Pharmacophore

ISSN-2229-5402

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

FROM MOLECULAR ACCURACY TO PHARMACOLOGICAL MEANING: AN EVIDENCE-CONSTRAINED THEORY OF INTERPRETABLE PREDICTION IN EARLY DRUG DISCOVERY

Ethan Blake[1], Isabella Russo[2]*, Klaus Richter[1], Maria Fernandez[3]

1. ***Department of Computational Chemistry and Molecular Informatics, Faculty of Pharmaceutical Sciences, University of Florida, Gainesville, United States.***
2. ***Department of Pharmacological Interpretation and Translational Modeling, Faculty of Pharmacy, University of Turin, Turin, Italy.***
3. ***Department of Drug Discovery and Evidence-Based Prediction, Faculty of Pharmacy, University of Barcelona, Barcelona, Spain.***

ARTICLE INFO

Received:
29 May 2024
Received in revised form:
02 October 2024
Accepted:
05 October 2024
Available online:
28 October 2024

Keywords: Interpretable machine learning, Molecular property prediction, Explainable artificial intelligence, Drug discovery, Molecular representation, Feature attribution

ABSTRACT

Machine-learning models can rank molecules, estimate molecular properties, and predict compound–target relations with increasing computational accuracy. Nevertheless, an accurate prediction does not necessarily reveal why a molecule behaves as predicted, whether the model has learned pharmacologically relevant information, or whether its explanation can support an experimental decision. This conceptual gap is especially consequential in early drug discovery, where model outputs are frequently interpreted across virtual screening, target assessment, lead optimization, and experimental prioritization. This Original Conceptual Theory Article develops evidence-constrained interpretability as a proposed theory for distinguishing molecular explanations that merely describe predictive behavior from interpretations that may support bounded pharmacological reasoning. The theory treats representation traceability, attribution faithfulness, independent evidence linkage, mechanistic-status qualification, uncertainty characterization, and decision-context alignment as interdependent conditions rather than interchangeable indicators of explanation quality. It further distinguishes benchmark performance from prospective usefulness, model attribution from biological mechanism, molecular plausibility from developability, and predictive confidence from calibrated uncertainty. Under the proposed theory, a molecular feature acquires pharmacological meaning only provisionally and only when the inferential step connecting that feature to a target, mechanism, or discovery decision is supported by evidence appropriate to the claim. The contribution is conceptual rather than empirically validated and does not establish causal explanation, clinical utility, regulatory acceptability, or operational readiness. Its principal implication is that interpretable molecular prediction should be evaluated as an evidence-governed scientific practice rather than as the production of visually plausible feature maps. This perspective may support more disciplined use of prediction and explanation in early discovery while identifying the experiments required to test, revise, or reject pharmacological interpretations.

To Cite This Article: Blake E, Russo I, Richter K, Fernandez M. From Molecular Accuracy to Pharmacological Meaning: An Evidence-Constrained Theory of Interpretable Prediction in Early Drug Discovery. Pharmacophore. 2024;15(5):59-68. https://doi.org/10.51847/xAGR5iorFC

Introduction

Machine learning has become embedded in several stages of pharmaceutical research, including molecular property prediction, compound prioritization, target-related modeling, toxicity assessment, and candidate generation. Its expanding role reflects the ability of computational models to extract regularities from chemical and biological data at scales that are difficult to examine manually. However, the scientific value of such models remains conditional on the quality, provenance, and relevance of the data from which their predictions are learned [1]. Computer-assisted drug discovery also encompasses heterogeneous tasks, representations, objectives, and validation strategies, meaning that a favorable result in one computational setting cannot be assumed to carry the same meaning in another [2]. The central problem is therefore not whether a model can produce an

Corresponding Author: Isabella Russo; Department of Pharmacological Interpretation and Translational Modeling, Faculty of Pharmacy, University of Turin, Turin, Italy. E-mail: isabella.russo@unito.it.

accurate molecular prediction, but what that prediction legitimately permits a scientist to infer about a molecule, target, mechanism, or experimental decision.

This distinction matters because early drug discovery depends on decisions made under substantial biological and chemical uncertainty. A model may assign a high probability of activity, predict a favorable property value, or rank a compound above alternatives without revealing whether the result arises from a transferable chemical relation, an assay-specific correlation, a data artifact, or a representation-dependent shortcut. Contemporary drug-design workflows consequently require more than improvements in architecture or benchmark performance: they require connections among model outputs, synthesis feasibility, experimental evidence, medicinal-chemistry reasoning, and iterative decision making [3]. Without those connections, computational accuracy may increase the efficiency of ranking while leaving the evidentiary basis of the ranking unresolved.

