a similar way.

Extensive information on the crystalline and electronic structure and on the physical and physicochemical properties of solids of stoichiometric composition is now available in standard handbooks. This allows us to choose as the selection criteria the most reliable data, determined by identical methods and under identical conditions, for the one-component catalysts. The most obvious choice of criteria includes the properties which are known (according to current theories of catalysis, or from experimental results) to be related to the ability of solids to accelerate chemical reactions. Some of these have already been discussed in Section II in relation to the use of simple correlation expressions for predicting catalyst behaviour. Some of the best known relationships between the catalytic activity and the physical characteristics of substances are discussed in Refs.6 and 18.

As well as experimental data, we can use as criteria the parameters of the electronic structure of solids, calculated by quantum mechanics or quantum chemistry methods, or various combinations of physical characteristics. Obviously, some of the properties of solids may be closely correlated, or even functionally related, if they have a common origin. The paramagnetism and the low-temperature specific heat of metals are typical examples; in the case of oxides the width of the forbidden energy gap is related to the colour and to other properties of the material. When choosing a system of criteria it is important to allow for these effects and to select characteristics which are as far as possible independent. The screening of superfluous (in this sense) properties can be made automatic by using the algorithms of mathematical statistics.

However, great care is needed in any preliminary statistical analysis of the system of criteria, because the exclusion of one or more members of a series of formally correlated properties may result in a serious lowering of the classifying and of the predictive power of the algorithm. This situation is illustrated in Fig.2, which shows a set of objects linearly separable in the space of two strongly correlated criteria (A and B). If one of these criteria is rejected the objects become unclassifiable by a simple binary classifier.

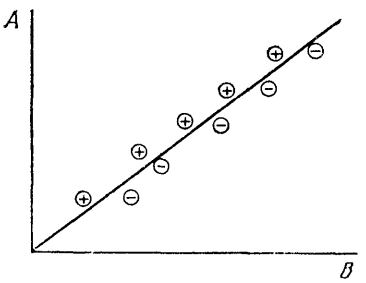


Figure 2. Examples of a linearly separable population in a space of correlated criteria A and B.

Even greater care is needed (for the same reason) in sifting the criteria on the grounds of weak correlation with the property of interest (catalytic activity, selectivity, etc.). It is intuitively obvious that some of the criteria may contribute only marginally to the catalytic activity towards certain reactions. These can be eliminated at the algorithm training stage, thus decreasing the number of dimensions of the recognition space. The "influence" of the criterion is a useful concept in this context. In the case of binary classifiers we can use as a measure of the influence the ratio

$$\beta_i = [\alpha_{cl} - \alpha_{cl}(i)]/\alpha_{cl}$$

where $\alpha_{cl}(i)$ is the classifying power of the algorithm in a space from which the ith criterion has been excluded. If the influence of the criterion is small, i.e. if its exclusion does not materially affect the classifying power, the criterion can legitimately be omitted.

The first reported use of recognition methods for catalyst prediction[39] dealt with the oxidation of CO on metal oxides, relying on the experimental information collected by Krylov[40]. The training sample consisted of 26 oxides: MoO_2, CoO, Co_3O_4, MnO, CdO, Ag_2O, NiO, CuO, SnO_2, Cu_2O, Co_2O_3, ZnO, TiO_2, Fe_2O_3, MgO, CeO_2, V_2O_5,, Cr_2O_3, ZrO_2, HgO, WO_3, ThO_2, BeO, GeO, Al_2O_3, and SiO_2. 20 physical properties of the catalysts, available in standard tabulations, were used as criteria. The catalysts were divided into three classes: good, moderately good, and poor, on the basis of their catalytic activity (listed in Ref.40). A success rate of 84% was achieved in the prediction for oxides taken out of the training sample by using successive dichotomies with a potential functions program[41]. By using methods to be discussed below it was shown that only 8 of the 20 criteria had any influence on the activity. Predictions of catalytic activity towards the oxidation of CO were made for a number of oxides not previously used in this reaction: in the case of HfO_2 the prediction was experimentally confirmed[42].

In the selection of a promoter for the catalytic formation of low-molecular aromatic hydrocarbons from alkanes of intermediate molecular weight (for example, dodecane) the training sample consisted of the oxides V_2O_5, MoO_3, CeO_2, Cr_2O_3, CaO, Bi_2O_3, and K_2O, which had been previously studied experimentally[29]. The objects for recognition included the oxides CoO, Mn_2O_3, CuO, Ag_2O, NiO, PdO, La_2O_3, ZrO_2, and WO_3.

Five "influential" properties were identified (Table 1) from the initial list of 12. Recognition was effected by the "Prognoz-73" program, using a two-category system. The validity of the prediction was tested experimentally, and the high activity which had been forecase for the oxides ZrO_2, Mn_2O_3, and WO_3 was confirmed.

Table 1. Physicochemical properties of oxide promoters[29].

Oxide	$Z(I+a)$	μ	r_M, (Å)	r_K, (Å)	n
V_2O_5	52.20	+128.00	0.91	0.80	0
MoO_3	51.72	+3.00	0.91	0.81	0
CeO_2	28.64	+26.00	0.94	0.99	0
K_2O	2.74	—29.00	1.39	1.08	0
CaO	9.60	—15.00	1.06	1.03	0
Bi_2O_3	23.88	—83.00	1.025	0.915	2
Cr_2O_3	28.59	+1 960.00	0.815	0.935	3
Ag_2O	8.23	—134.00	1.19	0.91	10
NiO	19.88	+660.00	0.875	0.95	6
CuO	20.00	+238.90	0.95	0.90	9
PbO	11.40	—42.00	1.11	0.935	2
Mn_2O_3	29.64	+14 100.00	0.81	0.945	4
La_2O_3	19.50	—78.00	1.04	0.99	0
ZrO_2	30.16	—13.80	0.88	0.94	0
WO_3	51.72	—15.80	0.91	0.82	0
CoO	18.58	+4 900.00	0.88	0.955	5

Symbols: Z is the valence of the metal, I is the ionisation potential (eV), a is the electron affinity (eV), μ is the magnetic susceptibility, r_M is the ionic radius of the metal, r_K is the ionic radius of the oxygen, n is the number of valence electrons not taking part in metal–oxygen bonds.

Table 2. Prediction of the promoting properties of elements added as oxides to vanadium–molybdenum catalysts[44].

Element	Δ_{exper}, %	Δ_{pred}, %	Element	Δ_{exper}, %	Δ_{pred}, %
Mg	35	32	Ce	—15	—11
Ti	5	11	Pr	10	6
Fe	—10	—8	Nd	5	10
Co	20	14	Gd	3	13
Ni	20	17	Tb	—3	2
Cu	—10	—8	Ho	10	14
Y	35	24	Er	20	10
Ag	—10	—7	Yb	35	13
Cd	3	4	Tl	—3	—4
La	3	13			

Symbols: experimental (Δ_{exper}) and predicted (Δ_{pred}) values of the change in the depth of oxidation of furfural caused by the addition of the promoter to the vanadium–molybdenum system.

