Topological Data Analysis in Materials Science: The Case of High-Temperature Cuprate Superconductors

I. Yu. Torshin[a],* and K. V. Rudakov[a],**

[a] *Dorodnicyn Computing Centre, Federal Research Center "Informatics and Control," Russian Academy of Sciences, ul. Vavilova 44, Moscow, 119333 Russia*

**e-mail: tiy1357@yandex.ru*

***e-mail: rudakov@ccas.ru*

Abstract—Adequate formalization of problems is the most important task that has to be solved in order to apply the modern methods of so-called "machine learning" to real problems. The effective application of the metric, logical, regression, and other algorithms of machine learning becomes possible only when feature generation procedures and classes of objects are adequately defined. In this study, the theory of topological analysis of poorly formalized problems and the theory of analysis of labeled graphs were applied to the problem of predicting numerical characteristics of crystalline materials. The methods developed were tested on the problem of predicting the critical temperature of superconducting transition (T_c) of high-temperature cuprate superconductors (1450 structures). As a result, in a tenfold 6 : 1 cross-validation, the best model with a linear recognition operator yielded quite high average value of the correlation coefficient ($r = 0.77$) between the predicted and experimentally determined values of T_c.

Keywords: algebraic approach to pattern recognition, theory of analysis of labeled graphs, data mining, superconductivity, materials science, solid state physics

DOI: 10.1134/S1054661820020157

1. INTRODUCTION

Poorly formalized problems (for which, by definition, there are no unambiguous methods for defining objects, features, and classes) are widely presented in biomedicine, chemoinformatics, bioinformatics, solid state physics, applied linguistics, and in other fields of modern science. The formalization of such a problem can be visualized as a transition from the set of initial descriptions of physical objects to a particular topology, then to a lattice, and then to a certain metric space [1].

Within the algebraic approach to the solution of the pattern recognition/classification/prediction problems, the formalization of a problem provides adequate definitions of the set of *initial information* (I_i) and the set of *final information* (I_f) [2–7]. Such definitions, in turn, allow one to form *a set of precedents* as a subset of the product ($I_i \times I_f$) and then to apply the algebraic approach to the solution of recognition/classification/prediction problems [2, 3]. Transition from the original descriptions of physical objects to the sets of precedents (i.e., to the subsets of $I_i \times I_f$) that are based on the topological approach to the formalization of problems [1] allows one to construct and "train" algorithms for solving problems $A(\theta) : I_i \to I_f$, where θ is the vector of internal parameters of the algorithm.

The algebraic approach necessarily implies the study of the sets of precedents $\mathrm{Pr} \subset I_i \times I_f$ and of the algorithms $A(\theta) : I_i \to I_f$ in respect to the fundamental properties of *solvability/regularity* of problems Z(Pr) and *correctness/completeness* of the corresponding algorithmic models $\{A(\theta)\}$. To this end, one applies the factorization [4] and the metric [5] approaches to the analysis of poorly formalized problems, which involve, in particular, the analysis of the compactness properties of the subsets of metric configurations [6, 7]. Obviously, the choice of the definitions of the sets I_i and I_f should be maximally adequate to the problem under study, because this naturally determines the performance quality of the algorithms $A(\theta) : I_i \to I_f$ developed within the so-called paradigm of "machine learning."

In this study, we apply a set of theoretical methods developed earlier for the analysis of poorly formalized problems to the problem of predicting the properties of crystalline materials. To this end, we use the concept of a chemograph—a special kind of a labeled graph—to describe the chemical structures of molecules and crystals [8, 9]. We introduce special types of labeling chemographs, "χ-chains," and "χ-nodules," on the basis of which we construct methods for the numerical evaluation of the similarity of structures and methods of feature generation. To the feature descriptions of chemical structures, we apply algorithms for predicting numerical variables [10]. The methods developed

Received August 20, 2019; revised August 20, 2019; accepted October 15, 2019

here will be tested on a sample of crystalline structures of high-temperature superconductors (HTSCs) with a view to predicting the critical temperature of superconducting transition (T_c). A brief description of the problem area is presented in the next section.

2. PHYSICAL MODELS OF HIGH-TEMPERATURE SUPERCONDUCTIVITY

The phenomenon of superconductivity has been known since the beginning of the 20th century. However, the fundamental physical principles of HTSCs are still one of the key problems of solid state physics. To date, the most fundamental theory of superconductors, the Bardeen–Cooper–Shriffer (BCS) theory, involves the formation of the so-called "bosonic condensate" (Bose–Einstein condensate). Bosons are particles with integer spin, an unlimited number of which can occupy the same quantum state, i.e., particles that form a bosonic condensate, which explains superconductivity. However, the BCS theory in its original strict version fails to explain the phenomenon of cuprate superconductors ($La_{2-x}Ba_xCuO_4$ and others), which are characterized by much higher values of superconducting transition temperature (Tc) than the theoretical limit of 30 Kelvin.

Further development of theoretical views on the phenomenon of superconductivity proceeded in several directions. First, a search is made for "intermediate" forms of fermions (particles with fractional spin such that only one particle can occupy a single quantum state) that are characterized by some properties of bosons. Such properties include, for example, the existence of antiparticles—particles with the same mass and spin but with opposite signs of all other interaction parameters (charges, etc.). A boson can be an antiparticle of itself, but a fermion, as a rule, is not. The researcher Ettore Majorana proposed a theoretical description of fermions that are antiparticles of themselves. Although such fermions have not been found experimentally, the so-called "Majorana states of fermions" are associated with one-dimensional models of superconductivity [11, 12] (see further). In [13], models of composite fermions were developed in which attempts were made to associate superconductivity with antiferromagnetic fluctuations that induce interactions between quasiparticles near the Fermi surface [14].

Second, there have been attempts to apply various definitions of bosons as "quasiparticles." For example, one defines polaron quasiparticles, each of which consists of an electron and a phonon (crystal lattice vibration). A *bipolaron*, a quasiparticle composed of polarons (literally, "*two polarons bound together by a phonon interaction*"), is a boson; it can form a Bose condensate and is characterized by translational invariance (i.e., can be described as a plane wave in a crystal lattice). In [15], Prof. Lakhno showed that translation-invariant bipolarons can form a Bose condensate even in one-dimensional systems (i.e., in the chains of sequentially interacting atoms). In this case, superconductivity is associated with the presence of "stripes," some local one-dimensional deformations (several nanometers long) of the crystal lattice, within each of which a superconducting bosonic condensate (a quantum mechanical unified wave–particle delocalized along the stripe) is formed.

Note that the data of individual experiments indirectly confirm the possibility of the existence of such "one-dimensional superconductivity" [16]. For example, X-ray scattering studies of the $YBa_2Cu_3O_{6.67}$ HTSC have shown that a superconducting transition occurs simultaneously with the suppression of the so-called charge density waves during one-dimensional compression of the crystal along the "a" axis of the unit cell of the crystal [17]. Possible evidence for the existence of one-dimensional "pair density waves" (polarons) was obtained by studying the Bi-2212 cuprate HTSC with the use of tunneling spectroscopy [18]. The studies of the thin-film single-crystal HTSC $La_{2-x}Sr_xCuO_4$ reveal the in-plane anisotropy of electron transfer upon cooling the crystal to T_c [19]. Since electric current flows in one direction, the existence of the in-plane anisotropy corresponds precisely to certain one-dimensional processes within the HTSC crystal.

Third, in the theoretical analysis of the phenomenon of HTSC, one can apply approaches borrowed wholesale from other fields of solid state physics. For example, superconductivity can be interpreted as a special case of the theory of metallic conductors and correspond to some "quasimetallic state" [20, 21]. Also, HTSC might be considered as a particular case of semiconductor theory and is interpreted as the result of some abstract "complex changes" in the Brillouin zone [22, 23], or even within the framework of simpler phenomenological models of semiconductors as the electron–hole model [24]. In all these directions, some quantum-mechanical calculations are used as a tool for argumentation. However, it is well known that the practical application of the available methods of quantum mechanics to calculations of predictive sort inevitably involves the introduction of a substantial set of assumptions, such as the use of only atoms with 100% occupancy, and is possible only for small structures (with at most 30 atoms in the unit cell) or for atoms without *f*-electrons, etc. [25].

Thus, despite the existence of several theoretical approaches to the phenomenon of HTSC, modern physics provides rather a limited toolbox for real predictions of the superconducting properties (as well as of many other properties) of crystalline materials. None of the existing approaches has been developed to an extent that would allow, for example, the calculation of the temperature T_c for an arbitrary crystal structure. Therefore, within the algebraic approach of

the scientific school of Academician Yu.I. Zhuravlev, materials science is a good example of the so-called "problem area" [26], in which mathematical methods can be applied to recognizing, classifying, and predicting the properties of objects by their feature descriptions.

Recognition/classification methods developed within the so-called paradigm of "machine learning" by precedents are becoming more and more popular in materials science [27], which is facilitated by the availability of the relevant databases [28]. For example, the algebraic approach was earlier applied to the classification of promising semiconductors with the general formula AB2X4 (X = S, Se, Te). Information on the properties of AB2X4 compounds, supplemented with feature descriptions of the individual elements (atomic radii, enthalpy of evaporation, melting and boiling points, ionization potentials, electronegativity, etc.), was used to form the initial sets of precedents. Then, the classification algorithms that allow the prediction of the existence of appropriate compounds and of the class of the crystal structure of the compounds (Heusler phases and compounds with structural classes of the type $YSiRh_2$, $MgCuAl_2$, $YSiPd_2$, etc.) were constructed [29]. The methods of graph theory can also be used to predict the properties of insulators and metalloids (semimetals) [30, 31], one-dimensional subgraphs can be used to theoretically substantiate the oxidation degree [32], and so on.

This brief overview of theoretical ideas regarding the HTSC phenomenon and the problems of materials science allows us to formulate several important conclusions:

1. Currently, there is no universal, unified theory of the properties of materials (including the properties of HTSCs), which would allow one to make nontrivial verifiable predictions. In particular, no theories have been proposed that would allow one to calculate the temperature Tc on the basis of an arbitrary crystalline structure, the oxygen "doping" level of a material, and some other parameters of the material.

2. Some of the existing theoretical views on superconductivity admit a one-dimensional character of the HTSC phenomenon.

3. The practical application of quantum-mechanical calculations for predicting the properties of materials (including HTSCs) is limited by many additional assumptions. As a rule, these calculations cannot be used for large-scale screening predictions of the properties of hundreds of thousands of materials in databases. In addition, these calculations do not allow the prediction of the values of parameters that are of interest to a practical researcher in materials science (the values of the critical temperature T_c, etc.).

