



Review

Explainable Artificial Intelligence: A Perspective on Drug Discovery

Yazdan Ahmad Qadri ¹, Sibhghatulla Shaikh ^{2,3}, Khurshid Ahmad ⁴, Inho Choi ^{2,3}, Sung Won Kim ¹
and Athansios V. Vasilakos ^{5,*}

¹ School of Computer Science and Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea; yazdan@yu.ac.kr (Y.A.Q.); swon@yu.ac.kr (S.W.K.)

² Department of Medical Biotechnology, Yeungnam University, Gyeongsan-si 38541, Republic of Korea; sibhghat.88@gmail.com (S.S.); inhochoi@ynu.ac.kr (I.C.)

³ Research Institute of Cell Culture, Yeungnam University, Gyeongsan-si 38541, Republic of Korea

⁴ Department of Health Informatics, College of Applied Medical Sciences, Qassim University, Buraydah 51452, Saudi Arabia; k.ahmad@qu.edu.sa

⁵ Department of Information and Communication Technology, University of Agder, 4879 Grimstad, Norway

* Correspondence: thanos.vasilakos@uia.no

Abstract

The convergence of artificial intelligence (AI) and drug discovery is accelerating the pace of therapeutic target identification, refining of drug candidates, and streamlining processes from laboratory research to clinical applications. Despite these promising advances, the inherent opacity of AI-driven models, especially deep-learning (DL) models, poses a significant “black-box” problem, limiting interpretability and acceptance within the pharmaceutical researchers. Explainable artificial intelligence (XAI) has emerged as a crucial solution for enhancing transparency, trust, and reliability by clarifying the decision-making mechanisms that underpin AI predictions. This review systematically investigates the principles and methodologies underpinning XAI, highlighting various XAI tools, models, and frameworks explicitly designed for drug-discovery tasks. XAI applications in healthcare are explored with an in-depth discussion on the potential role in accelerating the drug-discovery processes, such as molecular modeling, therapeutic target identification, Absorption, Distribution, Metabolism, and Excretion (ADME) prediction, clinical trial design, personalized medicine, and molecular property prediction. Furthermore, this article critically examines how XAI approaches effectively address the black-box nature of AI models, bridging the gap between computational predictions and practical pharmaceutical applications. Finally, we discuss the challenges in deploying XAI methodologies, focusing on critical research directions to improve transparency and interpretability in AI-driven drug discovery. This review emphasizes the importance of researchers staying current on evolving XAI technologies to realize their transformative potential in fully improving the efficiency, reliability, and clinical impact of drug-discovery pipelines.

Keywords: artificial intelligence; explainable artificial intelligence; drug discovery; molecular modeling; therapeutic innovation; personalized medicine



Academic Editor: Paolo Magni

Received: 24 June 2025

Revised: 5 August 2025

Accepted: 26 August 2025

Published: 27 August 2025

Citation: Qadri, Y.A.; Shaikh, S.; Ahmad, K.; Choi, I.; Kim, S.W.; Vasilakos, A.V. Explainable Artificial Intelligence: A Perspective on Drug Discovery. *Pharmaceutics* **2025**, *17*, 1119. <https://doi.org/10.3390/pharmaceutics17091119>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the Creative Commons Attribution (CC BY) license

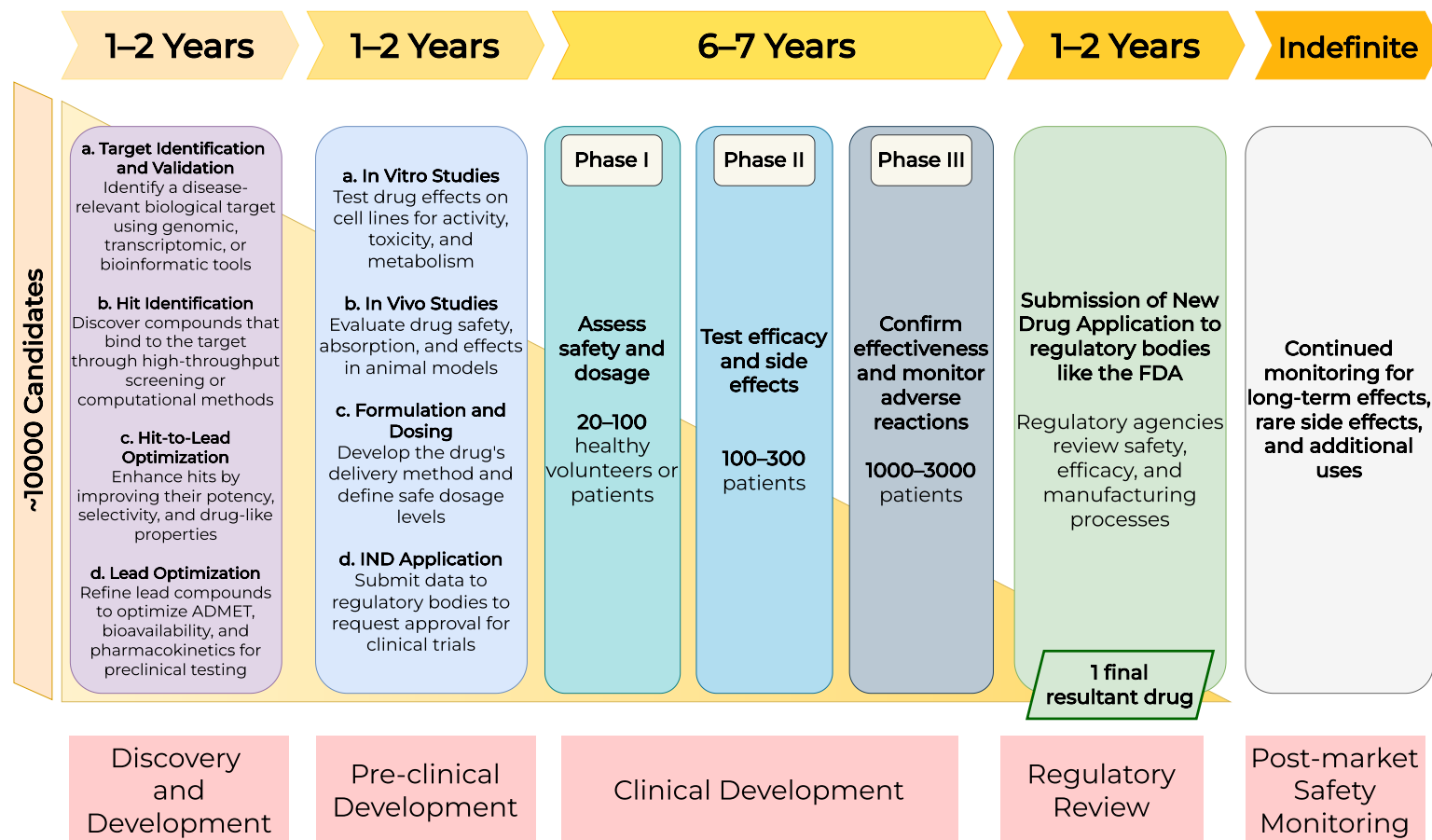
(<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Human Genome Project (HGP), completed in 2003, was a feat of human scientific ambition that took 13 years to understand the human genetic makeup. This process involved sequencing the human DNA to understand the genetic makeup [1]. HGP led

to new research offshoots, enabling the understanding of the genetic causes of diseases and possible interventions. Therefore, shedding light on the functional significance of the genes has led to an economic impact of \$800 billion. The emergence of novel diseases and persistent disorders has led to an increase in the use of pharmaceutical agents in daily life in recent years [2]. Furthermore, due to modern lifestyles that expose individuals to harmful pollutants and microorganisms, there is a pressing need for innovative and more reliable pharmacotherapeutic interventions. Consequently, rapid progress in drug discovery and development is unavoidable, and it is reasonable to expect the emergence of significant pharmaceutical solutions in a short time span. Presently, numerous research and development institutions, including government and commercial institutions, heavily rely on the expertise of pharmaceutical professionals for the development of these interventions.

The process of drug discovery evolved from the 1980s, when chemists began developing chemical compounds that specifically target distinct molecular entities such as receptors, enzymes, and ion channels. Structure-based drug development gained prominence in the 1990s, particularly in identifying lead compounds for drug development. Several technological developments, including high-throughput synthesis, genomics, structural biology, and computational chemistry, were brought into the drug-development process in the 2020s [3–6]. Given that the biological activity of a therapeutic molecule depends on its three-dimensional structure, medicinal chemistry plays a pivotal role in drug discovery. Thus, during the early stages of drug development, it is crucial to understand a drug's chemical properties through structure–activity relationship studies [7]. The lead optimization phase is a critical stage in the drug-discovery pipeline, wherein promising molecules identified during the hit-to-lead stage are systematically modified to improve their efficacy, selectivity, and drug-like properties. This process aims to enhance the therapeutic potential of the lead compounds while minimizing undesirable characteristics such as toxicity or poor bioavailability. During this phase, candidate compounds undergo a series of *in vitro* assays to assess their potency, physicochemical characteristics, and absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. Following this, preclinical *in vivo* studies are conducted to investigate the pharmacokinetic and pharmacodynamic properties of the selected molecules. Pharmacokinetics examines a drug's kinetics, which are primarily influenced by the body's ADME processes. In contrast, pharmacodynamics quantifies the drug's impact on the body, including various dynamics such as biomarker response, cytokine release, tumor progression, and other related factors [8]. Various physicochemical properties, such as molecular weight, lipophilicity, and permeability, influence the pharmacokinetic behavior of a drug [9]. Moreover, the drug's exposure and, consequently, its efficacy can be affected by the complex physiology of the body [10]. The data obtained during the research process is combined into a translational approach to predict a clinically appropriate and effective dosage and regimen that ensures safety [11,12]. Predicting clinical efficacy solely based on a compound's intrinsic properties or its behavior in preclinical *in vivo* studies can be challenging. However, a drug's ability to achieve a safe and effective exposure level is generally considered the primary determinant of its efficacy. The timeline and process of developing a new drug are illustrated in Figure 1. Typically, creating a new pharmaceutical drug takes about 12 to 15 years in the United States and requires continuous monitoring after its general rollout [13,14].



FDA: United States Food and Drug Administration
 IND: Investigational New Drug

ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity

Figure 1. Timeline of the conventional drug-discovery process. A typical drug undergoes five major phases. The process begins with target identification and validation, where disease-associated biological targets are identified and confirmed. This is followed by hit and lead discovery, in which compounds that interact with the target are identified and optimized for potency and selectivity. The preclinical phase involves in vitro and in vivo studies to assess the compound's safety, efficacy, pharmacokinetics, and pharmacodynamics. If successful, the drug enters clinical trials, conducted in three phases, each involving an increasing number of human participants to evaluate safety, dosage, and therapeutic efficacy. Finally, the post-marketing surveillance phase involves continuous monitoring of the drug's long-term safety and effectiveness in the broader population.

Artificial intelligence (AI) is a data-driven system that uses advanced tools and networks to mimic human intelligence [15]. The integration of AI in healthcare encompasses disease prediction and detection, genetic analysis and gene editing, drug discovery, radiography, and personalized medicine [16]. AI models demonstrate high accuracy and efficiency [17]. AI algorithms of varying complexity perform diverse functions at various levels of healthcare applications [18]. Neural networks (NN), such as convolutional neural networks (CNNs), have demonstrated a high degree of accuracy in biomedical image analysis [19]. In contrast, recurrent neural networks (RNNs) are adept at identifying anomalies in time-series biomedical data [20]. State-of-the-art large language models (LLMs) have revolutionized diagnosis, genomics, drug discovery, and personalized medicine [21,22]. Although these AI models yield highly accurate results, the basis for their reasoning is obscured by the highly complex mathematical processes that underpin these models. As of 2024, the United States Food and Drug Administration (FDA) had approved 950 artificial intelligence/machine-learning (AI/ML)-enabled devices for disease diagnosis [23]. The challenges in using AI for determining prognosis and developing treatment plans have slowed progress due to safety concerns. Therefore, clinical decisions must be founded on well-established principles. Although the conclusion is accurate, flawed reasoning is unacceptable, especially in safety-critical applications such as healthcare.

