Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(1):80-89
https://doi.org/10.51847/0r77qJPUoT

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

The Prediction–Explanation Divide in Interpretable Structure–Activity Modelling and How Pharmaceutical Decisions Should Cross It

Sara Al-Fahd[1]*, Noura Al-Khalifa[1], Omar Al-Turki[2]

[1]Department of Interpretable Structure–Activity Modeling, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia.
[2]Department of Prediction–Explanation Integration, College of Pharmacy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

***Email:** s.alfahd@ksu.edu.sa*

ABSTRACT

Interpretable structure–activity modelling is increasingly expected to provide not only accurate predictions but also scientifically credible reasons for them. Yet prediction, interpretation, explanation, and molecular mechanism represent different epistemic achievements, and their conflation can produce persuasive but unsupported pharmaceutical narratives. This Original Interpretability Architecture Article defines the prediction–explanation divide and proposes a Decision-Adaptive Prediction–Explanation Crossing Architecture for determining when and how model outputs may support progressively consequential pharmaceutical decisions. The architecture separates predictive validity from attribution faithfulness, chemical coherence, perturbational concordance, and decision governance. It treats feature attribution, fragment highlighting, counterfactual generation, and concept-based explanations as complementary probes of model behaviour rather than independent demonstrations of molecular causation. Medicinal-chemistry validation is organized around chemically meaningful perturbations, including analogue series, matched molecular pairs, activity cliffs, prospective synthesis, orthogonal assays, and mechanistic interventions where appropriate. The architecture also identifies failure modes arising from dataset shift, inadequate applicability domains, unstable explanations, implausible counterfactuals, uncalibrated confidence, retrospective rationalization, and human confirmation bias. Governance requirements include explicit intended-use statements, evidence-proportional escalation, independent challenge, documented overrides, monitoring, and requalification after material changes. The principal contribution is an original conceptual framework that links explanatory claims to decision-specific evidence burdens and prevents direct movement from prediction to mechanism without external corroboration. The proposed gates and relationships remain hypotheses for prospective evaluation and do not establish clinical suitability, regulatory acceptance, universal applicability, or deployment readiness.

Key words: *Pharmacoinformatics, QSAR, Chemical space, External validation, Applicability domain, Calibration*

Received: 19 November 2024; **Revised:** 25 February 2025; **Accepted:** 28 February 2025

INTRODUCTION

Machine-learning models in pharmacoinformatics can rank compounds, predict molecular properties, identify structural patterns, and generate visual or textual rationales. These capabilities have intensified interest in explainable artificial intelligence for drug discovery, where model outputs are expected to become intelligible to medicinal chemists and other decision-makers [1]. However, an explanation that appears chemically meaningful may describe only how a model processed its representation. It may not identify why a molecule produces an observed biological effect or what would occur after a real structural intervention.

This distinction matters because QSAR validity is not determined by predictive performance alone. Data quality, molecular representation, endpoint definition, validation design, chemical-space coverage, applicability domain, and the intended use of a model jointly determine whether a prediction can support a scientific claim [2]. Interpretability adds another layer of conditionality: an explanation can be coherent within the fitted model while remaining unstable across representations, unsupported outside the training domain, or disconnected from molecular causation.

The pharmaceutical consequence of this ambiguity depends on the decision. A feature map used to generate an exploratory hypothesis does not require the same evidence as an explanation used to eliminate a compound, modify a lead series, escalate a safety concern, or support translational reasoning. Machine learning already contributes across heterogeneous stages of pharmaceutical research and development, but the evidentiary meaning of its outputs varies across these contexts [3]. A general demand for "explainability" is therefore insufficient unless the required type of explanation and its decision boundary are specified.

This article develops an original, non-empirical interpretability architecture for crossing the prediction–explanation divide. It proposes a Decision-Adaptive Prediction–Explanation Crossing Architecture, abbreviated D-PECA, that separates prediction, interpretation, explanation, and mechanism while linking each level to explicit evidence gates. The architecture is intended to organize scientific reasoning, method evaluation, medicinal-chemistry validation, uncertainty qualification, and human review. It is a conceptual synthesis rather than an empirical study, exhaustive evidence review, validated standard, clinical guideline, regulatory framework, or deployment manual.

Prediction, interpretation, explanation, and mechanism

A prediction is a model-generated estimate, probability, class, ranking, or property value produced under a learned mapping. An interpretation characterizes how inputs, representations, parameters, or intermediate model states contributed to that output. An explanation is more demanding: it is a question-relative and audience-dependent account that connects model behaviour to a meaningful contrast, alternative, or scientific rationale. Explanation is commonly selective and contrastive rather than a complete disclosure of every computational process [4].

