Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

EXPLAINABLE AI FOR PHARMACEUTICAL PREDICTION: A CRITICAL REVIEW OF TRUST AND REPRODUCIBILITY

Yousef Al-Qahtani[1]*, Fahad Al-Salem[1], Abdullah Al-Harbi[2]

1. ***Department of Intelligent Pharmaceutical Engineering, Faculty of Pharmacy, King Saud University, Riyadh, Saudi Arabia.***
2. ***Department of Computational Drug Systems, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia.***

ARTICLE INFO

Received:
21 May 2025
Received in revised form:
08 September 2025
Accepted:
09 September 2025
Available online:
28 October 2025

Keywords: Explainable artificial intelligence, Pharmaceutical prediction, Drug discovery, ADMET, Trust, Reproducibility

ABSTRACT

Explainable artificial intelligence (XAI) has been widely advocated as a solution to the opacity of machine learning models in pharmaceutical prediction, yet the connection between explanation, trust, reproducibility, and scientific validity remains unresolved. A growing body of literature applies explanation methods across drug discovery, ADMET prediction, formulation design, and clinical pharmacology; however, much of this work assumes that making predictions visually or numerically interpretable inherently confers trustworthiness. This critical review examines the strengths, weaknesses, and ongoing challenges of XAI in pharmaceutical contexts, with particular focus on user trust, reproducibility of explanations, and suitability for regulated decision-making. The literature highlights persistent gaps, including a lack of human-centered evaluation, limited assessment of explanation stability, and a misalignment between common explanation outputs and regulatory expectations. While many studies present plausible explanations, far fewer demonstrate that these explanations meaningfully improve decisions. Without rigorous validation, XAI risks obscuring rather than clarifying model behavior, and in high-stakes pharmaceutical settings, intuitive but non-robust explanations may foster misplaced confidence. This review therefore proposes a framework for assessing the maturity of XAI in pharmaceutical prediction, emphasizing the need to advance from appealing explanatory artifacts toward reproducible, uncertainty-aware, and decision-tested explanation systems.

To Cite This Article: Al-Qahtani Y, Al-Salem F, Al-Harbi A. Explainable AI for Pharmaceutical Prediction: A Critical Review of Trust and Reproducibility. Pharmacophore. 2025;16(5):10-9. https://doi.org/10.51847/KeUPhBjlQn

Introduction

Machine learning has become increasingly visible across pharmaceutical sciences, including molecular property prediction, virtual screening, adverse event modeling, formulation development, and clinical pharmacology [1, 2]. Recent reviews portray explainable artificial intelligence as a necessary bridge between high-performing predictive models and the evidentiary expectations of scientific users [3, 4]. The appeal is understandable: if machine learning systems are to influence compound prioritization, safety assessment, or dose optimization, their outputs must be intelligible to experts who bear responsibility for downstream decisions [5]. Yet the rapid normalization of XAI in pharmaceutical papers has outpaced careful scrutiny of what these explanations actually prove.

A central assumption in the literature is that interpretability improves trust, safety, and adoption, but this assumption is rarely tested under realistic pharmaceutical decision conditions [6]. General critiques of explainable AI in health care warn that explanations can create false reassurance when they are plausible to humans but weakly connected to model behavior or clinical utility [7]. Similarly, arguments for inherently interpretable models in high-stakes domains challenge the routine use of post-hoc explanations as a substitute for transparent modeling [8]. In pharmaceutical prediction, this concern is amplified by small datasets, distribution shift, assay variability, and the frequent use of proxy endpoints.

The field now contains many examples of SHAP, counterfactual, attention-based, and graph-derived explanations for molecular and biomedical prediction tasks [9-11]. However, the number of papers using XAI is not equivalent to the amount of evidence that XAI improves pharmaceutical outcomes or satisfies regulatory expectations [12, 13]. Studies often show feature rankings, highlighted atoms, or plausible substructures, but rarely ask whether medicinal chemists, toxicologists,

Corresponding Author: Yousef Al-Qahtani; Department of Intelligent Pharmaceutical Engineering, Faculty of Pharmacy, King Saud University, Riyadh, Saudi Arabia. E-mail: yousef.qahtani@gmail.com.

formulators, or regulators would make better decisions because of those outputs [14, 15]. This creates a gap between technical explanation production and the higher standard of decision support required in regulated science.

The Landscape of XAI in Pharmaceutical Prediction

Dominant XAI Techniques

The most common XAI techniques in pharmaceutical prediction are post-hoc feature-attribution methods such as SHAP and LIME, gradient-based methods such as integrated gradients, attention mechanisms in neural networks, and counterfactual explanations [1, 4]. SHAP is especially popular because it produces ranked feature contributions that can be mapped onto molecular descriptors, assay variables, or clinical covariates [13, 14]. Attention and graph-based explanations are attractive in molecular modeling because they can be visualized on atoms, bonds, or subgraphs, but attention weights do not necessarily constitute faithful explanations of model reasoning [16, 17]. Counterfactual approaches are conceptually stronger for decision support because they ask how a molecule or input profile would need to change to alter a prediction, yet their feasibility and validity depend on whether the proposed changes are chemically, biologically, and operationally realistic [10, 18]. **Table 1** summarizes how major XAI methods differ in their explanation output, pharmaceutical interpretability value, and key limitations for model-supported prediction.

