Pattern-recognition Search for the Basic Regularities in the Stability of Complex Hydrides. Part 1. A Simplified Model

OLDŘICH ŠTROUF [a] and SVANTE WOLD [b]

[a] Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, 250 68, Rez near Prague, Czechoslovakia and [b] Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå, Sweden

The relationship between structural variables and the stability of complex hydrides ABH_nD_{4-n} (A = alkaline atom; B = the Group III B atom; n = number of hydrogen atoms and D = non-hydride substituents) is investigated by means of the pattern recognition (PR) method SIMCA.

A learning set consisting of 95 stable and 20 unstable complexes was characterized by 49 structural variables. SIMCA sorted out 28 of these as relevant for the classification of the complexes as stable or unstable.

The resulting relations between these 28 variables for the two classes makes possible a qualitative interpretation of which factors influence the stability of the complexes. The method classifies 75 % of the complexes correctly which is encouraging considering the crudeness of the model.

For inorganic as well as organic chemists, the complex hydrides ABH_nD_{4-n} (A = alkaline metal; B = the metal of the Group III B; n = the number of hydride atoms and D = a substituent other than hydride) represent an interesting and synthetically important group of chemical compounds. Theoretically, the combination of A, B, n and D could give an extremely high number of compounds due to practically unlimited set of the possible substituents D. Nevertheless, numerous empirical facts (see Ref. 1 and references therein) have shown that only some of the complex hydrides are stable, whereas the others exhibit a significant instability.

In this paper, we design a simplified model for the stability of complex hydrides ABH_nD_{4-n} ($n = 1 - 4$) based on structural data [2] for A, B and D. For the quantitative interpretation of these multivariate data we used the SIMCA method, one of the recent pattern recognition (PR) techniques.

PATTERN RECOGNITION

The search for and use of regularities in empirical data has always been of great importance in chemistry. Quantitative methods specifically designed to detect regularities in multivariate data are finding increased use in chemistry under the name of pattern recognition (PR). For reviews, see Kowalski,[3,4] Jurs and Isenhour,[5] Redl, Cramer and Berkoff [6] and others.[7] The principles of PR can be described as follows:

1. Formulate the problem as a classification problem where objects are to be assigned to one of several classes on the basis of multivariate data observed on these objects (see Fig. 1). In the present application, we wish to classify complex hydrides (the objects) of the general formula ABH_nD_{4-n} as "stable" (class 1) or "unstable" (class 2). Hence, the number of classes, Q, is two.

2. The values of M variables are observed or otherwise defined for a number of objects "known" to belong to the different classes. These objects are called the *training set* or the reference set.

In the present example, each complex hydride is characterised by 49 variables (see below). The training set consists of 95 complexes "known" to be stable and 20 complexes "known" to be unstable.

5

Fig. 1. Data used in the analysis of data by pattern recognition methodology.

3. Observe or define the values of the same variables for objects of unknown class assignment (see Fig. 1). These data constitute the *test set*. In the present application, this test set consists of the data for 109 complex hydrides for which information about the stability is rather incomplete or entirely lacking.

4. Find the "common pattern" in each class by analysing the training set by means of a mathematical procedure, We have used the SIMCA method which is specifically designed for chemical pattern recognition.[5,9]

5. Compare the objects in the test set with the common pattern of each class using the
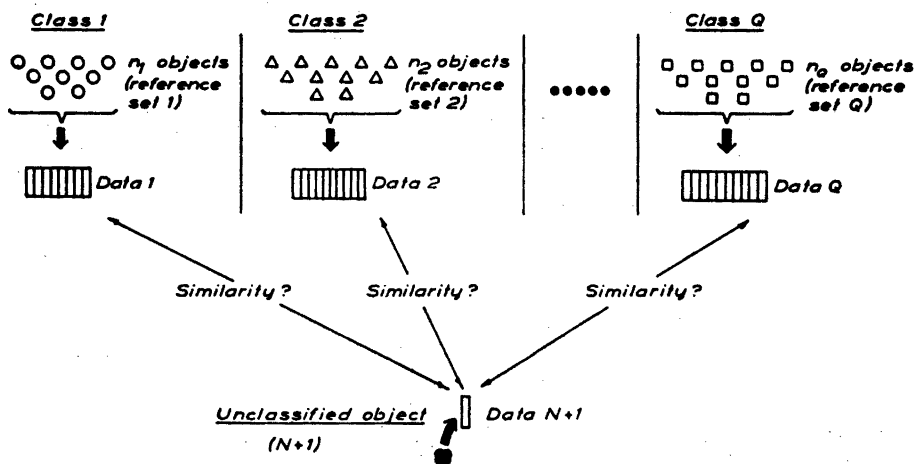

Fig. 2. The principle of pattern recognition as a methodology of assigning an object to the correct class on the basis of the values of M variables measured on the object.

mathematical procedure as a means of comparison. Each object is assigned to the class which shows the greatest similarity with the data of the object (Fig. 2).

Usually, PR methods are used to analyse empirical data such as spectra. In the present context, however, we are interested whether structural information can be used to predict the stability of complex hydrides. The variables used to characterise each hydride are therefore taken to be "theoretical", such as the electronegativity of atom A and the atomic weight of atom B. This use of PR methods to investigate structure reactivity relations has been applied to some extent to problems in biochemistry and pharmacology [6] but, to our knowledge, not in inorganic chemistry.

