Original article

UDC 544.15, 004.85

DOI: 10.15372/CSD2025688

EDN: CHQTQL

Meta-learning for estimating the energy gaps of aromatic molecules

L. S. PETROSYAN[1,2], I. P. KOSKIN[1] ✉, M. S. KAZANTSEV[1]

[1]*Vorozhtsov Novosibirsk Institute of Organic Chemistry, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[2]*Novosibirsk State University, Novosibirsk, Russia*

E-mail: l.petrosyan@g.nsu.ru, koskin@nioch.nsc.ru✉, kazancev@nioch.nsc.ru

Abstract

A novel predictive model based on meta-learning algorithms for estimating the energy gap between the frontier molecular orbitals of aromatic π-conjugated molecules is presented. The main goal of the study was to develop a highly accurate and robust model capable of replacing computationally expensive quantum chemistry calculations, in particular the methods based on density functional theory, for screening organic compounds in optoelectronic applications. The filtered subset of the publicly available PubChemQC PM6 database was used as the primary dataset. Molecular structure was encoded using Morgan fingerprints, which served as input for training three base models: Random Forest, Gradient Boosted Trees, and a fully connected Neural Network. Among these, the Gradient Boosted Trees model achieved the best performance (the mean absolute error was 0.1795 eV). To improve prediction accuracy, a meta-model was implemented and trained on the outputs of the above-mentioned base models. This approach demonstrated improved accuracy: the final mean absolute error was 0.1744 eV, which is 8 % better than simple averaging and 3 % better than the best-performing individual model. The proposed approach can be further enhanced by expanding the dataset and incorporating additional models, which pave the way for more accurate and efficient prediction of the properties of organic conjugated molecules in optoelectronics.

Keywords: machine learning, energy gap, conjugated molecules, meta-learning

Funding: the work was carried out within the State Assignment for the Novosibirsk Institute of Organic Chemistry SB RAS (Project No. FWUE-2025-0007).

Acknowledgements: authors thank the Siberian Supercomputer Centre of SB RAS (Novosibirsk) for providing computational resources to train machine learning models.

For citation: Petrosyan L. S., Koskin I. P., Kazantsev M. S., Meta-learning for estimating the energy gaps of aromatic molecules, *Chemistry for Sustainable Development*, 2025, Vol. 33, No. 5, P. 555–561. DOI: 10.15372/CSD2025688. EDN: CHQTQL.

INTRODUCTION

Machine learning and deep learning methods are winning ever-growing application in various areas of chemistry and in materials science [1]. Predictive models trained on available experimental and calculated data can provide a substantial decrease in the time necessary for predicting the target value in comparison with the classical approaches used in quantum chemistry

© Petrosyan L. S., Koskin I. P., Kazantsev M. S., 2025

(e.g. the density functional theory), retaining or even surpassing their characteristic levels of accuracy and error [2, 3].

An important task in the area of compiling predictive models on the basis of machine learning methods is a sequential search for new and optimised training procedures which would allow an enhancement in the accuracy and stability of the predictions of data under investigation (for example, physicochemical characteristics). One of the possible strategies is using an ensemble of various machine learning models: several models are sequentially trained with the same training data, and the results of their predictions are averaged [4]. Researchers have previously proposed an alternative version of ensemble learning: instead of averaging the results of predictions at the final stage, meta-learning is applied, within which the model takes the results of predictions from all models in the ensemble as input data and is trained to combine them into the final weighted prediction [5]. At present, meta-learning is widely used [6] as a method to search for the quantitative structure – property relationships [7] as a tool for finding molecular conformations and for molecular geometry optimisation [8], as well as to study the toxicity and biological activity of organic compounds [9].

Aromatic π-conjugated molecules are applied in organic optoelectronics as active materials for organic field-effect transistors, light-emitting diodes, lasers and photovoltaic cells [10, 11]. One of the most important characteristics of these compounds is the value of the energy gap between the frontier orbitals – the highest occupied and lower unoccupied molecular orbitals. At the same time, direct calculation of the energies of frontier orbitals and energy gap between them by means of the density functional theory can be associated with substantial time expenditure (from several hours to days), which, in turn, hinders screening of the hundreds of organic compounds for the purpose of searching the most promising ones among them.

Thus, the goal of the present work was to develop a highly precise and stable model that could replace resource-intensive quantum chemical calculations, in particular the methods of density functional theory.