Explainable artificial intelligence has been proposed as one route toward greater transparency in molecular prediction. Attribution maps, attention scores, surrogate models, counterfactual examples, and related methods can expose features associated with a model output, but their pharmaceutical interpretation remains conditional on domain-sensitive validation [4]. The rapid adoption of deep learning across drug-discovery tasks has widened the range of predictions that can be generated, yet methodological breadth does not establish that a model has acquired mechanistic or pharmacological understanding [5]. A highlighted substructure may identify a feature used by the model, for example, while remaining silent about whether that feature contributes to binding, selectivity, permeability, toxicity, assay interference, or an unmeasured confounding relation. Interpretability thus becomes scientifically meaningful only when the inferential distance between model behavior and pharmacological claim is made explicit.

Molecular prediction and generation also require confidence assessment and medicinal-chemistry interpretation before they can guide discovery decisions [6]. This article proposes evidence-constrained interpretability as a conceptual theory for governing that transition. The theory does not treat interpretability as a single model property or visualization technique. Instead, it defines an interpretation as pharmacologically admissible only when the prediction object is specified, the molecular representation is traceable, the explanation is faithful to the model, the attributed feature is linked to independent evidence, the mechanistic status of the claim is declared, relevant uncertainty is characterized, and the interpretation is aligned with a specific decision context. The need for such a theory follows from convergent recognition that pharmaceutical machine learning is constrained by data suitability, representation, evaluation design, experimental closure, and workflow integration [1-3]. The proposed contribution is therefore an organizing theory for disciplined interpretation, not an empirically validated framework, universal standard, or deployment-ready decision system.

Why Predictive Accuracy Can Fail To Produce Pharmacological Understanding

Predictive accuracy describes correspondence between model outputs and observed labels under a specified evaluation design. Pharmacological understanding, by contrast, concerns whether the relations inferred from a prediction are scientifically relevant to molecular behavior, biological interaction, mechanism, or a defined discovery decision. These constructs may overlap, but they are not equivalent. Ligand-based benchmarks can reward recognition of chemical similarity patterns that recur across training and test sets, allowing a model to appear generalizable while primarily reproducing memorized chemical neighborhoods [7]. Structure-based virtual-screening evaluations are likewise sensitive to the composition of active and inactive compounds, including differences that may be unrelated to the intended protein–ligand interaction [8]. Accuracy can therefore be genuine with respect to the benchmark while remaining misleading with respect to the scientific question that the benchmark is assumed to represent.

The problem becomes more pronounced when hidden dataset regularities provide predictive shortcuts. Analyses of DUD-E have shown that deep-learning models can exploit benchmark-specific biases and achieve apparently strong performance without necessarily learning the intended physical determinants of molecular recognition [9]. Collectively, benchmark memorization, chemical-data bias, and hidden dataset artifacts demonstrate that predictive success may reflect the structure of an evaluation resource rather than transferable pharmacological knowledge [7-9]. This does not make benchmark evaluation useless. It means that a benchmark result must be interpreted as evidence about performance under a defined data-generating and splitting procedure, not as automatic evidence of mechanism, target engagement, prospective enrichment, or therapeutic relevance.

A second source of failure is mismatch between the model endpoint and the decision for which the prediction is used. A model trained to reproduce an assay label may learn relations associated with assay conditions, measurement technologies, compound availability, or historical selection practices. Even when such a model predicts held-out observations accurately, the output may not answer whether a compound binds through the proposed mode, retains activity in a different biological system, possesses suitable selectivity, or can be developed into a useful therapeutic candidate. Realistic impact from artificial intelligence in drug discovery consequently depends on appropriate problem formulation, validation, and integration into scientific workflows rather than on predictive performance alone [10]. Pharmacological understanding requires a defensible account of what the label represents, what alternative explanations remain plausible, and which additional evidence would discriminate among them.

A third source of failure lies in the evidence substrate itself. Chemical and biological datasets frequently combine measurements generated under different protocols, laboratories, concentration ranges, biological systems, annotation practices, and quality controls. Data availability is therefore not equivalent to data suitability, and a large dataset is not necessarily informative for the intended pharmacological question. The provenance, measurement context, and limitations of the source

data constrain what a model can infer and what a scientist may responsibly claim from its output [11]. Evidence-constrained interpretation must accordingly begin before model explanation: it begins with the definition of the endpoint, the suitability of the data, the possibility of leakage or confounding, and the relationship between the computational task and the intended discovery decision. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why predictive accuracy can fail to produce pharmacological understanding within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for why predictive accuracy can fail to produce pharmacological understanding

