Interactive Research Toolbox for Chemical Compounds Analysis Based on Well-interpretable ML Methods ⋆⋆⋆

A.A. Glushko[1][0009-0007-1783-2507], A.A. Neznanov[1][0000-0001-9106-2298], G.M. Kuzmicheva[2][0000-0003-4458-8013], O.V. Maksimenkova[1][0000-0003-2467-730X], and S.O. Kuznetsov[1][0000-0003-3284-9001]

[1] HSE University
[2] MIREA-Russian Technological University

Abstract. This paper provides an overview of tool for computational experiments, which are utilized in the realm of compound property analysis. Moreover, it outlines the initial outcomes of developing a novel research toolbox system that concentrates on exploring the interpretability and explainability of machine learning outcomes. The proposed solution is based on open-source tools and offers a convenient approach to address specific material science issues. The efficacy of this solution is currently being tested on problems related to compound synthesis optimization, antibacterial activity analysis, and other related areas.

Keywords: Property prediction · quantitative similarity-activity analysis · machine learning · software tools.

1 Introduction

The development of artificial intelligence methods has facilitated the exploration of diverse machine learning techniques to analyze material properties and synthesize materials with predetermined characteristics. Notably, the AlphaFold[7][8] project, which automates protein folding, and the M3GNet[4][9] project, which automates compound property prediction, have gained widespread recognition.

This paper aims to review the preliminary stages and research directions for the development of well interpreted machine learning (ML) methods and their implementation in the form of research toolbox software (RTS) for solving specific problems of compound property analysis. For the test case the study examines the biocidal properties of reagents, as there is currently a lack of confirmed correlations between these properties and the characteristics of the reagents. Besides, there is no established relationship between the characteristics of the

[⋆] This work was financially supported by a grant from the Ministry of Science and Higher Education of the Russian Federation №0706-2020-0026

[⋆⋆] The article was prepared within the framework of the HSE University Basic Research Program

reagents and the composition and structure of microorganisms, particularly bacteria and viruses. This multifactorial problem can only be addressed through the formulation of optimal reagent parameters for specific microorganisms or groups of bacteria, fungi, and viruses, which can be achieved through the use of RTSs.

The rapid advancement of ML has prompted the authors to revise the requirements for the RTS. This is necessary to comprehend the emergence of promising implementations and to cater to the needs of conducting specific ML experiments conveniently, including the interpretation and design of their outcomes. The primary requirements for an RTS include: 1) support for computational experiment methodology; 2) user-friendliness; 3) extensibility; 4) utilization of open-source tools only; 5) ability to import/export data seamlessly at any stage of an experiment.

Furthermore, collaborative work based on modern cloud platforms is essential, while also supporting local experiments. This can be achieved through an appropriate RTS database and a knowledge base scheme. It may come as a surprise that there is still no reference implementation of RTS that fully meets the set of requirements, which were briefly outlined before. This, however, could be easily explained by the persistent technological advancements and that such a project needs a transdisciplinary team as one of the resources. Thus, one of the authors participated in the development of the RTS for Structural Analysis "Graph Model Workshop", which adheres to the stated methodology and emphasizes transdisciplinarity. However, the technology used in the previous work requires updating.

Simultaneously, the methodology of computational experiments is undergoing active updates, as highlighted in the article "Computational Experiments: Past, Present and Future"[15], which focuses on simulation. Additionally, the development of open platform projects has reached a threshold where knowledge of the methodology and integration features enables the implementation of powerful scenarios in a relatively short time while considering the fundamental needs of both analysts and subject matter experts.

As a reference system for particular cases, we consider the NOMAD[14] project, which open database has already accumulated more than 10 million compounds and 100 million results of computational experiments. To understand the vastness of the technological stack, one can explore the logical architecture of the NOMAND, where the main technologies are indicated (Fig. 1).

2 Interpretability

As the area of material properties research advances, researchers are increasingly interested in the appropriate application of artificial intelligence methods. With the development of more diverse and sophisticated mathematical methods[12][16], questions of interpretability and explainability have become more pressing. Moreover, it is important to note that the regulatory burden is expected to increase dramatically[10]. This trend has already been observed in medical informatics, where the accumulated experience with randomized clini-