4. Methods of machine learning, including those using the elements of graph theory, can be successfully used to predict the properties of various crystalline phases.

Note that items 1, 3, and 4 relate not only to the problem of predicting the properties of HTSCs, but also to almost any problem of modern materials science. The lack of a unified theory (item 1), significant limitations in the use of universal methods of physics (item 3), and a certain success in machine learning methods (item 4) allow one to consider materials science as a classical "problem area" containing a significant number of poorly formalized problems [1]. Accordingly, various tools of algebraic theory of recognition and, in particular, topological [1], combinatorial [4], and metric [5] approaches to the formalization of problems, to the generation of feature descriptions, and to the prediction of values of numerical features [10] can be used for predicting the properties of crystalline materials. The possible one-dimensional character of the HTSC phenomenon (item 2) serves as a basis for the application of the methods of the theory of chemograph analysis, including chemoinvariants based on "chains" and "nodules" of atoms [8, 9] (see the definitions below).

3. INITIAL DEFINITIONS

The topological theory of pattern recognition and classification [1–9] comprises the properly topological [1, 2, 4], the combinatorial [4, 5, 8, 9] and the metric [6, 7, 9] approaches to the analysis of the ill-stated problems. The first step of the application of the topological theory of pattern recognition is to define the primary feature descriptions, which are then used for generation of the relevant topologies and lattices [1], of the problem-specific metrics [7] and, then, of the "synthetic" numerical features of 2nd level, 3rd level etc [10].

In the case of molecules and crystalline substances the present formalism uses the concept of chemograph to introduce the primary features. A chemograph is a special graph for describing the chemical structure of substances. The need to introduce this concept has been associated with the fundamental features of the chemical structure of substances, which are well known in the quantum approach of molecular orbitals and in the chemical bond theory [8, 9].

Definition 1. A *graph G* is a set of vertices $V = V(G)$ plus a set of edges $E = E(G)$, $E \subset V^2$.

The set $\mathbf{\Gamma} = \{(V, E) \mid V \subset \mathrm{N}, E \subset \mathrm{N}^2\}$, which is the set of all subgraphs of an infinite complete graph $\mathrm{G} = (\mathrm{N}, \mathrm{N}^2)$, is the set of all possible graphs (N is the natural series).

Definition 2. A *chemograph* (*χ-graph*) is a finite, connected, nonoriented, labeled graph without loops.

The vertex set of a chemograph X is isomorphic to the set of atoms of a substance, and the set of edges, to the set of chemical bonds of a substance. Depending on the variant of the formalism, a matrix $M(X) = \{m_{ij}(X)\}$, $m_{ij} \in \mathrm{R}$, $i, j = 1,\ldots,|\mathrm{V(X)}|$, can (1) be an adja-

cency matrix, (2) contain the multiplicities of chemical bonds, or (3) contain interatomic distances.

Define the *set of all closed subgraphs* $\mathbf{\Pi}(X)$ of a chemograph X(V,E) as $\mathbf{\Pi}(X) = \{(\mathbf{v},\mathbf{e}) | \mathbf{v} \subseteq V,\ \mathbf{e} \subseteq E,\ \bigvee^{\mathbf{e}}(v_1, v_2) : v_1 \in \mathbf{v},\ v_2 \in \mathbf{v}\}$. The *union of the set of subgraphs* Π is a subgraph $\breve{\Pi}$, $\breve{\Pi} = \bigcup_{i=1}^{|\Pi|} \pi_i$; the *intersection* $\hat{\Pi}$ *of the set of subgraphs* Π is $\hat{\Pi} = \bigcap_{i=1}^{|\Pi|} \pi_i$. If the set of subgraphs $O \subseteq \mathbf{\Pi}(X)$ is such that $\breve{O} = X$, then $O \subseteq \mathbf{\Pi}(X)$ is called a *generator* of the chemograph X.

Definition 3. A *chain* $\langle v_0, v_1 \rangle$ between vertices v_0 and v_l is a route $\langle v_0, v_l \rangle = (v_0, e_1, v_1, e_2, v_2, \ldots, e_l, v_l)$, $\bigvee_{i=0}^{l-1} \bigvee_{j=i+1}^{l} v_i \neq v_j$, in which all vertices are different.

According to Definition 3, we calculate the set of all chains of a chemograph X as $\mathbf{C}(X) = \{c = (V_c, E_c) | c \in \mathbf{\Pi}(X),\ |E_c| = |V_c| - 1 > 0,\ \bigvee_{i=1}^{|Vc|} d(c, v_i) \leq 2,\ \bigvee_{i=1}^{|Vc|-1} \bigvee_{j=i+1}^{|Vc|} v_i \neq v_j\}$, where $d(c, \mathrm{v}_i)$ is the degree of the vertex v_i in the chemograph c. It is obvious that $\breve{\mathbf{C}}(X) = X$ (Theorem 6 in [8]). Accordingly, the structure of an arbitrary chemograph X can be described by a set of chains. The generators of X can be given by various subsets of $\mathbf{C}(X)$, for example, by sets of chains of fixed length, sets of chain that include a certain set of vertices, and so on.

Define an *operator* $\hat{c}$ *for calculating chains whose endpoint is a given vertex,* $\hat{c}v_i = \{\langle v_i, v \rangle \in \mathbf{C}(X), \langle v, v_i \rangle \in \mathbf{C}(X)\}$, and an operator $\hat{c}^n$ for *chains of length n* as $\hat{c}^n v_i = \{c | c \in \hat{c}v_i, |c| = n\}$, so that $\bigcup_{i=1}^{n_{\max}} \hat{c}^n v_i = \mathbf{C}(X)$, $n_{\max} = \max(|c|, c \in \hat{c}v_i)$. Then the following generators of the chemograph X are obvious: $\bigvee_{v \in V(X)} \breve{\hat{c}} v = X$ and $\bigcup_{v \in V(X)} \breve{\hat{c}}^n v = X,\ 2 \leq n \leq n_\chi,\ n_\chi(X) = \min_{v \in V(X)} \max_{c \in \hat{c}v} |c|$ (Theorems 7 and 8 in [8]). Denote $\mathbf{C} = \mathbf{C}(\mathrm{G})$.

Definition 4. A *connected subgraph* is a subgraph in which at least one elementary chain passes through any pair of vertices.

By Definition 4, the set of connected subgraphs of a chemograph X is defined as $\mathbf{S}(X) = \{\pi = (\mathbf{v}, \mathbf{e}) | \pi \in \mathbf{\Pi}(X),\ \bigvee_{i=1}^{|v|-1} \bigvee_{j=i+1}^{|v|} \exists c(V_c, E_c) \in \mathbf{C}(X) : (v_i, v_j) \subseteq V_c\} \subseteq V_c\}$. It is obvious that $\mathbf{C}(X) \subseteq \mathbf{S}(X)$, $\pi_1, \pi_2 \in \mathbf{S}(X)$, $\pi_1 \cap \pi_2 \neq \varnothing \Leftrightarrow \pi_1 \cup \pi_2 \in \mathbf{S}(X)$, $x \in \mathbf{S}(X) \Leftrightarrow x$ is a chemograph, and $\breve{\mathbf{S}}(X) = X$ (Theorems 4, 5, and 6 in [8], respectively).

Definition 5. *Nodules* are constructs of the form $(\Gamma(v), \hat{e}v)$, i.e., connected subgraphs consisting of the adjacency set $\Gamma(v) = \hat{v}\hat{e}v$ of a vertex v and all edges incident to v [8]. A *k-nodule* in a chemograph X is a subgraph consisting of a vertex v, *all k* edges incident to it, and all the other vertices incident to the k edges. All nodules of a chemograph X form a set $\mathbf{K}(X) = \{(\Gamma(v), \hat{e}v) | v \in V(X),\ d(v) > 1\}$, $\mathbf{K} = \mathbf{K}(\mathrm{G})$, which implies that $\breve{\mathbf{K}}(X) = X$ (Theorem 10 in [8]). Similarly, $\mathbf{K}'(X) = \{u | u \in \hat{\pi}\pi,\ \pi \in \mathbf{K}(X),\ |E(u)| = k,\ k > 1\} \cap \mathbf{S}(X)$.

According to Definition 1, chemographs are labeled graphs.

Definition 6. Suppose given a set of labels $Y = \{y_1, y_2, \ldots, y_{n(Y)}\}$. Then the *labeling* of a chemograph is performed by the corresponding vertex labeling function $\mu_V : V \to Y$. For a given label set, $\ddot{Y} = \bigcup_{n=1}^{\infty} Y^n$ is the set of all permutations over Y.

For most problems of chemoinformatics that deal with organic molecules, some relatively simple label sets based on the combinations of atom types (C, N, O, S), their charges, etc., are quite sufficient. In the case of chemoinformatics problems that arise in materials science, the label sets will be more complex, but can also be based on the combinations of atom types, oxidation degree, charge, and so on. In actual initial representations of substances (sets of internal coordinates of the unit cell of a crystal, sets of Cartesian coordinates), the types of atoms are always indicated, so that methods for determining the function μ_V are quite obvious to an expert in the problem area.

Definition 7. A χ-*chain* is an element α of the set of all χ-chains, $\alpha \in \tilde{Y} = \{\{y^1, y^2\}, y^1,\ y^2 \in \ddot{Y} \big| |y^1| = |y^2| = n, \forall i = 1 \ldots n : y_i^1 = y_{n-i+1}^2\}$.

Denote the set of all χ-chains of length n by $\tilde{Y}^n \subset \tilde{Y}$.

Definition 8. Let $\sigma \mathrm{Y}^k = \{\{y_1, y_2, \ldots, y_i \ldots, y_k\}, \bigvee_{i=1}^{k} y_i \in Y\}$. Then we define a χ-*k-nodule* as an element of the set $\hat{Y}(k) = \{Y \times \sigma Y^k\}$, and a χ-*nodule* κ, as an element of the set of all χ-nodules $\kappa \in \hat{Y} = \bigcup_{k=2}^{\max V} \hat{Y}(k)$, where maxV is the maximum possible valence of chemical elements in D.I. Mendeleev's periodic table.

Definitions 3, 7, and 8 of the sets $\mathbf{C}(X)$ and $\mathbf{K}(X)$ obviously imply the following lemma.

Lemma 1. $(\exists \mu_c : \mathbf{C} \to \tilde{Y}) \wedge (\exists \mu_\kappa : \mathbf{K} \to \hat{Y})$.