THE SIMCA METHOD

The PR method used here is called SIMCA (Simple Modelling Class Analogy). The method is based on the fact that data $y_{ik}^{(q)}$ (see Fig. 1) observed on a group of similar objects (class q) can, provided that a few continuity assumptions are fullfilled, be described by the principal components model [8,9]

$$y_{ik}^{(q)} = \alpha_i^{(q)} + \sum_{a=1}^{A} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)} \tag{1}$$

In eqn. (1), the parameters $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$ and $\theta_{ak}^{(q)}$ are determined as to make the sum of squared deviations $\varepsilon_{ik}^{(q)}$ minimal. The mathematical method to make this determination for a given matrix of data $(y_{ik}^{(q)})$ is known under the name principal components analysis.

The PR method based on eqn. (1) has the following steps:

A. Given data $y_{ik}^{(q)}$ for the objects in the training set of class q, the parameters in eqn. (1) are determined to minimise the sum of squares deviation (method of least squares).

B. The classification of non-assigned objects (the test set) is then accomplished by fitting the data of each such object (denoted by x_{ij}) to each of the Q class models (index q) with the parameters $\alpha_i^{(q)}$ and $\beta_{ia}^{(q)}$ fixed to the values obtained in step A. This corresponds to one linear regression for each object and each class, i.e. the determination of the coefficients $t_{aj}^{(q)}$ to minimise the residuals $e_{ij}^{(q)}$ in the least squares sense

$$x_{ij} - \alpha_i^{(q)} = \sum_{a=1}^{A} t_{aj}^{(q)} \beta_{ia}^{(q)} + e_{ij}^{(q)} \tag{2}$$

The object j is assigned to the class for which the residual variance $[s_j^{(q)}]^2$ is the smallest

$$[s_j^{(q)}]^2 = \sum_{i=1}^{M} \{e_{ij}^{(q)}\}^2 / (M - A) \tag{3}$$

Relevance of variables

The method allows the calculation of the *discrimination power* and the *modelling power* of each variable i. The former is computed as the sum of squared residuals for variable i when all objects in the training set are fitted to the class models of their "own" class divided by the sum of squared residuals for variable i when all objects in the training set are fitted to all other class models. The discrimination power gives valuable information about the importance of each variable for the prediction of the class assignment.

The modelling power of variable is essentially the residual (ε) variance of the variable divided by the variance in the original data (y) of the same variable, both variances computed over all objects in the training set. The modelling power indicates how important a variable is in describing the similarity within the classes.

Selection of variables

One important goal of the data analysis is a reduction of the number of variables. This is particularly important in the present problem since the variables were introduced with little knowledge of their actual relation to the stability of the complexes. Thus, the classification of a variable as relevant or irrelevant also indicates whether it is at all related to the stability of complex hydrides.

However, the selection of relevant variables (or, alternatively, the deletion of irrelevant variables from the set originally chosen) presents fundamental problems. Consider a training set of N objects, each "known" to belong either to class 1 or class 2. Let us then introduce a number of variables by means of which we wish to "explain" the classification of the training set. It is easily realized that *even if we pick these variables completely at random, we will sooner or later find a variable that is highly correlated*

with the classification of the training set. The problem is substantially worse if we also allow combinations of variables to be correlated with the classification.

Hence, it is evident that if we condition the selection of variables on their discrimination power — *i.e.* we search for variables that are highly correlated with the classification of the training set — we must limit the number of original variables. Usually, a ratio between the number of variables and the number of objects in the training set of less than 1 to 10 is recommended.[4] This would limit the number of variables to 12 in the present case being substantially smaller than we actually have.

However, if we instead condition the selection of the variables on their *modelling power*, the above problem of "contrivedness" is considerably less serious. We then search for variables being highly correlated with several other variables in each class submatrix. The chance of finding such variables in a bunch of random variables is sufficiently small when the number of original variables is of the same order as the average number of objects in each class. In the present case we have, on the average, 57 cases in each class, limiting the number of variables to $50-60$, more than the present number of 49. We have, in our selection of variables, gone one step further, deleting only variables showing *both* low discrimination and modelling power, thereby further reducing the change of this selection corresponding to a "random choice".

Furthermore, the set of retained variables is independently tested in a later step of the analysis (see below, verification). Hence, we feel that the deleted variables have little relation to the present classification problem.

In this manner, we have reduced the number of variables for the complex hydrides from 49 to 28, thus obtaining information on which theoretical variables are important for the prediction of the stability of a complex hydride.

Interpretation of the parameters

The values of the parameters $\alpha_i^{(q)}$ and $\beta_{ia}^{(q)}$ in eqn. (1) also give direct information about the "typical pattern" of the data in class q. Thus $\alpha_i^{(q)}$ is the average value of variable i in the

class and $\beta_{ia}^{(q)}$ contain information about the correlation structure. For example, if $\beta_{1a}^{(q)}$ is positive and $\beta_{2a}^{(q)}$ is negative, this tells us that an increase in the first variable can be compensated for by a decrease in the second one. Such information has a direct chemical interpretation as seen below and gives valuable clues to the factors influencing the stability of complex hydrides.

Scaling of data

In order to give each variable the same weight in the initial phase of the analysis, we have followed the usual procedure of subtracting the average of the variable (calculated over all complexes) and dividing by the standard deviation of the same variable. This makes each variable have the mean zero and the variance one. The scaling values used in the analysis of the hydride data are shown in Table 1.