A study[43] of the properties of the training sample of metallic promoters (consisting of metals from Groups VI and VII) has led to the prediction of a new, active promoter for nickel catalysts used in the hydrogenation of unsaturated compounds: niobium. The calculation was carried out with a potential functions program, using a three-category successive dichotomy scheme.

Promoters for multi-component vanadium–molybdenum catalysts used in the oxidation of furfural to maleic anhydride have been predicted[44]. The predictive experiments were organised as follows: out of 19 experimentally tested promoters, 17 were used to train the computer, and for the other 2 the selectivity of the reaction was predicted from the variation in the yield of side-products (resulting from full oxidation). An algorithm which forecasts the numerical value of the parameter of interest (in this case, the selectivity of the reaction) was used in this work. The algorithm, fully discussed in Ref. 44, consists in reproducing the shape of the function which describes the dependence of the selectivity on the physicochemical properties of the promoters. Effective criteria (and combinations of criteria) were chosen by an algorithm of the type used in Ref. 37. The results are shown in Table 2.

It is clear from Table 2 that the algorithm used in that work[44] had a high predictive power. Only four of the ten properties of the promoters arbitrarily selected were found to be effective: the distance between the metal atom and the oxygen atom, the electronegativity difference and the width of the energy gap in the oxide, and the colour of the oxide.

A factor analysis algorithm, based on the coefficients of the linear correlation between the output parameter Y (the logarithm of the exchange current) and a set of criteria chosen *a priori*, has been used[25] in the recognition and prediction of one-component metallic catalyst electrodes for electrolytic hydrogen evolution. The output parameter was defined by the expression

$$Y = \sum_{j=1}^{N} a_j \tilde{x}_j$$

where $\overline{x_j}$ are the average values of the normalised criterion, and a_j is a coefficient obtained by least squares. By using factor analysis for ranking the criteria[45] the initial set of seven criteria was reduced[25] to three: classifications based on these criteria (by the dichotomy method) agreed in 80% of the cases. The algorithm has also been used[26] to recalculate earlier results[39]: it gave a slightly worse agreement (80% instead of 84%), probably because the postulated linear dependence of the property of interest on each of the criteria is too rough an approximation. This work[25,26] shows that the factor analysis algorithms can be useful for catalyst recognition and prediction, though their power is lower than that of the "pure" recognition algorithms and their validity is more limited.

IV. PREDICTION OF MULTI-COMPONENT CATALYSTS

The most difficult (and yet, perhaps, the most important) problem of this type is the use of recognition methods in predicting multi-component catalysts at the component selection stage, i.e. when the true physicochemical properties of the working catalyst are still unknown. The only possible approach is to use the properties of the components used in the synthesis of the catalyst. Two paths are available: either to seek an effective method of calculating the physicochemical properties of multi-component solids, thus reducing the problem to the use of algorithms adequate for one-component catalysts, or to construct multi-level heuristic programs in which the recognition programs appear as sub-programs[46].

The simplest approach[21,22] to the prediction of multi-component catalysts relies exclusively on chemical composition data, using the following algorithm. The catalysts are classified into two opposite groups: one group catalyses the reaction for which a prediction is required, while the other group catalyses a reaction with the opposite mechanism. The identification (assigned to a class) problem is solved geometrically. The total set of initial data used as a training sample forms a matrix having $(M \cdot N)$ dimensions, where M is the number of objects and N is the number of criteria (in the

form of zeros and unities). In a row describing an object an entry of 1 denotes the presence in the object of a particular element. Standards are chosen for each class, and these act as "centres of gravity" for the class in N-dimensional criterion space. The object is considered to belong to one of the classes if its Euclidean distance from the standard for that class is less than the distance from the standard of the other class. The closer the unknown object is to the standard of the class of interest, and the further from the standard of the other class, the better are its prospects. A linear relationship between the activity of the catalyst and its geometric criterion (defined above) was demonstrated, as well as a high (>90%) probability of correct assessment of the activity or selectivity of catalysts.

An obvious disadvantage of this approach is that the search for possible new variants of catalysts is limited to elemental composition, since this is the only variable parameter in the training sample.

A more systematic approach to the prediction of multicomponent catalysts relies on a preliminary calculation of the physicochemical characteristics of the catalysts from data relating to the components. This type of problem is not uncommon in solid state chemistry, and it can be solved in a number of ways[47–49], all based on interpolation relationships in which the properties of the components are somehow averaged. Modern theories of the electronic structure of disordered solids also rely on this approach, though not always explicitly[50].

The use as criteria of quantities calculated by interpolation formulae offers two important advantages. One, already mentioned above, is the possibility of exploiting the extensive data available for one-component catalysts in predictions of multi-component catalysts. The other advantage is the very attractive possibility of predicting the properties of catalysts of complex composition from data on the catalytic properties of pure metals, oxides, etc.

The simplest way of averaging is with respect to mole fractions. In spite of its roughness this method is found experimentally to be useful for tentative predictions. It has been applied[29] in the prediction of the activity of mixed oxide catalysts for the aromatisation of the lower alkanes (the first step in the selection of a catalyst for the preparation of a "low aromatic" fraction from alkanes of moderately high molecular weight, described above). The correlations between the catalytic activity and the electronic and structural properties of the mixed oxide catalysts were established by the potential functions method. The catalytic activity was quantified (in terms of the selectivity s) from the total yield of aromatic compounds for the very large number of mixed dehydrocyclisation catalysts described in the literature, using the three-class system

class I ("good"): $67 < s < 100$
class II ("fairly good"): $33 < s < 67$
class III ("poor"): $0 < s < 33$

The effect of the energy of electron vacancies in the active centres and of the defect structure of the catalyst surface was calculated by the function

$$Q = \frac{\sum \nu_i [Z_i (I_1 + a)_i]}{1 + \bar{\mu}}$$

where ν_i is the mole fraction of the component in the mixture of oxides, Z_i is the valence of the metal in the ith oxide, and $(I_1 + a)$ is the sum of the first ionisation potential and of the electron affinity of the metal in the ith oxide. The numerator of the fraction is a measure of the acidity of the catalyst. The parameter $\bar{\mu}$ is the crystal-chemical similarity of the components of the oxide mixture, expressed on a point scale, which determines the defect structure of the surface. The averaged value of the magnetic susceptibility ($\bar{\chi}_\mu$) was used as the second criterion.

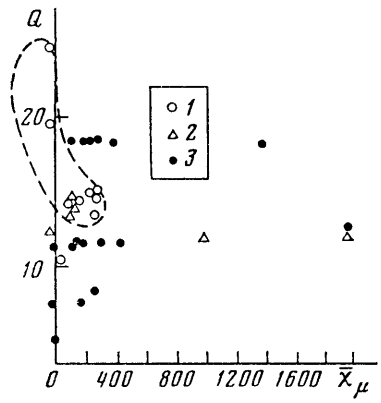


Figure 3. Distribution in criterion space of the catalysts for the aromatisation of the lower alkanes: *1*) good, *2*) average, *3*) poor catalysts.

It is clear from Fig.3 that in spite of the rough approximation of the initial assumptions the catalysts are well classified into three groups, even in two-dimensional space.