Corollary 1. $\exists \mu_c^{-1} \Rightarrow \bigcup_{v \in V(X)} \hat{c}v = \bigcup_{\alpha \in \tilde{Y}(X)} \mu_c^{-1}(\alpha) = X$.

Corollary 2. $\exists \mu_c^{-1} \Rightarrow \bigcup_{v \in V(X)} \hat{c}^n v = \bigcup_{\alpha \in \tilde{Y}(X), |\alpha| = n} \mu_c^{-1}(\alpha) = X$ *for* $n \leq n_\chi$ (*see the definition of the operator* $\hat{c}$).

Corollary 3. $\exists\mu_\kappa^{-1} \Rightarrow \bigcup_{\kappa\in\hat{Y}(X)}\mu_k^{-1}(\kappa) = X$ (*see Definition* 5).

According to Lemma 1, the set $\mathbf{C}(X)$ of chains of an arbitrary chemograph X is uniquely mapped to the subset of the set of χ-chains $\tilde{Y}(X)$ by the labeling function $\mu_c : \mathbf{C} \to \tilde{Y}$, and the set $\mathbf{K}(X)$ of all k-nodules of X, to the set of χ-nodules $\hat{Y}(X)$ by the function $\mu_\kappa : \mathbf{K} \to \hat{Y}$.

The existence of inverse mappings μ_c^{-1} and μ_κ^{-1} is an important issue for investigating the isomorphism of chemographs by establishing the completeness of their invariants (see below). In general, there is no question of the existence of surjective inverse mappings $\mu_c^{-1} : \tilde{Y} \to \mathbf{C}$ and $\mu_\kappa^{-1} : \hat{Y} \to \mathbf{K}$, because there exist obvious counterexamples (χ-chains constructed over the set of edges, etc.).

Nevertheless, some inverse functions that map the sets of χ-chains and χ-nodules into subgraphs are guaranteed if we operate not with the sets $\mathbf{C}$ and $\mathbf{K}$ but with the sets $\tilde{\mathbf{C}} = 2^{\mathbf{C}}$ and $\tilde{\mathbf{K}} = 2^{\mathbf{K}}$. Define the *operator* $\hat{\mu}_c^{-1} : \tilde{Y} \to \tilde{\mathbf{C}}$ *of constructing a preimage of a* χ*-chain* α as $\hat{\mu}_c^{-1}\alpha = \{c \in \mathbf{C} | \mu_c(c) = \alpha\}$, and the *operator* $\hat{\mu}_\kappa^{-1} : \hat{Y} \to \tilde{\mathbf{K}}$ *of constructing a preimage of a* χ-nodule κ as $\hat{\mu}_\kappa^{-1}\kappa = \{a \in \mathbf{K} | \mu_\kappa(a) = \kappa\}$. Thus, the operators $\hat{\mu}_c^{-1}$ and $\hat{\mu}_\kappa^{-1}$ construct subsets of kernel equivalence in the corresponding sets of subgraphs.

Note that the χ-chains (Definition 7) and χ-nodules (Definition 8), which are introduced on the basis of the given labeling function (Definition 6), are mathematical descriptions of the one-dimensional and the nodular substructures of crystals. As such, they can correspond to the one-dimensional physical character of the HTSC phenomenon mentioned in the Introduction. Indeed, in a perfect monocrystal, χ-chains correspond to fragments of chains of atoms that pass through the entire crystal. Lemma 1 and Definitions 6–8 provide a natural "bridge" between the tools of the theory of chemographs [8, 9] and the generation of the feature descriptions of crystals, required for the prediction of the properties of crystalline materials.

4. COMBINATORIAL ANALYSIS OF THE ISOMORPHISM OF CHEMOGRAPHS

In spite of the use of the infinite set $\mathbf{\Gamma}$ in the axiomatics of the theory of chemographs (Definitions 1–9), practical applications of this theory deal exclusively with finite samples of descriptions of physical objects. Hence, one can apply combinatorial and metric methods of the theory of solvability/regularity analysis [1–7] to the practical analysis of chemographs. Within this theory, sets of the feature descriptions and of the classes of objects to be tested are first defined and then the solvability/regularity of the classification problems and the correctness/completeness of the corresponding algorithms are analyzed. The problem of introduction of the feature descriptions of chemographs is closely related to the central problem of graph theory—establishing an isomorphism relation between two arbitrary graphs.

Definition 9. Graphs G_1 and G_2 are *isomorphic* ($G_1 \simeq G_2$) if there exists a one-to-one correspondence between their vertices and edges that preserves the adjacency of the vertices and the incidence of the edges.

Let $\mathbf{I}(G)$ be the set of all graphs isomorphic to a graph G. Then $G_1 \simeq G_2$ is equivalent to the statements $\mathbf{I}(G_1) = \mathbf{I}(G_2)$, $G_2 \in \mathbf{I}(G_1)$, and $G_1 \in \mathbf{I}(G_2)$. The verification of each of these conditions is characterized by high computational complexity ($O(n!)$ or, at best, $O(\sum_{i=1}^{m} n_i!) \ll O(n!)$). Therefore, for practical purposes, the analysis of some *necessary conditions of isomorphism* of arbitrary chemographs X_1 and X_2, such as, for example, $\tilde{Y}(X_1) = \tilde{Y}(X_2)$ or $\hat{Y}(X_1) = \hat{Y}(X_2)$ (which follow from Lemma 1), becomes important. The necessary conditions of isomorphism of graphs can be expressed as equalities of some numerical characteristics of graphs—of the so-called invariants.

Definition 10. An *invariant* of a graph is a function $\iota : \mathbf{\Gamma} \to R^n$, $n \in \mathbb{N}$, $\forall a \in \mathbf{\Gamma} : b \in \mathbf{I}(a) \Rightarrow \iota(b) = \iota(a)$. An *elementary* invariant is an invariant of the form $\iota : \mathbf{\Gamma} \to R$, $\iota \in \mathbf{E} = \{\iota : \mathbf{\Gamma} \to R | \forall a \in \mathbf{\Gamma} : b \in \mathbf{I}(a) \Rightarrow \iota(b) = \iota(a)\}$, a *tuple invariant* is an invariant of the form $\iota : \mathbf{\Gamma} \to R^n$, $n \geq 2$. An invariant is said to be *complete* if it satisfies the *condition of completeness of an invariant*: $\forall a \in \mathbf{\Gamma} : b \in \mathbf{I}(G) \Leftrightarrow \iota(a) = \iota(b)$. A pair of graphs that have the same value of an invariant are said to be *isomeric* with respect to a certain value of the invariant.

χ*-Invariants* are *invariants of chemographs*, that are based on the relations of membership of χ-chains and χ-nodules in chemographs. According to Lemma 1, the existence of μ_c implies the existence of relations between χ-chains, as elements of the set $\tilde{Y}$, and a given chemograph X.

We say that a χ*-chain* $\alpha \in \tilde{Y}$ *belongs to a chemograph* X, $\alpha \Subset X$, if $\alpha \in \tilde{Y}(X)$. Accordingly, the *membership of a* χ*-nodule* $\kappa \in \hat{Y}$ *in* X, $\kappa \Subset X$, corresponds to $\kappa \in \hat{Y}(X)$. It is obvious that tuples composed of these membership relations for arbitrary subsets of the sets $\tilde{Y}(X)$ and $\hat{Y}'(X)$ are also χ-invariants. The χ-invariants of a chemograph X can be both *binary* (i.e., they can map the relation itself ($\alpha \Subset X$), ($\kappa \Subset X$)) and *numerical* (such as the number of sub-

graphs in $\mathbf{\Pi}(X)$ that correspond to a given χ-chain α or a given χ-nodule κ).

Define the *operator of membership of a set of subgraphs* $\boldsymbol{\pi}$ *in a chemograph* X $\hat{\beta} : \tilde{\mathbf{\Pi}} \to \{0,1\}$ as $\hat{\beta}[X]\boldsymbol{\pi} = (\boldsymbol{\pi} \cap \mathbf{\Pi}(X) \neq \varnothing)$, $\boldsymbol{\pi} \in \tilde{\mathbf{\Pi}}$, and the *operator of the number of occurrences of the set of subgraphs* $\boldsymbol{\pi}$ in X $\hat{\eta} : \tilde{\mathbf{\Pi}} \to R$ as $\hat{\eta}[X]\boldsymbol{\pi} = |\boldsymbol{\pi} \cap \mathbf{\Pi}(X)|$, $\boldsymbol{\pi} \in \tilde{\mathbf{\Pi}}$. Since $\alpha \in \tilde{Y}(X)$ *and* $\kappa \in \hat{Y}(X)$, according to Lemma 1 and Definition 10, the expressions $\hat{\eta}[X]\hat{\mu}_c^{-1}\alpha$, $\hat{\eta}[X]\hat{\mu}_\kappa^{-1}\kappa$, $\hat{\beta}[X]\hat{\mu}_c^{-1}\alpha$, and $\hat{\beta}[X]\hat{\mu}_\kappa^{-1}\kappa$ are elementary χ-invariants. It is obvious that $\hat{\eta}[X]\hat{\mu}_c^{-1}\alpha$ and $\hat{\eta}[X]\hat{\mu}_\kappa^{-1}\kappa$ are numerical χ-invariants, and $\hat{\beta}[X]\hat{\mu}_c^{-1}\alpha$ and $\hat{\beta}[X]\hat{\mu}_\kappa^{-1}\kappa$ are binary χ-invariants. Denote the result of the application of the operator $\hat{\mu}_c^{-1}$ to the set of χ-chains $\boldsymbol{\alpha} = \{\alpha \in \tilde{Y}\}$ by $\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha} = \{\hat{\mu}_c^{-1}\alpha, \alpha \in \boldsymbol{\alpha}\}$ and the result of application of the operator $\hat{\mu}_\kappa^{-1}$ to the set of χ-nodules $\boldsymbol{\varkappa} = \{\kappa \in \hat{Y}\}$ by $\hat{\boldsymbol{\mu}}_\boldsymbol{\kappa}^{-1}\boldsymbol{\kappa} = \{\hat{\mu}_\kappa^{-1}\kappa, \kappa \in \boldsymbol{\varkappa}\}$. Denote the result of the application of the operator $\hat{\beta}$ to the set $\tilde{\boldsymbol{\pi}} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\}$, $\tilde{\boldsymbol{\pi}} \subset \tilde{\mathbf{\Pi}}$, by $\hat{\boldsymbol{\beta}}\tilde{\boldsymbol{\pi}} = \{\hat{\beta}\boldsymbol{\pi}_1, \hat{\beta}\boldsymbol{\pi}_2, \ldots, \hat{\beta}\boldsymbol{\pi}_n\}$, and of the operator $\hat{\eta}$ to the set $\tilde{\boldsymbol{\pi}}$, respectively, by $\hat{\boldsymbol{\eta}}\tilde{\boldsymbol{\pi}} = \{\hat{\eta}\boldsymbol{\pi}_1, \hat{\eta}\boldsymbol{\pi}_2, \ldots, \hat{\eta}\boldsymbol{\pi}_n\}$.

