Pharmacophore

ISSN-2229-5402

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

WHAT MAKES AN ARTIFICIAL INTELLIGENCE EXPLANATION PHARMACOLOGICALLY USEFUL BEYOND ATTRACTIVE HEATMAPS AND FEATURE-IMPORTANCE RANKINGS?

Sofia Morales[1]*, Naveen Kumar[2], Elena Georgiou[3], Carlos Garcia[4]

1. ***Department of Pharmacologically Useful Explainability, Faculty of Pharmacy, National University of La Plata, La Plata, Argentina.***
2. ***Department of Explanation Validity and Biological Relevance, Faculty of Pharmacy, Indian Agricultural Research Institute, New Delhi, India.***
3. ***Department of Beyond-Heatmap Interpretation, Faculty of Pharmaceutical Sciences, University of Patras, Patras, Greece.***
4. ***Department of Feature-Importance and Decision Utility, Faculty of Pharmacy, Polytechnic University of Valencia, Valencia, Spain.***

ARTICLE INFO

Received:
29 September 2024
Received in revised form:
20 November 2024
Accepted:
23 November 2024
Available online:
28 December 2024

Keywords: Explainable artificial intelligence, Computational pharmacology, Drug discovery, Pharmacometrics, Mechanism alignment, Explanation faithfulness

ABSTRACT

Artificial intelligence explanations are increasingly presented in pharmaceutical research as molecular saliency maps, ranked descriptors, pathway-level importance scores, covariate effects, counterfactual examples, and local prediction narratives. Although these outputs can make model behaviour more visible, visibility alone does not establish that an explanation is pharmacologically meaningful, scientifically dependable, or suitable for influencing a development decision. This article addresses the unresolved distinction between visually interpretable model output and pharmacologically useful explanation. It develops a proposed explanation-evaluation theory in which usefulness is defined relationally: an explanation must answer a specified scientific question for an identified user, support a bounded pharmaceutical decision, remain faithful to the model being interpreted, demonstrate context-appropriate stability, align cautiously with relevant chemical, biological, or pharmacometric knowledge, and communicate uncertainty without converting association into mechanism or prediction into experimental confirmation. The central contribution is the proposed Pharmacological Usefulness Test for AI Explanations, a non-compensatory evaluation construct in which visual attractiveness, predictive performance, user satisfaction, or apparent mechanistic plausibility cannot offset failure in a critical evidentiary dimension. The theory is designed to support comparative evaluation across molecular, omics, and pharmacometric models while preserving domain-specific evidence requirements. It also identifies failure modes arising from unstable attribution, inappropriate explanatory questions, mechanistic overinterpretation, uncalibrated confidence, automation bias, and insufficient human contestability. The proposed test is conceptual rather than empirically validated and does not establish clinical utility, regulatory acceptability, universal thresholds, or deployment readiness. Its intended value is to organize future explanation studies around pharmaceutical questions, evidentiary boundaries, and decision consequences rather than the persuasive appearance of explanatory graphics.

To Cite This Article: Morales S, Kumar N, Georgiou E, Garcia C. What Makes an Artificial Intelligence Explanation Pharmacologically Useful beyond Attractive Heatmaps and Feature-Importance Rankings? Pharmacophore. 2024;15(6):76-86. https://doi.org/10.51847/QdjN456XBW

Introduction

Artificial intelligence models are now capable of producing visually compelling accounts of why a prediction was generated, including colored molecular substructures, ranked physicochemical descriptors, highlighted omics features, influential patient covariates, and local response surfaces. Such outputs can create an immediate impression of transparency because they replace an opaque numerical prediction with a recognizable scientific representation. Yet post-hoc intelligibility cannot by itself make an opaque model dependable for consequential pharmaceutical decisions; present explanation techniques may also generate unwarranted confidence without establishing scientific validity, and their practical value remains dependent on whether human users can understand and critically assess the relevance of the information presented [1–3]. The unresolved problem is therefore

Corresponding Author: Sofia Morales; Department of Pharmacologically Useful Explainability, Faculty of Pharmacy, National University of La Plata, La Plata, Argentina. E-mail: sofia.morales@agro.unlp.edu.ar.

not whether an artificial intelligence system can generate an explanation-like artifact, but whether that artifact supports a defensible pharmacological interpretation within a specified context of use.

This distinction matters because pharmaceutical artificial intelligence does not operate within a single homogeneous problem class. Machine-learning applications span molecular property prediction, virtual screening, target identification, biomarker discovery, image-based phenotyping, formulation analysis, pharmacokinetic prediction, exposure–response modeling, trial enrichment, and other development activities. These applications involve different data-generating processes, scientific assumptions, uncertainty structures, evidentiary standards, and consequences of error [4]. A molecular attribution used to nominate a functional group for analogue design cannot be evaluated in the same manner as a patient-level feature explanation used to interpret an exposure prediction or a covariate ranking used to reconsider a pharmacometric model. Consequently, no undifferentiated explanation metric can determine usefulness across the pharmaceutical sciences.

The translational problem is equally important. An explanation may accurately describe how a model distributes importance while remaining disconnected from the biological or development question that motivated the analysis. Translational usefulness requires an interpretable relationship among the analytical output, the relevant scientific evidence, and the decision that the output is expected to support [5]. For example, identifying a molecular fragment as influential does not establish target engagement, causal activity, synthesizability, safety, or therapeutic value. Similarly, highlighting an omics feature does not demonstrate that it is a modifiable disease mechanism, and ranking a clinical covariate does not show that it should determine dose adjustment. In each case, a representational output must be converted into a bounded explanatory proposition that can be questioned, tested, and interpreted against domain knowledge.