Table 1. Practical interpretation value of common XAI techniques in pharmaceutical prediction

XAI technique	Typical explanation output	Added value in pharmaceutical prediction	Main caution
SHAP / LIME	Ranked feature contributions	Links predictions to descriptors, assay variables, or clinical covariates	May over-simplify correlated or interacting features
Integrated gradients	Input-level contribution scores	Useful for neural models using molecular fingerprints, sequences, or embeddings	Requires careful baseline selection
Attention or graph explanations	Highlighted atoms, bonds, subgraphs, or input tokens	Supports visually intuitive interpretation in molecular and biomedical models	Attention weights may not reflect true causal reasoning
Counterfactual explanations	Minimal input change needed to alter the prediction	Helps translate predictions into testable molecular or operational alternatives	Proposed changes must remain chemically and biologically plausible

Target Domains

XAI has been applied to diverse pharmaceutical prediction targets, including molecular activity, physicochemical properties, ADMET endpoints, drug–food interactions, adverse drug reactions, drug sensitivity, and formulation behavior [19, 20]. In drug discovery, explanations often identify molecular fragments or descriptors associated with predicted activity, as seen in activity-cliff analysis and compound optimization studies [9, 11]. In safety and clinical-facing tasks, interpretable models have been used to explain adverse drug reaction prediction, drug–drug interaction risk, drug-induced cardiotoxicity, and pathway-linked drug sensitivity [16, 17, 21-23]. Formulation work has also begun to use interpretable machine learning to connect excipient or process variables with in vitro performance, although this domain remains less developed than molecular prediction [24].

Common Evaluation Practices

The dominant evaluation practice in pharmaceutical XAI is to present explanations as feature rankings, highlighted substructures, or qualitative case studies and then judge whether they appear consistent with domain knowledge [9, 25]. This approach is useful for hypothesis generation but weak as evidence, because plausibility does not establish fidelity, stability, or decision utility [5]. Benchmarking work has begun to expose this weakness by showing that different attribution methods can behave inconsistently around activity cliffs, where small structural changes produce large activity changes [9]. The lack of standardized evaluation protocols means that two XAI studies may both claim interpretability while measuring fundamentally different properties of explanation quality [4, 12].

How Well Do We Trust? Evaluating User-Centered Impact

The Trust Hypothesis

The trust hypothesis in pharmaceutical XAI holds that users will be more willing to accept machine learning predictions when they can see the factors behind them [1, 6]. This hypothesis is often asserted in drug discovery and pharmaceutical development papers, but direct measurement of trust among medicinal chemists, formulators, pharmacometricians, or regulatory scientists remains rare [13, 14]. Broader human-computer interaction research shows that data scientists use interpretability tools unevenly and often treat them as aids for debugging rather than as proof of model validity [26]. Therefore, pharmaceutical XAI should be cautious about importing the language of trust without demonstrating how explanations affect expert judgment in realistic tasks.

Over-Trust and Misinterpretation

A major risk is that explanations may increase trust even when they are incomplete, unstable, or misleading [7]. Research on misleading black-box explanations shows that users can be manipulated into trusting models through persuasive but unfaithful explanation interfaces [27]. This concern is highly relevant to pharmaceutical prediction because SHAP plots, molecular heatmaps, and pathway diagrams can appear scientifically meaningful even when they reflect artifacts of data sampling, confounding, or leakage [15, 20]. If explanation interfaces are evaluated only by whether they satisfy users, they may reward rhetorical clarity rather than scientific reliability.

The Missing User Study

Despite frequent claims that XAI will help pharmaceutical experts, controlled user studies in this specific context are strikingly scarce [4, 6]. Many drug discovery studies demonstrate that an explanation aligns with known chemistry, but they do not test whether medicinal chemists make better compound-selection decisions after viewing it [11, 25]. Clinical pharmacology and population modeling work has begun to frame explainability as a practical workflow issue, as in SHAP-based covariate identification, but even there the evidence is stronger for workflow support than for measured improvement in expert decision quality [13, 14]. The missing user study is therefore not a minor evidence gap; it is central to whether XAI can claim to improve pharmaceutical practice.

Actionability vs. Explainability

Explaining a prediction is not the same as providing an actionable recommendation [5]. A model may identify lipophilicity, molecular weight, or a clinical covariate as influential, but this does not tell scientists whether changing that factor is feasible, safe, causal, or likely to generalize [19, 23]. Counterfactual and substructure-based explanation methods move closer to actionability by suggesting what changes might alter a prediction, yet they require stronger constraints to avoid proposing chemically implausible or pharmacologically unsafe modifications [10, 18]. The field therefore needs to distinguish interpretability that satisfies curiosity from explanation that supports defensible intervention.

