HELMHOLTZ
MUNICH

Helmholtz Zentrum München Deutsches Forschungszentrum für
Gesundheit und Umwelt (GmbH), Postfach 11 29, 85758 Neuherberg

Chemical Sciences Discipline Council
Gdańsk University of Technology

Dr. Igor Tetko
Group leader Chemoinformatics
Institute of Structural Biology

+49 89 31873575
igor.tetko@helmholtz-munich.de

*Re: evaluation of thesis of Karol Baran: "Few-shot machine learning for multimolecular systems modeling" submitted to Gdańsk University of Technology*

Dear Faculty,

In his doctoral dissertation, Mr Karol Baran investigates the application of machine learning (ML) methods, specifically graph neural networks and traditional approaches, to predict the properties and toxicity of ionic liquids. Given the limited data available in this field, the author focuses particularly on few-shot learning and meta-learning techniques. He also proposes novel methods that were developed specifically to improve the accuracy of chemical property predictions when adapting pre-trained models to very small datasets (domain-specific adaptation).

The aim of Baran's research was to use modern computational methods used in machine learning (ML) for modeling systems based on ionic liquids (ILs). The central hypothesis is that both traditional and modern machine learning methods can prove highly effective for this task, provided they are properly developed and applied. To achieve these objectives, the author formulated a series of research questions that formed the basis of his dissertation.

This dissertation has excellent originality. First and foremost, Mr. Baran developed a methodology for using graph neural networks (GNNs) to model ionic liquids, analyzing optimal ways to combine ions into a unified graph. To the best of my knowledge, his paper represents the first comprehensive comparative study in this field. He also proposed a number of innovative approaches in the realm of few-shot learning. In particular, he introduced the Attentive Model-Agnostic Meta-Learning (Att-MAML) method—an extension of the MAML algorithm that incorporates an auxiliary scaling layer; this layer modulates the significance of "meta-learned" structure-toxicity relationships in data-scarce scenarios (e.g., when the number of samples is very small). Furthermore, he proposed the TSA-Reptile (Task Similarity Aware Reptile) algorithm, which dynamically adjusts the learning rate based on the similarity between the current task and the distribution of tasks encountered during meta-learning (using the Tanimoto coefficient for solute

molecules as the similarity metric). Finally, he demonstrated that synthetic data (generated via the COSMO-RS method) can be used for model pre-training and the encoding of physically grounded relationships, thereby improving the results of subsequent fine-tuning on experimental data.

New methods based on graph neural networks (GNNs) were compared with traditional machine learning methods (utilizing descriptors) as well as other representation learning techniques (such as transformers) to identify the most effective approaches. The author correctly employed statistical tests (e.g., the Kruskal-Wallis test, t-test, etc.) to assess the statistical significance of the results. These tests confirmed that the author's conclusions are statistically grounded rather than the result of random correlations or “luck split” of data. A minor observation: confidence intervals were not provided for all tables; their inclusion would have facilitated an understanding of whether statistically significant differences exist between the presented values. Confidence intervals could have been calculated through repeated computations (e.g., using resampling methods, as the author did for certain other tables) or via bootstrapping (see, for instance, DOI: 10.1002/minf.201300030). Furthermore, when presenting numerical data in the text and tables, confidence intervals would have helped determine the appropriate number of significant figures (for example, it is likely that not all three digits in Table 5.6 are significant). It would be advisable to include such intervals in all tables for the public defense.

The author tested the developed methods using collected data covering various properties of ionic liquids (physicochemical characteristics, toxicity, etc.) and ranging in volume from small to extensive datasets. This enabled the validation of the proposed solutions against various target parameters and datasets of different sizes.

The dissertation is well-written, using clear language and maintaining a logical flow of presentation. The research process undertaken by the candidate is easy to follow. I would like to offer a minor suggestion: captions for figures and tables should not merely state what is depicted but should also direct the reader’s attention to specific, relevant information. For instance, the caption "Figure 5.6: Scenarios for the transition from a bimolecular structure to a graph" would be more informative—allowing the reader to immediately grasp the key conclusion associated with the figure—if phrased as follows: "Figure 5.6: Scenarios for encoding molecular structures as graphs; it is shown that the separate optimization of both ions within a single graph yielded the highest model accuracy (based on the coefficient of determination)." Similarly, a caption for Figure 5.7 reading "Pre-training models on synthetic COSMO-RS data was used to improve prediction accuracy during subsequent fine-tuning on experimental data" would be highly clear to readers. Although such captions are longer, they enable the reader to understand the content of a figure or table without having to consult the main text to determine its significance. I believe that employing such informative captions in the future will significantly enhance the quality of the author's publications. No changes to the dissertation text itself are required; however, I recommend using such detailed captions for tables and figures during the public defense.

The dissertation is a cumulative thesis based on five peer-reviewed research articles—with the candidate as the first author—accompanied by a comprehensive overview of the work. The overview follows a logical structure, encompassing an introduction to the field, an analysis of related research, a description of the methods employed, and a synthesis of the results and conclusions. The five articles are included in the appendix; the author's contribution to each is clearly defined, with the candidate playing a pivotal role in all the studies. The candidate is also a co-author of two additional articles not included in the dissertation. All articles have been published in prestigious, high-impact-factor journals, further attesting to the high quality of the research.

The results of this dissertation can find practical application in industry, primarily through the advancement of learning theory for small-data scenarios. The models developed by the author can be used to prioritize laboratory experiments specifically in studying carbon sorption to improve capture technologies, extracting mixtures (such as lignin) using optimal solvents, conducting toxicity screening, and so forth.

Naturally, the developed approaches are not a panacea for all problems; this is evidenced by their limited accuracy in predicting the properties of ionic liquids containing completely new ions absent from the training set. The author demonstrated that, in the case of toxicity, for instance, only quantitative predictions hold practical value. This may limit the applicability of the developed methods, e.g., in the virtual screening of novel anions and cations. Although Mr. Baran proposed addressing this challenge through the use of more extensive and diverse meta-training sets, I believe it is also worth considering molecular representations more closely linked to physical principles (approaches that are gaining popularity in machine learning). These include, for example, the pre-training of graph neural networks (GNNs) on quantum-mechanical atomic-level characteristics (see, e.g., DOI: 10.1186/s13321-025-00970-0) and/or the use of equivariant neural networks (e.g., DOI: 10.1021/acs.chemrestox.3c00032), as well as large language models (LLMs; e.g., DOI: 10.1039/d4sc04401k). I would be interested to hear the author's opinion on these ideas during the public defense as potential avenues for future work.

It would also be interesting to know the author's opinion on using various definitions of similarity when solving problems, e.g., those based on the properties of the solute rather than the Tanimoto coefficient.

One of the dissertation's shortcomings is the lack of experimental validation for the obtained predictions. Such verification, which could have been conducted in collaboration with colleagues specializing in experimental measurements, would have been highly significant for the practical implementation of the results and for gathering valuable feedback from experimentalists. While this observation does not diminish the value of the work, it aligns with current trends in the development of computational methods in chemistry.

HELMHOLTZ
MUNICH

Overall, I conclude that the doctoral dissertation meets the requirements established by Article 187 of the Law of July 20, 2018, "On Higher Education and Science." I request that the candidate be admitted to the public defense.

Sincerely,

Igor V. Tetko

Tuesday, 25 August 2026, Munich