This article develops an original explanation-evaluation theory for determining when an artificial intelligence explanation may be described as pharmacologically useful. Its proposed central construct, the Pharmacological Usefulness Test for AI Explanations, treats usefulness as a conditional relationship among the explanation, the underlying model, the user, the explanatory question, the pharmaceutical decision, the relevant scientific knowledge, and the unresolved uncertainty. The test is deliberately non-compensatory: strong predictive performance, visual clarity, mechanistic familiarity, or favorable user reaction cannot compensate for failed faithfulness, scientifically implausible instability, unsupported mechanism claims, inadequate uncertainty representation, or mismatch with the intended decision. The contribution is not presented as a validated score, regulatory standard, or deployment procedure. Rather, it provides a structured theory for designing, evaluating, comparing, and reporting explanations without confusing attractive visualization with pharmacological evidence.

The Difference between Visual Explanation and Pharmacological Explanation

A visual explanation is best understood as a representation of selected aspects of model behavior. It may show where a neural network attends, which inputs contribute to a prediction, how a response changes over a local feature range, or which examples resemble the case under examination. Explanation, however, is heterogeneous, selective, contrastive, and audience-dependent; heatmaps and feature rankings constitute only a subset of available explanatory forms, and different technical methods expose different properties of a model that require method-appropriate evaluation [6–8]. A saliency image may answer "where did the model concentrate?", while a feature ranking may answer "which encoded variables most affected this output under this analysis?" Neither automatically answers the pharmacological questions "why should this relationship occur?", "under what conditions should it persist?", "what alternative explanation is credible?", or "what follow-up action is justified?"

A pharmacological explanation must therefore do more than display importance. It must connect model behavior to a scientifically interpretable proposition about a molecular, biological, physiological, pharmacokinetic, pharmacodynamic, or development-relevant relationship. Scholarship on explainable artificial intelligence in drug discovery similarly emphasizes that useful explanation requires connection to chemically and biologically meaningful knowledge rather than termination at feature display [9]. This does not mean that every explanation must reveal a true causal mechanism. It means that the output should be formulated so that its scientific status is explicit. A substructure attribution may be described as a model-sensitive region associated with a prediction; it should not be relabeled as a pharmacophore, toxicophore, binding determinant, or causal structural mechanism without additional evidence. Pharmacological usefulness begins with disciplined claim classification.

The scientific representation used by the model also affects what can reasonably be interpreted. Chemically grounded descriptors can make learned relationships more accessible to chemical analysis, particularly when the representation preserves meaningful spatial or electronic information [10]. Nevertheless, chemically recognizable features are not self-validating. A descriptor may encode a plausible relationship while being correlated with an unobserved confounder, a dataset artifact, or another feature used by the model. Representation-level intelligibility must consequently be separated from explanatory validity. The relevant question is not simply whether a medicinal chemist recognizes the highlighted property, but whether the explanation faithfully characterizes the model, remains sufficiently stable, accords with the applicable scientific evidence, and is used within a claim boundary appropriate to the data and modeling task.

The proposed distinction can be summarized as follows: a visual explanation makes model output perceptible, whereas a pharmacological explanation makes a bounded scientific proposition examinable. The transition requires at least four changes. First, the displayed pattern must be translated into an explicit explanatory claim. Second, the claim must be tied to a user and decision. Third, it must undergo evaluation for faithfulness, stability, mechanism alignment, and uncertainty. Fourth, its permitted inference must be stated, including what the explanation does not establish. This distinction prevents the visual persuasiveness of an output from functioning as a substitute for pharmaceutical evidence. **Table 1** organizes the evidence,

constructs, and interpretation boundaries needed to develop the difference between visual explanation and pharmacological explanation within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for the difference between visual explanation and pharmacological explanation

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Visual attribution	What inputs, regions, examples, or encoded features does the explanation identify as influential?	Model-specific or post-hoc attribution, saliency, example-based explanation, or response visualization	Provides an inspectable representation of model behavior	Treating visibility, aesthetic clarity, or intuitive resemblance as evidence of validity	A visible pattern is an explanatory representation, not proof of pharmacological relevance
Explanatory proposition	What precise scientific statement is being made from the displayed output?	Explicit translation of model output into a falsifiable or contestable claim	Converts a graphic or ranking into a claim that can be evaluated	Allowing ambiguous language to shift between importance, association, mechanism, and causality	The proposition must state whether it concerns model behavior, association, hypothesis generation, mechanism, or decision support
Chemical or biological interpretability	Can the output be related to established molecular, biological, physiological, or pharmacological concepts?	Domain representations, prior knowledge, experimental literature, and mechanistic models	Supports scientific examination and generation of follow-up questions	Equating a recognizable feature with a causal determinant or validated mechanism	Plausibility may support interpretation but does not alone confirm causality or experimental truth
Model faithfulness	Does the explanation accurately reflect the behavior of the model it claims to explain?	Perturbation analysis, model-specific checks, counterfactual consistency, or comparison with known model structure	Establishes whether the representation describes the model rather than merely appearing reasonable	Producing a plausible narrative that is weakly connected to the actual prediction process	A pharmacologically plausible explanation that is unfaithful to the model is not a valid explanation of that model
Stability	Does the explanation remain appropriately consistent under scientifically irrelevant changes and respond to meaningful changes?	Repeated fitting, input perturbation, representation variation, and sensitivity assessment	Distinguishes robust explanatory structure from accidental attribution	Assuming either complete invariance or uncontrolled variability is acceptable	Stability must be evaluated relative to the scientific invariances and sensitivities expected in the application
Mechanism alignment	Is the explanatory proposition compatible with relevant chemical, biological, physiological, or pharmacometric knowledge?	External scientific knowledge and, where available, experimental or mechanistic evidence	Tests whether the output can be interpreted within a pharmacological knowledge system	Converting consistency with prior knowledge into confirmation of a mechanism	Alignment indicates compatibility or hypothesis value, not causal validation
Uncertainty	What is unknown about the prediction, explanation, data support, and scientific interpretation?	Predictive uncertainty, attribution variability, domain applicability, and evidentiary uncertainty	Limits the strength and consequence of the explanatory claim	Presenting a precise explanation around an uncertain prediction or unsupported extrapolation	Explanation precision must not exceed the certainty of the model, data, and supporting science
Decision boundary	What action, if any, may the explanation support?	Defined context of use, consequence analysis, and available corroborating evidence	Connects explanation to bounded pharmaceutical usefulness	Allowing an exploratory explanation to be treated as a recommendation, confirmation, or approval	Usefulness is conditional on the specified decision and does not establish clinical, regulatory, or deployment readiness

