Topological Chemograph Analysis Theory As a Promising Approach to Simulation Modeling of Quantum-Mechanical Properties of Molecules. Part II: Quantum-Chemical Interpretations of Chemograph Theory

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

[a] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333, Russia*
**e-mail: tiy1357@yandex.ru*

Abstract—An interpretation of the problem-oriented theory developed is given from different perspectives of quantum chemistry. It is shown that the results obtained within the developed formalism correspond to the solution of the single-electron Schrödinger equation on molecular fragments, to the additive scheme of electron density calculation in the density functional theory, and to the correction of the integrals of overlapping in the molecular orbital theory. The algorithms based on the developed formalism were tested on a sample of 134 thousand molecules, for which the highest occupied molecular orbital (HOMO) energy, the lowest unoccupied molecular orbital (LUMO) energy, the HOMO–LUMO gap energy, the rotational constants, etc., were calculated by the B3LYP/6-31G(2df,p) method of quantum-mechanical calculations. The cross-validation testing of linear and nonlinear models has resulted in rank correlations between calculated and experimental values within a range of 0.67–0.85. In this case, the speed of calculations by the developed algorithms is higher than for quantum-mechanical calculations by eight orders of magnitude. The developed algorithms can be used for large-scale screenings of molecules when solving the problems of molecular pharmacology and materials science.

Keywords: algebraic approach, labelled graph analysis theory, quantum mechanics, simulation modeling
DOI: 10.1134/S1054661821040258

INTRODUCTION

In the first part of this paper [11], a scheme for additive prediction of numerical values on the basis of initial chemograph descriptions was developed. The application of ideology of chemometric analysis in the form of matching pairs of metrics over a set of chemographs has allowed us to formulate the problem of generating the synthetic features of chemographs in the form

$$\arg\min_{\{\omega_k\}} \frac{1}{n}\sum_{m=1}^{|X|}\left|\sum_{k=1}^{n}\omega_k \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_{m_0}]\boldsymbol{\pi} \oplus \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_m]\boldsymbol{\pi} - T_m\right|, \quad (1)$$

where T_m is the value of the predicted kth numerical variable for object X_m, ω_k is the weight of chemoinvariants, and the expression $\hat{\imath}[k]\hat{\boldsymbol{\beta}}$ describes the method used for the generation of invariants for the description of chemographs [11]. In addition, let us consider the interpretations of obtained results from the perspective of chemical bonding theory, the single-electron approximation to solution of the Schrödinger equation (SE), and molecular orbital and density functional theories.

1. INTERPRETATION OF RESULTS FROM THE PERSPECTIVE OF CHEMICAL BONDING THEORY

The set of χ-invariants, into which the given χ-graph vertex corresponding to one of the atoms in a molecule is involved, describes a certain local context of this atom in a molecule as a connected χ-graph's subgraph, the edges of which correspond to the covalent bonds between the atoms. The descriptor of the nearest context of an atom is a χ-node, and the set of all the chains, which have fixed length m' and pass through the χ-graph's vertex, describes the order m'-vicinity of the atom.

In the chemical bonding theory, which is a composite of quantum-mechanical (QM) and classic notions, an efficient approach to the description of geometry for the local surrounding of an atom is the calculation of hybrid atom states, which can be used for the generation of alphabet Y and dictionaries $\tilde{Y}$, $\hat{Y}$. Over the set of χ-chains $\boldsymbol{\alpha} \subseteq \tilde{Y}$ and the set of χ-nodes $\boldsymbol{\kappa} \subseteq \hat{Y}$, there are χ-fragments $\boldsymbol{\pi} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\boldsymbol{\kappa} \cup \hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\boldsymbol{\alpha}$,

† Deceased.

Received July 18, 2021; revised July 18, 2021;
accepted July 18, 2021

$|\boldsymbol{\pi}| = n$, and the set of χ-invariants $\boldsymbol{\iota}_{\mathbf{e}} = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\boldsymbol{\alpha} \cup \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\boldsymbol{\kappa}$ is formed. Weights ω_k in Eq. (1) correspond to the summation of a molecule property (characterized by the kth target variable) over χ-fragments (each of which is characterized by a certain contribution to this property). Correspondingly, Eq. (1) entails two hypotheses about the molecule "structure–property" interrelation.

Definition 1. Additivity hypothesis: an entire molecule's property is the sum of contributions from χ-invariants, each of which corresponds to a molecule's χ-fragment.

Definition 2. Fixed contribution hypothesis: each χ-invariant makes an identical contribution to the studied properties into all the molecules that contain the corresponding χ-fragments.

2. INTERPRETATION IN SINGLE-ELECTRON APPROXIMATION TERMS

The adiabatic SE approximation (see Eq. (1.2) in [11]) implies that internuclear interaction term $V_{nn}(\mathbf{R})$ is fixed for given nuclear configuration $\mathbf{R}$ because the nuclei have a negligible small motion velocity as compared to the electrons. Correspondingly, the coordinates of nuclei $\mathbf{R}$ are steady-state SE parameters. The single-electron approximation implies the further simplification of the Hamiltonian form by revising the method of accounting for interelectron interaction term V_{ee}: it is assumed that interelectron interaction is approximated by the sum of single-electron operators (so-called Hartree–Fock approximation).

In the single-electron approximation, ψ-function $\psi(\mathbf{r},\mathbf{R})$ of a molecule is sought in the form of a product of "single-electron" ψ-functions like $\psi(\mathbf{r},\mathbf{R}) = \prod_i \psi_1(\vec{r}_i,\mathbf{R})$ such that every $\psi_1(\vec{r}_i,\mathbf{R})$ is a solution for the problem about the motion of a single electron in the field of all the nuclei, i.e., a solution for the "single-electron SE":

$$\begin{gathered}\hat{h}_i\psi_{1,k}(\vec{r}_i,\mathbf{R}) = e_{ik}\psi_{1,k}(\vec{r}_i,\mathbf{R}),\\ \hat{h}_i = -\frac{\hbar^2}{2m_e}\Delta_i - \sum_{\alpha}\frac{Z_\alpha}{R_{\alpha i}} + V_i(i),\\ \hat{H}_e = \sum_i \hat{h}_i, \quad E_k = \sum_i e_{ik},\end{gathered} \tag{2}$$

where $V_i(i)$ is the single-electron term of potential energy, i.e., the average electrostatic field acting on the ith electron from the other electrons in a molecule, and $V_i(i) = \frac{1}{2}\sum_{i\neq j}\frac{1}{d_{ij}}$. Every single-electron function $\psi_{1,k}$ must be a solution of Eq. (2) on the condition that functions $\psi_{1,k}$ are normalized and linearly independent [5].

In practice, eigenfunctions $\psi_{1,k}$ in Eq. (2) are presented in the form of a determinant composed of single-electron linearly independent functions (the so-called "Stater determinant"). Such a form for the solution of Eqs. (2) is due to the need to meet the quantum mechanics postulate that electrons are indistinguishable [11].

The single-electron approximation is reduced to the approximation of the precise electron Hamiltonian by the sum of single-electron operators the Fockians $\hat{h}_i$. Correspondingly, the energy of a system is written as the sum of the Fockians' eigenvalues; i.e., $E_k = \sum_i e_{ik}$. Although Fockian $\hat{h}_i$ incorporates coordinates of all nuclei $\mathbf{R}$ similarly to the procedure of calculating the field, which acts on a single electron, the motion of electron will be influenced only by the nearest nuclei with an increase in the size of a molecule. In other words, the larger a molecule, the more adequate its representation as a set of localized fragments.