How Reproducible Are Explanations? Measurement and Methodological Gaps

Explanation Instability

Explanation instability is one of the most serious but underreported problems in pharmaceutical XAI [9]. SHAP values and related feature-attribution outputs can vary when models are retrained on different splits, when correlated descriptors change, or when input perturbations alter local decision boundaries [13]. In small or heterogeneous pharmaceutical datasets, such instability can be especially consequential because feature rankings may be interpreted as mechanistic clues rather than model-contingent artifacts [15, 20]. Benchmarking molecular attributions against activity cliffs illustrates that explanation behavior can diverge precisely where medicinal chemistry decisions are most sensitive [9].

Data Leakage and Its Effect on Explanations

Data leakage undermines both predictions and explanations because a model that learns shortcuts can produce feature importances that appear convincing while reflecting invalid information flow [5]. Pharmaceutical datasets are vulnerable to leakage through scaffold overlap, duplicated compounds, assay batch effects, temporal splitting errors, and endpoint definitions that encode downstream knowledge [9, 11]. When such a model is explained, the resulting attribution can falsely appear to recover meaningful chemistry or biology, thereby reinforcing confidence in an invalid model [20, 28]. This makes leakage not only a predictive validation issue but also an epistemic threat to the interpretation of XAI outputs.

Lack of Reporting Standards

Most pharmaceutical XAI papers do not report explanation uncertainty, sensitivity to data resampling, confidence intervals for feature importance, or the effect of hyperparameter choices on explanations [4, 12]. Practical guides to SHAP in drug development have helped clarify implementation steps, but the broader literature still lacks shared reporting expectations for explanation robustness [13]. Without such standards, readers cannot determine whether a reported feature ranking is a stable property of the learned relationship or a fragile product of a single training run [14]. This weakness is particularly problematic in regulated settings, where claims about model behavior must be traceable, repeatable, and scientifically defensible. **Table 2** identifies the minimum robustness checks that should accompany pharmaceutical XAI results when explanations are used to support scientific interpretation or regulated decision-making.

Table 2. Minimum reporting checks for robust pharmaceutical XAI explanations

Robustness check	What should be reported	Why it adds value
Resampling stability	Whether feature rankings remain similar across train-test splits, bootstraps, or cross-validation folds	Shows whether explanations are reproducible rather than dependent on one dataset partition
Explanation uncertainty	Confidence intervals, variability ranges, or uncertainty bands for key feature effects	Helps readers judge whether top-ranked drivers are meaningfully distinguishable
Hyperparameter sensitivity	Whether explanation patterns change after reasonable model-tuning choices	Separates stable scientific signals from tuning artifacts

Method comparison	Agreement or disagreement between SHAP, LIME, gradients, attention, or counterfactual outputs	Reduces dependence on a single explanation method
Traceable reporting	Model version, dataset version, preprocessing steps, and explanation settings	Supports auditability, repeatability, and regulatory defensibility

Comparing Models vs. Comparing Explanations

A model can show reproducible predictive performance while producing unstable explanations, and this distinction is often overlooked [9]. Pharmaceutical papers usually compare models using accuracy, area under the curve, or error metrics, but they rarely compare explanations using stability, faithfulness, or sensitivity metrics [10, 21]. Graph neural network studies may present visually compelling subgraph rationales, yet there is no consensus on how to decide whether those rationales are reproducible across model seeds, datasets, or related chemical series [17, 22]. As a result, the community has made more progress in benchmarking predictions than in benchmarking the scientific reliability of the explanations attached to them.

The Illusion of Actionability: Bridging Explanation and Decision

Explanations That Tell You What, Not Why

Many XAI methods identify what features influenced a prediction without explaining why those features matter mechanistically [5]. In pharmaceutical science, this distinction is crucial because the desired output is often not a ranked list of descriptors but a reasoned basis for changing a molecule, formulation, dose, or process parameter [11, 24]. A SHAP value indicating that a descriptor increases predicted toxicity does not establish whether that descriptor is causal, modifiable, or merely correlated with an unmeasured liability [13, 23]. Therefore, feature attribution should be treated as a starting point for scientific interrogation rather than as evidence of mechanistic understanding.

Domain-Augmented XAI

Domain-augmented XAI is more promising because it attempts to connect explanations to biochemical pathways, known pharmacology, chemical substructures, or process knowledge [16]. Drug sensitivity models that combine chemical structures with gene pathways can make explanations more biologically interpretable, although interpretability still depends on the validity of the pathway representation and the data used to train the model [16]. Cardiotoxicity and adverse event studies similarly show the value of linking predictions to clinically meaningful or mechanistically relevant signals, but they also reveal how easily explanation can become a retrospective rationalization of model output [20, 23]. The most useful pharmaceutical explanations are likely to be those constrained by domain knowledge while still subjected to formal tests of fidelity and robustness.

The Counterfactual Promise

Counterfactual explanations offer a stronger bridge between explanation and decision because they express model behavior in the form of hypothetical changes [10]. In molecular design, counterfactuals can suggest structural modifications that might reduce toxicity or increase activity, and substructure-based approaches can make these suggestions more intelligible to chemists [18]. However, counterfactual plausibility is not guaranteed: a mathematically valid change may be synthetically inaccessible, biologically unsafe, or outside the model's domain of applicability [10, 18]. The promise of counterfactual XAI will remain limited until the field validates whether counterfactuals are chemically realistic, experimentally useful, and stable under model retraining.