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Prediction-object specification	What endpoint, label, unit of analysis, and use case is actually being predicted?	Assay definition, label provenance, measurement process, model output, and applicability domain	Establishes the precise object to which accuracy applies	Treating a proxy endpoint as evidence of target engagement, efficacy, safety, or therapeutic value	Accurate prediction of a proxy does not establish the downstream pharmacological phenomenon
Benchmark validity	Does the evaluation test transferable learning or reward recurrence of chemical similarity and dataset structure?	Split design, scaffold distribution, temporal separation, duplicate control, and prospective testing	Distinguishes benchmark success from generalizable molecular prediction	Claiming generalization from random or chemically permissive splits	Performance is valid only for the evaluated distribution and testing procedure
Chemical-data bias	Are active and inactive classes distinguishable through unintended chemical or database artifacts?	Property distributions, analogue density, source composition, decoy construction, and sampling history	Identifies non-pharmacological signals that can drive prediction	Interpreting shortcut learning as recognition of molecular interaction determinants	Dataset separability does not prove learning of the intended physical or biological relation
Biological and assay context	What biological system, assay technology, protocol, and measurement conditions produced the labels?	Target format, cell background, concentration, readout, controls, and experimental provenance	Connects the computational label to the scientific phenomenon it represents	Generalizing an assay-specific association to mechanism, efficacy, or clinical effect	An assay label supports only claims compatible with its experimental context
Molecular applicability domain	Is the compound sufficiently represented by the training evidence for the prediction to be interpretable?	Chemical similarity, scaffold coverage, representation distance, and out-of-domain assessment	Qualifies whether a prediction and explanation can be relied upon for a particular molecule	Assigning confident meaning to unfamiliar chemistry	Model confidence within an unfamiliar region is not equivalent to validated applicability
Evidence hierarchy	What independent evidence supports the transition from prediction to association, hypothesis, mechanism, or confirmation?	Structural data, structure–activity relationships, target assays, mutagenesis, orthogonal experiments, and prospective testing	Prevents semantic escalation of model outputs	Describing attribution as causal or mechanistic evidence	The strength of the claim must not exceed the strength and independence of the supporting evidence
Decision-context alignment	Which concrete discovery decision is the interpretation intended to support?	Screening capacity, optimization objective, experimental alternatives, cost of error, and human expertise	Makes usefulness conditional on a defined scientific action	Treating generic intelligibility as decision value	An interpretation useful for triage may be inadequate for mechanistic inference or lead optimization
Experimental closure	Which observation could corroborate, revise, or falsify the interpretation?	Prospective assays, synthesis, structural studies, target engagement, negative results, and model updating	Converts interpretation into a testable scientific hypothesis	Using explanation as an endpoint rather than the beginning of investigation	A prediction becomes stronger evidence only through appropriate independent testing

Molecular Representations, Attribution Methods, and the Limits of Post Hoc Explanation

A molecular model cannot learn from a molecule in the abstract; it learns from a representation. Fingerprints encode predefined fragments and topological patterns, descriptor vectors summarize selected physicochemical properties, graphs represent atoms and bonds with local neighborhoods, string encodings express molecules as symbol sequences, and three-dimensional approaches may incorporate conformers or spatial interaction features. Each representation preserves some chemical distinctions while suppressing others. Performance across molecular-learning benchmarks varies with task definition, representation, split strategy, and evaluation metric, so no benchmark can establish universal utility for a representation or model family [12]. This dependence has direct interpretive consequences. An explanation can identify only relations that are expressible within the model's input and learned state. A model that lacks explicit stereochemical, conformational, protonation, assay, or target-context information cannot recover those factors through post hoc visualization.

Learned representations may capture task-relevant chemical regularities more flexibly than fixed descriptors, but their contents are distributed across parameters and may not correspond to concepts recognized by medicinal chemists. Comparative analyses of molecular property prediction demonstrate that representation choice influences both performance and the information made available to the model [13]. Evidence-constrained interpretability therefore requires representation traceability: a documented account of what was encoded, what was omitted, how molecules were standardized, which alternative encodings were plausible, and whether the explanation is robust to chemically equivalent representations. A substructure highlighted in one

encoding should not be assumed to constitute a representation-independent explanation. Disagreement across randomized strings, graph constructions, conformers, or preprocessing pipelines may indicate interpretive instability, but agreement alone still does not establish biological validity.

Attribution methods add a second inferential layer. They estimate how model outputs respond to features, perturbations, internal attention patterns, local surrogates, or contrastive molecular changes. Attribution has been used to examine whether neural models capture binding-related molecular patterns, showing that model-used features can sometimes correspond to chemically meaningful hypotheses [14]. Nevertheless, attribution remains an explanation of model behavior. It does not independently prove that a highlighted atom, bond, substructure, residue, or interaction is responsible for a biological mechanism. Correlated features may receive interchangeable importance; perturbations may create chemically invalid molecules; attention may not measure causal contribution; and visually plausible maps may remain unstable across models or training runs. The proposed theory therefore separates model faithfulness from pharmacological faithfulness. The former asks whether an explanation accurately reflects the predictor. The latter asks whether the resulting interpretation is supported by independent chemical, structural, target, mechanistic, or experimental evidence. Post hoc explanation may contribute to pharmacological understanding only when this second evidentiary step is made explicit rather than assumed.