Other reactions have been studied on mixed oxide catalysts, including the hydrogenation of compounds with an ester group to alcohols and the gas-phase conversion of formaldehyde and isobutene into isoprene[27]. Owing to the lack of strict methods of calculating the characteristics of the electronic structure of mixed solid-phase systems there is no alternative to using an approximate additive scheme to obtain the properties of the mixed oxide catalyst: the property X is expressed as $X = \sum_i x_i \nu_i$, where x_i is the property for the pure component i and ν_i is its mole fraction. The following combinations of properties were identified as the most effective for catalysts used in these processes. For the hydrogenation of esters, $\sum_i (n_{d,f} + N_{d,f})_i \nu_i$ (where $n_{d,f}$ is the number of d and f valence electrons not taking part in bond formation with the oxygen, and $N_{d,f}$ is the number of fully occupied d and f shells in the metal atom) and $\sum_k J_k \nu_k$, an average value of the bond ionicity, and for the conversion of formaldehyde the average magnetic susceptibility $\sum \chi_i \nu_i$ and the parameter $\sum_k J_k$ (the sum of the bond ionicities in the active components).

Thus only two properties were found to be effective for each of these processes, and therefore the assignment to classes in property space can be displayed graphically (Fig.4). A separation is achieved, but we note some overlap of the classes (due to the lack of precision of the additive scheme used to calculate the properties).

Since many physical and chemical properties are not additive even in systems consisting of chemically similar components much care is needed in selecting the recognition criteria when averaging procedures are used. The most convenient and valid approach appears to be the use of microscopic characteristics, i.e. atomic parameters. Averaged

values of the atomic parameters, though physically meaningless, are often very accurately correlated to the properties of multi-component systems. Thus averaged values of the ionisation potential and of the electron affinity can be correlated with the work function of alloys, averaged radii with the molar volume, etc.

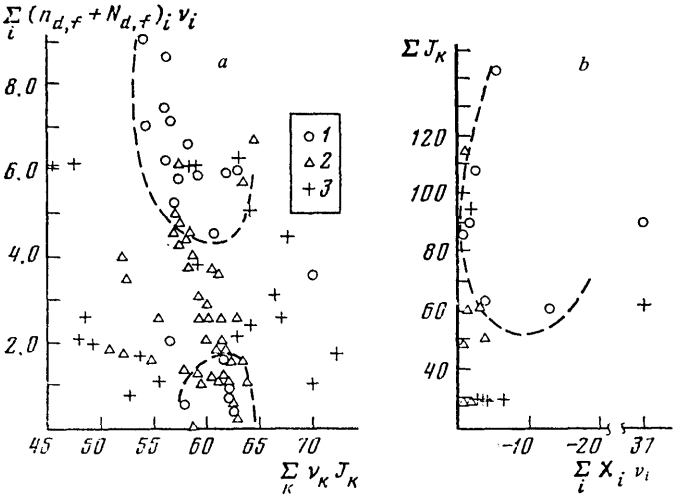


Figure 4. Distribution in criterion space of the catalysts for the hydrogenation of esters (*a*) and for the conversion of formaldehyde/isobutene into isoprene (*b*). Points *1–3* as in Fig. 3.

We have used averaged atomic parameters[28] as criteria for predicting hydrogen peroxide decomposition catalysts by a linear binary classifier with a non-zero threshold. The training sequence consisted of 93 alloys of silver or gold with Group VIII metals, for which catalytic activity data are given in a recent review[51]. The proportion of correctly classified catalysts was 85%, which appears to be the highest success rate which this algorithm can offer. Predictions were attempted for 38 catalysts. The predictive power of the algorithm was found to be slightly lower than the classifying power (80%).

The next improvement in the system of criteria is the replacement of atomic by averaged orbital characteristics, corresponding to the electrons in the valence sub-shells of the atoms. These include the quantum numbers, the orbital ionisation potentials, the average orbital radii, etc. Average values of whole classes of some of these characteristics can also be used as criteria. This allows us on the one hand to adopt a more flexible training program, since it can be shown[35] that the use of these characteristics is equivalent to changing from a linear classifier to one of a higher order. On the other hand the use of averages of a class of criteria (i.e. the moments of the criteria) justifies the departure from a purely additive scheme in the calculation of the parameters of multi-component catalysts.

The general formula for calculating the criteria thus becomes

$$X_{n\alpha} = \sum_i \nu_i \sum_j c_{ij\alpha} x_{ij}^n$$

where ν_i is the atom fraction of the *i*th component, *j* is the number of valence sub-shells, $c_{ij\alpha}$ is a weighting coefficient for averaging with respect to orbitals, and α is an index denoting the method of averaging used. This index is used because the orbital parameters for the atoms of each component can be averaged in different ways: over the number of orbitals in the valence sub-shells or over the number of valence electrons in the sub-shells (i.e. for the occupancy of the sub-shells, and hence also the number of unoccupied states ("holes")). All these methods of averaging give mutually independent results, with a different physical meaning.

Table 3. Classification, criteria, and level of influence of the criteria for bimetallic hydrogen peroxide decomposition catalysts.

[M], at.%	Activity, arbitrary units	Criteria				
		I (−0.387)	II (−0.375)	III (−0.349)	IV (−0.344)	V (+0.266)
		High activity				
15 Ir	2.971	64.319	7.998	488.54	524.41	0.470
11 Rh	2.371	63.317	7.920	504.85	540.84	0.364
11 Pt	2.114	62.770	7.895	498.89	533.15	0.200
25 Ru *	1.509	75.337	8.654	399.04	418.16	0.857
40 Os	1.486	63.728	7.970	379.28	423.74	1.091
40 Os *	1.463	76.221	8.705	376.20	399.04	0.750
25 Ru	1.386	62.908	7.917	466.26	506.97	0.857
		Low activity				
95 Rh *	0.314	79.465	8.860	405.17	420.04	1.312
92 Rh	0.314	78.706	8.816	412.17	431.47	1.294
78 Rh *	0.286	79.957	8.891	407.92	424.24	1.217
33 Pd	0.171	71.038	8.322	498.14	531.76	0.500
33 Pt	0.100	70.969	8.382	460.96	491.02	0.500
		Not classified				
17 Ir *	1.900	82.878	9.100	410.02	425.93	0.501
25 Pt	0.200	68.237	8.220	475.02	506.83	0.400

Symbols (criteria): I) square of the valance orbital ionisation potential (VOIP) of the neutral atom averaged over the unoccupied states of the valence sub-shells; II) VOIP of the neutral atom averaged over the unoccupied states of the valance sub-shells; III) square of the VOIP of the positively charged ion (+1) averaged over the number of atomic orbitals in the sub-shells; IV) square of the VOIP of the positively charged ion (+1) averaged over the number of valance electrons in the sub-shells; V) orbital quantum number averaged over the unoccupied states in the valence sub-shells. The level of influence of the criteria is shown in brackets.

*Au–M alloys; all the other alloys are Ag–M.

We have tested a system of 54 orbital criteria by producing a new set of predictions for the hydrogen peroxide decomposition catalysts mentioned above. Using the same algorithm, we raised its classifying power to 94.5% and its predictive power to 97.4%. Table 3 lists the most influential criteria for some of the compositions used in the classification, and Table 4 gives the corresponding values for the catalysts whose activity was predicted.