Accepting the hypothesis that an electron is predominantly delocalized around local structural fragments, Fockians $\hat{h}_i$ and single-electron ψ-functions $\psi_{1,k}(\vec{r}_i,\mathbf{R})$ can be written as sums over the local structural fragments of a molecule, i.e., over the local subsamples of atomic nuclei as subsets of joint vector $\mathbf{R}$. The latter corresponds to the summation of molecule's properties over χ-fragments in problem (1). In other words, the χ-fragments in prebasis $U(\mathbf{X})$ correspond to a certain set of single-electron wave functions, the linear combinations of which lead to the summation of eigenvalues over χ-invariants, and this corresponds to the fulfilment of additivity hypothesis (Definition 1).

3. INTERPRETATION IN MOLECULAR ORBITAL THEORY

Molecular orbital theory considers only the orbitals of valent electrons in the atoms composing a given molecule. For example, a molecular orbital may be presented as a linear combination (LC) of atomic orbitals (AOs). This assumption is associated with Fockians (2) corresponding to hydrogen-like ψ-functions at large (several angstroms) and small (fraction of angstrom) distances [1–5].

The LCAO approximation of molecular orbitals is rather convenient from the analytical perspective, as the orbitals of an atom, being the eigenfunctions of the Hermitian operator (atomic single-electron operator), are orthogonal to each other, thus enabling the total energy of a molecule to be approximated by the sum of energies over the orbitals of valent electrons. The following step in molecular orbital (MO) theory is the

construction of a molecular orbital for the entire molecule as a linear combination from the orbitals of individual fragments [12].

Let a random MO Ψ be presented as $\Psi = \sum_i c_i \psi_i$. Substituting this linear combination into the expression for calculating the average value of a physical parameter (see Postulate 2 in [11]) and introducing the normalization condition for Ψ for a pairwise nonorthogonal ψ_i, we obtain

$$\overline{A} = \frac{\sum_i \sum_j c_i c_j A_{ij}}{\sum_i \sum_j c_i c_j S_{ij}}, \quad (3)$$

where $A_{ij} = \int_{R^{3N}} \psi_i \hat{A} \psi_j d\mathbf{x}$ is the matrix elements of the operator for the calculation of a physical parameter (e.g., Hamiltonian) and $S_{ij} = \int_{R^{3N}} \psi_i \psi_j d\mathbf{x}$ is the matrix elements of the overlapping of ψ-functions (so-called "overlap integrals"), which characterize the degree of interelectron interaction for the functions ψ_i. If each of the functions ψ_i characterizes one or another AO of the ith atom in a molecule, the matrix (S_{ij}) can be unambiguously associated with an incidence matrix $(m_{ij}(X))$ of a chemograph X. In other words, incidence matrix $(m_{ij}(X))$ is a factorization of matrix (S_{ij}), in which higher values of overlapping integrals S_{ij} correspond to higher weights of edges m_{ij}. The latter implies that there is an interrelation between the overlapping integrals and the χ-invariant "weights" adjusted when solving machine-learning problem (1).

For orthogonal ψ_i and ψ_j, $S_{ij} = 0$. Within a single molecule, orthogonality is tested directly (when analytical expressions are known for all ψ_i and ψ_j) or indirectly (as the orthogonality of ψ-functions is identical to the distinguishability of their effects after inclusion into a LC).

To analyze the orbital overlapping effects from the perspective of chemograph theory, let us admit that there exists a finite, but rather large, set of chemographs $\mathbf{X}$, in which χ-fragments corresponding to one χ-invariant can be encountered in different chemographs (molecules). Let us consider two fragments of the same molecule from a set of precedents, every of which corresponds to a certain χ-invariant. Let χ-invariants be formed on the basis of χ-chains $\alpha, \beta \in \tilde{Y}$ such that χ-fragments $\hat{\mu}_c^{-1}\alpha$, $\hat{\mu}_c^{-1}\beta \in \tilde{\mathbf{\Pi}}$ are defined; the case with χ-nodes is considered in a similar way.

Let us consider all the molecular systems, which correspond to the chemographs in $\mathbf{X}$ and contain at least one of the two χ-fragments. Let us divide these systems into the three subsets: set of chemographs $AB = \{X \in \mathbf{X} | \hat{\mu}_c^{-1}\alpha \Subset X,\ \hat{\mu}_c^{-1}\beta \Subset X\}$ (which corresponds to the molecular systems containing two fragments) and subsets $A = \{X \in \mathbf{X} | \hat{\mu}_c^{-1}\alpha \Subset X\} \backslash AB$ and $B = \{X \in \mathbf{X} | \hat{\mu}_c^{-1}\beta \Subset X\} \backslash AB$. For certainty, let $|\hat{\mu}_c^{-1}\alpha| = 1$ and $|\hat{\mu}_c^{-1}\beta| = 1$ be fulfilled (the cases with $|\hat{\mu}_c^{-1}\alpha| > 1$, $|\hat{\mu}_c^{-1}\beta| > 1$ have no fundamental distinction, but require excessively complicated formalism).

According to the hypothesis of fixed contributions of χ-invariants (Definition 2) we assume that, in each of the molecular systems corresponding to sets of chemographs A, B, and AB, all χ-fragments $\hat{\mu}_c^{-1}\alpha$ are described by the same ψ-function ψ_A, while all χ-fragments $\hat{\mu}_c^{-1}\beta$ are described by the same ψ-function ψ_B. It is obvious that the interaction between ψ_A and ψ_B depends on the distance between them in each molecule from set AB: the longer is the distance between χ-fragments, the lower are the values of corresponding overlapping integral S_{AB}, and the more adequate is the description of ψ_{AB} as linear combinations of ψ_A and ψ_B.

Technically, distance d_{AB} between χ-fragments can be estimated by different methods, e.g., as an average distance between the atoms of fragments (centroids), as a minimum distance (i.e., distance to the nearest atom of another fragment), as a length of the shortest path in a chemograph between two χ-fragments (if there is a shared vertex, $d_{AB} = 0$), or as a length of the maximum path among the shortest paths when the number of shared vertices is taken into account ($d_{AB} = 0$ if all vertices are shared), etc. From the perspective of MO theory, the physical meaning of the hypothesis that the contributions from each of the χ-invariants have an additive character (i.e., they are independent of all the other contributions, Definition 1) is confirmed by the following theorem.

Theorem 1 (on the additive correction of interactions between the χ-fragments of a chemograph). *Let the quantitative estimate of interaction between two χ-fragments in a chemograph monotonically descend with an increase in the distance between the χ-fragments and be independent of the types of χ-fragments, the arrangement of χ-fragments in the context of a chemograph, and the selected method of measuring the distance between the fragments. The contributions of any two χ-fragments can then be considered as independent, and the correction for the interaction between a pair of χ-fragments can be taken into account as the contribution from the third χ-fragment, which forms a connected subgraph with this pair of fragments.*

Proof. Independently of the procedure used for the generation of χ-fragments and the method of measuring the distance between χ-fragments, if two fragments of a molecule have no the third fragment connecting them, the contributions of both χ-fragments to the overall property of a molecule may be consid-

ered independent so the overlap of such remote orbitals may be neglected ($S_{AB} = 0$). The feasibility of this condition depends on the method used for the generation of sets $\tilde{Y}(X)$ and $\hat{Y}(X)$ of chemoinvariants.