Evidence-Constrained Interpretability: Proposed Concepts and Definitions

Interpretability is often treated as an intrinsic quality of a model or as a property conferred by applying an explanation method. Such treatment is inadequate for pharmaceutical prediction because explanation techniques differ in their access to the underlying model, their local or global scope, the features they expose, and the form of explanation they produce [15]. This article therefore proposes evidence-constrained interpretability as a theory of interpretive admissibility rather than a new attribution algorithm. Under this theory, a molecular interpretation is admissible for pharmacological reasoning only when it states what the model predicted, traces the representation from which that prediction arose, demonstrates that the explanation reflects the model's behavior, links the explanation to evidence independent of the model, qualifies the mechanistic status of the claim, characterizes uncertainty, and specifies the decision for which the interpretation is intended. This definition aligns interpretability with predictive accuracy, descriptive accuracy, and relevance to the scientific audience rather than reducing it to visual simplicity [16].

The proposed theory distinguishes several levels of interpretive claim. A predictive association states that a molecular feature is statistically related to an output within the modeled data. A model rationale states that the feature materially contributed to the model's prediction. A pharmacological hypothesis links that feature provisionally to a target, process, property, or experimental outcome. A mechanistic claim requires independent evidence that the feature participates in the biological or physicochemical process being asserted. Experimental confirmation requires an appropriately designed observation that corroborates the specified relation while addressing plausible alternatives. These levels should not be collapsed because explanation purposes, stakeholders, and acceptable evidence differ across scientific contexts [17]. Evidence-constrained interpretability consequently requires every interpretive statement to carry an explicit mechanistic-status label rather than allowing association, model rationale, mechanism, and causality to merge rhetorically.

The internal logic of the theory consists of six connected conditions. First, prediction-object specification identifies the endpoint, unit of analysis, data-generating context, and intended use. Second, representation traceability documents the molecular and biological information available to the model. Third, attribution faithfulness tests whether the explanation reflects actual predictive dependence rather than a chemically persuasive narrative. Fourth, independent evidence anchoring links attributed features to structural, chemical, target, assay, or literature evidence. Fifth, uncertainty qualification records predictive, model, data, explanation, and out-of-domain uncertainty. Sixth, decision-context alignment determines whether the resulting interpretation is sufficient for a particular screening, optimization, or experimental-prioritization decision. This separation is necessary because the plausibility of an explanation does not establish its faithfulness to the predictor [18]. **Figure 1** presents the evidence-constrained molecular interpretation prism, showing how the article's key components and boundaries are connected within evidence-constrained interpretability: proposed concepts and definitions.

