https://doi.org/10.1038/s41576-025-00826-1

Beyond the black box with biologically informed neural networks

David A. Selby, Maximilian Sprang, Jan Ewald & Sebastian J. Vollmer

Machine learning models for multi-omics data often trade off predictive accuracy against biological interpretability. An emerging class of deep learning architectures structurally encode biological knowledge to improve both prediction and explainability. Opportunities and challenges remain for broader adoption.

Machine learning applied to data from high-throughput technologies has transformed biological research, facilitating integration of diverse omics datasets into multi-layered models of cellular systems. However, the predictive power of traditional 'black box' machine learning algorithms — and their ability to model complex nonlinear relationships — often comes at the expense of biological interpretability. Biologically informed neural networks (BINNs) offer a promising solution, combining predictive accuracy with explainability by incorporating decades of accumulated prior biological knowledge. This emerging paradigm is particularly well-suited for understanding model predictions based on complex, high-dimensional multimodal datasets found in multi-omics integration.

From black boxes to visible neural networks

BINNs are artificial neural networks whose architecture is explicitly constrained by biological pathway ontologies (Fig. 1). Unlike conventional fully connected deep learning models, which rely on arbitrarily chosen numbers of hidden nodes and layers, BINNs are designed using known pathway hierarchies from databases such as Reactome, Gene Ontology or KEGG[1]. Each node in the network represents a real-world biological entity — such as a gene, pathway or biological process — and edges reflect known relationships between these entities. For instance, an input node representing gene expression levels is only connected to a hidden pathway node if the gene is a known member of that pathway. This structure, which contrasts with the opacity of traditional black box models, has led to the term visible neural networks[2,3], or transparent neural networks[4].

The biologically informed architecture of BINNs tackles several challenges simultaneously. First, the incorporation of well-curated biological knowledge reduces the number of model parameters, thereby decreasing the amount of training data required. Second, the structure of the model is intuitive for biomedical researchers, even those with limited machine learning expertise. Third, the reduced dependence on training data and inductive bias mitigates overfitting and increases generalizability. Finally, by emulating cellular and genetic regulation, BINNs bridge the gap between data-driven models and mechanistic biological understanding.

Applications and success stories of BINNs

Since their introduction around 2018, BINNs have been widely applied in biomedicine, with notable successes in oncology, drug response prediction and survival analysis[1,4]. For example, models such as P-Net[5] have demonstrated efficacy in aligning molecular features with therapeutic outcomes. Other extensions have integrated genomic data with chemical structure data to predict therapeutic efficacy[3] or combined multi-omics and clinical data to predict patient survival in precision medicine[6]. BINNs are not limited to supervised learning tasks; biologically informed variational autoencoders — a form of unsupervised learning model — have also been used to analyse cellular processes and aid drug development[7].

Recent applications have extended BINNs to single-cell sequencing, uncovering cellular heterogeneity and regulatory networks. Although early works already used multi-omics data, the integration of multiple modalities has increased with time. These models have also been used to uncover novel pathway interactions, demonstrating their potential as discovery agents[8].

Why BINNs excel in multi-omics integration

Multi-omics datasets are inherently high-dimensional, heterogeneous and often limited in sample size relative to the number of features. BINNs leverage biological priors to reduce model complexity by constraining the hypothesis space early in the analysis pipeline, which — ideally — improves generalizability and predictive performance and makes them particularly effective in these scenarios.

Comparative studies reveal that BINNs perform comparably to, or better than, fully connected neural networks on various predictive tasks. For example, BINNs seem to excel in scenarios with small, high-dimensional datasets, which are suboptimal for dense neural networks but typical of omics studies. They also outperform traditional machine learning models in capturing non-linear, hierarchical relationships inherent to biological systems (Fig. 1), enabling meaningful insights beyond prediction, such as discovery of novel biomarkers[5,8]. Multiple omics fit in this hierarchy, because nodes in BINNs can represent any biological entity, for example, genes, metabolites or protein complexes. In genomic assays, such as mutation or copy number variation measurements, the features are mapped to the gene that contains the aberration. In transcriptomics and proteomics, multiple transcripts or proteins can be mapped to one gene. In metabolomics, a metabolite may be mapped to genes encoding enzymes that use or produce the respective molecule. To accommodate multiple inputs, a common entity is chosen (that is, genes) or specialized input layers can be crafted. As BINNs are special cases of multimodal deep learning, different data fusion strategies can be explored[9].

Moreover, BINNs integrate predictive and explanatory tasks seamlessly. Traditional machine learning models, including dense neural networks, often fail to provide biologically meaningful insights owing to the inexplicability of their internal nodes. Post-hoc, model-agnostic