The use of orbital parameters as criteria in recognition algorithms brings the statistical approach closer to the methods of quantum mechanics and quantum chemistry, since those parameters are used in the calculation of the electronic structure of solids. Further progress in this direction can

be expected. Thus it should be possible to express the structure of multi-component catalysts by the procedures used in the cluster component method[48,49], i.e. to characterise the catalyst in terms of a set of lattice fragments (clusters) corresponding to different types of lattice site. In this representation the formula of an ideal binary crystal with an NaCl type lattice becomes $[AB_6]_{1/2}[BA_6]_{1/2}$, whereas an ideal substitutional solution with the same lattice and the same proportion of components is described by the formula

$$[AA_6]_{\nu_6^A}[AA_5B]_{\nu_5^A}[AA_4B_2]_{\nu_4^A}\ldots[BB_6]_{\nu_6^B}[BB_5A]_{\nu_5^B}\ldots$$

where ν_i^A and ν_i^B are calculated by assuming a random distribution of the two types of atom over the lattice sites. Averaged values of the parameters of the electronic structure of the clusters, calculated by quantum-chemical methods, can conveniently be used as criteria for the multi-component system. The averaging over the cluster components of a whole series of electronic structure characteristics (local densities of states, their moments, atomic contributions to the total energy, etc.) could have real physical significance, since these characteristics are strictly additive.

Table 4. Prediction of bimetallic hydrogen peroxide decomposition catalysts.

{M}, at.%	Criteria				
	I	II	III	IV	V
			High activity		
52 Ru *	70.710	8.386	372.06	391.6	1.219
56 Os *	69.448	8.312	296.15	318.84	1.254
29 Ir *	81.281	9.011	399.28	415.22	0.734
76 Ru	66.393	8.133	368.49	395.05	1.390
72 Os	65.355	8.071	283.56	311.20	1.367
32 Ir	68.491	8.280	457.27	491.29	0.780
12 Pt	63.161	7.918	497.33	531.49	0.214
			Low activity		
75 Rh *	80.047	8.897	408.52	424.89	1.200
57 Rh	74.922	8.595	450.92	481.41	1.065
34 Pd	71.244	8.334	497.74	531.30	0.507
71 Pt	80.00	8.919	400.12	419.76	0.830
			Incorrect prediction		
14 Rh	64.539	7.991	501.40	537.31	0.438

Note: the criteria and their influence are as defined in Table 3.

*Au–M alloys; all the other alloys are Ag–M.

So far in our discussion we have treated multi-component catalysts as fairly homogeneous solids (effectively single-phase systems). In fact, a large group of industrial catalysts has a more complex structure, and must be treated as multi-phase systems. The most important members of this group are the supported catalysts, consisting of an active metal phase and a carrier. In spite of their greater complexity, their catalytic properties can be predicted by the methods discussed here. The only new requirement is for a unified criterion space, including the properties of the carrier and of the support phase. The properties of each phase separately can be calculated by averaging.

As an alternative to averaging, we can approach the problem of multi-component catalyst prediction by constructing multi-level heuristic programs[46]. This possibility was tested[9,22] by applying the "Prognoz-76" program to the set of data on catalysts for the oxidation of 2,6-dimethylpyridine to 2,6-pyridinedialdehyde. The program was based on the following algorithm. The first stage consists in classifying the objects of the training sample according to the amount of components, using the "Prognoz-73" program. In the second stage the same program is used to identify the influential conjunctions of component properties which determine the assignment of the catalyst to class A. At this stage a distinction is made between the main components, present in all catalysts of classes A and B, and the promoters. In the third stage the substances possessing criteria similar to those of the components of class A catalysts are selected. In the last stage consisting of components with the required criteria in the optimum relative amounts (as determined in the first stage) is synthesised.

We should stress that methods of predicting multi-component catalysts are not yet fully developed. Further tests are needed before syntactic algorithms can be used to rationalise the combination of components with different properties into a single complex catalyst. The initial "words" for the grammar of syntactic algorithms of this type could be provided by the proposed[6] scheme of logic attributes of a good catalyst, such as activity of the components in side-reactions, type of adsorption, influence of promoters on the geometry of the crystal lattice of the catalyst, etc. However, in computer-aided recognition schemes the predictive power of this scheme is enhanced by the presence within each logic unit of quantitative information from the data banks, which facilitates the gearing of logic couplings.

V. ANALYSIS OF CATALYTIC ACTION MECHANISMS (THE INVERSE RECOGNITION PROBLEM)

Pattern recognition methods can be used not only in the prediction and classification of catalysts, but also as an effective means of analysing the mechanism of catalytic reactions. For this purpose we use the "inverse pattern recognition problem", whose goal is not the prediction of new catalysts but the study and the ranking of the properties of catalysts, reactants, and products (in so far as they affect the rate and the selectivity of the stated reaction). By identifying these products and relating them semi-quantitatively to the output parameters of the reaction the investigator can acquire a basis for a theoretical model of the catalytic reaction and of the relationships between its mechanism and the structure of the reactant and product molecules. Stricter functional dependences can be obtained from the sem-quantitative relationships. The data required for analysing catalytic action mechanisms follow directly from the method of solving the prediction problem, and can be considered to be incidental results. This is because all recognition algorithms contain some provision for minimising the number of dimensions of the problem, i.e. for "rejecting" unimportant criteria, so that the criteria are ranked in terms of their influence on the output parameters of the objects.

Several algorithms for ranking the criteria and minimising their number have been used in the solution of catalysis problems. In the potential functions methods a simple "sliding recognition" algorithm (involving repeated recognition) has been effective. Each criterion is omitted in turn from the initial sample: if the success rate of the recognition is unchanged (or increased) by this omission the criterion is excluded from all further calculations. The other criteria (or properties) are ranked according to their contribution to the success of the recognition.

The pattern recognition problem for CO oxidation catalysts was solved by taking an initial set of 26 properties[39]. However, by using the sliding recognition algorithm the number of criteria was decreased to eight. We note that only four of these could be related to the catalytic activity by correlation analysis methods with a probability greater than 95%. The other criteria, taken singly, showed practically no correlation with the catalytic activity: this shows the importance of combinations of properties in creating an effective solid catlyst, as well as underlining the limitations of the predictive power of correlation methods in one dimension (or in a few dimensions).

Table 5. Results of the computer recognition of the selectivity of oxidation reactions on a V_2O_5 catalyst[41].

