https://doi.org/10.15407/ujpe71.7.573

A.D. TERETS',[1,2] T.YU. NIKOLAYENKO[1]

[1] Taras Shevchenko National University of Kyiv
(64/13, Volodymyrska Str., Kyiv 01601, Ukraine)
[2] O.O. Chuiko Institute of Surface Chemistry, Nat. Acad. of Sci. of Ukraine
(17, Generala Naumova Str., Kyiv 03164, Ukraine)

MODEL FOR PARAMETERIZATION OF INTERATOMIC INTERACTION POTENTIALS BY QUANTUM-MECHANICAL DESCRIPTORS ON THE BASIS OF A GRAPH NEURAL NETWORK

Based on graph neural networks, a machine learning model has been developed to predict the total energy of biomolecules from their structural formulas and quantum-mechanical descriptors by predicting the parameters of the functions that approximate interatomic potentials. The applicability of the created model for predicting the total energy of biomolecules, their interaction energy, and the ordering of conformers by energies has been proven. The physical validity of the obtained parameter values was demonstrated, which opens opportunities for further application of the model in molecular modeling problems.

Keywords: machine learning, interatomic interaction potentials, biomolecule conformers, quantum-mechanical descriptors, neural networks, biomolecular binding.

1. Introduction

The calculation of the energy of a system of atoms is a necessary step when applying physical methods to perform computer simulations of the structure and dynamics of both isolated molecules and condensed matter. In the framework of the quantum-chemical approach [1], such calculations are carried out on the basis of solving a quantum-mechanical problem aimed at finding the state of an electronic subsystem at given coordinates of the atoms under consideration. Despite their high reliability, such methods require a numerical solution of equations describing the electronic structure (for example, the Hartree–Fock equation or the density functional theory equation) at every step of the main simulation procedure, which leads to significant computational costs. This imposes a practical limitation on the application of those methods when simulating systems, in particular, biomolecular ones, which contain a substantial number of atoms and/or require a joint analysis of a considerable number of configurations using methods of statistical physics. Some reduction of the computational complexity is possible by approximating the system energy as a function of the coordinates of its atoms, in particular, by the sum of atom-to-atom potential functions. This approach, which is known as the force-field method, is the least computationally expensive and the basis of widely used implementations of the molecular dynamics method [2], docking [3], methods for finding the structures of atomic clusters [4–6], and protein structure modeling [7, 8].

A transition from the solution of complicated quantum-mechanical equations to the classical representation of force fields, which is approximate but computationally efficient, enables the modeling of systems on the time and size scales that are unattainable for direct quantum-mechanical calculations. Nevertheless, the reliability and efficiency of such modeling depend entirely on the correctness of the chosen interatomic potentials. This procedure includes both the choice of an appropriate functional form for their representation (in particular, with correct asymptotic behavior at large interatomic distances) and a proper calibration of the parameters in the chosen functions.

Citation: Terets' A.D., Nikolayenko T.Yu. Model for parameterization of interatomic interaction potentials by quantum-mechanical descriptors on the basis of a graph neural network. *Ukr. J. Phys.* **71**, No. 7, 573 (2026). https://doi.org/10.15407/ujpe71.7.573.

© Publisher PH "Akademperiodyka" of the NAS of Ukraine, 2026. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

In addition to standard approaches [9], where the functional form of the potential is taken in advance, interatomic potentials can also be approximated using machine learning (ML) methods. The ML methods, in which the functional dependence can be approximated on the basis of a set of examples of arguments and function values, are successfully applied in chemistry to predict reaction paths [10, 11], energies of excited states [12, 13], and formation energies [14, 15], as well as to search for new chemical compounds [16–18]. Recently, it has been shown [19] that ML potentials, in which ML methods are used to represent interatomic interaction potentials, can predict the energies of molecules with the accuracy of quantum-mechanical calculations, but at a much lower computational cost. From a computational point of view, modern ML models are based, as a rule, on neural networks. Some examples of such models are Grappa [20], TorchANI [21], and TorchMD-Net [22].

It is essential that the representation of interatomic potentials on the basis of ML models allows not only the approximation problem to be solved, but also avoids the need to calibrate the parameters of these potentials. For this purpose, we seek an ML model that could represent the energy of a system of atoms as a function of not only intermolecular distances but also the structure of the molecules themselves. A key precondition here is the choice of informative descriptors to represent the molecular structure, which facilitates the parameterization ("training") of the models and ensures their generalization ability. The relevant descriptors as numerical representations of molecular structure must be invariant to molecular rotations and translations, correctly take permutations of identical atoms into account, and unambiguously describe the molecular configuration.

However, despite the versatility of ML models based on the application of neural networks as a tool for approximating the energy of a system of atoms, their application in simulation methods requires more computations than the classical force-field method. Furthermore, the interpretation of data representations in the form of abstract multicomponent "vectors" operated by neural networks in intermediate calculations is difficult and does not allow the influence of the system structure on the system energy to be clearly traced. This circumstance can lead, in particular, to the fact that the model captures statistical correlations rather than cause-and-effect relations and, as a result, to the "error compensation" effect, when, for example, the influence of two different functional groups on the properties of a molecule is associated with those groups in an arbitrary way if those groups are mainly found together in the training data set of the model, and the approximation of the sum of the contributions made by those groups becomes sufficient.

Unlike existing approaches, in this paper, the possibility of a two-stage construction of an ML model of force fields on the basis of neural networks is analyzed. At the first stage, descriptors obtained from quantum-chemical calculations using the virial theorem and associated with individual atoms or atomic pairs, rather than a whole molecule, are applied. This allows the error compensation effect, which is characteristic of the direct total-energy approximation, to be reduced and, simultaneously, the efficiency of "transfer learning" [23] in the parameterization of interatomic potentials to be estimated. In addition, contrary to numerous neural network models that directly approximate the total energy or the potential of a molecule, our approach is aimed at predicting the parameters of classical interatomic potentials, which are afterward used to calculate the energy.

2. Virial Theorem and Its Application to the Construction of Quantum-Mechanical Descriptors

One of the main elements of the model proposed in this paper is the application of quantum-mechanical descriptors, which are based on the fundamental relation between the kinetic energy of electrons in a molecule and the molecular energy as a solution of a stationary Schrödinger equation; this relation is established by the virial theorem.