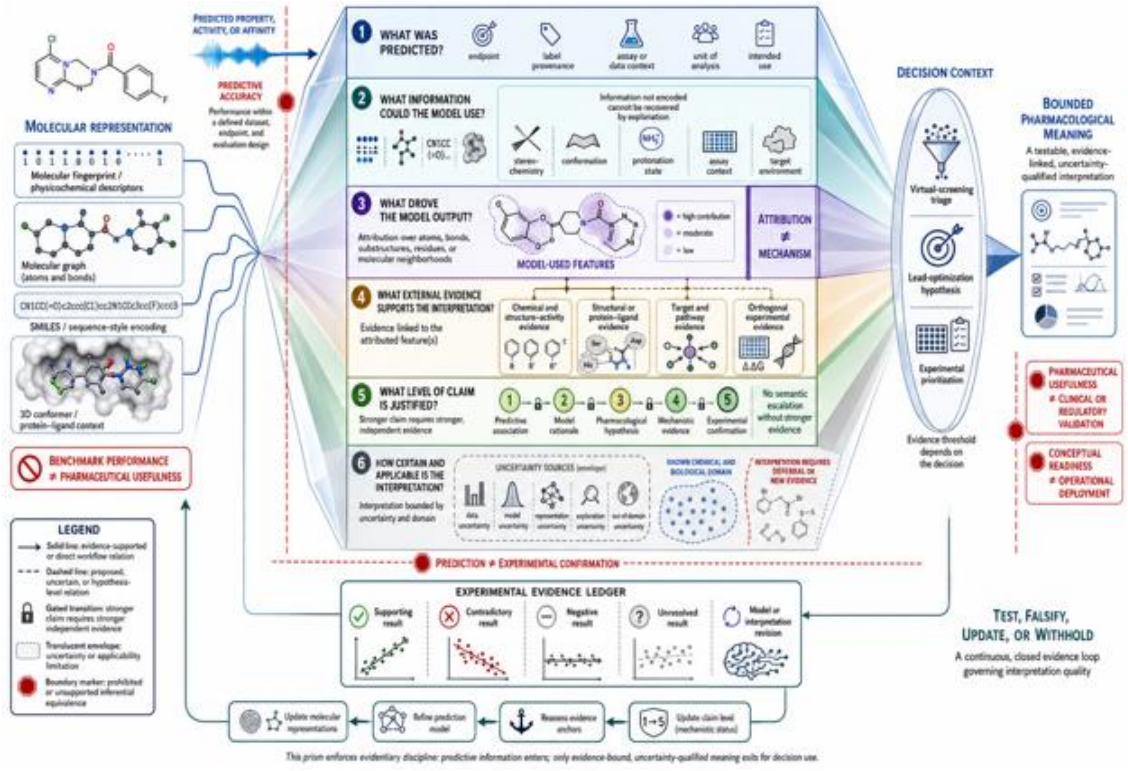

Figure 1. Evidence-Constrained Molecular Interpretation Prism.

The figure is an original conceptual synthesis that organizes the article's central contribution across predictive accuracy versus pharmacological understanding, molecular representations, attribution methods, and the limits of post hoc explanation, molecular representations, attribution methods, the limits of post hoc explanation, evidence-constrained interpretability: proposed concepts and definitions. 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.

The theory also rejects the assumption that its conditions can be collapsed into one universal interpretability score. An explanation could be faithful to a predictor yet unsupported by biological evidence, stable across retraining yet based on a biased endpoint, chemically plausible yet irrelevant to the intended decision, or useful for compound triage yet insufficient for a mechanistic claim. Explainability evaluation is therefore multidimensional and purpose-dependent rather than reducible to one metric [19]. Evidence-constrained interpretability should be applied as a structured scientific argument: each inferential step must be inspectable, its evidence source must be identifiable, uncertainty must remain visible, and unsupported steps must be marked as hypotheses. The theory is proposed as a basis for future testing and comparison, not as a validated checklist or certification system.

Linking Molecular Features to Targets, Mechanisms, and Decision Context

The transition from a molecular feature to a target-related interpretation begins with a distinction between predictive relevance and physical or biological participation. Sequence-based drug–target affinity models can learn associations between compound and protein encodings that support affinity prediction [20]. Such performance demonstrates that predictive structure exists in the input data, but it does not independently establish which atomic interaction occurs, whether the predicted relation survives assay transfer, or whether the feature highlighted by an explanation contributes causally to binding. Evidence-constrained interpretation therefore treats affinity prediction as one evidentiary layer. Structural observations, biochemical assays, mutagenesis, competition studies, selectivity profiles, and related evidence occupy separate layers that may strengthen, qualify, contradict, or overturn the model-derived hypothesis.

Interpretable compound–protein models can generate attention patterns connecting molecular substructures with protein residues, thereby offering candidate hypotheses about interaction determinants [21]. Graph-convolutional models likewise encode local molecular and protein structure for drug–target interaction prediction [22]. These methods can make a prediction more scientifically inspectable, but inspectability should not be confused with mechanistic confirmation. A residue or substructure identified by the model may correlate with activity because it represents direct interaction, scaffold membership, physicochemical context, analogue density, or a dataset-specific regularity. The interpretation becomes stronger only when independent evidence discriminates among those possibilities.

Graph-based affinity models further illustrate the difference between representation-level improvement and pharmacological explanation. Molecular graphs may improve prediction by preserving atom–bond neighborhoods that are obscured in simpler encodings [23]. Nevertheless, the semantic meaning assigned to a learned graph feature remains conditional on the endpoint,

training distribution, target representation, and explanation method. Evidence-constrained interpretability therefore requires a feature-to-evidence chain: the molecular feature must first be shown to influence the prediction, then connected to an independently supported target-related relation, and finally classified according to whether the evidence supports association, binding hypothesis, mechanistic participation, or experimental confirmation. Skipping any link risks converting a model rationale into an unsupported pharmacological narrative.

Decision context determines how complete that chain must be. A feature explanation used to diversify compounds selected for screening may require evidence of predictive faithfulness, applicability, and uncertainty, while a claim about binding mechanism requires substantially stronger structural and experimental support. Unified drug–target prediction libraries demonstrate that representations, architectures, training resources, and task choices materially affect model outputs [24]. The theory therefore makes interpretation conditional on the decision being supported, the alternatives under consideration, the cost of error, and the availability of confirmatory experiments. The same explanation may be admissible for low-consequence triage, insufficient for lead-series design, and unacceptable as evidence of target mechanism. Pharmacological meaning is consequently not an intrinsic property of the explanation; it is a bounded relation among the model, external evidence, and the scientific decision.