BUILDING A TRAINING DATASET

A public database PubChemQC PM6 [12], containing about 221 mln various chemical compounds, was used as the primary dataset for training. Among these records, about 2.6 mln contain the data on the energy gap between the frontier orbitals, calculated by the density functional theory (functional: B3LYP [13], basis set: 6-31+G* [14]). With the aid of RDKit library [15], the molecules corresponding to the following criteria were excluded from the database:

– inorganic compounds;
– aliphatic molecules;
– molecules containing not uniform conjugated system;
– ions and radicals;
– molecules in which the aromatic cycles are bound through a chain of three and more atoms;
– molecules containing no target atoms (C, Si, N, O, S, P, F, Cl, Br, I).

To refine the database, the functions of RDKit were used: determination of aromaticity, test for the presence of conjugated systems, selection of molecules by atomic composition, and treatment of charged structures.

After refining, the final database contained 260 thousand molecules, which corresponds to ~10 % of its initial volume. Then it was randomly divided into two parts: the training one (80 %) and the test one (20 %). The distribution of energy gaps in the training and test sets was similar to their distribution in the initial database (Fig. 1).

To transform the structural data of molecules from the SMILES format (Simplified Molecular Input Line Entry System) [16] into the formal suitable for machine learning models, a molecular descriptor characterising the nearest surroundings of atoms in the molecule was used. To generate these descriptors, we used the Morgan fingerprints methodology [17], implemented in the RDKit library (the radius of fragment construction was 3 nearest atoms, vector length was 2048 characters). A descriptor is unique for each individual molecule.

MACHINE LEARNING MODELS

The methods of machine learning used in the present work include: random forest (RF) [18], gradient boosted trees (GBT) [19], fully connected feed-forward neural network (NN) [20].

Within the RF model used for training, the number of decision trees was chosen to be 1000, and the maximal tree depth was 16. Other model parameters were not changed: their values were determined by default in the library of Yggdrasil Decision Forests.

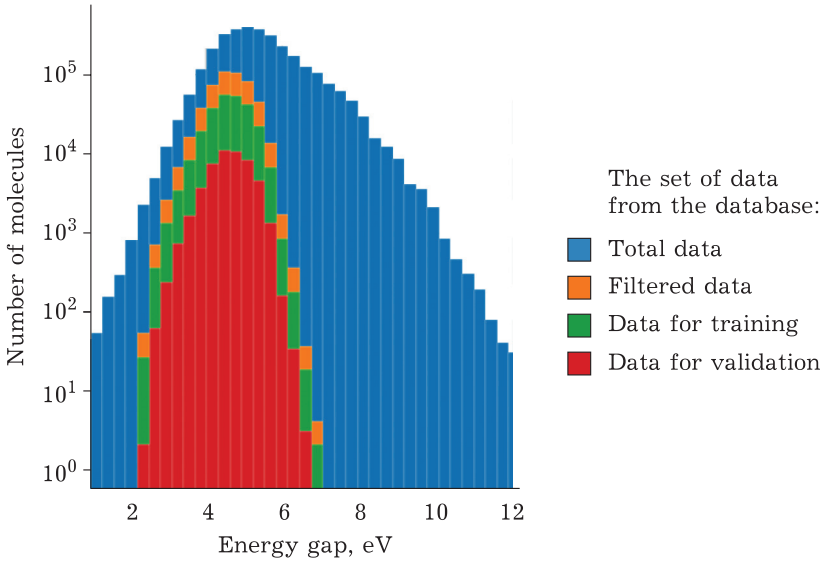

Fig. 1. Distribution of energy gaps for different sets from the PubChemQC PM6 database.

Within the GBT model used for training, the number of decision trees was chosen to be 1000, and the maximal tree depth was 10. Regularisation parameter L2 was chosen to be equal to 0.13. The learning rate (shrinkage) value was chosen to be 0.02. Other model parameters were not changed and remained equal to the default values in the library of Yggdrasil Decision Forests.