According to the virial theorem, the average value of the kinetic energy operator for electrons in the system can be written in terms of the total system energy [24] as follows:

$$\langle \hat{T}_e \rangle = -E. \tag{1}$$

In terms of the wave function Ψ, this relation looks like

$$\langle \Psi | \hat{T}_e | \Psi \rangle = \sum_i -\frac{\hbar^2}{2m_e} \langle \Psi | \Delta_{r_i} | \Psi \rangle =$$

$$= -N_e \frac{\hbar^2}{2m_e} \langle \Psi | \Delta_{r_1} | \Psi \rangle, \tag{2}$$

where $\hat{T}_e$ is the electron kinetic energy operator, $\hbar$ is the reduced Planck constant, m_e is the electron mass, N_e is the number of electrons in the system, and Δ_{r_i} is the Laplace operator in the coordinate r_i. The quantum average of the latter with the many-particle wave function Ψ of the electron subsystem in the molecule is given by

$$\langle\Psi|\Delta_{r_1}|\Psi\rangle = \int dr_1\, dr_2 \ldots dr_N \times \\ \times\, \Psi^*(r_1, r_2, ..., r_N)\Delta_{r_1}\Psi(r_1, r_2, ..., r_N). \quad (3)$$

(Hereafter, for brevity, the summation over the spin degrees of freedom is combined with the integration over the spatial coordinates.) It can be expressed in terms of the reduced single-particle density matrix

$$\gamma(r, \tilde{r}) = N_e \int dr_2 \ldots dr_N \times \\ \times\, \Psi^*(r, r_2, ..., r_N)\Psi(\tilde{r}, r_2, ..., r_N). \quad (4)$$

Then, expressing the kinetic energy in terms of $\gamma(r, \tilde{r})$, we obtain

$$-E = \langle\hat{T}_e\rangle = -N_e\frac{\hbar^2}{2m_e}\langle\Psi|\Delta_{r_1}|\Psi\rangle = \\ = -\frac{\hbar^2}{2m_e}\int (\Delta_{\tilde{r}}\gamma(r, \tilde{r}))_{\tilde{r}=r}\, dr. \quad (5)$$

Let us now represent the density matrix via its expansion in the basis functions $\chi_\mu(r)$, as is usually done when numerically solving equations in quantum-chemical methods [25]:

$$\gamma(r, \tilde{r}) = \sum_\mu \sum_\nu D_{\mu\nu}\chi_\mu(r)\chi_\nu(\tilde{r}), \quad (6)$$

where the coefficients $D_{\mu\nu}$ form the so-called electron density matrix, which contains the necessary information about the electronic structure of the system. Substituting this expansion into the expression for the energy, we obtain

$$-E = \langle\hat{T}_e\rangle = -\frac{\hbar^2}{2m_e}\sum_\mu \sum_\nu D_{\mu\nu} \times \\ \times \int \chi_\mu(r)\,(\Delta_{\tilde{r}}\chi_\nu(\tilde{r}))_{\tilde{r}=r}\, dr. \quad (7)$$

For further purposes, it is convenient to introduce the matrix element of kinetic energy, $K_{\mu\nu}$, for the basis functions,

$$K_{\mu\nu} = -\frac{\hbar^2}{2m_e}\int \chi_\mu(r)\,(\Delta_{\tilde{r}}\chi_\nu(\tilde{r}))_{\tilde{r}=r}\, dr. \quad (8)$$

Thus, the quantum average of the kinetic energy can be written as follows:

$$-E = \langle\hat{T}_e\rangle = \sum_\mu \sum_\nu D_{\mu\nu}K_{\mu\nu} = \mathrm{Tr}(D \cdot K), \quad (9)$$

where $\mathrm{Tr}(D \cdot K)$ is the matrix trace, which is a scalar quantity. As the kinetic energy of electrons in the given system, its counterpart in the auxiliary Kohn-Sham system was adopted; strictly speaking, this makes the statement of the virial theorem about the total energy approximate.

Next, let us distribute the sum (9) over individual atoms in order to separate the contribution of each of them (similarly to the introduction of Mulliken atomic charges [25]),

$$-E = \langle\hat{T}_e\rangle = \sum_A \sum_B \left(\sum_{\mu\in A}\sum_{\nu\in B} D_{\mu\nu}K_{\mu\nu}\right), \quad (10)$$

where A and B are atomic indices. The presented relationship between the matrices $D_{\mu\nu}$ and $K_{\mu\nu}$, on the one hand, and the total energy of the system, on the other hand, gives reason to expect that a neural network trained on the contributions of individual atoms or atomic pairs (in particular, chemical bonds) represented in those matrices will be able to form informative feature vectors. These vectors, in turn, can be used in other models to predict the energy parameters of molecules more accurately.

3. Characteristics and Preparation of Applied Datasets

A necessary precondition for training a neural network is the availability of a sufficiently large, diverse, representative, and balanced dataset, in particular, for tasks related to the modeling of molecular properties. In this work, the QMugs dataset [26] was used, which is a subset of the ChEMBL database [27] and contains only biologically and pharmacologically relevant molecules. This approach allows the model to be oriented towards predicting the properties of molecules that are similar to drugs, which enhances its practical value.

The original QMugs dataset contained information on about 665000 molecules. Given the limitations on computational resources, a subset of 304331 molecules was selected. These data were divided into three subsets: a validation set containing 19814 molecules

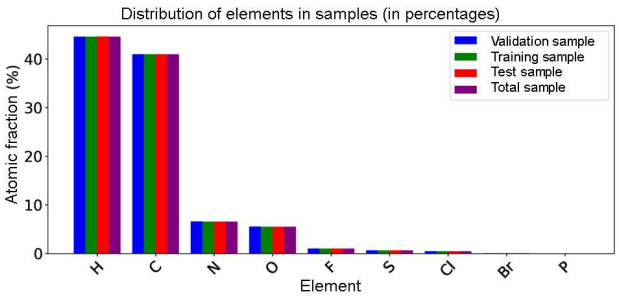


Fig. 1. Distributions of the numbers of atoms of various chemical elements in molecules in the training, validation, and test samples

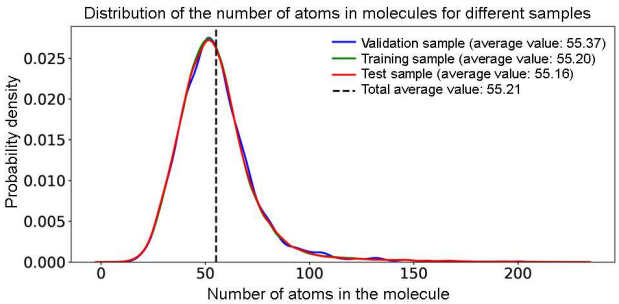


Fig. 2. Distributions of the numbers of atoms in a molecule in the training, validation, and test samples

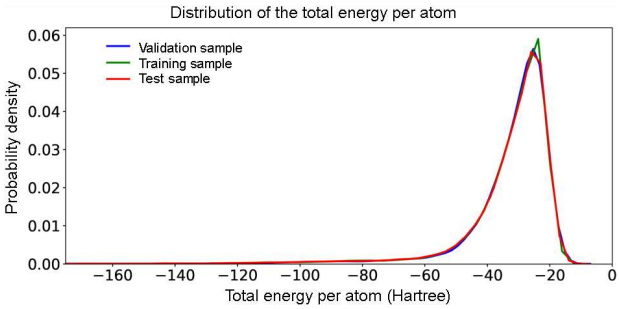


Fig. 3. Distribution of the total energy per atom among molecules in the training, validation, and test samples

(6.5%), a test set with 56904 molecules (18.7%), and a training set with 227613 molecules (74.8%).