PATTERN RECOGNITION RESULTS

1. Initial analysis. To give each variable the same weight in the analysis, the 49 variables were autoscaled[3,4] to uniform mean and variance by subtraction of the variable mean and division by its standard deviation times $\sqrt{223}$ (values in Table 1).

A SIMCA analysis of the scaled data using $A=5$ in eqn. (1) showed a fair separation of the two training sets. The classification was incorrect for 20 of the 95 "stable" (class 1) and 4 of the 20 "unstable" (class 2) complexes (see Table 2).

Table 1 shows the parameters α_i of the 49 scaled variables for class 1. The values of α_i for class 2 are the same but with opposite sign. These parameters, estimating the mean of each variable within the classes, are of the order 0.02 which is rather small compared to the residual standard deviation after one component in the final analysis, 0.06. Hence, the difference in each variable is insignificant between the two classes, only in combination do they make a contribution to the prediction of the stability of complex hydrides.

2. Reduction of the number of variables. By deleting variables with both low modelling and discrimination power, the number of

Table 1. Variables used in the PR analysis. A: Name of variable. B: Mean over all complexes. C: Standard deviation over all complexes. D: Modelling power with $A = 5$, 49 variables. E: Discrimination power, same model. F: α_i for class 1 (stable). G: β_i for class 1, $A = 1$, 28 variables. H: β_i for class 2, $A = 1$, 28 variables.

A		B	C	D	E	F	G	H
1 A, m.p. °C [a]		1184	499	.93	1.1	−.017	—	—
2 A, b.p. °C		1017	244	.88	1.2	−.017	—	—
3 A, density, g/ml [b]		83.5	29.5	.70	1.1	.012	—	—
4 A, atomic radius, Å [b]		188.7	32.3	.96	5.2	.018	.27	.08
5 A, coval. radius, Å [b]		155.0	32.3	.94	5.2	.017	.27	.07
6 A, ionic radius, Å [b]		92.26	30.2	.97	6.0	.017	.27	.08
7 A, 1st ion.energy, kcal/g mol		116.2	10.0	.83	4.0	−.016	−.28	−.06
8 A, electronegativity [a]		9.112	.809	.96	3.5	−.018	−.26	−.08
9 B, m.p. °C [a]		1320	921	.97	4.1	.004	.02	.26
10 B, b.p. °C		2988	778	.98	4.6	.006	.03	.28
11 B, density, g/ml [b]		301.3	169	.80	5.8	−.012	.05	−.34
12 B, atomic, radius, Å [b]		126.7	23.9	.99	4.6	−.008	−.04	−.30
13 B, coval. radius, Å [b]		105.7	20.1	.99	4.4	−.008	−.02	−.32
14 B, ionic radius, Å [b]		40.65	18.7	.98	6.4	−.009	−.01	−.34
15 B, 1st ion. energy, kcal/g mol		158.7	26.1	.96	6.6	.004	.04	.23
16 B, electronegativity [a]		17.13	2.38	.90	9.0	.001	.05	.14
17 D, mol.weight		54.68	37.4	.66	1.6	−.017	−.20	.07
18 D, No. of chain atoms		3.183	2.36	.74	3.1	−.014	−.19	.14
19 D, b.p. of DH at 760 Torr		16.92	138	.89	2.9	−.016	—	—
20 D, density of DH (20 °C)		756.6	455	.89	8.3	−.016	—	—
21 D, π-donor or acceptor (1,0 or −1)		.3348	.733	.57	4.7	−.028	−.08	.07
22 D, No. of substs. on 1st ligand atom		.7768	.896	.84	3.8	.001	−.24	.15
23 D, average electronegat. of these [a]		24.46	3.63	.52	2.1	−.006	—	—
24 D, coval. radius of 1st ligand atom [c]		788.2	214	.78	1.0	.001	—	—
25 D, ionic radius of 1st ligand atom [b]		72.15	17.7	.81	1.2	−.010	—	—
26 D, 1st ion.energy of 1st lig.atom		295.8	32.8	.76	2.0	−.012	—	—
27 D, electronegativ. of 1st lig.atom [a]		29.05	5.48	.83	3.3	−.024	−.19	.18
28 E, corresponding to var. 17		43.18	39.8	.72	2.1	−.008	−.23	.10
29 E,	18	2.509	2.50	.78	3.5	−.006	−.22	.16
30 E,	19	−47.40	167	.88	3.3	−.005	—	—
31 E,	20	604.8	524	.90	4.0	−.007	—	—
32 E,	21	.2768	.631	.63	2.9	−.022	−.07	.10
33 E,	22	.3438	1.08	.86	6.9	−.007	−.29	.16
34 E,	23	23.36	2.96	.63	1.4	−.007	—	—
35 E,	24	675.6	273	.79	2.7	.008	—	—
36 E,	25	62.76	23.5	.83	2.1	.001	—	—
37 E,	26	299.8	28.7	.77	4.0	−.012	—	—
38 E,	27	27.20	5.89	.87	2.1	−.011	−.22	.20
39 F,	17	28.87	37.1	.77	1.4	−.003	−.23	.11
40 F,	18	1.679	2.34	.79	2.9	−.001	−.22	.18
41 F,	19	−116.7	166	.90	2.8	.001	—	—
42 F,	20	426.6	497	.94	5.8	−.001	—	—
43 F,	21	.1741	.528	.64	3.1	−.013	−.07	.13
44 F,	22	−.0804	1.11	.89	5.2	.008	−.26	.16
45 F,	23	22.60	2.67	.69	1.7	−.001	—	—
46 F,	24	559.5	281	.91	3.1	.007	—	—
47 F,	25	52.58	24.1	.94	2.6	.003	—	—
48 F,	26	304.2	24.4	.73	2.7	−.003	—	—
49 F,	27	25.04	5.52	.89	3.1	−.002	−.21	.23

[a] Times 10; [b] Times 100; [c] Times 1000.