No.	Initial compound	Final compound	Success of prediction
	Class A (selectivity greater than 50%)		
1	*o*-Xylene	Phthalic anhydride	+
2	Naphthalene	Phthalic anhydride	+
3	Anthracene	Anthraquinone	+
4	Toluene	Benzoic acid	+
	Class B (selectivity between 50 and 10%)		
1	*m*-Xylene	Maleic anhydride	+
2	Pent-1-ene	Acetaldehyde	–
3	Pent-2-ene	Acetaldehyde	–
4	Butadiene	Maleic anhydride	+
5	Naphthalene	1,4-Naphthylquinone	+
6	Anthracene	Phthalic anhydride	+
7	*p*-Xylene	Maleic anhydride	+
8	*p*-Xylene	*p*-Toluic anhydride	+
9	*p*-Xylene	Terephthalic anhydride	+
10	2-Methylbut-2-ene	Acetone	+
11	*m*-Xylene	*m*-Toluic aldehyde	+
12	Toluene	Benzaldehyde	+
13	2-Methylbut-2-ene	Acetaldehyde	+
14	2-Methylbut-1-ene	Acetone	+
15	2-Methylbut-2-ene	2,3-Epoxy-2-methylbutene	–
16	2-Methylbut-1-ene	Acetaldehyde	+
17	But-1-ene	Maleic anhydride	+
18	But-2-ene	Maleic anhydride	+
19	3-Methylbut-1-ene	Acetone	+
20	Pent-1-ene	Acetone	+
21	Pent-2-ene	Propionaldehyde	–
22	Pent-2-ene	Methanol	–
23	2-Methylbut-2-ene	3-Methylbutan-2-one	–
	Class C (selectivity less than 10%)		
1	Propene	Formaldehyde	–
2	Pent-1-ene	Butanol	+
3	2-Methylbut-1-ene	2,3-Epoxy-2-methylbutane	+
4	Pent-2-ene	2,3-Epoxypentane	+
5	2-Methylbut-1-ene	Butanol	+
6	3-Methylbut-2-ene	Acetaldehyde	+
7	Pent-1-ene	Propionaldehyde	+
8	Pent-1-ene	Butyraldehyde	+
9	Pent-2-ene	2,3-Epoxypentane	+
10	Pent-1-ene	Methanol	+
11	Pent-1-ene	Crotonaldehyde	+
12	3-Methylbut-1-ene	Butanol	+
13	Propene	Acetaldehyde	+
14	Pent-2-ene	Butanol	+
15	Pent-2-ene	Acetone	–
16	2-Methylbutene	2,3-Epoxy-2-methylbutane	+
17	*p*-Xylene	*p*-Toluic acid	–
18	Butadiene	Formaldehyde	–
19	3-Methylbutane	3-Methylbutan-2-one	+
20	Pent-2-ene	Butan-2-ol	+
21	2-Methylbut-2-ene	Butanol	+
22	Pent-2-ene	Ethanol	+

In the "Prognoz-73" program the level of influence of the criteria is measured by the frequency with which a criterion occurs in the conjunctions of criteria leading to recognition. By using this algorithm we were able to lower the number of influential properties of the reactant and product molecules from 46 to five or six in the prediction of the selectivity of the oxidation of hydrocarbons on vanadium pentoxide. The problem was specially formulated to clarify the mechanism of oxidative conversions. The training sample consisted of a set of data[52,53] on the selectivity of the partial oxidation on vanadium pentoxide of more than 50 hydrocarbons. The results of the recognition experiment (Table 5) show a high proportion of successes, and therefore they confirm the validity of conclusions on the influence of the properties of reactant and product molecules on the selectivity of the reaction.

The properties of the reactants, catalysts, and products which affect the rate and the selectivity of the oxidation of hydrocarbons most markedly can be listed as follows. For the hydrocarbons being oxidised the properties (in decreasing order of influence) include: 1) the ionisation potential, 2) the number of symmetry operations, 3) the free valency index, 4) the magnetic susceptibility, 5) the dipole moment, 6) the steric accessibility of the C–C bond with maximum structural correspondence. For the oxidation products we have: 1) the π-electron density, 2) the highest π-electron level, 3) the ionisation potential, 4) maximum structural correspondence between the C–C bond and the crystal lattice of the catalyst, 5) steric accessibility of the C–C bond with maximum structural correspondence.

The most influential properties of the catalysts for the oxidation of carbon monoxide can be arranged in the following sequence: 1) density, 2) colour, 3) melting point, 4) ratio of the volume of the crystallographic holes in the oxide to the volume of the metal atom, 5) width of the forbidden energy gap, 6) molar magnetic susceptibility, 7) number of *d* electrons not involved in bond formation between metal and oxygen, 8) ionisation potential of the metal atom (averaged for the valence of the metal in the catalyst). The effective properties of the promoters[44] for vanadium–molybdenum catalysts used in the oxidation of furfural can also be listed (though the relative magnitude of their influence has not yet been determined): 1) distance between metal and oxygen atoms in the oxide, 2) electronegativity difference in the oxide, 3) width of the forbidden energy gap, 4) colour. Similarly, for the oxidation of dimethylpyridine: 1) width of the forbidden gap, 2) ionisation potential of the metal atom, 3) ratio of the volume of the crystallographic holes in the oxide to the volume of the metal atom, and 4) coordination number of the metal atom in the crystal lattice of the oxide.

These results allow us to make some general conclusions on the mechanism of oxidation reactions on heterogeneous catalysts, and can be used as evidence for the general validity of the inverse recognition problem.

The properties of the catalysts, and those of the organic molecules, can be classified into two groups: electronic and geometric. The selectivity of the oxidation reactions (i.e. the ratio of the rates for different reaction paths on the given catalyst) is a function of a combination of properties of both the reactant and the product molecules.

Let us examine more closely the electronic properties group of the solid. They all affect the mobility of the electrons in the solid, which is ultimately determined by the excitation energy of the electrons. These properites can be treated as enthalpic properties of the catalyst, which determine the activation energy of the reaction. The factors which affect the mobility of the electrons can themselves be divided into two groups: individual (determined by the electronic configuration of the metal atom, i.e. the ionisation potential, number of free *d*-electrons, etc.), and collective (electron density, width of the energy gap, colour, magnetic susceptibility). The latter group of properties includes also the melting point, which varies in the same sense as the energy of the interatomic

bonds in the lattice and therefore characterises the ease of "unpairing" of electrons in these bonds (in particular, the metal–oxygen bonds), and also the probability of forming lattice defects.

Thus, in contrast to the narrow view that catalytic properties result either from the local or from the collective properties of solids our analysis shows that the formation of an effective catalyst requires a combination of both types of property.

In addition to the electronic properties, the group of influential properties of catalysts includes parameters which characterise the geometry of the lattice: the ratio of the volume of the holes to the volume of the metal atom, and the distance between the metal and the oxygen atoms in the oxide. The mechanism of action of these parameters can be related both to the structural correspondence between the lattice of the catalyst and the organic molecule and to the ease of diffusion of oxygen atoms from the bulk to the surface, which can be very important in the partial oxidation of organic compounds.

Let us now turn to the properties of the organic molecules (reactants and products). The properties which affect the selectivity of the reaction now include the ionisation potential, the π-electron density and the π-electron level, the magnetic susceptibility, and the free valency index. This is not surprising, since the rate and the selectivity of the oxidation reaction are ultimately determined by electron transfer processes in the reactant–catalyst–product system.

