Asian Journal of Medical and Pharmaceutical Sciences
https://doi.org/10.30904/j.ajmps.2026.5026 | E-ISSN: 2348-0165

Explainable Artificial Intelligence (XAI) in Drug Discovery

Muneendra Madam*[1], Manjula Chella[2], M. Pradeep Kumar[3]

[1]*ScieGen Pharmaceuticals, 89 Arkay Drive, Hauppauge, New York 11788, USA*
[2]*Associate Professor, Department Pharmaceutical Technology, Vasavi Institute of Pharmaceutical Sciences, Vasavi Nagar, Peddapalli, Kadapa, A.P, India*
[3]*Principal and Professor, Vasavi Institute of Pharmaceutical Sciences, Vasavi Nagar, Peddapalli, Kadapa, A.P, India*
***Corresponding E-mail:** muneendrapharma2015@gmail.com*

Received: 28-03-2026 | Revised: 21-04- 2026 | Accepted: 25-05-2026 | Published: 20-06-2026

ABSTRACT
Artificial intelligence (AI) has transformed the drug discovery process, dramatically speeding up the identification of potential drug targets, the optimization of drug candidates, and the movement of compounds from the laboratory to the clinic. With the use of AI technologies, notably deep learning and machine learning methodologies, researchers have been able to analyze large datasets, find hidden biological patterns, and predict drug–target interactions with remarkable speed and precision. Yet, even with these advances, the intrinsic opacity of sophisticated AI models remains a fundamental barrier to broad use of AI in drug discovery. Specifically, deep learning systems tend to be "black boxes" that produce forecasts or recommendations but do not provide transparent explanations for how they arrived at those decisions. This lack of interpretability may impede scientific validation, regulatory approval and general confidence of researchers and doctors in AI-driven outcomes. This study gives a complete explanation of the main concepts of XAI, discusses important tools and methodologies now available and discusses a number of applications where XAI is being used to enhance results in drug development. It also discusses the challenges of implementing XAI, including technical, practical and regulatory hurdles, and considers possible future directions for research and development in this rapidly evolving area, with particular regard to its implications for pharmaceutical innovation.
Keywords: Artificial intelligence, explainable AI, drug discovery, target identification

Introduction

The completion of the Human Genome Project (HGP) marked a pivotal advancement in biomedical science by vastly expanding our understanding of the genetic underpinnings of disease. This milestone enabled the development of more precise and targeted therapeutic interventions and also generated significant economic benefits through innovation and biotechnology growth[1-4]. The drug discovery process has undergone major transformation since the 1980s. Initially characterized by simple chemical targeting of disease pathways, the field soon adopted structure-based drug design and, more recently, embraced high-throughput screening technologies that allow for the rapid evaluation of thousands of compounds. Modern drug development encompasses multiple critical phases:

Target Identification: Determining the molecular targets most relevant to a specific disease.

Hit Identification: Screening large libraries to find compounds that modulate the target.

Lead Optimization:
Refining active compounds to enhance efficacy, reduce toxicity, and improve pharmacokinetic properties.

Preclinical and Clinical Studies:
Conducting laboratory and animal testing, followed by phased human clinical trials to assess safety and effectiveness.

Regulatory Review & Post-Market Monitoring:
Gaining approval from health authorities and conducting ongoing safety surveillance once the drug is on the market.

The journey (Fig.1) from initial discovery to final approval is both lengthy and expensive, typically requiring **12–15 years** and substantial financial investment. Even after approval, continuous post-market monitoring is necessary to ensure long-term drug safety. The rapid expansion of biological data—driven by genomics, proteomics, and big data analytics—along with the increasing complexity of disease mechanisms, has elevated the importance of artificial intelligence (AI) in drug development. AI-driven approaches are now essential for efficiently analyzing large datasets, identifying novel therapeutic targets, and accelerating the pace of drug discovery.

Fig 1: Integration of AI and Explainable AI (XAI) Across the Drug Discovery Pipeline

Artificial Intelligence in Drug Discovery[5-8]

- **Artificial intelligence (AI)** encompasses a range of computational systems designed to simulate human

intelligence by leveraging advanced, data-driven models and interconnected networks. These systems are capable of learning from vast datasets, recognizing patterns, and making predictions or decisions with minimal human intervention.

- AI technologies have found diverse and impactful **applications** across biomedical research and healthcare, including:
  Disease prediction: AI models can analyze clinical and molecular data to forecast disease onset, progression, and patient outcomes.
  Genetic analysis: Machine learning tools assist in interpreting genomic data, identifying genetic variants linked to diseases, and understanding complex biological pathways.
  Drug discovery: AI accelerates the identification of novel drug candidates, predicts drug–target interactions, and assists in molecular design.
  Personalized medicine: By integrating patient-specific data, AI can help tailor therapeutic strategies to individual needs, improving efficacy and safety.
- **Neural Networks (NNs)**, including specialized architectures such as **Convolutional Neural Networks (CNNs)** for image analysis and **Recurrent Neural Networks (RNNs)** for time-series data, are at the forefront of AI applications in drug discovery. These networks excel at processing complex, high-dimensional biological data, such as medical imaging, genomics, and longitudinal patient records.
- AI-driven approaches enable the **efficient screening of enormous chemical spaces**, often exceeding 10^{60} potential small molecules. By rapidly evaluating and prioritizing compounds, AI significantly speeds up drug discovery, reduces experimental costs, and increases the likelihood of identifying effective therapeutics.
- Despite these advances, a **key challenge** remains: many state-of-the-art AI models, especially deep learning systems, function as "black boxes," providing little insight into their decision-making processes. This lack of interpretability poses barriers to building trust among researchers and clinicians, and it complicates **regulatory approval** for AI-assisted drug development.