The vast chemical space, estimated to encompass over 10^{60} potential molecules, offers a rich foundation for the discovery of novel drug candidates [24]. However, screening through such a large candidate list using rudimentary methods can significantly impede the drug-development process, making it time-consuming and financially burdensome. However, using AI-based methods has the potential to overcome these limitations as illustrated in Figure 2 [25]. AI can identify hit and lead compounds, enabling faster drug–target validation and optimization of drug structure design [26]. Incorporating large datasets into AI models has the potential to reduce the risk associated with introducing a new molecular entity, eliminating the need for extensive experimentation. Researchers can achieve an automated and more efficient screening and selection strategy by incorporating in-silico AI models, which differ from a ‘trial-and-error’ approach that relies solely on expert intuition. This paradigm increases the number of screened compounds while decreasing the screening times. While various efforts have been reported for the early phases of the drug-development pipeline, such as target identification and hit finding, the potential relevance of these techniques in the later stages of the process remains unclear. The use of AI tools is thought to significantly reduce the experimental burden and timelines currently required for characterizing drug response in vitro and in vivo [27]. The foundation of the outcomes of these models is uncertain due to the “black-box” nature of the AI models. Explainable AI (XAI) bridges the gap between the outcomes of an AI model and the underlying reasoning behind those outcomes. XAI techniques can establish a foundation for trusting the reliability of models that assist in the drug design pipeline. XAI techniques address these challenges by identifying which molecular features or descriptors contribute most significantly to a given prediction, or by estimating the marginal contribution of each feature to the output, or highlighting specific substructures that are strongly associated with a predicted outcome. These insights enable researchers to rationally prioritize or modify molecular scaffolds, improve candidate selection, and enhance lead optimization. Moreover, XAI can potentially enhance regulatory compliance and build confidence in AI-driven pipelines by offering human-interpretable explanations for model predictions, such as poor absorption, high distribution volume, metabolic instability, or toxicity during the ADMET prediction. With the adoption of multi-modalities, from SMILES strings and molecular graphs to transcriptomics and imaging data, XAI provides a necessary layer of

transparency, enabling the deployment of AI not only as predictive tools but rather as a reliable decision support system.

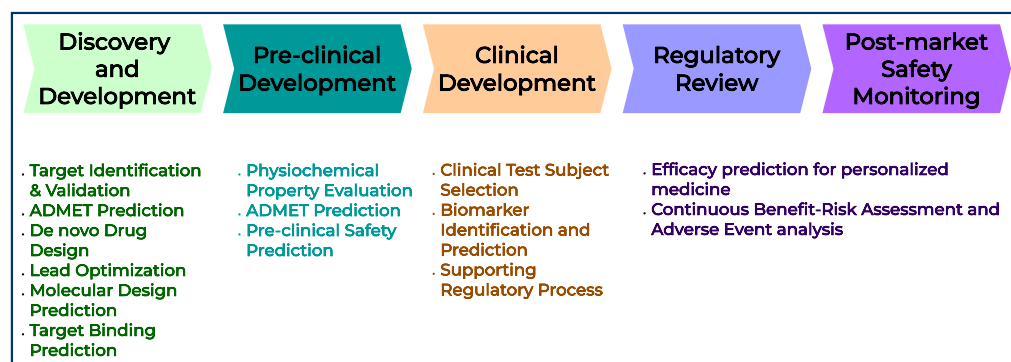


Figure 2. AI-supported drug-development pipeline. AI can potentially accelerate the progress of each stage, from target identification to post-market surveillance. This is achieved by enabling faster compound screening, predictive modeling, clinical trial optimization, and safety monitoring, thus improving efficiency and reducing development timelines.

The state-of-the-art AI models can potentially accelerate the drug-discovery process, particularly in the initial two stages. The published literature reviews have summarized the role of XAI in improving the drug-discovery pipeline. The two widely accepted explainability methods are the SHapley Additive exPlanations (SHAP) and the Local Interpretable Model-agnostic Explanations (LIME). The authors of [28] explore the role of SHAP and its variants in enhancing transparency in AI-driven drug-discovery processes. The authors outline the regulatory and practical importance of interpretability, emphasizing that explainability improves trust and reduces the downstream costs associated with opaque models. Their review outlines technical and regulatory challenges and future directions for XAI. Ding et. al. [29] systematically evaluates the literature on XAI applications in chemical and drug research, encompassing traditional Chinese medicine domains. However, this work predominantly relies on quantitative metrics without deep qualitative insights into the practical efficacy or impact of specific XAI techniques. In [30], authors deliver a structured taxonomy tailored specifically for medicinal chemistry, advocating for essential visualization and interactive methodologies. They outline clear guidelines for effectively integrating XAI into chemical research. The main limitation is that the recommendations primarily focus on structural visualization, rather than performance metrics or quantitative evaluations. Jiménez-Luna et al. [31] focus on the challenges associated with interpreting deep-learning (DL) models in drug discovery. The authors detail various feature attribution methods and gradient-based approaches to enhance the interpretability of models. They underscore that interpretability significantly impacts the practical application of DL, particularly when accuracy must be balanced with human comprehensibility and regulatory acceptance. Vo et al. [32] review XAI methodologies for predicting drug–drug interactions (DDIs). Given the clinical importance and high-risk nature of drug–drug interactions, the authors emphasize the necessity of transparent AI predictions to ensure reliability and clinical acceptance. It comprehensively categorizes ML/DL models, identifying gaps and limitations, and suggests pathways to strengthen model transparency and reliability. A comprehensive survey [33] covers various XAI frameworks and their applications, including target identification, compound design, and toxicity prediction. The authors identify key limitations, such as the interpretability versus performance tradeoff, and provide future research directions to guide the effective integration of XAI into drug-discovery processes. Although their work offers a clear understanding of XAI in drug discovery, their discussion

is presented from an AI-centric standpoint. Therefore, the existing literature for health science researchers is limited to brief reviews on this research domain.

To address the existing literature gap for biomedical science researchers, this comprehensive review examines the role of XAI in enhancing the interpretability and transparency of AI-driven drug-discovery methods. It summarizes the key XAI tools, models, and their applications in molecular modeling, target identification, molecular property prediction, clinical trial design, and personalized medicine. The review investigates how XAI addresses the opacity issues of traditional AI models, identifies current implementation challenges, and outlines key future research avenues for effectively incorporating XAI into pharmaceutical research. This article is organized as follows. Section 2 introduces XAI, its basic concepts, and its types. Section 3 elucidates the role of XAI in healthcare. Section 4 delves into the role of XAI in the drug-discovery process, while Section 5 discusses in detail the impact XAI has on the drug-discovery pipeline. The challenges and future research directions are outlined in Section 6. The discussion is concluded in Section 7.

2. Explainable AI

Explainable AI “explains” the output of an AI model. XAI constitutes a set of processes that explain the intent and reasoning for the output generated by an AI model. XAI elucidates the process and logical reasoning used by an AI model to arrive at a conclusion. Ensuring the accuracy along with safety in operation is crucial in critical applications such as autonomous vehicles, healthcare, and industrial Internet of Things (IIoT) [34]. Data-driven decision systems in critical applications should be both trustworthy and interpretable. Interpretability elucidates the inner workings and explains how an AI model makes a decision. While explainability takes into account all the interpretable factors that contribute to an AI model’s decision and allows the user to understand why the model made a particular decision [35]. The AI models can be classified into three categories based on their explainability: white-box, gray-box, and black-box models [34,36]. The white-box models are self-interpretable. Users can interpret the working logic of models, such as those in linear regression and decision trees. Still, there is a significant tradeoff in accuracy, as they assume the data to be linear or sub-linear, which is contrary to real-world data [37]. Additionally, self-interpretable models are not highly scalable and therefore do not meet the requirements for critical applications. The gray-box models aim to strike a balance between accuracy and interpretability. The gray-box models can support vital applications as they offer a level of interpretability by allowing analysis of the model’s inner workings and a higher accuracy [38]. However, the powerful AI models powering high-end applications are highly complex, making them difficult to interpret. Their ambiguous decision system makes them inappropriate for critical applications. However, for these black-box AI models with high obscurity, XAI tools can ensure trustworthiness [39]. The various types of AI models classified based on their interpretability are illustrated in Figure 3. XAI methods can be broadly classified into two groups, based on intrinsically interpretable models and post-hoc models [40]. The former category consists of models that are inherently easy to comprehend, while the latter requires a set of specialized methods to explain the model decisions [41]. The general outline of the XAI classification is illustrated in Figure 4. The following discussion follows the structure outlined in this figure.

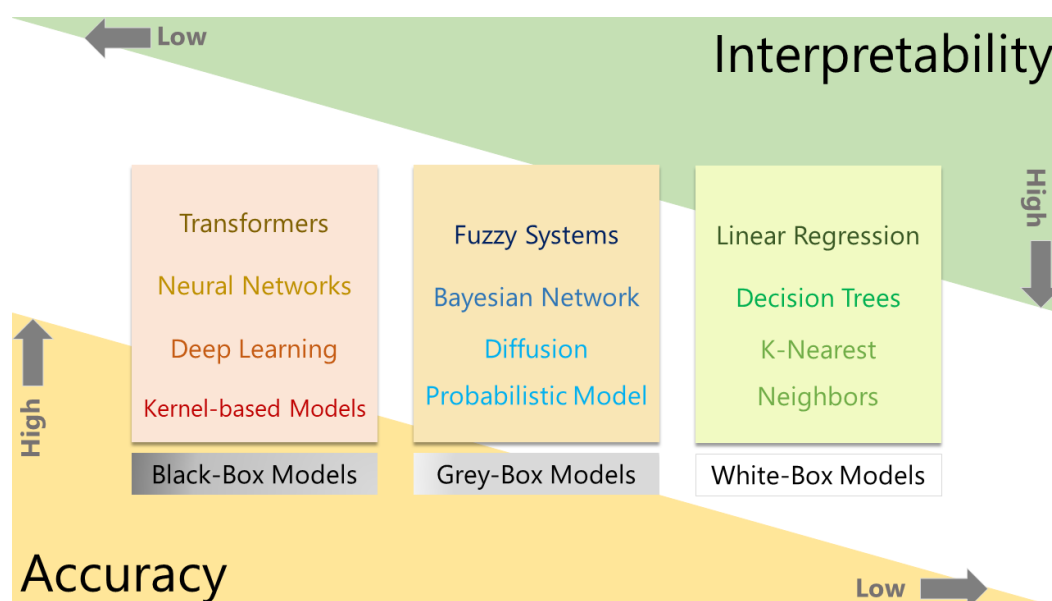


Figure 3. Types of AI models classified according to their interpretability. There is a tradeoff between the interpretability and accuracy of the AI model.

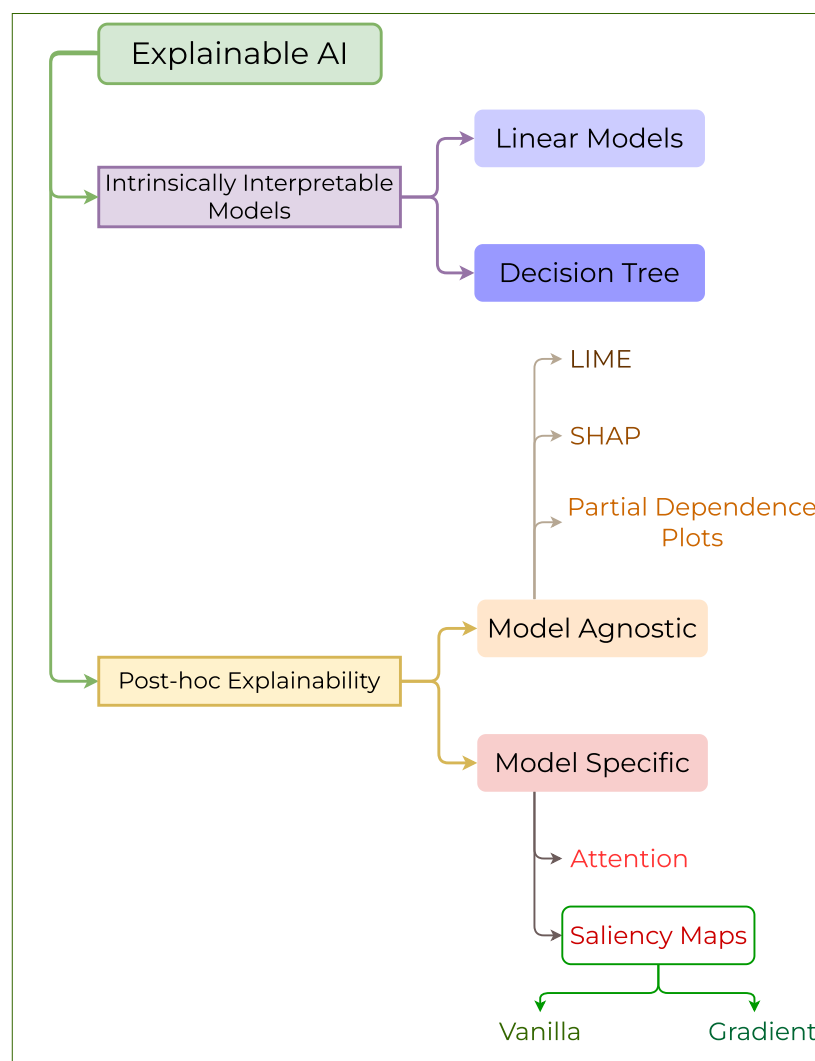


Figure 4. Outline of classification of XAI models.