Consider a more general case. Let $\iota : \mathbf{\Gamma} \to R^n$, $n \geq 2$, be a tuple invariant constructed over a *set of n elementary invariants* $\boldsymbol{\iota}_\mathbf{e} \subset \mathbf{E}$. For a given graph G, the expression $\boldsymbol{\iota}_\mathbf{e}(G) = \{\iota_i(G), i = 1 \ldots n\}$ denotes the set of values of invariants from $\boldsymbol{\iota}_\mathbf{e}$. Define a *function of enumeration of elementary invariants,* $\lambda : \boldsymbol{\iota}_\mathbf{e} \to \mathrm{N}$. Then *the operator of formation of a tuple invariant* $\hat{\imath} : 2^\mathbf{E} \to R^n$ by a given $\boldsymbol{\iota}_\mathbf{e}$ is defined as $\hat{\imath}\boldsymbol{\iota}_\mathbf{e} = (\iota_j, \iota_k, \ldots, \iota_l)$, $\iota_j, \iota_k, \ldots, \iota_l \in \boldsymbol{\iota}_\mathbf{e}$, $\lambda(\iota_j) = 1 \leq \lambda(\iota_k) < \ldots < \lambda(\iota_l) = n$. Denote the value of the ith element of the tuple $\hat{\imath}\boldsymbol{\iota}_\mathbf{e}$ for the graph G by $\hat{\imath}[i]\boldsymbol{\iota}_\mathbf{e}(G) = \iota(G) | \lambda(\iota) = i$.

Theorem 1. *If* $\forall a, b \in \mathbf{\Gamma} : \mathbf{I}(a) \cap \mathbf{I}(b) = \varnothing \Rightarrow \exists_{i=1..n} i : \hat{\imath}[i]\boldsymbol{\iota}_\mathbf{e}(a) \neq \hat{\imath}[i]\boldsymbol{\iota}_\mathbf{e}(b)$, *then* $\iota()$ *is a complete invariant.*

Let us write the condition of completeness of an invariant (Definition 10) in a form corresponding to the pairwise comparison of the graphs a and b: $\forall a, b \in \mathbf{\Gamma} : \iota(a) = \iota(b) \Leftrightarrow \mathbf{I}(a) = \mathbf{I}(b)$. The equivalence relation between the values of the invariants corresponds to the partition of the set of all graphs into classes of isomorphic graphs. The classes of isomorphic graphs over the set of all graphs $\mathbf{\Gamma}$ form the set of all classes of isomorphic graphs $\mathbf{I}(\mathbf{\Gamma}) = \{\mathbf{I}(a) | a \in \mathbf{\Gamma}\}$. For a pair of nonisomorphic graphs, the intersection of the corresponding elements from $\mathbf{I}(\mathbf{\Gamma})$ is empty. Since ι is an invariant by the hypothesis of the theorem (i.e., $\forall a, b \in \mathbf{\Gamma} : \mathbf{I}(a) = \mathbf{I}(b) \Rightarrow \iota(a) = \iota(b)$), we write the condition of completeness of an invariant in the form $\forall a, b \in \mathbf{\Gamma} : \mathbf{I}(a) \cap \mathbf{I}(b) = \varnothing \Rightarrow \iota(a) \neq \iota(b)$. It is obvious that the inequality of the values of two tuples corresponds to the existence of at least one position of these tuples at which there are two different values of the elementary invariant. Then, substituting ι in the form of the tuple $\hat{\imath}\boldsymbol{\iota}_\mathbf{e}$, we obtain a completeness criterion for tuple invariants in the assertion of the theorem. The theorem is proved.

Corollary 1. $\forall a, b \in \mathbf{\Gamma} : \mathbf{I}(\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(a)) \cap \mathbf{I}(\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(b)) = \varnothing \Rightarrow \exists_{i=1..|\boldsymbol{\alpha}|} i : \hat{\imath}[i]\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha} \neq \hat{\imath}[i]\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha}$, *where* $\boldsymbol{\alpha}_1(X) = \{\alpha \in \tilde{Y}(X) | \hat{\eta}[X]\hat{\mu}_c^{-1}\alpha = 1\}$.

Suppose that, for an arbitrary X, there exists a nonempty set $\boldsymbol{\alpha}_1(X)$ such that $\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(X) = X$. In view of the existence of the bijection, a single occurrence of each $\alpha \in \boldsymbol{\alpha}_1$ in X implies the one-to-one correspondence of each $\alpha \in \boldsymbol{\alpha}_1$ to a certain chain in X. Since $\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(X) = X$, it follows that $\mathbf{I}(a) = \mathbf{I}(\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(a))$. Therefore, both $\mathbf{I}(\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(a)) \cap \mathbf{I}(\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(b)) = \varnothing$ and $\boldsymbol{\alpha}_1(a) \neq \boldsymbol{\alpha}_1(b)$ are satisfied, which necessarily corresponds to the existence of a distinguishing element in the hypothesis of Corollary 1.

Corollary 2. *The assertion of the theorem is valid for a set of* χ*-nodules* $\boldsymbol{\kappa} \subseteq \hat{Y}$ $\boldsymbol{\kappa}_1(X) = \{\kappa \in \hat{Y}(X) | \hat{\eta}[X]\hat{\mu}_\kappa^{-1}\kappa = 1\}$, $\breve{\hat{\boldsymbol{\mu}}}_\boldsymbol{\kappa}^{-1}\boldsymbol{\kappa}_1(X) = X$, $\boldsymbol{\kappa}_1(X) \subseteq \boldsymbol{\kappa}$; *here* $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_1(X) | X \in \mathbf{\Gamma}\}$. *By construction, there is a bijection between the sets* $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\beta}}[X]\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha}$, and $\hat{\boldsymbol{\eta}}[X]\hat{\boldsymbol{\mu}}_\boldsymbol{\kappa}^{-1}\boldsymbol{\alpha}$.

The proof is analogous.

Corollary 3. *Over the sets* $\boldsymbol{\alpha}_1(X)$ *and* $\boldsymbol{\kappa}_1(X)$, *one can form irreducible covers of X.*

This is obvious from $\breve{\hat{\boldsymbol{\mu}}}_\mathbf{c}^{-1}\boldsymbol{\alpha}_1(X) = X$. Irreducible covers can be found by a complete enumeration of combinations of elements $\boldsymbol{\alpha}_1(X)$ or by reduced enumeration methods (for example, within the metric approach to data analysis [6, 7]).

Corollary 4. *Complete invariants can be formed over* $\boldsymbol{\alpha} \subseteq \tilde{Y}$ *and* $\boldsymbol{\kappa} \subseteq \hat{Y}$ *if, for every chemograph X, there are* $\boldsymbol{\kappa}'(X) = \{\kappa \in \boldsymbol{\kappa} | \hat{\eta}[X]\hat{\mu}_\kappa^{-1}\kappa = 1\}$ *and* $\boldsymbol{\alpha}'(X) = \{\alpha \in \boldsymbol{\alpha}' | \hat{\eta}[X]\hat{\mu}_c^{-1}\alpha = 1\}$ *such that* $\breve{\boldsymbol{\pi}}'(X) = X$ *for* $\boldsymbol{\pi}'(X) = \hat{\boldsymbol{\mu}}_\boldsymbol{\kappa}^{-1}\boldsymbol{\kappa}'(X) \cup \hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha}'(X)$.

This assertion follows from the uniqueness of the occurrence of elements in $\boldsymbol{\kappa}'(X)$ and $\boldsymbol{\alpha}'(X)$. Note that, in the case of existence of $\breve{\boldsymbol{\pi}}'(X) = X$, $\boldsymbol{\alpha}_1(X)$ and $\boldsymbol{\kappa}_1(X)$ may be empty.

Corollary 5. *Suppose that each graph G in a finite set* $\mathrm{Pr} \subset \mathbf{\Gamma}$ *is labeled by an isomorphism label* $\mathrm{iso}(G) : \mathbf{I}(\mathbf{\Gamma}) \to \mathrm{N}$ *so that the solvability criterion of a classification problem is expressed as* $\bigvee_{a,b \in \mathrm{Pr}} \mathrm{iso}(a) \neq \mathrm{iso}(b) \Rightarrow \iota(a) \neq \iota(b)$. *Then the solvability criterion is*

equivalent to the condition of local completeness of an invariant: $\forall_{a,b\in \mathrm{Pr}} \mathrm{iso}(a) \neq \mathrm{iso}(b) \Rightarrow \exists_{i=1..|\chi|} i : \hat{\imath}[i]\boldsymbol{\chi}(a) \neq \hat{\imath}[i]\boldsymbol{\chi}(b)$.

Corollary 6. *The problem of recognition of isomorphic graphs is solvable if and only if the sum over the set Pr of the cardinalities of the differences between the class of isomeric graphs and the class of isomorphic graphs is zero.*

Define μ_ι—*the operator of constructing a class of isomeric graphs* (Definition 10) for a given graph G and an invariant ι—as $\mu_\iota(G,\iota) = \{g \in \boldsymbol{\Gamma} \,|\, \iota(g) = \iota(G)\}$, so that the condition of completeness of ι is $\forall a \in \boldsymbol{\Gamma}$: $\mu_\iota(a,\iota)\backslash\mathbf{I}(a) = \varnothing$. Suppose that graphs from Pr, isomorphic to G, form a *local set of isomorphic graphs* $\mathbf{i}(G,\mathrm{Pr}) = \{g \in \mathrm{Pr} \,|\, \mathrm{iso}(g) = \mathrm{iso}(G)\}$, $\mathbf{i}(G,\mathrm{Pr}) \subset \mathbf{I}(G)$, and the graphs isomeric to G form a *local set of isomeric graphs* $\mathbf{i\mu}(G,\iota,\mathrm{Pr}) = \{g \in \mathrm{Pr} \,|\, \iota(g) = \iota(G)\}$, $\mathbf{i\mu}(G,\iota,\mathrm{Pr}) \subset \mu_\iota(G,\iota)$. Then $\forall_{a\in \mathrm{Pr}} \mathbf{i\mu}(a,\iota,\mathrm{Pr})\backslash\mathbf{i}(a,\mathrm{Pr}) = \varnothing$ and, respectively, $\sum_{a\in \mathrm{Pr}} |\mathbf{i\mu}(a,\iota,\mathrm{Pr})\backslash\mathbf{i}(a,\mathrm{Pr})| = 0$.