The NN model contained 5 hidden layers composed of 1500, 1000, 500, 250, 125 neurons. The neural network had a pyramidal structure with decreasing number of neurons while layer depth increased. This structure allows retrieving more and more complicated dependences from input data while moving deeper into NN. ReLu was chosen as an activation function. For each hidden layer, the parameter value was equal to 0.01. Each hidden layer in NN was followed by a dropout layer. This technique of NN construction is used for pruning connections between neurons in adjacent layers. All direct and inverse edges with a dropout layer are temporarily pruned, thus creating a new network architecture based on the parent network. The connections are removed with a fall probability equal to the parameter value. For each dropout layer, the parameter value was equal to 0.05. A feed-forward fully connected NN with similar parameters was used as the meta-model.

For all machine learning models described in the present work, including meta-model, cross-validation was performed, which implies multiple total set partitioning into a set of samples for training and testing, with sequential independent model training over each set. This allows us to generalise the developed model independently of the total set partitioning, and to decrease the probability of overtraining. For the listed internal parameters of machine learning models, variation was performed to find the optimal values that would allow obtaining the best accuracy of the estimates of the energy gap between the frontier orbitals.

RESULTS AND DISCUSSION

The use of meta-learning is efficient in machine learning: its essence is in bringing together two and more models to achieve higher prediction accuracy, which allows using the strengths of individual models compensating for their shortcomings, which often gives better results in comparison with those obtained using one model. In our work, we employed the model with parallel meta-learning, where several primary machine-learning models are trained independently of each other. This approach is especially useful when one of the models works well with describing various internal classes of compounds in the total set better than other models in the ensemble do.

There were three primary models of machine learning: RF, GBT, NN. The choice of machine-learning models and NN architecture is due to their prevalence, frequent application for estimating the energy gas between frontier orbitals, as well as performance simplicity and high rate.

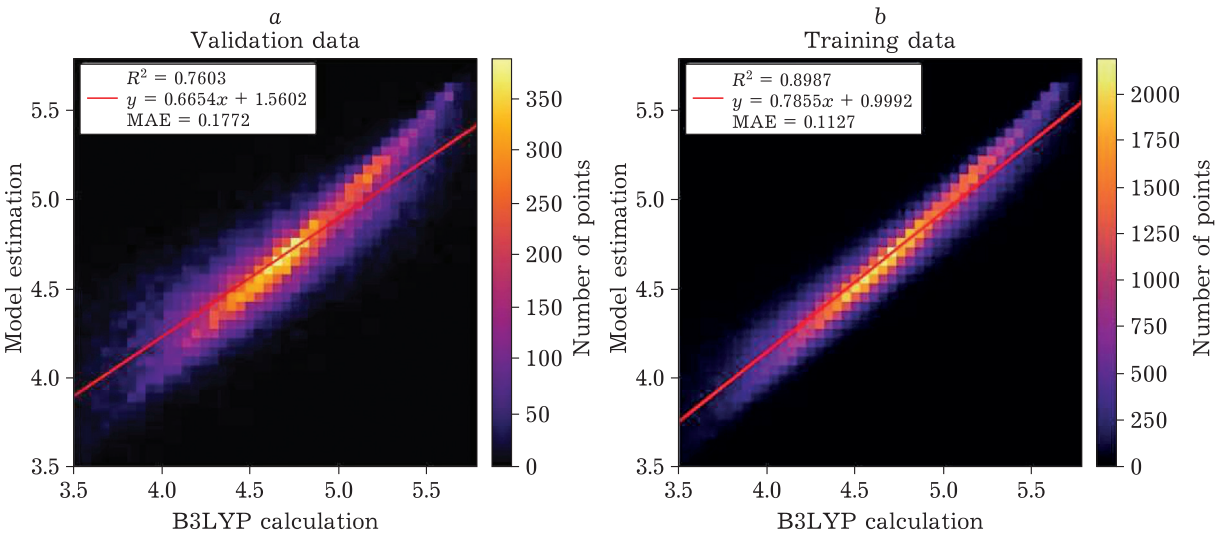

Fig. 2. Results of training and validation for the model based on random forest (RF): *a* – data for test set; *b* – data for training set. Here and in Fig. 3–5: the values predicted with the density functional theory are plotted at the *X* axis, while predictions generated by the model are plotted along the *Y* axis. The red line is a linear approximation of the regression obtained. MAE is the mean absolute error.

