Databases on the Properties of Substances and Computer-Assisted Design of Inorganic Compounds

Nadezhda Kiselyova[1][0000-0001-7243-9096], Victor Dudarev[2][0000-0002-3583-0704] and Andrey Stolyarenko[1]

[1] A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS), Moscow, 119334, Russia

[2] National Research University Higher School of Economics, Moscow, 109028, Russia

kis@imet.ac.ru

Abstract. The virtually integrated distributed system of databases on the properties of inorganic substances and materials of the A.A. Baikov Institute of Metallurgy and Materials Science, Russian Academy of Sciences is considered. The information-analytical system for automation of process of new inorganic compounds computer-assisted design based on machine learning methods usage for search for regularities in information of the databases on inorganic substances and materials properties is discussed. The results this system application for compound design that have not yet been synthesized are presented.

Keywords: Database, Inorganic Substance and Material, Machine Learning.

1 Introduction

Modern information technologies have made it possible to systematize and make available a huge array of data accumulated by chemistry over the centuries. Chemists and materials scientists make extensive use of the rich capabilities provided by numerous databases (DB), including the database on the properties of inorganic substances and materials (DBs PISM) [1], containing not only publications [2-5], but also data on the properties of substances [1,6-11]. More detailed information on the information resources of inorganic chemistry is given in the IRIC database developed by us [12].

Information service does not limit the capabilities of the developed databases. One of the ways to make rational use of information on substances is the search for regularities that connect the properties of substances with the parameters of components. The objective existence of such regularities is a consequence of the Periodic Law. However, numerous attempts to present the desired complex regularities in an analytical form, as a rule, were unsuccessful, especially in the case of multicomponent substances. The methods for finding such complicated regularities in the data, based on the ideas of machine learning, were developed. In the mid-sixties, the idea of using machine learning to find regularities, that relate the properties of inorganic compounds to the parameters of components, was first proposed in our Institute of Metallurgy and Materials Science (IMET) [13]. Already the first calculations allowed us to find the relationship

between the properties of binary compounds and the parameters of chemical elements, as well as to use the found regularities to predict compounds not yet obtained with an accuracy of about 90% [14]. Our further research in this area was associated with the use of more advanced machine learning programs [15–17] and complication of the composition of the compounds being predicted [18–19].

2 Integrated database system of IMET RAS on the properties of inorganic substances and materials

The source of information for the use of machine learning methods is the DBs PISM. In contrast to the databases usually used for these purposes, the information systems developed by the Institute of Metallurgy and Materials Science, Russian Academy of Sciences [1, 6, 11], by their functional structure, are focused on the selection of information for machine learning, which significantly reduces the time for preparation and analyzing the necessary data.

One of the most important problems in the application of machine learning to inorganic chemistry is the inconsistency of data obtained by different researchers. In this regard, the selection of information for machine learning is carried-ot by qualified experts in this subject area. This procedure is facilitated by providing the experts with the full texts of publications contained in our DBs PISM, from which examples are selected for machine learning, as well as through special programs for detecting sharply distinguished objects (outliers).

Now the integrated system of the DBs PISM includes the information systems developed in the IMET [1, 6, 11]: on the phase diagrams of semiconductor systems (Diagram), the properties of the acoustooptical, electro-optical, and nonlinear optical substances (Crystal), the band gap of inorganic substances (Bandgap), the properties of inorganic compounds (Phases), and the properties of chemical elements (Elements), the AtomWork database on the properties of inorganic substances, developed at the National Institute for Materials Science (NIMS, Japan) [8], and the TKV on substances thermal constants, developed in the Joint Institute for High Temperatures of RAS and Lomonosov Moscow State University cooperation.

The Phases database on the properties of inorganic compounds currently contains information on the properties of approximately 54000 ternary compounds and more than 34000 quaternary compounds, collected using more than 36000 publications. It includes brief information about the most common properties of inorganic compounds: crystal chemical (the type of crystal structure with indication of the temperature and pressure above which this structure is implemented, the crystal system, the space group, the number of formula units in the unit cell, and the lattice parameters) and thermophysical (the melting type and temperature, the temperature of decomposition of the compound in solid or gaseous phases, and the boiling point at atmospheric pressure) data. In addition, the database contains information on the superconducting properties of compounds. This database is available on the Internet for registered users [11].