Let two χ-fragments $\hat{\mu}_c^{-1}\alpha$ and $\hat{\mu}_c^{-1}\beta$, $\left|\hat{\mu}_c^{-1}\alpha\right| = 1$, $\left|\hat{\mu}_c^{-1}\beta\right| = 1$, according to Definitions 1 and 2, make contributions ω_α and ω_β to the property of a molecule, and $\omega_{\alpha\beta}$ describes the contribution to the property of a molecule from the interaction between χ-fragments (i.e., $\omega_{\alpha\beta}$ is the correction for the interaction between $\hat{\mu}_c^{-1}\alpha$ and $\hat{\mu}_c^{-1}\beta$). Let us note that the case of $\left|\hat{\mu}_c^{-1}\alpha\right| > 1$ is trivial and merely corresponds to $\omega_\alpha\left|\hat{\mu}_c^{-1}\alpha\right|$.

According to the theorem, let the interaction between χ-fragments be determined by distance $d_{\alpha\beta}$ between them such that correction $\omega_{\alpha\beta}$ is calculated as monotonically descending function f^- of $d_{\alpha\beta}$ $\omega_{\alpha\beta} = f^-(d_{\alpha\beta})$; i.e., the longer distance $d_{\alpha\beta}$, the smaller $\omega_{\alpha\beta}$. The contribution of both fragments and their correction for their interaction to the property of a molecule can then be described as $\omega_\alpha + \omega_\beta + f^-(d_{\alpha\beta})$.

Let chemograph X contain m χ-fragments, $\mu_1, \mu_2, \ldots, \mu_m$, $\mu_i \in \mathbf{S}(X)$. Let us consider the distances from ith χ-fragment μ_i to all the other χ-fragments μ_j, $i \neq j$. The correction introduced for the i,jth pair of contributions is equal to $f^-(d_{ij})$ and, taking into account all the other fragments, depends on distances $\{d_{ij}\}$ to the ith fragment. For fragments with identical distributions of $\{d_{ij}\}$, the correction is the same and equal to $s_i = \sum_{j\neq i} f^-(d_{ij})$.

The distributions of distances $\{d_{ij}\}$ depend on the centrality of a fragment: the closer is it to the center of a molecule, the lower the number of longer paths (corresponding to lower $f^-(d_{ij})$), and the higher the number of shorter paths (corresponding to higher $f^-(d_{ij})$). By definition, the center of a graph is the set of the ith vertices with minimum values of ith chemograph vertex "centrality" sums $c_i = \sum_{j=1,m} d_{ij}$. Since the fragments with the same centrality c_i have equal corrections s_i, let us consider two fragments with different centralities (i.e., with different distributions of $\{d_{ij}\}$). Let i_1 be more peripheral and i_2 be more central, so that $c_{i_1} \geqslant c_{i_2}$ and $s_{i_1} \leqslant s_{i_2}$. If the procedure for the generation of χ-fragments is such that there exists third fragment i_3 from $\mathbf{S}(X)$ to connect χ-fragments μ_{i_1} and μ_{i_2}, $c_{i_1} \geq c_{i_3} \geq c_{i_2}$ and, correspondingly, $s_{i_1} \leq s_{i_3} \leq s_{i_2}$.

According to the aforesaid, the contributions of χ-fragments μ_{i_1} and μ_{i_2} to the property of molecules are estimated as $\omega_{i_1} + \omega_{i_2} + f^-(d_{i_1i_2})$. At the same time, for χ-fragment i_3, weight ω_{i_3} is defined such that the contribution all the three fragments is equal to $\omega_{i_1} + \omega_{i_2} + \omega_{i_3} + f^-(d_{i_1i_2}) + f^-(d_{i_1i_3}) + f^-(d_{i_2i_3})$. However, according to the theorem, the interaction between χ-fragments is independent of the arrangement of χ-fragments in the context of a chemograph; so, the description of interactions between random χ-fragments μ_{i_1} and μ_{i_2} must be independent of their centrality. The latter corresponds to $s_{i_1} = s_{i_3} = s_{i_2} = s$. The sum of corrections $\omega_{i_1i_2}$ over all the pairs of χ-fragments μ_{i_1} and μ_{i_2} is then equal to $\sum_{i=1,m} s_i$, i.e., is a product of the number of chemograph χ-fragments (m) and a constant (s). Correspondingly, contribution ω_{i_3} of the third fragment μ_{i_3} can be considered as a correction for the interaction between two fragments μ_{i_1} and μ_{i_2}. The theorem is proven.

Corollary 1. The condition of theorem corresponds to the sets of χ-fragments, which are formed via the complete enumeration of subgraphs of corresponding type (χ-chains and χ-nodes).

Corollary 2. The property of the entire molecule can be calculated by summation over χ-fragments. The sum remains unchanged after permutation of the summands.

Corollary 3. According to the theorem, when the interactions between χ-fragments are taken into account, the calculated property of a molecule incorporates term ms equal to the product of the number of chemograph χ-fragments m and constant s. This follows from the condition that $s_{i_1} = s_{i_3} = s_{i_2} = s$.

Corollary 4. When the theorem condition is fulfilled, property W of molecule X is calculated by the additive scheme as $W = \sum_{i=1,m} \omega_i + ms$, *i.e.,* by summation over m χ-fragments, $\mu_i \in \mathbf{\Pi}(X)$. It follows from this that ψ_i are orthogonal (i.e., their overlap integrals are zero).

Theorem 1 with corollaries shows that even the simplest additive scheme of calculating the properties of a molecule in compliance with problem formulation (1) makes it possible to take into account the overlap integrals within the framework of MO theory. This becomes possible under certain conditions imposed on the construction of sets $\tilde{Y}(X)$ and $\hat{Y}(X)$ of chemoinvariants. Moreover, the addition of χ-invariants corresponding to the "third" χ-fragments (i.e., the fragments composing a connected subgraph for a random pair of χ-fragments) to sets $\tilde{Y}(X)/\hat{Y}(X)$ allows us to presume the orthogonality of ψ-functions corresponding to χ-invariants in such $\tilde{Y}(X)/\hat{Y}(X)$ (corollary 4).

There are several approaches to experimental verification of the entire complex of hypotheses under the condition of Theorem 1.

First, it is possible to estimate the computational accuracy of the additive scheme, which is implied in Corollary 4 (Theorem 1) and corresponds to the formulation of chemometric analysis problem (1). The degree of correlation between the calculated and experimental QM parameters of molecules is an indirect method of estimating the reasonability of the introduced complex of hypotheses (Definitions 1 and 2).

Second, the interactions of functions ψ_A and ψ_B corresponding to χ-invariants in sets $\tilde{Y}$ and $\hat{Y}$ can be estimated by measuring the distances between the corresponding χ-fragments of set AB. Let us point out that overlap integral S_{AB} of ψ-functions ψ_A and ψ_B, which correspond to two certain χ-invariants, is calculated within a single molecule, and the distances between the corresponding χ-fragments are determined for sets of molecules A, B, and AB. (The latter, by the way, implies a kind of "averaging" for functions ψ_A and ψ_B and S_{AB} over a sample.)

To estimate the distances between χ-invariants α and β over sample of molecules AB, let us calculate the empirical distribution function (e.d.f.) of distances $\{d_{AB}\}$ between χ-fragments $\hat{\mu}_c^{-1}\alpha$ and $\hat{\mu}_c^{-1}\beta$. Let us note that neither the value of $\left|\hat{\mu}_c^{-1}\alpha\right|$ ($\left|\hat{\mu}_c^{-1}\alpha\right| = 1$ or $\left|\hat{\mu}_c^{-1}\alpha\right| > 1$) nor the type of an χ-invariant (χ-chain or χ-node) is of fundamental significance for e.d.f. construction. The obtained e.d.f. is used as a basis to construct metric ρ_L, which calculates the average distance between two χ-invariants (e.g., as the mathematical expectation of e.d.f.). In addition, some other definitions of metric ρ_L can also be used for the analysis of "distances" between χ-invariants (see [11]).