Users, Decisions, and Explanatory Questions in Pharmaceutical Research

Explanation quality cannot be assessed independently of the person expected to use the explanation and the purpose for which it is requested. Pharmaceutical users may include medicinal chemists, computational chemists, biologists, toxicologists, bioinformaticians, pharmacometricians, clinicians, statisticians, safety scientists, and development decision-makers. Each role brings different prior knowledge, evidentiary expectations, and authority to act. Explanation requirements should therefore be specified through the stakeholder, the intended purpose, and the evaluation objective before an explanation method is selected [11]. An explanation suitable for model debugging may be inappropriate for mechanistic hypothesis generation; an explanation intended to compare candidate compounds may be insufficient for interpreting a patient-level prediction. User-centeredness does not mean tailoring graphics merely to preference. It means defining what the user must understand, challenge, and decide. The explanatory question provides the link between user and method. Questions may be descriptive, contrastive, counterfactual, mechanistic, uncertainty-focused, or decision-oriented. "Which features contributed?" differs from "Why was compound A ranked above compound B?", "What minimal structural change would reverse the prediction?", "Would the explanation persist under an alternative molecular representation?", or "Which assumption is driving the projected exposure difference?" The form of explanation and the information accompanying it can alter perceived helpfulness and acceptance,

particularly when users are informed about model errors [12]. Perceived clarity or satisfaction therefore cannot serve as proof of faithfulness, correctness, or pharmaceutical value. User evaluation must examine whether the explanation enables an appropriate inference and whether users recognize its limitations, not merely whether they report that the explanation is understandable.

Decision consequence adds a further requirement. An exploratory explanation used to select a hypothesis for laboratory testing may tolerate greater uncertainty than an explanation used to exclude a candidate, revise a clinical model, or support an individualized treatment judgment. Evidence from decision studies shows that machine-learning recommendations can influence expert selections, making behavioral effects, reliance, and potential overreliance relevant components of explanation evaluation [13]. The central unit of analysis should consequently be the **user–question–decision triad**. For every explanation, investigators should specify who receives it, what exact question it addresses, what decision could change because of it, what evidence would justify that change, and what harm could follow from an incorrect interpretation. These specifications determine the required explanatory depth and the intensity of validation.

Method-specific choices must then be made within this triad. In drug-development applications, SHAP-based attribution depends on decisions concerning the background distribution, modeled output, feature dependence, aggregation level, and interpretation of global versus local effects [14]. These choices can materially alter what a ranking appears to say and must therefore be documented as part of the explanation rather than hidden as implementation detail. Pharmacometric contexts impose additional constraints because explanatory claims must respect structural model assumptions, longitudinal and hierarchical data-generating processes, covariate relationships, inferential objectives, and the distinction between predictive flexibility and mechanistic interpretation [15]. A useful explanation in this setting may need to identify whether an apparent covariate effect reflects model structure, data support, extrapolation, or a plausible physiological relationship. Across pharmaceutical domains, the correct explanatory method is thus not the most visually impressive one, but the one capable of answering the specified question without exceeding the evidence appropriate to the decision.

Faithfulness, Stability, Mechanism Alignment, and Uncertainty

Explanation evaluation is inherently multidimensional because different properties answer different validity questions. Faithfulness asks whether an explanation accurately reflects the model process it claims to describe; stability asks whether explanatory conclusions persist under changes that should be scientifically irrelevant and respond appropriately to changes that should matter; mechanism alignment asks whether the interpretation is compatible with relevant chemical, biological, physiological, or pharmacometric knowledge; and uncertainty asks how confidently the prediction and its explanation may be interpreted. These properties are related but not interchangeable. An explanation can be faithful to a model that learned a spurious relationship, stable around an unsupported extrapolation, mechanistically plausible but disconnected from the actual model, or visually uncertain while the underlying uncertainty remains uncharacterized. Explanation evaluation must therefore distinguish fidelity, robustness, comprehensibility, usefulness, and related concepts instead of treating “interpretability” as a single measurable quality [16].

Faithfulness should be tested through deliberate interventions on the model, inputs, representations, or explanatory procedure. Depending on the application, these tests may include feature ablation, controlled perturbation, counterfactual consistency, comparison with transparent reference models, parameter randomization, retraining, or evaluation against known model structure. Stability testing should similarly be conditional rather than absolute. Small changes in an irrelevant formatting variable, equivalent molecular encoding, or resampled training set should not produce arbitrary explanatory reversals, whereas a meaningful structural modification, pathway perturbation, or covariate change may appropriately alter the explanation. The growing evaluation literature shows why illustrative examples and subjective inspection are insufficient: explanation properties require operational definitions, explicit tests, and criteria matched to the claim being made [17]. A stable but systematically wrong explanation is not useful, and a faithful explanation may remain too unstable for the decision under consideration.

Uncertainty introduces at least four analytically distinct concerns: uncertainty in the model prediction, uncertainty produced by limited data support or distributional shift, variability in the explanation itself, and uncertainty in the scientific interpretation assigned to the output. Deep-learning uncertainty is multifaceted and cannot be replaced by a softmax value, confidence score, attribution magnitude, or sharply rendered heatmap [18]. Pharmaceutical explanation reports should therefore distinguish predictive uncertainty from explanation uncertainty and both from evidentiary uncertainty. The consequences of uncertainty must also be related to the proposed use. An uncertain structural attribution may still be useful for generating alternative compounds for testing, but it may be inadequate for excluding a candidate or asserting a toxicity mechanism. Uncertainty communication should be designed around the consequences and limits of the decision rather than appended as a generic warning [19].