As part of the preliminary analysis, a statistical analysis of the key dataset characteristics was performed. The distributions of molecules over the chemical elements (H, C, N, O, F, S, Cl, Br, and P; Fig. 1), the number of atoms in a molecule (Fig. 2), and the energy per atom (Fig. 3) were analyzed. The similarity of the obtained distributions for the validation, test, and training samples confirms the correctness of the performed partition, which is a necessary condition for constructing reliable and generalizable ML models.

For each molecule, using the Psi4 program [28], we calculated the kinetic energy matrix elements using formula (10) and the same Def2-SVP basis set as that used in creating the initial QMugs dataset.

Since our ultimate goal is to construct a mapping of the molecular structure onto the force field parameters, it is necessary to represent the molecular structure in a form suitable for machine learning methods. In such problems, a molecule is often represented as a graph [29], which allows its structure and properties to be effectively taken into account when solving machine learning problems. In the graph, the nodes correspond to atoms, and the edges to chemical bonds or non-covalent interactions between the atoms.

In the ML model developed in this work, every node of the graph (an atom) contains information about the atom as a chemical element, which is represented as a vector with a single non-zero component, and the index of this component determines the chemical element (the so-called one-hot coding).

One of the stages of graph formation is the determination of chemical bonds between the atoms. Based on the data on the molecular structure, the multiplicity of every bond (single, double, *etc.*) is determined. Additionally, spatial "interactions" are taken into account: the distance between the atoms in every pair is calculated, and if it exceeds a given threshold, the interaction for this pair is omitted. For atoms within this threshold, the bond is classified as non-covalent.

The graph edges are described using a set of parameters. Among them are the coordinates of the atoms and the bond type: covalent (taking into account the bond multiplicity) or non-covalent. This parameter set allows the graph to store information about both covalent and non-covalent interactions within the molecule.

On the basis of those data, a graph is created using the Deep Graph Library (DGL) [30]. The graph stores both structural and physicochemical information about the molecule, which makes it suitable for further analysis using ML methods.

4. Architectures of Created Neural Networks

The created ML model included two neural networks. They corresponded to the proposed two-stage

approach to predicting the energy of a system of atoms as a function of atomic coordinates.

The first network was aimed at encoding the configuration of the chemical environment of every atom and bond in the form of a vector of a given dimension, the so-called embedding. For its implementation, the Graph Neural Network (GNN) approach was used, in accordance with the chosen method for representing the molecular structure. Among the existing variants of GNN, the Message Passing Neural Network (MPNN) architecture [31] was chosen. The architecture and features of training this network are described below in more detail; see Section 4.1. To implement all ML models in this work, the PyTorch framework [32] was used.

After analyzing the structural formula of the molecule using the created MPNN, separate fully connected neural networks were created for atoms and atomic pairs to predict the parameters of interatomic potentials. Since the corresponding networks do not use information about interatomic distances, the parameters of such potentials in the proposed approach can be predicted only on the basis of the structural formulas of molecules. At the same time, interatomic distances affect the obtained molecular energies only through the expressions for interatomic potentials. These expressions are described in more detail in Section 4.2.

4.1. Graph Neural Network for creating embeddings

To construct embeddings as vector representations of the local environment of atoms and bonds in a molecule, a graph neural network with the MPNN architecture was developed, which included 4 message passing steps and a hidden layer of 350 neurons. When training this network, it was used as a predictor of the contributions made by the elements of the kinetic energy and electron density matrices [see Eq. (10)] associated with the individual atoms,

$$\mathrm{KD}_i^{\mathrm{atom}} = \sum_{\mu,\nu\in i} K_{\mu\nu} D_{\mu\nu}, \tag{11}$$

and the chemical bonds,

$$\mathrm{KD}_{ij}^{\mathrm{edge}} = \sum_{\mu\in i}\sum_{\nu\in j} K_{\mu\nu} D_{\mu\nu}, \tag{12}$$

in a molecule to the total kinetic energy of its electronic subsystem. In doing so, when training the MPNN, auxiliary modules were added to its architecture in the form of two separate fully connected neural networks: one for atoms ($\mathrm{MLP}_{\mathrm{atom}}$) and the other for bonds ($\mathrm{MLP}_{\mathrm{edge}}$). More specifically, at each iteration of the training process, the embedding

$$h_i = \mathrm{MPNN}\big(x_i, \{(x_j, e_{ij}) \mid j \in \mathcal{N}(i)\}\big) \tag{13}$$

was initially calculated for every i-th atom using the MPNN. Here, x_i is the vector of input atomic features, e_{ij} is the vector of features for the bond between atoms i and j, and $\mathcal{N}(i)$ is the set of neighboring atoms (connected by the given chemical bond). Then, the auxiliary fully connected network $\mathrm{MLP}_{\mathrm{atom}}$ takes the obtained embedding h_i as input data and calculates the corresponding contribution for the atoms,

$$\widehat{\mathrm{KD}}_i^{\mathrm{atom}} = \mathrm{MLP}_{\mathrm{atom}}(h_i). \tag{14}$$

The network for atoms included three linear layers with LeakyReLU activation functions and a hidden-layer width of 160 neurons.

The $\mathrm{MLP}_{\mathrm{edge}}$ network for bonds takes as input the concatenation of the summed embeddings of two atoms forming an edge in the molecular graph and the bond feature vector e_{ij},

$$\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}} = \mathrm{MLP}_{\mathrm{edge}}\left((h_i + h_j) \parallel e_{ij}\right), \tag{15}$$

where $\parallel$ denotes the vector concatenation operation. The $\mathrm{MLP}_{\mathrm{edge}}$ architecture consisted of four linear layers with LeakyReLU activations between them and a hidden-layer width of 199 neurons. When training the MPNN, only covalent bonds were taken into account, whereas the contribution from non-covalent interactions was neglected.