Afterwards, all the pairs of χ-invariants found in the chemographs from set $\mathbf{X}$ are used as a basis to form the ρ_L-configuration of distances between χ-invariants and analyze the distribution of distances in this ρ_L-configuration and the existence of metric concentrations [6–10]. In particular, the proximity of the ρ-spectra of a constructed ρ_L-configuration to the Gaussian distribution or negligible deviations of each i-spectra of a ρ_L-configuration from the ρ-spectra evidence that the ρ_L-configuration is close to a space of isolated points (which correspond to the same values of all distances ρ_L in the ideal case). The latter is evidence for the mutual orthogonality of χ-invariants and their corresponding "averaged" ψ-functions.

On the whole, from the perspective of MO theory, the delocalization of electrons about the χ-fragments corresponding to χ-chains and χ-nodes is implied. The sharing of an electron in the case of χ-node is quite feasible, as the latter represents the nearest covalent environment of any atom of a molecule. Preliminary experiments have shown that, in most cases, χ-chains correspond to π-systems (chains of conjugated double bonds) or fragmented π-systems (e.g., two double bonds spaced apart by two ordinary bonds). That there is delocalization of electrons in π-systems is commonly known.

4. INTERPRETATION OF THE RESULTS FROM THE PERSPECTIVE OF DENSITY FUNCTIONAL THEORY

Density functional theory (DFT) is a very important contemporary area in QM, which not only provides the possibility to derive physically interpretable SE forms, but also makes it possible to increase the precision of QM calculations. The central idea of DFT is the reformulation of quantum-mechanical Postulates 1–4 [11] in the forms corresponding to the transition from the Ψ-functions of a molecular system to the distribution of its electron density:

$$\rho(\vec{r}) = N \int_{R^{3N}} \left|\Psi(\mathbf{x})\right|^2 d\mathbf{x}. \quad (4)$$

From the physical perspective, the advantages of such an approach are obvious: electron density $\rho(\vec{r})$ has a clear physical meaning and is experimentally measurable (e.g., in diffraction experiments). The transition from highly dimensional configuration space R^{3N} to the "physical" Cartesian space, in which vectors $\vec{r}$ are defined, essentially improves the interpretability of QM functionals in density functional theory.

Within the framework of density functional theory, it has been shown that a state with a minimum energy (i.e., the ground state) has transform F reciprocal to (4) such that the Ψ-function can be written as an electron density functional; i.e., $\Psi(\mathbf{x}) = F(\rho(\vec{r}))$. Correspondingly, the average value of physical parameter A is calculated as $\overline{A} = \left\langle F(\rho(\vec{r}))\middle|\hat{A}\middle|F(\rho(\vec{r}))\right\rangle$ (Postulate 2). For instance, the total system's energy (in the Born–Oppenheimer approximation) can be expressed as a functional of $\rho(\vec{r})$ as $E(\rho(\vec{r})) = F_{HK}(\rho(\vec{r})) + V_{en}(F(\rho(\vec{r})))$, $F_{HK}(\rho(\vec{r})) = \hat{T}F(\rho(\vec{r})) + V_{ee}(F(\rho(\vec{r})))$. The problem of searching the form for functional $F_{HK}(\rho(\vec{r}))$ (Hohenberg–Kohn) is the central problem in density functional theory [2].

From the perspective of analyzing the results from the application of topological chemograph theory (Eq. (1) and others), the electron density calculation procedures naturally admit the additive models, which factorize the contribution of different energy terms to summary $\rho(\vec{r})$ (in contrast to the Ψ-functions, which describe the entire system as an integer whole). Chemical bonding theory [3, 8, 9], which gives a list of rules for the generation of chemograph incidence matrix, also makes it possible to obtain a rough map of elec-

tron density $\rho(\vec{r})$ in molecules. It is quite clear that the covalent bonds corresponding to chemograph edges are characterized by an increase in the electron density in the internuclear space in comparison with the sum of the electron densities of free atoms. The precision of reproducibility for $\rho(\vec{r})$ depends on the selection of covalent radii, polarization, and a number of other quantitative parameters used in chemical bonding theory [2, 5].

Correspondingly, the representation of a molecule as a set of χ-fragments is a method for the additive fragment-wise description of the electron density map of a molecule. The averaging of $\rho(\vec{r})$ over the fragments surrounding every atoms (χ-node) makes it possible to develop schemes for the prediction of atom-wise molecule properties (e.g., the partial charge, reaction centers of molecules, etc.) in the process of learning on the set of labelled chemographs with numerical estimates of their vertices/edges.

5. ON THE APPLICATION OF FORMALISM FOR SOLUTION OF MOLECULAR SCREENING PROBLEMS

Additive chemometric analysis scheme (1) representing the summation of weighed feature values with the function of losses in the form of a module may have several QM interpretations. From the perspective of chemical bonding theory, the problem (1) corresponds to the introduction of hypotheses about the additivity and constancy of contributions from χ-invariants. Within the single-electron approximation, the χ-fragments in prebasis $U(\mathbf{X})$ correspond to a certain set of single-electron wave functions. From the perspective of MO theory, the defined methods for constructing the sets of χ-invariants make it possible to take into account the overlapping integrals of χ-fragments for delocalization of electrons χ-fragments. From the perspective of density functional theory, scheme (1) can be interpreted as a method for the additive description of the electron density map of a molecule.

From all these facts, it is quite clear that the precision of calculations by additive scheme (1) can hardly approach that of results obtained by earlier-developed computational schemes of quantum mechanics due to rather strong assumptions about the additivity and the constancy of contributions of χ-invariants (Definitions 1 and 2). Nevertheless, a number of specific features characteristic of the correlation cloud of points $O(\mathbf{X}) = \{(W_m(X), T_m(X)), X \in \mathbf{X}, m = 1,\ldots,|\mathbf{X}|\}$, where $W_m(X)$ is calculated for all $X \in \mathbf{X}$ (e.g., in compliance with the solution of problem (1)), make it possible to estimate the practical applicability of the obtained results.

First, cross-validation estimates of the correlation coefficient and the other functionals of estimating the adequacy of models on $O(\mathbf{X})$ (standard deviation, determination coefficient, robust linear smoothing functionals, different statistical functionals, etc.) provide a comprehensive quality estimation of simulation modeling values of T by means of model W. Second, the relevance of characteristics can be estimated for cloud $O(\mathbf{X})$ of points for solving the corresponding problems of the classification and high-performance screening of molecules in silico.

Theorem 2 (screening theorem). *The precision in the classification of chemographs from* $\mathbf{X}$ *is proportional to the degree of covering correlation cloud* $O(\mathbf{X}) = \{(W_m, T_m)\}$ *of points by the cells of the major diagonal of a coordinate grid formed by the corresponding percentiles of values T and W.*

Proof. Let target variable T in lattice $L(T(\mathbf{X}))$ correspond to chain $A(\mathbf{X})$), and let the estimates obtained by means of model W correspond to chain $A'(\mathbf{X})$). Let us construct e.d.f. for both chains $\mathrm{cdf}(A(\mathbf{X})) = \{(\lambda_i, P_i)\}$ and $\mathrm{cdf}(A'(\mathbf{X})) = \{(\lambda'_j, P'_j)\}$, $i, j = 1,\ldots,|\mathbf{X}|$. For a specified number of percentiles N_P ($N_P = 2, 3, \ldots, 10$, etc.), let's determine the subsets, which belong to set $\mathbf{X}$ and correspond to the pth percentiles of values T and W, as $\Pi(p, \mathrm{cdf}(A(\mathbf{X}))) = \left\{(\lambda_i, P_i)\middle|\frac{p-1}{N_P} < P_i \le \frac{p}{N_P}\right\} \subset \mathrm{cdf}(A(\mathbf{X}))$ and $\Pi(p, \mathrm{cdf}(A'(\mathbf{X}))) = \left\{(\lambda'_j, P'_j)\middle|\frac{p-1}{N_P} < P'_j \le \frac{p}{N_P}\right\}$, respectively.