The Elements database includes information about 90 of the most common properties of chemical elements: the thermal (the melting and boiling points at 1 atm and the

standard values of thermal conductivity, molar heat capacity, enthalpy of atomization, entropy, etc.), size (the ionic, covalent, metal, and pseudopotential radii, the atomic volume, etc.), and other physical properties (the magnetic susceptibility, electrical conductivity, hardness, density, etc.); etc. The database is available on the Internet [11].

The Diagram database contains data on phase P,T,x-diagrams of binary and ternary semiconductor systems and the physicochemical properties of phases formed in them, collected and evaluated by highly qualified experts. The Diagram database is available on the Internet for registered users [11].

The Bandgap database includes information about the band gap of more than 3600 inorganic substances and is available on the Internet [11]. It has English version only.

The Crystal database includes information about the piezoelectric (piezoelectric coefficients, elastic constants, etc.), nonlinear optical (nonlinear optical coefficients, the Miller tensor components, etc.), crystal chemical (the type of the crystal structure, crystal system, space and point groups, the number of formula units per unit cell, and the crystal lattice parameters), optical (refractive indices, the transparency band, etc.), and thermal (melting point, specific heat, thermal conductivity, etc.) properties of more than 140 acousto-optical, electro-optical, and nonlinear optical materials, collected and evaluated by highly qualified experts in the subject area. It has Russian and English versions available for registered users on the Internet [11].

The AtomWork Inorganic Material Database (NIMS, Japan) contains information about more than 82000 crystal structures, 55000 values of the properties of materials, and 15000 phase diagrams; it is also available on the Internet [8].

The TKV DB on substances thermal constants contains information, available online from the Internet, on about 27 thousand substances formed by all chemical elements.

The complex integration approach that combines integration at data and user interfaces level is applied to these database integration [20]. The special single entry point allows a search for the all data on certain substance from different DBs.

3 Inorganic Compounds Computer-Assisted Design System

Machine learning procedure involves several stages:

1. The objects selection for machine learning.
2. The attribute description formation (including the most informative attributes selection and filling attribute values gaps also).
3. The best ML algorithms selection.
4. Machine learning including application of algorithms ensembles and collective solution synthesis in a case of several algorithms usage.
5. ML quality estimation.
6. New objects prediction and results interpretation.

The special information-analytical system (IAS), which, in addition to the information service for professionals, is designed to search for regularities in big chemical data and computer design of inorganic compounds was developed in IMET [21]. It includes (Fig. 1), along with the integrated system of DBs PISM, a subsystem of information analysis

and predictions, bringing together a set of programs of machine learning, a base of found regularities (the knowledge base), a base of predictions of the possibility of forming and properties of inorganic compounds that have not been yet synthesized, and a management subsystem.

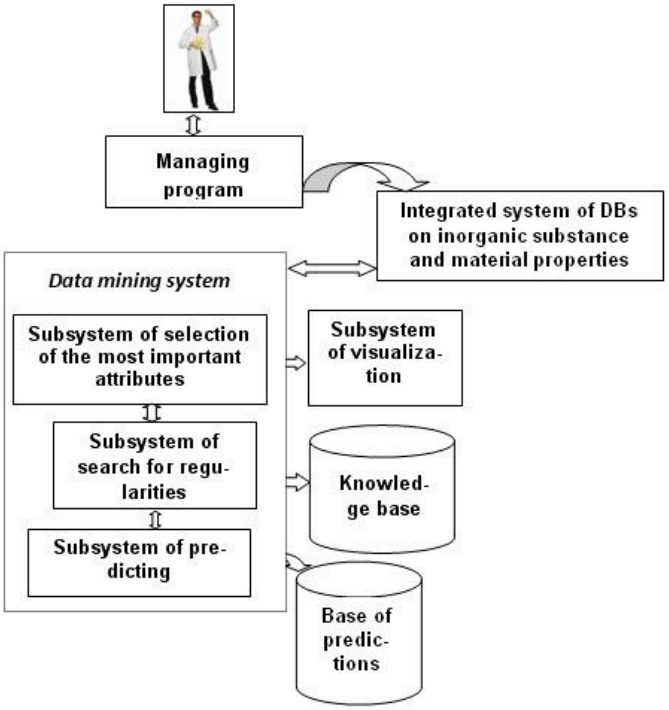

Fig. 1. The information-analytical system structure for inorganic compounds design.