Corollary 7. Let ι *be an invariant and* $r_\iota(\iota,\mathrm{Pr}) = 1 - \frac{1}{|\mathrm{Pr}|^2}\sum_{a\in \mathrm{Pr}} |\mathbf{i\mu}(a,\iota,\mathrm{Pr})\backslash\mathbf{i}(a,\mathrm{Pr})|$. *The invariant* ι *is complete if and only if* $r_\iota(\iota,\mathrm{Pr}) = 1$.

This is obvious from Corollary 6.

Corollary 8. *Define an operator* $\phi(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ *of selecting the subset of the pairs of chemographs for each of which there is a difference in the i-th position of the tuple invariant* $\hat{\imath}\boldsymbol{\chi}$, $\phi(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr}) = \{(a,b) \,|\, a,b \in \mathrm{Pr},\ \hat{\imath}[i]\boldsymbol{\chi}(a) \neq \hat{\imath}[i]\boldsymbol{\chi}(b)\}$. *Suppose that a set* $\boldsymbol{\chi}$ *is such that* $\varphi(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr}) \cap \varphi(\hat{\imath}\boldsymbol{\chi}, j, \mathrm{Pr}) = \varnothing$ *for every i and j*, $i \neq j$. *Then* $r_\iota(\iota,\mathrm{Pr}) = \sum_{i=1}^{|\chi|} \varphi_\iota(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$, *where* $\varphi_\iota(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr}) = |\varphi(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})| \,/\, |\mathrm{Pr}|^2$.

This is derived from Corollaries 6 and 7.

Thus, Theorem 1 states that a tuple invariant is a complete invariant if there is a distinguishing element for an arbitrary pair of nonisomorphic graphs. The equivalence relation on the set of initial descriptions, defined by the *isomorphism label* iso(), corresponds to the expert assessment of the equivalence of two arbitrary crystal structures (for example, the coincidence of the partial coordinates of atoms, the space group, all parameters of the lattice, and occupancy of atoms to within the error of the method).

Within the present formalism, tuple invariants of chemographs are considered as vectors of feature descriptions of objects on the basis of which the solvability/regularity properties of the problems of classification of chemographs are tested. The indicator $r_\iota(\iota,\mathrm{Pr})$ obtained in Theorem 1 is a quantitative combinatorial estimate of the completeness of the tuple invariant ι over the set of precedents Pr, and the indicators $\varphi_\iota(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ allow one to estimate the contribution of each element of the tuple. It is expedient to construct the enumeration function $\lambda : \boldsymbol{\iota}_\mathbf{e} \to \mathrm{N}$ of tuple elements on the basis of functionals that estimate "informativity." Such functionals can be used to select informative values of features [2, 33, 34], which allows one to find locally complete invariants for a relatively small number of elementary invariants included in the tuple.

5. DISTANCE FUNCTIONS BETWEEN CHEMOGRAPHS AND THE METRIC APPROACH TO THE ANALYSIS OF THE ISOMORPHISM OF CHEMOGRAPHS

Theorem 1 is not only necessary for the computational testing of the local completeness of various tuple invariants of chemographs. Theorem 1, together with the corollaries, provides a fundamental basis for constructing metric distance functions between chemographs (that is, metrics on a set of chemographs). Within the present formalism, the distance function $d_\chi(X_1, X_2)$ between chemographs X_1 and X_2 is a function of the type $d'_\chi(\iota(X_1), \iota(X_2))$ that depends on the given tuple invariant (ι). The value of the tuple invariant ι for a certain chemograph represents a vector in the space R^n, and let us recall that the generation, analysis, and application of distance functions over vectors in R^n is one of the most important directions in data mining [35].

To introduce a metric, one can use binary and numerical elementary invariants ($\hat{\eta}[X]\hat{\mu}_c^{-1}\alpha$, $\hat{\eta}[X]\hat{\mu}_\kappa^{-1}\kappa$, $\hat{\beta}[X]\hat{\mu}_c^{-1}\alpha$, $\hat{\beta}[X]\hat{\mu}_\kappa^{-1}\kappa$ etc.) arranged in some order (defined by the function $\lambda : \boldsymbol{\iota}_\mathbf{e} \to \mathrm{N}$); each tuple element can be assigned a certain weight, etc. In view of Theorem 1, of special interest is the use of the tuple invariants based on elementary χ-invariants over various sets of χ-chains and χ-nodules.

It is obvious that the *property of completeness of an invariant* ι considered in Theorem 1 is needed for the invariant to be used in the construction of the *metric distance functions*. Indeed, the value of $d'_\chi(\iota(X_1), \iota(X_2))$ should be zero for isomorphic X_1 and X_2, while, for nonisomorphic graphs, the value of $d'_\chi(\iota(X_1), \iota(X_2))$ should be strictly greater than 0 (the first metric axiom). Indeed, this axiom holds for complete (or, at least, for locally complete) invariants.

Suppose given an arbitrary set of χ-chains $\boldsymbol{\alpha} \subseteq \tilde{Y}$ and an arbitrary set of χ-nodules $\boldsymbol{\kappa} \subseteq \hat{Y}$. Define a set of subgraphs $\boldsymbol{\pi} = \hat{\boldsymbol{\mu}}_\boldsymbol{\kappa}^{-1}\boldsymbol{\kappa} \cup \hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\boldsymbol{\alpha}$, $|\boldsymbol{\pi}| = n$, over which χ-invariants will be formed.

Introduce a metric over the set of binary χ-invariants. Let us form a set of elementary χ-invariants $\boldsymbol{\iota}_\mathbf{b}(X) = \hat{\boldsymbol{\beta}}[\mathbf{X}]\boldsymbol{\pi} = \{\hat{\beta}[X]\pi_i \,|\, \pi_i \in \boldsymbol{\pi},\ i = 1..n\}$. Take an enumeration function λ and form a binary tuple

invariant $\hat{\imath}\boldsymbol{\iota}_{\mathbf{b}}(X) = \hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\boldsymbol{\pi}$. Define a distance function $d_{\chi b}(\{\omega_i\}, X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\hat{\imath}[i]\hat{\boldsymbol{\beta}}[X_1]\boldsymbol{\pi} \oplus \hat{\imath}[i]\hat{\boldsymbol{\beta}}[X_2]\boldsymbol{\pi}$ (ω_i are the weights of elementary invariants; in the simplest case, $\omega_i = 1$) over binary χ-invariants, which, by construction, is the *Hamming metric* (provided that the invariant $\hat{\imath}\boldsymbol{\iota}_{\mathbf{b}}$ is complete).

Introduce a metric over a set of numerical χ-invariants. Define a set $\boldsymbol{\iota}_{\boldsymbol{\eta}}(X) = \hat{\boldsymbol{\eta}}[\mathbf{X}]\boldsymbol{\pi} = \{\hat{\eta}[X]\pi_i \,|\, \pi_i \in \boldsymbol{\pi}, i = 1...n\}$, take a function λ, and formulate (locally) a complete invariant $\hat{\imath}\boldsymbol{\iota}_{\boldsymbol{\eta}}(X) = \hat{\imath}\hat{\boldsymbol{\eta}}[\mathbf{X}]\boldsymbol{\pi}$. By construction, the function $d_{\chi\eta}(\{\omega_i\}, X_1, X_2) = \sqrt[p]{\sum_{i=1}^{n}\omega_i\left|\hat{\imath}[i]\hat{\boldsymbol{\eta}}[X_1]\boldsymbol{\pi} - \hat{\imath}[i]\hat{\boldsymbol{\eta}}[X_2]\boldsymbol{\pi}\right|^p}$ is the Minkowski metric, whose particular case is given by the Euclidean metric ($p = 2$, $\omega_i = 1$).

The introduction of a metric on the set of objects (i.e., chemographs) provides a basis for the metric analysis of the solvability/regularity criteria for problems and the correctness/completeness criteria for algorithmic models [5–7]. The metric forms of these criteria for a two-class problem over the set of precedents $Q = \{(q_i, b_i \,|\, q_i \in R^n,\ b_i \in (0,1))\}$ with the classes C^+ and C^- ($C^+ \cup C^- = Q$ and $C^+ \cap C^- = \varnothing$) are formulated with the use of the metrics ρ_q, which can be represented by the metrics $d_{\chi b}$ and $d_{\chi\eta}$ proposed above. Then the solvability criterion for the problem $Z(C^+/C^-)$ is calculated as $\underset{Q}{\forall} q_1, q_2 : b_1 \neq b_2 \Rightarrow \rho_q(q_1, q_2) > 0$ and the regularity criterion, as $\underset{Q}{\forall} q_1, q_2 : \rho_q(q_1, q_2) > 0$. For the existence of a correct algorithm, it is sufficient that the condition $r_K(C^+, Q) = 1$ is satisfied, where $r_K(K, M) = \frac{1}{N}\sum_{i=1}^{N}(\min_{j\notin K}\rho_{ij} > \min_{k\in K}\rho_{ik})$ is a combinatorial functional that estimates "ε-compactness" of the sets of objects $K \subset M$ in respect to the entire set M. The values of the functional $r_K()$ can be estimated by the subquadratic methods of analysis of metric condensations [5].

6. METHODS FOR PREDICTING NUMERICAL TARGET VARIABLES

To solve materials science problems, including the quantitative prediction of the properties of HTSCs, one needs, first, the methods for generating feature descriptions of chemographs and, second, the algorithms for predicting the corresponding numerical variables. Such algorithms have been developed within topological and lattice-theoretical interpretations of the generation of synthetic numerical features [10].