Regulatory Expectations vs. Current XAI Practice

Emerging Regulatory Guidance

Regulatory discussion of artificial intelligence in drug development increasingly emphasizes transparency, lifecycle control, documentation, and the need for valid scientific evidence rather than mere algorithmic novelty [29, 30]. Landscape analyses of AI and machine learning in regulatory submissions show that sponsors are already using these methods across drug development, but the evidentiary expectations remain heterogeneous and context-dependent [30]. Recent regulatory commentary argues that AI-enabled therapeutic ecosystems require governance across data, models, users, and post-deployment monitoring, which is broader than the explanation of any single prediction [31]. This creates a demanding standard for pharmaceutical XAI: explanations must be embedded in an accountable decision process, not simply appended as visual outputs.

The Compliance Gap

Typical pharmaceutical XAI studies provide local feature attributions, molecular heatmaps, or ranked descriptors, but these outputs do not by themselves satisfy expectations for uncertainty quantification, traceability, validation, and lifecycle management [13, 29]. The compliance gap is especially clear when post-hoc explanations are presented without evidence that they are stable across datasets, robust to distribution shift, or understandable to the intended regulated user [4, 7]. Even practical SHAP workflows in drug development require careful framing because a clear feature ranking can still be scientifically weak if it lacks sensitivity analysis or decision relevance [13]. The current literature therefore offers useful exploratory tools but rarely the level of documented assurance needed for high-stakes regulatory reliance.

Industry Perspectives

Industry and regulatory-facing analyses generally view AI as promising but not yet mature enough to be accepted on the basis of opaque performance claims or superficial transparency [29, 31]. The strongest industry-relevant papers stress that explainability must be paired with model validation, domain expertise, auditability, and governance across the model lifecycle [2, 30]. This is a more demanding position than the one implied by many XAI application papers, where an explanation plot is treated as evidence of interpretability without demonstrating user benefit or regulatory sufficiency [1, 6]. The emerging consensus is therefore cautious: XAI may support submissions, but it is not a substitute for scientific justification, prospective validation, and transparent risk management.

Emerging Solutions and Methodological Advances

Uncertainty-Aware Explanations

Uncertainty-aware explanations are an important next step because they would indicate not only which features influenced a prediction, but also how reliable that explanation is under data and model uncertainty [3, 13]. This is particularly important in pharmaceutical datasets, where sparse endpoints, noisy assays, and chemical-series bias can make confident-looking explanations misleading [9, 15]. ADMET and cardiotoxicity applications show the value of interpretable outputs, but they also highlight the risk of presenting model-derived explanations without explicit uncertainty around the explanatory claim [23]. A mature XAI workflow should therefore communicate when an explanation is unstable, outside the model's domain of applicability, or insufficient for decision-making.

Rigorous Evaluation Frameworks

Rigorous XAI evaluation requires benchmarking faithfulness, sensitivity, stability, human understandability, and actionability rather than relying on qualitative plausibility alone [4, 9]. Molecular activity-cliff benchmarking is a useful example because it tests whether attribution methods behave sensibly in chemically difficult regions where explanations matter most [9]. Counterfactual frameworks add another evaluation burden: they must be judged not only by whether they change the model output, but also by whether the proposed molecular change is chemically valid, synthesizable, and pharmacologically meaningful [10, 18]. Pharmaceutical-specific XAI challenges should therefore include expert review, perturbation testing, uncertainty reporting, and prospective decision tasks.

Inherently Interpretable Models

A growing critique argues that high-stakes decisions should preferentially use inherently interpretable models when their performance is adequate, instead of explaining black boxes after the fact [8]. In pharmaceutical prediction, sparse models, decision trees, rule lists, and explainable boosting machines may be attractive for tasks where transparency, auditability, and reproducibility outweigh marginal gains in predictive accuracy [6, 11]. This does not mean that deep learning and graph neural networks should be abandoned, because they can capture complex molecular and biological relationships [17, 22]. Rather, the burden of justification should shift: when black-box models are used, authors should demonstrate why a simpler interpretable alternative is insufficient and why the post-hoc explanation is faithful.

Cross-Domain Lessons From Other Safety-Critical Fields

Aviation and Medical Devices

Safety-critical fields such as aviation and medical devices treat trust in automation as a property of systems, procedures, training, monitoring, and certification rather than as a feature of an explanatory graphic [31]. This lesson is directly relevant to pharmaceutical XAI because drug development decisions are embedded in institutional workflows, quality systems, and accountability structures [29]. Health care critiques warn that explainability can create false hope if it is not linked to clinical utility, workflow integration, and measurable user outcomes [7]. Pharmaceutical XAI should therefore move away from one-off explanation demonstrations and toward validated human-AI work systems.

Finance and Law

Finance and law offer useful analogies because both fields confront algorithmic accountability, audit trails, contestability, and the need to justify decisions to affected stakeholders [26, 27]. Studies showing that users can be misled by persuasive explanations are especially relevant to governance because they reveal that explanation interfaces can shape trust independently of truth [27]. Pharmaceutical organizations should therefore treat XAI outputs as auditable claims requiring documentation, version control, and challenge mechanisms, not merely as communication aids [30]. The lesson is that transparency without accountability can become another form of opacity.