Random forest is a machine-learning algorithm with the general idea of constructing a set of decision trees which are independent of each other, and their estimates are averaged to give the final RF estimate. Each decision tree is a method to represent decision rules in the hierarchical structure composed of the elements of two types: nodes and leaves. The nodes contain decision rules and check the compliance of examples with a certain rule in one of the attributes of the training set. Decision trees also form the basis of GBT, however, in this case, each tree acquires the information from the previous one: what were the samples with which it had mistaken most substantially? These samples are given a higher weight in training a current tree. Thus, a gradual decrease in estimate error proceeds. The choice of RF and GBT as primary models is due to their ability to work efficiently with the input data containing a large number of attributes.

A neural network, being initially arranged similarly to the biological system, is composed of nodes connected with each other, called artificial neutrons, which generally model brain neurons. The connections between them imitate synapses in the brain. Each artificial neuron receives signals from the neurons connected to it, processes these signals and sends them to other connected neurons. A signal is a real number, and the output signal of each neuron is calculated using a certain nonlinear function of the sum of its input signals, which is called the activation function. Signal strength for each linking is determined by the statistical weight, which is corrected during training. A neuronal network can transform input data into a more abstract and complex representation, which allows it to discover unevident dependences in the data.

The vector representation of chemical structure obtained from the SMILES descriptor through Morgan fingerprints procedure was used as the input data for all primary machine learning models. The target value for machine learning is the energy gap between frontier orbitals for the compounds obtained from the processed database PubChemQC PM6. Testing was performed with 20 % randomly chosen compounds from the final database. Validation criterion was the mean absolute error of the value predicted by the machine learning model with respect to the value calculated using the density functional theory.

The results of model training on the basis of RF are shown in Fig. 2. One can see that the mean absolute error for training data (MAE = 0.1127 eV) is substantially smaller than for validation data (0.1772 eV). This behaviour is characteristic of so-called overfitting, as a result of which the machine learning model overestimates the influence of local dependences in the training dataset, and this leads to worsening of prediction quality for the data that were not used in model training. For this reason, the RF model was not used to construct the meta-model.

The data of models training on the basis of GBT and NN (a fully connected network with five hidden layers of 1500, 1000, 500, 250, 125 neu-

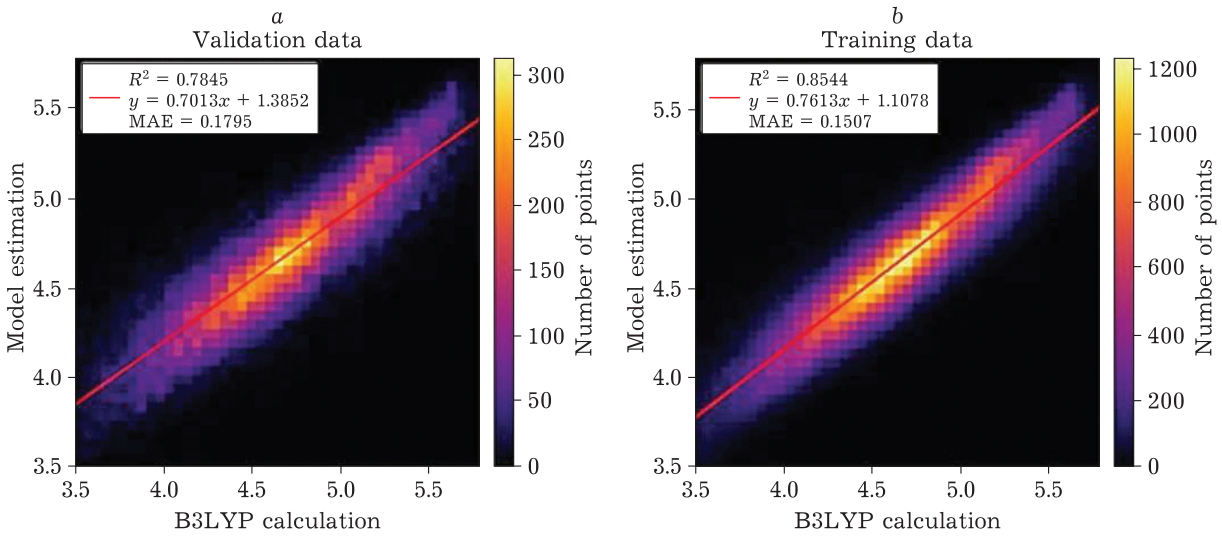

Fig. 3. Results of training and validation for the gradient boosted tree (GBT) model: *a* – data for test set; *b* – data for training set. For designations, see Fig. 2.

