

# Toward Explainable Biomedical Deep Learning

Training and Explaining Neural Networks in Bioinformatics and Medicinal Chemistry

Andrea Mastropietro<sup>1,2,\*</sup>

<sup>1</sup>Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

<sup>2</sup>Lamarr Institute for Machine Learning and Artificial Intelligence, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

## Abstract

Deep learning is a powerful tool for biomedical applications. However, it has a shortcoming that cannot be underestimated: the absence of interpretability. Therefore, the aim of this doctoral research is to propose a comprehensive biomedical deep learning pipeline enriched with explainable artificial intelligence components. This pipeline, which goes from the discovery of disease-associated genes to the development of novel drugs, by opening the black box and rationalizing predictions, can enable a more effective and transparent usage of neural network-based models in real-world bioinformatics and chemoinformatics scenarios.

## Keywords

deep learning, explainable artificial intelligence, bioinformatics, chemoinformatics

## 1. Introduction

Deep learning has been extensively used in bioinformatics and chemoinformatics, delivering promising results in tasks such as disease gene identification and molecular activity prediction. However, its widespread adoption in biomedicine is hindered by the inherent black-box character of neural networks, whose complex, nonlinear mechanisms often make their decisions hard to rationalize. In fields where understanding the underlying biological rationale is critical, such as determining the genetic basis of diseases or the efficacy of therapeutic compounds, this lack of transparency undermines trust and limits practical utility.

To overcome this challenge, explainable artificial intelligence (XAI) is needed to reveal which input features influence predictions and how those features interact. This research contributes to this goal by developing and applying novel XAI techniques specifically designed for deep learning models in biomedical contexts. These methods are integrated into a comprehensive biomedical deep learning pipeline, enabling explainable outputs at each stage, from gene prioritization to drug repurposing and design.

In addition to deep models, the research deals with classical machine learning and network-based algorithms, which are relevant in the life sciences and therefore hold a place within this work. The resulting framework demonstrates that deep learning can be both powerful and explainable, offering scientists tools that are not only accurate but also trustworthy.

## 2. The Explainable Biomedical Deep Learning Pipeline

This doctoral research introduces an explainable biomedical deep learning pipeline, illustrated in Figure 1, which integrates multiple components across the two main domains of bioinformatics and chemoinformatics, leveraging large-scale biological and chemical data. In the bioinformatics area, the pipeline begins with the training of a gene discovery model (block 1), whose predictions must be

ITADATA2025: The 4<sup>th</sup> Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

\*Corresponding author.

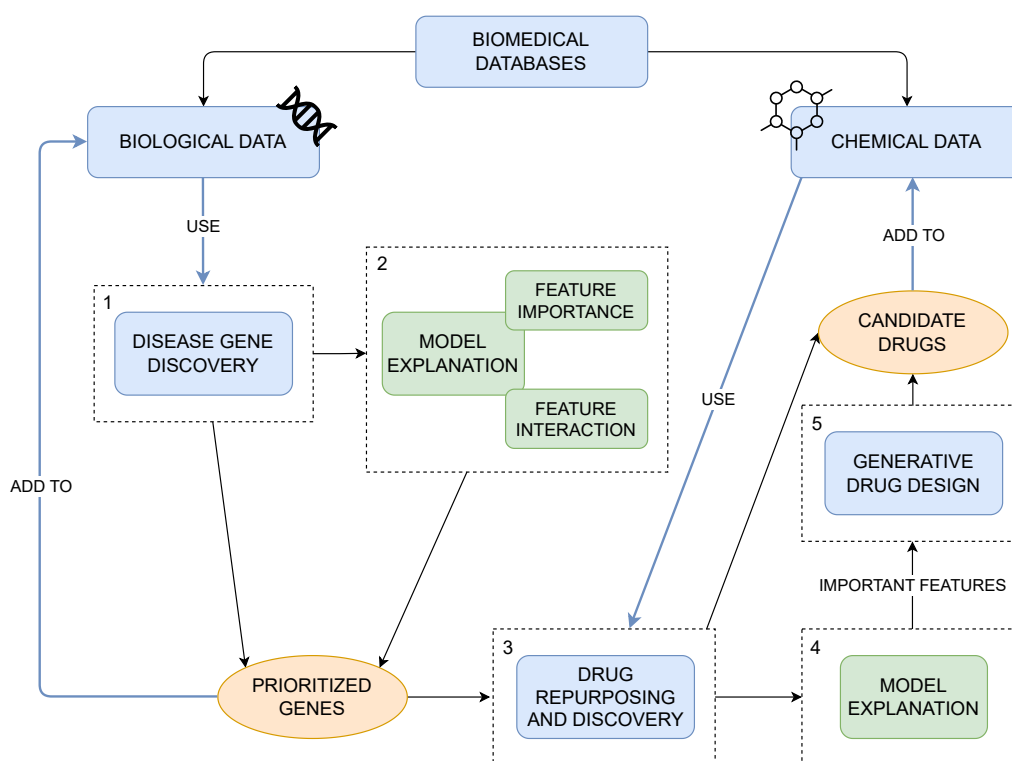
✉ mastropietro@bit.uni-bonn.de (A. Mastropietro)