Interpretability and explainability should not be treated as interchangeable labels for any method that renders a model output visible. They encompass different goals, audiences, levels of transparency, and methodological assumptions [5]. Global summaries describe broad model behaviour, whereas local explanations concern particular predictions. Intrinsic interpretability arises from the model structure itself, whereas post-hoc methods approximate or interrogate a fitted model. None of these categories independently establishes that the highlighted feature is responsible for the underlying molecular effect.

A mechanism is an externally supported account linking a molecular intervention to a biological or physicochemical consequence through an evidenced process. It therefore differs from a model mechanism, which describes how a computational system transforms its inputs. In high-consequence settings, reliance on post-hoc explanations can be problematic when an intrinsically interpretable alternative could provide a more direct account of the prediction process [6]. Even intrinsic transparency, however, does not convert a predictive association into a molecular mechanism.

D-PECA consequently proposes an ordered but non-automatic continuum: prediction, interpretation, explanation, and mechanism. The sequence represents increasing evidentiary burden rather than a guaranteed progression. A prediction may be interpreted without becoming scientifically explanatory, and an explanation may remain useful without establishing mechanism. Because explanations are contrastive and audience-relative, their adequacy must be evaluated against the precise pharmaceutical question being asked [4]. The proposed continuum requires empirical testing and does not assign universal thresholds to any level.

Why model explanation may not explain molecular causation

Predictive accuracy and interpretation quality are separable properties. A model may classify compounds successfully while assigning unstable, chemically implausible, or method-dependent importance to their structural features. Conversely, a technically simple or locally faithful explanation may accompany a prediction that lacks adequate external validation. Benchmarks developed for QSAR interpretation show that explanation methods require evaluation independently from conventional predictive performance [7]. The first divide therefore lies between obtaining a correct output and identifying a defensible account of how that output arose.

A second divide concerns plausibility and evidence. Explanations often resemble familiar scientific reasoning: they highlight a functional group, emphasize an aromatic region, or propose that replacing one fragment would

change activity. Such narratives can reassure users because they align with expectations, even when the explanation method has not been tested against the phenomenon that matters for the decision. Comparable concerns in high-consequence modelling show that plausible post-hoc explanations may create false confidence when they lack decision-relevant validation [8].

A third divide concerns association and intervention. A predictive model estimates patterns under the distribution represented in its data. A causal claim instead concerns what would change under an intervention, subject to assumptions about confounding, selection, measurement, and the causal structure of the system. Predictive and counterfactual questions are therefore not equivalent [9]. In molecular modelling, changing an encoded feature, deleting an atom in silico, or generating a nearby counterfactual does not by itself establish what synthesis, exposure, binding, signalling, metabolism, or toxicity would follow from the corresponding physical change.

The prediction–explanation divide can thus arise through three conflations: model sensitivity is mistaken for chemical importance; chemical association is mistaken for biological causation; and explanatory plausibility is mistaken for decision readiness. Interpretation benchmarks provide evidence that the first conflation is methodologically testable [7]. The latter two require progressively stronger external evidence. Alternative explanations may include correlated substructures, dataset artefacts, scaffold imbalance, representation dependence, latent assay conditions, or model multiplicity, in which different models produce similar predictions for different reasons.

Figure 1 presents the original conceptual synthesis for prediction–interpretation–explanation–mechanism continuum, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