Evaluating Faithfulness, Stability, Uncertainty, and Pharmacological Usefulness

Evaluation should first determine whether an explanation is faithful to the prediction it purports to explain. In molecular property prediction, uncertainty methods vary in how well they calibrate confidence, rank likely errors, and identify cases requiring caution [25]. A comparable principle applies to explanations: a chemically intuitive attribution is not necessarily faithful, and a faithful attribution is not necessarily pharmacologically valid. Faithfulness can be examined through model-sensitive perturbation, feature removal, local surrogate fidelity, retraining, and comparison with known model dependencies. Molecular validity must be preserved during such tests because deleting atoms, breaking aromatic systems, or generating implausible analogues may measure response to invalid inputs rather than response to a scientifically meaningful intervention.

Stability addresses whether similar predictions produce similar interpretations under changes that should not alter the scientific conclusion. Relevant tests include retraining with different initializations, bootstrapping the data, changing equivalent molecular encodings, varying conformers, and examining matched molecular pairs. Scalable uncertainty methods exhibit different task-dependent strengths and weaknesses, making method selection itself part of the evaluation problem [26]. In drug design, predictive confidence must also be distinguished from calibrated uncertainty arising from noise, model limitations, sparse chemical coverage, and distribution shift [27]. Explanation stability and uncertainty should therefore be evaluated jointly: a stable explanation may still be systematically wrong, while an unstable explanation may reveal that apparently confident predictions rest on poorly determined internal rationales.

The available uncertainty literature indicates that calibration, scalability, and interpretation cannot be represented by a single confidence value [25-27]. Evidence-constrained interpretability extends this observation by proposing an interpretive uncertainty profile containing at least four elements: uncertainty in the prediction, uncertainty arising from the molecular representation, uncertainty in the explanation method, and uncertainty in the external evidence used to support a pharmacological claim. Evidential approaches can jointly model molecular predictions and uncertainty and may assist prioritization when their assumptions and calibration are made explicit [28]. They do not, however, establish that the explanation is mechanistically correct. Uncertainty qualification should modify the permissible strength of the interpretation rather than function as a decorative confidence interval around an otherwise unbounded claim.

Pharmacological usefulness is the most application-dependent evaluation dimension. It asks whether the explanation improves a defined scientific decision compared with prediction alone, conventional chemical reasoning, or another appropriate baseline. Molecular counterfactuals can identify local structural changes associated with altered predictions and thereby clarify model sensitivity [29]. Their usefulness depends on chemical validity, accessibility, applicability to the relevant series, and whether the contrast suggests an informative experiment. A useful explanation may reveal a model shortcut, identify an analogue that distinguishes competing hypotheses, expose an out-of-domain prediction, or justify deferral. User satisfaction or visual plausibility is insufficient. Evaluation should instead examine whether the interpretation improves hypothesis quality, experimental discrimination, error detection, compound selection, or justified refusal to act.

Implications for Virtual Screening, Lead Optimization, and Experimental Prioritization

In virtual screening, evidence-constrained interpretability can help distinguish ranking efficiency from evidentiary significance. Deep-learning-assisted docking can reduce the number of compounds subjected to computationally intensive evaluation and thereby expand the chemical space that can be searched [30]. The resulting predictions remain prioritization devices rather than proof of binding, potency, selectivity, or therapeutic value. Interpretability may add value when it reveals why a model favors particular chemotypes, whether those rationales persist across related compounds, and whether the ranked candidates occupy unfamiliar chemical regions. It may also identify cases in which high scores are supported by unstable, biased, or representation-dependent features and should therefore be deprioritized or tested cautiously.

Prospective screening provides a stronger test because computational ranking is closed by synthesis, acquisition, and experimental evaluation. Ultra-large-library docking has produced new chemotypes through workflows in which predictions were followed by empirical testing and characterization [31]. Evidence-constrained interpretation would not replace that closure. It would document which molecular features drove prioritization, what structural or target evidence supported those

features, what uncertainty remained, and which observations could falsify the working hypothesis. The objective is not to require mechanistic certainty before screening, but to ensure that a screening explanation is represented as a bounded selection rationale rather than as an established mechanism.

In lead optimization, the theory supports comparison of competing molecular modifications rather than unrestricted acceptance of a feature map. Rapid computational and experimental identification of kinase inhibitors illustrates the potential value of tightly connected design–test cycles [32]. Within such a cycle, explanations may help identify features associated with potency, selectivity, or another modeled endpoint, but optimization decisions must remain multiparametric. A substructure associated with improved predicted activity may simultaneously affect solubility, permeability, metabolic stability, toxicity, synthetic accessibility, or intellectual novelty. Evidence-constrained interpretability therefore requires endpoint-specific evidence and uncertainty, explicit identification of unmodeled objectives, and human review of the tradeoffs introduced by each proposed modification.