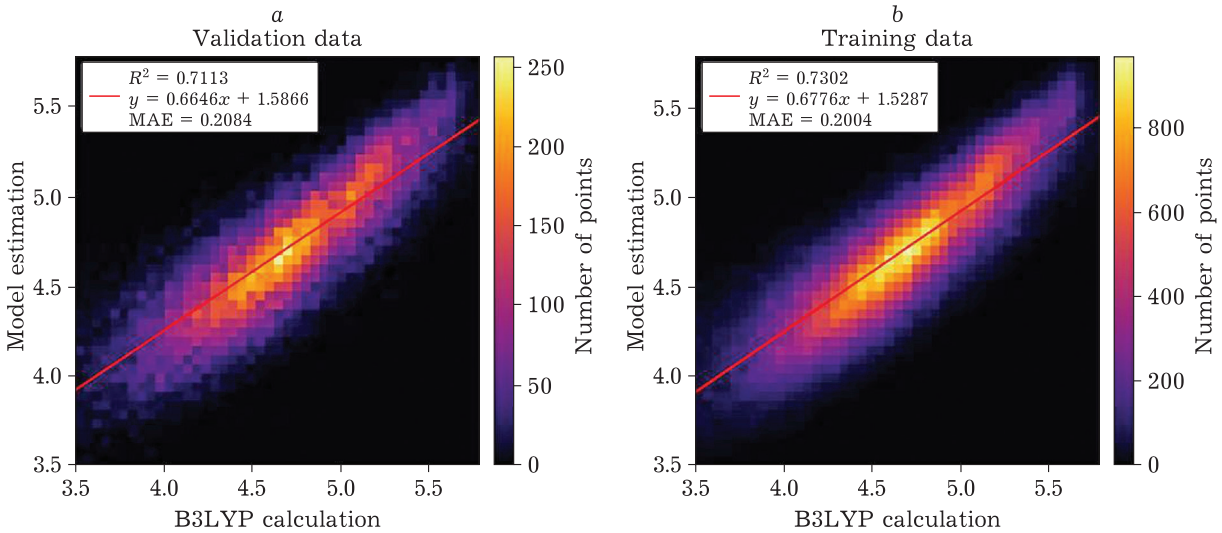

Fig. 4. Results of training and validation for the model based on neural network (NN): *a* – data for test set; *b* – data for training set. For designations, see Fig. 2.

rons) are presented in Fig. 3 and 4. The mean absolute error of prediction for the validation data of the indicated models is 0.1795 and 0.2084 eV, respectively, which is slightly more than the error for training data (0.1507 and 0.2004 eV). This behaviour is characteristic of machine learning and is not an evidence of overfitting. So, these two models were chosen for further training of the meta-model.

A standard and most frequently used approach to meta-model construction from different parallel machine learning models is averaging. In other words, each prediction of the primary model contributes equally to the value finally predicted by the ensemble. However, in the case of the data obtained by us, averaging causes an improvement of the quality of final prediction: the mean absolute error becomes larger than the smallest error for the models employed (0.1893 against 0.1795 eV).

An alternative approach is so-called meta-learning. Within this approach, to combine the predictions of primary models, an additional model is used: a meta-model which is trained over the output data of primary models. The main task of the meta-model is to determine the best way in which the predictions should be weighted and combined to enhance the total accuracy.