2.1. Intrinsically Interpretable Models

Intrinsically interpretable models are designed in such a way that humans can readily understand their structure, parameters, and decision-making processes. They provide interpretations organically due to their structure, and there is no need for extra post-hoc approaches for interpretability [42]. These models are significant in critical domains as healthcare, finance, and law, where understanding the underlying causes for a certain decision is equally important to model accuracy. Inherently interpretable models can provide stakeholders with unparalleled insights into decision-making, enabling trust, transparency, and accountability to flourish. The following features define intrinsically interpretable models. (a) Simplicity, as they rely on uncomplicated mathematical models, including linear systems, rule sets, or tree ensembles. (b) Transparency, as each action or decision made by this model can be delineated and elucidated. (c) Feature importance clarifies the contribution of each feature to the final prediction. These models typically possess fewer parameters than black-box models, making them more comprehensible and interpretable [43]. The intrinsically interpretable models can be classified into the following categories:

2.1.1. Linear Models

Linear models represent the simplest type of intrinsically interpretable models, where the output is the result of a linear combination of the input features. These models assume a linear relationship between input features and target variables. Therefore, the impact of an input variable on the output is directly interpreted by its coefficient [44]. Linear Regression models calculate a continuous dependent variable from a set of predictor variables. The model, in turn, fits a linear equation to the empirical data. Each coefficient provides a clear interpretation of how a one-unit change in a feature influences the target variable, assuming all other features are held constant [45]. Logistic regression is a classification technique used for binary problems in a manner quite similar to linear regression. It generates a model to compute the probability that an instance belongs to a specific class. In the logistic regression model, this output is transformed through a sigmoid function that ranges between 0 and 1. In logistic regression, the coefficients signify the log-odds of a one-unit change in the respective feature. Although still not as intuitive as the coefficients from linear regression, directly interpreting the output with probabilities and odds ratios does it [46].

The linear models are simple, as they are directly interpretable and can be trained easily to perform accurately using a moderately sized dataset. The use of regularization can also help improve explainability in linear models. Regularization techniques help reduce overfitting by adding a penalty to the model's loss function, encouraging simpler weights [47]. This reduces the influence of less important input variables by moving the variable coefficients to near zero, improving model interpretability. Regularization also enhances generalization, making the model more robust to unseen data. However, it requires careful hyperparameter tuning to determine the optimal strength of the penalty. Additionally, regularization is particularly effective in high-dimensional data, where it helps mitigate the risk of overfitting due to the large number of input variables. On the other hand, regularization can also overly simplify models if the regularization parameter is too large, causing the model to miss important patterns. It also assumes equal importance for all input variables, which may not hold true, so incorporating domain knowledge can enhance performance. Regularization may be computationally expensive, especially for high-dimensional data, making it less suitable for real-time applications [48]. Additionally, it presumes that data are linearly separable and independent and identically distributed (IID), which may not apply to complex datasets like time-series or spatial data, requiring more advanced models [49].

Generalized Additive Models (GAMs) extend traditional linear models by allowing the relationship between each feature and the target variable to be modeled with smooth, nonlinear functions, while still maintaining an additive structure. Unlike linear models, which assume a straight-line relationship between features and the target, GAMs are more flexible, as they can capture complex, nonlinear relationships [50]. However, despite this flexibility, GAMs remain interpretable because each feature's contribution to the prediction remains independent of the contributions of the others. This structure enables the visualization of the effect of each feature, often in the form of individual plots for each constituent function. The advantages of GAMs lie in their balance of flexibility and interpretability [51]. They can model nonlinear patterns in the data without sacrificing transparency, as the additive nature of the model ensures that each feature's influence is clear and separable from the others. This makes GAMs particularly useful for applications where both accuracy and interpretability are essential. However, a key limitation of GAMs is that they are restricted to additive relationships between features and the target, meaning they cannot model interactions between features. Consequently, while GAMs are powerful in capturing individual feature effects, they may be less suitable for datasets where feature interactions play a critical role [52].

2.1.2. Decision Tree

A decision tree is a hierarchical ML model that partitions data into subsets based on feature values, using a series of if-then rules to make decisions [53]. The model is structured with three primary components: the root node, representing the initial feature used for data splitting; internal nodes, indicating subsequent decision points; and leaf nodes, which provide the final prediction or classification. This structure allows for high interpretability, as each path from the root to a leaf can be easily understood as a sequence of decision rules. One of the key advantages of decision trees is their transparency and ease of interpretation. Additionally, they do not require feature scaling, allowing them to work effectively with unprocessed data [54]. Furthermore, decision trees are capable of modeling complex, nonlinear relationships, making them versatile in capturing a wide range of diverse patterns. However, these models have limitations, such as their susceptibility to overfitting, particularly when the trees are deep, which can lead to the capture of noise rather than meaningful patterns in the data. Decision trees are also known for their instability, as small changes in the input data can result in significant alterations to the tree structure, making them sensitive to data variations [55].

Similar to a decision tree, rule-based association is an ML method that identifies relationships or patterns between variables in large datasets through the use of if-then rules. The explicit association rules make them particularly useful in applications such as market basket analysis, where relationships between items can be easily extracted and interpreted. Rule-based models are inherently interpretable, as users can assess the relevance and validity of each rule and modify them, if necessary, based on domain knowledge [56]. This interpretability is crucial for applications in sensitive fields, such as healthcare and finance, where understanding the rationale behind decisions is essential for ensuring fairness, accountability, and trustworthiness in AI systems.

In healthcare, decision trees and rule-based models are employed for medical diagnoses due to their transparent decision-making processes, such as diagnosing diseases based on symptoms and test results. In finance, linear models such as logistic regression are used for credit scoring to predict default risk, while rule-based models aid in fraud detection. In the legal domain, rule-based models and decision trees are utilized in risk assessments, including determining parole eligibility and predicting recidivism. These models' interpretability makes them valuable in fields where transparency is crucial.

2.2. Post-Hoc Explainability

Models that are not inherently interpretable require additional tools to enable a human to understand them. Post hoc methods, applied after a model is trained, aim to explain the decisions of a complex, already trained model. These methods are often indispensable when using black-box models, as the internal working mechanisms of the model are too complex to be understood without additional aids. Examples of post-hoc interpretative methods include backpropagation-based methods within neural networks, which quantify the features at the input, and model-agnostic approaches such as LIME and SHAP. These provide approximate decision processes for black-box models and, in this manner, may be used to provide insight into the interpretation of different inputs about predictive outputs [57]. While these post hoc interpretations are indispensable in the study of black-box models, they suffer from issues of complexity and domain specificity. In addition to identifying essential input features, these methods can often only roughly approximate the kind of relationships that exist between features and outputs in many domains, excluding image and text analysis [41].

2.2.1. Model-Agnostic XAI

These techniques are used to explain the output of machine-learning models, without regard for the underlying AI model. They are model-agnostic because they do not depend on the architecture or inner workings of the model, which allows them to be applied to any machine-learning model. The primary motive behind model-agnostic methods is to provide interpretability to predictions or insights into how the model arrives at a decision [58]. Since the complexity of the underlying AI can vary, a surrogate model that is inherently interpretable can be used to explain the model's decisions. A surrogate model is an interpretable model, such as a decision tree or linear model, used to approximate the predictions of a more complex "black-box" model. By analyzing the surrogate, insights can be gained into how the black-box model makes decisions. The advantage of this approach is that it provides a global understanding of the complex model's behavior. However, it may not fully capture the behavior of the original model, particularly in highly nonlinear problems. Surrogate models are often used to obtain high-level explanations of complex models, such as in DL applications in healthcare.

LIME

LIME is a widely used method that provides interpretability for individual predictions of any ML model, irrespective of its complexity. LIME is particularly useful for complex, black-box models such as deep neural networks and ensemble methods, which often produce accurate predictions but are difficult to interpret. The core idea behind LIME is to generate a local surrogate model, typically a simpler and more interpretable model like linear regression or decision trees, to approximate the behavior of the black-box model in the neighborhood of a specific instance. This surrogate model allows for detailed local explanations, making the black-box model's predictions more transparent on a case-by-case basis [59]. To explain an individual prediction, LIME first samples data points around the instance in question by perturbing the input features and generating new examples similar to the original instance. Perturbation involves modifying the input data slightly to see how the black-box model's predictions change. By observing these changes, LIME can understand how sensitive the model's prediction is to individual features. It then feeds these perturbed instances into the black-box model to obtain corresponding predictions. LIME assigns more weight to perturbed instances that are closer to the original instance and fits a simple interpretable model to this weighted data, approximating the black-box model's decision-making process in that local region. The surrogate model thus provides

insights into the contribution of each feature toward the prediction, allowing users to understand which features were most influential in the model's decision for that particular instance.

LIME's primary advantage is its model-agnostic nature, meaning it can be applied to any machine-learning model, regardless of complexity. This makes it particularly effective in explaining nonlinear models [60]. It also provides detailed explanations for individual predictions, which is crucial in high-stakes decision-making areas such as healthcare, finance, and legal systems. However, LIME is not without limitations. One notable drawback is its instability; small changes in data can lead to different explanations for similar instances, especially when the black-box model's decision boundary is highly nonlinear. Furthermore, the local explanations provided by LIME may not generalize well to the entire model, limiting the scope of the explanations to the neighborhood around the instance being explained. LIME can also be computationally intensive as it requires generating numerous perturbed samples and running predictions for each, which can be challenging for large datasets or complex models.

Despite these limitations, LIME has demonstrated broad applicability in various fields. In credit scoring, for instance, LIME can explain why a loan application was approved or denied, providing users with detailed reasons based on their financial features. Similarly, in healthcare, LIME has been used to explain diagnostic models, helping clinicians understand which patient features contributed most to a given prediction. LIME has also been applied to image classification tasks, where it can highlight the specific parts of an image that were most influential in the model's decision. In summary, LIME offers a flexible, interpretable solution for understanding the behavior of complex machine-learning models at the local level, although it is essential to consider its limitations in terms of stability and generalizability.

Shapley Additive Explanations

SHAP is a method rooted in cooperative game theory that provides a comprehensive and theoretically sound framework for explaining individual predictions by quantifying the contribution of each feature. It assigns each feature an importance value, known as the SHAP value, which represents the feature's contribution to the model's prediction. SHAP values offer a unified measure of feature importance by considering the contribution of each feature in the context of all possible feature subsets. This ensures that the assigned SHAP values accurately reflect each feature's role in the prediction [61].

The primary advantage of SHAP is its solid theoretical foundation, which guarantees consistency in attributing feature importance at local (individual prediction) and global (overall model behavior) levels. SHAP values ensure additivity, meaning that the sum of all feature contributions equals the model's prediction, which provides a clear and interpretable breakdown of how each feature influences the outcome. However, the method's robustness comes with the disadvantage of high computational cost, particularly when applied to large and complex models, due to the need to compute contributions across all possible feature subsets.

SHAP operates by calculating the contribution of each feature through the lens of Shapley values, a concept from cooperative game theory. Shapley values represent a fair allocation method, originally designed to distribute payouts among players based on their contribution to a coalition's total value. In machine learning, each feature is considered a "player," and the model's prediction is the total "payout." SHAP values ensure that each feature's contribution is fairly evaluated by averaging the marginal contribution of the feature across all possible combinations of features. This approach provides a comprehensive and equitable distribution of the prediction value among the input features.

SHAP delivers a robust and interpretable method for understanding the behavior of complex models, ensuring fairness and transparency in decision-making processes [62].

Partial Dependence Plots

Partial Dependence Plots (PDPs) illustrate the marginal effect of one or two features on a model's prediction by varying the target feature(s) while keeping other features constant. This approach helps visualize how changes in a specific feature impact the predicted outcome, providing an intuitive and straightforward means of interpreting feature influence. However, PDPs assume feature independence, which may not hold in real-world scenarios where features are often correlated. This can lead to misleading interpretations, particularly in complex models. However, Accumulated Local Effects (ALE) plots offer an alternative to PDPs by addressing the limitations related to feature dependencies. ALE plots estimate the local effect of a feature on the model's predictions by computing the changes in the prediction within small intervals of the target feature and then accumulating these effects over the feature's range [43].

2.2.2. Model-Specific XAI

Model-specific XAI methods are tailored to leverage the internal structures and characteristics of specific types of machine-learning models to provide detailed and context-sensitive interpretations. These methods are inherently linked to the particular architecture or operational principles of the models they are designed for, enabling a deeper and more precise understanding of model behavior than generic, model-agnostic approaches [63]. This category of XAI techniques is particularly relevant for complex models such as deep neural networks, where interpretability is crucial for understanding the decision-making process, building trust, and ensuring compliance in critical applications like healthcare, finance, and autonomous systems [64]. Some of the major model-specific XAI methods are described as follows.