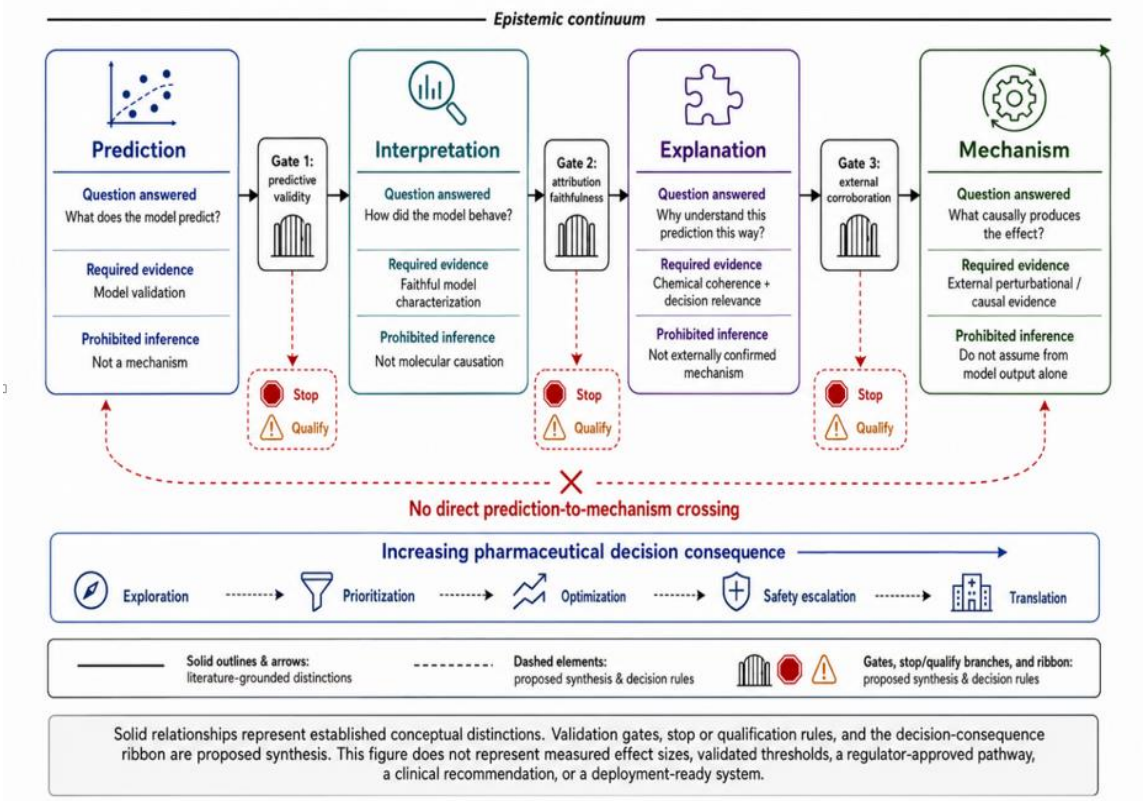

Figure 1. Prediction–Interpretation–Explanation–Mechanism Continuum.

The figure is an original conceptual synthesis developed for "The Prediction–Explanation Divide in Interpretable Structure–Activity Modelling and How Pharmaceutical Decisions Should Cross It." Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

Proposed prediction–explanation decision architecture

D-PECA is proposed as a decision architecture rather than a new explanation algorithm. Its purpose is to determine what additional evidence is required before a model output can support a stronger scientific or pharmaceutical

claim. Explanations used during compound optimization must connect model behaviour to chemically meaningful questions, such as which structural change is proposed, which endpoint may change, and what uncertainty accompanies that inference [10]. The architecture therefore begins with the decision and works backward to the required explanatory evidence.

The first gate, G0: predictive validity, asks whether the underlying prediction is suitable for interpretation. It includes endpoint integrity, data quality, appropriate validation, chemical-space coverage, applicability-domain assessment, calibration where probabilistic claims are made, predictive uncertainty, and sensitivity to anticipated dataset shift. The second gate, G1: attribution faithfulness, asks whether the explanation reflects the fitted model rather than human expectations or incidental data patterns. Scientific explainability methods differ in assumptions and actionability, requiring tests tailored to the explanation claim [11].

The third gate, G2: chemical coherence, evaluates whether an interpretation respects molecular identity, valence, stereochemistry, protonation, tautomerism, fragment context, and plausible chemical transformations. Atom- or fragment-level coloring may help identify regions influencing a prediction, but such coloring remains a model-relative interpretation rather than direct evidence of molecular mechanism [12]. G2 therefore asks whether the explanatory object corresponds to a chemically meaningful entity and whether proposed alternatives remain within a relevant chemical neighbourhood.

The fourth gate, G3: perturbational concordance, asks whether explanation-derived claims agree with observed consequences of meaningful changes. Potential tests include equivalent-representation perturbations, feature-removal experiments, matched molecular pairs, activity cliffs, analogue series, prospective synthesis, orthogonal assays, and mechanistic interventions. The distinction between prediction and intervention motivates this escalation [9]. D-PECA does not assume that every project requires the strongest experiment, but it prohibits mechanistic language when intervention-sensitive evidence is absent.