Table 2. Resulting classification of the hydrides. Prior classification: Compounds 1 − 95, stable (1); compounds 96 − 115, unstable (2); compounds 116 − 224, test set (0). A: Name of hydride. B: Prior classification in verification step. C: Resulting classification with $A = 5$ and 49 variables. D: Resulting classification with $A = 1$ and 28 variables. E: Resulting classification in the verification step.

A + sign after the resulting class number indicates that the ratio between the standard deviations for the two class models is larger than $\sqrt{2}$.

A	B	C	D	E	A	B	C	D	E
1 LiBH₄	1	1	1	1	53 LiAlH(OEt)₃	1	1	2	2
2 LiBH₃CH₃	0	2	1	1	54 LiAlH(OBu-tert)₃	0	1	2	1
3 LiBH₂(2,4,6-PhMe₃)₂	1	1+	1	1	55 LiAlH(OPent-tert)₃	1	1	1	1
4 LiBH₂SH₂	0	2	2	2	56 NaAlH₄	0	1	1	2
5 LiBHEt₃	1	1+	1	2	57 NaAlH₃(iso-Bu)	1	2+	1	1
6 LiBH(N₂C₃H₃)₃	0	1+	1	1	58 NaAlH₃(C≡C − C₄H₉-n)	0	2	1	1
7 LiBH(N₂C₃HMe₂-3,5)₃	1	1+	1	1	59 NaAlH₃(OBu-tert)	1	2+	2	2
8 LiBH(SMe)₃	0	2	2	2	60 NaAlH₂Me₂	0	1+	1	1
9 NaBH₄	1	1+	1+	1+	61 NaAlH₂Et₂	1	1+	1	1
10 NaBH₃Et	0	2	1+	1+	62 NaAlH₂(n-Pr)₂	0	1+	1	1
11 NaBH₃(n-Pr)	1	2	1+	1+	63 NaAlH₂(iso-Bu)₂	1	1+	1	1
12 NaBH₃CN	0	1+	1+	1+	64 NaAlH₂(C≡C₄H₉-n)₂	0	1+	1	1
13 NaBH₃CONHMe	1	1	1	1	65 NaAlH₂(OBu-tert)₂	1	1	2	2
14 NaBH₃(2,4,6-PhMe₃)₂	0	1+	1	1	66 NaAlH₂(OCH₂CH₂NMe₂)₂	0	1	2	1
15 NaBHMe₃	1	1+	1	1	67 NaAlH₂(OCH₂CH₂OMe)₂	1	1	2	1
16 NaBHEt₃	0	1+	1	1	68 NaAlHMe₃	0	1+	1	1
17 NaBH(n-Pr)₃	1	1+	1	1	69 NaAlHEt₃	1	1+	1	1
18 NaBH(OMe)₃	0	2	2+	2+	70 NaAlH(n-Pr)₃	0	1+	1+	1+
19 NaBH(O-Bu-t)₃	1	2+	2+	2+	71 NaAlH(iso-Bu)₃	1	1+	1+	1+
20 NaBH(OOCH)₃	0	2	2+	2+	72 NaAlHPh₃	0	1+	1+	1
21 NaBH(OCH₂CH₂OMe)₃	1	2+	2+	2+	73 NaAlH(NEt₂)₃	1	1	1	1
22 NaBH(N₂C₃H₃)₃	0	1+	1	1	74 NaAlH(OEt)₃	0	1	2	2
23 NaBHF₃	1	1+	2+	2	75 NaAlH(OBu-tert)₃	1	1	2	1
24 NaBH(n-Bu)₃	0	1+	1	1	76 NaAlH(OCH₂CH₂NEt₂)₃	0	2	1	1
25 NaBH(OPr-iso)₃	1	2+	2+	2+	77 NaAlH(OCH₂CH₂OCH₃)₃	1	2	2	1
26 NaBH₃NMe₂	0	2	1	1	78 KAlH₄	0	1+	1+	1+
27 KBH₄	1	1+	1+	1+	79 KAlH₂Me₂	1	1+	1+	1+
28 KBH₃CN	0	1+	1+	1+	80 KAlH₂Et₂	0	1+	1+	1+
29 KBH₃CONHMe	1	1+	1+	1+	81 KAlH₂(n-Pr)₂	1	1+	1+	1+
30 KBH₃CONMe₂	0	1+	1+	1+	82 KAlH₂(iso-Bu)₂	0	1+	1+	1+
31 KBH₃(COOEt)	1	1	1+	1+	83 KAlHMe₃	1	1+	1+	1+
32 KBH₃F	0	1	1+	1+	84 KAlHEt₃	0	1+	1+	1+
33 KBH₂(N₂C₃H₃)₂	1	1+	1	1	85 KAlH(n-Pr)₃	1	1+	1+	1+
34 KBHEt₃	0	1+	1+	1+	86 KAlH(iso-Bu)₃	0	1+	1+	1+
35 KBH(NC₄H₄)₃	1	1+	1	1	87 KAlHPh₃	1	1+	1	1+
36 KBH(N₂C₃H₃)₃	0	1+	1	1	88 KAlHCl₃	0	2+	1	1
37 KBH(N₂C₃HMe₂-3,5)₃	1	1+	1	1	89 RbAlH₄	1	1+	1+	1+
38 RbBH₄	0	1+	1+	1+	90 CsAlH₄	0	1+	1+	1+
39 CsBH₄	1	1+	1+	1+	91 LiGaH₄	1	2+	2+	2+
40 CsBH(NC₄H₄)₃	0	1+	1	1+	92 NaGaH₄	0	2	2	2+
41 LiAlH₄	1	1	2	2+	93 KGaH₄	1	1	1+	1
42 LiAlH₂Me₂	0	1+	2	2+	94 RbGaH₄	0	1+	1+	1+
43 LiAlH₂(iso-Bu)₂	1	1+	2	2	95 CsGAH₄	1	1+	1+	1+
44 LiAlH₂(CN)₂	0	1	2	2	96 LiBH₃SH	0	2	1	1
45 LiAlH₂(OMe)₂	1	2	2	2	97 LiBHMe₃	2	1+	1	1
46 LiAlHMe₃	0	1+	2+	2+	98 LiBH(OMe)₃	0	1	2+	2+
47 LiAlHEt₃	1	1+	2	2	99 LiBH(OCH₂CH₂OMe)₃	2	2	2+	2+
48 LiAlH(iso-Bu)₃	0	1+	1	1	100 NaBH(OEt)₃	0	2+	2+	2+
49 LiAlH(C₅H₉-cyclo)₃	1	1+	1+	1+	101 NaBH(OCH₂CH₂NMe₂)₃	2	2+	2+	2+
50 LiAlH(C₇H₁₃-cyclo)₃	0	1+	1+	1+	102 KBH(OCH₂CH₂OCH₃)₃	0	2	2+	2
51 LiAlHPh₃	1	1+	1+	1	103 LiAlH₂(OC₃H₇-iso)₂	2	2	2	2
52 LiAlH(NEt₂)₃	0	1	1	1+	104 LiAlH₂(OBu-sec)₂	0	2	2	2