The Role of XAI[9-12]

- **Explainable Artificial Intelligence (XAI)** is an emerging field dedicated to making AI systems' decision-making processes accessible, transparent, and understandable to humans. The primary goal of XAI is to demystify how complex algorithms arrive at specific predictions or recommendations, thereby fostering greater trust, accountability, and adoption among end-users.
- XAI is especially vital in **safety-critical domains** like healthcare and drug discovery, where clinicians, pharmaceutical researchers, and regulatory agencies must be able to comprehend and validate the logic behind AI-driven outcomes. Without explainability, these stakeholders may be reluctant to rely on or approve AI-generated insights for patient care or therapeutic development.
- The adoption of XAI methods offers several important benefits:
  Transparency & interpretability: XAI clarifies internal model processes, allowing users to understand how specific inputs lead to outputs.
  Trust & reliability: By making AI's reasoning visible, XAI increases user confidence in model recommendations, which is crucial for clinical and regulatory acceptance.
  Regulatory compliance: Regulatory bodies often require transparent evidence on how predictions are generated, making XAI essential for model approval and real-world deployment.
  Rational decision-making: By providing clear explanations for molecular design choices or property predictions, XAI enables more informed and rational decision-making in drug discovery.
- **Types of AI Models by Explainability[13-15]:**
  White-box models:
  These models are inherently interpretable and provide transparent reasoning (e.g., linear regression, decision trees, rule-based systems). They are easy to understand but may offer limited accuracy when dealing with intricate, high-dimensional data.
  Gray-box models:
  These aim to balance interpretability and predictive performance, providing some level of insight into their operations while handling more complexity than white-box models.
  Black-box models:
  These models (Fig 2), such as deep learning networks and transformers, achieve state-of-the-art accuracy but are notoriously difficult to interpret, making their inner workings opaque to users.

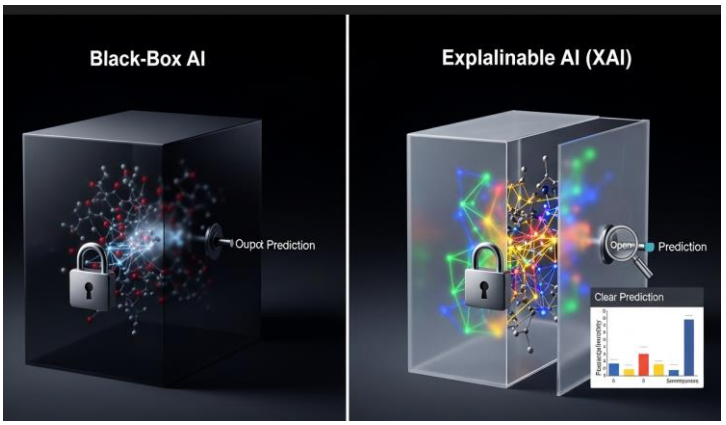

Fig 2: Comparison of Black-Box AI and Explainable AI (XAI) in Drug Discovery

Types of XAI Approaches:

Intrinsically interpretable models:

These are designed with transparency in mind from the outset, ensuring that the logic behind decisions is straightforward (examples include linear models, decision trees, and rule-based algorithms).

Post-hoc explainability:

These methods are applied after training more complex, less interpretable models to provide explanations for their predictions. Post-hoc approaches include:

Model-agnostic techniques:

Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which can be used with any model to interpret predictions.

Model-specific techniques:

Approaches tailored to particular architectures, such as attention mechanisms and saliency maps in neural networks, which highlight the most influential features or regions driving model outputs.

Key XAI Techniques

LIME (Local Interpretable Model-agnostic Explanations): LIME creates simple, interpretable models (such as linear regressions) around individual predictions made by complex models. By approximating the black-box model locally, LIME helps users understand which input features most strongly influenced a specific prediction, making the model's decision logic more transparent for individual cases.

SHAP (SHapley Additive exPlanations):

SHAP leverages principles from cooperative game theory to assign each input feature a quantitative value representing its contribution to the difference between the actual prediction and the average prediction. This method offers a consistent and theoretically sound approach to interpreting the impact of each variable on a model's outputs, both for single predictions and across datasets.

Partial Dependence Plots (PDPs):

PDPs visualize the relationship between a selected feature (or features) and the predicted outcome, while averaging out the effects of all other input variables. This technique helps reveal how changing one feature influences the model's predictions, making it easier to interpret complex models' behavior at a global level.

