Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 2024, **34**, 780–782

Mendeleev Communications

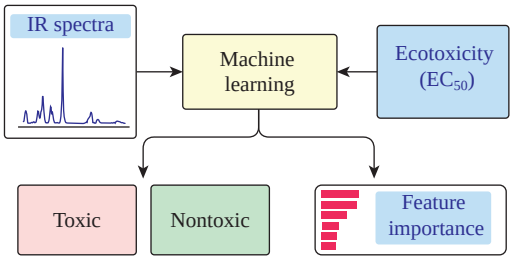
Machine learning-enabled prediction of ecotoxicity (EC_{50}) of diverse organic compounds *via* infrared spectroscopy

Maksim Yu. Sidorov,[a] Mikhail E. Gasanov,*[b] Artur A. Dzeranov,[a,c] Lyubov S. Bondarenko,[a] Anastasiya P. Kiryushina,[d] Vera A. Terekhova,[e] Gulzhian I. Dzhardimalieva[a,c] and Kamila A. Kydralieva[a]

[a] *Institute of General Engineering Training, Moscow Aviation Institute (National Research University), 125993 Moscow, Russian Federation*
[b] *Skolkovo Institute of Science and Technology, 121205 Moscow, Russian Federation. E-mail: gasanov.mikchail@gmail.com*
[c] *Federal Research Center of Problems of Chemical Physics and Medicinal Chemistry, Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russian Federation*
[d] *A. N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, 119071 Moscow, Russian Federation*
[e] *Department of Soil Science, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation*

DOI: 10.1016/j.mencom.2024.10.004

A new, less time-consuming and resource-intensive approach to predicting the EC_{50} ecotoxicity index, which is crucial for assessing the impact of compounds on ecosystems, is proposed. Efficient EC_{50} prediction based on infrared spectroscopy data and EC_{50} values from the EcoTOX database is achieved using machine learning. The best results with an F1-score of 0.83 were obtained with the SVC and XGBoost models.



Keywords: ecotoxicology, effective concentration, EC_{50}, feature importance, infrared spectroscopy, algae, machine learning.

Chemical regulation aims to protect both human health and the environment, with ecotoxicological research focusing on the latter. Regulatory hazard assessment is based on extensive animal testing, for example in the European Union, acute toxicity testing of chemicals is required by the REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) legislation.[1] With over 200 million substances in the Chemical Abstracts Service (CAS) archive[2] and over 350 000 chemicals and mixtures currently registered on the global market, chemical hazard assessment is a major challenge.[3] Traditional methods for assessing their potential impact on ecosystems are ethically controversial, often resource-intensive (time, personnel, required test material) and inherently unable to meet the rapid testing requirements of the ever-growing chemical universe in need of evaluation.[4] These ethical and financial considerations are the drivers for finding alternatives to animal testing that include computational (*in silico*) methods.[5,6] This study proposes a novel approach that uses machine learning and infrared (IR) spectroscopy techniques to efficiently and accurately predict the half-maximal effective concentration (EC_{50}). IR spectral data for a diverse set of compounds were collected and the corresponding EC_{50} values were compiled from the open source EcoTOX database.[7] Entries for an important taxonomic and trophic level (producers) in aquatic ecotoxicology, algae, are included, representing one of the largest subsets of data available in EcoTox. For effects that are not lethality-based, toxicity is typically characterized by an EC_{50} value, which is the concentration of a substance that produces 50% of the maximal effect level, often compared to a negative and/or positive control treatment.

The open EcoTOX database was used as a source of ecotoxicity data. The CSV data were obtained from the project website[8] and contained the following parameters: Trophic Level, Effect Value, Name, Test Statistics, Duration (days), Chemical Name, CAS and others. The original dataset was filtered by the following criteria: test statistics is EC_{50}, and the experiment duration is 96 h. SMILES were obtained by Chemical Name using the CirPy Python package.[9] The number of compounds was then compared across different taxonomic groups and the most represented group, ALGAE, was selected (Figure 1). Each compound was classified as Toxic (1) or Nontoxic (0) based on an EC_{50} threshold level of 100 mg dm^{-3}, following the standard procedure for classifying toxic compounds.[10] Random Undersampling was applied to address class imbalance in the final dataset. The final training dataset contained 225 observations and 36 features.