3.1 Subsystems for searching for classifying regularities and predictions

In the development of this subsystem, the most important task was the selection of the most appropriate mathematical methods for searching regularities in chemical data. Typically, this task is performed by the trial-and-error method. In the selection of machine learning methods for analysis of chemical information, many years' experience in the application of these methods to inorganic compounds design was taken into account [18]. The following methods and programs have been selected:

- a wide range of algorithms of the Recognition multifunctional system, developed at the Computing Center of the Russian Academy of Sciences [16] and bringing together, in addition to well-known techniques, the algorithms of pattern-recognition (based on calculation of estimates), voting algorithms based on deadlock tests, voting algorithms based on logical regularities, weighted statistical voting algorithms, etc.;

— a ConFor computer system for training a computer in the procedure for concept formation [15], which is based on an original organization of data in the computer memory in the form of growing pyramidal networks.

As a rule, it is not possible to specify in advance which algorithm would be the most efficient for solving a particular problem. Therefore, it seems promising to apply the methods of prediction by algorithms ensembles. In a collective decision creation, the possible prediction errors of individual algorithms can be compensated in many cases by correct results of other algorithms. Based on this, we included programs that implement different strategies for collective decision-making, for example, the Bayesian method, methods using clustering and selection, decision templates, logical correction, the method of a convex stabilizer, the Woods dynamic method, committee methods, etc., [16] in the developed IAS [21].

3.2 Subsystem for searching the classifying properties of components

For the selection of informative properties of the chemical compound's components, we included programs based on algorithms [22-24] in the IAS. The selection of the properties of the components, the most informative for the classification of substances, has a double meaning. On the one hand, it drastically decreases the volume of the information analyzed, which for multicomponent substances comprise hundreds of properties of elements and simpler compounds, as well as functions of these properties. On the other hand, the selection of properties of the components most important for the classification of chemical substances, enables the physical interpretation of the classifying regularities, which enhances the credibility of the predictions obtained and finding substantial causal links between the parameters of the objects and the development of the physical and chemical models of phenomena.

3.3 Visualization subsystem

This subsystem facilitates the results interpretation, which constructs the projections of the points corresponding to the compounds in two-dimensional spaces of the properties of components, including not only the initial parameters but also user-specified algebraic functions of these parameters.

3.4 Knowledge and prediction bases

The knowledge base contains the obtained classifying regularities. The prediction base contains the results of previous computer experiments, as well as links to operation information stored in the knowledge base. Using the prediction base helped to improve the functionality of the databases on the properties of inorganic substances and materials, developed at the IMET, by providing the user with not only known data about already studied substances but also predictions for inorganic compounds not yet synthesized and evaluations of their properties.

3.5 Management subsystem

The management subsystem organizes the computing process, ensures interaction between the functional subsystems of the IAS, and provides access to the system on the Internet. In addition, the management subsystem provides the expert with software for data preparation for analysis, outputting reports, and implementation of other service functions. In particular, we developed a special subsystem to retrieve information from the database, which, after evaluation of the expert, is used to learn the computer, and to prepare it for further analysis. It gives the expert the capability to edit the found information and to form training samples for analysis. In the latter case, the expert marks only the selected properties of the components in a special table (menu), and the subsystem for the sample preparation for analysis retrieves the selected property values from the Elements database. If needed, the algebraic functions of the initial properties are formed in the subsystem and the description of the compounds is assembled in the form of an Excel table, which is then input to the prediction subsystem. The subsystem of result delivery is intended to make predictions in a tabular form conventional to chemists and materials scientists.

4 Use the IAS for predicting new compounds and evaluation of their properties

The machine learning application allowed a search for inorganic compounds formation regularities, a prediction of thousands not yet synthesized substances and some their properties evaluation using obtained regularities. This approach efficiency to inorganic compounds design can be illustrated by comparison of the predictions results with newer experimental data obtained after publication of our predictions.

