



Review Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Interpretable Pharmacoinformatics Beyond Descriptor Importance: A Validation Framework for Mechanistically Grounded, Uncertainty-Aware, and Decision-Relevant Molecular Explanations

Kenji Tanaka^{1*}, Yui Kobayashi¹, Takeshi Nakamura²

¹Department of Pharmacoinformatics and Interpretable Models, Graduate School of Pharmaceutical Sciences, University of Tokyo, Tokyo, Japan.

²Department of Mechanistic Validation and Decision-Relevant AI, Graduate School of Pharmacy, Kyoto University, Kyoto, Japan.

***Email:** tanaka.kenji@mol.f.u-tokyo.ac.jp

ABSTRACT

Feature importance has become a common language for explaining quantitative structure–activity relationship and molecular machine-learning models, yet an attribution is only one kind of interpretive claim. A descriptor, atom, bond, fragment, or learned concept can receive high importance without establishing that the explanation is stable across model specifications, faithful to the model's actual decision process, representative of global behaviour, mechanistically causal, or useful for a medicinal-chemistry decision. This critical review examines evidence from 2017–2024 on molecular explainability, QSAR validation, counterfactuals, graph explanations, uncertainty, activity cliffs, representation dependence, and interpretable model design. The evidence supports a central distinction between prediction performance and explanation validity, and shows that chemical plausibility cannot substitute for tests of faithfulness, stability, applicability, and scope. We propose an interpretability claim ladder in which descriptor attribution, model-behaviour explanation, mechanistic hypothesis, and decision-relevant interpretation are progressively stronger claims with progressively stronger validation requirements. The ladder is not an empirically validated scoring system; it is an evidence-bounded synthesis for preventing inferential escalation. Pharmacoinformatics becomes more interpretable not when more importance maps are produced, but when each explanation is matched to its explanatory object, molecular representation, chemical-space support, validation design, uncertainty, and intended decision.

Keywords: Explainable artificial intelligence, Pharmacoinformatics, QSAR, Molecular representation, Feature attribution, Medicinal chemistry

INTRODUCTION

Modern pharmacoinformatics increasingly combines classical QSAR reasoning with machine-learning architectures whose predictive logic is difficult to inspect directly. Explainable artificial intelligence has therefore become an important methodological theme in drug discovery, spanning feature attribution, counterfactual reasoning, example-based explanations and interpretable model structures [1]. Deep QSAR expands the representational and functional flexibility available to molecular modelling, but it also inherits familiar requirements concerning validation, chemical-space support and the distinction between prediction and inference [2]. These issues are not peculiar to deep learning: QSAR interpretation remains inseparable from applicability-domain and validation logic [3].

The central problem is that predictive success does not validate an explanation. Deep neural networks have achieved strong performance in benchmarked ChEMBL bioactivity tasks [4], while MoleculeNet demonstrated how markedly molecular-learning conclusions depend on datasets, tasks, splitting regimes and representations [5]. An attribution produced after such a prediction therefore answers a narrower question than is often implied. This review argues that descriptor importance should be treated as an entry point into interpretation, not its endpoint. The proposed interpretability claim ladder distinguishes four escalating claims: which input features influence a prediction; how the fitted model behaves in a defined neighborhood or domain; whether that behaviour supports a mechanistic hypothesis; and whether the explanation is sufficiently reliable and relevant to inform a drug-discovery decision.

Scope and literature selection

This critical review focuses on peer-reviewed work published from 2017 through 2024 that directly informs interpretation of molecular prediction models, with emphasis on QSAR, molecular machine learning, graph neural networks, counterfactuals, feature attribution, representation, uncertainty and medicinal-chemistry decision support. General XAI scholarship was retained only where it supplies concepts that molecular studies use but do not independently establish. Contemporary XAI taxonomies distinguish multiple explanatory families rather than a single construct called “interpretability” [6], and interpretable-machine-learning scholarship further separates predictive accuracy, descriptive accuracy and relevance to an intended audience [7]. These distinctions define the review’s analytical boundary.