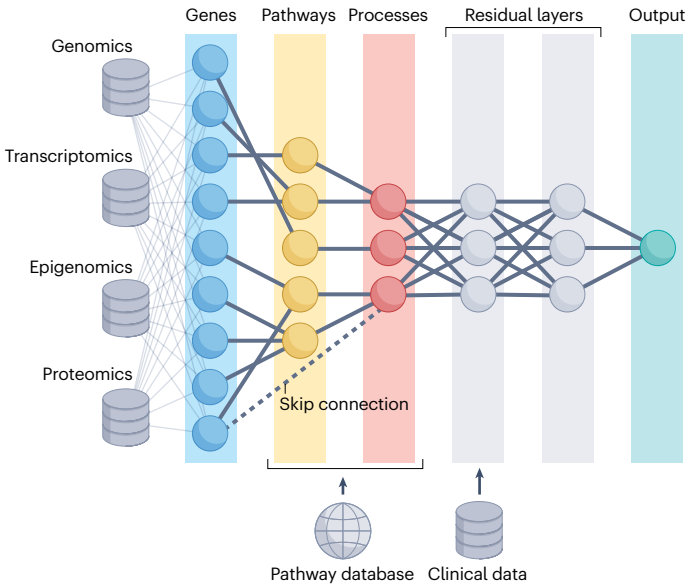

Fig. 1 | A biologically informed neural network architecture. Multiple omics are used as input and their respective features are linked to genes via known mappings. The genes are, in turn, connected to a hierarchy of biological ontology from a database (for example, genes, pathways and higher-order processes). To accommodate uneven hierarchies, skip connections (dotted line) or dummy nodes can be used. Fully connected residual nodes may capture interactions not included in the structural ontology used to build the architecture, which otherwise embeds a strong inductive bias in the predictions. Clinical measurements, or other data which cannot be linked directly to the pathway ontology, can be included via late fusion into the neural network or via dummy pathways.

interpretability methods offer input-level explanations but are prone to instability and can fail to reflect highly non-linear relationships[10], such as between genes and processes. By contrast, BINNs enforce interpretability as an intrinsic property, allowing predictions to be directly linked to specific genes or pathways. This ante-hoc approach enhances robustness by incorporating known biological constraints, making BINNs ideal for tasks that require both prediction and inference, such as biomarker discovery and drug target validation.

Advancing BINNs for biomedical discovery

Despite their promise, BINNs face several challenges. Most studies evaluate BINNs within narrow datasets and tasks, limiting insights into their generalizability across domains and conditions. The reasons for their apparently superior performance — whether due to biological inductive bias, multi-omics data fusion strategies or the introduced sparsity — remain unclear. Additionally, the lack of standardized benchmarks and tools hampers accessibility and reproducibility.

To fully realize their potential, future research should focus on developing robust frameworks for BINN construction and evaluation. Expanding the use of flexible architectures capable of handling various kinds of biological knowledge, incorporating advanced multimodal fusion strategies and systematically exploring the impact of different ontologies will be essential. Furthermore, leveraging BINNs for hypothesis generation, such as predicting novel pathway relationships, represents an exciting research opportunity.

BINNs may represent a transformative approach in computational biology, uniting predictive accuracy with biological interpretability. By embedding domain knowledge gathered over decades of genetic research, these architectures provide more transparent, data-driven biomedical models that reduce computational costs and enable built-in interpretability. However, to fully harness their potential, the field must address key challenges:

- Standardization: develop common benchmarks and tools to improve accessibility, reproducibility and study comparability.
- Rigorous evaluation: conduct more comprehensive evaluations and ablation studies to understand the mechanisms behind the performance of BINNs and their generalizability relative to alternative approaches, such as graph neural networks and classic machine learning.
- Flexible architectures: explore architectures that can incorporate diverse biological knowledge and advanced data fusion strategies.
- Hypothesis generation: combining modern neural architecture search methods with BINNs could unlock discovery of novel pathway interactions and regulatory mechanisms.
- Focus on the core: systematically investigate the choice of knowledge databases and the hierarchy level to build BINNs in close relation to their application.

Overcoming these hurdles may unlock the full potential of multi-omics and BINNs, paving the way for more explainable, data-driven discoveries in genomics, drug development and precision medicine.

David A. Selby [1] ✉, Maximilian Sprang [2], Jan Ewald [3] & Sebastian J. Vollmer [1,4]

[1]Data Science and its Applications, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. [2]Department of Dermatology, University Medical Center of the Johannes Gutenberg University, Mainz, Germany. [3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Leipzig, Germany. [4]University of Kaiserslautern–Landau, Kaiserslautern, Germany.

✉e-mail: david.selby@dfki.de

Published online: 04 March 2025

References

1. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinformatics* **24**, 198 (2023).
2. van Hilten, A. et al. Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *npj Syst. Biol. Appl.* **10**, 81 (2024).
3. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684 (2020).
4. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* **24**, 125–137 (2023).
5. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
6. Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H. & Kang, M. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med. Genomics* **12** (Suppl. 10), 189 (2019).
7. Seninge, L., Anastropoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat. Commun.* **12**, 5684 (2021).
8. Hou, Z., Leng, J., Yu, J., Xia, Z. & Wu, L. Y. PathExpSurv: pathway expansion for explainable survival analysis and disease gene discovery. *BMC Bioinformatics* **24**, 434 (2023).
9. Nguyen, T. et al. Optimal fusion of genotype and drug embeddings in predicting cancer drug response. *Brief. Bioinform.* **25**, bbae227 (2024).
10. Molnar, C. et al. in *xxAI – Beyond Explainable AI* (eds Holzinger, A. et al.) 39–68 (Springer, 2022).

Acknowledgements

D.A.S. and S.J.V. acknowledge support from the German Federal Ministry of Education and Research within project curATime (03ZU1202JA). J.E. acknowledges support of the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus under ERA PerMed (MIRACLE, 2021-055).

Competing interests

The authors declare no competing interests.