Experimental prioritization is the context in which the proposed theory may provide its clearest scientific benefit. Prediction-guided antibiotic discovery has shown that computational prioritization becomes pharmacologically meaningful when candidates undergo extensive biological evaluation [33]. Explainable deep learning has also been used to identify a structurally distinct antibiotic class whose predictions were followed by experimental testing [34]. These examples support the importance of prospective closure, but they do not validate a universal interpretability theory. Under the proposed approach, an explanation should identify the next experiment most capable of separating competing interpretations—for example, an orthogonal assay, a matched molecular pair, a selectivity test, structural analysis, or a target-engagement experiment. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop implications for virtual screening, lead optimization, and experimental prioritization within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from From Molecular Accuracy to Pharmacological Meaning An Evidence-Constrained Theory of Interpretable Prediction

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Prediction-object specification	Endpoint definition, labels, assay context, model output, intended use	Defines what the prediction represents and what it does not represent	Bounded predictive claim	Proxy mismatch, label ambiguity, measurement noise	Independent review of endpoint suitability and data provenance
Representation traceability	Molecular encoding, preprocessing, conformers, target information, omitted variables	Identifies what information was available to the model	Representation record and omission statement	Encoding dependence and incomplete molecular context	Robustness testing across scientifically defensible representations
Attribution-faithfulness assessment	Predictor, explanation method, perturbations, retraining runs	Tests whether highlighted features materially influence the prediction	Model-rationale statement	Correlated features, invalid perturbations, method disagreement	Model-sensitive and chemistry-preserving faithfulness tests
Independent evidence anchoring	Structural evidence, SAR, assays, target annotations, literature, negative findings	Links model-used features to evidence external to the predictor	Evidence-linked molecular hypothesis	Sparse, conflicting, or biased evidence	Orthogonal evidence review and prospective falsification
Mechanistic-status qualification	Attribution result and external evidence tier	Distinguishes association, model rationale, hypothesis, mechanism, and confirmation	Explicit claim-status label	Ambiguous or mixed evidence	Predefined criteria for movement between evidence levels
Uncertainty qualification	Calibration data, ensembles, out-of-domain measures, explanation variance	Limits reliance on uncertain predictions and explanations	Case-specific uncertainty profile	Calibration drift and incomplete uncertainty decomposition	Evaluation under scaffold, temporal, and distribution shifts
Virtual-screening use	Ranked molecules, applicability information, feature rationales, assay capacity	Supports compound selection and diversification	Bounded screening-priority rationale	Ranking may exploit artifacts or omit relevant properties	Prospective comparison with prediction-only and standard workflows
Lead-optimization use	Confirmed series data, multiparameter endpoints, counterfactual analogues	Supports generation of testable structural hypotheses	Prioritized analogue or modification hypothesis	Activity cliffs, property tradeoffs, synthesizability limits	Series-specific synthesis and experimental evaluation
Experimental prioritization	Competing hypotheses, uncertainty, experimental cost, expected discrimination	Selects experiments that can confirm, revise, or reject interpretations	Testable next-experiment recommendation	Information gain may be estimated incorrectly	Prospective assessment against conventional expert prioritization
Governance and revision	Versioned evidence, negative results, model updates, human decisions	Preserves provenance and supports correction	Auditable interpretation record	Selective reporting and automation bias	Independent oversight, challenge procedures, and periodic reassessment

Limitations and Boundary Conditions

The proposed theory has conceptual limitations. Its components are analytically separable but may interact differently across endpoints, representations, model classes, and discovery organizations. Representation traceability may be relatively straightforward for fixed molecular descriptors and substantially more difficult for multimodal learned embeddings.

Independent evidence may be abundant for well-characterized targets but sparse for phenotypic screens, novel targets, or poorly studied chemical matter. Decision-context alignment also depends on institutional resources, available assays, and the consequences of false-positive and false-negative decisions. The theory therefore cannot be assumed to have one universal implementation or one fixed evidence threshold.

The supporting evidence base also has limitations. Much of the available literature evaluates prediction, explanation, uncertainty, or discovery workflows separately rather than testing their integration within a common prospective design. Retrospective benchmarks may not capture temporal changes, project-specific selection pressures, assay transfer, or the compound-quality constraints encountered in real discovery programs. Explainability studies frequently evaluate plausibility, feature recovery, or local sensitivity without establishing improved pharmaceutical decisions. Uncertainty methods are similarly conditional on the evaluated tasks and calibration procedures [27]. These limitations mean that the proposed relations among faithfulness, evidence linkage, uncertainty, and usefulness remain theoretical until tested prospectively.

The theory must not be used to certify a molecular mechanism, clinical benefit, regulatory acceptability, or deployment readiness. It cannot convert weak labels into strong biological evidence, recover information omitted from the representation, eliminate distribution shift, or ensure that human reviewers will identify every misleading explanation. A structured interpretation may still be wrong, and a well-supported molecular hypothesis may fail because of unmodeled pharmacokinetic, toxicological, formulation, synthetic, or biological factors. The theory is intended to constrain claims and organize testing, not to replace medicinal chemistry, structural biology, pharmacology, toxicology, experimentation, or accountable scientific judgment.