The search and selection process was designed for critical synthesis rather than meta-analysis. Concept-based searching and citation chaining were combined with bibliographic and DOI verification; no pooled effect, universal algorithm ranking or systematic-review completeness claim is made. Evidence was read according to the specific model, representation, dataset, chemical series, benchmark and decision context in which each conclusion was obtained. This is important because a molecular explanation can be technically valid for one model while remaining non-transferable to another representation, endpoint or chemical region.

What interpretability means in pharmacoinformatics

Interpretability is not a single property residing inside a model. Rudin’s distinction between intrinsically interpretable models and post-hoc explanations is especially important because an explanation of a black box and a transparent predictive rule carry different evidential meanings [8]. In pharmacoinformatics, the distinction is sharpened by the fact that the same molecule can be represented by physicochemical descriptors, fingerprints, graphs, fragments, learned embeddings or physically grounded quantities. What is being explained must therefore be declared before an explanation is judged successful.

Human explanation research adds a second constraint: explanations are selective, contrastive and audience-dependent rather than exhaustive reproductions of a decision process [9]. A medicinal chemist asking why analogue A is predicted to be more potent than analogue B is posing a different question from a modeller asking which variables globally drive predictions. Molecular-XAI scholarship correspondingly spans counterfactuals, local structural rationales and representation-level explanations [10]. Chemical intelligibility is useful, but it is neither identical to model faithfulness nor evidence of molecular causation.

We therefore propose an interpretability claim ladder. At the first level, attribution identifies model-associated importance for a descriptor or molecular unit. At the second, model-behaviour explanation requires evidence that the reported dependence is faithful, stable and correctly scoped to a local or global region. At the third, a mechanistic hypothesis requires external chemical or physical coherence but remains a hypothesis unless intervention or independent experimental evidence supports causation. At the fourth, decision-relevant interpretation requires a specified user, action, uncertainty tolerance and applicability boundary. Movement up this ladder is not automatic; each step adds a validation burden.

Descriptor attribution and its limitations

Feature attribution is attractive in medicinal chemistry because it can convert a numerical prediction into a structural rationale that appears compatible with analogue design. In compound optimization, this rationalization can help interrogate why a model favors one molecular change over another [11]. The problem begins when the visual salience of an attribution is mistaken for reliability. Sheridan showed that atom-level coloring can change substantially across descriptor and model combinations even when predictions are comparatively robust [12]. The

important methodological implication is not that atomic heatmaps are useless, but that prediction robustness and explanation robustness are separate properties.

Descriptor importance is also contingent on the available training support. Across many activity-classification settings, comparatively small datasets and limited feature sets can sometimes support useful prediction, but these observations do not imply a universal minimum sample requirement or a universally sufficient feature subset [13]. More stringent tests arise when closely related compounds behave differently. Activity cliffs provide such a challenge because similar structures with marked activity differences force an explanation method to recover local structural distinctions rather than broad correlational regularities; benchmarking on these cases has shown substantial method dependence [14].

The proposed claim ladder therefore treats attribution as a provisional state. An attribution may be retained at the first level when its direction or location changes materially after plausible model, descriptor or neighborhood perturbations. It can support a stronger model-behaviour claim only when the explanation remains sufficiently stable for the intended use, tracks the fitted model rather than an analyst's chemical expectation, and is interpreted within the model's supported chemical region. For medicinal-chemistry use, the relevant question is not simply which feature is important, but under what perturbations, neighboring structures, model specifications and decision stages that importance remains consistent.

Because these failure modes are difficult to compare in prose, the following table organizes descriptor attribution as a sequence of model-support states rather than as a ranked list of explanation methods.

Table 1. Temporal states in which descriptor importance can remain usable, become unstable, or require interpretive downgrading