To any variable with a finite set of values $I_k = \{\lambda_{k_1}, \lambda_{k_2}, ..., \lambda_{k_b}, ..., \lambda_{k_{|I_k|-1}}, \Delta\}$, there corresponds a set of disjoint subsets $\Gamma_k^{-1}(\lambda_{k_b})$ of a finite set of precedents Q. In the case of a numerical kth variable, the set I_k is linearly ordered ($\lambda_{k_{b-1}} \le \lambda_{k_b} \le \lambda_{k_{b+1}}$), so that, to each value of λ_{k_b}, there corresponds a subset $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b}\Gamma_k^{-1}(\lambda_{k_\beta})$ of the set of precedents. The definition of a pair of classes $C_{kb}^+ = u(\lambda_{k_b})$, $C_{kb}^- = \neg u(\lambda_{k_b})$, $C_{kb}^+ \cup C_{kb}^- = Q$, $C_{kb}^+ \cap C_{kb}^- = \varnothing$, for each λ_{k_b} defines a classification problem solved by the algorithm $\hat{A}_{kb}(\theta)$.

The prediction of the kth numerical variable can thus be made by the algorithm $\hat{A}(k, \theta) : J_{ob} \to R$, where J_{ob} is a set of admissible feature descriptions (typically, $J_{ob} \subseteq \bigcup_{k=1}^{n} I_k$) and θ is a vector of parameters. The algorithm $\hat{A}(k, \theta)$ allows one to calculate, on the basis of the data in the information matrix $\hat{M}(m_i)$ ($i = 1, ..., N$), the column of the corresponding values of the kth variable. The algorithm $\hat{A}(k, \theta)$ can either "directly" predict the values of the kth variable (regression algorithm or "neural networks," for instance), or it can be made as a composition of the classification algorithms $\hat{A}_{kb}(\theta)$. One can apply various methods to determine the values of the vectors $\theta \in R^n$: methods of computational linear algebra (singular decomposition, etc.), neural networks, stochastic approximation, and so on.

The prediction of numerical variables can be made within *chemometric approach* as the problem of matching the values of a certain "expert" metric d_e and a "feature" metric with weights (i.e., $d_{\chi b}$ or $d_{\chi\eta}$) according to the following criteria:

$$\arg\min_{\{\omega_i\}} \sum_{m=1}^{N-1}\sum_{j=m+1}^{N}\left|d_{\chi b}(\{\omega_i\}, X_m, X_j) - d_e(X_m, X_j)\right|, \quad (1.1)$$

$$\arg\min_{\{\omega_i\}, p} \sum_{m=1}^{N-1}\sum_{j=m+1}^{N}\left|d_{\chi\eta}(\{\omega_i\}, X_m, X_j) - d_e(X_m, X_j)\right|. \quad (1.2)$$

A practically important particular case of an expert metric is given by a metric based on a scalar (that represents the numerical variable to be predicted). This "one-dimensional" metric satisfies all three metric axioms (since they are valid for any three different points on the real axis). In this case, each of the criteria (1.1) and (1.2) is, actually, equivalent to an additive scheme that involves the summation of feature values with weights followed by the application of a correcting operation (a corrector function). Indeed, suppose that the zero element {0} appears in all the sets I_k, so that one can determine the distance from the zero element to any other element of the set I_k by a scalar

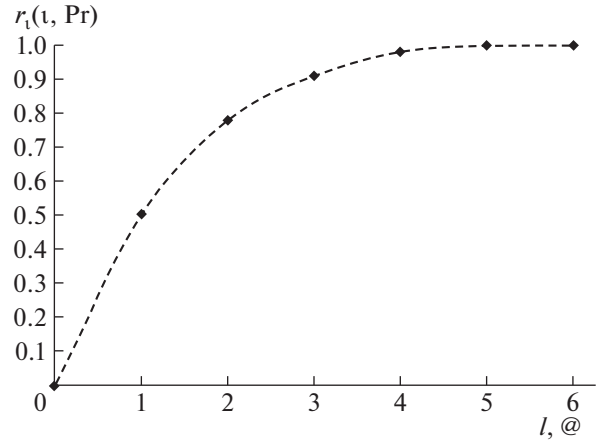

Fig. 1. Calculation of the local completeness ($r_\iota(\iota, \mathrm{Pr})$) of tuple invariants over χ-chains of fixed length.

expert metric d_e. Then one can reformulate criteria (1.1) and (1.2) in terms of the distance from the zero element and make a transition from the estimate of the pairwise distances to the summation over all objects; i.e., one can calculate the value of the corresponding sum of residuals. In this case, the "machine learning" problem is formulated as $\arg\min_{\{\omega_i\},p} \sum_{m=1}^{N} \left| d_{\chi\eta}(\{\omega_i\},\{0\},X_m) - T(X_m) \right|$, where $T(X_j)$ is the value of the predicted numerical variable for object X_j.

7. APPLICATION OF THE CHEMOGRAPH ANALYSIS FORMALISM IN THE FRAMEWORK OF THE TOPOLOGICAL APPROACH TO DATA ANALYSIS TO THE PROBLEMS OF PREDICTING THE PROPERTIES OF CRYSTALLINE MATERIALS: AN EXAMPLE OF PREDICTION OF T_c FOR CUPRATE HTSCs

The formalism developed allows one to pass from the set of initial descriptions of the structures of materials to feature descriptions acceptable for "machine learning" algorithms. In this study, by the initial descriptions are meant the internal coordinates of atoms in a unit cell together with the space group and the lattice parameters. From this representation, one passes to the Cartesian coordinates of atoms and then to the tuple invariants based on χ-chains and χ-nodules, developed in this study (Theorem 1). To test the algorithms following from the formalism proposed, we used a set of labels Y comprised of Cartesian products of the chemical types of the elements by their admissible oxidation degrees ($|Y| = 548$).

Before the application of the prediction algorithms, one should determine the optimal values of the parameters n and k that are used to generate the feature descriptions of chemographs. To this end, we carried out combinatorial testing of the completeness of invariants from the families $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_c^{-1}\tilde{Y}^n$ ($n = 1...7$), $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_\kappa^{-1}\hat{Y}(k)$ ($k = 3...7$), and $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}](\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\tilde{Y}^n \cup \hat{\boldsymbol{\mu}}_\kappa^{-1}\hat{Y}(k))$. For each tuple invariant, we calculated local completeness estimate $r_\iota(\iota, \mathrm{Pr})$ (Corollary 7 of the Theorem 1).

The completeness of the tuple invariants (i.e., the solvability of the appropriate problems, see Theorem 1 with corollaries) was tested on a sample of 125000 pairwise different structures of substances from the ICSD database [36]. The coordinates of atoms in the unit cell of each crystalline structure were transformed from partial to Cartesian coordinates. Then, by translations of the corresponding Fedorov's group, we obtained Cartesian coordinates for the atoms of a cube consisting of 27 unit cells. In this array of Cartesian coordinates, we determined all χ-chains and χ-nodules by the complete enumeration algorithm. The criterion of contact of atoms was the overlap of the ionic radii with a tolerance of 0.1 Å. The experiments have shown the potential of the use of all three families of χ-invariants for generating binary tuple invariants. The results of experiments with the family of invariants $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\tilde{Y}^n$ (χ-chains of fixed length) are presented in Fig. 1.

The estimate of the completeness of the binary tuple invariants over the sets of all χ-chains of length n has shown that the "qualitative composition" of a chemograph ($n = 1$, i.e., the presence of labels of atoms of specific type) is similar for 50% of pairs of nonisomorphic chemographs ($r_\iota(\iota, \mathrm{Pr}) = 0.50$). The use of a qualitative edge composition of a chemograph ($n = 2$, i.e., the presence of paired combinations of atomic labels) has significantly increased the accuracy of distinguishing between isomorphic and isomeric chemographs ($r_\iota(\iota, \mathrm{Pr}) = 0.79$). An increase in the length of a χ-chain led to a monotonic increase in the combinatorial estimate of the local completeness of the corresponding tuple invariants, and the estimate of $r_\iota(\iota, \mathrm{Pr})$ reached a value of 0.98 already at $n = 4$. Therefore, from a practical viewpoint, chains of length 4 are quite sufficient for the existence of locally complete tuple invariants. The difference from the molecules of organic substances (where the completeness of an invariant is reached only at $n \geq 7$) is quite understandable: the dimension of the dictionary for labeling the atoms of organic molecules was much lower ($|Y| = 20$) [9].

On the basis of the invariants $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_\mathbf{c}^{-1}\tilde{\mathrm{Y}}^4$, we have formalized the problem of quantitative prediction of temperature T_c for a sample of 1450 structures of cuprate HTSCs for which the temperatures T_c were known (data from the ATOMWORKS database [37]). The efficiency of various algorithms for solving the problem was estimated in cross-validation experiments, which included ten random partitions of the entire sample into "6 : 1 training-test" sets of objects. As $\lambda : \mathfrak{t}_\mathbf{e} \to \mathrm{N}$, we used functionals based on the exact Fisher's test [10].

Table 1. Examples of atomic chains with maximum weights in a linear model for calculating T_c. The chains are arranged in decreasing order of weights ω_i.

Chain	$\varphi_t(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$	ω_i, K
Ba–O–Ca–O	0.108	7.43
Bi–Cu–O–O	0.048	6.49
Pb–O–O–Sr	0.048	6.31
Hg–O–Ba–O	0.036	6.08
Bi–O–O–Ca	0.008	5.92
Cu–O–Ca–O	0.607	5.49
Cu–O–Ba–O	0.382	5.27
Y–O–Cu–O	0.575	5.04
Pb–O–Cu–O	0.061	4.92
Y–O–Y–O	0.595	4.72
Eu–O–Eu–O	0.016	4.33
Y–O–Ni–O	0.011	4.3
Bi–Sr–O–O	0.178	4.16
Bi–O–Bi–O	0.269	4.11
Cu–Ba–O–O	0.047	−1.42
Ba–Ba–O–O	0.045	−1.48
Hg–O–Hg–O	0.019	−1.77
Pb–O–Sr–O	0.085	−1.83
Tl–O–Ba–O	0.013	−2.11

To generate synthetic numerical features, we used the algorithm $\hat{A}_{kb}(\theta(\mathrm{Pr})) = B_{kb}(\theta(\mathrm{Pr})) \circ C_{kb}(\theta(\mathrm{Pr}))$, which is a composition of the recognition operator B_{kb} and the correction operation C_{kb} over the set of precedents Pr. For linear recognition operators B_{kb}, the correction operation C_{kb} was either a linear transformation, a logarithm, an exponential function, a power law function, a neural network of certain configuration, etc. [10]. To calculate the vector θ of the algorithm $\hat{A}_{kb}(\theta(\mathrm{Pr}))$ and the metric $d_{\chi b}$, we used a singular decomposition, neural networks, and multistart stochastic approximation [10]. Preliminary experiments showed that an optimal solution of the problem of predicting the critical temperature Tc can be obtained for linear operations B_{kb} and C_{kb} with the use of stochastic approximation to determine the vector θ ($r = 0.77$ in cross validation, Fig. 2).