Consider now the geometric properties of the organic molecules which are found to be influential. On the whole, their correlation with the mechanism of the catalytic action is self-evident: steric accessibility of the bond and maximum structural correspondence between the C–C bond and the lattice of the catalyst The importance of the number of symmetry operations of the reacting molecule is not coincidental: the influence of this factor arises from the increased probability of interaction between the reacting molecule and the catalyst surface, since interaction occurs during any type of collision if the molecule is symmetrical. The presence of a dipole moment in the molecule also increases the probability of effective collision by orienting the incident molecule in the electrostatic field of the catalyst. It is interesting to note that this analysis fully confirms Balandin's principle of geometric correspondence in heterogeneous catalysis.

We should also stress another result, which is often ignored. In order to show good selectivity a catalytic reaction requires a definite correspondence in the combination of properties not only between the reactant and the catalyst but also between the reactant ("substrate") and the product. Thus in the separation of the above set of hydrocarbon oxidation reactions into two classes with a selectivity threshold of 25% different criteria of the reactants and the products are found to be influential (Table 6), but their interplay is logical and easily interpreted. Clearly, a hydrocarbon of class A must be a good electron donor, and therefore it should have a low ionisation potential and a well delocalised system of π-electrons (a low minimum free valency index according to Coulson). Correspondingly, the reaction product should be stable, i.e. it should be a poor electron donor (with a high ionisation potential) and it should not have a highly developed π-electron system (in order to facilitate desorption of the product).

For the class B hydrocarbons we expect the opposite properties, as well as a high structural correspondence between the C–C bond of the reaction product and the oxygen of the lattice, which favours breakdown of the product molecule. It can be shown that the predominant properties (and their relative values) for the product and the reactant vary with the value of the threshold chosen for the classification.

This example of the analysis of a reaction mechanism has established that the effectiveness of the catalytic act in the oxidation of hydrocarbons depends on a favourable conjunction of two types of factor: enthalpic (related to the energetics of the electronic transitions) and entropic (related to the geometry and symmetry of the molecules and of the lattice). The catalyst, reactant, and product make equally important contributions to the process, and the outcome is determined by an appropriate combination of their properties.

Table 6. Combinations of properties in molecules which determine the selectivity of the reaction.

Molecule	Properties	
	Class A ("good")	Class B ("poor")
Oxidised hydrocarbon	Low value of I, low value of κ	high value of κ
Oxidation product	High value of I, optimum ρ_π, high value of ΔE	high level of structural correspondence between the C–C bond and the O-atom of the catalyst, high value of ρ_π, low value of ΔE

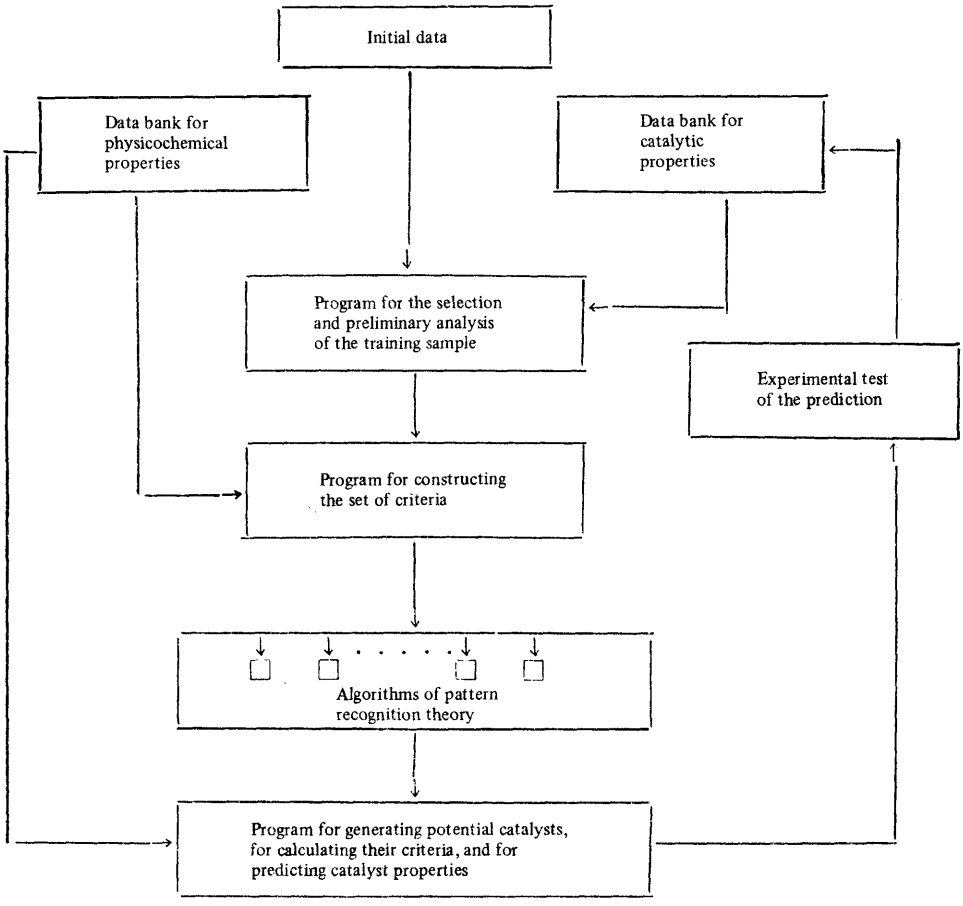
Symbols: I ionisation potential, κ minimum free valence index (Coulson), ρ_π minimum π-electron density, ΔE difference between the sums of the dissociation energies of the initial and the final compound.

VI. PROSPECTS OF FUTURE DEVELOPMENTS

Pattern recognition theory was formulated as a new topic in applied mathematics towards the end of the 1950's, but its practical application in science and technology had to await the development (in the 1960's) of the second generation of computers, which made the laborious calculations involved a practical possibility. The earliest applications of recognition methods to catalyst prediction were reported during that period. However, until specially adapted computational methods became available the problem remained intractable because it tended to make unreasonable demands on computer time.

The appearance and wide adoption in the 1970's of third-generation computers offered entirely new possibilities of automating all the steps in the prediction of catalyst behaviour by pattern recognition methods, including the choice of components of the training sequence, the search for and calculation of the physical properties, etc., and also the automatic "splicing" of all the steps. An efficient way of exploiting these possibilities is to set up an automatic system for predictions by pattern recognition methods, consisting of data banks, problem modules, and training programs. The significance of these concepts will be explained below: at this point we shall only stress the fact that the work which initially had to be done manually is now handled by the algorithms of the system.

The most general type of system for the prediction of heterogeneous catalysts is illustrated in the diagram. Without entering into its details, we shall examine the basic principles which should be borne in mind when designing a system of this type.



The first step in the creation of such a system is the organisation of renewable data banks, i.e. tables of properties, stored in the archives of the external memory of the computer. One of these should contain data on the state of the catalysts, their activity and selectivity, and other "desired" properties, while the other is dedicated to data on the physicochemical properties of solids. As stated above, a comprehensive and well documented handbook on catalyst properties is now available[18], with regularly published supplements. This can provide the basis of the catalyst data bank, to be updated as new information becomes available.