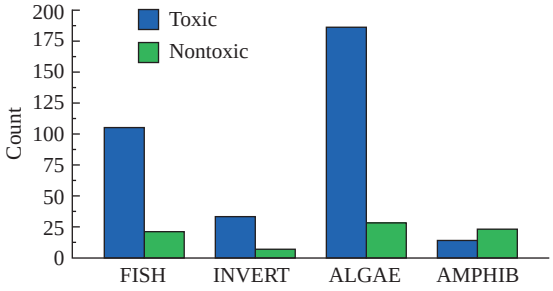


Figure 1 Number of samples of each taxonomic level group in the EcoTOX dataset.

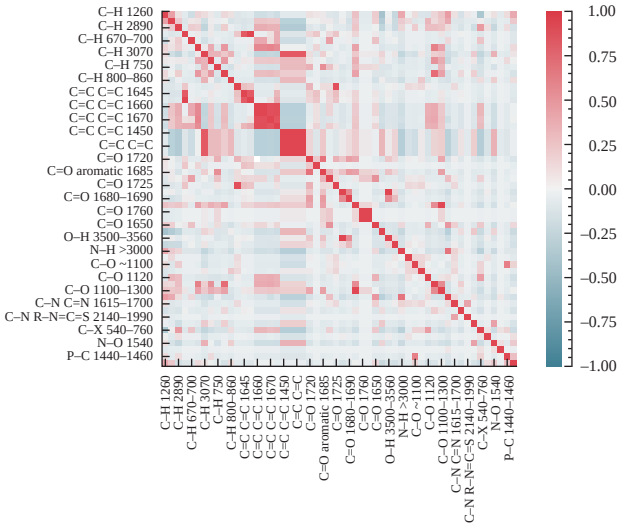


Figure 2 Correlation matrix illustrating the relationships between parameters derived from IR spectra and used in classification.

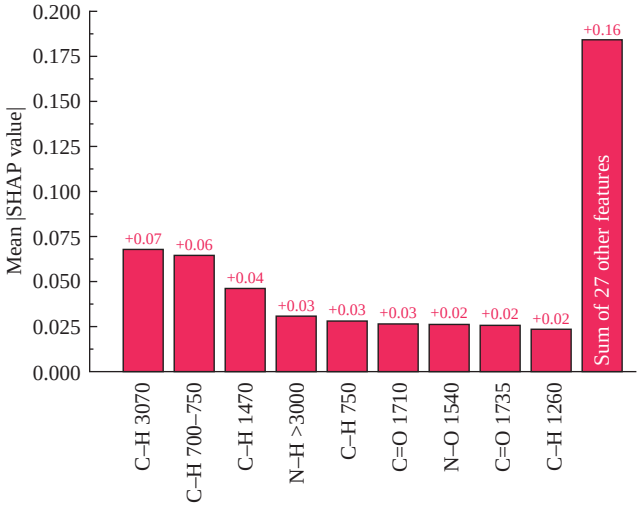


Figure 4 Feature importance of the RandomForest model for toxic/nontoxic classification of compounds using SHAP analysis.

IR spectroscopy provides information on the presence of molecular bonds that were used as predictors to classify compounds as Toxic or Nontoxic. The dataset comprises 12 200 records, each identified by a SMILES string and containing 67 binary values indicating the presence (1) or absence (0) of specific molecular bonds based on IR spectroscopy data.[11] SMILES were used as keys to merge the toxicity dataset and IR spectra. To reduce dimensionality and minimize the risk of overfitting, features with the uniform labels in all records and features with a correlation higher than 0.75 were removed. Figure 2 shows the correlation matrix indicating the bond type and its corresponding wavelength.