Mechanism alignment is the most easily overstated component. It evaluates whether an explanation is consistent with established or reasonably hypothesized pharmacology, not whether the explanation has discovered or verified a mechanism. Alignment may be examined against medicinal-chemistry knowledge, binding evidence, pathway biology, physiological constraints, known pharmacokinetic relationships, experimental perturbations, or mechanistic models. Conflicts may reveal model failure, insufficient evidence, previously unrecognized biology, or limitations in the prior knowledge itself; they should trigger investigation rather than automatic rejection or acceptance. Within model-informed drug development, interpretation is conditioned by model assumptions, qualification evidence, uncertainty, and context of use [20]. The same discipline should

govern artificial intelligence explanations. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop faithfulness, stability, mechanism alignment, and uncertainty within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Faithfulness, stability, mechanism alignment, and uncertainty

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Predictive adequacy context	Model outputs, target definition, validation evidence, applicability domain	Establish whether the prediction being explained is sufficiently supported for the intended analysis	A bounded statement of predictive capability and known failure conditions	Strong aggregate performance may conceal subgroup, scaffold, assay, or temporal failure	Evaluate performance under conditions relevant to the intended pharmaceutical use
Model faithfulness	Model architecture, learned parameters, inputs, explanatory method	Determine whether the explanation tracks the actual behavior of the model	Evidence that identified features or relationships influence the model as claimed	A scientifically plausible narrative may be unfaithful to the model	Use perturbation, ablation, randomization, transparent comparators, or model-specific checks
Input stability	Equivalent encodings, minor perturbations, measurement variability	Test invariance to changes that should not alter the scientific conclusion	Consistent explanatory conclusions under scientifically irrelevant variation	The definition of "irrelevant" depends on the application and representation	Predefine plausible invariances and assess explanation sensitivity to each
Training stability	Resampled data, repeated fitting, alternative seeds, model classes	Determine whether the explanation reflects reproducible structure or one fitted instance	Distribution of explanatory conclusions across credible fitted models	Agreement can be low even when predictive performance is similar	Repeat model fitting and report explanatory variability, not only a selected example
Mechanism alignment	External chemical, biological, physiological, or pharmacometric knowledge	Examine compatibility between the explanation and relevant scientific relationships	A graded statement of consistency, conflict, or insufficient evidence	Prior knowledge may be incomplete, context-specific, or itself uncertain	Compare with independent evidence and state whether the test concerns plausibility or confirmation
Prediction uncertainty	Data density, model ensemble, posterior or predictive distribution, calibration evidence	Characterize uncertainty in the model output to which the explanation is attached	Context-specific uncertainty statement for the prediction	Confidence scores may be uncalibrated or insensitive to distribution shift	Assess calibration, applicability, and uncertainty behavior in the intended domain
Explanation uncertainty	Alternative explanation methods, perturbations, repeated fits, background distributions	Characterize variability in the explanatory output	Range or distribution of plausible attributions or relationships	A single explanation may conceal substantial methodological dependence	Compare credible analytical choices and report convergence or disagreement
Interpretive uncertainty	Scientific evidence, alternative hypotheses, confounding structure	Limit the inference that can be drawn from the explanation	Explicit statement of competing interpretations and unresolved questions	Association may be rhetorically converted into mechanism or causality	Require claim classification and identify the additional evidence needed for stronger inference
Decision-conditioned adequacy	User role, explanatory question, decision consequence, corroborating evidence	Determine whether the combined evidence is adequate for the proposed use	Conditional suitability, revision requirement, or rejection for the specified decision	Adequacy thresholds are context-dependent and not yet empirically standardized	Validate thresholds prospectively for defined users and decisions

Proposed Pharmacological Usefulness Test for AI Explanations

The proposed Pharmacological Usefulness Test for AI Explanations, abbreviated PUT-AI, evaluates whether a particular explanation is suitable for a particular pharmaceutical question and decision. It does not ask whether an explanation method is universally good. Instead, it asks whether the explanation satisfies six non-compensatory dimensions: decision–question fit, model faithfulness, scientifically plausible stability, mechanism alignment, uncertainty adequacy, and human interpretability and contestability. This proposal reflects the need to distinguish interpretation of a molecular prediction model from claims about chemical or causal truth [21]. PUT-AI begins with a written explanatory proposition and ends with a bounded decision-use statement. The test is failed at the outset when no identifiable user, question, decision, or permissible inference has been specified.

The first operational principle is that explanation robustness must be evaluated rather than inferred from graphical smoothness. Atom-coloring and related local structural interpretations can vary with perturbation design, model behavior, molecular context, and analytical implementation [22]. Under PUT-AI, a robustness test must therefore be selected according to the scientific claim. A medicinal-chemistry explanation might be challenged through matched molecular changes, alternative representations, repeated model fitting, or comparison across relevant local neighborhoods. The objective is not to demand identical colors for every representation, but to determine whether the conclusion used for decision-making survives scientifically credible analytical variation. A failure may indicate that the explanation remains exploratory, that the model or question requires revision, or that corroborating evidence is necessary.

The second principle is that mechanism alignment must be externalized. Attribution can support mechanistically informative hypotheses when model-identified features are compared with independent chemical or binding evidence [23]. However, the explanatory output is only one element of that comparison. PUT-AI therefore asks whether the proposed mechanism claim is

supported by evidence independent of the explanation, whether alternative mechanisms remain plausible, and whether the claim is phrased at the correct level. An attribution that repeatedly highlights a molecular region may justify a hypothesis about a binding determinant; it does not by itself establish binding mode, target engagement, causal activity, or therapeutic action. Mechanism alignment is passed only relative to the stated claim and evidence level, not because the explanation resembles a familiar pharmacological narrative.

The third principle is that outcomes should be reported as an evaluation profile rather than a composite usefulness score. Human-readable structure–property relationships may facilitate chemical interrogation, but their scientific value remains conditional on validation against model behavior and chemical knowledge [24]. PUT-AI therefore permits three criterion-level judgments: sufficient for the specified bounded use, conditional pending additional evidence or revision, and not sufficient for the specified use. A failure in faithfulness cannot be offset by strong user satisfaction; missing uncertainty cannot be compensated by mechanistic plausibility; and visual appeal cannot override instability. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop proposed pharmacological usefulness test for ai explanations within the article's central argument.

