Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN LIFE SCIENCES: NAVIGATING THE PATH TO CLINICAL TRUST

Nilesh S Kulkarni[1], Kabir G Kharade[1] and K. Vengatesan[2]

[1]*Department of Computer Science, Shivaji University, Kolhapur (MS), India*

[2]*Department of Computer Science and Engineering,*

School of Engineering, Dayanand Sagar University, Bangalore, India

**Corresponding author E-mail: nsk.rs.csd@unishivaji.ac.in, kgk_csd@unishivaji.ac.in, vengiecse2005@gmail.com*

Received: 15 January 2025	Revised: 15 February 2026	Accepted: 10 March 2026	Published: 27 March 2026

DOI: **https://doi.org/10.5281/zenodo.19271611**

Abstract:

Artificial intelligence is rapidly transitioning from a supportive tool to an active partner in biomedical research, generating hypotheses and predicting clinical risks. However, patient-specific clinical integration requires mechanistic insight into model reasoning, not merely confidence estimates. This review evaluates current Explainable AI (XAI) methodologies—from SHAP and LRP to Concept Bottleneck Models—across genomics, drug discovery, and diagnostics. While post-hoc methods surface biologically plausible features, they often suffer from faithfulness issues, instability, and scalability limitations in real-world workflows. We propose a domain-aware XAI framework designed to satisfy clinical workflow constraints and evolving regulatory expectations including FDA guidance. Closing the translational gap requires intrinsically interpretable systems grounded in biological pathway logic and evaluated for equity across diverse populations. Ultimately, only by developing systems prioritizing biological faithfulness, equity-conscious evaluation, and regulatory-standardized benchmarking can the field bridge the persistent gap between predictive performance and genuinely accountable clinical trust and safety in healthcare settings.

Keywords: *Algorithmic Bias, Clinical Trust, Drug Discovery, Explainable AI, Foundation Models, Graph Neural Networks.*

1. Introduction

There is something almost paradoxical about the state of AI in medicine right now. Models have become genuinely useful — in some narrow domains, better than experts — and yet the richer and more capable those models become, the harder it gets to explain what they are actually doing. AlphaFold 3, for example, can predict the joint structure of proteins, nucleic acids, and small molecules with accuracy that was unimaginable a decade ago (1).

Yet the diffusion-based architecture that makes this possible encodes its reasoning across billions of parameters in a way that resists straightforward inspection.

In most engineering fields, a black-box system that works reliably is good enough. Medicine is different. A clinician who receives a mortality risk score needs to interrogate it — to check whether it aligns with what they know about the patient, and to catch cases where the model is generalizing from a population that does not match the patient in front of them. Regulatory bodies have started to reflect this: both the FDA and the European Medicines Agency have signaled that transparency is becoming a prerequisite for AI-based software used as a medical device, not a feature added at the end of development (2,3). XAI, in this context, is less about making AI look trustworthy and more about making it genuinely accountable. The stakes are high. Obermeyer *et al.* demonstrated empirically that a widely deployed clinical algorithm systematically underestimated the care needs of Black patients — not because the algorithm was designed to discriminate, but because it was trained on biased proxy variables (4). This is the foundational problem that XAI must solve: surfacing the reasoning behind a model's predictions in enough detail that such failures can be detected and corrected before they reach patients.

2. Overview of explainable ai approaches

XAI methods fall into two groups. Post-hoc tools apply after training. Intrinsic architectures build transparency in. Neither is perfect. Choice depends on clinical needs. Post-hoc methods are flexible. They layer onto existing models. No retraining is required.

- SHAP relies on game theory. It calculates feature contributions (5). Genomics uses this widely. Clinicians trace risk to specific variants. Opaque scores become interpretable (6,7). Reviews link this to higher trust in decision support.
- LIME probes models locally. It perturbs inputs to observe shifts. Linear surrogates explain specific data points (8). ICU teams use this for sepsis alerts. Clinicians identify driving physiological signals.
- LRP backpropagates output through layers. It redistributes relevance to inputs. Pfeifer *et al.* found LRP outperforms SHAP for gene selection in breast cancer (9). AttnLRP now covers transformers (10). This explains the full model behavior.