Let us define the class of chemographs interesting to a researcher in terms of percentiles $\Pi(p, \mathrm{cdf}(A(\mathbf{X})))$ and introduce the class membership function f_C such that $f_C(X, \Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = 1$, if chemograph X belongs to a "positive" class of chemographs; $f_C(X, \Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = -1$, if X belongs to a "negative" class; and $f_C(X, \Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = 0$ otherwise. Function f_C may be monotonical or non-monotonical.

Let us postulate that there is one-to-one correspondence between percentiles $\mathrm{cdf}(A(\mathbf{X}))$ and $\mathrm{cdf}(A'(\mathbf{X}))$; i.e., at the same value of p, percentile $\Pi(p, \mathrm{cdf}(A(\mathbf{X})))$ is always associated with percentile $\Pi(p, \mathrm{cdf}(A'(\mathbf{X})))$. The values $p = 1, \ldots, N_P$ correspond to the sequence of cells on the major diagonal of a coordinate grid formed by corresponding percentiles $\Pi(p, \mathrm{cdf}(A(\mathbf{X})))$ and $\Pi(p, \mathrm{cdf}(A'(\mathbf{X})))$. Hence, a chain of percentile cells has been formed on correlation cloud $O(\mathbf{X})$ of points.

Let us consider the ideal case, in which the correlation cloud is a straight line (i.e., correlation coefficient $r(O(\mathbf{X})) = 1.0$, and standard deviation $\mathrm{std}(O(\mathbf{X})) = 0$). The above defined mutually equivalent percentiles then cover all the correlation cloud points, $\Pi(p, \mathrm{cdf}(A(\mathbf{X}))) = \Pi(p, \mathrm{cdf}(A'(\mathbf{X})))$ for all p, and the classification error is zero.

Let us consider the more general case with $r(O(\mathbf{X})) < 1$ and $\mathrm{std}(O(\mathbf{X})) > 0$. Let us select the percentiles of values T with $f_C(X, \Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = 1$. For each of these percentiles, all the correlation cloud

points lying below the cell of a percentile on the planar diagram of cloud $O(\mathbf{X})$ correspond to false positive classification errors (values T are lower than for W). The cloud points "on the left" from the percentile rectangle are false negative errors (values T are higher than the predicted W). Summing the values $|\Pi(p, \mathrm{cdf}(A(\mathbf{X})))\backslash\Pi(p, \mathrm{cdf}(A'(\mathbf{X})))|$, $|\Pi(p, \mathrm{cdf}(A'(\mathbf{X})))\backslash\Pi(p, \mathrm{cdf}(A(\mathbf{X})))|$ over all the percentiles with $f_C(X,\Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = 1$ and further over the percentiles with $f_C(X,\Pi(p, \mathrm{cdf}(A(\mathbf{X})))) = -1$, we find classification errors of the first and second types. Based on the calculation of the numbers of errors, it is possible to find different accuracy estimates for a classification algorithm. The theorem is proven.

Corollary 1. Correlation coefficient $r(O(\mathbf{X}))$ is an indirect characteristic of accuracy of classification by the percentiles of values.

Corollary 2. The difference between the correlation coefficients for learning and control indirectly characterizes the overfitting of a classification algorithm.

The use of Theorem 2 may be useful for the design of computational experiments and the analysis of obtained data. For example, if $O(\mathbf{X})$ in a cross-validation experiment is such that the first and last quartiles do not overlap, the precision of classification by the first and last quartiles will be close to 100%. Such SM precision is quite satisfactory for the problems of the large-scale screenings of molecules and crystals.

6. ON CONTEMPORARY METHODS OF CALCULATING THE QUANTUM-MECHANICAL PARAMETERS OF MOLECULES

The testing of developed SM procedures requires maximally precise estimates for the quantum-mechanical properties of molecules. For instance, the use of single-electron approximation (2) is characterized by a very low precision, as it is necessary to take into account the electron correlation effects. DFT methods are most acceptable for the precise calculation of molecular geometry, and the errors in the calculation of thermodynamic parameters are high (2–3 kcal/mol) [2]. High-precision calculations (error, ~0.1 kcal/mol) are characterized by the use of hybrid schemes incorporating the elements of single-electron approximation, MO theory, and DFT approach [4].

For example, at rather long (several angstroms) or, vice versa, very small (several fractions of angstrom) distances between particles, Fockians $\hat{h}_i$ (2) correspond to the hydrogen-like ψ-functions and can be written in the form of polynomials, in which only the senior term multiplied by an exponent is essential alone. This allows us to introduce so-called Gaussian-type orbitals (GTOs), in which the radial components of Ψ-functions are specified in the form of parametric functions $R_v(r) = r^{k_v}e^{-\xi_v r^2}$, $k_v, \xi_v \in R$, $v = 1,\ldots,N_b$, where N_b is the size of a basis. When searching for the best MO approximations, the bases composed of linear combinations of such radial components are used in combination with semi-empirical potentials, DFT functionals, etc. [2]. For the sample of molecules analyzed in the present study (see the following section), the values of different quantum-mechanical parameters of molecules were determined by using the B3LYP/6-31G(2df,p) Gaussian basis of "mixed" semi-empirical potential B3LYP and the Gaussian-9 software [1]. The calculations incorporated the optimization of molecular geometry (coordinates of nuclei $\mathbf{R}$) by the B3LYP/6-31G(2df,p) method from DFT in the single-electron approximation [4].

7. RESULTS OF SIMULATION MODELING OF QUANTUM-MECHANICAL PARAMETERS

The models for the generation of the informative numerical features for chemographs via the solution of the problem (1) were tested together with the corresponding algorithms for the prediction of numerical variables [10] on a sample of 134 thousand stable small organic molecules with a maximum of nine "heavy" atoms C, O, N, and F (no more than 20 atoms; hereinafter, the sample will be denoted as 134K). For this sample, the geometric, energy, electronic, and thermodynamic properties were calculated earlier using QM techniques [4].

The initial descriptions of chemographs in set $\mathbf{X}$ were presented by interatomic chemical bond-multiplicity matrices $\mathbf{M}(X)$ with specified types of chemical atoms. To obtain set Q of precedents with the further application of Theorem 1, the transition from matrices $\mathbf{M}(X)$ to tuple invariants on the basis of χ-chains and χ-nodes was performed as described in the first part of this paper [1]. To test the algorithms following from the proposed formalism, alphabet Y incorporating the elements of the Cartesian product of the chemical type of an element and the charge and admissible hybridization states of atoms was used. The size of this alphabet was $|Y| = 44$ when the effects of hydrogen atoms were taken into account, and much smaller when hydrogen atoms were ignored ($|Y| = 14$). Additional examples of Y-alphabets were detailed in [9].