Table 3. Components, relationships, uncertainties, and validation needs within Proposed pharmacological usefulness test for AI explanations

Evaluation dimension	What must be tested	Suitable evidence or assessment	Failure signal	Interpretive limitation	Decision relevance
Decision–question fit	Whether the explanation answers the stated pharmaceutical question for the identified user and decision	Predefined user–question–decision statement; task analysis; comparison between requested and delivered information	Explanation is understandable but answers a different question or supports no identifiable action	User preference does not establish scientific or decision validity	Determines whether further evaluation is meaningful for the proposed use
Model faithfulness	Whether the explanation accurately represents the behavior of the model	Perturbation, ablation, randomization, counterfactual testing, transparent comparator, or model-specific analysis	Highlighted features can be removed or altered without the expected model response; explanation remains unchanged after model disruption	Faithfulness to a model does not establish that the model learned a valid scientific relationship	Required before the explanation can be used to interpret the model
Scientifically plausible stability	Whether the conclusion persists under irrelevant changes and responds to meaningful ones	Alternative encodings, repeated fitting, local perturbations, measurement variation, background distributions	Explanatory conclusion reverses under equivalent representations or arbitrary analytical choices	Complete invariance is neither expected nor desirable when scientific inputs genuinely change	Determines whether an explanation can reliably support comparison or follow-up
Mechanism alignment	Whether the proposition is compatible with relevant independent scientific evidence	Literature, structural biology, assay evidence, pathway knowledge, physiological constraints, mechanistic models	Mechanistic language is unsupported, contradictory evidence is ignored, or association is presented as causation	Alignment is compatibility, not confirmation; prior knowledge may be incomplete	Determines whether the explanation may support hypothesis generation or mechanistic investigation
Uncertainty adequacy	Whether predictive, explanatory, domain, and interpretive uncertainties are characterized	Calibration analysis, applicability assessment, explanation ensembles, sensitivity analysis, alternative hypotheses	Precise recommendation is presented despite extrapolation, unstable attribution, or unresolved interpretation	No universal uncertainty threshold applies across decisions	Limits the permissible consequence of action based on the explanation
Human interpretability and contestability	Whether the intended user can understand, question, reproduce, and override the explanation	Task-based user study; documentation review; error-detection exercise; challenge and override procedure	Users cannot identify assumptions, errors, alternatives, or conditions under which the explanation should be rejected	Satisfaction and perceived clarity are not substitutes for appropriate reasoning	Supports accountable use and prevents explanation-induced automation bias
Cross-evidence coherence	Whether model, explanation, experimental, and domain evidence converge without being treated as interchangeable	Triangulation across independent evidence sources and explicit contradiction analysis	One evidence source dominates despite material conflict with others	Convergence strengthens a bounded claim but does not establish universal truth	Supports prioritization of follow-up rather than final confirmation
Conditional outcome	Whether all critical requirements are adequate for the specified decision	Criterion-level evaluation profile with written justification	One critical failure is hidden by averaging or a composite score	Proposed categories require future empirical threshold validation	Produces a bounded conclusion: suitable, conditional, or unsuitable for the specified use

Figure 1 presents the pharmacological explanation stress test, showing how the article's key components and boundaries are connected within proposed pharmacological usefulness test for ai explanations.

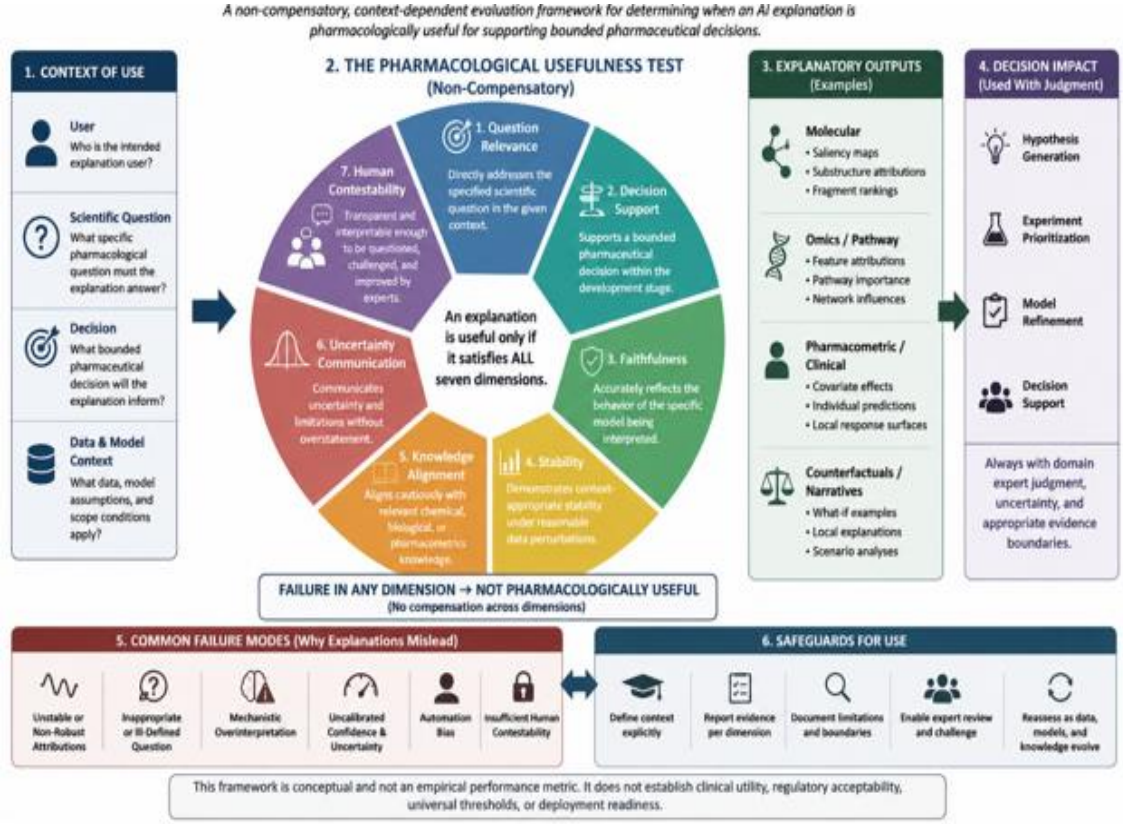

Figure 1. Pharmacological Explanation Stress Test.