The fifth gate, G4: decision fit and governance, links explanatory burden to pharmaceutical consequence. Exploratory hypothesis generation may tolerate qualified, model-relative explanations, whereas compound elimination, safety escalation, or translational use requires stronger validation, independent challenge, and explicit residual uncertainty. In consequential settings, the availability of a more interpretable model should also be considered before relying on post-hoc rationalization [6]. These gates, their sequence, and their stop, qualify, or escalate rules are proposed constructs requiring prospective evaluation across endpoints, models, organizations, and decision contexts.

Feature attribution, fragments, counterfactuals, and concept-based explanations

Feature-attribution methods assign portions of a predicted output to encoded inputs or learned components. Local approximations and Shapley-value approaches can reveal which molecular features influenced a particular model prediction [13]. Their proper interpretation is model-relative: a large contribution indicates influence within the fitted mapping, not a causal effect of the corresponding atom or fragment in a biological system.

Molecular counterfactuals describe structural alternatives for which the model predicts a different outcome. They can make an explanation contrastive and potentially useful for lead optimization, but proposed changes must be screened for molecular validity, synthetic feasibility, applicability-domain membership, and unintended property changes [14]. A small representational edit is not necessarily a realistic chemical intervention.

Concept-based explanations use predefined abstractions, such as pharmacophores or functional patterns, to connect learned representations with medicinal-chemistry language. Pharmacophore-informed graph models demonstrate how concept-level information can support property prediction and interpretation [15]. Nevertheless, concept recognition remains conditional on the selected vocabulary and does not establish that the concept mediates the observed activity.

D-PECA therefore treats attribution, fragments, counterfactuals, and concepts as complementary probes. Agreement may increase confidence that a model uses a chemically coherent signal, whereas disagreement should trigger investigation of representation dependence, local instability, correlated features, or model multiplicity. No individual method, or agreement among methods, is sufficient to establish molecular mechanism.

Stability, faithfulness, chemical validity, and actionability

Stability concerns whether materially equivalent inputs, retrained models, or small irrelevant perturbations produce appropriately similar explanations. Activity cliffs provide a demanding local test because minor structural changes can accompany major activity changes, allowing assessment of whether attribution methods recognize

decision-relevant structure–activity discontinuities [16]. Stability should be interpreted locally rather than assumed across an entire chemical space.

Faithfulness concerns whether an explanation reflects the fitted model rather than human expectation. Tests using alternative molecular encodings, model controls, or test-time augmentation show that intuitive explanations can remain inconsistent across representations or insufficiently responsive to model changes [17]. An explanation may therefore be visually persuasive yet technically unfaithful.

Chemical validity requires chemically meaningful entities, plausible transformations, and appropriate treatment of valence, stereochemistry, protonation, tautomerism, and molecular context. Actionability further requires that the explanation can inform a defined decision without concealing uncertainty. Molecular uncertainty methods vary across datasets and tasks, so explanatory confidence should be conditioned on validated predictive uncertainty rather than inferred from appearance [18].

D-PECA proposes that an explanation should be assessed across all four dimensions rather than accepted after passing only one. Stability does not prove correctness, faithfulness does not prove causation, chemical validity does not guarantee developability, and actionability remains decision-specific. **Table 1** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses in the prediction–explanation divide in interpretable structure–activity modelling and how.

Table 1. Definitions, components, relationships, and intended uses in The Prediction–Explanation Divide in Interpretable Structure–Activity Modelling and How.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Predictive validity	Invalid outputs being explained	Endpoint, data, split, external validation, chemical space	**Proposed:** explanation begins only after prediction suitability is established	Prevents interpretation of unsupported predictions	External validation and applicability-domain evidence [2]	Leakage, shift, poor calibration	Valid prediction does not establish explanation
Attribution faithfulness	Explanations reflecting expectation rather than model behaviour	Model, representation, attribution method	**Proposed:** explanations must respond appropriately to model and input perturbations	Distinguishes model diagnosis from visual plausibility	Sanity, perturbation, and invariance tests [17]	Stable but unfaithful narratives	Faithfulness does not establish molecular mechanism
Chemical coherence	Invalid fragments or transformations	Molecular identity, valence, stereochemistry, local context	**Proposed:** explanatory objects must correspond to chemically meaningful entities	Improves medicinal-chemistry interpretability	Chemical review and feasible counterfactuals [14]	Impossible or out-of-domain changes	Chemical plausibility does not prove biological consequence
Perturbational concordance	Correlational interpretation presented as causation	Analogues, activity cliffs, matched pairs, assays	**Proposed:** stronger claims require observed response to meaningful perturbation	Connects explanation to external evidence	Cliff-sensitive and prospective tests [16]	Confounding or assay-specific effects	Concordance remains system- and endpoint-specific
Predictive uncertainty	Confidence inferred from explanation appearance	Calibration, ensembles, uncertainty estimator, shift	**Proposed:** explanation use is conditioned on uncertainty quality	Supports qualification and escalation	Comparative uncertainty validation [18]	Miscalibration and false certainty	Uncertainty estimates are not guarantees
Decision fit	One explanation	Intended action, consequence,	**Proposed:** evidence	Aligns explanations	Predefined decision and	Overuse beyond	The proposed tiers are not