Attention Mechanisms:

Widely used in neural network architectures, attention mechanisms automatically learn to focus on the most relevant parts of the input data when generating predictions. In applications like natural language processing or molecular property prediction, attention maps can highlight which words, sequence elements, or molecular features were most influential in the model's decision-making.

Saliency Maps & Grad-CAM (Gradient-weighted Class Activation Mapping): These visualization tools are particularly valuable for interpreting deep learning models applied to images. Saliency maps identify which pixels or regions in an image had the greatest impact on a model's prediction, while Grad-CAM provides a heatmap showing the spatial areas most important for a specific class decision, offering intuitive insights into the model's focus.

Deep LIFT (Deep Learning Important Features) and Integrated Gradients:

Both are attribution methods for deep learning models. Deep LIFT compares the activation of each input feature to a reference value, assigning importance scores based on how much each feature contributed to the prediction. Integrated Gradients gradually interpolate from a baseline input to the actual input, accumulating gradients to determine feature importance. These techniques provide robust, theoretically grounded explanations for predictions made by complex neural networks.

Table 1: XAI Techniques and Their Features

Technique	Type	Purpose	Use Case in Drug Discovery
LIME	Post-hoc	Local explanations for individual predictions	Explaining a model's choice of drug candidate
SHAP	Post-hoc	Feature attribution, global and local explanations	Understanding variable impact on predictions
Attention Mechanism	Model-specific	Highlights important input regions/features	Molecular feature focus in design
Grad-CAM	Model-specific	Visualizes key image regions impacting decisions	Imaging-based drug discovery
Partial Dependence Plots	Post-hoc	Shows feature effect on prediction outcomes	Analyzing drug property changes

XAI in Healthcare Applications

XAI (table1) enhances the interpretability of AI for diagnosis, prognosis, and treatment planning.

Examples:

SHAP: Interpreting microbiome and EHR-based predictions.

LIME: Explaining Parkinson's and glioblastoma diagnoses.

Grad-CAM: Interpreting deep learning for imaging-based COVID-19 detection.

XAI tools help clinicians understand and trust AI-driven decisions, supporting adoption in clinical settings.

XAI in Drug Discovery Applications

XAI bridges AI-predicted outcomes with biological interpretation, crucial for:

- Target identification
- Molecular property prediction
- Toxicity assessment
- Drug–target and drug–drug interaction modeling
- Drug repositioning and combination therapy

Key platforms and tools:

- **Chemprop + SHAP**: Interpretable ADMET predictions.
- **GraphIX**: Explains biopharmaceutical network predictions.
- **AlphaFold 3**: Adds confidence scoring for protein structure predictions.
- **InstructMol**: Aligns natural language and chemical features for molecule design.

Challenges in XAI for Drug Discovery

- **Data limitations**: Many datasets are small, biased, or incomplete; synthetic data and transfer learning can help.
- **Complexity-Interpretability Tradeoff**: Balancing the accuracy of complex models with the need for transparency.
- **Ethical Concerns**: Preventing bias and ensuring fairness across demographic groups.
- **Regulatory Compliance**: Models must provide clear, scientifically relevant explanations.

Future Directions

- **Multimodal Data Integration**: Combining genomics, proteomics, and clinical data for better models.
- **Next-Gen XAI Frameworks**: Integrating graph neural networks and attention mechanisms for deeper biological insight.
- **Experimental Validation**: Using XAI to guide laboratory experiments and refine predictions.
- **Collaborative Platforms**: Federated learning projects (e.g., MELLODDY) for secure, large-scale data sharing and model training.
- **Ethical-by-Design Systems**: Embedding fairness, privacy, and accountability in XAI from the outset.
- XAI is essential for the responsible, effective deployment of AI in drug discovery.
- It enables transparency, trust, and actionable insights, fostering innovation and safer, more personalized therapies.
- Continued research and collaboration are required to address data, ethical, and regulatory challenges and to fully realize XAI's potential in pharmaceutical sciences.

Conflict of Interests

The authors declare no conflict of interest

Ethics Approval

Not applicable

Funding

This study received no specific funding from public, commercial, or not for profit funding agencies.

AI Tool Declaration

The authors declare that no AI and related tools are used to write the scientific content of this manuscript.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444.

[2] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018; 23(6): 1241–1250.

[3] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.

[4] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135–1144.

[5] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017;30:4765–4774.

[6] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825. 2017.

[7] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589.

[8] Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370–3388.

[9] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477.

[10] European Medicines Agency. Guideline on computerised systems and electronic data in clinical trials. EMA/226170/2021.

[11] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: The system causability scale (SCS). KI-Künstliche Intelligenz. 2020;34(2):193–198.

[12] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); 2018. p. 80–89.

[13] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793–4813.

[14] Kundu S, AI in medicine must be explainable. Nat Med. 2021;27(8):1328.

[15] Xie F, Lu L, Dong Z, et al. Explainable artificial intelligence for molecular property prediction in drug discovery. Brief Bioinform. 2022;23(1):1.