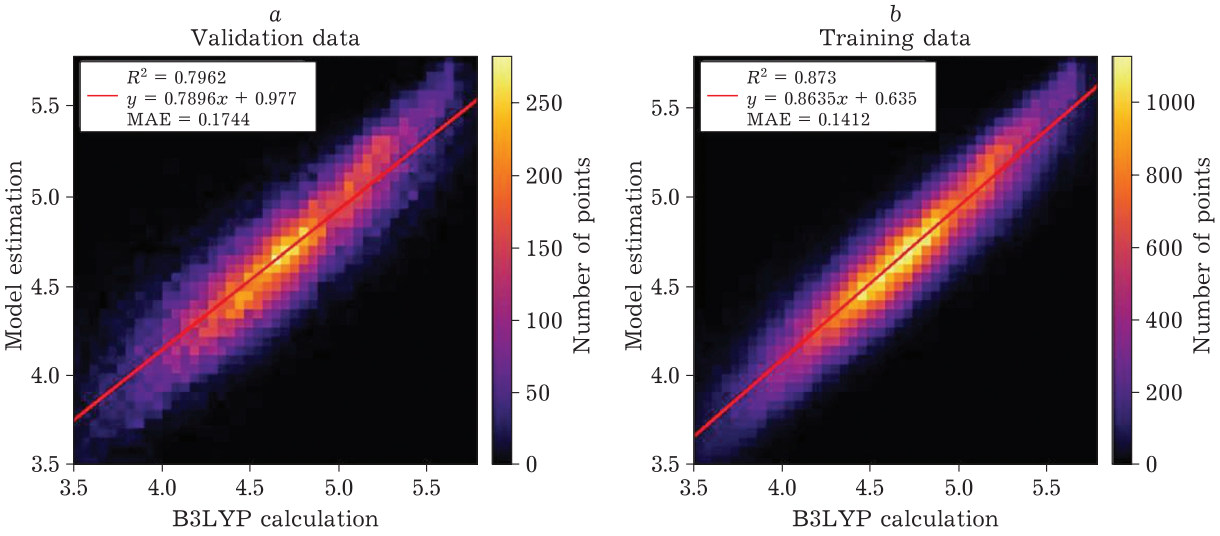

Fig. 5. Results of training and validation for the model based on meta-learning: *a* – data for test set; *b* – data for training set. For designations, see Fig. 2.

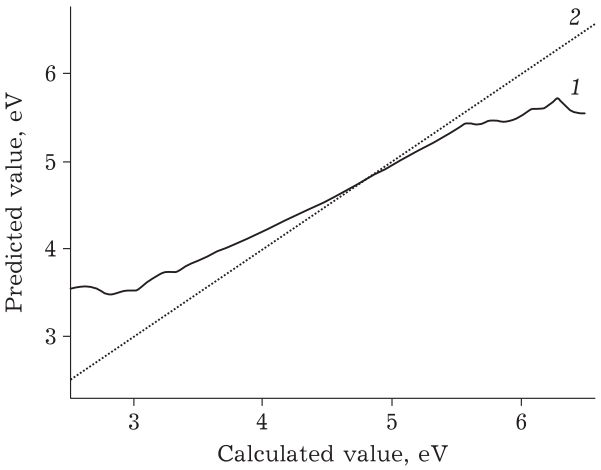

Fig. 6. Deviation of the values of the energy gap between frontier orbitals estimated using the meta-model (*1*) plotted with respect to the calculated values (*2*).

One can see in the data presented in Fig. 5 that the application of meta-model leads to a decrease in the mean absolute error by 8 % in comparison with simple averaging (0.1744 against 0.1893 eV) and by 3 % in comparison with the best model – GBT (0.1744 against 0.1795 eV). In addition, the model obtained by meta-learning exhibits the best predictive capacity for the fragment within the range of 3.5–4.5 eV (Fig. 6), containing most of the molecules from the database used for training. So, it may be assumed that the reason of an insignificant increase in accuracy is in the insufficient number of molecules in the set within the ranges corresponding to small and large energy gaps. In turn, this means that database expansion and its better uniformity in the distribution of energy gap values would likely lead to further improvement of the predictive capacity of the model based on meta-learning. Another potential direction for the development of the present work is the construction of a more complex ensemble using other machine learning methods and other methods of chemical structure vectorisation.

Thus, within the present work, meta-learning was used to predict the values of energy gaps in aromatic π-conjugated molecules. The final mean absolute error of prediction for new molecules that were not involved in training was 0.1744 eV, which corresponds to ~5 % of the mean relative error. This error is comparable to the accuracy of frontier orbital energy prediction using the density functional theory. The approach used herein, though providing lower estimation accuracy in comparison with more complex methods (for example, convolutional graph neural networks), is an easier and faster one for training and subsequent use.

CONCLUSION

It has been demonstrated in the present work that, among three machine learning models, the GBT model demonstrates the best accuracy in estimating the energy gap in aromatic compounds within the used set of aromatic π-conjugated compounds. The model based on meta-learning has been constructed for the first time and applied, and it has demonstrated higher prediction accuracy in comparison with both the direct av-

eraging of primary models and the best primary model.