Table 2. Continued.

A	B	C	D	E	A	B	C	D	E
105 LiAlH(OPr-iso)₃	2	1+	2	1	163 LiAlH₃(OC₁₀H₁₉)			2	
106 LiAlH(OBu-sec)₃	0	1	1	1	164 LiAlH₃Cl			2+	
107 NaAlH₃Et	2	2	1	1	165 LiAlH₂(Et)₂			2	
108 NaAlH₃(OCH₂CH₂NMe₂)	0	2+	2	2	166 LiAlH₂(Pr-n)₂			2	
109 NaAlH₃(OCH₂CH₂OCH₃)	2	2+	2	1	167 LiAlH₂(C₅H₉-cyclo)₂			1	
110 NaAlHCl₃	0	2	2	2	168 LiAlH₂(PhMe)₂			1	
111 LiInH₄	2	2+	2+	2+	169 LiAlH₂(OEt)₂			2	
112 LiInHBr₃	0	2+	2+	2+	170 LiAlH₂(OBu-tert)₂			2	
113 LiInHCl₃	2	2+	2+	2+	171 LiAlH₂(OC₁₀H₁₉)₂			1	
114 LiInHI₃	0	2+	2+	2	172 LiAlH₂Cl₂			2+	
115 LiTlH₄	2	2+	2+	2+	173 LiAlH₂Br₂			2+	
116 LiBH₃CN			1		174 LiAlH₂I₂			2	
117 LiBH₃NC₄H₄			1		175 LiAlH(C₂H₃)₃			1	
118 LiBH₃OH			2		176 LiAlH(C₁₀H₂₁-n)₃			1	
119 LiBH₃OMe			2		177 LiAlH(NPh₂)₃			1+	
120 LiBH₃OBu-tert			2		178 LiAlH(OMe)₃			2	
121 LiBH(Bu-n)₃			1		179 LiAlH(OPr-n)₃			1	
122 LiBH₃OH₂			2		180 LiAlH(OBu-n)₃			1	
123 LiBH(Bu-sec)₃			1		181 LiAlH(C₅H₁₁-n)₃			1	
124 LiBH(Bu-iso)₃			1		182 LiAlH(OC₆H₁₃-n)₃			1	
125 LiBHPh₃			1		183 LiAlH(OC₆H₁₁-cyclo)₃			1	
126 LiBH(OH)₃			2+		184 LiAlH(OPh)₃			1	
127 LiBHF₃			2+		185 LiAlHCl₃			2+	
128 NaBH₃D			1+		186 LiAlHBr₃			2+	
129 NaBH₃(C₅H₅)			1		187 LiAlHI₃			2	
130 NaBH₃COOH			1		188 NaAlH₃(OMe)			2	
131 NaBH₃COOMe			1		189 NaAlH₃(O-furfuryl)			2	
132 NaBH₃OH			1		190 NaAlH₂(OMe)₂			2	
133 NaBH₃(OPh)			1		191 NaAlH₂(OEt)₂			2	
134 NaBH₃(OPhCl-*p*)			1		192 NaAlH₂(OPr-n)₂			2	
135 NaBH₃(OPhCl-*m*)			1		193 NaAlH₂(OPr-iso)₂			2	
136 NaBH₃F			1		194 NaAlH₂(OPh)₂			1	
137 NaBH₃(PMe₂)			1		195 NaAlH₂(OCH₂CH₂OEt)₂			2	
138 NaBH₂(OH)₂			2		196 NaAlH₂-(OCH₂CH₂OPr-n)₂			1	
139 NaBH₂F₂			2		197 NaAlH₂-(OCH₂CH₂OPr-iso)₂			2	
140 NaBHD₃			1+		198 NaAlH₂-(OCH₂CH₂OBu-n)₂			1	
141 NaBHPh₃			1		199 NaAlH₂(OCH₂CH₂OPh)₂			1	
142 NaBH(OH)₃			2+		200 NaAlH₂-(OCHMeCH₂OMe)₂			2	
143 NaBH(OBu-n)₃			2+		201 NaAlH₂-(OCMe₂CH₂OMe)₂			2	
144 KBH₃(NMe₂)			1+		202 NaAlH₂(O(CH₂)₄OMe)₂			1	
145 KBH₃(NEt₂)			1+		203 NaAlH(C₆H₁₁-cyclo)₃			1+	
146 KBH₃(OH)			1+		204 NaAlH(C₅H₅)₃			1+	
147 KBH₂(N₂C₃HMe₂-3,5)₂			1		205 NaAlH(CH₂Ph)₃			1	
148 KBH₂(OH)₂			1		206 NaAlH(OMe)₃			2	
149 KBH(Pr-iso)₃			1		207 NaAlH(OPr-n)₃			2	
150 KBH(Bu-sec)₃			1		208 NaAlH(OPr-iso)₃			2	
151 KBH(C₅H₁₁-sec)₃			1		209 NaAlH(OBu-sec)₃			2	
152 KBH(C₅H₉-cyclo)₃			1		210 NaAlH(OPh)₃			1	
153 KBH(C₆H₁₁-cyclo)₃			1		211 NaAlH(OCH₂CH₂OEt)₃			1	
154 KBH(OH)₃			2		212 NaAlH(OCH₂CH₂OPh)₃			1	
155 RbBH₃(N(Bu-n)₂)			1		213 NaAlH(OCMe₂CH₂OMe)₃			1	
156 RbBH(C₅H₅)₃			1		214 NaAlH-(OCH₂CH₂CH₂OMe)₃			1	
157 LiAlH₃Me			2						
158 LiAlH₃Et			2						
159 LiAlH₃(C₂H₃)			2						
160 LiAlH₃(OMe)			2+						
161 LiAlH₃(OEt)			2+						