Rather than applying explanations after training, intrinsic architectures embed transparency directly into the model structure. Protein Language Models utilize attention maps to highlight residues driving binding affinity, although the field still debates whether these weights represent genuine mechanism or simply convenient proxies. For stricter accountability, Concept Bottleneck Models force predictions through intermediate layers of human-defined concepts—like metabolic pathway activity—so domain experts can validate the reasoning chain before accepting an output (11). This approach translates to small-molecule design as well, where Jiménez-Luna *et al.* applied concept whitening to graph neural networks, enabling the direct identification of structural motifs responsible for property predictions instead of relying on opaque latent states (12).

3. Applications across the life sciences

3.1 Genomics and multi-omics integration

Omics data presents significant modelling challenges due to high dimensionality, sparse samples, and complex molecular dependencies. XAI helps separate biological signal from noise in these environments. Chereda *et al.* used Graph Layer-wise Relevance Propagation (GLRP) to pinpoint patient-specific molecular subnetworks in breast cancer metastasis, identifying druggable drivers consistent with clinical knowledge (13). Similarly, Usman *et al.*

applied SHAP to Huntington's disease single-nucleus RNA sequencing, revealing altered genes distinct from standard differential expression analysis—indicating XAI provides complementary rather than redundant insights (7). While Partin *et al.* outline broader deep learning trends in drug response, LRP-based methods specifically show promise for unravelling cross-modal signalling relationships in integrated transcriptomic and proteomic data (14).

3.2 Accelerating drug discovery

Costly late-stage failures plague the Drug Design-Make-Test-Analyse (DMTA) cycle. Graph Neural Networks (GNNs) now standardize property prediction by representing molecules as atomic graphs, where XAI pinpoints toxicity-driving substructures for immediate medicinal chemistry intervention. Lavecchia confirms that integrating XAI bridges high-performance prediction with mechanistic understanding across ADMET and lead optimization workflows (15). Bibliometric analysis by Yao *et al.* further identifies XAI integration as a rapidly expanding research frontier within GNN-based drug discovery (16).

3.3 Clinical diagnostics

In clinical diagnostics, the disconnect between model accuracy and clinician trust remains critical. Jin *et al.* reported that generic XAI techniques frequently appear out of context or unhelpful to medical users, underscoring the necessity for domain-aware solutions rather than off-the-shelf explainability tools (17). Their guidelines for medical image analysis represent a concrete step toward aligning technical outputs with clinical workflow realities.

4. Critical challenges to clinical trust

4.1 The faithfulness problem

Generated explanations sometimes diverge from the model's actual decision boundaries. Post-hoc attributions can look biologically coherent while hiding reliance on spurious correlations rather than true signal. In high-dimensional settings, SHAP values shift with the reference distribution, and LIME approximations destabilize under small perturbations. The clinical risk is that a convincing explanation might validate a fundamentally unreliable model.

4.2 Algorithmic bias and health equity

When models train predominantly on one demographic, predictions fail for others—and XAI faithfully explains those biased outputs. Fixing the explanation without addressing data bias achieves nothing. Obermeyer *et al.* demonstrated this empirically when a health algorithm assigned lower risk scores to Black patients by optimizing for healthcare expenditure, a proxy confounded by access inequities (4). Nazer *et al.* catalogue mitigation strategies (18), while the STANDING Together consensus mandates proactive performance evaluation across population groups (19).

4.3 Scalability

Computing SHAP attributions for whole-genome sequences or 3D medical imagery incurs latency often incompatible with time-sensitive workflows. Benchmarking studies frequently underrepresent this constraint by utilizing curated, lower-dimensional datasets. The gap between benchmark performance and real-world utility remains a recurring theme in clinical AI.