Model-support state	Attribution comparison anchor	Dominant uncertainty source	Validation discriminator over the next state	Decision consequence	Transfer boundary	Failure signature	Proposed claim-ladder endpoint
Initial fitted-model state	Importance assigned to a descriptor, atom, bond or fragment for the current prediction	Model specification and representation choice	Refit or re-explain under plausible alternative specifications	Treat as a provisional rationale only	Current endpoint, training domain and representation	Importance location or sign changes while prediction remains similar	Attribution level
Re-specified model state	Persistence of the same explanatory object after descriptor or model change	Explanation instability despite predictive robustness	Compare explanation direction, locality and rank under controlled respecification	Avoid using unstable features as design rules	Models sharing the tested data support, not arbitrary new tasks	Chemically different rationales support similar predictions	Attribution level, or model-behaviour level if stable
Local challenge state	Explanation behaviour around close analogues and activity-cliff-like cases	Local discontinuity and sparse neighborhood support	Test whether the explanation distinguishes structurally small but activity-relevant changes	Restrict interpretation to the local series when discrimination is supported	Defined analogue neighborhood and assay context	Smooth explanation persists across discordant close analogues	Local model-behaviour level
Later chemical-series extension	Reuse of a prior attribution after new chemistry enters the project	Chemical-space shift and changing model support	Re-estimate explanation stability and uncertainty after domain expansion	Reassess earlier design rationales before reuse	Newly sampled chemistry must be supported rather than assumed equivalent	Historical attribution survives only because the model is extrapolating	Return to attribution level pending revalidation

The same logic can be represented as competing interpretation paths rather than a single progression. The following figure makes explicit where an apparently persuasive attribution should advance, remain provisional, or return to a weaker claim.

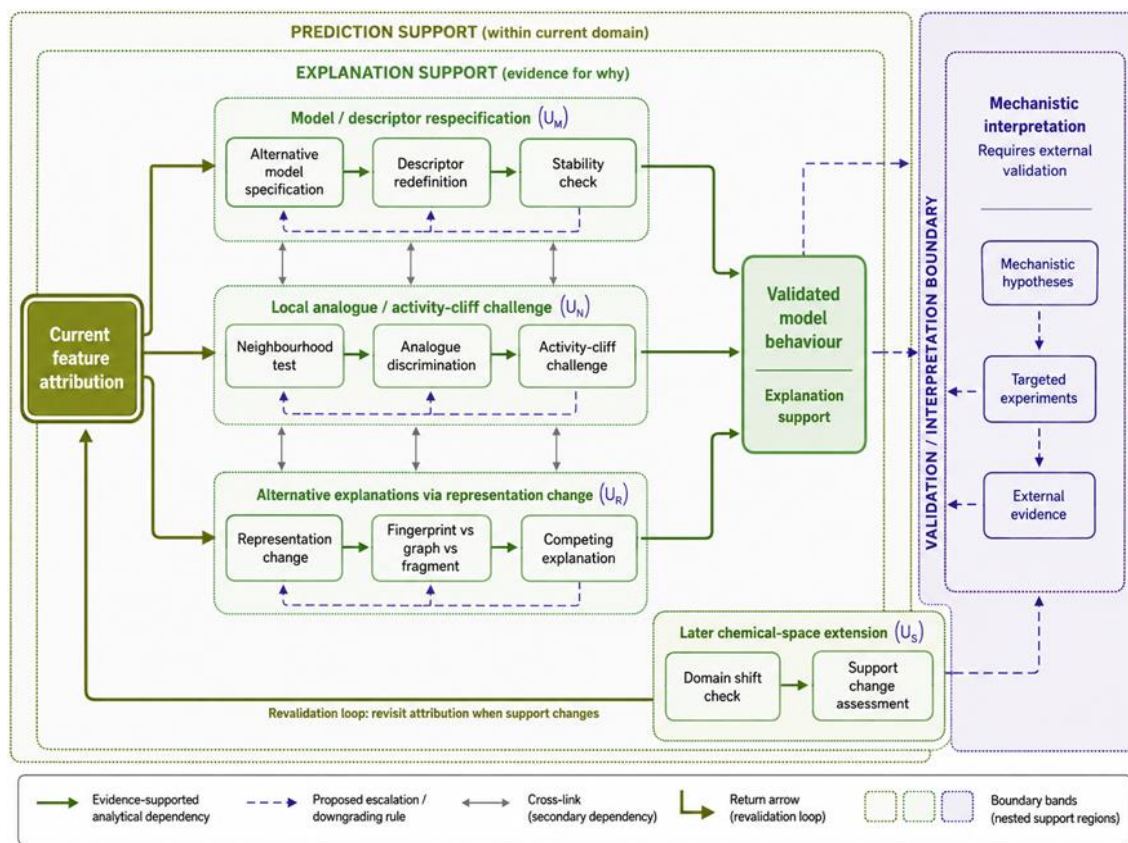


Figure 1. Descriptor importance under successive tests of stability, locality and chemical-space support