	standard applied to all uses	alternatives, reversibility	burden increases with decision consequence	with pharmaceutical use	residual-uncertainty review [6]	validated purpose	regulatory readiness levels

Decision-specific explanation requirements

Explanation adequacy depends on the model, endpoint, representation, and intended QSAR use. Architectures that combine prediction and interpretation can support different scientific questions, but their explanatory outputs must still be assessed in the context in which they will be used [19]. A generic declaration that a model is "interpretable" provides insufficient information for a pharmaceutical decision.

Molecular representations encode different information and inductive assumptions. Fingerprints, graphs, sequences, descriptors, conformations, and learned embeddings may emphasize different structural relationships [20]. Consequently, explanation requirements should specify which molecular entity is represented, what invariances are expected, which chemical states are excluded, and whether the explanation remains meaningful across relevant representations.

Pharmaceutical decisions also emerge from multidisciplinary workflows rather than isolated model outputs. AI-supported drug design requires integration with chemical knowledge, biological evidence, experimental feasibility, and expert judgement [21]. An explanation should therefore clarify what it contributes, what evidence remains external to the model, and which decision-maker retains responsibility.

D-PECA proposes four practical use tiers. Exploratory explanations may generate hypotheses with explicit qualification. Prioritization requires predictive validity, attribution faithfulness, and basic chemical coherence. Lead optimization requires stronger local stability and perturbational evidence. Safety escalation or translational use requires independent review, documented uncertainty, external corroboration, and a clear stop rule. These tiers organize evidence burdens but are not validated readiness classifications.

Validation using medicinal chemistry and perturbational evidence

Explanation validation should test whether model-derived rationales correspond to chemically meaningful structure–activity relationships. Substructure-aware learning objectives can improve alignment between graph-model explanations and predefined chemical substructures [22]. Such alignment supports interpretation of model behaviour but does not establish that a highlighted substructure causes the observed biological effect.

Activity cliffs offer a stringent test because closely related compounds may display sharply different activities. Their presence exposes local generalization failures that aggregate performance measures can conceal [23]. Cliff analysis can therefore reveal whether a model and its explanation respond to structural distinctions that matter within a medicinal-chemistry series.

Models may also be evaluated explicitly for their ability to recognize activity-cliff behaviour [24]. Additional evidence can arise from matched molecular pairs, analogue series, prospective synthesis, orthogonal assays, target perturbations, or rescue experiments. The required test should correspond to the claim: an attribution claim requires faithfulness testing, whereas a mechanistic claim requires external intervention-sensitive evidence.

D-PECA proposes a validation ladder from computational diagnostics to chemical and biological perturbation. Progression should stop when evidence fails, remain qualified when uncertainty is unresolved, and escalate only when the intended decision demands stronger support. **Table 2** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for testable propositions, evidence requirements, and validation criteria for the prediction–explanation divide in interpretable structure–activity modelling and how.

Table 2. Testable propositions, evidence requirements, and validation criteria for The Prediction–Explanation Divide in Interpretable Structure–Activity Modelling and How.

Architecture layer or process stage	Core function	Information or material flow	Interaction with other components	Validation criterion	Potential failure mode	Human or experimental responsibility	Readiness boundary
G0: Predictive validity	Establish prediction suitability	Data and representation to validated output	Conditions every later gate	External and domain-relevant performance	Leakage, shift, miscalibration	Modeller and validation scientist	No explanation-based use after G0 failure