4.1 Prediction of the TiNiSi crystal structure type for compounds with the composition ABAl

The equiatomic aluminides are of interest for the search for new magnetic materials. Thirty years ago, the prediction of new compounds of this type was carried out by us [25]. The algorithm based on the growing pyramidal networks learning (GPNL) [15] was used in the search for the criteria of this crystal structure type formation at ambient conditions. The learning set contained 39 examples of the compounds ABAl (hereinafter, A and B are various chemical elements) with the TiNiSi crystal structure type and 57 examples of the compounds with the structures different from TiNiSi. The following properties of elements A and B (attributes) were chosen for description of intermetallics: the distribution of electrons in the energy levels of isolated atoms of the chemical elements, the first three ionization potentials, the metal radii by Bokii and Belov, the standard entropies of individual substances, the melting points, the number of complete electronic shells, the number of electrons in incomplete s-, p-, d- or f-electronic shells for the atoms of elements.

Table 1. Part of a table illustrating the prediction of the crystal structure type TiNiSi for compounds with the composition ABAl [25].

A B	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Ru	-	O	-	-	-	+	-	+	+	+	+	+	+	-	-
Rh	O	O	∅	∅	-	+	-	©	+	+	©	+	©	∅	-
Os	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ir	©	©	©	©	+	©	+	©	©	©	©	©	©	+	©
Pt	©	©	©	©	+	⊕	+	⊕	⊕	⊕	⊕	⊕	⊕	©	

Table 1 is a result of comparing the predictions for each sets of properties and contains the comparison results of our predictions with newer experimental data. The following notations are used: + - prediction of the TiNiSi crystal structure type; - - prediction of the absence of the TiNiSi crystal structure type; ⊕ - a compound with composition ABAl has the TiNiSi crystal structure type and this fact was used for machine learning; © - prediction of the TiNiSi crystal structure type was confirmed experimentally; O - prediction of the crystal structure type different from TiNiSi was confirmed experimentally; ∅ - prediction of the crystal structure type different from TiNiSi was not confirmed experimentally; here and in other Tables the blank spaces correspond to the disagreement of the predictions with the use of different sets of the component's properties; all data and predictions are given for the substances under ambient conditions. A comparison of our predictions with newer experimental data has shown that the prediction error is lower than 12 %.

4.2 Design of compounds with composition ABX_2 (X – S, Se, or Te)

The chalcogenides with composition ABX_2 are a class of compounds that is promising for the search for new semiconducting and nonlinear optical materials. Taking into account the perspective of these compounds practical use the design of their not yet synthesized analogues was made [26]. Previously we predicted new compounds of this composition [27] also.

The task solution was subdivided into to stages: (1) prediction of the formation of compounds with composition ABX_2; and (2) prediction of the crystal structure type of these compounds under ambient conditions.

Prediction of the formation of compounds with composition ABX_2. Data on 667 examples of the formation of ABX_2 (X = S, Se, or Te) and 504 examples of the absence of this composition compounds in the systems $A_2X–B_2X_3$ and AX–BX under ambient conditions were used for machine learning. The data were taken from the DB Phases. 84 properties of the elements A, B, and X, whose values were taken from the DB Elements, were used for the compounds representation in the computer memory.

For the data analysis, several machine learning algorithms that are included into IAS were used. The learning quality was estimated on the basis of examination recognition in the mode of cross-validation. The analysis of the results using various algorithms has

shown that the best predictions under cross-validation have been obtained for the decision tree method (DT) [16] (accuracy of prediction being 72%), the logical regularities voting algorithm (LoReg) [16] (accuracy of prediction being 67.3%), and the deadlock test algorithm [16] (accuracy of prediction being 67.6%). These algorithms have been used for collective decisions using the committee method, in which the resulting prediction is calculated as an average arithmetic value of predictions obtained using different algorithms [16]. Using this procedure, the compound's formation predictions were obtained.

Prediction of the crystal structure type of compounds ABX_2. Data on 158 examples of the formation of ABX_2 with the crystal structure under ambient conditions α-$NaFeO_2$, 44 compounds with NaCl structure, 47 compounds with chalcopyrite structure, and 24 compounds with TlSe structure were used for machine learning. The same properties were used for the compound's representation.