Attention

Attention mechanisms were developed to address the challenges traditional neural network architectures posed in processing long data sequences, particularly within natural language processing (NLP) tasks [65]. Initially introduced for machine translation, attention mechanisms enable models to focus selectively on different parts of the input sequence when generating each segment of the output. This approach effectively mitigates the limitations of RNNs, which rely on fixed-size context vectors that struggle with long-term dependencies. The core functionality of attention mechanisms lies in the computation of attention weights, which quantify the relevance of each input element to a specific output element. In sequence-to-sequence models, these weights are derived by measuring the similarity between the current state of the decoder and each state of the encoder. The resulting attention weights are then used to generate a weighted sum of the encoder states, forming a context vector that guides the model's current prediction [37]. Various forms of attention mechanisms exist, including self-attention, where each element in the input sequence attends to all others, a method particularly effective in models like Transformers that have set new benchmarks in NLP. Soft attention assigns differentiable weights to all input elements, while hard attention selects a single element in a non-differentiable manner, often requiring reinforcement learning for optimization. Attention mechanisms enhance interpretability by highlighting which parts of the input data are most influential in the model's predictions, typically visualized through attention heatmaps. This transparency facilitates a clearer understanding of the decision-making process. Consequently, attention mechanisms are widely utilized in NLP applications such as machine translation, text summarization, and question answering. They are also employed in computer vision tasks,

where they enable models to concentrate on specific regions of an image, thereby improving performance in tasks such as object detection and image captioning.

Saliency Maps

Saliency maps are a widely used visualization technique designed to interpret the decision-making process of DL models, particularly CNNs. They identify and highlight the regions of an input, such as specific pixels in an image, that are most influential in driving the model's predictions. The theoretical basis for saliency maps lies in the observation that the gradient of a model's output with respect to its input features can indicate the sensitivity of the prediction to changes in those features. In essence, these gradients can reveal which parts of the input the model considers most significant when making a prediction [37].

Generating a saliency map involves computing the gradient of the model's output score for a specific class with respect to each input pixel. This gradient is then visualized as a heatmap, where the magnitude of the gradient at each pixel denotes its importance to the prediction. A higher gradient value suggests that minor alterations in that pixel would lead to a substantial change in the model's output, thereby identifying the critical regions of the input that the model relies on. This method provides an intuitive understanding of the model's focus and decision-making process, particularly in complex image recognition tasks. Several variations of saliency maps offer different perspectives on feature importance.

1. **Vanilla Saliency Maps:** These use the absolute value of the gradient of the output class with respect to each input pixel, providing a basic visualization of feature relevance. Guided Backpropagation enhances this approach by allowing only the gradients that positively influence the target class to flow back, thus filtering out irrelevant information and offering a more refined view of feature importance. Integrated Gradients further refine the attribution process by calculating the cumulative gradient as the input transitions from a baseline to the actual input, resulting in a more stable and comprehensive measure of feature contribution. Gradient saliency methods constitute a category of XAI techniques that utilize the gradients of a model's output with respect to its input features to determine the contribution of each feature to the model's predictions. These methods are grounded in the principle that the gradient of the output with respect to the input can indicate how sensitive the model's prediction is to small changes in the input variables. By analyzing these gradients, one can infer which features are most influential in driving the model's decisions [37]. The operational process of gradient saliency methods involves computing the derivative of the model's output with respect to each input feature, resulting in a gradient vector. This vector captures the direction and magnitude of change in the prediction for infinitesimal variations in each feature. The gradients are then used to generate visualizations or attribution scores that highlight the relative importance of the input features. There are several notable gradient-based attribution techniques, each tailored to provide unique insights into model behavior:
2. **Gradient Saliency Maps:** These use the raw gradients to generate a visual representation of feature importance. The saliency map indicates which input features, such as pixels in an image or words in a text, have the most significant impact on the model's prediction. This visualization allows for a straightforward interpretation of the model's focus and decision-making process.
 - **Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM):** CAM and Grad-CAM extend the concept of saliency maps by integrating class-specific gradient information with spatial feature maps from convolutional layers. CAM works by leveraging the linear relationship between convolutional feature maps and the output layer in CNNs with global

average pooling (GAP). Specifically, it computes the weighted sum of the feature maps in the last convolutional layer using the weights from the output layer corresponding to a particular class [66]. This yields a coarse localization map that indicates the most discriminative regions used by the model for a given prediction. Grad-CAM computes the gradient of the class score with respect to the feature maps of a target convolutional layer, and then performs a GAP on these gradients to obtain importance weights for each feature map. Grad-CAM utilizes the gradients of any target concept flowing into the final convolutional layer to produce a localization map, making it compatible with a variety of CNN-based models without architectural modifications [67]. By combining the spatial awareness of CNNs with gradient information, Grad-CAM provides a more interpretable and class-discriminative visualization, which is particularly valuable for complex image-based models [37].

- **Deep-Learning Important Features (DeepLIFT):** DeepLIFT assigns contribution scores to each input feature by comparing the network’s output to a baseline or reference output. Unlike simple gradient methods, DeepLIFT propagates these differences backward through the network, providing a more stable and interpretable measure of feature importance. This approach addresses some limitations of gradient-based methods, such as zero gradients in saturated regions of activation functions, thereby offering a more comprehensive view of feature contributions [37].

The primary advantage of gradient-based attribution methods is their ability to provide both local, instance-specific, and global, model-wide interpretability, making them versatile tools for understanding complex models. Gradient-based methods have broad applicability across various domains. In computer vision, they are employed to visualize feature importance in image classification, object detection, and segmentation tasks, providing insights into which parts of an image contribute most to the model’s predictions. In NLP, they help identify the significance of individual words or phrases in tasks such as text classification and sentiment analysis, facilitating a deeper understanding of how models process linguistic information. In the health-care sector, gradient-based methods are employed to evaluate the impact of clinical variables on model predictions, facilitating medical diagnosis and prognosis by identifying the factors that most significantly influence the model’s decision-making process. Overall, gradient saliency methods are powerful tools for elucidating the inner workings of complex machine-learning models, offering interpretable explanations that can enhance trust, transparency, and accountability in high-stakes applications.

We summarize the key features of post-hoc methods discussed in this section in Table 1.

Table 1. Comparison of the key post-hoc XAI techniques used in drug discovery.

Technique	Basic Working Principle	Input Type	Requirements
SHAP	Uses cooperative game theory. Assigns each feature an importance value for a prediction	Tabular, molecular descriptors, genomic data	High

Table 1. Cont.

Technique	Basic Working Principle	Input Type	Requirements
LIME	Perturbs input locally and fits a simple interpretable model to approximate the prediction	Tabular, image, text	Moderate
Partial Dependence Plots	Shows average predicted outcome as a function of one or two features, marginalizing others	Tabular	Low to moderate
Attention	Allocates weights to input elements, indicating their contribution	Sequences like SMILES, molecular graphs	Moderate
Saliency Maps	Computes gradients of the output with respect to input features	Image, 2D/3D molecular structures	Moderate
Gradient Saliency	Measures the sensitivity of output to small perturbations in input by computing gradients	Text, image, sequence	Moderate

3. XAI in Healthcare

On average, global healthcare expenses per capita are increasing due to longer life expectancy. Thus, it increases the burden on those suffering from chronic diseases. Therefore, questions about the long-term viability of current healthcare systems are growing. AI has the potential to help address these issues by improving care quality and cost effectiveness [68]. However, because of the potential fatal consequences of inaccurate predictions by an AI model, these models must be transparent and explainable. Clinicians must understand the AI decision-making processes to develop trust and enable adoption. Thus, healthcare decision-makers must be reliable, accurate, and transparent in their actions. To overcome this challenge, research efforts are ongoing to make ML and DL models interpretable [69]. AI systems should provide clinicians with explicit explanations of their results, such as highlighting crucial aspects that influence diagnostic decisions in disease identification [70].

To elucidate the link between microbial communities and phenotypes, the SHAP method was used, which interprets model predictions depending on the contribution of each feature [71]. Positive SHAP values suggest characteristics that support the projected outcome. Dopaminergic imaging modalities, such as SPECT DaTscan, have been investigated for early diagnosis of Parkinson's disease [72], with the LIME algorithm used to classify cases and provide interpretable explanations. XAI has also been used to diagnose acute critical illnesses. An early warning score system uses SHAP to explain predictions based on Electronic Health Record (EHR) data [73]. Furthermore, XAI approaches have been investigated in Glioblastoma diagnosis, with models using fluid-attenuation inversion recovery data validated for multiform classification and LIME used to assess local feature significance in test samples [70].

An explainable computer-assisted approach for lung cancer diagnosis has been presented, which uses the LIME method to generate natural language explanations from important features [74]. An ensemble clustering-based XAI model for traumatic brain injury diagnosis improved interpretability by combining expert knowledge and automated analysis [75]. COVID-NET, a model for COVID-19 detection using chest X-rays, obtained 93.3% accuracy and 91.1% sensitivity after interpreting its data using GSInquire, which

audits the network's internal decision-making by identifying the most influential internal features and mapping them to specific regions in chest X-ray images. This ensures the model bases its COVID-19 predictions on clinically relevant patterns rather than spurious correlations or artifacts [76]. Additionally, an interpretable ML model has been constructed to predict post-stroke hospital discharge disposition [77].

XAI-enabled classification models for COVID-19 have been presented to produce accurate predictions and credible explanations [78]. The model utilizes 380 positive and 424 negative CT volumes, aiding radiologists in localizing lesions and enhancing diagnostic insight. Early detection of sepsis is crucial, as delays can lead to irreparable organ damage and higher mortality, which is addressed by analyzing health information from the Cardiology Challenge 2019 [79]. An XAI model based on 168 hourly characteristics was developed, utilizing a gradient boosting model (XGBoost) with K-fold cross-validation to predict sepsis risk and provide interpretable results in the ICU setting. A study used brain MRI scans from 1901 participants from the IXI, ADNI, and AIBL datasets to classify Alzheimer's Disease by training a model on chronological and brain age data [80]. This model outperformed the existing ML approaches, with 88% accuracy for females and 92% for males. It can support both regression and classification tasks while preserving the morphological semantics of the input space and assigning feature scores to quantify the contribution of each region to the final result. Table 2 tabulates the XAI tools used in recent proposals to understand the outcomes of AI-based disease detection networks.

Table 2. Summary of XAI models used for healthcare applications.

XAI Tool	Modality	Applications	Reference
CAM	Bone X-ray	The model was intended to estimate knee damage severity and pain level based on X-ray images.	[81]
CAM	Lung Ultrasound and X-ray	The model uses three types of lung ultrasound images and VGG-16 and VGGCAM networks to classify three pneumonia subtypes.	[82]
CAM	Breast X-ray	A globally aware multiple instance classifier (GMIC) was proposed, which uses CAM to find the most informative regions by combining local and global data.	[83]
CAM	Lung CT	It trains the DRE-Net model on data from both healthy and COVID-19 patients.	[84]
Grad-CAM	Lung CT	A deep feature fusion method was proposed, with higher performance compared to a single CNN.	[85]
Grad-CAM	Chest Ultrasound	A semi-supervised model integrating an attention mechanism and disentanglement was proposed, with Grad-CAM used to improve explainability.	[86]
Grad-CAM	Colonoscopy	It uses DenseNet121 to predict the presence of ulcerative colitis in patients.	[87]
Grad-CAM	Chest CT	A neighboring-aware graph neural network was suggested for COVID-19 detection based on chest CT images.	[88]

Table 2. Cont.

XAI Tool	Modality	Applications	Reference
Grad-CAM and LIME	Lung X-ray and CT	The study examines five deep-learning models and uses a visualization technique to interpret NASNetLarge.	[89]
Attention	Breast X-ray	The study uses the A ³ Net model with triple-attention learning to diagnose 14 chest illnesses.	[90]
SHAP	EHR	It proposes a predicted length-of-stay strategy to solve imbalanced EHR datasets.	[91]
SHAP	Lung CT	It introduces a model for predicting mutations in individuals with non-small cell lung cancer.	[92]
LIME and SHAP	Chest X-ray	It provides a single pipeline to improve CNN explainability using several XAI approaches.	[93]

4. XAI in Drug Discovery

In biological systems, there are intricate layers of regulation. These layers encompass dynamic interactions among genes, proteins, signaling networks, and metabolic pathways. Therapeutic targeting and drug response prediction are challenging due to the inherent variability of diseases, particularly cancer, neurodegeneration, and metabolic disorders. Despite the strong predictive capabilities of AI and ML methodologies when working with large datasets, their opaque and black-box nature frequently limits the biological interpretability of their results. XAI is an essential tool that allows researchers to understand the reasoning behind a model's specific prediction by connecting these decisions to biologically significant variables. This interpretability is crucial for enhancing trust and reproducibility, as well as for developing new hypotheses based on mechanisms that will inform future phases of drug discovery.

Recent advancements in explainable and interpretable AI have markedly improved the reliability and acceptance of AI models in drug discovery and healthcare. Numerous newly established frameworks illustrate the effective integration of XAI principles into drug–target interaction (DTI) prediction and molecular property modeling. DeFuseDTI and DTRE utilize advanced DL architectures in conjunction with feature attribution methods to enhance the precision and interpretability of DTI predictions, thereby facilitating more informed therapeutic decisions. ARGENT further refines this methodology by integrating attention mechanisms and interpretable embeddings, enabling researchers to correlate model predictions with distinct biological or chemical characteristics. DCGAN-DTA utilizes generative adversarial networks to predict drug–target affinity, ensuring transparency via interpretable outputs. These models underscore the growing emphasis on integrating XAI into complex predictive systems, highlighting the importance of transparency, trust, and actionable insights in modern drug-development processes.