162 LiAlH₃(OBu-tert)			2						

Table 2. Continued.

A	B	C	D	E	A	B	C	D	E
215 NaAlH(O(CH$_2$)$_4$OMe)$_3$			1		220 CsInH$_4$			1+	
216 LiGaH$_3$(Me)			2+		221 NaTlH$_4$			2+	
217 NaInH$_4$			2+		222 KTlH$_4$			2	
218 KInH$_4$			1		223 RbTlH$_4$			2	
219 RbInH$_4$			1		224 CsTlH$_4$			1	

variables was decreased first to 38 and then to 28 variables without an increase in the error rate in the classification of the training set. For the ligands, such variables were deleted which showed low relevance for all three ligands. When further deletions were made, the error rate increased and we therefore stopped at the number of 28 (see Table 1).

3. Final analysis. The classification with these 28 variables was essentially the same with $A = 5$, 4, 3, 2 and 1. For simplicity of interpretation we have preferred the results with $A = 1$. The resulting parameters β_i for class 1 and 2 are shown in Table 1. The parameters α_i for the retained 28 variables are, naturally, the same as for the 49 variable case. The residual standard deviation with $A = 1$ is 0.06.

The classification of the training and test sets with the 28 variables and $A = 1$ using the parameters in Table 1 is shown in Table 2; for interpretation, see below.

4. Verification. To study the significance of the above classification, we divided the training set into two halves, letting the first half continue being the training set and making the second half constitute the test set with "known answers". These data were then analysed with $A = 1$ and the same 28 variables as above. Thus, the parameters α and β were calculated on the basis of the first half only and the classes of complexes in the second half were "predicted" on the basis of these parameters. The results shown in Table 2 are encouraging. The error rate in the classification of the test set was 10 of 47 for the first class and 2 of 10 for the second class. This is significantly better than chance ($\chi^2 = 19.1$, $p < 0.001$) which indicates that the parameters in Table 2 indeed have a real prediction power for the stability of complex hydrides.

DISCUSSION

Extraction of relevant variables (features). The results of computation show that the variables related to fundamental atomic properties of alkaline metals (atomic, covalent and ionic radii, the first ionization energy and electronegativity) exhibit significantly higher relevance than the variables related to derived properties (melting point, boiling point and density). The relevance of the variables proposed for the metals of the Group III B does not differ markedly in this respect.