Local explanations versus global model behaviour

Local explanations answer questions about a prediction or neighborhood, not necessarily about the fitted function everywhere. Molecular counterfactuals exemplify this distinction: they identify feasible structural changes associated with a changed prediction and can be generated across several model families, but they do not by themselves describe the model's global response surface [15]. Bond-centric Shapley explanations similarly provide fine-grained local information for graph neural networks [16]. Their resolution is valuable only if the analyst preserves the scope of the claim.

Quantitative studies of molecular GNN explanations demonstrate why this scope cannot be assumed. Explanation quality varies across models, tasks and explanation methods, and can be evaluated rather than judged only from appealing molecular coloring [17]. Chemistry-aware substructure masking can improve alignment between graph explanations and chemically coherent units [18], yet a highlighted substructure still describes a model-related explanatory object, not a proven causal pharmacophore. The proposed claim ladder therefore requires every local explanation to state its neighborhood, comparator and representation, while global claims require separate evidence that the reported relation persists across the intended model domain.

Explanation stability and faithfulness

A faithful explanation should reflect the decision process of the fitted model; a stable explanation should remain sufficiently consistent under perturbations that ought not to change its meaning. These properties are related but not interchangeable. Molecular attribution benchmarks have explicitly evaluated accuracy, sparsity, fidelity or infidelity, stability and sensitivity, showing that explanation quality is multidimensional [19]. A visually coherent map can therefore fail because it is unstable, because it does not track the model, or because it is faithful to a model whose learned association is itself scientifically uninformative.

Uncertainty adds another layer. Molecular property models can decompose and explain sources of predictive uncertainty, making it possible to ask not only what drove a prediction but why the model is uncertain about it

[20]. Explanation behaviour can also be altered by the learning objective: introducing substructure-aware constraints has improved benchmarked molecular explanations in a defined GNN setting [21]. Training-instance and atom-level contributions available from interpretable graph-kernel approaches further show that an explanation can depend on which training examples support the prediction as well as on which molecular units are highlighted [22].

For this review, these findings motivate a proposed three-part validation requirement before an attribution is promoted to a model-behaviour explanation: fidelity to the fitted predictor, stability under scientifically plausible perturbation, and disclosure of support or uncertainty. None alone establishes mechanism. Conversely, failure of one dimension should not be concealed by strength in another; a stable but unfaithful explanation is not rescued by reproducibility, while a faithful but highly unstable explanation may be unsuitable for a design decision that requires consistent rationale.

Mechanistic plausibility is not causal proof

The strongest interpretive overreach occurs when a chemically plausible explanation is treated as evidence that the highlighted feature causes the measured biological outcome. Activity cliffs make this error visible because very similar molecules can show large activity differences, exposing local limitations in machine-learning smoothness and representation [23]. A model may generate a tidy structural narrative through such a region while missing the local relationship that matters for the assay. Conversely, correctly identifying a cliff-associated substructure still does not establish why that structural change alters activity.

Physically grounded representations can strengthen interpretation at a different evidential level. Machine learning of real-space chemical descriptors can reproduce physically defined atomic and interatomic quantities and thereby support explanations anchored to explicit chemical observables [24]. Such grounding is stronger than a purely arbitrary latent coordinate, but the inferential boundary remains. A physically meaningful descriptor can support a hypothesis about electronic or structural contribution; it does not, without independent evidence, establish target engagement, cellular mechanism, pharmacodynamic consequence, clinical response or safety effect.

The proposed claim ladder therefore inserts a validation gate between model-behaviour explanation and mechanistic hypothesis. Crossing that gate requires evidence external to the explanatory output itself and matched to the claimed level of mechanism. The decision-state ledger below records what can be concluded at different levels of explanatory support, what remains unresolved, and what observation should force reassessment.