Critical Framework for Evaluating XAI in Pharma

Figure 1 summarizes the review's critical framework by showing how common pharmaceutical XAI outputs must pass faithfulness, stability, leakage, user-impact, actionability, uncertainty, and regulatory-alignment tests before they can support decision-grade trust.

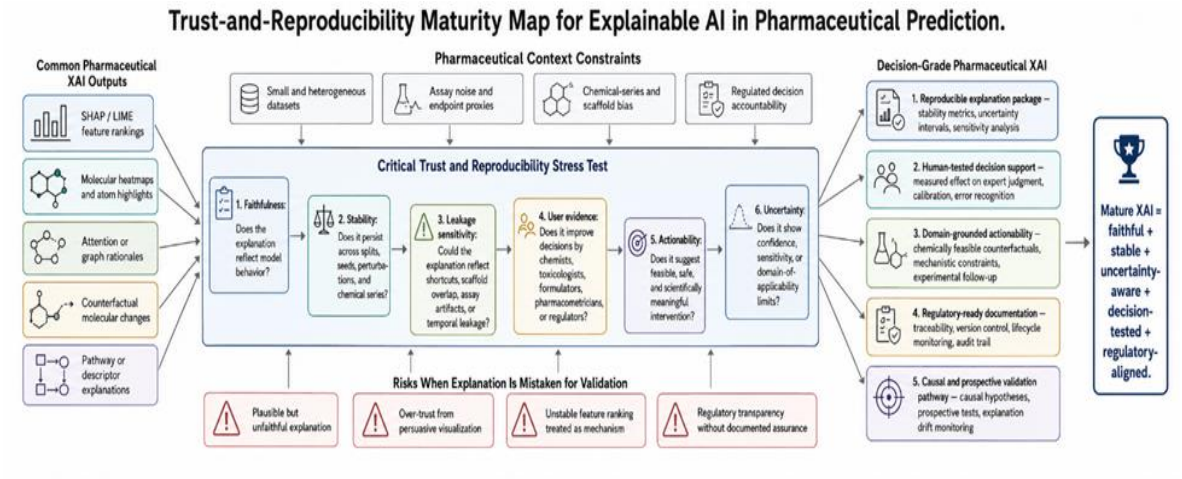

Figure 1. Trust-and-Reproducibility Maturity Map for Explainable AI in Pharmaceutical Prediction.

Proposed Dimensions

A useful evaluation framework for pharmaceutical XAI should include faithfulness, stability, actionability, user comprehension, uncertainty communication, and regulatory alignment [4, 13]. Faithfulness asks whether the explanation reflects the model's actual reasoning, while stability asks whether similar models or perturbations produce similar explanatory claims [9]. Actionability asks whether the explanation supports a scientifically feasible intervention, which is especially important in formulation, ADMET, and molecular optimization contexts [18, 24]. Regulatory alignment asks whether the explanation contributes to a defensible evidence package rather than merely improving the appearance of transparency [30, 31].

Table 3 presents a maturity matrix that distinguishes exploratory explanation artifacts from decision-grade pharmaceutical XAI requirements based on faithfulness, reproducibility, actionability, user validation, uncertainty communication, and regulatory alignment.

Table 3. Conceptual Maturity Matrix for Explainable AI in Pharmaceutical Prediction

XAI maturity dimension	Exploratory XAI practice	Decision-support XAI practice	Decision-grade pharmaceutical XAI standard	Evidence gap identified in the review
Explanation purpose	Produces visually or numerically interpretable outputs such as SHAP rankings, heatmaps, or attention maps.	Links explanation outputs to a specific scientific decision, such as compound triage, ADMET prioritization, formulation adjustment, or safety review.	Defines the intended user, decision context, acceptable uncertainty, and consequence of acting on the explanation before model deployment.	Many studies imply usefulness without specifying the decision the explanation is meant to improve.
Faithfulness to model behavior	Assumes that feature importance or highlighted substructures represent the model's reasoning.	Uses perturbation tests, ablation, counterfactual checks, or explanation-method comparisons to evaluate whether explanations track model behavior.	Demonstrates that the explanation remains faithful under realistic pharmaceutical data conditions, including correlated descriptors, scaffold effects, and endpoint noise.	Fidelity testing remains inconsistent, especially for molecular heatmaps, attention mechanisms, and graph rationales.
Reproducibility and stability	Reports a single explanation from one trained model or one data split.	Tests explanation consistency across random seeds, train–test splits, descriptor sets, and related chemical series.	Provides explanation stability metrics, uncertainty intervals, and sensitivity analysis as part of the model evidence package.	Pharmaceutical XAI papers often benchmark prediction performance more rigorously than explanation reproducibility.
Data-leakage protection	Treats leakage as a predictive validation problem only.	Examines whether explanation outputs may reflect scaffold overlap, duplicate compounds, assay-batch artifacts, or temporal leakage.	Documents leakage-control procedures and verifies that explanations do not reinforce invalid shortcuts learned by the model.	Leakage can make explanations appear chemically meaningful even when the predictive model is invalid.
Human-centered validation	Uses expert plausibility checks or visual inspection.	Conducts task-based studies with medicinal chemists, toxicologists, formulators, pharmacometricians, or regulatory reviewers.	Measures whether explanations improve decision accuracy, trust calibration, error recognition, time-to-decision, and resistance to misleading explanations.	Controlled user studies in pharmaceutical XAI remain scarce despite frequent claims about trust.