The performance of four machine learning models was compared to solve the problem of toxic/nontoxic substance classification based on IR spectra. The following models were used: RF (RandomForest),[12] SVC (Support Vector Classifier),[13] kNN (*k*-Nearest Neighbors)[14] and XGB (Extreme Gradient Boosting).[15]

The RandomizedSearchCV method was used to uncover the optimal configuration of hyperparameters for the compared machine learning models. Unlike exhaustive methods such as GridSearchCV, RandomizedSearchCV randomly samples a subset of hyperparameter combinations, efficiently navigating through large parameter spaces.[16] The hyperparameters were optimized using RandomizedSearchCV with a cross-validation parameter value of 5. Using this approach, it was possible to find a balance between exploration and exploitation, identifying the most effective hyperparameter settings for the considered models across diverse datasets and algorithms.

Metrics such as F1-score and recall were used to evaluate the performance of the applied classification model.[17] F1-score, calculated as the harmonic mean of precision and recall, provides a balanced assessment of the model's ability to classify instances into different classes, especially in the presence of class imbalance. Meanwhile, recall measures the ability of the model to correctly identify all positive instances, irrespective of false negatives. These metrics offered comprehensive insights into the effectiveness of the applied classifier, guiding the understanding of its performance in diverse classification scenarios.

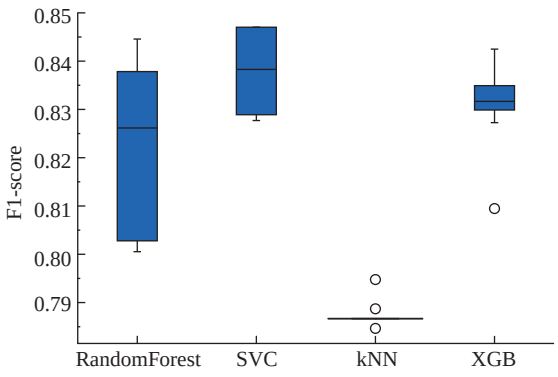


Figure 3 Performance of machine learning models during hyperparameter optimization.

To evaluate the predictive performance of the models, a series of computational experiments were conducted. The training process was repeated 15 times, with each iteration including a split into test and training datasets. The resulting F1-score metrics are depicted as boxplots in Figure 3. As shown in the graph, all simulation runs on the test data achieved an F1-score of at least 0.78. Notably, decision tree-based models such as XGB and RF demonstrated more stable results. Additionally, SVC exhibited the highest recall metric with a value of 0.79 ± 0.01.

In this work, the importance of various molecular bonds in predicting EC_{50} values was analyzed using machine learning models, with feature contributions quantified by SHAP (SHapley Additive exPlanations) values.[18] The results of the SHAP analysis are presented as a ranked bar chart in Figure 4. The SHAP analysis revealed that the C–H stretch at 3070 cm^{-1} has the largest impact with a mean absolute SHAP value of 0.07, indicating its significant role in the model predictions. It is closely followed by the C–H stretch in the 700–750 cm^{-1} range with a mean SHAP value of 0.06. The C–H stretch at 1470 cm^{-1} and the N–H stretch at >3000 cm^{-1} also make notable contributions to the model, each with a mean SHAP value of 0.04 and 0.03, respectively.

Importantly, the cumulative effect of the 27 other molecular features was substantial, collectively providing a mean SHAP value of 0.16, highlighting the complexity and multiple factors influencing EC_{50} predictions.

All computational experiments were performed on a PC with 16 Gb of RAM and 12 CPU cores. The source code and data are freely available on Github.[19] These studies to verify the proposed model will be continued further for synthesized and published hybrid organo-inorganic compounds, in which magnetite nanoparticles were modified with compounds of different nature (alkoxysilanes, natural polyelectrolytes and metal-organic frameworks). For some compounds, the microstructure and EC_{50} were determined by IR spectroscopy.[20–22]

This work was supported by the Russian Science Foundation (project no. 23-23-00621).

References

1 [dataset] European Commission, *Regulation (EC) No 1907/2006, version 06/06/2024*, EUR-Lex, 2024; http://data.europa.eu/eli/reg/2006/1907/2014-04-10.

2 [dataset] Chemical Abstracts Service, *CAS REGISTRY: The Authoritative Source for Chemical Substance Data*, American Chemical Society, 2024; https://www.cas.org/cas-data/cas-registry.