XAI Tools Enabling Interpretability in Drug Discovery

In recent years, there has been an increase in the number of XAI tools designed to elucidate the predictions generated by complex models in drug discovery. The following tools offer case-specific interpretability, including structure–activity modeling, toxicity prediction, and molecular property analysis. A plethora of XAI tools has significantly enhanced the interpretability of complex models in drug discovery. SHAP utilizes game theory to assess the contribution of individual input features and has been extensively applied in models such as random forests, support vector machines (SVMs), and deep neural networks to identify molecular substructures affecting compound activity in Quantitative Structure–Activity Relationship (QSAR) studies [61,94]. LIME creates simple surrogate

models tailored to specific predictions, enabling chemists to understand the crucial structural elements that influence a compound's expected activity or toxicity [57]. Combined with attention mechanisms, Graph Neural Networks (GNNs) effectively recognize critical atoms and bonds in molecular graphs, facilitating optimization guided by substructures. Integrated Gradients and DeepLIFT provide gradient-based attributions that are vital in omics-driven research, pinpointing genes or features that impact drug response classifications. Furthermore, Chemprop, a framework for predicting molecular properties, has been integrated with SHAP to clarify ADMET predictions by linking pharmacokinetic properties with specific atomic and structural features, thereby enabling informed lead optimization. Drug repositioning can be facilitated by identifying and elucidating biologically plausible compound-disease associations using GraphIX, which combines GNN with SHAP-like methods [95]. InstructMol is a multimodal model that employs natural language prompts and molecular structures to create new compounds [96]. It achieves this by ensuring that textual and chemical features align in an interpretable manner, enabling rationale-driven molecule design. AlphaFold 3 includes confidence scoring to identify uncertain areas in predicted protein structures [97]. This makes structural drug design more reliable. Furthermore, platforms like PandaOmics and ID4 utilize explainable analytics and visualization components to assist with target discovery, disease mechanisms, and lead prioritization, enhancing transparency in AI-driven pharmaceutical processes [98]. Table 3 lists the XAI tools used in the current drug-discovery processes.

Table 3. Summary of XAI models used in identifying interactions for the development of drugs.

Tool/Platform	Description	Applications in Drug Discovery	Reference
SHAP	A model-agnostic method that assigns each feature an importance value for a particular prediction	Interpreting ML predictions in QSAR and SAR studies, identifying key molecular features influencing compound activity, and increasing transparency in model-guided drug design	[61,94]
LIME	Explains the predictions of any classifier by approximating it locally with an interpretable model	Understanding model decisions in compound activity prediction and toxicity assessments	[60]
DeepLIFT	Attributes importance scores to each input feature by comparing the activation to a reference activation	Interpreting DL models in genomics and proteomics data analysis	[37]
Integrated Gradients	Assigns feature importance by integrating gradients of the model's output with respect to the inputs	Explaining deep neural networks in molecular property prediction	[99]
AlphaFold 3	Predicts protein structures and their interactions with high accuracy using AI	Accelerating target identification and understanding protein-ligand interactions.	[97]
GraphIX	A graph-based XAI framework for drug repositioning using biopharmaceutical networks	Identifying potential new uses for existing drugs by analyzing biological networks	[95]
InstructMol	Integrates molecular graph data and SMILES sequences with natural language by fine-tuning a pretrained LLM	Enhances the foundation for XAI in drug discovery by aligning molecular structures with natural language through instruction tuning	[96]
PandaOmics	An AI-driven platform for target discovery and biomarker identification	Discovering novel therapeutic targets and biomarkers in various diseases	[98]

5. Impact of XAI on Drug Discovery

The development of AI and ML has transformed drug research. Transparency and interpretability become increasingly crucial as the complexity of these models grows. XAI solves this issue by providing a better understanding of the predictions provided by ML algorithms.

5.1. Data Analysis

XAI algorithms facilitate the analysis of large and diverse datasets containing chemical, biological, and clinical information to find novel drug targets, predict medication efficacy and toxicity, and improve drug design [100]. Advanced computational approaches and ML algorithms are utilized in XAI drug discovery to process and evaluate large datasets from multiple sources, such as molecular structures, biochemical tests, high-throughput screening (HTS), and preclinical and clinical trials [33].

AI and ML models have shown promising outcomes in areas such as lead optimization, virtual screening, chemical design, and medication repurposing [101–104]. As these models evolve, they have the potential to significantly increase drug-discovery success rates while reducing time and costs. However, their predictive capacities frequently lack interpretability, making it difficult for academics, clinicians, and regulatory authorities to trust and validate the results. Without insights into model decision-making, it is difficult to evaluate and prioritize targets or compounds. XAI addresses this issue by providing clear explanations of model predictions [28], which increases trust, enables the detection of biases or inaccuracies, and facilitates a deeper understanding of model behavior [33].

5.2. Molecular Property Prediction

XAI can optimize lead compounds to enhance effectiveness, pharmacokinetics, and drug-like features, resulting in the development of more effective medications with fewer adverse effects [105]. XAI in drug development improves the transparency and accountability of AI models, which are critical for lead optimization and toxicity prediction [106]. This increases trust in AI-generated outcomes, encouraging their use in the pharmaceutical industry. XAI also identifies and mitigates biases, resulting in fair and accurate predictions, which are critical for avoiding the development of ineffective or harmful medications [31]. Various XAI investigations have focused on unraveling molecular substructures using the gathered data in drug discovery. The authors in [94] utilize SHAP to interpret key characteristics and substructures for predicting chemical activity. Jiménez-Luna et al. [107] also used integrated gradient attribution to highlight key chemical characteristics and structural aspects in graph neural network models.

5.3. Personalized Medicine

XAI algorithms help to analyze patient data and predict individual responses to treatments, allowing for the development of personalized and effective medications [108]. In drug research, XAI facilitates personalized medicine by utilizing AI to analyze large datasets for evidence-based decision-making, drug repurposing, and real-time monitoring [109]. XAI methodologies, such as SHAP, LIME, and attention mechanisms, help researchers understand the molecular or biological features that influence predictions, allowing them to correlate model outputs with domain expertise and refine compound design decisions.

5.4. Unraveling Drug–Drug and Drug–Target Interactions

Drug–drug interactions (DDIs) are common in polypharmacy, when the effects of one drug might influence the actions of another in a combined therapy regimen. Ideally, such interactions produce synergistic effects as well as therapeutic advantages. However, in

the treatment of multiple diseases, adverse drug events that result in toxicity or reduced efficacy may occur, thereby increasing patient morbidity and death [110,111]. The current growth in the approval of new medications and indications has increased the possibility of DDIs [112,113]. While wet-lab investigations to verify DDIs are time-consuming and resource-intensive, rendering them unsuitable for routine use, AI models have been used to predict DDIs better [114–116]. Efforts have been made to improve drug database models to aid clinical decision-making. Effective DDI management is critical for maintaining pharmacovigilance and patient safety. The application of XAI in predicting DDIs has recently been extensively reviewed elsewhere [116].

Biomedical experiments to investigate DTIs are resource-intensive. To reduce costs and time, ML algorithms have been used to predict these interactions. The abundance of drug and target data, advances in computing technology, and the distinct capabilities of multiple ML algorithms have made them the primary tools for predicting drug–target interactions. This prediction approach aids in screening out inappropriate compounds, which is an important stage in novel drug development [117]. Modeling cellular networks in cancer using AI provides a quantitative framework for investigating the association between network properties and disease, allowing the identification of potential new anticancer targets and drugs [118–120]. The use of XAI in identifying novel anticancer targets, the ideas underlying common algorithms, and its applications in biological investigation have recently been reviewed elsewhere [121].

5.5. Facilitating Drug Repositioning and Combination Therapy

Drug repositioning entails identifying new therapeutic applications for FDA-approved drugs. This strategy focuses on assessing the efficacy of existing drugs or those under development in various pathological conditions [122,123]. Since 1995, new drug approvals have been declining due to the traditional drug-development procedure, which is costly and time-consuming. Hence, drug repositioning has emerged as a potential alternative, using XAI to expedite drug discovery while lowering costs and risks [122]. The significant benefits of this strategy include knowledge of drug pharmacokinetics and toxicity, as well as the low cost of implementation, which benefits low- to middle-income nations where traditional therapies may be too expensive [124].

Drug repositioning strategies combine computational and experimental techniques to uncover new therapeutic applications for current drugs [125,126]. ML, network analysis, and NLP are three critical computing methodologies [127]. These methods are classified as disease-centric, drug-centric, or combinations of both [128]. Disease-centric techniques identify new applications for drugs by grouping diseases based on phenotypic commonalities, molecular markers, and genetic variants [129,130]. Drug-centric techniques seek similarities in molecular action between drugs to identify new potential applications [131]. Combination techniques integrate both strategies by creating drug–drug and disease–disease similarity networks, assigning drugs based on meta-path scores, and predicting disease–drug relationships by correlating disease expression patterns with genes affected by drugs [132,133].

5.6. Clinical Trial Design

XAI enhances clinical trial design by identifying appropriate patient demographics, predicting trial success, and detecting possible adverse effects. This enables a more accurate assessment of the safety and efficacy of novel medications in humans [134]. XAI can also aid in predictive modeling, patient selection, and safety precautions during drug development.

5.7. Ethics and Regulatory Implications

Lack of transparency in AI systems raises significant ethical concerns, particularly in healthcare and drug development, where decisions must be both interpretable and justifiable. Recent frameworks emphasize the importance of fairness, accountability, and human oversight in the deployment of AI. XAI contributes to these goals by exposing decision logic, identifying potential biases, and facilitating more transparent communication with regulatory bodies, clinicians, and interdisciplinary teams.

6. Key Challenges and Future Research Directions in XAI for Drug Discovery

6.1. Key Challenges

XAI is rapidly becoming a crucial element in AI-assisted drug discovery. It promotes informed decision-making in drug screening, biomarker identification, clinical trial design, and personalized medicine by enhancing model transparency, interpretability, and reliability. In clinical and biomedical settings, XAI enhances interpretability for healthcare practitioners, facilitates bias detection, improves patient communication, promotes ethical adherence, and ensures regulatory compliance. Nonetheless, despite its potential, several significant challenges must be addressed to harness XAI's capabilities in drug discovery to their maximum.

6.1.1. Data Limitations

To discover significant patterns, XAI models require large, high-quality datasets with diverse and varied sample spaces. However, many drug-discovery datasets are limited, incomplete, or biased, compromising model performance and interpretability. Innovative technologies, such as data augmentation, synthetic data production, and transfer learning, will be crucial in overcoming data scarcity and enhancing generalizability.

6.1.2. Complexity and Interpretability Tradeoff

Highly accurate models, particularly deep NNs, often operate as black boxes, offering little insight into their decision-making processes. In contrast, interpretable models may lack the predictive effectiveness necessary for complex biomedical applications. Striking a balance in this tradeoff presents a significant challenge. Developing hybrid XAI frameworks that combine predictive power with intuitive interpretability is a viable strategy and a challenge for widespread adoption.

6.1.3. Ethical and Bias Concerns

It is crucial to thoroughly assess the ethical implications of XAI models, particularly in terms of bias and fairness across different demographic groups. Predictions based on biased training data may exacerbate existing health disparities. The responsible use of AI in drug discovery requires rigorous validation procedures and models that are created with fairness in mind.

6.1.4. Regulatory Compliance

To be used in clinical or regulatory settings, XAI systems must provide clear, scientifically relevant explanations that align with expert knowledge. Regulatory bodies seek models that can be comprehended to evaluate their safety and reliability. The current application of XAI requires further development to produce understandable results that meet high safety and regulatory standards.

6.2. Future Research Directions

6.2.1. Multimodal Data Integration and Augmentation

Drug responses depend on multiple factors, including genetic variations, protein expression, metabolic pathways, and clinical phenotypes. Consequently, future research should prioritize the integration of multimodal data—including genomics, proteomics, transcriptomics, metabolomics, and real-world clinical data—to build comprehensive and context-aware models of drug action. Integrating gene expression profiles with chemical structure data significantly enhances the performance of drug sensitivity prediction models, while also improving biological plausibility and interpretability [135]. Other studies demonstrate how multimodal fusion not only enhances predictive performance but also provides mechanistic insights into drug action [136]. However, aligning heterogeneous data types remains a significant challenge due to differences in data scale, format, and biological context. To address this, research should also explore data augmentation strategies such as generative modeling, cross-modal embeddings, and transfer learning to enrich underrepresented data domains and improve generalization. By advancing data integration and augmentation methodologies, XAI frameworks can evolve into more resilient and biologically grounded systems, ultimately supporting safer and more personalized therapeutic development.