The problem was solved in two ways. In the first case, multi-class learning and prediction, where the cumulative information on the four above-mentioned crystal phases has been used, was applied. In the second case, four problems of the dichotomy were solved - division into two classes, e.g., class 1, compounds with chalcopyrite crystal structure, and class 2, compounds with another structure. The results of predictions were compared, and a decision was made if the predictions obtained by multi-class prediction and dichotomies did not contradict each other. The results are summarized in Table 2.

Table 2. Part of a table illustrating the prediction of the crystal structure type for compounds with the composition ABX_2 (X – S, Se, or Te) [26].

X	**S**								**Se**								**Te**							
A B	**Li**	**Na**	**K**	**Cu**	**Rb**	**Ag**	**Cs**	**Tl**	**Li**	**Na**	**K**	**Cu**	**Rb**	**Ag**	**Cs**	**Tl**	**Li**	**Na**	**K**	**Cu**	**Rb**	**Ag**	**Cs**	**Tl**
B	#5	#5	#5	©3	#5	3	#5	#5	1	1		#5	1	3	1	4	1	1	4		1		1	4
Al	#5			#3		#3		©4	#5	#4	#5	#3		#3		#5		#4	#4	#3		#3	4	#4
Sc	#1	#1	©1	#5	©1			#1	1	1	1		1	#5	1		1		1		1		1	
Cr	#5	#1	#1	#5	#1	#5		#5	$2	#1	©1	#5	#1	#5	5				5	3	5	#5	5	#5
Fe	#5		#5	#3	#5	#3		#5		1	#5	#3	#5	#3		#5			1	#3	1	#3	1	©5
Ga	#5	#5	#5	#3	#5	#3	©5	#5	©5	4	#5	#3	4	#3		#5	#3	#4	#5	#3		#3		#4
As		#5		#5		#5	5	#5		#5	#5	#5	#5	#2	#5	©5	2		4	#6	$1	#2	5	
Y	#2	#1	#1	©5	©1	#5	5	#1	#1	#1	1		1	#5	1	#1	1	1		#5	1	#5	1	#1
In	#5	#1	#5	#3	#5	#3	#5	#4	#5	#1	©5	#3	#5	#3		#4	#3	#4	#4	#3		#3	$4	#4
Sb		©5	#5	#5	#5	$2	©5	©2	©2	©5	#5	#5	#5	#2		©5	#2	#2	4	#5	#5	#2	#5	#1
La		#2	#1		#1		#1	#6	#5	#1	©1	#5	#1		1	#6	2		©1		1		1	6
Ce		#2	#1	#5	#1		#1	#6	#5	#1	1	#5	#1	#6	1	#6			©1		©1		1	#6
Pr	#2	#1	#1	#5	#1		#5		1	#1	1	#5	#1	#6	1	#1	2		©1	#5	1		1	#1
Nd	#2	#1	#1	#5	#1		#5			#1	1		#1		1	#1	1	1	©1	#5	©1		©1	#1

X	S								Se								Te							
A B	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl
Pm		1	1	5	1	5		1	1	1	1		1		1	1	1	1			1		1	
Sm	#2	#1	#1	#5	#1	#5	#5	#1		#1	©1	#5	#1	#6	1	©1		©1	©1		©1		1	#1
Gd	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1	#5	1	#1	1	1	©1		1	#5	1	#1
Tb	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1		1	#1	1	1			1	©5	1	#1
Ho	#1	#1	#1	#5	#1	#5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Er	#1	#1	#1	#5	#1	©5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1	#1		1	#5	1	#1
Tm	#1	#1	#1	#5	#1	#5	#5	#1	1	1	1		1	#5	1	#1	1	1			1	#5	1	#1
Yb	#1	#1	#1	#5	#1	#5	#5	#1		#1	#1		1	#5	#5	©1	2	1	1	#5	1		1	
Lu	#1	#1	#1	#5	#1	#2	#5	#1	1	1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Bi	#2	#2	#2	#5	#1	#1		#1	#2	#2	#2	#2	#5		#5	#2	#2	#2		#5		$1		#1

In Table 2, the following notations were used: 1 – prediction of the structure of the α-$NaFeO_2$ type; 2 – prediction of the structure of the NaCl type; 3 – prediction of the structure of the chalcopyrite type; 4 – prediction of the structure of the TlSe type; 5 – prediction of the structure different from the ones mentioned above; 6 – prediction of the absence of ABX_2; the symbol # is used for objects for the machine learning; © - predictions was confirmed experimentally; $ - predictions was not confirmed experimentally.