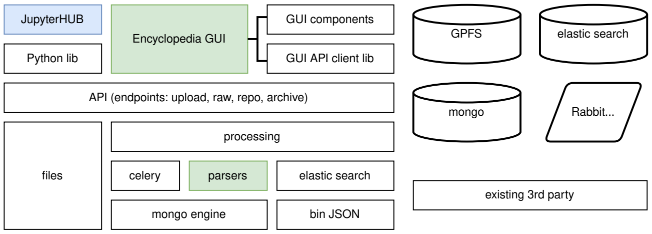


Fig. 1. Architectural composition of the NOMAD system

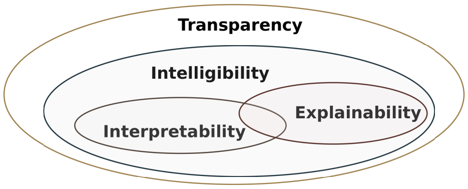


Fig. 2. Basic Concepts of Interpretability

cal trials has led to greater regulatory oversight[5]. Let us emphases that *transparency* has been considered a crucial addition to *reproducibility* for a long time. So far, a successful explanation of the meaning of the terminology remains the work ”A Survey of Explainable AI Terminology”[11], where Figure 2 is given. Moreover, the terminological system has shifted away from the original concept of ”interpretability,” which remains central. The serious research on interpretability of ML methods in its modern form began in 2015, after the full implementation of deep ML and the first generative models had appeared. The papers, which not only consider ML, but take into account the needs of experts as well, appeared in 2018 (g.e. the open work ”The Building Blocks of Interpretability” with interactive illustrations[6]). For a comprehensive collection of recipes for ”standard” ML methods, Molnar’s constantly updated work[3] is the best resource.

3 Solution architecture

The proposed RTS architecture largely resembles many of today’s collaborative RTS, but with a simplified business logic layer (backend), a more specific user interface (frontend), and a focus on the work as ML experts as subject matter experts. The system places a conscious emphasis on ease of use and integration with other researcher tools, including Microsoft Excel.

From a user’s perspective, the system presents two main workplaces:

1. Material Scientist (Domain Expert) can access a web application that utilizes Apache Foundation community components;

2. Data Analyst can access additional workplaces based on the JupyterLab.

The primary RTS data repository is a relational database (PostgreSQL) that was extended with ML capabilities integrated into tables (project MindsDB) and feature engineering (project Featuretools). The database was extended with a (the Apache AGE that provides graph database functionality for PostgreSQL).

The RTS data model can be divided into four parts:

1. General ontologies, property (feature) descriptions, and standard data sets (units, physical constants, etc.).
2. Data on the studied entities of a particular subject area (e.g., bacteria in the study of antibacterial effects).
3. Data on investigated chemical compounds.
4. Data from experiments and the data sets used in them (runs of computational experiments, results obtained, training, test, and validation samples).

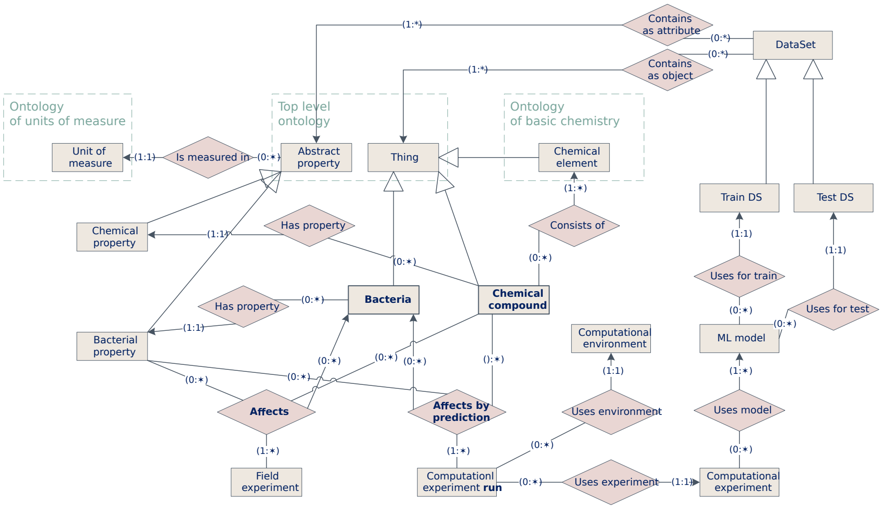
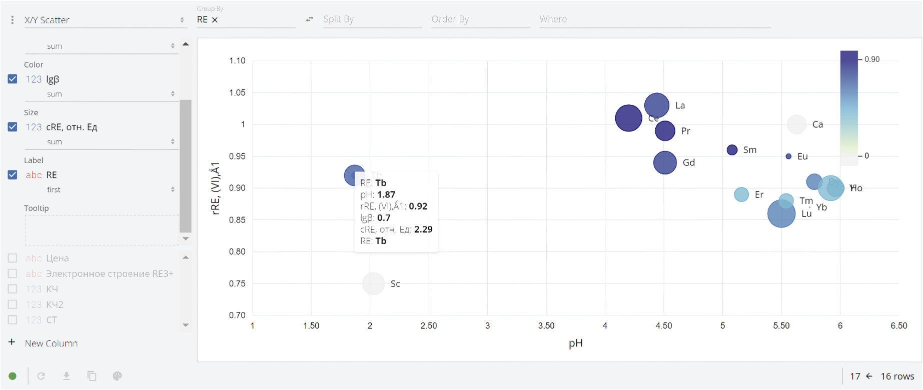