5. Looking ahead: building domain-aware xai for biology and medicine

Future progress in XAI in biology mostly involves developing well-structured and robust methods rather than technical adaptation of existing methods. Pathway-based explanations is one way to achieve this, linking the model's output to familiar biological structures, rather than individual features. While it is technically feasible to incorporate knowledge graphs, their integration in routine clinical use has not yet been realized. Human-in-the-Loop (HITL) systems, which treat clinician feedback as a form of iterative training data, are another approach. Salloch and Eriksen describe this as co-reasoning, emphasizing active judgment over passive acceptance (20). Finally, there is a lack of standardized benchmarks for the quality of explanations. Recently, the FDA issued guidance indicating the move toward requiring explanation standards in addition to performance thresholds (2,3). Developing 'ground truth' explanations remain difficult but domestically necessary for regulatory acceptance.

Conclusion

The argument for XAI in clinical AI is sometimes framed primarily as a regulatory compliance requirement — which it increasingly is. But the deeper motivation is epistemic. Medicine has always required that clinical decisions be justifiable: to patients, to colleagues, to oversight bodies. AI that cannot participate in that justification process cannot be a genuine partner in clinical decision-making, regardless of its accuracy on benchmark datasets. The goal of domain-aware XAI is not to make models appear transparent. It is to make them genuinely accountable to the knowledge and values of the people they are meant to serve. By prioritizing biological faithfulness, equity-conscious evaluation, and regulatory alignment, XAI will be the bridge between computational performance and clinical utility — and ultimately between machine intelligence and human trust.

References

1. Abramson, J., Adler, J., Dunger, J., *et al.* (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*, 493–500.
2. U.S. Food and Drug Administration. (2024). *Transparency for machine learning-enabled medical devices: Guiding principles*.
3. U.S. Food and Drug Administration. (2025). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations*.
4. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
5. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS, 30*, 4768–4777.
6. Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting artificial intelligence models. *Brain Informatics, 11*(1), 10.
7. Usman, M., Varea, O., Radeva, P. I., *et al.* (2025). Explainable AI model reveals disease-related mechanisms in single-cell RNA-seq data. *arXiv preprint*.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" *KDD*, 1135–1144.
9. Pfeifer, B., *et al.* (2024). Stable feature selection utilizing GNN and LRP. *Artificial Intelligence in Medicine*.
10. Achtibat, R., *et al.* (2024). AttnLRP. *ICML*.
11. Koh, P. W., Nguyen, T., Tang, Y. S., *et al.* (2020). Concept bottleneck models. *ICML*.

12. Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2023). Explainable AI in drug discovery. *Machine Learning*.
13. Chereda, H., Bleckmann, A., Menck, K., *et al.* (2021). Explaining decisions of GNNs. *Genome Medicine, 13*, 42.
14. Partin, A., *et al.* (2023). Deep learning methods for drug response prediction in cancer. *Frontiers in Medicine, 10*, 1086097.
15. Lavecchia, A. (2025). Explainable AI in drug discovery. *WIREs Computational Molecular Science*.
16. Yao, R., Shen, Z., Xu, X., *et al.* (2024). Knowledge mapping of GNNs for drug discovery. *Frontiers in Pharmacology, 15*, 1393415.
17. Jin, W., Li, X., Fatehi, M., & Hamarneh, G. (2023). Guidelines and evaluation of clinical XAI. *Medical Image Analysis, 84*, 102684.
18. Nazer, L. H., *et al.* (2023). Bias in AI algorithms. *PLOS Digital Health, 2*, e0000278.
19. Alderman, J. E., *et al.* (2024). STANDING Together consensus recommendations. *The Lancet Digital Health, 7*(1), e64–e88.
20. Salloch, S., & Eriksen, A. (2024). Co-reasoning and practical judgment. *The American Journal of Bioethics, 24*(9), 67–78.