🌐 <https://www.mastro.me/> (A. Mastropietro)

🆔 0000-0002-3456-9428 (A. Mastropietro)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



**Figure 1:** The proposed explainable biomedical deep learning pipeline.

explained (block 2). To this end, the research explores two complementary aspects of explainability: feature importance, which highlights the influence of individual inputs (e.g., specific gene mutations), and feature interaction, which uncovers how features act jointly (e.g., gene–gene epistasis).

To obtain explainable gene–disease predictions, we developed **XGDAG** (eXplainable Gene–Disease Associations via Graph Neural Networks) [1], a framework that combines graph neural networks (GNNs) with explainability techniques such as GNNExplainer [2], GraphSVX [3], and SubgraphX [4]. This approach not only explains predictions but also aids in discovering novel gene associations through explainable outputs.

However, meaningful explanations require robust training, particularly challenging in bioinformatics due to the prevalence of positive–unlabeled (PU) data. To address this, we propose **NIAPU** (Network-Informed Adaptive PU Learning) [5], a novel method that uses a Markov diffusion process on biological networks to assign pseudo-labels with varying degrees of *positiveness*. These pseudo-labels allow GNNs to learn effectively, enabling XGDAG to produce accurate and interpretable results.

For feature interaction, we developed **EpiDetect**, a method designed to detect epistatic interactions from genome-wide data. By analyzing neural network weights, EpiDetect estimates the extent to which combinations of genetic variants influence a phenotype, revealing complex trait mechanisms beyond single-gene effects.

Once disease-associated genes are identified, they can inform therapeutic strategies (block 3) and be used as targets for drug treatments. We demonstrate this with a case study on primary biliary cholangitis (PBC) [6, 7], where NIAPU-identified genes were used to guide **drug repurposing**. In this bioinformatics-driven approach, gene discovery guides the identification of new treatments.

The pipeline then transitions into chemoinformatics (block 4), where the graph-like nature of molecules is exploited. This makes GNNs a natural fit for predicting compound activity against target proteins. To explain these models, we developed **EdgeSHAPer** [8, 9], the first edge-centric Shapley value-based method for GNNs. Using a tailored Monte Carlo sampling approach, EdgeSHAPer efficiently identifies molecular substructures most responsible for activity prediction outcomes.

We further extended EdgeSHAPer to regression tasks (specifically, compound potency prediction in **protein–ligand interactions**) [10] to investigate whether GNNs truly capture meaningful biochemical interaction patterns. Our findings showed mixed results, revealing that GNNs tend to memorize ligand structures and can learn interaction-relevant information only when supported by high-quality graph representations: an unanticipated novel finding.

In addition, we address limitations in classical machine learning models. Approximate Shapley values often fail in support vector machines (SVMs) used for molecular activity prediction. To overcome this, we developed **SVERAD** (Shapley-Value Expressed Radial Basis Function) [11, 12], a method that computes exact Shapley values from binary molecular fingerprints in quadratic time, providing reliable feature attributions for SVMs.

The final component of the pipeline (block 5) opens to future research directions, focusing on **generative drug design**. Using the features identified in earlier stages, generative models can be guided to create new molecules with desired properties. These candidate drugs can then be validated and integrated into biomedical databases, closing the loop in a pipeline that is not only data-driven and predictive but also explainable and biologically grounded.

### 3. Results and Conclusions

This doctoral research presents a comprehensive explainable biomedical deep learning pipeline, beginning with the challenge of detecting disease-associated genes. This is at first addressed with NIAPU, a network-informed adaptive PU learning method that enables effective training in PU settings by propagating pseudo-labels through biological networks. NIAPU not only facilitates gene prioritization but also supports downstream explainability.

In the second stage of the pipeline, explainability comes into play through the XGDAG framework. By explaining GNN-based predictions, XGDAG generates explanation subgraphs highlighting candidate disease genes. Proper GNN training is made possible thanks to NIAPU’s pseudo-labeling. The results were validated through enrichment analysis, confirming that XAI methods can be used not just for post hoc explanation but as active tools for gene discovery. XGDAG represents the first approach to integrate PU learning and GNN explainability for this task.

Beyond single-gene associations, the research addresses the complexity of epistatic interactions through EpiDetect, a novel method that uses neural network weights to detect gene–gene interactions influencing diseases and traits. This deepens the biological insights offered by the pipeline and highlights the multifactorial nature of disease mechanisms.