40 compositions have been experimentally tested and only in five cases the predictions turned out to be incorrect, i.e., the prediction error was about 12.5 %. Beyond that the melting point and bandgap were evaluated for compounds with the chalcopyrite crystal structure type [28].

From ternary to quaternary compounds. Prediction of the crystal structure type of compounds A_2BCHal_6. Searching for and studying halide compounds having the composition A_2BCHal_6 (Hal = F, Cl, Br, or I) with the elpasolite crystal structure type is related to the development of new luminescent, laser, and magnetic materials.

The set for computer-assisted analysis included information about 289 ($A \neq C$) compounds having the elpasolite structure; 20 compounds with Cs_2NaCrF_6 type of crystal structure; 57 compounds with crystal structures another than the ones given above under ambient conditions; and 81 $AHal–BHal_3–CHal$ systems where compounds are not formed [19]. The 134 properties of chemical elements A, B, C, and Hal were included in the initial set of component parameters.

The problem of predicting new halo-elpasolites included solving three intermediate tasks. Formation of compounds with composition A_2BCHal_6 was predicted in the first of them (task 1). The next task included searching for regularities and predicting the formation of compounds with given composition and the most common types of crystal structures (elpasolite or Cs_2NaCrF_6). The latter task was divided into two smaller ones. When solving the first of them, the multi-class prediction of belonging to four classes (elpasolites, compounds with the Cs_2NaCrF_6 structure, compounds with the structure

different from those shown above, and the systems containing no compounds with composition A_2BCHal_6 (task 2)) was performed. Next, halide systems were consecutively divided into three classes: the target class, e.g., 1 - elpasolites; class 2 - compounds with non-elpasolite structure; and class 3 - the AHal–$BHal_3$–CHal systems containing no compounds with composition A_2BCHal_6 (task 3). The final decision regarding the class that a compound being predicted belongs to, was made by comparing the predictions obtained when solving all three tasks. If the results were inconsistent, the prediction was regarded to be uncertain and the prediction table cell was left empty.

The algorithms LoReg, artificial neural network learning (ANN), K-nearest neighbor (KNN), and support vector machine (SVM) ensure the best accuracy of prediction of compound formation (task 1) in the cross-validation mode and the collective decision-making software based on the algorithm of generalized polynomial corrector [16] provided the best estimate for prediction accuracy, namely 95%.

When solving task 2 of multi-class prediction, the set of algorithms including DT, KNN, SVM, ANN, learning a multilayer perceptron, and the algorithm of the convex stabilizer [16] for collective decision-making, ensured the best accuracy of examination prediction: 89%. When forming the regularity that allows one to demarcate elpasolites from compounds with differing crystal structures and from systems where no A_2BCHal_6 compounds are formed (task 3), the best accuracy (80%) was provided by the set of algorithms that included the algorithms LoReg, ANN, KNN, SVM, and the Bayesian method of collective decision-making [16].

Some results of comparing the predictions found by solving all three classification tasks are summarized in Table 3. The following notations are used: 1 - prediction of compounds with the elpasolite crystal structure; 2 - prediction of compounds with the Cs_2NaCrF_6 structure type; 3 - prediction of compounds having crystal structure another than the abovementioned ones; and 4 - prediction of the absence of an compound in the ACl–BCl_3–CCl system; the # symbol is used to denote previously studied compounds; the information about them was used for machine learning.

Table 3. Part of a table illustrating the prediction of the crystal structure type for compounds with the composition A_2BCCl_6 [19].