6.2.2. Next-Generation XAI Frameworks

The complex biochemical interactions and the effects of pharmaceuticals on various targets necessitate the development of innovative XAI models and frameworks. Future research should focus on models that integrate GNNs with attention-based architectures to model and interpret complex biochemical interactions accurately. GNNs are well-suited for representing molecular structures, while attention mechanisms can highlight the most dominant molecular substructures that contribute to biological activity, binding affinity, or toxicity [31,137,138]. By leveraging these techniques, XAI models can provide intuitive and interpretable explanations that align with pharmacological principles, offering clarity in identifying functional groups responsible for specific pharmacological effects and highlighting structural alerts linked to adverse outcomes. When combined with cross-modal attention between molecular and protein representations, these models could also clarify drug–target binding mechanisms. Incorporating domain knowledge, such as known reaction rules, toxicophore databases, or protein–ligand interaction motifs, further enhances the biological plausibility of the explanations. This direction enables the creation of transparent, mechanism-aware AI systems that not only predict outcomes but also generate actionable hypotheses, supporting critical decision-making in hit-to-lead optimization, multitarget drug design, and safety profiling.

6.2.3. Experimental Validation and Hybrid Models

The integration of XAI with experimental methodologies, including molecular dynamics simulations (MDS) and high-throughput screening, can facilitate the confirmation and enhancement of computational predictions. Research efforts should focus on integrating XAI with MDS and HTS to validate and refine predictions generated by the AI models. Attention maps and feature attributions used in [138] can be used to highlight critical substructures involved in drug–target interactions. These predictions can then be evaluated using MDS to test the stability and conformational dynamics of the predicted binding modes, thereby offering physicochemical validation of model outputs. Similarly, XAI-guided compound prioritization can inform HTS experiments by narrowing the chemical search space to biologically plausible candidates, enhancing hit rates and reducing false positives [137]. Experimental feedback from such validation efforts can be reintegrated into

training datasets to fine-tune model weights and improve generalizability, establishing a feedback loop between computation and experimentation. Furthermore, XAI can support drug repurposing by identifying alternative binding sites or off-target effects, which may then be verified through in vitro assays or biochemical profiling. This hybrid approach not only augments model performance but also advances the interpretability and scientific validity of AI-driven drug discovery, enabling the generation of testable hypotheses that are both biologically plausible and experimentally verifiable.

6.2.4. Collaborative Open Platforms

The MELLODDY project is a large-scale federated learning (FL) initiative in which several pharmaceutical companies collaboratively train advanced AI models without explicitly sharing their proprietary data. It leverages over 2.6 billion activity records from 21 million molecules across 40,000 assays. This enables improved predictive modeling for drug discovery while preserving data privacy and intellectual property rights. The MELLODDY project serves as a benchmark for collaborative ecosystems that facilitate the exchange of data, models, and tools, thereby expediting the development and validation of XAI frameworks. Open research platforms can enhance reproducibility, transparency, and regulatory compliance among stakeholders.

6.2.5. Ethical-by-Design Frameworks

Incorporating ethical considerations into the design of XAI systems is crucial for ensuring safety across diverse demographic groups. To ensure the ethical utilization of AI in healthcare and pharmaceutical development, it is imperative to integrate fairness constraints, safeguard data privacy, and promote stakeholder accountability. Fairness constraints are a primary consideration and critical in drug-discovery applications involving patient data or population-specific models, as algorithmic bias can lead to unequal access to treatment or inaccurate predictions across demographic subgroups. Racial bias in healthcare algorithms can significantly impact treatment prioritization, underscoring the need to incorporate fairness-aware modeling techniques into biomedical AI pipelines [139]. Similarly, safeguarding data privacy through methods such as FL can enable large-scale collaboration without compromising sensitive information. Moreover, the development of XAI systems must be coupled with mechanisms for stakeholder accountability, ensuring that domain experts, data custodians, and AI developers are collectively responsible for model decisions and their consequences. Future research must therefore prioritize the co-design of XAI systems with ethics experts, clinicians, and regulatory bodies to create frameworks that not only explain model behavior but also align with broader safety and ethical values. This ethical-by-design approach is foundational to building trustworthy AI systems that can be safely and equitably deployed in the pharmaceutical and healthcare sectors.

7. Conclusions

As AI transforms drug research, the incorporation of explainability has become a fundamental requirement rather than an ancillary attribute. XAI reconciles predicted accuracy with scientific confidence by providing openness, accountability, and biological interpretability in AI models. This study highlights the growing importance of XAI tools and frameworks, which clarify the reasoning behind complex predictions, allowing researchers to make more informed, ethical, and practical decisions when designing and developing novel therapies. The future of drug development relies on integrating advanced AI models with strong interpretability, robust ethical protections, and interdisciplinary collaboration. Confronting existing limitations, such as data integrity, model complexity, and regulatory

requirements, while adopting emerging technical breakthroughs, will ensure that AI technologies are both effective and trustworthy in clinical contexts. Progressing this research area necessitates a purposeful transition to next-generation XAI research emphasizing transparency, inclusion, and fairness. XAI can expedite drug-development timeframes, mitigate risks, and facilitate more tailored and accountable therapeutic approaches. The meticulous implementation of this approach will characterize the forthcoming epoch of pharmaceutical research, whereby data-driven discovery is both insightful and comprehensible.

Author Contributions: Conceptualization, Y.A.Q. and K.A.; Investigation, Y.A.Q. and S.S.; Writing—Original Draft Preparation, Y.A.Q. and S.S.; Writing—Review and Editing, K.A.; Funding Acquisition, A.V.V.; Supervision, I.C. and S.W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Norwegian Research Council under the grant name SecureIoT.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ADME	Absorption, Distribution, Metabolism, and Excretion
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
ALE	Accumulated Local Effects
CAM	Class Activation Mapping
CNN	Convolutional Neural Networks
DDI	Drug–Drug Interactions
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DTI	drug–target interaction
EHR	Electronic Health Records
FDA	Food and Drug Administration
FL	Federated learning
GAM	Generalized Additive Models
GAP	Global Average Pooling
GNN	Graph Neural Networks
HGP	Human Genome Project
HTS	High-Throughput Screening
ICU	Intensive Care Unit
IID	Independent and Identically Distributed
LIME	Local Interpretable Model-Agnostic Explanation
LLM	Large Language Model
MDS	Molecular Dynamics Simulations
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Networks
PDP	Partial Dependence Plots
QSAR	Quantitative Structure–Activity Relationship
RNN	Recurrent Neural Networks
SHAP	Shapley Additive exPlanations
SVM	Support vector Machine
XAI	Explainable Artificial Intelligence

References

- Hood, L.; Rowen, L. The Human Genome Project: Big science transforms biology and medicine. *Genome Med.* **2013**, *5*, 79. <https://doi.org/10.1186/gm483>.
- Meganck, R.M.; Baric, R.S. Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases. *Nat. Med.* **2021**, *27*, 401–410. <https://doi.org/10.1038/s41591-021-01282-0>.
- Lombardino, J.G.; Lowe, J.A. The role of the medicinal chemist in drug discovery—then and now. *Nat. Rev. Drug Discov.* **2004**, *3*, 853–862. <https://doi.org/10.1038/nrd1523>.
- Ali, S.; Ahmad, K.; Shaikh, S.; Chun, H.J.; Choi, I.; Lee, E.J. Mss51 protein inhibition serves as a novel target for type 2 diabetes: A molecular docking and simulation study. *J. Biomol. Struct. Dyn.* **2024**, *42*, 4862–4869. <https://doi.org/10.1080/07391102.2023.2223652>.
- Ali, S.; Ahmad, K.; Shaikh, S.; Lim, J.H.; Chun, H.J.; Ahmad, S.S.; Lee, E.J.; Choi, I. Identification and Evaluation of Traditional Chinese Medicine Natural Compounds as Potential Myostatin Inhibitors: An In Silico Approach. *Molecules* **2022**, *27*, 4303. <https://doi.org/10.3390/molecules27134303>.
- Ahmad, S.S.; Ahmad, K.; Lee, E.J.; Shaikh, S.; Choi, I. Computational Identification of Dithymoquinone as a Potential Inhibitor of Myostatin and Regulator of Muscle Mass. *Molecules* **2021**, *26*, 5407. <https://doi.org/10.3390/molecules26175407>.
- Maryanoff, B.E. Drug Discovery and the Medicinal Chemist. *Future Med. Chem.* **2009**, *1*, 11–15. <https://doi.org/10.4155/fmc.09.2>. PMID: 21426067.
- Glassman, P.M.; Muzykantov, V.R. Pharmacokinetic and Pharmacodynamic Properties of Drug Delivery Systems. *J. Pharmacol. Exp. Ther.* **2019**, *370*, 570–580. <https://doi.org/10.1124/jpet.119.257113>.
- Shaikh, S.; Ali, S.; Lim, J.H.; Ahmad, K.; Han, K.S.; Lee, E.J.; Choi, I. Virtual Insights into Natural Compounds as Potential 5 α -Reductase Type II Inhibitors: A Structure-Based Screening and Molecular Dynamics Simulation Study. *Life* **2023**, *13*, 2152. <https://doi.org/10.3390/life13112152>.
- Velkov, T.; Bergen, P.J.; Lora-Tamayo, J.; Landersdorfer, C.B.; Li, J. PK/PD models in antibacterial development. *Curr. Opin. Microbiol.* **2013**, *16*, 573–579. <https://doi.org/10.1016/j.mib.2013.06.010>.
- Cook, D.; Brown, D.; Alexander, R.; March, R.; Morgan, P.; Satterthwaite, G.; Pangalos, M.N. Lessons learned from the fate of AstraZeneca’s drug pipeline: A five-dimensional framework. *Nat. Rev. Drug Discov.* **2014**, *13*, 419–431. <https://doi.org/10.1038/nrd4309>.
- Morgan, P.; Brown, D.G.; Lennard, S.; Anderton, M.J.; Barrett, J.C.; Eriksson, U.; Fidock, M.; Hamrén, B.; Johnson, A.; March, R.E.; et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* **2018**, *17*, 167–181. <https://doi.org/10.1038/nrd.2017.244>.
- Singh, N.; Vayer, P.; Tanwar, S.; Poyet, J.L.; Tsaioun, K.; Villoutreix, B.O. Drug discovery and development: Introduction to the general public and patient groups. *Front. Drug Discov.* **2023**, *3*, 1201419. <https://doi.org/10.3389/fddsv.2023.1201419>.
- Matthews, H.; Hanison, J.; Nirmalan, N. “Omics”-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes* **2016**, *4*, 28. <https://doi.org/10.3390/proteomes4030028>.
- Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80–93. <https://doi.org/10.1016/j.drudis.2020.10.010>.
- Bhardwaj, A.; Kishore, S.; Pandey, D.K. Artificial Intelligence in Biological Sciences. *Life* **2022**, *12*, 1430. <https://doi.org/10.3390/life12091430>.
- Lawrence, E.; El-Shazly, A.; Seal, S.; Joshi, C.K.; Liò, P.; Singh, S.; Bender, A.; Sormanni, P.; Greenig, M. Understanding Biology in the Age of Artificial Intelligence. *arXiv* **2024**, arXiv:2403.04106.
- Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. <https://doi.org/10.1136/svn-2017-000101>.
- Yu, H.; Yang, L.T.; Zhang, Q.; Armstrong, D.; Deen, M.J. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* **2021**, *444*, 92–110. <https://doi.org/10.1016/j.neucom.2020.04.157>.
- Minic, A.; Jovanovic, L.; Bacanin, N.; Stoean, C.; Zivkovic, M.; Spalevic, P.; Petrovic, A.; Dobrojevic, M.; Stoean, R. Applying Recurrent Neural Networks for Anomaly Detection in Electrocardiogram Sensor Data. *Sensors* **2023**, *23*, 9878. <https://doi.org/10.3390/s23249878>.
- He, K.; Mao, R.; Lin, Q.; Ruan, Y.; Lan, X.; Feng, M.; Cambria, E. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *Inf. Fusion* **2025**, *118*, 102963. <https://doi.org/10.1016/j.inffus.2025.102963>.
- Ali, S.; Qadri, Y.A.; Ahmad, K.; Lin, Z.; Leung, M.F.; Kim, S.W.; Vasilakos, A.V.; Zhou, T. Large Language Models in Genomics—A Perspective on Personalized Medicine. *Bioengineering* **2025**, *12*, 440. <https://doi.org/10.3390/bioengineering12050440>.
- U.S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2025. Available online: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices> (accessed on 21 May 2025).
- Mak, K.K.; Pichika, M.R. Artificial intelligence in drug development: Present status and future prospects. *Drug Discov. Today* **2019**, *24*, 773–780. <https://doi.org/10.1016/j.drudis.2018.11.014>.

25. Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J.; Yang, S. Artificial intelligence in drug development. *Nat. Med.* **2025**, *31*, 45–59. <https://doi.org/10.1038/s41591-024-03434-4>.
26. Sellwood, M.A.; Ahmed, M.; Segler, M.H.S.; Brown, N. Artificial intelligence in drug discovery. *Future Med. Chem.* **2018**, *10*, 2025–2028. <https://doi.org/10.4155/fmc-2018-0212>.
27. Pillai, N.; Dasgupta, A.; Sudsakorn, S.; Fretland, J.; Mavroudis, P.D. Machine Learning guided early drug discovery of small molecules. *Drug Discov. Today* **2022**, *27*, 2209–2215. <https://doi.org/10.1016/j.drudis.2022.03.017>.
28. Kirboğa, K.K.; Abbasi, S.; Küçüksille, E. Explainability and White Box in Drug Discovery. *Chem. Biol. Drug Des.* **2023**, *101*, 560–572. <https://doi.org/10.1111/cbdd.14262>.
29. Ding, Q.; Yao, R.; Bai, Y.; Da, L.; Wang, Y.; Xiang, R.; Jiang, X.; Zhai, F. Explainable Artificial Intelligence in the Field of Drug Research. *Drug Des. Dev. Ther.* **2025**, *19*, 4501–4516. <https://doi.org/10.2147/DDDT.S525171>.
30. Ponzoni, I.; Capoferri, L.; Reis, P.A.B.; Holliday, J.D.; Bender, A. Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1681. <https://doi.org/10.1002/wcms.1681>.
31. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. <https://doi.org/10.1038/s42256-020-00236-4>.
32. Vo, T.; Nguyen, N.; Kha, Q.; Le, N. On the Road to Explainable AI in Drug–Drug Interactions Prediction: A Systematic Review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2112–2123. <https://doi.org/10.1016/j.csbj.2022.04.021>.
33. Alizadehsani, R.; Oyelere, S.S.; Hussain, S.; Jagatheesaperumal, S.K.; Calixto, R.R.; Rahouti, M. Explainable Artificial Intelligence for Drug Discovery and Development—A Comprehensive Survey. *IEEE Access* **2024**, pp. 35796–35812. <https://doi.org/10.1109/ACCESS.2024.3373195>.
34. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>.
35. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>.
36. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>.
37. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371.
38. Seddik, B.; Ahlem, D.; Hocine, C. An Explainable Self-Labeling Grey-Box Model. In Proceedings of the 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), Oum El Bouaghi, Algeria, 12–13 October 2022; pp. 1–7. <https://doi.org/10.1109/PAIS56586.2022.9946912>.
39. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. <https://doi.org/10.1007/s12559-023-10179-8>.
40. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Rana, O.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques and solutions. *ACM Comput. Surv.* **2023**, *55*, 194. <https://doi.org/10.1145/3561048>.
41. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1379. <https://doi.org/10.1002/widm.1379>.
42. Vollert, S.; Atzmueller, M.; Theissler, A. Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. In Proceedings of the 2021 IEEE 26th International Conference on Emerging Technologies and Factory Automation (ETFA), Online, 7–10 September 2021; pp. 1–8. <https://doi.org/10.1109/ETFA45728.2021.9613467>.
43. Hanif, A.; Zhang, X.; Wood, S. A Survey on Explainable Artificial Intelligence Techniques and Challenges. In Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Conference Workshops (EDOCW), Gold Coast, Australia, 25–29 October 2021; pp. 81–89. <https://doi.org/10.1109/EDOCW52865.2021.00036>.
44. Salih, A.M.; Wang, Y. Are Linear Regression Models White Box and Interpretable? *arXiv* **2024**, arXiv:2407.12177.
45. Abu-Faraj, M.; Al-Hyari, A.; Alqadi, Z.A.A. Experimental Analysis of Methods Used to Solve Linear Regression Models. *Comput. Mater. Contin.* **2022**, *72*, 5699–5712. <https://doi.org/10.32604/cmc.2022.027364>.
46. Hope, T.M.H. Linear regression. In *Machine Learning: Methods and Applications to Brain Disorders*; Mechelli, A.; Vieira, S., Eds.; Academic Press: London, UK, 2020; pp. 67–81. <https://doi.org/10.1016/B978-0-12-815739-8.00004-3>.
47. Tian, Y.; Zhang, Y. A comprehensive survey on regularization strategies in machine learning. *Inf. Fusion* **2022**, *80*, 146–166. <https://doi.org/10.1016/j.inffus.2021.11.005>.
48. Pargent, F.; Pfisterer, F.; Thomas, J.; Bischl, B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput. Stat.* **2022**, *37*, 2671–2692. <https://doi.org/10.1007/s00180-022-01207-6>.
49. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.

50. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Röttger, R.; Müller, H.; Holzinger, A. Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. <https://doi.org/10.1016/j.cogsys.2024.101243>.
51. Agarwal, R.; Melnick, L.; Frosst, N.; Zhang, X.; Lengerich, B.; Caruana, R.; Hinton, G. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *arXiv* **2020**, arXiv:2004.13912.
52. Wood, S.N. Inference and computation with generalized additive models and their extensions. *TEST* **2020**, *29*, 307–339. <https://doi.org/10.1007/s11749-020-00711-5>.
53. Oviedo, F.; Lavista Ferres, J.; Buonassisi, T.; Butler, K.T. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Accounts Mater. Res.* **2022**, *3*, 597–607. <https://doi.org/10.1021/accountsmr.1c00244>.
54. Lötsch, J.; Kringel, D.; Ultsch, A. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics* **2022**, *2*, 1–17. <https://doi.org/10.3390/biomedinformatics2010001>.
55. Izza, Y.; Ignatiev, A.; Marques-Silva, J. On Explaining Decision Trees. *arXiv* **2020**, arXiv:2010.11034.
56. Kozielski, M.; Sikora, M.; Wawrowski, Ł. Towards consistency of rule-based explainer and black box model: fusion of rule induction and XAI-based feature importance. *arXiv* **2024**, arXiv:2407.14543.
57. Cesarini, M.; Malandri, L.; Pallucchini, F.; Seveso, A.; Xing, F. Explainable AI for Text Classification: Lessons from a Comprehensive Evaluation of Post Hoc Methods. *Cogn. Comput.* **2024**, *16*, 3077–3095. <https://doi.org/10.1007/s12559-024-10325-w>.
58. Gianfagna, L.; Di Cecco, A. *Explainable AI with Python*; Springer: Cham, Switzerland, 2021. <https://doi.org/10.1007/978-3-030-68640-6>.
59. Dieber, J.; Kirrane, S. Why model why? Assessing the strengths and limitations of LIME. *arXiv* **2020**, arXiv:2012.00093.
60. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You? Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>.
61. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
62. Salih, A.M.; Raisi-Estabragh, Z.; Boscolo Galazzo, I.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2024**, *7*, 2400304. <https://doi.org/10.1002/aisy.202400304>.
63. Speith, T. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea, 21–24 June 2022; pp. 1–12. <https://doi.org/10.1145/3531146.3534639>.
64. Weber, L.; Lapuschkin, S.; Binder, A.; Samek, W. Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement. *Inf. Fusion* **2023**, *92*, 154–176. <https://doi.org/10.1016/j.inffus.2022.11.013>.
65. Dehimi, N.E.H.; Tolba, Z. Attention Mechanisms in Deep Learning: Towards Explainable Artificial Intelligence. In Proceedings of the 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), Oum El Bouaghi, Algeria, 24–25 April 2024. <https://doi.org/10.1109/PAIS62114.2024.10541203>.
66. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. *arXiv* **2015**, arXiv:1512.04150.
67. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. <https://doi.org/10.1007/s11263-019-01228-7>.
68. Higgins, D.; Madai, V.I. From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Adv. Intell. Syst.* **2020**, *2*, 2000052. <https://doi.org/10.1002/aisy.202000052>.
69. Nasarian, E.; Alizadehsani, R.; Acharya, U.R.; Tsui, K.L. Designing Interpretable ML System to Enhance Trust in Healthcare: A Systematic Review to Proposed Responsible Clinician-AI-Collaboration Framework. *Inf. Fusion* **2024**, *108*, 102412. <https://doi.org/10.1016/j.inffus.2024.102412>.
70. Rucco, M.; Viticchi, G.; Falsetti, L. Towards Personalized Diagnosis of Glioblastoma in Fluid-Attenuated Inversion Recovery (FLAIR) by Topological Interpretable Machine Learning. *Mathematics* **2020**, *8*, 770. <https://doi.org/10.3390/math8050770>.
71. Carrieri, A.P.; Haiminen, N.; Maudsley-Barton, S.; Gardiner, L.J.; Murphy, B.; Mayes, A.E.; Paterson, S.; Grimshaw, S.; Winn, M.; Shand, C.; et al. Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* **2021**, *11*, 4565. <https://doi.org/10.1038/s41598-021-83922-6>.
72. Magesh, P.R.; Myloth, R.D.; Tom, R.J. An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Comput. Biol. Med.* **2020**, *126*, 104041. <https://doi.org/10.1016/j.combiomed.2020.104041>.
73. Lauritsen, S.M.; Kristensen, M.; Olsen, M.V.; Larsen, M.S.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **2020**, *11*, 3852. <https://doi.org/10.1038/s41467-020-17431-x>.
74. Meldo, A.; Utkin, L.; Kovalev, M.; Kasimov, E. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. *Artif. Intell. Med.* **2020**, *108*, 101952. <https://doi.org/10.1016/j.artmed.2020.101952>.
75. Yeboah, D.; Steinmeister, L.; Hier, D.B.; Hadi, B.; Wunsch, D.C.; Olbricht, G.R.; Obafemi-Ajayi, T. An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups. *IEEE Access* **2020**, *8*, 180690–180705. <https://doi.org/10.1109/ACCESS.2020.3027453>.

76. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *Sci. Rep.* **2020**, *10*, 19549. <https://doi.org/10.1038/s41598-020-76550-z>.
77. Yao, L.; Syed, A.R.; Rahman, M.H.; Rahman, M.M.; Foraker, R.E.; Banerjee, I. Predicting Post-stroke Hospital Discharge Disposition Using Interpretable Machine Learning Approaches. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Angeles, CA, USA, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2955–2961. <https://doi.org/10.1109/BigData47090.2019.9006592>.
78. Ye, Q.; Xia, J.; Yang, G. Explainable AI For COVID-19 CT Classifiers: An Initial Comparison Study. *arXiv* **2021**, arXiv:2104.14506.
79. Reyna, M.A.; Josef, C.S.; Jeter, R.; Shashikumar, S.P.; Westover, M.B.; Nemat, S.; Clifford, G.D.; Sharma, A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge. *Crit. Care Med.* **2020**, *48*, 210–217. <https://doi.org/10.1097/CCM.0000000000004145>.
80. Varzandian, A.; Razo, M.A.S.; Sanders, M.R.; Atmakuru, A.; Fatta, G.D. Classification-Biased Apparent Brain Age for the Prediction of Alzheimer’s Disease. *Front. Neurosci.* **2021**, *15*, 673120. <https://doi.org/10.3389/fnins.2021.673120>.
81. Pierson, E.; Cutler, D.M.; Leskovec, J.; Mullainathan, S.; Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **2021**, *27*, 136–140. <https://doi.org/10.1038/s41591-020-01192-7>.
82. Born, J.; Wiedemann, N.; Cossio, M.; Buhre, C.; Brändle, G.; Leidermann, K.; Goulet, J.; Aujayeb, A.; Moor, M.; Rieck, B.; et al. Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis. *Appl. Sci.* **2021**, *11*, 672. <https://doi.org/10.3390/app11020672>.
83. Shen, Y.; Wu, N.; Phang, J.; Park, J.; Liu, K.; Tyagi, S.; Heacock, L.; Kim, S.G.; Moy, L.; Cho, K.; et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **2021**, *68*, 101908. <https://doi.org/10.1016/j.media.2020.101908>.
84. Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Wang, R.; Zhao, H.; Zha, Y.; et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT Images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 2775–2780. <https://doi.org/10.1109/TCBB.2021.3065361>.
85. Wang, S.H.; Govindaraj, V.V.; Górriz, J.M.; Zhang, X.; Zhang, Y.D. COVID-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network. *Inf. Fusion* **2021**, *67*, 208–229. <https://doi.org/10.1016/j.inffus.2020.10.004>.
86. Fan, Z.; Gong, P.; Tang, S.; Lee, C.U.; Zhang, X.; Song, P.; Chen, S.; Li, H. Joint localization and classification of breast masses on ultrasound images using an auxiliary attention-based framework. *Med. Image Anal.* **2023**, *90*, 102960. <https://doi.org/10.1016/j.media.2023.102960>.
87. Sutton, R.T.; Zaiane, O.R.; Goebel, R.; Baumgart, D.C. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Sci. Rep.* **2022**, *12*, 2748. <https://doi.org/10.1038/s41598-022-06726-2>.
88. Lu, S.; Zhu, Z.; Górriz, J.M.; Wang, S.H.; Zhang, Y.D. NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network. *Int. J. Intell. Syst.* **2022**, *37*, 1572–1598. <https://doi.org/10.1002/int.22686>.
89. Pun, N.S.; Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Appl. Intell.* **2021**, *51*, 2689–2702. <https://doi.org/10.1007/s10489-020-01900-3>.
90. Wang, H.; Wang, S.; Qin, Z.; Zhang, Y.; Li, R.; Xia, Y. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Med. Image Anal.* **2021**, *67*, 101846. <https://doi.org/10.1016/j.media.2020.101846>.
91. Alsinglawi, B.; Alshari, O.; Alorjani, M.; Mubin, O.; Alnajjar, F.; Novoa, M.; Darwish, O. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci. Rep.* **2022**, *12*, 607. <https://doi.org/10.1038/s41598-021-04608-7>.
92. Le, N.Q.K.; Kha, Q.H.; Nguyen, V.H.; Chen, Y.C.; Cheng, S.J.; Chen, C.Y. Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer. *Int. J. Mol. Sci.* **2021**, *22*, 9254. <https://doi.org/10.3390/ijms22179254>.
93. Abeyagunasekera, S.H.P.; Perera, Y.; Chamara, K.; Kaushalya, U.; Sumathipala, P.; Senaweera, O. LISA: Enhance the explainability of medical images unifying current XAI techniques. In Proceedings of the 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Pune, India, 7–9 April 2022; pp. 1–9. <https://doi.org/10.1109/I2CT54291.2022.9824840>.
94. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2020**, *63*, 8761–8777. <https://doi.org/10.1021/acs.jmedchem.9b01101>.
95. Takagi, A.; Kamada, M.; Hamatani, E.; Kojima, R.; Okuno, Y. GraphIX: Graph-based In silico XAI (explainable artificial intelligence) for drug repositioning from biopharmaceutical network. *arXiv* **2022**, arXiv:2212.10788.
96. Cao, H.; Liu, Z.; Lu, X.; Yao, Y.; Li, Y. InstructMol: Multi-Modal Integration for Building a Versatile and Reliable Molecular Assistant in Drug Discovery. *arXiv* **2024**, arXiv:2311.16208v2.
97. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