Actionability	Identifies influential features, descriptors, atoms, pathways, or covariates.	Connects explanations to feasible changes in molecules, formulations, doses, process parameters, or follow-up experiments.	Demonstrates that recommended actions are chemically realistic, biologically plausible, safe, and within the model's domain of applicability.	Many explanations answer "what influenced the prediction" but not "what should be done next."
Uncertainty communication	Presents explanation outputs as fixed and visually authoritative.	Reports uncertainty around predictions and, where possible, around explanations.	Communicates when an explanation is unstable, weakly supported, outside domain, or insufficient for decision-making.	Most studies do not report explanation uncertainty or confidence in explanatory claims.
Regulatory alignment	Adds explanation plots to model results as transparency artifacts.	Links explanation outputs to documentation, traceability, validation, and risk-management needs.	Embeds XAI into lifecycle governance, audit trails, version control, monitoring of explanation drift, and decision accountability.	Current XAI practice often falls short of regulatory expectations for documented scientific assurance.

Minimum Viable XAI for Pharma

Minimum viable XAI for high-stakes pharmaceutical decisions should demonstrate that explanations are reproducible, understandable to the intended expert users, linked to the decision being made, and accompanied by uncertainty or sensitivity analysis [13, 14]. Studies that merely show a plausible SHAP plot or molecular heatmap should be treated as exploratory unless they test explanation robustness and user impact [7, 9]. For drug safety, cardiotoxicity, and adverse event prediction, this standard is especially important because an attractive explanation may influence risk perception even when the model is poorly calibrated or trained on biased data [20, 21, 23]. A minimum viable standard would not eliminate uncertainty, but it would prevent explanation from being mistaken for validation.

A Call for Collaborative Benchmarking

Community-driven benchmarking is needed because isolated case studies cannot establish whether XAI methods generalize across endpoints, chemical series, institutions, or regulatory contexts [4, 9]. Public benchmarks should include molecular property prediction, ADMET, formulation, pharmacometric, and safety tasks, along with predefined tests for explanation stability and user comprehension [14, 23, 24]. Graph neural network and counterfactual approaches would especially benefit from shared benchmarks because their explanations are often visually compelling but difficult to compare objectively [17, 18, 22]. Collaborative benchmarking would also help separate methods that are merely interpretable in presentation from those that are scientifically reliable.

Future Directions and Grand Challenges

Toward Causally Grounded Explanations

The most important scientific shift is from correlational feature attribution toward causally grounded explanation [5, 11]. Pharmaceutical decisions often require knowing what will happen if a molecule, excipient, dose, or process parameter is changed, but SHAP and related methods generally describe associations learned by the model rather than causal mechanisms [13, 24]. Domain-augmented approaches that incorporate pathways, mechanisms, or chemical constraints are a step forward, yet they still need stronger causal validation before they can support intervention [16, 23]. Future XAI should therefore integrate causal inference, mechanistic modeling, and experimental feedback rather than treating model interpretation as an endpoint in itself.

Human-AI Co-Decision Research

The field urgently needs controlled studies testing whether XAI improves decisions made by medicinal chemists, toxicologists, formulators, pharmacometricians, and regulatory reviewers [26]. Such studies should measure not only user satisfaction, but also decision accuracy, calibration of trust, time-to-decision, error recognition, and resistance to misleading explanations [7, 27]. Pharmaceutical XAI has relied too heavily on expert plausibility checks, which are informative but insufficient for demonstrating behavioral benefit [6, 25]. Human-AI co-decision research would directly test whether explanations help experts act better or merely feel better informed.

Integration with Model Lifecycle Management

Future pharmaceutical XAI must be integrated with model lifecycle management, including monitoring of model drift, explanation drift, version control, and retraining effects [30, 31]. This is critical for manufacturing, formulation, and clinical pharmacology settings where data distributions may shift as processes, populations, or measurement systems change [14, 24]. Explanation drift should be treated as a model-risk signal because a model may preserve average predictive performance while changing the reasons behind its predictions [9]. Robust XAI will therefore require governance procedures that track explanations over time, not just model outputs.

Table 4 consolidates the major failure modes through which pharmaceutical XAI can generate misplaced confidence and pairs each risk with corrective standards for stronger, decision-relevant, and reproducible manuscripts.