The figure is an original conceptual synthesis that organizes the article's central contribution across difference between visual explanation and pharmacological explanation, users, decisions, and explanatory questions in pharmaceutical research, explanatory questions in pharmaceutical research, faithfulness, stability, mechanism alignment, and uncertainty, mechanism alignment, proposed pharmacological usefulness test for AI explanations. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Evaluation Across Molecular, Omics, and Pharmacometric Models

PUT-AI contains invariant questions, but its evidence must be adapted to the model class. In molecular property prediction, the representation, learning architecture, task definition, data partitioning strategy, and evaluation setting can alter what relationships a model learns [25]. Molecular explanations should therefore be assessed across chemically meaningful neighborhoods and alternative representations where appropriate. A highlighted atom or fragment should be tested against matched structural changes, known activity cliffs, scaffold transitions, stereochemical distinctions, and applicability boundaries. When the explanation is intended to guide analogue design, it should also distinguish model-sensitive structural features from claims about synthesis, physicochemical suitability, selectivity, metabolism, toxicity, or developability. A molecular explanation can support prioritization of experiments, but it cannot substitute for them.

Omics explanations require a different evidentiary design. High dimensionality, correlated variables, batch effects, data-processing choices, multiple plausible biological levels, and limited external validation can make a feature ranking unstable or biologically ambiguous [26]. A gene, transcript, protein, metabolite, or pathway may appear important because it is a proxy for a correlated process rather than an independent driver. Evaluation should therefore examine stability across preprocessing pipelines, resampling, feature-grouping strategies, cohorts, and biologically coherent aggregation levels. Pathway-level coherence may improve interpretability but still does not establish causal pathway activity. Explanations intended for biomarker discovery should specify whether they support prediction, subgroup description, biological hypothesis generation, or intervention prioritization, because each use requires different corroborating evidence.

Pharmacometric models raise a further distinction between prediction and parameterized scientific interpretation. Machine-learning approaches and conventional pharmacometric approaches may both predict pharmacokinetic observations, but they encode assumptions and expose relationships differently [27]. A feature attribution from a flexible predictor may identify variables associated with an exposure prediction without providing the structural interpretation offered by a parameterized pharmacokinetic model. Conversely, a mechanistic or population model can be interpretable in structure while remaining misspecified or inadequately supported by data. PUT-AI therefore asks whether the explanation concerns a prediction, a parameter, a structural assumption, an individual deviation, or a proposed physiological relationship. It should also determine

whether covariate effects are stable across sampling designs, patient distributions, model specifications, and clinically relevant ranges.

Hybrid and scientific machine-learning models combine explicit pharmacokinetic structure with learned components, creating both opportunities and new explanation boundaries. Neural-network terms can add flexibility to established pharmacokinetic models while preserving selected structural relationships [28]. Their explanations must identify which behavior arises from the mechanistic component, which arises from the learned component, and whether interactions between the two remain scientifically interpretable. Evaluation should test whether the learned component corrects a reproducible residual pattern, compensates for structural misspecification, or exploits dataset-specific correlations. Across molecular, omics, and pharmacometric applications, PUT-AI therefore preserves common validity dimensions while rejecting identical metrics or thresholds.

Failure Modes, Misuse, and Requirements For Human Interpretation

Artificial intelligence explanations can create risks that are not captured by predictive performance or visual quality. Machine-learning systems may generate unintended consequences, fail when transferred into real workflows or populations, and produce persuasive explanations that do not justify the confidence placed in them [29–31]. In pharmaceutical research, these risks include using a salient molecular fragment as though it were a validated mechanism, interpreting a correlated omics feature as a causal biomarker, treating an influential covariate as an established dose determinant, or assuming that a mechanistically styled hybrid model is correct because part of its structure is familiar. The explanation may be technically faithful yet operationally harmful when the decision context, alternative evidence, or consequence of error is ignored.

Fairness and representational adequacy create additional boundaries. Technical fairness adjustments cannot resolve all ethical and contextual problems arising from healthcare data, social structure, clinical practice, and unequal consequences [32]. A locally faithful explanation may reveal how a model uses a variable without establishing that using the variable is scientifically appropriate, ethically acceptable, or equitable. In pharmaceutical applications involving patients or population subgroups, explanation evaluation should therefore investigate data coverage, subgroup performance, proxy relationships, missingness, measurement practices, and differential consequences. A feature's influence may be accurately described while the underlying data-generating process remains biased or unsuitable for the proposed use.

Methodological misuse occurs when explanation settings are selected to produce a preferred narrative, when only visually coherent examples are reported, or when disagreement among explanation methods is concealed. Instability can be hidden by presenting a single fitted model, one background distribution, one molecular representation, or one local example. Uncertainty can be obscured by attaching a definitive narrative to a model operating beyond its applicability domain. Mechanistic overclaiming occurs when consistency with prior knowledge is treated as confirmation, whereas mechanistic underinterpretation occurs when a conflict with prior knowledge is dismissed without investigating model failure or potentially novel science. These failure modes should be treated as reportable evaluation outcomes rather than inconveniences to be removed from the presentation.

Human interpretation remains necessary but must be structured. Domain expertise can identify implausible chemistry, contradictory biology, unrealistic covariate behavior, or inappropriate decision consequences, yet expert intuition is also vulnerable to confirmation bias and automation effects. Human oversight should therefore include an explicit opportunity to challenge the explanation, inspect alternative explanations, review uncertainty, document disagreement, and decline the recommended inference. The responsible user should know what the explanation addresses, how it was generated, where it is unstable, what evidence supports its interpretation, and what additional work is required before action. **Table 4** organizes the evidence, constructs, and interpretation boundaries needed to develop failure modes, misuse, and requirements for human interpretation within the article's central argument.