An important consideration in the assembly of the catalyst data bank is the reliability of the initial data. In any recognition algorithm the inclusion of unreliable data in the training sample must lower the validity of the prediction. Unfortunately, in the field of catalysis unreliable data are especially numerous, with many deliberately misleading ("diversionary") data in open publications as well as in patent applications. To some extent this can be guarded against by introducing weighted reliability coefficients: for example, 1 for original and well substantiated data in reputable publications, 0.8 for other publications, and 0.5 for patents. Allowance for the reliability coefficients can be made in the training of the recognition algorithm. Thus if the potential functions method is used the contributions of the objects to the total potentials of the classes are simply multiplied by the appropriate reliability coefficient. Other methods of allowing for the reliability of the data either directly or indirectly are available for the other algorithms.

The physicochemical properties bank should include data on a wide range of solids of various classes. The choice of these properties is a fundamental problem in the development and application of recognition methods. At present it is helpful to subdivide the criteria into two classes (basic, and derived). Since chemical reactions take place at the atomic–molecular level the basic criteria should include data on the crystallographic and electronic characteristics of the catalysts, and also on the atoms of their constituent elements. The criteria should be, as far as possible, quantitative, though the use of qualitative characteristics (of the "yes" or "no" type) is not excluded.

Criteria related to the electronic structure of solids (width of the energy bands, Fermi level, electron density and density of states, strength of interatomic bonds, etc.) determine the energetics of the chemisorption and of the overall catalytic act. In this sense they can be termed enthalpic. The geometric parameters of the crystal lattice (in particular, its symmetry) can influence the probability of various elementary acts, i.e. they can determine the entropy factor, as well as being naturally related to the electronic structure.

The derivative criteria are obtained by functional relationships from the basic criteria. Practically any of the physical properties of solids can be used as an example. Thus the density is related to the crystal lattice parameters and to the composition of the substance, the depth of colour to the width of the forbidden energy gap, etc. Supplementing the

basic criteria with derived criteria can be helpful in the solution of the inverse recognition problem.

The first step in the operation of a prediction system is to search the catalyst bank for catalysts which accelerate the reaction of interest and to classify them according to activity (or other specified property), number of phases (presence of a carrier, etc.), and chemical composition. These catalysts are the training sample. If the bank contains few data (or no data) on the reaction under study a sample of reasonable size can nevertheless be assembled by using data for analogous reactions[27,29].

In the next step the properties of the components of the catalysts in the training sample are extracted from the bank of physicochemical properties. These are used to generate the criteria to be applied in the recognition. The formulation of these criteria may require special calculations (for example, quantum-chemical), and therefore this part of the system may include a whole set of programs. Before compiling tables of criteria it is advisable to sift out the criteria which are mutually correlated. Superfluous criteria can be excluded by ranking them according to their influence, using the fastest (though not necessarily the most effective) recognition algorithm. These algorithms can be termed first-level algorithms. They can conveniently be realised by using the programs of the potential functions method.

When the required correspondence between the size of the training sample and the number of criteria has been achieved, the main step in the operation of the system can take place: the training of the second-level recognition algorithms. The purpose of this step is to find the optimum recognition procedure, offering the highest classifying power. If the required classifying power cannot be achieved by any of the algorithms we must go back to the criterion formulation step and find additional criteria (and criterion correlations).

When a classifier offering an acceptable effectiveness has been defined, a prediction of the target property can be made for all the proposed objects. In the previous methods of prediction the target properties were listed by the experimenter, but the automated scheme includes a block which generates possible compositions of (hypothetical) effective catalysts. This block operates in two steps. At first, by using repeated dichotomies or other decision processes, it outlines a region of effective criterion space corresponding to the optimum value of the catalyst property of interest. Then it generates hypothetical catalysts whose criteria lie within the selected region. In the case of one-component catalysts or promoters this is equivalent to choosing from the catalyst bank the compounds with the most favourable set of properties. In the case of multi-component catalysts the system identifies combinations of substances having criteria close to the optimum values.

If the dependence of the physicochemical properties of the catalysts on the concentration of the components can be described by a simple functional relationship (for example, some form of averaging) the problem of finding the best composition can be formulated mathematically as the search for a minimum in some function of the composition. The most obvious choice of this function is the sum of the squared deviations of the criteria for the chosen catalysts from their optimum value:

$$\Phi(\nu) = \sum_i \rho_i [\chi_{oi} - \chi_i(\nu)]^2$$

The weighting coefficients ρ_i reflect the degree of influence of the criterion. The final result from the operation of this system is a recommendation of the relative amounts of active components likely to yield the best catalyst or promoter.

The results of experimental tests of the predictions can be used to supplement the catalytic properties bank. If a catalyst offering the required effectiveness has still not been found, a new prediction is made with the expanded data bank, and the search becomes an iterative procedure. In this variant the pattern recognition method can be used for planning experiments on catalyst selection when the initial information is insufficient for "training" the computer[9].

REFERENCES

1. I.I.Ioffe and S.Z.Roginskii, Zhur.Fiz.Khim., 1957, **31**, 612.
2. K.Tanabe, "Solid Acids and Bases" (Translated into Russian), Izd.Mir, Moscow, 1973.
3. V.A.Roiter and G.I.Golodets, "Vvedenie v Teoriyu Kinetiki i Kataliza" (Introduction to the Theory of Kinetics and Catalysis), Izd.Naukova Dumka, Kiev, 1971.
4. G.K.Boreskov, Kinetika i Kataliz, 1967, **8**, 1020.
5. D.A.Dowden, Catalysis Rev., 1971, **5**, 1.
6. D.L.Trimm, Pure Appl.Chem., 1978, **50**, 1147.
7. D.L.Trimm, Chimia, 1979, **33**, 415.
8. D.L.Trimm, Chem.Technol., 1979, **9**, 571.
9. I.I.Ioffe, V.A.Reshetov, and A.M.Dobrotvorskii, "Raschetnye Metody v Prognozirovanii Geterogennykh Katalizatorov" (Computation Methods for the Prediction of Heterogeneous Catalysts), Izd.Khimiya, Leningrad, 1977.
10. S.Beran and R.Zagradnik, Kinetika i Kataliz, 1977, **18**, 359.
11. M.B.Kuz'minskii and A.A.Bagatur'yants, "Itogi Nauki i Tekhniki. Kinetika i Kataliz" (Advances in Science and Technology. Kinetics and Catalysis), Vol.8, Izd.VINITI, Moscow, 1980, p.99.
12. V.Medeleanu, N.Doca, O.Dragomir and I.Motoc, Bul. Sti.si Teh.Inst.Politehn.Timisoara, Ser.Chim., 1978, **23**(2), 152.
13. K.Kuchynka, J.Fusek, and O.Strouf, Coll.Czech. Chem.Comm., 1981, **46**, 52.
14. N.Doca, V.Medeleanu, O.Dragomir, and I.Motoc, Bul. Sti.si Teh.Inst.Politehn.Timisoara, Ser.chim., 1978, **23**(2), 156.
15. P.C.Jurs and T.L.Isenhour, "Pattern Recognition in Chemistry" (Translated into Russian), Izd.Mir, Moscow, 1977,
16. E.M.Savitskii and V.B.Gribulya, "Prognozirovanie Neorganicheskikh Soedinenii s Pomoshch'yu EVM" (Computer-aided Prediction of Inorganic Compounds), Izd.Nauka, Moscow, 1977.
17. V.E.Golender and A.B.Rozenblit, "Vychislitel'nye Metody Prognozirovaniya Lekarstv" (Computational Methods of Drug Prediction), Izd.Zinatne, Riga, 1978.
18. V.A.Roiter (Editor), "Kataliticheskie Svoistva Veshchestv. Spravochnik" (Handbook on the Catalytic Properties of Substances), Vols.1–4, Izd. Naukova Dumka, Kiev, 1968-1977.
19. E.G.Derouane, Ind.chim.belge, 1971, **36**(5), 359.
20. L.A.Currie, J.J.Filliben, and J.R.De Voe, Analyt. Chem., 1972, **44**. 438R.
21. Z.K.Amirova, R.M.Masagutov, B.F.Morozov, and S.I.Spivak, Kinetika i Kataliz, 1980, **21**, 1174.