To determine the optimal values for parameters n and k, which are used to generate the feature descriptions of chemographs, the combinatorial testing of invariants from families $\hat{\mathfrak{i}}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{Y}^n$ ($n = 1$–7), $\hat{\mathfrak{i}}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(k)$ ($k = 3$–7), and $\hat{\mathfrak{i}}\hat{\boldsymbol{\beta}}[\mathbf{X}](\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{Y}^n \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(k))$ for completeness was carried out. The functionals based on the precise Fisher test were used to produce the functions for enumeration of the elementary invariants, $\lambda : \boldsymbol{\iota}_{\mathbf{e}} \to N$. For each tuple invariant, local completeness estimates $r_{\mathfrak{i}}(\mathfrak{i}, Q)$ (see [11]) were calcu-

Table 1. Results of cross-validation testing of simulation modeling algorithms for 15 QM parameters (r and $r(c)$ are the average rank correlation coefficients for learning and control, respectively)

Constant	QM parameter	Units	r	$r(c)$
A	Rotation constant A	GHz	0.77	0.73
B	Rotation constant B	GHz	0.74	0.73
C	Rotation constant C	GHz	0.72	0.71
M	Dipole moment	Debye	0.72	0.72
A	Isotropic polarizability	$Bohr^3$	0.69	0.67
HOMO	Energy of highest occupied molecular orbital	Hartree	0.82	0.79
LUMO	Energy of lowest unoccupied molecular orbital	Hartree	0.85	0.83
Gap	Gap difference between LUMO and HOMO	Hartree	0.86	0.83
r2	Electronic spatial extent	$Bohr^2$	0.67	0.67
ZPVE	Zero point vibrational energy	Hartree	0.85	0.85
U0	Internal energy at 0 K	Hartree	0.69	0.67
U	Internal energy at 298.15 K	Hartree	0.69	0.67
H	Enthalpy at 298.15 K	Hartree	0.69	0.67
G	Free energy at 298.15 K	Hartree	0.69	0.67
Cv	Heat capacity at 298.15 K	cal/M K	0.75	0.75

lated on a regular set of chemographs from the 134K set. The results show that $k = 4$ and $n = 5$ correspond to an acceptably high values of the local completeness of family invariants, i.e., $r_\iota(\iota, Q) = 0.97$. In other words, the set of χ-invariants $\hat{\imath}\hat{\boldsymbol{\beta}}\mathbf{[X]}(\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{Y}^5 \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(4))$ makes it possible to distinguish 97% of the pairs of the chemographs produced from the 134K dataset.

Similarly to the study [10], the prediction of numerical values was performed via algorithmic compositions $\hat{A}(\theta(\mathrm{Pr})) = B(\theta(\mathrm{Pr})) \circ C(\theta(\mathrm{Pr})) \circ D(\theta(\mathrm{Pr}))$, which are commonly adopted in Zhuravlev's science school and incorporate a recognizing operator B, a corrector operation C and a decisive rule D, $\theta(\mathrm{Pr})$ being the vector of inner parameters of the algorithm. *Linear recognizing operator B*, which generates the synthetic numerical features of chemographs, was constructed in compliance with the additive scheme (1) on the basis of the set $\hat{\imath}\hat{\boldsymbol{\beta}}\mathbf{[X]}(\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{Y}^5 \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(4))$ of χ-invariants. The synthetic features produced by the operator B were the predictions of 15 QM molecular parameters determined on the basis of high-precision QM calculations for molecules from sample 134K [11] (see the list in the Table 1). The optimal weights of chemoinvariants in scheme (1) were determined by means of multistart stochastic optimization [10].

The corrector C was constructed as the Cartesian product of the set of synthetic features obtained in compliance with scheme (1) and the vector of six correcting operations (linear transformation, logarithm, exponent, power function of three types $y = x^v$, $v = 1.5, 2, 3$) such that the summary number of synthetic features after application of the operator $B(\theta(\mathrm{Pr})) \circ C(\theta(\mathrm{Pr}))$ was equal to 90.

Linear models (in which the vector of parameters θ was optimized by means of multistart stochastic optimization or singular value decomposition) or nonlinear models (so-called neural network of three levels) were used as *decisive rule D*. The efficiency of the application of different compositions $\hat{A}(\theta(\mathrm{Pr}))$ or of the methods for fitting the inner parameters θ was estimated in cross-validation experiments, in which sampling operator $\hat{\zeta}$ (see the first part of this paper [11]) was determined such as to form set of samples $\hat{\zeta}\mathbf{X}$ incorporating ten arbitrary divisions of the set of chemographs $\mathbf{X}$ into "case-control" pairs of groups at the size ratio of 6 : 1.

The results of cross-validation experiments have shown that the "topological" algorithms $\hat{A}(\theta(\mathrm{Pr}))$ optimal for the prediction of the 15 studied QM molecular parameters correspond to the following requirements: (1) the correction for the effects of the hydrogen atoms (the use of the enhanced alphabet, $|Y| = 44$), (2) the use of the linear decisive rule D, (3) the use of stochastic optimization to find parameters θ, and (4) the use of correction for the number of chemograph χ-fragments (see Corollaries 3 and 4 from Theorem 5 in the first part

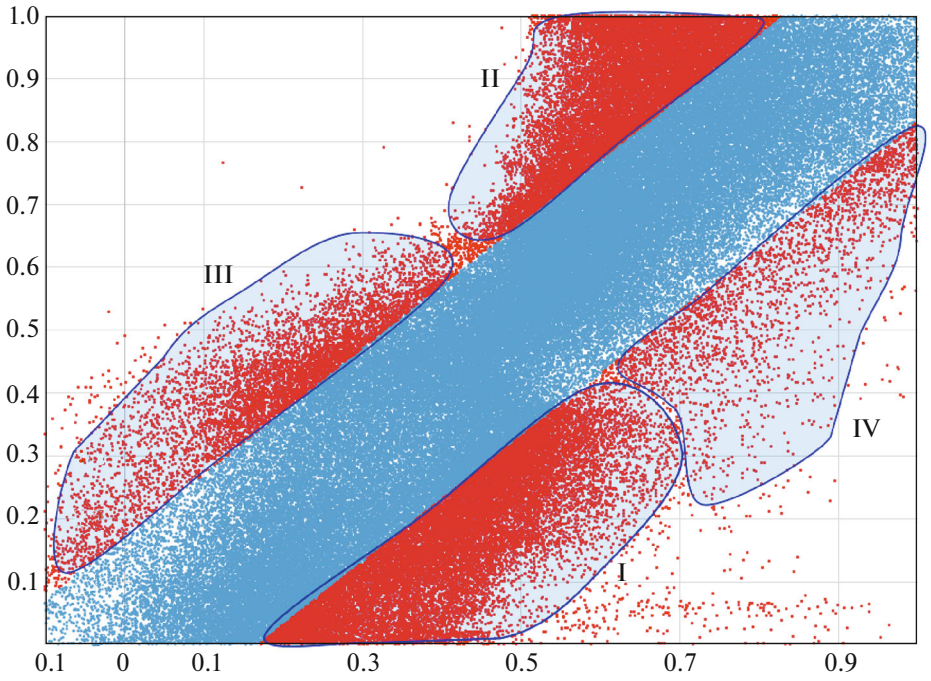


Fig. 1. Analysis of errors in LUMO–HOMO calculations (arb. un.) with regions of the most typical errors. The band in the figure center corresponds to the standard deviation of the topological SM algorithm.

of this paper [11]). The results of the numerical experiments are summarized in Table 1.

It is clear from the data of Table 1 that the rank correlations between the calculated and experimental values were in range 0.67–0.85. Despite such essential distinctions between the accuracies of the simulation modeling of different QM properties of molecules, the algorithms of the simulation modeling of all 15 properties were characterized by acceptable generalizing ability. The latter can be indirectly characterized by the differences between the rank correlation coefficient values (Theorem 2, Corollary 2), which were obtained in the course of learning (r) and control ($r(c)$) and attained only 0.016 in the average (95% CI of 0.003–0.041).