REFERENCES

1. Sarker I. H., Machine learning: algorithms, real-world applications and research directions, *SN Comput. Sci.*, 2021, Vol. 2, No. 3, Art. 160.
2. Dral P. O., Barbatti M., Molecular excited states through a machine learning lens, *Nat. Rev. Chem.*, 2021, Vol. 5, No. 6, P. 388–405.
3. Bendifallah S., Puchar A., Suisse S., Delbos L., Poilblanc M., Descamps P., Golfier F., Touboul C., Dabi Y., Daraï E., Machine learning algorithms as new screening approach for patients with endometriosis, *Sci. Rep.*, 2022, Vol. 12, Art. 639.
4. Wang Z., Wen H., Su Y., Shen W., Ren J., Ma Y., Li J., Insights into ensemble learning-based data-driven model for safety-related property of chemical substances, *Chem. Eng. Sci.*, 2022, Vol. 248, Part A, Art. 117219.
5. Littlestone N., Warmuth M. K., The weighted majority algoritm, *Inf. Comput.*, 1994, Vol. 108, No. 2, P. 212–261.
6. Kwon S., Bae H., Jo J., Yoon S., Comprehensive ensemble in QSAR prediction for drug discovery, *BMC Bioinf.*, 2019, Vol. 20, No. 1, Art. 521.
7. Wang J., Zheng S., Chen J., Yang Y., Meta learning for low-resource molecular optimization, *J. Chem. Inf. Model.*, 2021, Vol. 61, No. 4, P. 1627–1636.
8. Rodríguez-Belenguer P., March-Vila E., Pastor M., Mangas-Sanjuan V., Soria-Olivas E., Usage of model combination in computational toxicology, *Toxicol. Lett.*, 2023, Vol. 389, P. 34–44.
9. Deshsorn K., Sirisaksoontorn W., Hirunpinyopas W., Iamprasertkun P., Exploring how base model combination affects the results of a "stacking" ensemble machine learning model: an applied study on optimization of heteroatom doped carbon data, *FlatChem*, 2025, Vol. 50, Art. 100827.
10. Li Q., Zhang Y., Xie Z., Zhen Y., Hu W., Dong H., Polycyclic aromatic hydrocarbon-based organic semiconductors: ring-closing synthesis and optoelectronic properties, *J. Mater. Chem. C*, 2022, Vol. 10, No. 7, P. 2411–2430.
11. Galabov B., Ilieva S., Cheshmedzhieva D., Nikolova V., Popov V. A., Hadjieva B., Schaefer H. F., Mini-review on structure–reactivity relationship for aromatic molecules: recent advances, *ACS Omega*, 2022, Vol. 7, No. 10, P. 8199–8208.
12. Nakata M., Shimazaki T., Hashimoto M., Maeda T., PubChemQC PM6: data sets of 221 million molecules with optimized molecular geometries and electronic properties, *J. Chem. Inf. Model.*, 2020, Vol. 60, No. 12, P. 5891–5899.
13. Becke A. D., Density-functional thermochemistry. I. The effect of the exchange-only gradient correction, *J. Chem. Phys.*, 1992, Vol. 96, No. 3, P. 2155–2160.
14. Ditchfield R., Hehre W. J., Pople J. A., Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules, *J. Chem. Phys.*, 1971, Vol. 54, No. 2, P. 724–728.
15. Bento A. P., Hersey A., Félix E., Landrum G., Gaulton A., Atkinson F., Bellis L. J., de Veij M., Leach A. R., An open source chemical structure curation pipeline using RDKit, *J. Cheminf.*, 2020, Vol. 12, No. 1, Art. 51.
16. Weininger D., SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, Vol. 28, No. 1, P. 31–36.
17. Capecchi A., Probst D., Reymond J.-L., One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome, *J. Cheminf.*, 2020, Vol. 12, No. 1, Art. 43.
18. Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P., Feuston B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci.*, 2003, Vol. 43, No. 6, P. 1947–1958.
19. Hagg A., Kirschner K. N., Open-source machine learning in computational chemistry, *J. Chem. Inf. Model.*, 2023, Vol. 63, No. 15, P. 4505–4532.
20. Käser S., Vazquez-Salazar L. I., Meuwly M., Töpfer K., Neural network potentials for chemistry: concepts, applications and prospects, *Digital Discovery*, 2023, Vol. 2, No. 1, P. 28–58.

Submitted 01.04.2025
Approved 23.06.2025
Accepted 15.07.2025