Table 2. Decision states for preventing chemically plausible explanations from being promoted prematurely to causal mechanism

Decision state	Information currently available	Unresolved uncertainty	Validation discriminator	Permissible interpretive action	Reassessment trigger	Decision consequence	question
Attribution agrees with known chemistry	Model highlights a descriptor or substructure that is chemically credible	Whether the model actually relies on that relation and whether it is causal	Test faithfulness and stability before external mechanistic inference	Retain as a model-associated rationale	Explanation changes under equivalent model or representation choices	Do not convert plausibility into a mechanism claim	What evidence is sufficient to move beyond attribution?
Local analogue pattern is recovered	Explanation distinguishes close analogues, including difficult local changes	Whether the recovered pattern generalizes beyond the tested neighborhood	Challenge with additional local chemistry and explicit domain checks	Use as a local model-behaviour explanation	A new close analogue breaks the explanation while remaining in assay scope	Restrict design inference to the supported neighborhood	When does local explanatory success justify broader SAR interpretation?
Physical descriptor is interpretable	Highlighted quantity has explicit physical meaning	Whether that quantity mediates the biological endpoint	Independent experimental or mechanistic evidence at	Formulate an explicitly qualified mechanistic	Physical association persists but biological perturbation	Separate chemical explanation from	Which validation level matches the mechanistic

			the claimed biological level	c hypothesis	does not support the proposed pathway	pharmacologic al causation	claim being made?
Multiple explanation methods agree	Distinct methods identify a similar molecular region or feature	Shared bias from data, representatio n or model family	Cross- representatio n and external validation rather than consensus alone	Treat converge nce as supportive, not decisive	Agreement disappears after data or representatio n shift	Avoid reasoning that consensus itself establishes causality	How much independent evidence is needed before convergent explanations become decision- relevant?

Interpretation across molecular representations

Interpretation is representation-contingent because a model cannot attribute importance to information that its representation does not expose. Comparative work across public and proprietary molecular-property datasets has shown that fixed and learned representations differ in performance according to endpoint, dataset and chemical-space regime rather than exhibiting universal superiority [25]. Consequently, two models predicting the same endpoint may offer apparently conflicting explanations because one reasons through fingerprint bits, another through graph neighborhoods, and another through learned latent features. Agreement or disagreement between their explanations is therefore uninterpretable unless the explanatory unit is specified.

Domain knowledge can make those units more chemically recognizable. Group-contribution concepts have been incorporated into graph neural networks to produce explanations organized around chemically meaningful groups, while preserving predictive modelling within a learned architecture [26]. Alternative fragmentation schemes such as pBRICS similarly alter the molecular units available for property interpretation [27]. Molecular segmentation methods extend this idea by learning group-level partitions that can support structure–property analysis without requiring a predetermined fragment vocabulary [28]. These strategies enlarge the interpretability toolkit, but none establishes that a chosen group, fragment or segment corresponds to a causal pharmacological mechanism.

Representation also determines which uncertainty is visible. A graph-kernel model can expose influential training instances and atomic contributions [22], whereas real-space descriptors can connect learned predictions to physically defined quantities [24]. The appropriate comparison is therefore not simply which representation is “more interpretable,” but whether the explanatory object matches the scientific question. An atom-level attribution may be appropriate for a local structural perturbation, a fragment may better express analogue-series reasoning, and a physically grounded quantity may better support a hypothesis about electronic or intermolecular contributions.

We propose a representation-translation gate within the interpretability claim ladder. Before transferring an explanation from one molecular representation to another—or from a model-derived unit to a medicinal-chemistry statement—the analyst should identify what information survives the translation, what is lost, and whether the new explanatory unit has independent chemical meaning. The table below separates representation-dependent decisions that can otherwise appear equivalent in prose.

