Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw
Banacha 2c Street, 02-097 Warsaw, Poland

mobile: +48504726203, e-mail: Dariusz.Plewczynski@pw.edu.pl, www: https://plewczynski-lab.org

Warszawa,
06.06.2025

Warsaw, 06/06/2025

Prof. dr hab. Dariusz Plewczyński
Laboratory of Bioinformatics and Computational Genomics
Faculty of Mathematics and Information Science,
Warsaw University of Technology
Laboratory of Functional and Structural Genomics,
Centre of New Technologies, University of Warsaw

REVIEW of PhD dissertation of Łukasz Maziarka, MSc

Adapting Deep Learning Architectures
for Drug Discovery

Completed at the Faculty of Mathematics and Computer Science

Jagiellonian University

under the supervision of
Prof. dr hab. Jacek Tabor
and dr Stanisław Jastrzębski

in the field of technical sciences
in the discipline of technical informatics and telecommunication

The work presented to me is the result of a successful drug discovery analysis in the data-driven artificial intelligence paradigm. The author managed to combine the experimental data with cheminformatics algorithms; artificial intelligence methods development; focusing on the process of drug discovery from computational, AI-driven perspective.

Cheminformatics is an interdisciplinary field that applies computational methods and data science techniques to chemical information, playing a critical role in modern drug discovery. It encompasses methods for managing chemical data, predicting molecular properties, conducting virtual screenings, and generating new chemical entities through computer-aided drug design (CADD). At its core, cheminformatics seeks efficient ways to handle the immense complexity and scale of chemical space, facilitating the identification and optimization of potential drug candidates. However, challenges persist in accurately representing molecules digitally due to their complex and dynamic structures, including capturing their 3D conformations, subtle chemical interactions, and combinatorial variability of chemical space.

To address these challenges, *deep learning* methods have recently emerged as transformative tools within cheminformatics. Techniques such as Graph Neural Networks (GNNs) effectively model molecules as interconnected nodes and edges, enabling automated learning of meaningful molecular representations and the direct prediction of properties without predefined descriptors. Attention-based architectures, adapted from natural language processing, further enhance these models by focusing selectively on chemically relevant molecular features and interactions, thereby improving predictive accuracy and interpretability. Additionally, deep generative models, including variational autoencoders (VAEs) and generative adversarial networks (GANs), enable the systematic exploration and synthesis of new molecules, significantly expanding the capacity to design compounds with targeted properties.

Generally, the integration of deep learning in cheminformatics substantially streamlines and accelerates the drug discovery pipeline. By rapidly predicting molecular properties and facilitating high-throughput virtual screening, these AI methods significantly reduce the reliance on labor-intensive laboratory experiments and decrease the likelihood of late-stage failures in drug development. Moreover, generative approaches allow for efficient lead optimization and rational drug design by proposing novel chemical entities that align closely with desired pharmacological profiles. Ultimately, the ongoing advancement of deep learning applications within cheminformatics holds

promise for transforming pharmaceutical research, greatly enhancing the speed, efficiency, and effectiveness of drug discovery processes.

Lukasz Maziarka in his thesis entitled "*Adapting Deep Learning Architectures for Drug Discovery*" addresses the innovative adaptation and optimization of deep learning architectures for drug discovery tasks within cheminformatics, particularly emphasizing computer-aided drug design. The research is structured around two core areas: predicting molecular properties and generating novel chemical compounds.

For molecular property prediction, Maziarka introduces multiple novel models, including the Set Aggregation Network (SAN) for efficient feature aggregation, the Spatial Graph Convolutional Network (SGCN) integrating 3D spatial data, and transformer-based architectures like the Molecule Attention Transformer (MAT) and Relative Molecule Attention Transformer (R-MAT), leveraging attention mechanisms to capture molecular interactions. Additionally, he developed HuggingMolecules, an open-source library facilitating the use of pre-trained molecular prediction models, and EmBERT, an adaptation of BERT for multitask learning.

In the molecular generation domain, the dissertation introduces Mol-CycleGAN, employing generative adversarial networks for molecule optimization, and PluGeN, a flow-based network for controlled molecular generation. These contributions significantly improve the prediction and optimization of molecular structures, thus accelerating drug discovery processes.

The overall contributions provide substantial advancements in the integration of artificial intelligence with drug discovery, offering improved predictive accuracy, efficiency, and innovative methods to generate novel molecules with desired pharmaceutical properties.