Fig. 3. Part of the RTS data model in terms of predicting antibacterial properties

The last two parts of the RTS data model are unified by storing data as triples ⟨*object*, *property*, *value*⟩. In such triple, an object represents an instance of an entity, a property – a semantic relation, and a value – any structured data element. For the convenience of use basic ML methods, it is recommended to work with *numeric* values. The part of the conceptual data model is shown in Fig. 3 (the entity occurrences in common ontologies are indicated). In the current development stage we have chosen to use "ChemFont"[1] ontology for chemical compounds, "NCBITaxon"[2] ontology for bacteria and "Ontology of units of Measure"[13] for units interpretation and conversion. The "affects" ternary relationships link the agent, object of influence, object properties, an experiment, and have several properties, such as the characteristics of influence (e.g., ”6% increase”). To predict the effect, a similar ”affects by prediction” relationship has been introduced, replacing a real experiment with a computational experiment. This allows the ontology-based cross-cutting comparison of results in subgroups, which serves as a basis for prediction-guided chemical experimentation.

4 Visualization

Classical visualizers of ML experiments are now being complemented by specific visualizers of spectrograms, diffractograms, and other variants of both input data and results. One of the directions for further research is the development, comparison, and integration of specific interpretation and visualization methods. Solutions from well-known companies such as OriginLab and WaveMetrics serve as a reference for this purpose. It is important for us to understand ”first principals” (physics, chemistry, and mathematics) behind a particular solution to ensure adequate comparison and selection of methods. To interactively investigate preliminary results we use Perspective, while Apache Superset is utilized to build final interactive reports. Figure 4 illustrates one of the 15 visualization options available for all user’s roles.



X/Y Scatter

Group By: RE ×

Split By

Order By

Where

sum

Color: ☑ 123 lgβ — sum

Size: ☑ 123 cRE, отн. Ед — sum

Label: ☑ abc RE — first

Tooltip

☐ abc Цена

☐ abc Электронное строение RE3+

☐ 123 КЧ

☐ 123 КЧ2

☐ 123 СТ

+ New Column

RE: **Tb**
pH: **1.87**
rRE, (VI),Å1: **0.92**
lgβ: **0.7**
cRE, отн. Ед: **2.29**
RE: **Tb**

17 ← 16 rows

Fig. 4. Example of an interactive study of a sample of lanthanides

5 Antibacterial activity analysis

In our paper, we provide an illustrative example of the analysis of antibacterial activity of Rare-earth elements (REE). This task holds immense value due to its potential application in the pharmaceutical industry. The study primarily employs rule-based models due to their interpretability and ability to provide insights into feature importance and relationships. The dataset used in the study comprises information on chemical compounds being tested and bacteria affected by REE. The target variable is diameter of the bacterial growth retardation. Chemical compounds data contains: ionic radii according to Shannon system; electronegativity of REE; structural type or crystal structure; coordination number of REE in the salt; coordination number of REE in solution; pH of solution; etc.

The inclusion of data on bacterial properties that confer resistance to active compounds is expected to enhance the understanding of the mechanisms underlying antibacterial activity. Bacteria data contain:

1. gram-test [gram-positive / gram-negative];
2. wall thickness;
3. wall structure [homogeneous / layered];
4. wall composition 1 [teixoic, lipoic and teixoic acids - up to 50% of the dry weight / absent];
5. wall composition 2 [high peptidoglycan (murein) content - 90% dry weight / small murein content - 5-10% dry weight];
6. outer membrane [no lipopolysaccharides, some strains have toxic glycolipids / lipopolysaccharides, phospholipids, lipoproteins];
7. membrane pore diameter;
8. proteins [antigen specificity determining proteins / porins];
9. penicillin sensitivity [high / low];
10. rigid and plastic cell wall layers [bound covalently / bound labilely].