The computation extracts five relevant variables for each ligand from eleven variables originally proposed. Three features (molecular weight, number of atoms and number of substituents on the first atom of ligand) represent the bulkiness of the ligand, whereas the remaining two, π-accepting ability and electronegativity, characterize the nature of the first ligand atom. In the formulation of the model we have not included the nature of the atoms of the ligand chain, other than the first one. This might play a role in the classification of some special type of complex hydrides (see below).

Factors influencing the performance of the classification. First, the results of the classification are based on the characterization of the atoms interacting directly with the central atom B. In the present model, the influence of the atoms located further in the chain of a ligand is described only by their bulkiness, not by their nature. Nevertheless, the high correctness of the classification (about 75 %) shows, that important characteristics of the crucial atoms have been included. An increase in the degree of correctness may be expected when the variables for nature of the other ligand atoms are considered (as well as the other additional variables).

Secondly, the correctness of the classification is surprisingly high in spite of the fact that the selection of the hydrides used for training as "stable" and "unstable" is not based on the rigid definition of "stability" and "instability", respectively. The situation in complex hydride chemistry forced us to use as samples such hydrides which were proclaimed by authors on the basis of more or less direct observations as "stable" or "unstable" compounds. But we are not able to check the situation. Often it is difficult to make a decision at all, e.g. some hydrides are stable in the solid state and unstable when dissolved in a solvent. But PR is a powerful method for classification of systems with low information level. The present work shows that it is sufficient if the learning set is correctly classified in the majority of cases; the individual details are less important as long as the average contains the correct information.

Furthermore, the correctness of the classification is surprisingly high if we also consider the shortage of "unstable" hydrides for training which may exert a substantial influence on the learning procedure.

Finally, possible errors in the data used and an information noise arising during their working up might also negatively influence the performance of the classification.

The total effect of the above four factors is probably responsible for the main part of about 25 % incorrectness in our classification case.

Note. When only a half of the original training set is used, the percentage of correct classification does not markedly change, but about 10 % of cases are re-classified.

General trends. The simplified model, MODEL-1, reveals some general fundamental trends in "stability" of hydrides ABH_nD_{4-n} as related to the ligand D:

1. The hydrides with ligands attached to the central atom B by means of atom with free electron pair(s) are significantly less stable than those where D are alkyls. The stability of alkoxo derivatives $(D = OR)$ might be to some extent overestimated in literature.

2. On the other hand, hydrides with ligands bonded to the atom B by an atom with unsaturated bond are more stable than alkyl hydrides of comparable size. This finding favours the study of hydrides of this type.

3. The hydrides ABH_nD_{4-n} where $A = Li$, Na, K and B equals boron or aluminium seem to be stable when D is small. Very interesting exceptions in this series are exhibited by $LiAlH_nD_{4-n}$ hydrides which, using our MODEL-1, are "unstable" in the case of small size of D. This is evident, e.g., from the "instability" of the derivatives with $D = CH_3$ and C_2H_5. The destabilizing effect of small size of the ligands in $LiAlH_nR_{4-n}$ seems to dominate even over the above stabilizing effect 2. The naturally extreme case is $LiAlH_4$ with very small substituents H^- which is classified into class of "unstable" compounds. But this classification can be correct if we consider that $LiAlH_4$ is thermodynamically metastable.[10] Some other ABH_4 are also classified as "unstable" ones (Table 3).

Non-generality of MODEL-1. The classification of hydrides in the training set with some donor atom in an appropriate position in the chain of the ligands is in some cases dubious because of our *a priori* knowledge. Thus, $NaBH(OCH_2CH_2OCH_3)_3$,[11] $NaAlH_2(OCH_2CH_2OCH_3)_2$[12] and $NaAlH(OCH_2CH_2OCH_3)_3$[12] classified as "unstable" are to a significant extent stable. On the other hand, a majority of included items, namely $LiBH(OCH_2CH_2OCH_3)_3$,[11] $NaBH(OCH_2CH_2N(CH_3)_2)_3$,[11] $KBH(OCH_2CH_2OCH_3)_3$,[11] $NaAlH_3OCH_2CH_2OCH_3$[12] and $NaAlH_3OCH_2CH_2N(CH_3)_2$[13] are classified "correctly" as "unstable" hydrides. These facts speak for a non-generality of our MODEL-1 in this respect.

Very interesting is also the classification of the above type of hydrides of "unknown stability". The subset of the hydrides of the general formula $NaAlH_2(O\text{-}\text{-}\text{-}OR)_2$ (where --- represents a carbon chain) is split by the MODEL-1 into "stable" and "unstable" classes, respectively. This classification is evidently

Table 3. Classification of ABH_4 according to the MODEL-1. Hydrides under the border line are classified as "unstable" compounds, otherwise as "stable" ones.

$LiBH_4$	$NaBH_4$	KBH_4	$RbBH_4$	$CsBH_4$
$LiAlH_4$	$NaAlH_4$	$KAlH_4$	$RbAlH_4$	$CsAlH_4$
$LiGaH_4$	$NaGaH_4$	$KGaH_4$	$RbGaH_4$	$CsGaH_4$
$LiInH_4$	$NaInH_4$	$KInH_4$	$RbInH_4$	$CsInH_4$
$LiTlH_4$	$NaTlH_4$	$KTlH_4$	$RbTlH_4$	$CsTlH_4$

sensitive even to very small changes of R. Thus, $NaAlH_2(OCH_2CH_2OC_3H_7-n)_2$ [14] is classified as "stable", whereas very similar $NaAlH_2(OCH_2CH_2OC_2H_5)_2$ [15] and $NaAlH_2(OCH_2CH_2-OC_3H_7-iso)_2$ [14] are classified as "unstable" ones. This fact calls for consideration of some secondary variables for the MODEL-2 as additional ones to the variables representing the effect of a donor atom in the chain of ligand.