The subject of my assessment, i.e. the doctoral dissertation, in my opinion, is in full accordance with the conditions set out in Art. 187 of the Act of July 20, 2018 Law on Higher Education and Science (Journal of Laws of 2021, items 478, 619, 1630). It presents the originality of the solved scientific problem, general theoretical and technical knowledge candidate in the field of technical informatics, as well as the ability to conduct scientific work.

The doctoral dissertation of Lukasz Maziarka, MSc was prepared in the in the Department of Machine Learning at the Institute of Computer Science and Computational Mathematics, Faculty of Mathematics and Computer Science, Jagiellonian University, under the supervision of Prof. Jacek Tabor, PhD and Stanisław Jastrzębski, PhD. The thesis outcomes are highly recognized, evidenced by multiple high-impact publications and over 800 citations according to Google Scholar by December 2024. Maziarka's research resulted in nine papers, published in prestigious venues including CORE rank A and A* conferences and cheminformatics journals. Additionally, he served as the principal investigator of the NCN *Preludium* grant related to his thesis research.

The author presents the theses of the PhD in two pages summaries of the dissertation in Polish and English, the *Introduction and Motivation*, Part I dedicated to *Adapting Neural Networks for Molecular Property Prediction*, and Part II to *Generative Models for Molecular Design*, followed by the PhD candidate research profile, *Conclusions* of the thesis, information on publications, *Bibliography*, finally, the full text of the publications included in the thesis.

In the first chapter, the author presents an introduction to the general characteristics of molecules, motivation for the thesis. The *Introduction* section of Łukasz Maziarka's thesis emphasizes the transformative impact of deep learning (DL) across various fields such as computer vision and natural language processing, due to the structured nature of their data. However, applying DL techniques to cheminformatics and drug discovery poses distinct challenges due to the complex and highly sensitive nature of chemical structures, where minor structural differences significantly influence biological activity. The thesis outlines the drug discovery process as lengthy, costly, and prone to high failure rates, highlighting the critical need for improved computational predictive methods. DL has the potential to substantially enhance various aspects of this process, such as virtual screening, molecular docking, molecular design, property prediction, and optimization. Furthermore, recent successes in AI, notably the 2024 Nobel Prize-winning AlphaFold2, underscore the importance and feasibility of integrating AI-driven approaches into chemistry and medicine. A central challenge identified is the lack of a universally agreed molecular representation, with common methods including SMILES strings, fingerprints, molecular graphs, and 3D spatial representations each presenting their own advantages and limitations. Selecting appropriate molecular representations and adapting DL architectures to effectively handle these challenges are critical to advancing drug discovery.

In the second chapter, the author focuses on adapting Neural Networks for molecular property prediction. Part I of the thesis addresses the critical challenge in cheminformatics of accurately predicting molecular properties, such as biological activity, solubility, and toxicity. It introduces several deep learning models specifically adapted for drug discovery, emphasizing effective molecular representation. Key contributions include the Set Aggregation Network (SAN), which optimizes feature aggregation from molecular graphs; the Spatial Graph Convolutional Network (SGCN), incorporating 3D spatial information to capture molecular conformations and stereochemistry; the Molecule-augmented Attention Transformer (MAT) and Relative Molecule Attention Transformer (R-MAT), employing self-attention mechanisms to dynamically capture molecular interactions; and HuggingMolecules, an open-source library facilitating the use of these pre-trained transformer-based models. Additionally, the research highlights the significance of atomic representation choices as crucial hyperparameters and explores multitask learning adaptations using a BERT-based architecture, significantly enhancing prediction performance by efficiently leveraging shared parameters across tasks.

The third chapter covers the topics related to the generative models for molecular design. Part II of the thesis focuses on generative deep learning models for molecular design, essential for drug discovery tasks such as lead optimization and exploring novel chemical spaces. The thesis introduces two key generative methods: Mol-CycleGAN, which adapts CycleGAN in the latent space of a pre-trained graph-based Variational Autoencoder (JT-VAE) to selectively modify existing molecules while preserving their core structures, and PluGeN, a Plugin Generative Network employing a flow-based module to transform latent spaces of existing generative models into disentangled representations. Both methods effectively generate molecules with specific targeted properties or enhance existing molecules, demonstrating significant potential for accelerating early-stage drug discovery through innovative molecular generation approaches.

The fourth chapter is dedicated to the PhD candidate research profile. He has contributed to numerous publications in prestigious venues, proving his expertise in areas such as deep learning applications in drug discovery, molecular property prediction, and anomaly detection using flow-based models and graph neural networks. His contributions have been recognized with high impact factors and commendable MEiN points across various conferences and journals.