Table 3. Decision states for translating molecular explanations across descriptors, graphs, fragments and physically grounded representations

Representati on decision state	Explanatory object currently available	Competing interpretation	Principal uncertainty	Permissible translation	Observation requiring reassessment	Proposed claim-ladder endpoint
Fixed descriptor or fingerprint model	Descriptor value, bit or encoded substructure	Importance may reflect correlated encoded information rather than a unique chemical driver	Descriptor construction and feature correlation	Interpret within the declared encoding and endpoint	Equivalent prediction receives materially different rationale under another valid encoding	Attribution level

Atom- or bond-resolved graph model	Node, edge or local graph neighborhood	Salient graph element may reflect learned contextual dependence rather than an isolated group effect	Neighborhood definition and message-passing context	Use for model-local structural reasoning	Salience disappears when chemically equivalent context is represented differently	Local model-behaviour level
Fragment or group-based representation	Chemically defined fragment, group or molecular segment	Group boundaries may be imposed or learned differently across methods	Fragment granularity and segmentation rule	Compare explanations only after matching explanatory units	Alternative chemically defensible fragmentation changes the conclusion	Representation-qualified model-behaviour level
Physically grounded representation	Explicit atomic, interatomic or other physically defined quantity	Physical correlation may be mistaken for biological mechanism	Transfer from physical quantity to assay-level consequence	Use to formulate a chemically grounded hypothesis	Biological or pharmacological evidence fails to support the inferred pathway	Mechanistic-hypothesis level only
Cross-representation convergence	Similar structural region implicated by distinct representations	Agreement may arise from shared data bias or correlated encodings	Independence of models, data and explanatory units	Treat convergence as supportive evidence	Convergence fails under domain shift or external validation	Proposed escalation only after external validation

Figure 2 extends this translation problem into chemical-space support: representation changes, domain boundaries and validation gates must remain visible simultaneously rather than being treated as independent modelling choices.

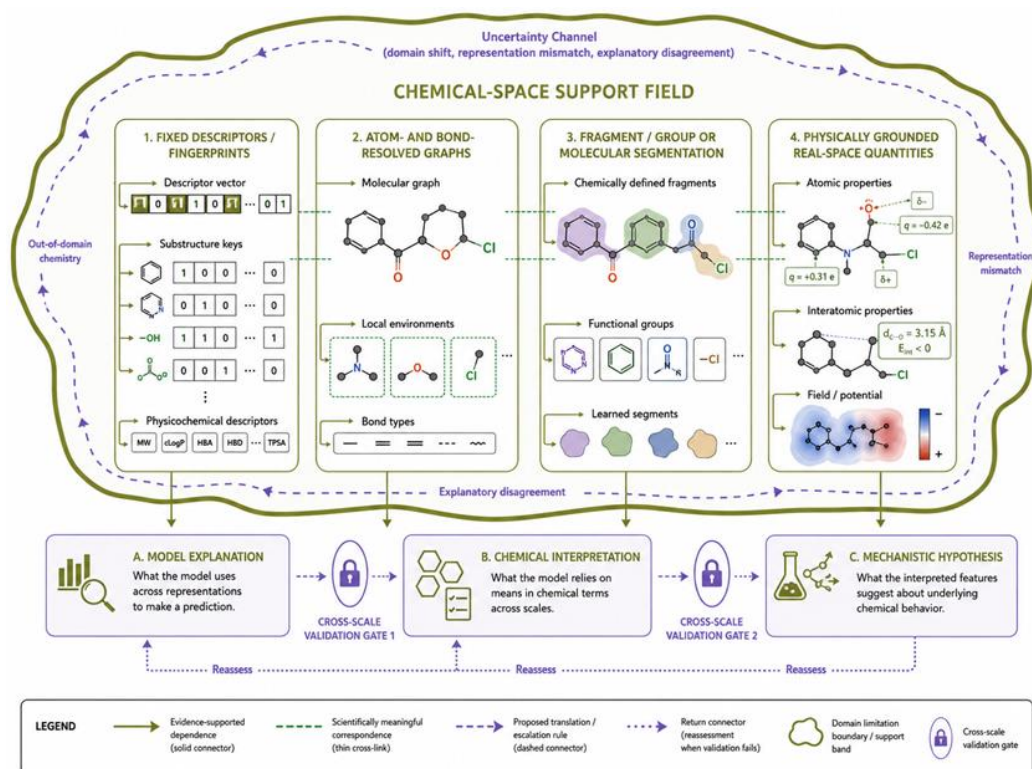


Figure 2. Representation-dependent explanations across chemical-space support and validation boundaries

Human-centred interpretability for drug discovery decisions

An explanation becomes decision-relevant only when its intended user and action are specified. Concept-whitening approaches demonstrate that molecular GNN representations can be aligned with selected human-readable molecular concepts, providing an intrinsically inspectable alternative to purely post-hoc attribution [29].