The discovered genes can become targets for drug repurposing (block 3). In a case study on PBC, NIAPU was used to expand the set of target genes, leading to meaningful candidate drugs with an approach that is advantageous from both development time and safety profile perspectives.

Bridging into chemoinformatics, the pipeline employs GNNs to predict compound activity, supported by EdgeSHAPer, the first edge-centric Shapley value explanation method for GNNs. EdgeSHAPer identifies relevant molecular substructures and outperforms existing tools in both explanation accuracy and chemical relevance. It was further extended to predict compound potency in protein–ligand interactions. Surprisingly, we found that while GNNs often struggle to learn interaction patterns from overly simplistic graph representations, some models prioritize meaningful interaction edges when trained on high-quality data. These results emphasize the critical role of high-quality graph representations in enabling effective learning and consequent explainability.

To address explanations in classical models, we introduced SVERAD, a method for exact Shapley value computation in SVMs using binary molecular fingerprints. SVERAD provides reliable feature attributions in quadratic time rather than exponential, further enhancing trust in models used for compound activity prediction.

As future research directions, the output of these models, i.e., important molecular features and substructures, can be fed into the final stage of the pipeline: generative drug design (block 5). By guiding generative models with features deemed important by the previous steps of the pipeline, we can

enable the creation of molecules that are both novel and effective, filling the last gap in the proposed framework.

In summary, this doctoral research presents a complete and explainable deep learning pipeline for biomedicine, from gene discovery to drug design. Each component plays a pivotal role in enabling a transparent and trustworthy usage of deep learning models. We presented different XAI solutions working at every step of the pipeline, thereby enhancing the trustworthiness of neural networks in bioinformatics and chemoinformatics, and going toward explainable biomedical deep learning.

## Acknowledgments

I would like to thank Aris Anagnostopoulos, Paolo Tieri, and Jürgen Bajorath for their instrumental help and support during my doctoral studies. The Ph.D. thesis this extended abstract summarizes was defended on January 31<sup>st</sup>, 2024, at Sapienza University of Rome.

## Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-4 in order to: Grammar and spelling check. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

- [1] A. Mastropietro, G. De Carlo, A. Anagnostopoulos, XGDAG: explainable gene–disease associations via graph neural networks, *Bioinformatics* 39 (2023) btad482.
- [2] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: generating explanations for graph neural networks, *Advances in Neural Information Processing Systems* 32 (2019).
- [3] A. Duval, F. D. Malliaros, GraphSVX: Shapley value explanations for graph neural networks, in: *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Proceedings, Part II* 21, Springer, 2021, p. 302–318.
- [4] H. Yuan, H. Yu, J. Wang, K. Li, S. Ji, On explainability of graph neural networks via subgraph explorations, in: *International Conference on Machine Learning*, PMLR, 2021, p. 12241–12252.
- [5] P. Stolfi, A. Mastropietro, G. Pasculli, P. Tieri, D. Vergni, NIAPU: network-informed adaptive positive-unlabeled learning for disease gene identification, *Bioinformatics* 39 (2023) btac848.
- [6] E. Shahini, G. Pasculli, A. Mastropietro, P. Stolfi, P. Tieri, D. Vergni, R. Cozzolongo, G. Giannelli, F. Pesce, Network proximity-based drug repurposing strategy for primary biliary cholangitis, *Digestive and Liver Disease* 54 (2022) S106.
- [7] E. Shahini, G. Pasculli, A. Mastropietro, P. Stolfi, P. Tieri, D. Vergni, R. Cozzolongo, F. Pesce, G. Giannelli, Network proximity-based drug repurposing strategy for early and late stages of primary biliary cholangitis, *Biomedicines* 10 (2022) 1694.
- [8] A. Mastropietro, G. Pasculli, C. Feldmann, R. Rodríguez-Pérez, J. Bajorath, EdgeSHAPer: bond-centric Shapley value-based explanation method for graph neural networks, *iScience* 25 (2022) 105043.
- [9] A. Mastropietro, G. Pasculli, J. Bajorath, Protocol to explain graph neural network predictions using an edge-centric Shapley value-based approach, *STAR Protocols* 3 (2022) 101887.
- [10] A. Mastropietro, G. Pasculli, J. Bajorath, Learning characteristics of graph neural networks predicting protein–ligand affinities, *Nature Machine Intelligence* 5 (2023) 1427–1436.
- [11] A. Mastropietro, C. Feldmann, J. Bajorath, Calculation of exact Shapley values for explaining support vector machine models using the radial basis function kernel, *Scientific Reports* 13 (2023) 19561.
- [12] A. Mastropietro, J. Bajorath, Protocol to explain support vector machine predictions via exact shapley value computation, *STAR Protocols* 5 (2024) 103010.