Table 4. Evaluation, implementation, and research priorities arising from What Makes an Artificial Intelligence Explanation Pharmacologically Useful beyond Attractive Heatmaps and Feature-Importance Rankings?

Research or implementation priority	Unresolved gap	Required methodological work	Evidence needed	Responsible actors	Expected contribution
Decision-specific explanation protocols	Explanations are often selected before users, questions, and decisions are defined	Develop protocols that begin with a user–question–decision specification and permissible inference	Task analyses, decision studies, consequence mapping, and domain-specific case evaluations	Pharmaceutical scientists, model developers, decision scientists, intended users	Prevents mismatch between explanation format and pharmaceutical purpose
Faithfulness benchmarks	Many explanation methods are demonstrated visually without showing that they track model behavior	Construct model-specific and model-agnostic faithfulness tests for pharmaceutical tasks	Perturbation studies, transparent reference models, simulated ground-truth mechanisms where appropriate	Machine-learning researchers, chemoinformaticians, bioinformaticians, pharmacometricians	Distinguishes plausible narratives from valid model explanations
Stability and reproducibility testing	Explanations may depend on model seed, encoding, background sample, preprocessing, or local perturbation	Define scientifically relevant invariance and sensitivity tests for each domain	Repeated fitting, alternative representations, preprocessing	Model developers, data scientists, domain experts	Reveals whether explanatory conclusions are robust enough for the intended decision

			comparisons, external datasets		
Mechanism-alignment evaluation	Familiar scientific language can convert association into unsupported mechanism claims	Develop graded alignment criteria separating compatibility, hypothesis support, experimental confirmation, and causality	Independent assays, structural evidence, pathway perturbations, physiological constraints, mechanistic models	Experimental scientists, medicinal chemists, systems pharmacologists, pharmacometricians	Preserves the distinction between explanation and mechanistic evidence
Explanation-uncertainty reporting	Predictive uncertainty is frequently reported without uncertainty in attribution or interpretation	Develop explanation ensembles, sensitivity profiles, and reporting standards for interpretive uncertainty	Calibration analysis, applicability assessment, repeated explanations, competing scientific hypotheses	Statisticians, machine-learning researchers, pharmaceutical modelers	Prevents false precision and supports appropriately bounded use
Cross-domain evaluation studies	Shared explanation language conceals major differences among molecular, omics, and pharmacometric models	Compare invariant PUT-AI dimensions using domain-specific tests rather than one universal benchmark	Multi-domain case studies, preregistered evaluation plans, task-specific external evidence	Interdisciplinary pharmaceutical research teams	Determines which components generalize and which require domain-specific adaptation
Human contestability studies	User satisfaction is often measured without testing error detection, challenge, or override	Evaluate whether users identify misleading explanations and respond appropriately to uncertainty	Controlled user studies, error-seeded tasks, override behavior, qualitative reasoning analysis	Human-factors researchers, domain users, ethicists	Measures whether explanations improve appropriate reasoning rather than passive acceptance
Documentation and audit infrastructure	Explanation generation choices and disagreements are often insufficiently recorded	Create versioned explanation records, model cards, evidence logs, and challenge histories	Reproducible workflows, metadata standards, audit trails, governance evaluation	Research institutions, sponsors, software teams, quality and governance groups	Supports reproducibility, accountability, and post hoc investigation
Staged implementation boundaries	Conceptual frameworks may be treated as operational tools before validation	Define exploratory, evaluative, prospective, and decision-influencing stages with explicit transition evidence	Prospective studies, workflow evaluations, external validation, consequence monitoring	Sponsors, investigators, governance committees, intended decision owners	Prevents premature deployment and clarifies readiness requirements

Limitations and Boundary Conditions

The proposed theory is a conceptual synthesis rather than an empirically validated measurement instrument. PUT-AI does not supply universal thresholds for faithfulness, stability, mechanism alignment, uncertainty adequacy, or human contestability. The relevance and acceptable strength of each dimension depend on the data, model, explanatory question, user, and consequence of the decision. The proposed conditional outcomes should therefore not be interpreted as certification categories. Empirical studies are required to determine whether the dimensions can be measured reliably, whether critical failures can be identified consistently, and whether their use improves scientific reasoning or pharmaceutical decisions.

The available evidence also spans heterogeneous disciplines. General explainable-artificial-intelligence research provides evaluation concepts, pharmaceutical studies provide domain examples, and pharmacometric or translational scholarship supplies context-of-use principles. These literatures do not establish that one evaluation procedure will perform consistently across molecular graphs, high-dimensional omics data, population pharmacokinetic models, or hybrid scientific machine-learning systems. Some proposed relationships are therefore theoretical extensions rather than direct findings from a single source. The article deliberately retains this distinction: evidence supports the need for multidimensional, decision-conditioned evaluation, while the exact integration of the dimensions into PUT-AI remains proposed.

Mechanism alignment presents a particularly important boundary. An explanation may be compatible with known chemistry or biology and still be wrong, while an apparent conflict may reflect model error, incomplete knowledge, context dependence, or a potentially novel observation. PUT-AI cannot adjudicate these possibilities without external investigation. It also cannot establish synthesizability, developability, safety, clinical benefit, regulatory acceptability, or therapeutic value from explanation quality. Human review likewise does not guarantee correctness. The proposed test can structure scrutiny and make assumptions visible, but it cannot eliminate epistemic uncertainty, institutional bias, inadequate data, or misuse by decision-makers.

Research Agenda and Implementation Priorities

The first research priority is to test whether PUT-AI dimensions can be operationalized reproducibly in specific pharmaceutical tasks. Studies should begin with narrow contexts, such as local molecular explanations for analogue prioritization, pathway explanations for biomarker hypothesis generation, or covariate explanations for pharmacometric model review. Investigators should predefine the user, explanatory question, decision, acceptable inference, failure conditions, and external evidence. Comparisons should test whether PUT-AI-based evaluation identifies misleading explanations that would be missed by visual inspection, predictive metrics, or user satisfaction alone. The objective should not be to prove one universal score, but to establish criterion-specific validity and decision consequences.