After the networks had predicted the corresponding contributions at the atomic ($\widehat{\mathrm{KD}}_i^{\mathrm{atom}}$) and bond ($\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}}$) levels, their sums were compared with the reference values calculated using the quantum-chemical method. Hence, the loss function was defined as the mean absolute error (MAE)

$$\mathrm{MAE}_{\mathrm{MPNN}} = \\ = \frac{1}{n}\sum_{M=1}^{n}\left|\sum_{i\in M}\widehat{\mathrm{KD}}_i^{\mathrm{atom}} + \sum_{(i,j)\in M}\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}} - KD_M^{\mathrm{true}}\right|, \tag{16}$$

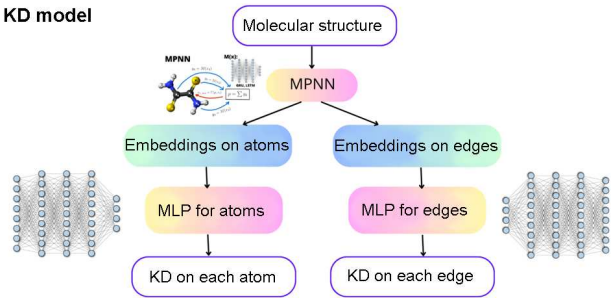


Fig. 4. Architecture of the model for creating embeddings

where KD_M^{true} is the true value calculated using the quantum-chemical methods. Next, standard error backpropagation was carried out to simultaneously update the weights of the MPNN, MLP_{atom}, and MLP_{edge} in order to minimize this error. The model for creating embeddings was trained on an Nvidia GeForce GTX 1080 video card, using the Adam optimization algorithm [33] and a batch size of 64.

4.2. Fully connected network for predicting the total molecular energy

As a result of using the MPNN network trained in the above-described way, the embeddings h_i, which encode the atomic chemical "context" and are determined by the structural formula of the molecule but do not depend on its spatial structure, were calculated for each atom in the selected molecules. The next step was to create two separate fully connected networks, the input data for which were the obtained embeddings and the edge feature vectors.

Let one of those networks (denoted as `charge_net`) take the embeddings h_i as input and assume the partial charges of atoms to be

$$Q_i = \text{charge\_net}(h_i). \tag{17}$$

Let the other network (denoted as `feats_net`) operate at the level of atomic pairs. For each edge of the molecular graph, it takes the concatenation

$$x_{ij} = [h_i \parallel h_j \parallel e_{ij}] \tag{18}$$

of the embeddings of atoms i and j as input; here, the symbols $\parallel$ denote the sequential union (concatenation) of the vector elements h_i and h_j, and the vector of edge features e_{ij}. On the basis of this vector, `feats_net` assumes a set of force field parameters

$$C = \text{feats\_net}(x_{ij}) = [C_0, C_1, C_2, ..., C_{13}] \tag{19}$$

for each atomic pair in the molecule. Specifying the distance between the atoms as

$$r_{ij} = |R_i - R_j|, \tag{20}$$

the long-range component of the interatomic interaction is calculated using the predicted parameters as

$$E_{ij}^{\text{LR}} = \frac{Q_i Q_j}{r_{ij}} - \frac{|C_0|}{r_{ij}^6} + \frac{|C_1|}{r_{ij}^{12}}, \tag{21}$$

and the short-range component as

$$E_{ij}^{\text{SR}} = \sum_{k=1}^{6} C_{2k} \times \\ \times \exp\left(-\frac{\big(r_{ij} - (1.2 + 0.6(k-1))\big)^2}{C_{2k+1}^2 + \varepsilon}\right), \tag{22}$$

where $\varepsilon = 10^{-5}$ is a term introduced to avoid division by zero during training. Equation (21) is used only for pairs of atoms that do not share a covalent bond; for covalently bound atoms, we put $E_{ij}^{\text{LR}} = 0$. Finally, the total interaction energy of atoms i and j is given by the sum

$$E_{ij} = E_{ij}^{\text{SR}} + E_{ij}^{\text{LR}}. \tag{23}$$

The predicted total energy contributions are read out from the graph's edges and nodes to obtain the total energy. It is compared with the energy E^{QC} calculated using the quantum chemical method for this molecule according to the MAE (mean absolute error) metric. To facilitate the training of the neural network, the model was used not to predict the total energy directly, but to make a correction to the estimate E^{lin} for the energy obtained using linear regression. The input data for such a regression were the vector N of the atomic numbers of all chemical elements in the molecule,

$$E^{\text{lin}} = w^\top N + b = \sum_{\alpha=1}^{m} w_\alpha N_\alpha + b, \tag{24}$$

where $N = (N_1, N_2, ..., N_m)$, m is the number of chemical elements, w is the vector of linear regression coefficients, and b is the free term of linear regression. Thus, the loss function during the training and testing of the networks was calculated using the formula

$$\text{MAE} = \frac{1}{n_{\text{mols}}} \sum_{M=1}^{n_{\text{mols}}} \left| \sum_{(i,j)\in M} E_{ij} + E_M^{\text{lin}} - E_M^{QC} \right|,$$

where the index M enumerates the molecules in the sample, and the expression $(i, j) \in M$ means that all atomic pairs in the M-th molecule are taken into account in the sum.

During the training of networks, to avoid the overtraining of neural networks and increase their resistance to small changes in the spatial positions of atoms, normally distributed random values ("Gaussian noise") with a mean value of 0 and the standard deviation $\sigma = 0.005$ Å were added to the atomic coordinates.

When training the `charge_net` and `feats_net` models, the Adam algorithm with an initial training rate of 0.001, together with a training rate reduction scheduler, was used, which reduced the rate by a factor of 0.8 every 20 iterations ("epochs"). The models were trained using an Nvidia A100 graphics card and a batch size of 400. This approach allows the model to approach the minimum of the loss function more effectively at later training stages. In total, the training lasted 400 iterations.

It is worth noting that only individual molecules, rather than their complexes, were used in the training. This feature makes the training process faster, but, generally speaking, changes it from the interpolation mode to the extrapolation one when applying the parameters of the interatomic interaction potentials (21) and (22), obtained using Eqs. (17) and (19), to the most practically important case of intermolecular interaction. Therefore, the validation results presented below for the models demonstrate their generalization ability in this case.

Besides the `charge_net` and `feats_net` models for predicting the total energy of molecules, their analogs were also created for predicting the kinetic energy of electrons in molecules. Their non-equivalence to the models for predicting the total energy of molecules is associated with the approximate character of the virial theorem in the case when, for the kinetic energy of electrons, the corresponding value for the Kohn–Sham auxiliary system in the DFT method is adopted. However, since according to the virial theorem, the average value of the kinetic energy operator is equal to the total system energy (1) to within a sign, then, when training the networks for the kinetic energy prediction, a similar parameterization in the neural network architecture was used, with the sign change to the opposite in front of the sum of the terms E_{ij}^{LR} and $E_{ij}^{\mathrm{SR}} = 0$.