Thus, the application of the simplest model with linear operators $B_{kb}(\theta(\mathrm{Pr}))$ and $C_{kb}(\theta(\mathrm{Pr}))$ and stochastic approximation for finding the vector of parameters θ has allowed us to obtain a noticeable correlation between the predicted and the experimental values of the critical temperature of the superconducting transition. Note that this linear model was implemented under the assumption of a perfect crystalline structure, without taking into account powder microcrystallinity or the occupancies of atomic positions, which are related to the oxygen doping of the HTSC crystal. The presence of horizontal strips in Fig. 2 is partially attributed to the fact that the occupancies of atomic positions have not been taken into account.

In contrast to the methods of "deep learning," neural networks, etc., the present formalism allows one to obtain interpretable results. In other words, the analysis of the features generated by the methods of analysis of the completeness of the tuple invariants studied has allowed us to obtain results that are potentially important for deeper understanding of the problem area. For example, the calculation of the values of the indicator $\varphi_t(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ (see Corollary 8 to Theorem 1) and of the weights ω_i of the elementary invariants has allowed us to reveal the types of atomic chains with the maximum absolute contribution to the calculated values of T_c (Table 1).

For example, in mercuric HTSCs ($HgBa_2CuO_{4+\delta}$ [38, 39], $HgBa_2CaCu_2O_{6+\delta}$, and others), the atomic chain Hg–O–Ba–O made the maximum contribution to T_c (+6.08 K), while the chain "Hg-O-Hg-O," conversely, promoted the reduction of T_c (−1.77 K). In the case of bismuth-containing HTSCs $Bi_2Sr_2CuO_{6+\delta}$ [40], $Bi_2Sr_2CaCu_2O_{8+x}$, and others, the maximum contribution was made by the "Bi-Cu-O-O" (+6.49 K) and "Bi–O–O–Ca" (+5.92 K) chains, whereas more widespread "Bi–Sr–O–O" and "Bi–O–Bi–O" chains made close contributions (on the order of +4.1 K). In the case of lead-containing HTSCs $Pb_3Sr_4Ca_3Cu_6O_x$, $Pb_2Sr_2(Y_xCa_{1-x})Cu_3O_{8+\delta}$, and others [41], the "Pb–O–O–Sr" and "Pb–O–Cu–O" chains made a positive contribution to the increase of T_c, while the chain "Pb–O–Sr–O" made a negative contribution (−1.83 K). In the case of all Ba, Ca-containing cuprate HTSCs, the widespread chains "Ba–O–Ca–O," "Cu–O–Ba–O," and "Cu–O–Ca–O" made positive contributions to T_c (+5.27...7.43 K), while the less widespread chains "Ba–Ba–O–O" and "Cu–Ba–O–O" made negative contributions (on the order of −1.4 K). In this case, the contributions of structurally similar chains "Cu–O–Ba–O" and "Cu–O–Ca–O" were sufficiently close (+5.27 K and +5.49 K, respectively).

The generalization of the results of algorithm adjustment (first of all, the weights ω_i for specific chains) to other chemical elements, oxidation degrees, etc., is an important problem for the practical application of the developed algorithms for predicting T_c.

First, one can develop special alphabets of labels Y for specific problems that include some abstract "meta-elements." Recall that D.I. Mendeleev's periodic table is a discrete system based on the monotonicity of variation of the properties of elements within groups/periods. The property of monotonicity allows one to define various types of "meta-elements" that may contain both whole groups/periods of elements and certain subsets of elements from the same group/period.

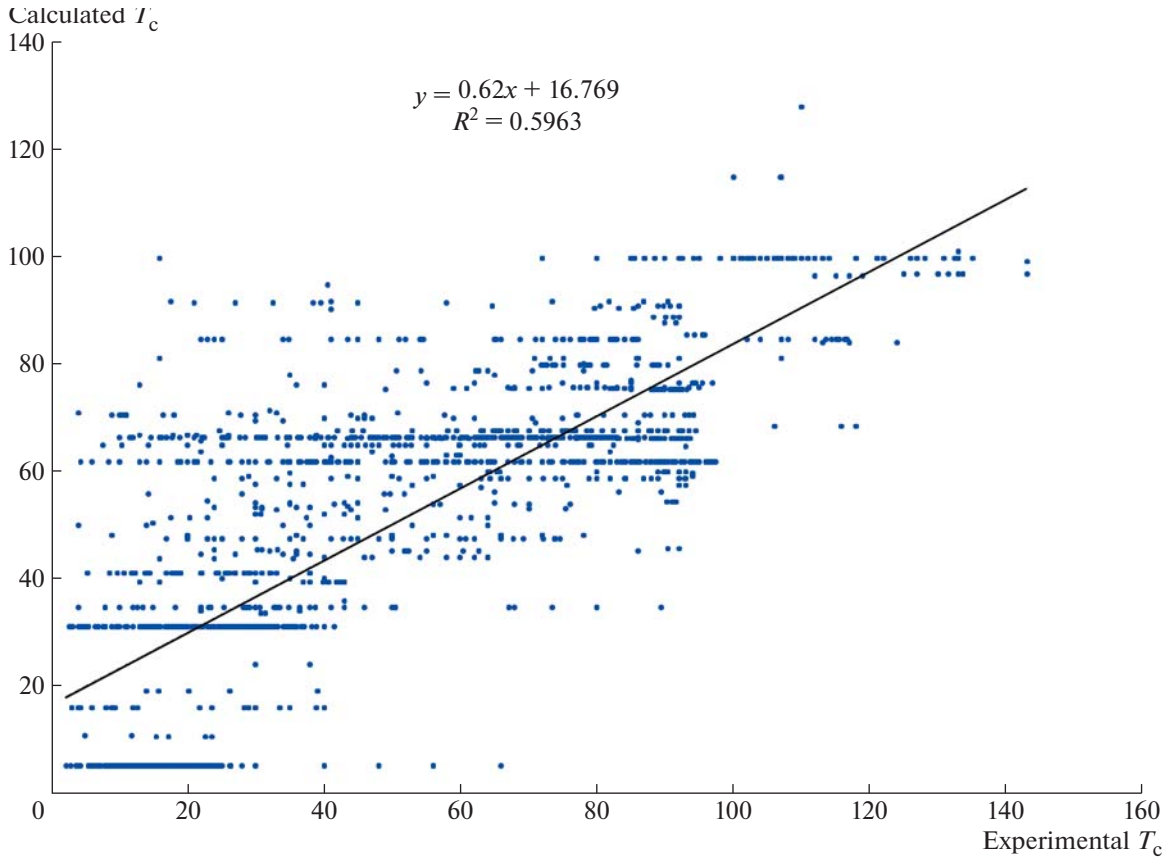

Fig. 2. Example of correlation between calculated and experimentally determined values of T_c of cuprate HTSCs. For this example, the values of the correlation coefficient on the training dataset and on the testing dataset were 0.82 and 0.75, respectively.

Second, the already obtained results allow us to juxtapose chains differing only by a single atom in a certain position (see above the example of the chains "Cu–O–Ba–O" and "Cu–O–Ca–O"). In our opinion, a comparative analysis of the weights of such chains is a quite promising direction of research.

8. CONCLUSIONS

Within the application of the mathematical methods of recognition and "data mining" [35] to problems of the materials science, the choice of the primary and "synthetic" feature descriptions of the object is of fundamental importance. In this paper, we have presented the results of application of the methods of topological analysis of poorly formalized problems (including metric data analysis and of the theory of chemographs) to the problems of predicting the properties of crystalline materials. The formalism developed and the corresponding machine learning algorithms have been experimentally tested on a sample of crystalline structures of cuprate HTSCs for each of which the temperature T_c was measured. Note that the applicability of the methods developed is by no means restricted to predicting T_c or other parameters of high-temperature superconductivity. The existence of the so-called "topological materials" or "topological phases" suggests that many electrical, mechanical, and other properties of materials can be attributed to processes that admit a two-dimensional or even a one-dimensional mathematical description (which corresponds to elementary chain invariants investigated in the present cycle of studies on the theory of chemographs). The theoretical results obtained imply that, under the condition of the completeness of a "chain" of invariants, the latter allows one to generate universal feature descriptions of crystalline structures, which can be used for solving the problems of predicting various properties of crystalline materials.

ACKNOWLEDGMENTS

We are grateful to Prof. O.A. Gromova for useful discussions on expert data analysis.

FUNDING

This work was supported by the Russian Foundation for Basic Research, project nos. 19-07-00356, 18-07-01022, 17-07-01419, 16-07-01129, and 18-07-00944.

COMPLIANCE WITH ETHICAL STANDARDS

CONFLICT OF INTEREST

We declare that we have no conflicts of interest related to the preparation and publication of this article.