As a principal investigator, Łukasz secured funding from the National Science Center of Poland under the NCN Preludium grant, focusing on

transformer-based methods for active chemical compound discovery. Additionally, he actively participated as a co-investigator in other research grants, collaborating on projects exploring deep processing of structured data and investigating the robustness of molecular latent representations in autoencoders.

Łukasz Maziarka has established fruitful research collaborations with recognized institutions and organizations, including partnerships with NVIDIA and researchers from the Institute of Pharmacology of the Polish Academy of Sciences and Wrocław University of Technology. These collaborations have yielded significant outcomes, including innovative methodologies in molecule self-attention transformers and multi-label conditional generation from pre-trained models, published in high-impact venues such as the Journal of Cheminformatics and IEEE Transactions on Pattern Analysis and Machine Intelligence. In addition to his research contributions, Łukasz actively promotes science through organizing workshops and serving as a reviewer for conferences and journals in the field of machine learning and cheminformatics. His research profile proves his commitment to advancing the intersection of deep learning and drug discovery.

The fifth chapter contain the conclusions and discussion of thesis results. Lukasz Maziarka's PhD thesis concludes with significant contributions in adapting deep learning architectures for cheminformatics, particularly in molecular property prediction and molecule generation. His research has demonstrated the transformative potential of tailored deep neural network models and appropriate chemical representations to enhance both the efficiency and accuracy of drug discovery tasks. Widely recognized publications in conferences and journals underscore his impact, with innovative model architectures like MAT and R-MAT advancing molecular property prediction, and novel methods such as Mol-CycleGAN and PluGeN crossing boundaries in molecular generation. Beyond academic success, Lukasz Maziarka has secured substantial research funding, including a NCN Preludium grant, and contributed to other research projects.

Finally, the sixth chapter provides the information on publications constituting the collection of research papers.

The results of the PhD thesis were already published in nine papers:

- [P1] **Łukasz Maziarka**, Marek Smieja, Aleksandra Nowak, Jacek Tabor, Łukasz Struski, and Przemysław Spurek. Set Aggregation Network as a Trainable Pooling Layer. International Conference on Neural Information Processing (ICONIP), pp. 419–431, 2019. CORE A, 140 MEiN points
- [P2] Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Smieja, Łukasz Struski, Agnieszka Słowik, **Łukasz Maziarka**. Spatial Graph Convolutional Networks. International Conference on Neural Information Processing (ICONIP), pp. 668–675, 2020. CORE A, 140 MEiN points.
- [P3] **Łukasz Maziarka**, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, Stanisław Jastrzebski. Molecule-augmented attention transformer. Workshop on Graph Representation Learning, Neural Information Processing Systems (NeurIPS), 2019.
- [P4] **Łukasz Maziarka**, Dawid Majchrowski, Tomasz Danel, Piotr Gainski, Jacek Tabor, Igor Podolak, Paweł Morkisz, Stanisław Jastrzebski. Relative Molecule Self-Attention Transformer. J. Cheminf, 16, 3, 2024. 100 points.
- [P5] Piotr Gainski, **Łukasz Maziarka**, Tomasz Danel, Stanisław Jastrzebski. HuggingMolecules: An Open-Source Library for Transformer-Based Molecular Property Prediction (Student Abstract). AAAI Conference on Artificial Intelligence (AAAI), pp. 12949-12950, 2022. CORE A*, 200 points.
- [P6] Agnieszka Pocha, Tomasz Danel, Sabina Podlewska, Jacek Tabor, **Łukasz Maziarka**. Comparison of Atom Representations in Graph Neural Networks for Molecular Property Prediction. International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2021. CORE A, 140 MEiN points.
- [P7] **Łukasz Maziarka**, Tomasz Danel. Multitask learning using BERT with task-embedded attention. International Joint Conference on Neural Networks (IJCNN), pp. 1-6, 2021. CORE A, 140 MEiN points.
- [P8] **Łukasz Maziarka**, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, Michał Warchoł. Mol-CycleGAN: a generative model for molecular optimization. J. Cheminformatics, 12, 2, 2020. 100 MEiN points.
- [P9] Maciej Wołczyk, Magdalena Proszewska, **Łukasz Maziarka**, Maciej Zieba, Patryk Wielopolski, Rafał Kurczab, Marek Smieja. PluGeN: Multi-Label Conditional Generation from Pre-trained Models. AAAI Conference on Artificial Intelligence (AAAI), pp. 8647-8656, 2022. CORE A*, 200 points.