Research Agenda and Implementation Priorities

The first research priority is to operationalize pharmacological meaning without reducing it to subjective plausibility. Prospective studies should compare prediction-only outputs, conventional post hoc explanations, and evidence-constrained interpretations in clearly defined tasks such as compound triage, analogue selection, target-hypothesis testing, and assay prioritization. Evaluation should measure whether interpretations improve decision quality, identify model failure, produce more discriminating experiments, or support justified deferral. Studies should preregister the intended decision, evidence thresholds, comparison condition, failure criteria, and handling of negative results. The proposed theory should be revised or rejected when its conditions do not improve scientific reasoning.

A second priority is the development of chemistry- and pharmacology-sensitive evaluation infrastructure. Benchmark resources should include representation variants, matched molecular pairs, activity cliffs, stereochemical alternatives, temporal splits, unfamiliar scaffolds, contradictory assays, and negative controls. Explanation evaluation should distinguish model faithfulness from evidence-link validity and should record disagreement among attribution methods rather than concealing it. Data and explanation provenance should be versioned so that interpretations can be reconstructed after model updates or new experiments. Such infrastructure would support cumulative testing of when an explanation remains stable, when it changes appropriately, and when it creates an unjustified mechanistic narrative.

Implementation should proceed in stages. Initial use should be limited to low-consequence research settings in which explanations are treated as hypotheses, human oversight is available, and all outputs can be challenged. A second stage may incorporate prospective evaluation within defined virtual-screening or optimization workflows, with predefined stopping rules and explicit comparison against existing practice. Broader use should depend on replicated evidence that the approach improves a specified decision without increasing automation bias, workload, or false confidence. At every stage, the implementation record should preserve the prediction, representation, explanation method, evidence links, uncertainty assessment, human decision, experimental outcome, and subsequent revision.

Conclusion

Evidence-constrained interpretability is proposed as a theory for governing the transition from molecular prediction to bounded pharmacological meaning. Its central claim is that predictive accuracy, feature attribution, chemical plausibility, and model confidence are individually insufficient to establish pharmacological understanding. An admissible interpretation must specify the prediction object, trace the molecular representation, demonstrate faithfulness to the predictor, connect attributed features to independent evidence, declare the mechanistic status of each claim, characterize uncertainty, and align the interpretation with a defined discovery decision. The theory does not validate molecular mechanisms, guarantee prospective success, or establish clinical, regulatory, or deployment readiness. Its intended contribution is to make inferential boundaries visible and to convert explanations into testable, revisable scientific hypotheses. By treating interpretation as an evidence-governed practice rather than a visualization endpoint, the proposed theory offers a basis for evaluating when molecular predictions may inform virtual screening, lead optimization, and experimental prioritization—and when the scientifically responsible outcome is to withhold interpretation and obtain additional evidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520-94. doi:10.1021/acs.chemrev.8b00728
3. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
4. 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
5. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50. doi:10.1016/j.drudis.2018.01.039
6. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res. 2021;54(2):263-70. doi:10.1021/acs.accounts.0c00699
7. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
8. Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
9. Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One. 2019;14(8). doi:10.1371/journal.pone.0220113
10. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
11. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
12. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
13. Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
14. 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
15. Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):93. doi:10.1145/3236009
16. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. 2019;116(44):22071-80. doi:10.1073/pnas.1900654116
17. 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
18. 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
19. 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
20. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17). doi:10.1093/bioinformatics/bty593
21. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019;35(18):3329-38. doi:10.1093/bioinformatics/btz111
22. Torng W, Altman RB. Graph convolutional neural networks for predicting drug-target interactions. J Chem Inf Model. 2019;59(10):4131-49. doi:10.1021/acs.jcim.9b00628
23. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: Predicting drug-target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-47. doi:10.1093/bioinformatics/btaa921
24. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: A deep learning library for drug-target interaction prediction. Bioinformatics. 2020;36(22-23):5545-7. doi:10.1093/bioinformatics/btaa1005

25. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
26. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
27. Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today. 2021;26(2):474-89. doi:10.1016/j.drudis.2020.11.027
28. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546
29. Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. Chem Sci. 2022;13(13):3697-705. doi:10.1039/D1SC05259D
30. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep Docking: A deep learning platform for augmentation of structure-based drug discovery. ACS Cent Sci. 2020;6(6):939-49. doi:10.1021/acscentsci.0c00229
31. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019;566(7743):224-9. doi:10.1038/s41586-019-0917-9
32. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
33. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.e13. doi:10.1016/j.cell.2020.01.021
34. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2024;626(7997):177-85. doi:10.1038/s41586-023-06887-8