REFERENCES

1. I. Yu. Torshin and K. V. Rudakov, "On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification," Pattern Recogn. Image Anal. **25** (4), 577–587 (2015).
2. Yu. I. Zhuravlev, K. V. Rudakov, and I. Yu. Torshin, "Algebraic criteria for local solvability and regularity as an instrument for researching amino acid sequence morphology," Trudy Mosk. Fiz.-Tekhn. Inst. **3** (4), 45–54 (2011).
3. Yu. I. Zhuravlev, "Correct algebras over sets of incorrect (heuristic) algorithms," I: Cybern. **13** (4), 489–497 (1977).
4. I. Yu. Torshin and K.V. Rudakov, "Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 1: Factorization approach," Pattern Recognition and Image Anal. **27** (1), 16–28 (2017).
5. I. Yu. Torshin and K.V. Rudakov, "Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 2: Metric approach within the framework of the theory of classification of feature values," Pattern Recogn. Image Anal. **27** (2), 184–199 (2017).
6. I. Yu. Torshin and K. V. Rudakov, "On metric spaces arising during formalization of recognition and classification problems. Part 1: Properties of compactness," Pattern Recogn. Image Anal. **26** (2), 274–284 (2016).
7. I. Yu. Torshin and K. V. Rudakov, "On metric spaces arising during formalization of problems of recognition and classification. Part 2: Density properties," Pattern Recognit. Image Anal. **26** (3), 483–496 (2016).
8. I. Yu. Torshin and K. V. Rudakov, "On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 1: Fundamentals of modern chemical bonding theory and the concept of the chemograph," Pattern Recogn. Image Anal. **24** (1), 11–23 (2014).
9. I. Yu. Torshin and K. V. Rudakov, "On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 2. Local completeness of invariants of chemographs in view of the combinatorial theory of solvability," Pattern Recogn. Image Anal. **24** (2), 196–208 (2014).
10. I. Yu. Torshin and K. V. Rudakov, "On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables," Pattern Recognit. Image Anal. **29** (4), 654–667 (2019).
11. H. Fröhlich, "On the theory of superconductivity: the one-dimensional case," Proc. R. Soc. Lond. Ser. A **223** (1154), 296–305 (1954). https://doi.org/10.1098/rspa.1954.0116
12. F. von Oppen, Y. Peng, and F. Pientka, "Topological superconducting phases in one dimension," in *Topological Aspects of Condensed Matter Physics, École de Physique des Houches, Session CIII, 4–29 August 2014*, Ed. by C. Chamon, M. O. Goerbig, R. Moessner, and L. F. Cugliandolop (Oxford University Press, Oxford, 2017), pp. 387–447. https://doi.org/10.1093/acprof:oso/9780198785781.003.0009
13. K. Nishi, "Possible higher temperature superconductivity in the modulation-doped superlattice structure of cuprate superconductors," Phys. Lett. A **382** (45), 3293–3297 (2018). https://doi.org/10.1016/j.physleta.2018.09.024
14. V. A Khodel., J. W. Clark, and M. V. Zverev, "Toward a topological scenario for high-temperature superconductivity of copper oxides," Phys. Lett. A **382** (45), 3281–3286 (2018). https://doi.org/10.1016/j.physleta.2018.09.017
15. V. Lakhno, "A translation invariant bipolaron in the Holstein model and superconductivity," SpringerPlus **5**, Article **1277**, 1–18 (2016). https://doi.org/10.1186/s40064-016-2975-x
16. Y. Li, J. Terzic, P. G. Baity, D. Popović, G. D. Gu, Q. Li, A. M. Tsvelik, and J. M. Tranquada, "Tuning from failed superconductor to failed insulator with magnetic field," Sci. Adv. **5** (6), eaav7686, 1–5 (2019). https://doi.org/10.1126/sciadv.aav7686
17. H.-H. Kim, S. M. Souliou, M. E. Barber, E. Lefrancois, M. Minola, M. Tortora, R. Heid, N. Nandi, R. A. Borzi, G. Garbarino, A. Bosak, J. Porras, T. Loew, M. Konig, P. M. Moll, A. P. Mackenzie, B. Keimer, C. W. Hicks, and M. Le Tacon. "Uniaxial pressure control of competing orders in a high-temperature superconductor," Sci. **362** (6418), 1040–1044 (2018). https://doi.org/10.1126/science.aat4708
18. W. Ruan, X. Li, C. Hu, Z. Hao, H. Li, P. Cai, X. Zhou, D.-H. Lee, and Y. Wang. "Visualization of the periodic modulation of Cooper pairing in a cuprate superconductor," Nat. Phys. **14** (12), 1178–1182 (2018). https://doi.org/10.1038/s41567-018-0276-8
19. J. Wu, A. T. Bollinger, X. He, and I. Bozovic, "Spontaneous breaking of rotational symmetry in copper oxide superconductors," Nat. **547** (7664), 432–435 (2017). https://doi.org/10.1038/nature23290
20. P. Giraldo-Gallo, J. A. Galvis, Z. Stegen, K. A. Modic, F. F. Balakirev, J. B. Betts, X. Lian, C. Moir, S. C. Riggs, J. Wu, A. T. Bollinger, X. He, I. Bozovic, B. J. Ramshaw, R. D. McDonald, G. S. Boebinger, and A. Shekhter, "Scale-invariant magnetoresistance in a cuprate superconductor," Sci. **361** (6401), 479–481 (2018). https://doi.org/10.1126/science.aan3178
21. Y. He, M. Hashimoto, D. Song, S.-D. Chen, J. He, I. M. Vishik, B. Moritz, D.-H. Lee, N. Nagaosa, J. Zaanen, T. P. Devereaux, Y. Yoshida, H. Eisaki, D. H. Lu, and Z.-X. Shen, "Rapid change of superconductivity and electron-phonon coupling through critical doping in Bi-2212," Sci. **362** (6410), 62–65 (2018). https://doi.org/10.1126/science.aar3394
22. H. C. Po, A. Vishwanath, and H. Watanabe, "Symmetry-based indicators of band topology in the 230 space groups," Nat. Commun. **8**, Article 50 (2017). https://doi.org/10.1038/s41467-017-00133-2
23. K. Gotlieb, C.-Y. Lin, M. Serbyn, W. Zhang, C. L. Smallwood, C. Jozwiak, H. Eisaki, Z. Hussain, A. Vishwanath, and A. Lanzara A. "Revealing hidden spin-momentum locking in a high-temperature cuprate superconduc-

tor," Sci. **362** (6420), 1271–1275 2018. https://doi.org/10.1126/science.aao0980

24. P. Popčević, D. Pelc, Y. Tang, K. Velebit, Z. Anderson, V. Nagarajan, G. Yu, M. Požek, N. Barišić, and M. Greven, "Percolative nature of the direct-current paraconductivity in cuprate superconductors," Quantum Mater. **3**, Article **42**, 1–6 (2018). https://doi.org/10.1038/s41535-018-0115-2
25. M. G. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, and Z. Wang, "A complete catalogue of high-quality topological materials," Nat. **566** (7745), 480–485 (2019). https://doi.org/10.1038/s41586-019-0954-4
26. Yu. I. Zhuravlev, "Correct algebras over sets of incorrect (heuristic) algorithms," I: Cybern. **13** (4), 489–497 (1977); II: Cybern. **13** (6), 814–821 (1977); III: Cybern. **14** (2), 188–197 (1978).
27. K. T Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nat. **559** (7715), 547–555 (2018). https://doi.org/10.1038/s41586-018-0337-2
28. J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, and B. Meredig, "Materials science with large-scale data and informatics: Unlocking new opportunities," MRS Bull. **41** (5), 399–409 (2016). https://doi.org/10.1557/mrs.2016.93
29. Yu. I. Zhuravlev, N. N. Kiselyova, V. V. Ryazanov, O. V. Sen'ko, and A. A. Dokukin, "Design of inorganic compounds with the use of precedent-based pattern recognition methods," Pattern Recogn. Image Anal. **21** (1), 95–103 (2011). https://doi.org/10.1134/S1054661811010135
30. P. V. Balachandran, J. Theiler, J. M. Rondinelli, and T. Lookman, "Materials prediction via classification learning," Sci. Rep. **5**, Article 13285, 1–16 (2015). https://doi.org/10.1038/srep13285
31. B. Bradlyn, L Elcoro, J. Cano, M. G. Vergniory, Z. Wang, C. Felser, M. I. Aroyo, and B. A. Bernevig, "Topological quantum chemistry," Nat. **547** (7663), 298–305 (2017). https://doi.org/10.1038/nature23268
32. F. Grasselli and S. Baroni, "Topological quantization and gauge invariance of charge transport in liquid insulators," Nat. Phys. **15**, 967–972 (2019). https://doi.org/10.1038/s41567-019-0562-0
33. I. Yu. Torshin, "The study of the solvability of the genome annotation problem on sets of elementary motifs," Pattern Recogn. Image Anal. **21** (4), 652–662 (2011). https://doi.org/10.1134/S1054661811040171
34. I. Yu. Torshin, "On solvability, regularity, and locality of the problem of genome annotation," Pattern Recogn. Image Anal. **20** (3), 386–395 (2010).
35. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction*, (Springer, New York, 2001).
36. A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design," Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **B58**, Part 3 (1), 364–369 (2002). https://doi.org/10.1107/S0108768102006948
37. Y. Xu, M. Yamazaki, and P. Villars, "Inorganic materials database for exploring the nature of material," Jpn. J. Appl. Phys. **50** (11S), Article 11RH02 (2011). https://doi.org/10.1143/JJAP.50.11RH02
38. I. Yu. Torshin, V. A. Alyoshin, and E. V. Antipov, "Synthesis and properties of the high-temperature superconductor $HgBa_2CuO_{4+d}$," Sverkhprovodimost: Fiz., Khim., Tekh. **7** (10-12), 1579–1587 (1994).
39. S. N. Putilin, E. V. Antipov, O. Chmaissem, and M. Marezio, "Superconductivity at 94 K in $HgBa_2Cu0_{4+\delta}$," Nat. **362**, 226–228 (1993). https://doi.org//10.1038/362226a0
40. H. Maeda; Y. Tanaka; M. Fukutomi, and T. Asano, "A new high-T_c oxide superconductor without a rare Earth element," Jpn. J. Appl. Phys. **27**, Part 2 (2), L209–L210 (1988). https://doi.org/10.1143/JJAP.27.L209
41. Ch. Chen, B. M. Wanklyn, E. Dieguez, A. J. Cook, J. W. Hodby, A. Schwartzbrod, A. Dabkowski, and H. Dabkowska, "Phase diagram and crystal growth of $Pb_2Sr_2(Y_xCa_{1-x})$ Cu_3O_{8+y}," J. Cryst. Growth **118** (1–2), 101–108 (1992). https://doi.org//10.1016/0022-0248(92)90054-M

Translated by I. Nikitin

Ivan Yur'evich Torshin was born in 1972. He graduated from the Department of Chemistry, Moscow State University, in 1995, received candidates degrees in chemistry in 1997 and in physics and mathematics in 2011. Currently he is an associate professor at Moscow Institute of Physics and Technology, lecturer at the Faculty of Computational Mathematics and Cybernetics, Moscow State University, a senior researcher at the Federal Research Center Computer Science and Control, Russian Academy of Sciences, and a researcher at the Center for Big Data Storage and Analysis, Moscow State University (https://bigdata-msu.ru). He is the author of 485 publications in peer-reviewed journals in informatics, medicine, chemistry, and biology, 9 monographs: 5 in Russian and 4 in English (in the series "Bioinformatics in Post-genomic Era", Nova Biomedical Publishers, NY, 2006-2009).

Konstantin Vladimirovich Rudakov was born in 1954. He is a Russian mathematician, academician of the Russian Academy of Sciences, Deputy Director of the Federal Research Center Computer Science and Control, Russian Academy of Sciences, Head of Department "Intelligence Systems" at the Moscow Institute of Physics and Technology, and academic advisor at the Center for Big Data Storage and Analysis, Moscow State University.

SPELL: 1. surjective