5. Results and Discussion

5.1. Training and validation results for models predicting the total energy of molecules

The dynamics of the loss function reduction during the training of the model for creating embeddings are shown in Fig. 5, *a*, and the corresponding results for the models for predicting the kinetic and total energies are depicted in Figs. 5, *b* and 5, *c*, respectively.

For the test set, the average errors for the predicted kinetic energy of electrons and total energy of molecules were 6.1 kcal/mol (Fig. 6) and 2.2 kcal/mol (Fig. 7); in the latter case, the obtained value is close to the generally accepted chemical accuracy threshold (1 kcal/mol). Such an accuracy level is considered acceptable for most tasks in computational chemistry, in particular, for the high-quality reproduction of thermodynamic characteristics and the adequate modeling of chemical processes.

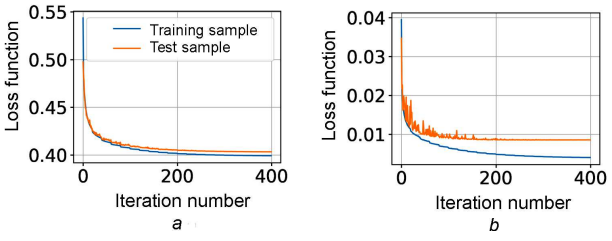


Fig. 5. Dependences of the loss function on the iteration number for the training and test sets of models for creating embeddings (*a*), predicting the kinetic energy (*b*), and predicting the total energy (*c*)

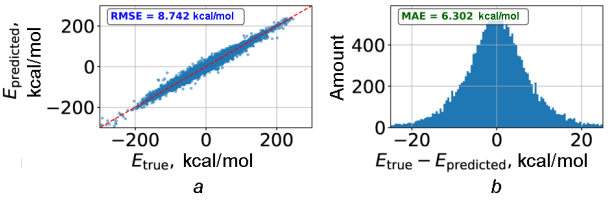


Fig. 6. Estimation of model predictions for the kinetic energy on the validation data set: comparison of true and predicted values (in kcal/mol) (*a*); distribution of the error in determining the kinetic energy (in kcal/mol) (*b*). E_{true} are the reference values of kinetic energy (in kcal/mol); $E_{\mathrm{predicted}}$ are the model predictions

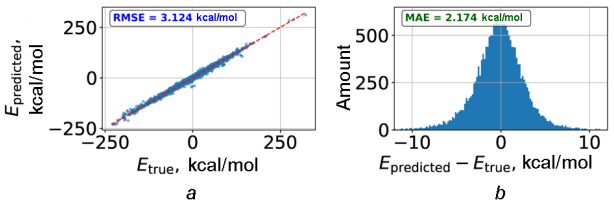


Fig. 7. Evaluation of model predictions for the total energy on the validation dataset: comparison of true and predicted values (in kcal/mol) (*a*); distribution of the error in determining the total energy (in kcal/mol) (*b*). E_{true} are the reference values of total energy (in kcal/mol); $E_{\text{predicted}}$ are the model predictions

Intermolecular interaction energies (kcal/mol) calculated for the A–T and G–C base pairs using the created ML model and other methods

Method	A–T pair	G–C pair
Reference value [34]	12.4	25.4
ML model	14.8	30.5
ANI	10.7	17.2
GFN2-xTB	16.1	29.2

Taking into account the small error produced by the created models when they are applied to individual molecules (in the interpolation mode), their operation in the extrapolation mode was tested for task types different from those used to train the model. In particular, the prediction of the intermolecular interaction energy was considered; this quantity was not included in the training data set. Such an application also corresponds to a more practical case, where the key role is played not by the absolute values of the system energy, but by the energy differences arising when the system configuration changes.

5.2. Application of models for evaluating interaction energies in DNA nucleotide base pairs

The interaction energy between the base pairs of deoxyribonucleic acid (DNA), namely, the adenine (A)–thymine (T) and guanine (G)–cytosine (C) pairs, is an important characteristic that affects the strength of pairing between its polymer chains. These energy parameters are important for understanding the thermodynamic properties of the DNA macromolecule, its melting point, its ability to form stable secondary structures, and the mechanisms of chain breakage and damage under various physical and chemical factors. Furthermore, the procedure for evaluating the interaction energy, as an example of how a certain model of the interatomic interaction potential can be applied, is important for the development of accurate molecular dynamics models for biotechnological applications.

Two strong hydrogen bonds are formed in the adenine–thymine pair, and three in the guanine–cytosine pair, which makes the G–C pair generally more stable than the A–T one. The interaction energy in the gas phase is usually evaluated to be about 12.4 kcal/mol for the A–T pair, and about 25.4 kcal/mol for the G–C pair [34]. The values of the interaction energy may vary depending on the chosen quantum chemical calculation method and the specific simulation conditions.

Within the framework of the developed models, the interaction energy of two molecules, E_{int}, was determined using the formula

$$E_{\text{int}} = (E_{\text{mol 1}} + E_{\text{mol 2}}) - E_{\text{complex}}, \qquad (25)$$

where E_{complex} is the energy of the whole system (the complex) of two molecules, and $E_{\text{mol 1}}$ and $E_{\text{mol 2}}$ are the energies of isolated molecules in their equilibrium geometry. Since the QMugs molecular set was used to train the ML models, the equilibrium geometries of which were found by the semi-empirical quantum-chemical method GFN2-xTB [35], the same method was used to determine the geometries of the molecular complex and its components, for which the created ML models were applied to calculate the energies E_{complex}, $E_{\text{mol 1}}$, and $E_{\text{mol 2}}$ in Eq. (25).

To compare with other modeling methods applied to intermolecular interactions, the energies E_{int} for the A–T and G–C pairs were also found using the TorchANI neural network and the semi-empirical method GFN2-xTB. The results obtained are quoted in Table. The obtained values testify that the created ML model demonstrates results for the interaction energy that are comparable with the results obtained using other methods and quite close to the reference values: in the case of the adenine–thymine and guanine–cytosine DNA base pairs, its deviation from the known values equals 2.4 and 5.1 kcal/mol, respectively.

5.3. Prediction of the interaction energy of molecules in noncovalently bound complexes

For a more complete estimation of the accuracy of the created ML model, it is appropriate to test it on a

wider set of molecules and obtain a statistically more substantiated characteristic of its performance in predicting molecular interaction energies. For this purpose, the GMTKN55 data set [36] was used, namely, its subset S66 [37], which contains 66 chemically representative dimers formed by noncovalently bound molecules. The interaction energies in this set vary from 2.82 to 19.49 kcal/mol, with an average value of 5.47 kcal/mol.