22. Z.K.Amirova, V.A.Semenov, R.M.Masagutov, and B.F.Morozov, "Doklady na Chetvertuyu Mezhdunarodnuyu Konferentsiyu po Primeneniyu EVM v Khimii" (Papers Presented to the Fourth International Conference on the Application of Computers in Chemistry), Pt.2, Izd.Sibirsk.Otd.Akad.Nauk SSSR, Novosibirsk, 1979, p.559.
23. Z.K.Amirova and R.M.Masagutov, Symposium, "Neftekhimicheskie Protsessy v Mnogofaznykh Sistemakh" (Petrochemical Processes in Multi-phase Systems), Moscow, 1980, p.133.
24. Z.K.Amirova and S.I.Spivak, "Matematicheskie Metody v Khimii. Tret'ya Vsesoyuznaya Konferentsiya. Tom 3. Matematicheskaya Obrabotka Eksperimentov" (Mathematical Methods in Chemistry. Third All-Union Conference. Vol.3. Mathematical Treatment of Experiments), TsNIITEneftekhim, Moscow, 1980, p.53.
25. N.V.Korovin, D.G.Naryshkin, and N.V.Fadeeva, Zhur.Fiz.Khim., 1979, **53**, 1487 [Russ.J.Phys.Chem., 1979(6)].
26. D.G.Naryshkin, N.V.Korovin, and N.V.Fadeeva, Zhur.Fiz.Khim., 1980, **54**, 514 [Russ.J.Phys.Chem., 1980(2)].
27. I.I.Ioffe, V.S.Fedorov, A.L.Shapiro, D.N.Grinberg, and B.Yu.Gurevich, Zhur.Fiz.Khim., 1971, **45**, 1878 [Russ.J.Phys.Chem., 1971(7)].
28. I.I.Ioffe, A.M.Dobrotvorskii, and V.A.Belozerskikh, see Ref.24, p.61.
29. I.I.Ioffe, N.M.Danyushevskaya, and V.S.Fedorov, Neftekhimiya, 1977, **17**, 34.
30. A.G.Arkad'ev and E.M.Braverman, "Obuchenie Mashiny Klassifikatsii Ob"ektov" (The Training of Machines for Object Classification), Izd.Nauka, Moscow, 1971.
31. V.I.Vasil'ev, "Raspoznayushchie Sistemy. Spravochnik" (Handbook on Recognition Systems), Izd.Naukova Dumka, Kiev, 1969.
32. J.T.Chou and R.Gonzalez, "Principles of Pattern Recognition" (Translated into Russian), Izd.Mir, Moscow, 1978,
33. K.Fu, "Structural Methods in Pattern Recognition" (Translated into Russian), Izd.Mir, Moscow, 1977.
34. E.M.Braverman (Editor), "Automatic Analysis of Complex Systems" (Translated into Russian), Izd. Mir, Moscow, 1969, p.22.
35. N.Nilson, "Learning Machines" (Translated into Russian), Izd.Mir, Moscow, 1967.
36. M.A.Aizerman, E.M.Braverman, and L.I.Rozenoer, "Metod Potentsial'nykh Funktsii V Teorii Obucheniya Mashin" (The Potential Functions Method in the Theory of Machine Training), Izd.Nauka, Moscow, 1970.
37. M.M.Bongard, "Problema Uznavaniya" (The Recognition Problem), Izd.Nauka, Moscow, 1967.
38. D.N.Anderson and T.L.Isenhour, Pattern Recognition, 1973, **5**. 249.
39. I.I.Ioffe, V.S.Fedorov, and B.Yu.Gurevich, Symposium, "Osnovy Predvideniya Kataliticheskogo Deistviya" (Fundamentals of the Prediction of Catalytic Action), Izd.Nauka, Moscow, 1977, p.210.
40. O.V.Krylov, Kinetika i Kataliz, 1962, **3**, 502.
41. I.I.Ioffe, V.S.Fedorov, and K.M.Mukhenberg, Dokl. Akad.Nauk SSSR, 1969, **189**, 1290.
42. Y.Onishi, Y.Iisuka, and T.Hamamura, Bull.Chem.Soc. Japan, 1980, **53**, 263.
43. N.S.Barinov, G.M.Strashnova, I.I.Ioffe, D.V.Mushenko, V.S.Fedorov, and V.S.Zavgorodnyi, USSR P. 191 400/1975; Byull.Izobr., 1975, No.42, p.14.
44. V.E.Golender, A.B.Rozenblit, M.V.Shimanskaya, and I.I.Ioffe, Izv.Akad.Nauk LatvSSR, Ser.Khim., 1975(2), 224.
45. D.N.Pauli and A.E.Maxwell, "Factor Analysis as a Statistical Method" (Translated into Russian), Izd. Mir, Moscow, 1967.
46. E.V.Aleksandrov, "Osnovy Teorii Evristicheskikh Reshenii" (Fundamentals of the Theory of Heuristic Solutions), Izd.Sovetskoe Radio, Moscow, 1975.
47. M.D.Balmakov and Z.U.Borisova, Fiz.i Khim.Stekla, 1976, **2**(3), 234.
48. A.N.Men', R.Yu.Bogdanovich, V.M.Kamyshov, and V.B.Fetisov, "Sostav–Defektnost'–Svoistvo Tverdykh Tel. Metod Klasternykh Komponentov" (Composition–Defect–Property Relations in Solids. The Cluster Component Method), Izd.Nauka, Moscow, 1977.
49. G.P.Popov and G.N.Il'inova, Khim.i Tekhnol. Oksidnykh Magnitnykh Materialov, 1977(3), 3.
50. R.Elliott, J.Kramhansl, and P.Lis, Symposium, "Theory of the Properties of Disordered Materials" (Translated into Russian), Izd.Nauka, Moscow, 1977, p.11.
51. I.A.Zubovich, Symposium, "Kataliticheskie Reaktsii v Zhidkoi Faze" (Catalytic Reactions in the Liquid Phase), Izd.Nauka, Alma-Ata, 1967, p.7.
52. S.K.Battacharyya and I.B.Gulati, Ind.Eng.Chem., 1958, **50**, 1719.
53. S.K.Battacharyya and R.Krishnamurthy, Current Sci., 1959, **28**, 363.

All-Union Petrochemical
Processes Research Institute,
Leningrad