One of the best SM algorithms developed within this topological approach was the algorithm to estimate the gap between the highest occupied molecular orbital and the highest occupied molecular orbital (LUMO–HOMO gap width): r = 0.86 for learning and $r(c)$ = 0.83 for control at a standard deviation of 0.14–0.17. Despite the apparently smeared character of respective correlation cloud $O(\mathbf{X})$ (see an example in the Figure 1), 92% of the points in the quartile of maximum "topological" values for the LUMO–HOMO gap width corresponded to the quartile of the largest "quantum" LUMO–HOMO gap widths. Here, the quartiles of the largest "topological" and "quantum" LUMO–HOMO gap widths overlap each other up to 77% of points. In other words, in a large-scale screening of molecules by the LUMO–HOMO gap width, the quartile of maximum values found by the SM algorithm makes it possible to select 77% of compounds with the highest values of this molecular property.

The analysis of correlation between the number of atoms in a molecule and the error of "topological" calculations has shown that a maximum error was observed for molecules with 8–12 atoms and descended with an increase in the number of atoms ($r = -0.33$). Let us note that maximum errors (of 50% and higher) were observed for the molecules containing five hydrogen atoms (as a rule, aliphatic hydrocarbons with the general formula C_nH_{2n+2}). Contrarywise, a higher number of double bonds in a molecule corresponded to a lower error. A decrease in the error with an increase in the number of atoms in a molecule indicates that "topological" calculations may be extended to more complicated molecular systems.

The four regions of the most typical errors of the resulting SM algorithm of calculation of the LUMO–HOMO gap are detailed in the Fig. 1. The central band of points in Fig. 2 (region 0) corresponds to the chemographs, for which the differences between the "topological" and the "quantum" LUMO–HOMO gap widths were less than one standard deviation. The highest number of errors (8% of the 134K dataset) was in the region I. Hereinafter, "error" will mean the SM result, which differs from the "quantum" value by more than one standard deviation.

Region I corresponds to the domain of rather low LUMO–HOMO gap-width values, which were overestimated (on average, by 0.15) by the "topological"

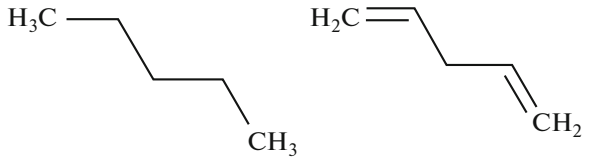


Fig. 2. Model configurations: aliphatic C–C–C–C–C χ-chain (left) and χ-chain corresponding to the C=C–C–C=C π-system (right), for which QM calculations were carried out by the B3LYP/6-31G method.

SM algorithm. The analysis of the atomic composition and of the structural formulas of chemographs related to the region I has shown that they contain twice as many hydroxyl groups (–OH), ten times more NH_4^+ (ammonium) derivatives, and 11 times more aldehydes (–CHO) in comparison with the chemographs from region 0.

Region II corresponds to the underestimation of high LUMO–HOMO gap widths (0.7–1.0) (by –0.20). The chemographs in the region II have five to ten times fewer double bonds and 2 times more fluorine atoms as compared to the other regions. Hence, aliphatic (saturated) fluorine-containing compounds are more frequently encountered in region II.

The fact that aliphatic chains represent a significant source of error corresponds to the considerations stated in the section on MO theory related to the delocalization of electrons around χ-fragments. Aliphatic chains are characterized by lower rotation thresholds than are chains with double bonds, and a fluorine atom (the most electronegative chemical element) makes an essential contribution to the redistribution of local charges. Both of these considerations prevent delocalization of the electrons around the relevant χ-fragments.

In order to estimate the delocalization of electrons in aliphatic and double-bond systems we performed QM calculations of model chains C–C–C–C–C (aliphatic, i.e., containing only ordinary carbon–carbon bonds, with all the others being carbon–hydrogen bonds) and C=C–C–C=C (a π-system, the unsaturated chain which has two double C=C bonds, while the others are ordinary carbon–carbon bonds). QM calculations were performed by the B3LYP/6-31G method (Gaussian-9 software [1, 4], Fig. 2). The results of QM calculations have shown that the partial charges on the carbon atoms in the aliphatic C–C–C–C–C chain are nearly identical (their absolute values ~0.01, and the signs of charges alternate as "–", "+", "–", "+", "–"). At the same time, the unsaturated C=C–C–C=C system is observed to have a distinctly different distribution of charges: the absolute values of the charges differed from each other by an order of magnitude (from 0.001 for the terminal C atoms to 0.046 for the central C atom) with retention of the sign-alternation pattern. In this case, the equilibrium conformation of the aliphatic C–C–C–C–C system corresponds to a plane (Fig. 2) and the equilibrium conformation of the C=C–C–C=C π-system, the two planes of the double bonds are perpendicular to each other. These specific features of the electronic and geometric structure of the model chains indicate that there are essential distinctions that are observed in the degree of sharing the electrons around the aliphatic and unsaturated χ-fragments of the corresponding χ-chains and lead to the predominance of aliphatic chains in the error region II.

Region III is opposite to the region I and corresponds to the area of rather low LUMO–HOMO gap values underestimated by the SM algorithm. Compared to the other regions of errors, molecules in the region III contained from four to six times fewer triple bonds and from four to ten times more hydrocarbons with sp^2 hybridization and no more than one hydrogen atom. Here, there are three to eight times more chains of carbon atoms corresponding to π-systems (C=C–C=C, etc.) and, vice versa, two to five times fewer aliphatic chains (C–C–C, etc.). Hence, the chemographs describing aromatic and other π-systems are predominant among the points in region III.

Region IV is opposite to the region II and corresponds to the overestimation of average and high LUMO–HOMO gap values (0.5–0.8). The molecules containing three or more rings are from three to ten times more frequently encountered in the region IV than in the other regions. Hence, the errors in region IV are associated with the predominance of polycyclic compounds.

The above-presented results of an expert analysis of the errors of the SM algorithm allowed us to draw several conclusions of constructive character. First, the chemographs in regions I–IV of errors correspond to the predominance of molecules with fundamentally different chemical structures: ammonium derivatives and aldehydes (region I), aliphatic (fluorine-containing) molecules (region II), aromatic and other π-systems (region III), and polycyclic compounds (region IV). Second, these classes of molecules can be determined on the basis of the structures of labelled chemographs and the postulates of chemical-bonding theory prior to calculations. Third, the analysis of these classes of molecules independently may fundamentally improve the accuracy of SM algorithms.

The separate analysis of the classes of molecules in connection with the errors in regions I–IV can further be performed within the framework of expert data analysis or with the use of more flexible approaches to machine learning (e.g., on the basis of the method of committees, boosting, etc.). In the case of expert analysis, the division of the entire 134K sample into two subgroups containing specific chemographs (corresponding to polycyclic, aromatic, and aliphatic compounds, n = 26 765) and all the other chemographs has allowed us to increase the correlation coefficient

Table 2. Examples of χ-chains with maximum absolute values of weights for calculation of LUMO–HOMO gap width. Chains are arranged in the order of decreasing weights ω_i. ^C is a carbon atom with steric stress (e.g., in a three-membered ring), C* is a chiral center of type D (as in D-alanine, etc.), $C is a carbon atom in a ring, and R, R1, R2, and R3 are random substituting groups (so-called "radicals")