In this work, only those dimer geometries from the S66 data set were used that were optimized for further analysis with the GFN2-xTB method. Afterwards, the interaction energies of the molecules in the complexes were calculated using the DFT method, the exchange-correlation functional ωB97X-D [38], and the def2-SVP basis set, taking into account the effect of basis function overlapping (Basis Set Superposition Error, BSSE [39]). The latter was evaluated and compensated using the *counterpoise correction* procedure implemented in the `Psi4` program.

Although in the original work [37], the interaction energies were obtained by the high-precision CCSD(T)/CBS method [40], which is considered the reference in quantum chemical calculations, the applied ωB97X-D/def2-SVP variant of the DFT method allowed us to obtain results close to the reference ones. The corresponding comparison of energies is presented in Fig. 8. It is essential that the same variant of the DFT method was also used to calculate the total energy of molecules when creating the training set for the neural networks. This correspondence demonstrates that the chosen method is appropriate and correct for our study.

The results obtained by applying the developed ML model to calculate the energy of intermolecular interaction in the examined complexes are shown in Fig. 9. They demonstrate that the created model can predict the interaction energy values corresponding to the reference ones, with a root-mean-square error of 1.7 kcal/mol. Thus, for the problem of determining the energy of intermolecular interaction, the created model produces correct results, despite the fact that this parameter was not the main target metric during the model training.

5.4. Prediction of relative conformer energies

The variability in the spatial arrangement of atoms in biomolecules (such as proteins, nucleic acids, or ligands), provided that their graphs of chemical bonds are fixed, is determined by the molecular conformation. Due to the "flexibility" of their structure, biomolecules can acquire various conformations via rotating atomic groups around single σ-bonds. The conformational diversity substantially affects the functional properties of molecules: including their ability to interact with other molecules, and their catalytic activity [41]. In particular, protein α-helices and β-sheets represent different ways of folding the same polypeptide chain. When analyzing experimental data of rotational spectroscopy [42], as well as absorption and Raman scattering spectra [43], it is important to take into account the existence of many conformers in the specimen composition, because all available conformers contribute to the total spectrum measured in the gas phase. In this case, the intensity of the contribution made by each conformer is proportional to the probability of its existence under specific thermodynamic conditions. This probability is determined by

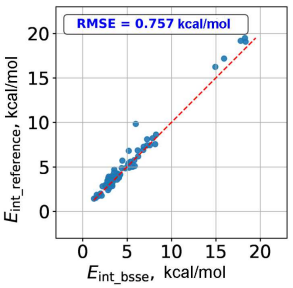
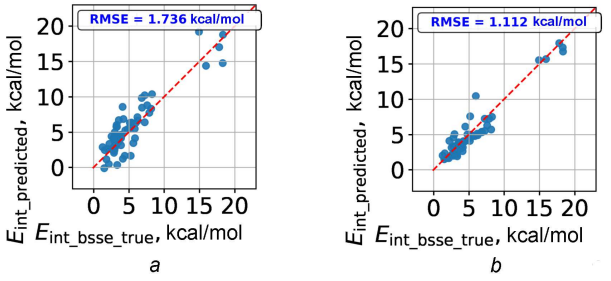


Fig. 8. Comparison of the molecular interaction energies E_{int_bsse} obtained using the DFT method (ωB97X-D/def2-SVP) with the $E_{int\_reference}$-values reported in the original work [37]



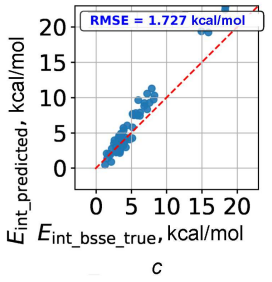


Fig. 9. Comparison of the reference $E_{int\_bsse\_true}$-values of the molecular interaction energies in the complexes of the S66 set with the corresponding energies $E_{int\_predicted}$ predicted by the created model (*a*), the GFN2-xTB method (*b*), and the ANI method (*c*)

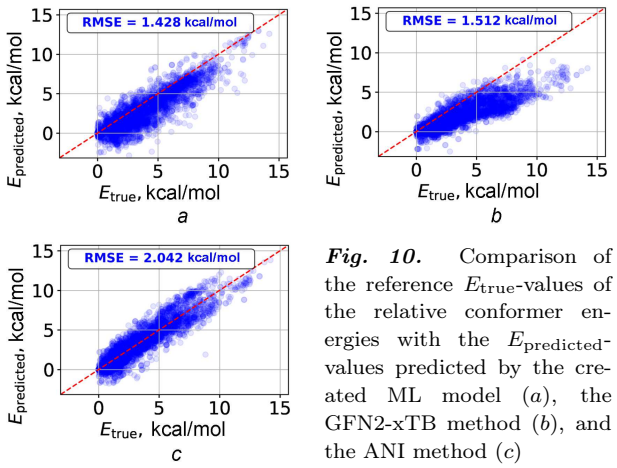


Fig. 10. Comparison of the reference E_{true}-values of the relative conformer energies with the $E_{\text{predicted}}$-values predicted by the created ML model (a), the GFN2-xTB method (b), and the ANI method (c)

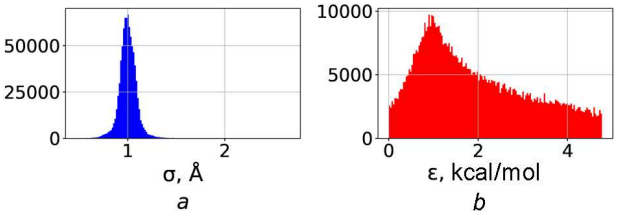


Fig. 11. Distributions of the Lennard-Jones interatomic potential parameters, evaluated on the basis of the predictions by the created ML model: the characteristic position σ of the minimum (a) and the potential well depth ε (b)

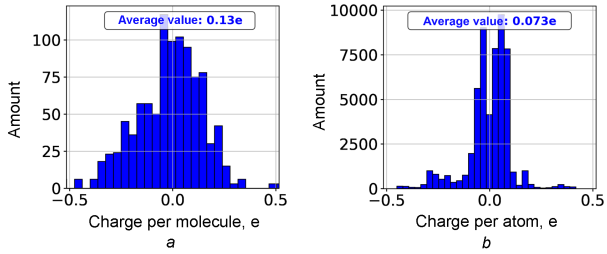


Fig. 12. Predicted charge distributions for whole molecules (a) and for individual atoms (b)

the relative energy of each conformer according to the Gibbs distribution. That is why methods for evaluating the relative energies of molecular conformers are of practical importance.