The dissertation is organized into two primary sections: the first addresses adapting neural network architectures for the prediction of molecular properties, covering research presented in the initial seven publications, while the second investigates neural network-based methods for the generation of novel chemical compounds with targeted properties, as detailed in the final two publications. As I described earlier, structurally, the thesis consists of five chapters: Chapter 2 covers adapted neural network models for molecular property prediction, Chapter 3 describes novel generative models for molecule design, Chapter 4 outlines the candidate's research profile, including grants and

additional contributions, and Chapter 5 provides a summary of key findings, contributions, and future research directions. The dissertation concludes by emphasizing the significance and broad impact of its contributions, demonstrated through publications in prestigious machine learning and cheminformatics venues and extensive recognition within the drug design community. Specifically, the Mol-CycleGAN and MAT models are highlighted due to their considerable citation counts and practical adoption in academia and industry.

The author in his PhD thesis therefore demonstrated significant advancements in deep learning methods for computational modeling of molecular properties, the adaptation and optimization of deep learning architectures for cheminformatics, particularly in computer-aided drug design. Lukasz Maziarka developed advanced neural network models for molecular property prediction (SAN, SGCN, MAT, R-MAT, and EmBERT) and innovative generative methods (Mol-CycleGAN and PluGeN) for molecule generation and optimization. Overall, the thesis presents a comprehensive strategy for addressing cheminformatics challenges through specialized deep learning models, significantly enhancing both predictive accuracy and efficiency in drug discovery.

The thesis by Lukasz Maziarka asserts that despite the inherent challenges in representing molecular structures, deep learning architectures can effectively be adapted to address key tasks in cheminformatics. Specifically, the research focuses on enhancing the prediction accuracy of molecular properties and advancing methods for generating novel molecules. This includes the development of specialized neural network models tailored to diverse chemical data representations, such as graph-based architectures that integrate 3D spatial information. These models are designed to capture intricate molecular interactions and improve the precision of property predictions. Additionally, the thesis introduces innovative generative models capable of synthesizing new chemical compounds with specified characteristics, thereby accelerating the discovery of potential drug candidates. Moreover, the study contributes accessible tools and frameworks aimed at fostering continued innovation and practical implementation in the field. By enhancing precision, efficiency, and cost-effectiveness in early-stage drug discovery, these advancements are poised to significantly impact the pharmaceutical research landscape.

Below I will try to briefly summarize the key research achievements of the PhD student, which were presented by the author and the reviewer:

- *Adaptation of Deep Learning for Cheminformatics:* Maziarka successfully adapted deep learning architectures to tackle challenges in cheminformatics, particularly in drug discovery. This involved developing specialized neural network models tailored to different types of chemical data representations, such as graph-based architectures incorporating 3D spatial information. These models significantly enhanced the accuracy of predicting molecular properties, addressing the complexities of molecular structure and activity prediction.
- *Innovative Generative Models for Molecular Design:* The thesis introduced generative models capable of creating new chemical compounds with specific desired characteristics. This included methods utilizing generative adversarial networks (GANs) and other deep learning techniques to explore and optimize chemical space efficiently. By generating molecules that meet defined criteria, these models accelerated the process of lead optimization and novel compound discovery.
- *Development of Practical Tools and Frameworks:* Maziarka's research also contributed practical tools and software frameworks designed to facilitate further development in cheminformatics and drug discovery. These resources aim to make it easier to implement the advanced deep learning models in pharmaceutical research, therefore enhancing precision, efficiency, and cost-effectiveness in early-stage drug development.
- *Recognition and Impact:* The PhD student's work has been well-received within the scientific community, evidenced by substantial citations and publications in international journals and conferences. The research not only advances the theoretical understanding of deep learning in cheminformatics but also offers practical solutions that support drugs' discovery and optimization, especially molecular representation, property prediction, and molecule generation.

Suggested future research directions:

- *Enhanced Neural Architectures*: Develop more sophisticated neural network architectures tailored to specific challenges in cheminformatics, such as handling complex and time-dependent molecular interactions and improving predictive accuracy across diverse datasets.
- *Advanced Molecular Representations*: Explore other methods for representing molecular structures, including 3D-aware embeddings that

capture spatial relationships and equivariant networks that respect molecular symmetries. These approaches can enhance the quality of molecular property predictions and generative modeling.