Table 4. Critical Failure Modes and Corrective Standards for Pharmaceutical XAI

Failure mode	How it appears in pharmaceutical XAI papers	Why it threatens trust or reproducibility	Pharmaceutical example context	Corrective standard for stronger manuscripts
Plausibility mistaken for validity	Authors show feature rankings, substructure highlights, or pathway diagrams that appear consistent with domain knowledge.	A plausible explanation may still be unfaithful, unstable, biased, or driven by leakage.	A molecular heatmap highlights a known toxicophore, but the model may rely on scaffold artifacts or assay-source bias.	Pair qualitative plausibility with perturbation tests, ablation analysis, stability checks, and comparison across explanation methods.
Post-hoc explanation treated as transparency	Black-box models are presented as interpretable because SHAP, LIME, attention, or graph explanations are attached after training.	Post-hoc explanation does not guarantee transparent reasoning or regulatory sufficiency.	A deep ADMET model reports SHAP values but does not justify why a simpler interpretable model was inadequate.	Require authors to compare against interpretable baselines and justify the use of black-box models when decision stakes are high.
Unstable explanations treated as mechanistic insight	A single feature ranking is interpreted as evidence of chemical, biological, or clinical mechanism.	Explanations can change across training seeds, descriptor sets, data splits, or nearby chemical series.	A descriptor appears important for solubility prediction in one split but disappears after scaffold-based splitting.	Report explanation stability across resampling, random seeds, model versions, and chemically meaningful subgroups.
Leakage-amplified interpretability	Explanations appear scientifically coherent because the model has learned invalid shortcuts.	Leakage can produce both high predictive performance and convincing but false explanatory narratives.	A drug-safety model learns duplicated compound families or post-outcome coding artifacts, then attributes risk to misleading variables.	Use temporal, scaffold, external, and leakage-aware validation before interpreting explanation outputs.
User satisfaction mistaken for calibrated trust	Studies assume that understandable explanations increase appropriate trust.	Explanations may increase confidence even when predictions or explanations are wrong.	A medicinal chemist may accept a visually persuasive atom-level rationale without knowing that the explanation is unstable.	Measure trust calibration, error recognition, decision accuracy, and resistance to misleading explanations in controlled user tasks.
Actionability gap	Explanations identify influential variables but do not indicate feasible interventions.	Decision-makers need to know whether changing a feature is possible, safe, causal, and generalizable.	A model identifies lipophilicity as influential but does not indicate whether modification will preserve potency, permeability, or safety.	Evaluate counterfactuals and recommendations using chemical feasibility, biological plausibility, synthesis constraints, and experimental follow-up.
Uncertainty-free explanation	Explanation plots are presented without confidence intervals, sensitivity ranges, or domain-of-applicability warnings.	Users may overinterpret weak explanatory claims as stable knowledge.	A SHAP plot for a sparse cardiotoxicity dataset appears definitive despite endpoint noise and limited external validation.	Report prediction uncertainty, explanation uncertainty, sensitivity to perturbation, and out-of-domain flags.
Regulatory checkbox XAI	Explanation is included as a presentational add-on rather than as part of governance.	Visual transparency alone does not satisfy requirements for validation, documentation, lifecycle control, or accountability.	A model intended for regulated safety review provides heatmaps but lacks version control, audit trail, or explanation-drift monitoring.	Integrate XAI with model-risk management, lifecycle monitoring, documentation standards, and predefined decision-use criteria.

Strengths and Limitations of This Review

Strengths

The main strength of this review is its critical focus on trust and reproducibility across the pharmaceutical XAI landscape, rather than treating explanation methods as automatically beneficial [1, 4]. By bringing together work on drug discovery, ADMET, adverse events, formulation, clinical pharmacology, and regulation, it highlights recurring weaknesses that are not visible within single-domain case studies [14, 20, 24]. The review also connects pharmaceutical XAI to broader critiques of post-hoc explanation, over-trust, and high-stakes interpretability [7, 8, 27]. This cross-cutting perspective supports a more demanding standard for evaluating whether XAI is truly fit for pharmaceutical decision-making.

Limitations

This review is qualitative and critical rather than systematic, so it may not capture every relevant paper, especially rapidly emerging grey literature, preprints, proprietary industry work, and regulatory case examples [29, 30]. Its emphasis on trust and reproducibility may understate the exploratory value of XAI in early discovery, where explanation can still be useful for hypothesis generation even if it is not yet decision-grade [11, 25]. The review also reflects a skeptical interpretation of post-hoc explanation methods, whereas some proponents may place greater weight on their practical utility in debugging and

scientific communication [3, 13]. Nevertheless, the limitations of the review do not weaken its central argument that pharmaceutical XAI requires stronger evidence before being treated as a source of justified trust.

Conclusion

XAI has become a widely adopted component of pharmaceutical prediction workflows, yet its ability to engender trust and ensure reproducibility remains largely unproven. The field has generated many explanation artifacts, but fewer demonstrations that those artifacts improve scientific decisions.

The critical gaps are a paucity of human-centered evaluation, instability of popular explanation methods, and a disconnect with regulatory expectations. These gaps matter because pharmaceutical prediction is not a low-stakes visualization exercise; it can influence decisions about efficacy, safety, quality, and patient risk.

Without rigorous validation and domain-specific standards, XAI risks becoming a compliance checkbox rather than a genuine enabler of safe, data-driven pharmaceutical decisions. Explanations that look persuasive but lack fidelity, stability, and actionability may undermine the very trust they are intended to create.