To analyze the applicability of the developed ML model to this problem, a subsample of molecules containing 7–9 atoms (excluding H atoms) was selected from the PC9 data set [44]. A total of 117 molecules were selected, and a search for possible conformations was performed for each of them using the CREST program [45]. In total, 4253 conformers were obtained. Then, the molecular geometries were further optimized using the GFN2-xTB method. After additional optimization using the DFT method ωB97X-D/def2-SVP, the Psi4 program was used to calculate the relative energies of the conformers as the difference between the energy of a given molecular conformation and the lowest energy of its conformers.

To comparatively analyze the correctness of the predictions produced by the created ML model, the relative energies were found using the GFN2-xTB method and the ANI neural network. The results of this analysis, shown in Fig. 10, indicate that the created model demonstrates good correlation in determining the relative energies of conformers, with RMSE = 1.4 kcal/mol. Since all conformers of the same molecule have the same graph and, accordingly, the same force field parameters, this test additionally verifies the correctness of the analytical formulas themselves for calculating the system energy, in which both the interatomic distances and the force field parameters are present.

5.5. Physical content of the parameters of the created interatomic potential

In the created potential (21)–(23), the functional forms of the terms in Eq. (21), which describe long-range interactions, correspond to the electrostatic interaction and the Lennard-Jones potential

$$V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{26}$$

However, their choice does not automatically guarantee that those terms correspond to the indicated physical mechanisms of interatomic interaction because the effects of error compensation between different terms are possible during the training of the ML models. It is therefore necessary to verify separately the physical correctness of the obtained parameters in those terms.

The Lennard-Jones potential (26) includes two quantities: the parameter σ corresponds to the distance at which the potential reaches its minimum, and the parameter ε characterizes the depth of the potential well. These parameters can be calculated on the basis of the coefficients obtained for model (23)

using the following formulas:

$$\sigma = \frac{C_0}{C_1}, \quad \varepsilon = \frac{C_0^2}{4C_1}. \tag{27}$$

The distributions of the σ- and ε-values obtained in this way are plotted in Fig. 11.

The calculated distributions demonstrate that, by their order of magnitude, typical σ- and ε-values are comparable with the corresponding parameters used in classical force fields, in particular, in GROMOS [46], where typical values of σ are about 3 Å, and those of ε are about 0.13 kcal/mol. This fact confirms the physical validity of the proposed approach, applied to training the `feats_net` network.

In Fig. 12, the distribution of total charges in 1000 electroneutral molecules from the validation dataset is shown. The total charges were obtained by summing up Q_i-values [see Eq. (17)] predicted by the `charge_net` network as the charges of individual atoms in each molecule. Although the network training process did not impose any special restrictions on Q_i, except for the use of these quantities in the term corresponding to Coulomb's law, the obtained results indicate that the values of the total molecular charge are mainly within the limits of $\pm 0.25e$, where e is the electron charge, which is acceptable for electroneutral molecules. Thus, the predicted values of the atomic charges are not arbitrary but consistent with one another according to the structural formula of the molecule. The total molecule charge is $0.13e$ (Fig. 12, *a*) and the atomic charge is $0.073e$ (Fig. 12, *b*), these values are of the same order of magnitude as the fractional charges assigned to atoms in other interatomic potentials.

6. Conclusions

To summarize, a two-stage approach to training neural networks for reproducing the dependence of the energy of biomolecules on their spatial structure and quantum-mechanical descriptors has been proposed and implemented. It was demonstrated that, based on the structural formula and the components of the electron kinetic energy of the molecule, which were determined using the density functional theory methods and are localized at atoms and atomic pairs, it is possible to obtain feature vectors that are analogous to the atomic types used in classical force fields for molecular dynamics. For this purpose, a graph neural network model was created to calculate such feature vectors for all atoms in an arbitrary electroneutral molecule consisting of the chemical elements H, C, N, O, F, S, Cl, Br, and P. A functional form for the interatomic interaction potential has been proposed; its parameters are functions of the atomic feature vectors found in the framework of the graph neural network model. To approximate this function, a model based on a fully connected neural network was created. Using this model and the proposed potentials, it is possible to determine the total energy of the molecule and the kinetic energy of its electrons in the pairwise interaction approximation, with mean absolute deviations of 2.2 kcal/mol and 6.1 kcal/mol, respectively. It was confirmed that the created models can be generalized to molecules not included in the training sample, as well as to problems beyond the scope of direct training. In particular, the applicability of the interatomic interaction potentials obtained using the created models on the basis of the structural formulas of molecules for predicting their interaction energies was confirmed. For the S66 set of 66 molecular complexes, the root-mean-square error in determining this energy equals 1.7 kcal/mol. It was found that the proposed two-stage neural network model allows the relative energies of biomolecular conformers from the PC9 set to be determined with a root-mean-square error of 1.4 kcal/mol. It was found that the parameters of the Lennard-Jones-type interatomic potentials, determined by the created neural network models, are consistent in order of magnitude with the corresponding parameters of the classical GROMOS force field.

The authors are sincerely grateful to the Friedrich Schiller University of Jena for providing access to the resources of the University Computing Center, which substantially accelerated our research.

1. R. Parr, Y. Weitao. *Density-Functional Theory of Atoms and Molecules, International Series of Monographs on Chemistry* (Oxford University Press, 1994).
2. D. Frenkel, B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications, Computational Science* (Academic Press, 2001).
3. D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Disc.* **3**, 935 (2004).