Yet concept alignment does not prove that the selected concepts are complete, causally correct or optimal for a medicinal chemist.

Human-centred interpretability should therefore be evaluated against the decision being supported. A counterfactual may be useful when the question is which feasible structural modification could alter a prediction [15]; an attribution map may help interrogate an analogue series [11]; and a global model summary may be preferable when portfolio-level model behaviour is under review. Human preference alone cannot establish faithfulness, while technically faithful explanations may still be irrelevant to the design decision. We therefore propose that decision-relevant interpretation requires four declared elements: user, intended action, applicable chemical region and uncertainty that would change the action.

This requirement also limits translational rhetoric. None of the reviewed molecular-XAI studies establishes that explanation use improves clinical outcomes, regulatory acceptance or prospective drug-development success. The defensible claim is narrower: appropriately validated explanations may improve the inspectability of computational reasoning and make assumptions or failure modes easier to interrogate during model-informed decision making.

Benchmarking requirements

Interpretability should be benchmarked against competitive predictive models rather than rewarded merely because a model is easy to visualize. Sparse topological regression has shown that an explicitly interpretable QSAR strategy can remain competitive across a large collection of human-target ChEMBL datasets, demonstrating that interpretability and predictive utility need not always be opposing objectives [30]. That result nevertheless remains specific to the evaluated tasks and does not establish universal superiority.

Operational benchmarking should also reflect the workflow in which models are used. Automated QSAR platforms can integrate model development, validation, uncertainty estimation and explainability within design–make–test–analyze cycles [31]. Integration, however, does not validate an explanation by itself. A credible benchmark should separately examine predictive performance, applicability support, explanation faithfulness, stability, uncertainty, sensitivity to representation and performance on difficult local cases such as activity cliffs [14].

The proposed claim ladder therefore requires a benchmark to match the intended interpretive claim. Feature attribution requires perturbational or attribution-focused tests; local explanation requires neighborhood fidelity; global interpretation requires evidence across the relevant model domain; mechanistic hypotheses require external evidence; and decision-relevant explanations require evaluation against the actual decision process. No single score can substitute for this layered validation.

Unresolved challenges and research priorities

The first unresolved problem is explanation disagreement. Atom coloring can vary across valid model specifications [12], attribution metrics capture different properties [19], and representation changes alter what counts as an explanatory object [25]. Future studies should therefore report disagreement rather than hiding it through a single preferred visualization. Cross-representation concordance is informative only when the compared explanations operate on commensurable molecular units.

Second, explanation uncertainty requires more explicit treatment. Prediction uncertainty can itself be interpreted [20], but uncertainty in the explanation—its sensitivity to retraining, neighborhood choice, representation or data shift—should be reported separately. This is particularly important around activity cliffs, where apparently minor structural differences can expose local failure [23].

Third, the field needs prospective decision studies. Current evidence supports methods for explanation generation and benchmarking, but does not establish that these explanations improve compound-selection efficiency, medicinal-chemistry reasoning or downstream development outcomes. Studies should compare decisions made with and without specific explanation types while preserving the underlying predictive information, thereby separating explanation utility from model accuracy.

Finally, causal validation must remain distinct from model explanation. Physically grounded descriptors [24], chemically coherent fragments [26], learned segments [28] and convergent attributions may strengthen a mechanistic hypothesis, but none should be promoted automatically to causal evidence. The most productive research agenda is therefore not a search for one universally interpretable architecture. It is the construction of claim-matched validation designs in which the strength of an explanation is proportional to the evidence supporting the exact inference being made.

CONCLUSION

Interpretable pharmacoinformatics requires more than locating important descriptors, atoms or fragments. Attribution can initiate scientific interrogation, but stronger claims require evidence of model faithfulness, stability, local or global scope, representation dependence, chemical-space support and uncertainty. Mechanistic interpretation demands an additional evidential boundary, and decision relevance requires a defined user and action.