- *Integration of Physics-Informed Models*: Incorporate physical constraints and principles into deep learning models, enhancing their ability to simulate realistic molecular behaviors and interactions. This fusion could lead to more robust predictions of drug-target interactions and pharmacokinetic properties.
- *Scalability and Efficiency*: Address scalability challenges by optimizing training procedures for large-scale molecular datasets. This includes exploring distributed learning techniques and hardware acceleration to handle increasingly complex models and vast chemical spaces efficiently.
- *Interdisciplinary Applications*: Extend deep learning models to integrate with fields like structural biology, genomics, and materials science. For instance, combining protein-ligand modeling with molecular design to predict binding affinities and optimize drug candidates.
- *Foundation Models in Chemistry and Biology:* Investigate the development of domain-specific foundation models analogous to BERT or GPT. These models could learn comprehensive representations of chemical and biological data, hopefully facilitating transfer learning and enhancing predictive capabilities across various molecular tasks.
- *Explainable AI in Drug Design:* Enhance interpretability of deep learning models to provide insights into molecular predictions. Develop methodologies for explaining how model decisions are influenced by molecular features, aiding chemists in understanding structure-activity relationships and guiding experimental design.
- *Active Learning and Human-in-the-Loop Systems:* Implement active learning strategies that intelligently select informative data points for model refinement. Explore human-in-the-loop systems where domain experts interact with AI models to iteratively improve predictions and generate novel hypotheses.

Questions to PhD student:

- Can you explain the rationale behind the choice of graph neural networks (GNNs) and attention mechanisms for molecular property prediction in your thesis? What were the advantages and limitations of these approaches and how you compare them with other deep architectures?
- Your thesis introduces novel deep learning architectures like MAT and R-MAT for molecular property prediction. Could you detail the architectural

innovations and how they address challenges specific to cheminformatics in general?

- In the context of molecule generation, what were the main challenges you encountered with unconditional (Mol-CycleGAN) and conditional (PluGeN) approaches? How did these methods advance the generation of the novel chemical compounds?
- Your collaboration with NVIDIA resulted in the development of the Relative Molecule Self-Attention Transformer. Can you describe how this model differs from traditional transformers and its impact on molecular modeling tasks?
- How did your research on transformer-based methods for active chemical compound discovery address scalability challenges in handling large chemical datasets?
- Could you discuss the implications of your work on multi-task learning for molecular property prediction? How do you envision this approach evolving to handle increasingly complex chemical datasets and tasks?
- What role do you see for explainable AI techniques in enhancing the interpretability of deep learning models in cheminformatics, particularly for your methods and their applications?
- How can you manage to integrate physical constraints and principles into your deep learning models, such as with physics-informed neural networks?
- Can you elaborate on the methodology and results of your study comparing atom representations in graph neural networks for molecular property prediction? What were the implications for model performance and generalizability, and their impact on explainability?
- Looking ahead, what are the emerging challenges in applying Large Language Models (LLMs) to drug design and discovery, and do you plan to address these challenges in your future research endeavors?

Conclusion

In the summary of my assessment of the doctoral dissertation of Lukasz Maziarka, MSc under the title "*Adapting Deep Learning Architectures for Drug Discovery*", I have to say that I very highly rate the presented work.

Considering the readability and scientific value of the doctoral dissertation, the successful combination of carefully described cheminformatics pipelines and deep learning tools, as well as chemically relevant computational

models, I value the doctoral dissertation of Lukasz Maziarka as an important contribution to drug discovery and informatics.

The doctoral dissertation meets the conditions set out in Art. 187 of the Act of July 20, 2018 Law on Higher Education and Science (Journal of Laws of 2021, items 478, 619, 1630). Moreover, I believe that this dissertation exceeds all customary and statutory requirements for doctoral dissertations, constitutes an original solution to a scientific problem, demonstrates the candidate's general informatics knowledge in cheminformatics and drug discovery techniques and demonstrates the ability to independently conduct scientific work.

Therefore, I am pleased to submit my review to the Council of the Scientific Discipline of Technical Informatics and Telecommunication of the Faculty of Mathematics and Computer Science, Jagiellonian University in Cracow to admit Lukasz Maziarka to the next stages of his doctoral dissertation.

In addition, considering the high substantive and extensive level of the dissertation (Part I and Part II detailed descriptions), nine published contributions, clear and transparent way of presenting the research topic, methodology, and main results, I would like to request that the dissertation should be distinguished with an appropriate award.

Dariusz Plewczynski, PhD, Professor of Exact and Natural Sciences; Principal Investigator
Phone: +48 22 554 36 54 or +48 22 234 7219
e-mail: d.plewczynski@cent.uw.edu.pl or Dariusz.Plewczynski@pw.edu.pl www: https://plewczynski-lab.org

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw; Banacha 2c Street, 02-097 Warsaw, Poland
Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology; ul. Koszykowa 75, 00-662 Warsaw, Poland