98. Kamya, P.; Ozerov, I.V.; Pun, F.W.; Tretina, K.; Fokina, T.; Chen, S.; Naumov, V.; Long, X.; Lin, S.; Korzinkin, M.; et al. PandaOmics: An AI-Driven Platform for Therapeutic Target and Biomarker Discovery. *J. Chem. Inf. Model.* **2024**, *64*, 3961–3969. <https://doi.org/10.1021/acs.jcim.3c01619>.
99. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Precup, D., Teh, Y.W., Eds.; PMLR: New York, NY, USA, 2017; pp. 3319–3328.
100. Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *J. Chem. Inf. Model.* **2022**, *62*, 447–462. <https://doi.org/10.1021/acs.jcim.1c01263>.
101. Maia, E.H.B.; de Souza, L.H.M.; de Souza, R.T.; Andricopulo, A.D. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8*, 343. <https://doi.org/10.3389/fchem.2020.00343>.
102. Schneider, P.; Walters, W.P.; Plowright, A.T.; Sieroka, N.; Listgarten, J.; Goodnow, R.A., Jr.; Fisher, J.; Jansen, J.M.; Duca, J.S.; Rush, T.S.; et al. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* **2020**, *19*, 353–364. <https://doi.org/10.1038/s41573-019-0050-3>.
103. Sahoo, B.M.; Kumar, B.V.V.R.; Sruti, J.; Mahapatra, M.K.; Banik, B.K.; Borah, P. Drug Repurposing Strategy (DRS): Emerging Approach to Identify Potential Therapeutics for Treatment of Novel Coronavirus Infection. *Front. Mol. Biosci.* **2021**, *8*, 628144. <https://doi.org/10.3389/fmolb.2021.628144>.
104. Ali, S.; Shaikh, S.; Ahmad, K.; Choi, I. Identification of active compounds as novel dipeptidyl peptidase-4 inhibitors through machine learning and structure-based molecular docking simulations. *J. Biomol. Struct. Dyn.* **2025**, *43*, 1611–1620. <https://doi.org/10.1080/07391102.2023.2292299>.
105. Danishuddin; Kumar, V.; Faheem, M.; Lee, K.W. A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. *Drug Discov. Today* **2022**, *27*, 529–537. <https://doi.org/10.1016/j.drudis.2021.09.013>.
106. Rao, J.; Zheng, S.; Lu, Y.; Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns* **2022**, *3*, 100628. <https://doi.org/10.1016/j.patter.2022.100628>.
107. Jiménez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *J. Chem. Inf. Model.* **2021**, *61*, 1083–1094. <https://doi.org/10.1021/acs.jcim.0c01344>.
108. Shen, L.; Bai, J.; Jiao, W.; Shen, B. The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead. *Precis. Clin. Med.* **2021**, *4*, 80–84. <https://doi.org/10.1093/pcmedi/pbab007>.
109. Drancé, M. Neuro-Symbolic XAI: Application to Drug Repurposing for Rare Diseases. In *Database Systems for Advanced Applications*; Bhattacharya, A., Lee Mong Li, J., Agrawal, D., Reddy, P.K., Mohania, M., Mondal, A., Goyal, V., Uday Kiran, R., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 539–543.
110. Askari, M.; Eslami, S.; Louws, M.; Wierenga, P.C.; Dongelmans, D.A.; Kuiper, R.A.; Abu-Hanna, A. Frequency and nature of drug–drug interactions in the intensive care unit. *Pharmacoepidemiol. Drug Saf.* **2013**, *22*, 430–437. <https://doi.org/10.1002/pds.3415>.
111. Bories, M.; Bouzillé, G.; Cuggia, M.; Le Corre, P. Drug–Drug Interactions in Elderly Patients with Potentially Inappropriate Medications in Primary Care, Nursing Home and Hospital Settings: A Systematic Review and a Preliminary Study. *Pharmaceutics* **2021**, *13*, 266. <https://doi.org/10.3390/pharmaceutics13020266>.
112. Reis, A.M.M.; Cassiani, S.H.D.B. Evaluation of three brands of drug interaction software for use in intensive care units. *Pharm. World Sci.* **2010**, *32*, 822–828. <https://doi.org/10.1007/s11096-010-9445-2>.
113. Vonbach, P.; Dubied, A.; Krähenbühl, S.; Beer, J.H. Evaluation of frequently used drug interaction screening programs. *Pharm. World Sci.* **2008**, *30*, 367–374. <https://doi.org/10.1007/s11096-008-9191-x>.
114. Cheng, F.; Zhao, Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J. Am. Med. Inform. Assoc.* **2014**, *21*, e278–e286. <https://doi.org/10.1136/amiajnl-2013-002512>.
115. Ryu, J.Y.; Kim, H.U.; Lee, S.Y. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4304–E4311. <https://doi.org/10.1073/pnas.1803294115>.
116. Vilar, S.; Uriarte, E.; Santana, L.; Lorberbaum, T.; Hripcsak, G.; Friedman, C.; Tatonetti, N.P. Similarity-based modeling in large-scale prediction of drug–drug interactions. *Nat. Protoc.* **2014**, *9*, 2147–2163. <https://doi.org/10.1038/nprot.2014.151>.
117. Xu, L.; Ru, X.; Song, R. Application of Machine Learning for Drug–Target Interaction Prediction. *Front. Genet.* **2021**, *12*, 680117. <https://doi.org/10.3389/fgene.2021.680117>.
118. Ideker, T.; Nussinov, R. Network approaches and applications in biology. *PLoS Comput. Biol.* **2017**, *13*, e1005771. <https://doi.org/10.1371/journal.pcbi.1005771>.
119. Lai, X.; Gupta, S.K.; Schmitz, U.; Marquardt, S.; Knoll, S.; Spitschak, A.; Wolkenhauer, O.; Pützer, B.M.; Vera, J. MiR-205-5p and miR-342-3p cooperate in the repression of the E2F1 transcription factor in the context of anticancer chemotherapy resistance. *Theranostics* **2018**, *8*, 1106–1120. <https://doi.org/10.7150/thno.19904>.
120. Lai, X.; Eberhardt, M.; Schmitz, U.; Vera, J. Systems biology-based investigation of cooperating microRNAs as monotherapy or adjuvant therapy in cancer. *Nucleic Acids Res.* **2019**, *47*, 7753–7766. <https://doi.org/10.1093/nar/gkz638>.

121. You, Y.; Lai, X.; Pan, Y.; Zheng, H.; Vera, J.; Liu, S.; Deng, S.; Zhang, L. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct. Target. Ther.* **2022**, *7*, 156. <https://doi.org/10.1038/s41392-022-00994-0>.
122. Peyvandipour, A.; Saberian, N.; Shafi, A.; Donato, M.; Drăghici, S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **2018**, *34*, 2817–2825. <https://doi.org/10.1093/bioinformatics/bty133>.
123. Würth, R.; Thellung, S.; Bajetto, A.; Mazzanti, M.; Florio, T.; Barbieri, F. Drug-repositioning opportunities for cancer therapy: Novel molecular targets for known compounds. *Drug Discov. Today* **2016**, *21*, 190–199. <https://doi.org/10.1016/j.drudis.2015.09.017>.
124. Pantziarka, P.; Bouche, G.; Meheus, L.; Sukhatme, V.; Sukhatme, V.P. Repurposing drugs in your medicine cabinet: Untapped opportunities for cancer therapy? *Future Oncol.* **2015**, *11*, 181–184. <https://doi.org/10.2217/fon.14.244>.
125. Park, K. A review of computational drug repurposing. *Transl. Clin. Pharmacol.* **2019**, *27*, 59–63. <https://doi.org/10.12793/tcp.2019.27.2.59>.
126. Al-Taie, Z.; Liu, D.; Mitchem, J.B.; Papageorgiou, C.; Kaifi, J.T.; Warren, W.C.; Shyu, C.R. Explainable Artificial Intelligence in High-Throughput Drug Repositioning for Subgroup Stratifications with Interventionable Potential. *J. Biomed. Inform.* **2021**, *118*, 103792. <https://doi.org/10.1016/j.jbi.2021.103792>.
127. Xue, H.; Li, J.; Xie, H.; Wang, Y. Review of Drug Repositioning Approaches and Resources. *Int. J. Biol. Sci.* **2018**, *14*, 1232–1244. <https://doi.org/10.7150/ijbs.24612>.
128. Lotfi Shahreza, M.; Ghadiri, N.; Mousavi, S.R.; Varshosaz, J.; Green, J.R. Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning. *J. Biomed. Inform.* **2017**, *68*, 167–183. <https://doi.org/10.1016/j.jbi.2017.03.006>.
129. Xu, R.; Wang, Q. PhenoPredict: A disease phenome-wide drug repositioning approach towards schizophrenia drug discovery. *J. Biomed. Inform.* **2015**, *56*, 348–355. <https://doi.org/10.1016/j.jbi.2015.06.027>.
130. Xu, R.; Wang, Q. A genomics-based systems approach towards drug repositioning for rheumatoid arthritis. *BMC Genom.* **2016**, *17*, 518. <https://doi.org/10.1186/s12864-016-2910-0>.
131. Lamb, J. The Connectivity Map: A new tool for biomedical research. *Nat. Rev. Cancer* **2007**, *7*, 54–60. <https://doi.org/10.1038/nrc2044>.
132. Lee, B.K.B.; Tiong, K.H.; Chang, J.K.; Liew, C.S.; Abdul Rahman, Z.A.; Tan, A.C.; Khang, T.F.; Cheong, S.C. DeSigN: Connecting gene expression with therapeutics for drug repurposing and development. *BMC Genom.* **2017**, *18*, 934. <https://doi.org/10.1186/s12864-016-3260-7>.
133. Tian, Z.; Teng, Z.; Cheng, S.; Guo, M. Computational drug repositioning using meta-path-based semantic network analysis. *BMC Syst. Biol.* **2018**, *12*, 134. <https://doi.org/10.1186/s12918-018-0658-7>.
134. Wang, Q.; Huang, K.; Chandak, P.; Zitnik, M.; Gehlenborg, N. Extending the nested model for user-centric XAI: A design study on GNN-based drug repurposing. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 1266–1276. <https://doi.org/10.1109/TVCG.2022.3209435>.
135. Zhang, B.; Huang, Z.; Zheng, H.; Li, W.; Liu, Z.; Zhang, Y.; Huang, Q.; Liu, X.; Jiang, H.; Liu, Q. EFMSDTI: Drug–target interaction prediction based on an efficient fusion of multi-source data. *Front. Pharmacol.* **2022**, *13*, 1009996. <https://doi.org/10.3389/fphar.2022.1009996>.
136. Huang, A.; Xie, X.; Wang, X.; Peng, S. A Multimodal Data Fusion-Based Deep Learning Approach for Drug–Drug Interaction Prediction. In *Bioinformatics Research and Applications; Lecture Notes in Computer Science; Springer, Cham, Switzerland*, 2023; Volume 13760, pp. 275–285. https://doi.org/10.1007/978-3-031-23198-8_25.
137. Sturm, H.; Teufel, J.; Isfeld, K.; Friederich, P.; Davis, R. Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI. *Angew. Chem. Int. Ed.* **2025**, *137*, e202503259. <https://doi.org/10.1002/ange.202503259>.
138. Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 9244–9255.
139. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* **2019**, *366*, 447–453. <https://doi.org/10.1126/science.aax2342>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.