4. R. Johnston. *Atomic and Molecular Clusters* (Taylor and Francis, 2002).
5. Y. Xiang, D.Y. Sun, X.G. Gong. Generalized simulated annealing studies on structures and properties of Nin (n = = 2–55) clusters. *J. Phys. Chem. A* **104**, 2746 (2000).
6. P.N. Day, R. Pachter, M.S. Gordon, G.N. Merrill. A study of water clusters using the effective fragment potential and Monte Carlo simulated annealing. *J. Chem. Phys.* **112**, 2063 (2000).
7. K.A. Dill, J.L. MacCallum. The protein-folding problem, 50 years on. *Science* **338**, 1042 (2012).
8. S. Mayewski. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. *Proteins Struct. Func. Bioinform.* **59**, 152 (2005).
9. M. Allen, D. Tildesley. *Computer Simulation of Liquids*, *Computer Simulation of Liquids* (Clarendon Press, 1989).
10. Q. Zhao, D.M. Anstine, O. Isayev, B.M. Savoie. Δ^2 machine learning for reaction property prediction. *Chem. Sci.* **14**, 13392 (2023).
11. U.V. Ucak, I. Ashyrmamatov, J. Ko, J. Lee. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature Commun.* **13**, 1186 (2022).
12. R. Souza, J. Duarte, R. Goldschmidt, I. Borges, Jr. Machine learning prediction of electronic molecular excited state properties. *J. Braz. Chem. Soc.* **36**, 1 (2025).
13. Š. Sršeň, O.A. von Lilienfeld, P. Slavíček. Fast and accurate excited states predictions: Machine learning and diabatization. *Phys. Chem. Chem. Phys.* **26**, 4306 (2024).
14. D. Zhang, Q. Chu, D. Chen. Predicting the enthalpy of formation of energetic molecules via conventional machine learning and GNN. *Phys. Chem. Chem. Phys.* **26**, 7029 (2024).
15. M.R. Dobbelaere, I. Lengyel, C.V. Stevens, K.M. Van Geem. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *J. Cheminform.* **16**, 99 (2024).
16. U.K. Ghosh, F. Al Abir, N. Rifaat, S.M. Shovan, A. Sayeed, M.A.M. Hasan. Most dominant metabolomic biomarkers identification for lung cancer. *Inform. Med. Unlock.* **28**, 100824 (2022).
17. X. Zhang, I. Jonassen, A. Goksøyr. *Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data* (Exon Publications, 2021).
18. Z. Zhang, Z.-P. Liu. *Intelligent Computing Theories and Application: Proceedings of the 15th International Conference ICIC 2019, Nanchang, China, August 3–6, 2019* (Springer 2019), Part II, p. 517 [ISBN: 978-3-030-26968-5, 978-3-030-26969-2].
19. J. Behler. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
20. L. Seute, E. Hartmann, J. Stühmer, F. Gräter. Grappa – a machine learned molecular mechanics force field. *Chem. Sci.* **16**, 2907 (2025).
21. X. Gao, F. Ramezanghorbani, O. Isayev, J.S. Smith, A.E. Roitberg. TorchANI: A free and open source pytorch-based deep learning implementation of the ANI neural network potentials. *J. Chem. Inform. Model.* **60**, 3408 (2020).
22. S. Doerr, M. Majewski, A. Pérez, A. Krämer, C. Clementi, F. Noe, T. Giorgino, G. De Fabritiis. TorchMD: A deep learning framework for molecular simulations. *J. Chem. Theor. Comput.* **17**, 2355 (2021).
23. S.J. Pan, Q. Yang. A survey on transfer learning. *IEEE Trans. Knowledg. Data Eng.* **22**, 1345 (2010).
24. E. Weislinger, G. Olivier. The classical and quantum mechanical virial theorem. *Int. J. Quant. Chem.* **8**, 389 (1974).
25. F. Jensen. *Introduction to Computational Chemistry* (Wiley, 2017).
26. C. Isert, K. Atz, J. Jiménez-Luna, G. Schneider. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
27. D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M. Magariños, J. Mosquera, P. Mutowo, M. Nowotka *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucl. Acids Res.* **47**, D930 (2018).
28. D.G.A. Smith, L.A. Burns, A.C. Simmonett, R.M. Parrish, M.C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan *et al.* Psi4 1.4: Open-source software for high-throughput quantum chemistr. *J. Chem. Phys.* **152**, 184108 (2020).
29. Z. Guo, K. Guo, B. Nan, Y. Tian, R.G. Iyer, Y. Ma, O. Wiest, X. Zhang, W. Wang, C. Zhang, N.V. Chawla. Graph-based molecular representation learning. In: *Proc. of the 32nd International Joint Conference on Artificial Intelligence IJCAI '23* (Publisher, 2023), p. XXX.
30. M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, Z. Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv:1909.01315.
31. J. Gilmer, S.S. Schoenholz, P.F. Riley. Neural Message Passing for Quantum Chemistry. In: *Proc. of the 34th International Conference on Machine Learning*. Edited by D. Precup,Y.W. Teh (PMLR, 2017), p. 1263.
32. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Publisher, 2019), p. 8024.
33. D.P. Kingma, J. Ba. A method for stochastic optimization. arXiv:1412.6980.
34. Y. Mo. Probing the nature of hydrogen bonds in DNA base pairs. *J. Mol. Model.* **12**, 665 (2006).
35. C. Bannwarth, S. Ehlert, S. Grimme. GFN2-xTB – an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theor. Comput.* **15**, 1652 (2019).

36. L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19**, 32184 (2017).
37. J. Řezáč, K.E. Riley, P. Hobza. S66: A Well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theor. Comput.* **7**, 2427 (2011).
38. J.-D. Chai, M. Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615 (2008).
39. S. Boys, F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **19**, 553 (1970).
40. K. Raghavachari, G.W. Trucks, J.A. Pople, M. Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **157**, 479 (1989).
41. H.N. Motlagh, J.O. Wrabl, J. Li, V.J. Hilser. The ensemble nature of allostery. *Nature* **508**, 331 (2014).
42. A. Kovács, A.Y. Ivanov. Vibrational analysis of α-D-glucose trapped in Ar matrix. *J. Phys. Chem. B* **113**, 2151 (2009).
43. I. Peña, E.J. Cocinero, C. Cabezas, A. Lesarri, S. Mata, P. Écija, A.M. Daly, Á. Cimas, C. Bermúdez, F.J. Basterretxea, S. Blanco, J.A. Fernández, J.C. López, F. Castaño, J.L. Alonso. Six pyranoside forms of free 2-deoxy-D-ribose. *Angew. Chem. Internat. Edit.* **52**, 11840 (2013).
44. M. Nakata, T. Shimazaki. PubChemQC Project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model* **57**, 1300 (2017).
45. P. Pracht, F. Bohle, S. Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169 (2020).
46. N. Schmid, A.P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A.E. Mark, W.F. van Gunsteren. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **40**, 843 (2011).

Received 08.09.25.
Translated from Ukrainian by O.I. Voitenko

А.Д. Терець, Т.Ю. Ніколаєнко

МОДЕЛЬ ДЛЯ ПАРАМЕТРИЗАЦІЇ ПОТЕНЦІАЛІВ МІЖАТОМНОЇ ВЗАЄМОДІЇ ЗА КВАНТОВО-МЕХАНІЧНИМИ ДЕСКРИПТОРАМИ НА ОСНОВІ ГРАФОВОЇ НЕЙРОННОЇ МЕРЕЖІ

На основі графових нейронних мереж розроблено модель машинного навчання для передбачення повної енергії біомолекул за їхньою структурною формулою та квантово-механічними дескрипторами шляхом прогнозування параметрів функцій, що апроксимують міжатомні потенціали. Доведено застосовність створеної моделі для передбачення повної енергії біомолекул, енергії їхньої взаємодії, а також для прогнозування впорядкування конформерів за енергіями. Показано фізичну обґрунтованість отриманих параметрів, що відкриває можливості для подальшого використання моделі в задачах молекулярного моделювання.

Ключові слова: машинне навчання, потенціали міжатомної взаємодії, конформери біомолекул, квантово-механічні дескриптори, нейронні мережі, зв'язування біомолекул.