A second priority is the creation of explanation-evaluation infrastructure. Pharmaceutical artificial intelligence studies should preserve model versions, training data provenance, preprocessing decisions, explanation settings, background distributions, random seeds, alternative representations, uncertainty analyses, and records of domain-expert challenge. Reporting should distinguish the prediction being explained, the explanatory method, the proposition inferred, the supporting evidence, and the prohibited inference. Repositories of failed, unstable, contradictory, or non-actionable explanations would be particularly valuable because publication practices currently favor visually coherent examples. Cross-domain benchmark collections should include explanation failure cases rather than only prediction tasks.

Implementation should proceed in stages. The initial stage should use PUT-AI prospectively as a research-design and reporting aid without allowing explanations to determine consequential actions. A subsequent evaluative stage should examine reproducibility, external validity, user reasoning, and failure detection in realistic pharmaceutical workflows. Only after context-specific evidence is available should an explanation influence bounded decisions, and even then it should remain subject to documentation, independent evidence, human challenge, and post-use monitoring. This staged approach prevents a conceptual theory from being mistaken for an operational standard while creating a pathway for rigorous future assessment.

Conclusion

An artificial intelligence explanation becomes pharmacologically useful not when it is visually attractive, intuitively familiar, or attached to a high-performing model, but when it answers a defined scientific question for an identified user and bounded decision while surviving appropriate tests of faithfulness, stability, mechanism alignment, uncertainty, and human contestability. The proposed Pharmacological Usefulness Test for AI Explanations organizes these requirements as non-compensatory dimensions and distinguishes model interpretation from chemical truth, biological mechanism, causality, experimental confirmation, clinical utility, and regulatory acceptability. Its contribution is a theory of explanation evaluation rather than a validated score or deployment procedure. By requiring explicit explanatory propositions, external evidence boundaries, domain-specific tests, and transparent failure outcomes, the approach may help pharmaceutical researchers replace persuasive explanatory appearance with disciplined scientific interrogation. Its ultimate value depends on future empirical evaluation across molecular, omics, pharmacometric, and hybrid modeling contexts.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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. doi:10.1038/s42256-019-0048-x
2. 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). doi:10.1016/S2589-7500(21)00208-9
3. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(4). doi:10.1002/widm.1312
4. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
5. Terranova N, Venkatakrishnan K, Benincosa LJ. Application of machine learning in translational medicine: Current status and future opportunities. AAPS J. 2021;23(4):74. doi:10.1208/s12248-021-00593-x
6. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007
7. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012
8. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. Proc IEEE. 2021;109(3):247-78. doi:10.1109/JPROC.2021.3060483
9. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
10. Gallegos M, Vassilev-Galindo V, Poltavsky I, Martín Pendás Á, Tkatchenko A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. Nat Commun. 2024;15(1):4345. doi:10.1038/s41467-024-48567-9

11. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. doi:10.1016/j.jbi.2020.103655
12. Kenny EM, Ford C, Quinn M, Keane MT. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. Artif Intell. 2021;294:103459. doi:10.1016/j.artint.2021.103459
13. Jacobs M, Pradier MF, McCoy TH Jr, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. Transl Psychiatry. 2021;11(1):108. doi:10.1038/s41398-021-01224-x
14. 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). doi:10.1111/cts.70056
15. McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. Br J Clin Pharmacol. 2022;88(4):1482-99. doi:10.1111/bcp.14801
16. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion. 2021;76:89-106. doi:10.1016/j.inffus.2021.05.009
17. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Comput Surv. 2023;55(13s):295. doi:10.1145/3583558
18. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion. 2021;76:243-97. doi:10.1016/j.inffus.2021.05.008
19. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. 2019;1(1):20-3. doi:10.1038/s42256-018-0004-1
20. Marshall S, Madabushi R, Manolis E, Krudys K, Staab A, Dykstra K, et al. Model-informed drug discovery and development: Current industry good practice and regulatory expectations and future perspectives. CPT Pharmacometrics Syst Pharmacol. 2019;8(2):87-96. doi:10.1002/psp4.12372
21. Wellawatte GP, Gandhi HA, Seshadri A, White AD. A perspective on explanations of molecular prediction models. J Chem Theory Comput. 2023;19(8):2149-60. doi:10.1021/acs.jctc.2c01235
22. Sheridan RP. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: How robust is it? J Chem Inf Model. 2019;59(4):1324-37. doi:10.1021/acs.jcim.8b00825
23. McCloskey K, Taly A, Monti F, Brenner MP, Colwell LJ. Using attribution to decode binding mechanism in neural network models for chemistry. Proc Natl Acad Sci U S A. 2019;116(24):11624-9. doi:10.1073/pnas.1820657116
24. Wellawatte GP, Schwaller P. Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models. Commun Chem. 2025;8(1):11. doi:10.1038/s42004-024-01393-y
25. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
26. Toussaint PA, Leiser F, Thiebes S, Schlesner M, Brors B, Sunyaev A. Explainable artificial intelligence for omics data: A systematic mapping study. Brief Bioinform. 2024;25(1). doi:10.1093/bib/bbad453
27. Keutzer L, You H, Farnoud A, Nyberg J, Wicha SG, Maher-Edwards G, et al. Machine learning and pharmacometrics for prediction of pharmacokinetic data: Differences, similarities and challenges illustrated with rifampicin. Pharmaceutics. 2022;14(8):1530. doi:10.3390/pharmaceutics14081530
28. Valderrama D, Ponce-Bobadilla AV, Mensing S, Fröhlich H, Stodtmann S. Integrating machine learning with pharmacokinetic models: Benefits of scientific machine learning in adding neural network components to existing PK models. CPT Pharmacometrics Syst Pharmacol. 2024;13(1):41-53. doi:10.1002/psp4.13054
29. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797
30. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
31. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-6. doi:10.1126/science.abg1834
32. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health. 2020;2(5). doi:10.1016/S2589-7500(20)30065-0