C	**Li**					**Na**					**K**				**Rb**			
A B	**Na**	**K**	**Rb**	**Cs**	**Tl**	**Li**	**K**	**Rb**	**Cs**	**Tl**	**Li**	**Na**	**Rb**	**Cs**	**Li**	**Na**	**K**	**Cs**
Al	4	#4	4	4	1	4	#4	4	#4	1	#4	#4	4	4	4	4	4	4
Sc		#3	#1	#2	#1			#1	#1	#3		4	1	#1	4	4		#4
Ti		3	3	3			1	1	#1	1	4			1	4	4	3	1
V		3	#3	#3			#1	#1	1	1	4	4		1	4	4		1
Cr			3	#3			#1	#1		#1	4		#3	#1	4	#4	3	3
Fe	4	#4		3		4	#4		#1		#4	#4			4	4	4	#4
Y			#1	#1			#4	#3	#1	3	4	#4	1	#1	4	4	1	1
In			#1	#3	3		1	#1	#3	3			3	#3	4			3
La	#4	#4		#1		#4	4		#1		#4	4	1	1	4	4	1	1
Ce	4	#4	1	#1		4	4		#1		#4	4	1	1	4	4	1	1
Pr	#4	#4	#4	#1		#4	#4	#4	#1		#4	#4	1	#1	#4	#4	1	1

C	Li					Na					K				Rb			
A B	Na	K	Rb	Cs	Tl	Li	K	Rb	Cs	Tl	Li	Na	Rb	Cs	Li	Na	K	Cs
Nd	4	4	4	#1		4	#4	4	#1		4	#4	1	#1	4	4	1	1
Pm			3	1				3	1		4	4	1	1		4	1	1
Sm	4		#3	#1		4	#4		#1			#4	1	#1	4	4		1
Eu		3	#3	#1	3			#3	#1	3		4	1	#1		4	1	1
Gd		3	#3	#1	3			#3	#1	3	4		1	#1	4	4		1
Tb		3	#1	#1	3			3	#1	3			1	#1		4	1	1
Dy		3	#1	#1	3		#4	#3	#1	3		#4	1	#1		4	1	1
Ho		3	#1	#1	3			#3	#1	3			1	#1		4	1	1
Er		3	#1	#1	3			#3	#1	3			1	#1		4	1	1
Tm		#3	#1	#1	#3			#3	#1	#3			1	#1		4	1	1
Yb		3	#1	#1	#3		3	#3	#1	3			1	#1		4	1	1
Lu			#1	#3	#3		3	#3	#1	3			1	1	4	4	1	1
Tl		1	1				1	1	#1		4		1	1	4	4	1	1
U	#4	#4	4	#1		#4		#3	#1		#4	4		1	4	4		1
Pu	4	4		1		4	4	3	#1		4	4	1	1	4	4		1

5 Conclusions

During half of the century the predictions of thousands of inorganic compounds in binary, ternary and more complicated chemical systems were obtained and some their properties (melting point, critical temperature of superconductivity, band gap energy, etc.) were estimated in IMET. The obtained predictions usage allows an essential progress provision in a search for new magnetic, semiconductor, superconductor, nonlinear optical, electro-optical, acousto-optical and other materials. Hundreds of predicted compounds were synthesized and our results experimental verification shows that the average prediction accuracy is higher than 80%. Machine learning methods application to search for regularities in big chemical data of DB PISM gives an opportunity for theoretic design of new inorganic compounds that allows substantially reduce the costs for search for new materials with predefined properties, replacing them by computations. It is important to note that only information on components properties (chemical elements or more simple compounds) is used in prediction process.

This work was partially supported by the Russian Foundation for Basic Research (project nos. 17-07-01362 and 18-07-00080) and the State task № 075-00746-19-00. We are grateful to V.V. Ryazanov, O.V. Sen'ko, and A.A. Dokukin for long-term help and collaboration.