3 C. Schür, L. Gasser, F. Perez-Cruz, K. Schirmer and M. Baity-Jesi, *Sci. Data*, 2023, **10**, 718; https://doi.org/10.1038/s41597-023-02612-2.

4 C. Rovida and T. Hartung, *Alternatives to Animal Experimentation*, 2009, **26**, 187; https://doi.org/10.14573/altex.2009.3.187.

5 A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori and M. S. Rao, *Chem. Res. Toxicol.*, 2020, **33**, 20; https://doi.org/10.1021/acs.chemrestox.9b00227.

6 M. A. Pukalchik, A. M. Katrutsa, D. Shadrin, V. A. Terekhova and I. V. Oseledets, *J. Soils Sediments*, 2019, **19**, 2265; https://doi.org/10.1007/s11368-019-02253-2.

7 J. H. Olker, C. M. Elonen, A. Pilli, A. Anderson, B. Kinziger, S. Erickson, M. Skopinski, A. Pomplun, C. A. LaLone, C. L. Russom and D. Hoff, *Environ. Toxicol. Chem.*, 2022, **41**, 1520; https://doi.org/10.1002/etc.5324.

8 [dataset] United States Environmental Protection Agency, *ECOTOX Knowledgebase*, 2024; https://cfpub.epa.gov/ecotox/.

9 [dataset] M. Swain, *CIRpy, Python wrapper for the NCI Chemical Identifier Resolver (CIR)*, v1.0.2, GitHub, 2016; https://github.com/mcs07/CIRpy.

10 United Nations Economic Commission for Europe, *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 10th edn., United Nations, New York, 2023; https://doi.org/10.18356/9789210019071.

11 D. S. Koshelev, *Appl. Spectrosc.*, 2024, **78**, 387; https://doi.org/10.1177/00037028241226732.

12 A. Liaw and M. Wiener, *R News*, 2002, **2** (3), 18; https://journal.r-project.org/articles/RN-2002-022/.

13 C. Cortes and V. Vapnik, *Machine Learning*, 1995, **20**, 273; https://doi.org/10.1007/BF00994018.

14 P. Cunningham and S. J. Delany, *ACM Computing Surveys*, 2022, **54** (6), 128; https://doi.org/10.1145/3459665.

15 [dataset] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li and J. Yuan, *xgboost: Extreme Gradient Boosting*, version 0.4-2, CRAN Repository, 2015; https://doi.org/10.32614/CRAN.package.xgboost.

16 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825; https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

17 D. V. Carvalho, E. M. Pereira and J. S. Cardoso, *Electronics*, 2019, **8**, 832; https://doi.org/10.3390/electronics8080832.

18 S. M. Lundberg and S.-I. Lee, in *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, eds. U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, Curran Associates, Red Hook, NY, 2018, pp. 4766–4775; https://doi.org/10.48550/arXiv.1705.07874.

19 [dataset] M. Gasanov and M. Sidorov, *pyMakSid/EcoToxicityMachine Learning*, GitHub, 2024; https://github.com/pyMakSid/EcoToxicityMachineLearning.

20 L. Bondarenko, A. Kahru, V. Terekhova, G. Dzhardimalieva, P. Uchanov and K. Kydralieva, *Nanomaterials*, 2020, **10**, 2011; https://doi.org/10.3390/nano10102011.

21 L. Bondarenko, E. Illés, E. Tombácz, G. Dzhardimalieva, N. Golubeva, O. Tushavina, Y. Adachi and K. Kydralieva, *Nanomaterials*, 2021, **11**, 1418; https://doi.org/10.3390/nano11061418.

22 L. Bondarenko, Y. Saveliev, D. Chernyaev, R. Baimuratova, G. Dzhardimalieva, A. Dzeranov, E. Kelbysheva and K. Kydralieva, *Phys. Chem. Chem. Phys.*, 2023, **25**, 15862; https://doi.org/10.1039/D3CP01404E.

Received: 3rd June 2024; Com. 24/7519