Chain	$\varphi_t(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$	ω_i, arb. units
–^C(R,R1)^C(R2,R3)^CH_2^CH_2^CH_2–	0.0043	0.402
–HC(R)–H_2C–H_2C–H_2C–H_2C–	0.0063	0.369
–H_2C–H_2C–HC(R)–H_2C–CH_3	0.0077	0.337
–^CH_2–H^C*(R1)–H_2C–H^C*(R2)–O–	0.0042	0.333
–H^C*(R1)–H^C(R2)–O–H_2C–O–	0.0042	0.326
CH_3–HC(R)–H_2C–O–CH_3	0.0053	0.305
CH_3–H$C*(R1)–H$C*(R2)–O–$H_2$$C*–	0.0050	0.277
–^C(R1,R2)–$H_2$$C*–H$C*(R3)–$H_2$$C*–H^C(R)–	0.0046	0.277
NH_2–C(R1)=N–C(R2)=O	0.0047	–0.330
–$H_2$$C*–$C(R)=$C(R1)–H$C=H$C–	0.0095	–0.340
–C(R1)=HC–C(R2)=HC–HC=	0.0163	–0.340
–HC=C(R1)–N=C(R2)–HC=	0.0075	–0.341
–C(R1)=HC–HC=C(R2)–NH_2	0.0072	–0.425
NH_2–$C(R)=$C(R1)–HC=O	0.0065	–0.444
=$C(R)–N=$C(R1)–HC=O	0.0043	–0.463
–C(R1)=C(R2)–O–HC=C(R3)–	0.0045	–0.469
–C(R1)=C(R2)–HN–C(R3)=HC–	0.0051	–0.471
=C(R1)–N=HC–C(R2)=O	0.0043	–0.482

for control from $r(c) = 0.83$ to 0.88 at a decrease in the standard deviation from 0.17 to 0.11.

Analysis of weights ω_i for χ-invariants in Eq. (1) makes it possible to obtain interpretable results comparable with considerations of an expert chemist. For example, the calculation of the values of $\varphi_t(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ (see Theorem 1 in the first part of our paper [11]) and of the weights of elementary χ-invariants has allowed us to reveal the chemoinvariants that make the greatest contributions to the calculated LUMO–HOMO gap widths (Table 2).

Let us note that the aforementioned χ-invariants were rather rarely encountered: $\varphi_t(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ was less than 0.0077 (i.e., 0.77% of all the chemographs in sample 134K, Table 2). The greatest contribution to an increase in the LUMO–HOMO gap width was introduced by the chemoinvariants containing *carbon atoms with pronounced steric stress* (–^C(R,R1)^C(R2,R3)^CH_2^CH_2^CH_2–, etc.) and *aliphatic chains* (–HC(R)–H_2C–H_2C–H_2C–H_2C–, –H_2C–H_2C–HC(R1)–HC(R2)–CH_3, etc.), i.e., the fragments of molecules with obviously increased HOMO energies. On the contrary, a maximum contribution to a decrease in the gap width was made by π-*systems* (=C(R1)–N=N–C(R2)=HC–, =C(R1)–N=HC–C(R2)=O), which imply the sharing of an electron along the chains of π-bonds by definition. The latter facilitates the transfer of electrons between the LUMO and HOMO orbitals.

In brief, the SM algorithms developed are characterized by a precision and a high generalizing ability of models acceptable for conducting large-scale screenings of the quantum-mechanical properties of molecules in silico.

CONCLUSIONS

The proposed procedures for simulation modeling of the quantum-mechanical properties of molecules can be important for solving various problems in theoretical and applied chemistry. In theoretical chemistry, it is very important to develop theories and computational models that are acceptable for all classes of compounds and that make it possible to establish the semiquantitative interrelations between the spatial structures of molecules and their properties. Such models should allow the researcher to identify the

structural features that determine particular properties of the molecules, the possible reactions of the molecules and to predict the effects of structural modifications of the molecule. The need for computationally efficient models of this kind is obvious since the number of the known chemical compounds is measured in several hundreds of billions. For datasets of this size it seems quite impossible to obtain the data of precise quantitative QM calculations for each compound.

The cross-validation testing of the SM algorithms developed here allowed us to estimate the accuracy and the extent of overfitting of the proposed models and algorithms. In this case, the "topological" models of machine learning are characterized by good physicochemical interpretability (including interpretability in the terms of the quantum theory) and computational efficiency when compared with high-precision calculations on the basis of density functional theory. In fact, according to the models proposed the calculations involve a simple summation of the floating point numbers, the amount of which is comparable to the number of edges in the chemograph involved. These peculiarities of the algorithms developed provide the possibility of their application to the solution of a broad spectrum of problems: estimation of the quantum-mechanical properties of metabolites and of oligopeptides, the search/design of molecules with specified quantum-mechanical properties, solving various problems of materials science, the design of new drugs, repurposing of already-known medicines, etc.

ACKNOWLEDGMENTS

The authors are grateful to O.A. Gromova (Federal Research Center "Computer Science and Control," Russian Academy of Sciences) for useful discussions on expert data analysis.

FUNDING

This study was supported by the Russian Foundation for Basic Research, grant no. 19-07-00356.

COMPLIANCE WITH ETHICAL STANDARDS

This paper is an original scientific product by its authors, has not been published earlier, and will not be submitted to any other journals before receiving a *PRIA* Editorial Board resolution on its nonacceptance for publication.

Conflict of Interest

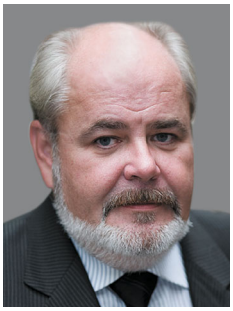
The authors declare that they have no conflicts of interest.

REFERENCES

1. L. A. Curtiss, P. C. Redfern, and K. Raghavachari, "Gaussian-4 theory," J. Chem. Phys. **126**, 084108 (2007). https://doi.org/10.1063/1.2436888
2. W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, 2nd ed. (Wiley-VCH Verlag, 2001). https://doi.org/10.1002/3527600043
3. V. I. Minkin, B. Ya. Simkin, and R. M. Minyaev, *Theory of Structure of Molecules* (Feniks, Rostov-on-Don, 1997).
4. R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014). https://doi.org/10.1038/sdata.2014.22
5. N. F. Stepanov, *Quantum Mechanics and Quantum Chemistry* (Mir, Moscow, 2001).
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 Recognit. Image Anal. **26**, 274–284 (2016). https://doi.org/10.1134/S1054661816020255
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**, 483–496 (2016). https://doi.org/10.1134/S1054661816030202
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 Recognit. Image Anal. **24**, 11–23 (2014). https://doi.org/10.1134/S1054661814010209
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 Recognit. Image Anal. **24**, 196–208 (2014). https://doi.org/10.1134/S1054661814020151
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**, 654–667 (2019). https://doi.org/10.1134/S1054661819040175
11. I. Yu. Torshin and K. V. Rudakov, "Topological theory of chemograph analysis as a promising approach to simulation of quantum mechanical properties of molecules. Part 1: On the generation of feature descriptions of molecules," Pattern Recognit. Image Anal. **31** (2021).
12. R. B. Woodward and R. Hoffmann, *The Conservation of Orbital Symmetry* (Academic Press, New York, 1970).

Translated by E. Glushachenkova

Ivan Yur'evich Torshin. Born in 1972. Candidate of Physics and Mathematics, Candidate of Chemistry, Associate Professor of the Moscow Institute of Physics and Technology, teacher in the Faculty of Computational Mathematics and Cybernetics of Moscow State University, senior researcher of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, and researcher of the Big Data Storage and Analysis Center of Moscow State University. Author of 520 papers in journals on informatics, medicine, chemistry, and biology and nine monographies, of which five are in Russian and four in English (in the series "Bioinformatics in the Post-Genomic Era," Nova Biomedical Publishers, New York, 2006–2009).



Konstantin Vladimirovich Rudakov (1954–2021). Russian mathematician, Academician of the Russian Academy of Sciences, associate director of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (Dorodnitsyn Computational Center), head of the Faculty of Intelligent Systems of the Moscow Institute of Physics and Technology, and scientific supervisor of the Big Data Storage and Analysis Center of Moscow State University.