Due to the nature of the domain and a limited number of experiments, the dataset currently has a large number of features as it contains both bacteria parameters and parameters of compounds. At the same time we due with a small sample with only 99 objects. This leads to the fact that traditional supervised learning approaches are not particularly appealing to domain experts. Rather than focusing on the accuracy of model predictions, they are more interested in obtaining evidences, hypotheses, and directions for further experiments. Therefore, rule-based models were selected as the initial models for computational experiments, as they are quite interpretable and enable to generate a range of hypotheses for future work.

Note that REE is used in the experiments in two forms: salt and solution. It has been observed that the choice of form significantly impacts on the obtained results (fig. 5). To support this, the training sample was employed in its entirety, as well as separately for salts and solutions. The training process involved the use of three models, namely RandomForest, GradientBoosting, and XGBoost. The performance of these models was evaluated using mean absolute error, and the results are presented in Table 1.

Parameter importance charts were generated (figures 5, 6, 7) based on the analysis of the best models. The results indicate that for the mixed dataset, the feature with the highest importance is "Is salt". Additionally among others, the model takes into account such features as crystal structure type, wall thickness, electronegativity, etc. However, these features are assigned relatively low weights, less than 0.03. This means that a state of active compound strongly affects results

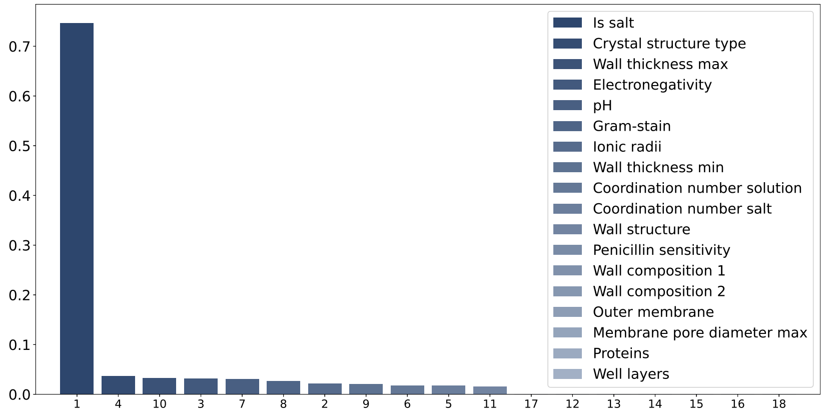


Fig. 5. Features importance for XGBoost model on mixed dataset

Model	Mixed dataset		Salts dataset		Solutions dataset	
Measures	MAE	R2	MAE	R2	MAE	R2
RandomForest	4.44	0.38	**4.80**	0.079	3.89	-0.98
GradientBoosting	4.52	0.49	4.97	0.007	3.41	-0.49
XGBoost	**4.42**	0.51	4.90	0.046	**3.40**	-0.52

Table 1. Model training results measured in MAE and R2

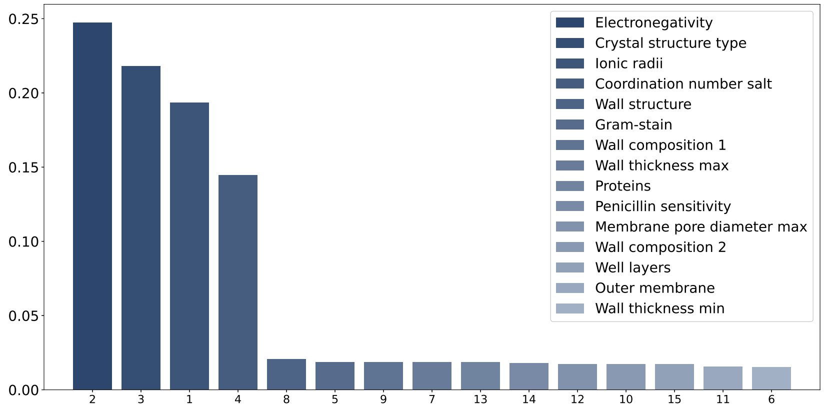


Fig. 6. Features importance for RandomForest model on salts dataset

and we should analyse them separately. After splitting the dataset, the results for solutions shew a significant improvement and results for salts become poorer, which, however, can be explained by the small size of dataset. After splitting the models begin to assign higher importance to other features, with coordination number being the most important feature for solutions, while electronegativity is the most important feature for salts. It is worth noting that the model trained on the solutions dataset does not account for bacterial properties. In contrast, the model trained on the salts dataset takes into account bacterial properties such as Gram-stain and wall structure, as well as other properties with less weight. The obtained results were used to confirm the hypothesis of a significant effect of electronegativity on antibacterial activity and facilitated the development of an additional set of experiments. Model hyper parameters we have used are listed in the table 2.