The field must pivot toward causally grounded, uncertainty-aware, and behaviorally tested explanation systems if it is to fulfill its promise. Only then can XAI mature from an interpretive accessory into a scientifically reliable component of pharmaceutical decision-making.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Qadri YA, Shaikh S, Ahmad K, Choi I, Kim SW, Vasilakos AV. Explainable artificial intelligence: a perspective on drug discovery. Pharmaceutics. 2025;17(9):1119.
2. Ding Q, Yao R, Bai Y, Da L, Wang Y, Xiang R, et al. Explainable artificial intelligence in the field of drug research. Drug Des Devel Ther. 2025;4501-16.
3. Lavecchia A. Explainable artificial intelligence in drug discovery: bridging predictive power and mechanistic insight. Wiley Interdiscip Rev Comput Mol Sci. 2025;15(5):e70049.
4. Ponzoni I, Páez Prosper JA, Campillo NE. Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. Wiley Interdiscip Rev Comput Mol Sci. 2023;13(6):e1681.
5. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh CY, et al. From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. J Chem Inf Model. 2023;63(24):7617-27.
6. Kırboğa KK, Abbasi S, Küçüksille EU. Explainability and white box in drug discovery. Chem Biol Drug Des. 2023;102(1):217-33.
7. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50.
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.
9. Jiménez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. J Chem Inf Model. 2022;62(2):274-83.
10. Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. Chem Sci. 2022;13(13):3697-705.
11. Rodriguez-Perez R, Bajorath J. Explainable machine learning for property predictions in compound optimization: miniperspective. J Med Chem. 2021;64(24):17744-52.
12. Preuer K, Klambauer G, Rippmann F, Hochreiter S, Unterthiner T. Interpretable deep learning in drug discovery. In: Explainable AI: interpreting, explaining and visualizing deep learning. Cham: Springer International Publishing; 2019. p. 331-45.
13. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clin Transl Sci. 2024;17(11):e70056.
14. Brooks L, Harun R, Jin JY, Lu J. Shap-Cov: An Explainable Machine Learning Based Workflow for Rapid Covariate Identification in Population Modeling. CPT Pharmacometrics Syst Pharmacol. 2025;14(8):1322-31.
15. König C, Vellido A. Understanding predictions of drug profiles using explainable machine learning models. BioData Min. 2024;17(1):25.

16. Wang Y, Yu X, Gu Y, Li W, Zhu K, Chen L, et al. XGraphCDS: an explainable deep learning model for predicting drug sensitivity from gene pathways and chemical structures. Comput Biol Med. 2024;168:107746.
17. Wang C, Kumar GA, Rajapakse JC. Drug discovery and mechanism prediction with explainable graph neural networks. Sci Rep. 2025;15(1):179.
18. Amara K, Rodríguez-Pérez R, Jiménez-Luna J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. J Cheminform. 2023;15(1):67.
19. Kha QH, Le VH, Hung TN, Nguyen NT, Le NQ. Development and validation of an explainable machine learning-based prediction model for drug–food interactions from chemical structures. Sensors. 2023;23(8):3962.
20. Dey S, Luo H, Fokoue A, Hu J, Zhang P. Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics. 2018;19(Suppl 21):476.
21. Zhou F, Uddin S. Interpretable drug-to-drug network features for predicting adverse drug reactions. In: Healthcare. 2023;11(4):610.
22. Zhou F, Khushi M, Brett J, Uddin S. Graph neural network-based subgraph analysis for predicting adverse drug events. Comput Biol Med. 2024;183:109282.
23. Mukherjee S, Swanson K, Walther P, Shivnaraine RV, Leitz J, Pang PD, et al. ADMET-AI enables interpretable predictions of drug-induced cardiotoxicity. Circulation. 2025;151(3):285-7.
24. Ye Z, Yang W, Yang Y, Ouyang D. Interpretable machine learning methods for in vitro pharmaceutical formulation development. Food Front. 2021;2(2):195-207.
25. Rudrapal M, Kirboga KK, Abdalla M, Maji S. Explainable artificial intelligence-assisted virtual screening and bioinformatics approaches for effective bioactivity prediction of phenolic cyclooxygenase-2 (COX-2) inhibitors using PubChem molecular fingerprints. Mol Divers. 2024;28(4):2099-118.
26. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI conference on human factors in computing systems. 2020. p. 1-14.
27. Lakkaraju H, Bastani O. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. p. 79-85.
28. Liu K, Sun X, Jia L, Ma J, Xing H, Wu J, et al. Chemi-net: a molecular graph convolutional network for accurate drug property prediction. Int J Mol Sci. 2019;20(14):3389.
29. Zhang K, Yang X, Wang Y, Yu Y, Huang N, Li G, et al. Artificial intelligence in drug development. Nat Med. 2025;31(1):45-59.
30. Liu Q, Huang R, Hsieh J, Zhu H, Tiwari M, Liu G, et al. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. Clin Pharmacol Ther. 2023;113(4):771-4.
31. Singh R, Paxton M, Auclair J. Regulating the AI-enabled ecosystem for human therapeutics. Commun Med. 2025;5(1):181.