The classification of the $NaAlH(O\cdots OR)_3$ hydrides as "stable" species is in accordance with the generally found higher stability of these monohydrido derivatives over corresponding dihydrido compounds.

Parameters alfa *

1. Alfa 4, 5 and 6 (for radii) are positive for class 1 and negative for class 2; the average size of metal A is larger for "stable" hydrides than four "unstable" ones.

2. Alfa 7 (for the first ionization energy) is negative for class 1 and positive for class 2; more easily ionizable alkaline metal gives (on average) more stable complex hydrides. This is a natural consequence of point 1 and the correlations in a group in the periodic system.

3. Alfa 8 (for electronegativity) is negative for class 1 and positive for class 2; on average, the complex hydrides with low electronegativity of alkaline metal are relatively more stable than those with high electronegativity. This is also a correlation of the same nature as point 2. Hence, considering the connection between variables 1 through 5 in a group in the periodic system, these variables mainly express the place of atom A in the group.

4. Alfa 11 (for density of the central atom) is negative for class 1 and positive for class 2; this shows that the central metal atoms with low density give generally more stable hydrides than specifically heavier ones.

5. Alfa 32 (π-acceptor ability of the first atom of the second ligand) is negative for class 1 and positive for class 2; the influence of the difference in accepting ability of the first ligand atom is highest for the 2nd ligand and stability is higher in the case of acceptors.

* Only significant parameters were used for interpretation.

Parameters beta *

1. For class 2, the parameters beta 12, 13, 14 (size) for central atoms B are negative, while the parameters beta 27, 38, 49 (electronegativity of the first atom), beta 22, 33, 44 (number of substituents on the first atom) and beta 18, 29, 40 (number of atoms in the chain) for the ligands are positive; thus, a larger atom B can be compensated not only by a less electronegative first atom, but also by a less bulky ligand. In the same way as for atom A, the parameters beta 15 and 16 representing the nature of the central atom B are positive as a consequence of the correlations in the periodic system.

2. For class 1, the parameters beta 4, 5, 6 expressing the size of alkaline metals A are positive, whereas the parameters beta 27, 38, 49, beta 22, 33, 44, beta 18, 29, 40 and beta 17, 28, 39 (molecular weight of D) for ligands are negative; thus, a larger atom A is compensated by a less electronegative first ligand atom as well as by a less bulky ligand.

Further, the negative signs of the parameters beta 7 and 8 for the first ionization energy and electronegativity, representing the nature of the alkaline metals A, show that a decrease in these two factors potentiates the trends found for the ligands.

Conclusion

We feel that PR methodology gives an interesting opportunity to investigate complex structure-reactivity relationships. Considering the crudeness of the present model, exhibited by the rather simple nature of the variables used, the results of the present investigation is encouraging. We are presently working on an extension of MODEL-1 to include more variables and on an extension of the data set to more cases in the training set.

* Only significant parameters were used for interpretation.

REFERENCES

1. Hajos, A. *Komplexe Hydride.* In Kirsten, W., Ed., *Organisch-praeparative Methoden,* Band 4, VEB Deutscher Verlag der Wissenschaften, Berlin 1966.
2. Štrouf, O. and Wold, S. *Data Base 1. Stability of Complex Hydrides (MODEL-1),* Umeå University, Umeå 1977.
3. Kowalski, B. R. *Anal. Chem. 47* (1975) 1152 A.
4. Kowalski, B. R. In Klopfenstein, C. E. and Wilkins, C. L., Eds., *Computers in Chemical and Biochemical Research,* Vol. 2, Academic, New York 1974.
5. Jurs, P. C. and Isenhour, T. L. *Chemical Applications of Pattern Recognition,* Wiley-Interscience, New York 1975.
6. Redl, G., Cramer, IV, R. D. and Berkoff, C. E. *Chem. Soc. Rev. 3* (1974) 273.
7. Kanal, L. *IEEE Trans. Inf. Theory 20* (1974) 697.
8. Wold, S. *Pattern Recognition 8* (1976) 127.
9. Wold, S. and Sjöström, M. In Chapman, N. B. and Shorter, J., Eds., *Advances in Linear Free Energy Relationships,* Vol. 2, Plenum, London 1977.
10. Ref. 1., p. 80.
11. Sochor, P., Kadlecová, H. and Štrouf, O. *Collect. Czech. Chem. Commun. 40* (1975) 3177.
12. Čásenský, B., Macháček, J. and Abrham, K. *Collect. Czech. Chem. Commun. 36* (1971) 2648.
13. Macháček, J., Čásenský, B. and Abrham, K. *Collect. Czech. Chem. Commun. 38* (1973) 343.
14. Křiž, O., Čásenský, B. and Štrouf, O. *Collect. Czech. Chem. Commun. 38* (1973) 2076.
15. Čásenský, B., Macháček, J. and Vít, J. *Czech. Pat.* 134 720 (15 Jan. 1970).