Parameter	XGBoost mixed	XGBoost solutions	RandomForest salts
colsample_bytree	0.8	0.8	-
gamma	1	0.5	-
max_depth	3	3	4
min_child_weight	17	15	-
subsample	1.0	0.6	-
learning_rate	0.02	0.02	-
n_estimator	600	600	900

Table 2. Hyper parameters of different models

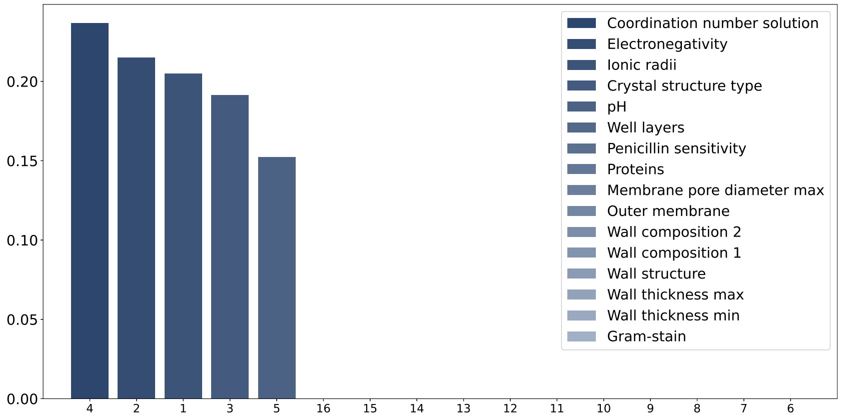


Fig. 7. Features importance for XGBoost model on solutions dataset

6 Results and prospects

The paper presented an environment for computational experiments in material science and related fields. It is designed to create and compare *well-interpretable* ML artifacts with ontology-based explanation. As a result, we applied our system to analyze antimicrobial activity. We identified the parameters that have the greatest impact on the final result of the substance's antimicrobial activity by using rule-based model. This information seems to be valuable in guiding future experiments and selecting the most promising parameters for testing. In addition to analyzing antibacterial activity, we are also testing the system's efficacy in optimizing the synthesis of materials, specifically titanium sols.

Currently, our team is engaged in the active enhancement of the user interface and data model of our research toolbox. A crucial aspect of this development is the incorporation of additional ontologies (chemical, biological, etc.), that are expected to substantially augment the system's capabilities. We are optimistic that the system for simultaneous conducting/analyzing "*virtual*" experiments and interpreting/directing "*real*" experiments will prove to be a valuable asset to domain researchers in material science.

References

1. ChemFOnt. https://www.chemfont.ca/ontology_brows
2. NCI Term Browser. https://nciterms.nci.nih.gov/ncitbrowser
3. C. Molnar: Interpretable machine learning (2022)
4. Chen C., Ong S.P: A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science **2**, 718–728 (2022)
5. Chen H., Gomez C., Huang C.M., Unberath M: Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. NPJ Digital Medicine **5**(156) (2022)
6. Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, Alexander Mordvintsev: The building blocks of interpretability. Distill Journal (2018)
7. DeepMind EMBL-EBI: AlphaFold protein structure database. http://alphafold.ebi.ac.uk
8. Jumper J., Evans R., Pritzel A.: Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021)
9. Lab, M.V.: Matterverse.ai. http://matterverse.ai
10. Matus K.J.M., Veale M: Certification systems for machine learning: Lessons from sustainability. Regulation & Governance **16**(1), 177–196 (2022)
11. Miruna-Adriana Clinciu, Helen Hastie: A survey of explainable AI terminology. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019). pp. 8–13. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-8403
12. Oviedo F., Ferres J.L., Buonassisi T., Butler K.T: Interpretable and explainable machine learning for materials science and chemistry. Accounts of Materials Research **3**(6), 597–607 (2022)
13. Rijgersberg, H.: OM - Ontology of Units of Measure. https://github.com/HajoRijgersberg/OM (May 2023)
14. Sbailò L., Fekete Á., Ghiringhelli L.M., E: The NOMAD Artificial-Intelligence Toolkit: Turning materials-science data into knowledge and understanding. NPJ Computational Materials **8** (2022)
15. Xue X., Yu X.N., Zhou D.Y., Wang X., Zhou Z.B., Wang F.Y: Computational experiments: Past. Present and Future **2202**(13690) (2022)
16. Zhong X., Gallagher B., Liu S., Kailkhura B..H.A.H.T.Y: Explainable machine learning in materials science. NPJ Computational Materials **8** (2022)