References

1. Kiselyova, N.N., Dudarev, V.A., and Zemskov, V.S.: Computer information resources in inorganic chemistry and materials science. Russ. Chem. Rev., 79(2), 145-166 (2010).
2. ACS Publications Homepage, https://pubs.acs.org/, last accessed 2019/04/22.
3. ScienceDirect Homepage, https://www.sciencedirect.com/, last accessed 2019/04/22.
4. Springer Nature Homepage, https://link.springer.com/, last accessed 2019/04/22.
5. Wiley Online Library Homepage, https://onlinelibrary.wiley.com/, last accessed 2019/04/22.
6. Kiselyova, N.N., Dudarev, V.A., Stolyarenko, A.V.: Integrated system of databases on the properties of inorganic substances and materials. High Temperature, 54(2), 215-222 (2016).
7. NIST Data Gateway, https://www.nist.gov/srd, last accessed 2019/04/22.
8. NIMS Materials Database (MatNavi) Homepage, http://mits.nims.go.jp/index_en.html, last accessed 2019/04/22.
9. SpringerMaterials Homepage, https://materials.springer.com/, last accessed 2019/04/22.
10. Blokhin, E., Villars, P.: The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome. In: Andreoni, W., Yip, S. (eds.) Handbook of Materials Modeling, pp.1-26. Springer, Heidelberg (2018).
11. IMET RAS DBs Homepage, http://www.imet-db.ru/, last accessed 2019/04/22.
12. DB IRIC (Information Resources on Inorganic Chemistry) Homepage, http://iric.imet-db.ru/, last accessed 2019/04/22.
13. Savitskii, E.M., Devingtal', Yu.V., and Gribulya, V.B.: Prediction of metallic compounds with composition A_3B using computer. Doklady Physical Chemistry, 183(5), 1110-1112 (1968).
14. Savitskii, E.M., Gribulya, V.B.: Application of computer techniques in the prediction of inorganic compounds. Oxonian Press Pvt., Ltd., New Delhi-Calcutta (1985).
15. Gladun V.P.: Processes of formation of new knowledge. SD "Pedagog 6", Sofia (1995).
16. Zhuravlev, Yu.I., Ryazanov, V.V., and Sen'ko, O.V.: RECOGNITION. Mathematical methods. Software system. Practical solutions. Phasis, Moscow (2006).
17. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. J. Machine Learning Research 12(Oct.), 2825-2830 (2011).
18. Kiselyova, N.N.: Komp'yuternoe konstruirovanie neorganicheskikh soedinenii. Ispol'zovanie baz dannykh i metodov iskusstvennogo intellekta (Computer Design of Inorganic Compounds: Use of Databases and Artificial Intelligence Methods). Nauka, Moscow (2005).
19. Kiselyova, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: Prediction of New Halo-Elpasolites. Russ. J. Inorg. Chem. 61(5), 604-609 (2016).
20. Dudarev, V.A.: Information systems on inorganic chemistry and materials science integration. Krasand, Moscow (2016).
21. Kiselyova, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: A system for computer-assisted design of inorganic compounds based on computer training. Pattern Recognition and Image Analysis, 21(1), 88-94 (2011).
22. Senko, O.V.: An Optimal Ensemble of Predictors in Convex Correcting Procedures. Pattern Recognition and Image Analysis. 19(3), 465-468 (2009.
23. Yuan, G.-X., Ho, C.-H., Lin, C.-J.: An Improved GLMNET for L1-regularized Logistic Regression. J. Machine Learning Research. 13, 1999-2030 (2012).
24. Yang, Y., Zou, H.: A Coordinate Majorization Descent Algorithm for L1 Penalized Learning. J. Statistical Computation & Simulation. 84(1), 1-12 (2014).

25. Kiseleva, N.N., Burkhanov, G.S.: Search for new ternary phases with Al, Ga, and In using an information-prediction system. Russian Metallurgy, 1, 223-226 (1989).
26. Kiselyova, N.N., Podbel'skii, V.V., Ryazanov, V.V., Stolyarenko, A.V.: Computer-aided design of new inorganic compounds with composition ABX_2 (X = S, Se or Te). Inorg. Mater.: Applied Researches. 1(1), 9-16 (2010).
27. Savitskii, E.M., Kiseleva, N.N.: Cybernetic prediction of the existence of ABX_2 phases. Inorg. Mater. 15(6), 866-868 (1979).
28. Kiselyova, N.N., Stolyarenko, A.V., Gu, T., et al.: Computer-aided design of new inorganic compounds promising for search for electronic materials. In: Proc. The Sixth Int.Conf. on Computer-Aided Design of Discrete Devices (CAD DD 07). vol. 1, pp. 236-24.2 UIPI NASB, Minsk (2007).