The proposed interpretability claim ladder organizes these escalating obligations without treating them as an empirically validated score. Its principal purpose is restraint: explanations should advance only as far as their validation permits. This shift preserves the usefulness of molecular XAI while reducing the risk that attractive visualizations are mistaken for robust model behaviour, causal mechanism or demonstrated drug-development benefit.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. 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
2. Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: The emergence of deep QSAR. *Nat Rev Drug Discov.* 2024;23(2):141-55. doi:10.1038/s41573-023-00832-0
3. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. *Chem Soc Rev.* 2020;49(11):3525-64. doi:10.1039/d0cs00098a
4. Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform.* 2017;9(1):45. doi:10.1186/s13321-017-0232-0
5. 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
6. 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
7. 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
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
9. 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
10. 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
11. Rodríguez-Pérez R, Bajorath J. Explainable machine learning for property predictions in compound optimization. *J Med Chem.* 2021;64(24):17744-52. doi:10.1021/acs.jmedchem.1c01789
12. 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
13. Siemers FM, Feldmann C, Bajorath J. Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity. *Cell Rep Phys Sci.* 2022;3(11):101113. doi:10.1016/j.xcrp.2022.101113

14. Jiménez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. *J Chem Inf Model.* 2022;62(2):274-83. doi:10.1021/acs.jcim.1c01163
15. 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
16. Mastropietro A, Pasculli G, Feldmann C, Rodríguez-Pérez R, Bajorath J. EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks. *iScience.* 2022;25(10):105043. doi:10.1016/j.isci.2022.105043
17. Rao J, Zheng S, Lu Y, Yang Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns (N Y).* 2022;3(12):100628. doi:10.1016/j.patter.2022.100628
18. Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun.* 2023;14(1):2585. doi:10.1038/s41467-023-38192-3
19. Wang Y, Huang M, Deng H, Li W, Wu Z, Tang Y, et al. Identification of vital chemical information via visualization of graph neural networks. *Brief Bioinform.* 2023;24(1). doi:10.1093/bib/bbac577
20. Yang CI, Li YP. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *J Cheminform.* 2023;15(1):13. doi:10.1186/s13321-023-00682-3
21. Amara K, Rodríguez-Pérez R, Jiménez-Luna J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *J Cheminform.* 2023;15(1):67. doi:10.1186/s13321-023-00733-9
22. Xiang Y, Tang YH, Lin G, Reker D. Interpretable molecular property predictions using marginalized graph kernels. *J Chem Inf Model.* 2023;63(15):4633-40. doi:10.1021/acs.jcim.3c00396
23. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model.* 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
24. 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
25. Yang K, Swanson K, Jin W, Coley C, 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
26. Aouichaoui ARN, Fan F, Mansouri SS, Abildskov J, Sin G. Combining group-contribution concept and graph neural networks toward interpretable molecular property models. *J Chem Inf Model.* 2023;63(3):725-44. doi:10.1021/acs.jcim.2c01091
27. Vangala SR, Krishnan SR, Bung N, Srinivasan R, Roy A. pBRICS: A novel fragmentation method for explainable property prediction of drug-like small molecules. *J Chem Inf Model.* 2023;63(16):5066-76. doi:10.1021/acs.jcim.3c00689
28. Chen Z, Li D, Liu M, Liu J. Graph neural networks with molecular segmentation for property prediction and structure-property relationship discovery. *Comput Chem Eng.* 2023;179:108403. doi:10.1016/j.compchemeng.2023.108403
29. Proietti M, Ragno A, La Rosa B, Ragno R, Capobianco R. Explainable AI in drug discovery: Self-interpretable graph neural network for molecular property prediction using concept whitening. *Mach Learn.* 2024;113(4):2013-44. doi:10.1007/s10994-023-06369-y
30. Zhang R, Nolte D, Sanchez-Villalobos C, Ghosh S, Pal R. Topological regression as an interpretable and efficient tool for quantitative structure-activity relationship modeling. *Nat Commun.* 2024;15(1):5072. doi:10.1038/s41467-024-49372-0
31. Mervin L, Voronov A, Kabeshov M, Engkvist O. QSARtuna: An automated QSAR modeling platform for molecular property prediction in drug design. *J Chem Inf Model.* 2024;64(14):5365-74. doi:10.1021/acs.jcim.4c00457