G1: Attribution faithfulness	Test model-relative explanation	Prediction to attribution or local rationale	Precedes chemical interpretation	Response to model, feature, and encoding controls	Human-intuitive but unfaithful output	Model developer and independent challenger	Supports model diagnosis only
G2: Chemical coherence	Test chemical meaning	Attribution to fragments, concepts, or counterfactuals	Connects technical output to chemistry	Valid molecular states and feasible changes	Invalid structures or arbitrary fragments	Medicinal chemist	Does not establish biological effect
G3: Perturbational concordance	Compare explanation with observed change	Proposed structural change to experiment	Links explanation to external evidence	Consistency with cliffs, matched pairs, or assays [23]	Confounding or non-transferable local evidence	Experimental scientist	Mechanistic language remains qualified
G4: Decision fit and governance	Control consequential use	Evidence package to accountable decision	Integrates all earlier gates	Intended-use match, residual uncertainty, review	Unsupported escalation or automation bias	Decision owner and reviewer	Not clinical or regulatory authorization
Requalification loop	Reassess after material change	New data, model, endpoint, or context back to G0	May invalidate prior explanations	Repeated validation under changed conditions	Stale explanations and hidden shift	Change-control owner	Prior passage does not transfer automatically

Failure modes and misleading explanatory narratives

Uncertainty estimates may appear reassuring while failing to identify or rank prediction errors reliably. Comparative evaluations show that uncertainty-estimation performance depends on the molecular task and method [25]. A narrow interval, high confidence score, or stable explanation should therefore not be treated as evidence that structural uncertainty, dataset shift, or model misspecification has been resolved.

Evaluation results can also change with dataset size, split strategy, molecular representation, and model architecture. These elements interact rather than operating as independent technical choices [26]. An explanation derived under a random split may therefore appear credible while failing under scaffold, temporal, or externally shifted conditions relevant to pharmaceutical use.

Atom-level coloring illustrates the narrative risk. Highlighted atoms can change with descriptors, models, or perturbation procedures even when the resulting visualization appears chemically reasonable [27]. Similar risks apply to fragments, attention maps, counterfactuals, and concepts when users interpret familiarity as evidence of faithfulness.

D-PECA identifies recurring failure narratives: explanation without valid prediction, interpretation outside the applicability domain, chemically impossible counterfactuals, attribution driven by confounding, retrospective rationalization, and agreement caused by shared method assumptions. Human confirmation bias can amplify each failure. Passing selected checks reduces specific risks but does not guarantee correctness, causality, or decision readiness.

Governance and human review

Responsible use requires more than a technically interpretable output. Consequential model integration should define the workflow, anticipated harms, monitoring arrangements, escalation pathways, and accountable human roles [28]. Within D-PECA, governance begins when the intended decision is specified rather than after an explanation has already influenced action.

Structured reporting supports review by making data provenance, intended use, model design, evaluation, and limitations traceable. Reporting frameworks for clinical AI illustrate how minimum-information requirements can strengthen scrutiny of modelling claims [29]. D-PECA adapts this general principle without presenting an existing checklist as a pharmaceutical regulatory standard.

The proposed governance structure assigns distinct responsibilities to the model developer, validation scientist, medicinal chemist or endpoint expert, decision owner, independent challenger, and change-control owner. Review records should document explanation disagreement, applicability-domain status, uncertainty, unresolved alternatives, override reasons, and evidence that would trigger reassessment.

Human review is not an automatic remedy. Experts may share dataset assumptions, prefer familiar chemical narratives, or defer to apparently sophisticated models. Governance should therefore include independent challenge, explicit stop authority, and requalification after changes to data, endpoint, representation, model, or intended use. These elements are proposed safeguards rather than validated operating procedures.

Limitations and research priorities

Applicability-domain methods can identify compounds that are dissimilar to reference data, but their performance depends on the molecular representation, similarity definition, training set, and endpoint [30]. Domain membership is therefore conditional evidence, not a universal certificate of reliability. Local explanations may fail even for nominally in-domain compounds, while some out-of-domain predictions may remain informative but require stronger qualification.

Large chemical language models can learn transferable representations that encode aspects of molecular structure and properties [31]. Such capability may expand the scope of predictive modelling, yet embeddings, attention patterns, or learned concepts do not automatically provide faithful explanations or molecular mechanisms. Foundation-scale models remain subject to the same decision-specific evidence boundaries.

Research should compare explanation methods independently from prediction, test invariance across equivalent molecular representations, develop feasible counterfactual benchmarks, and prospectively evaluate explanatory claims through medicinal-chemistry perturbations. Studies should also examine explanation behaviour under dataset shift, integrate uncertainty with applicability-domain assessment, and test whether structured human review improves decisions rather than documentation alone. **Figure 2** presents the original conceptual synthesis for decision-adaptive interpretability architecture for pharmaceutical use, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

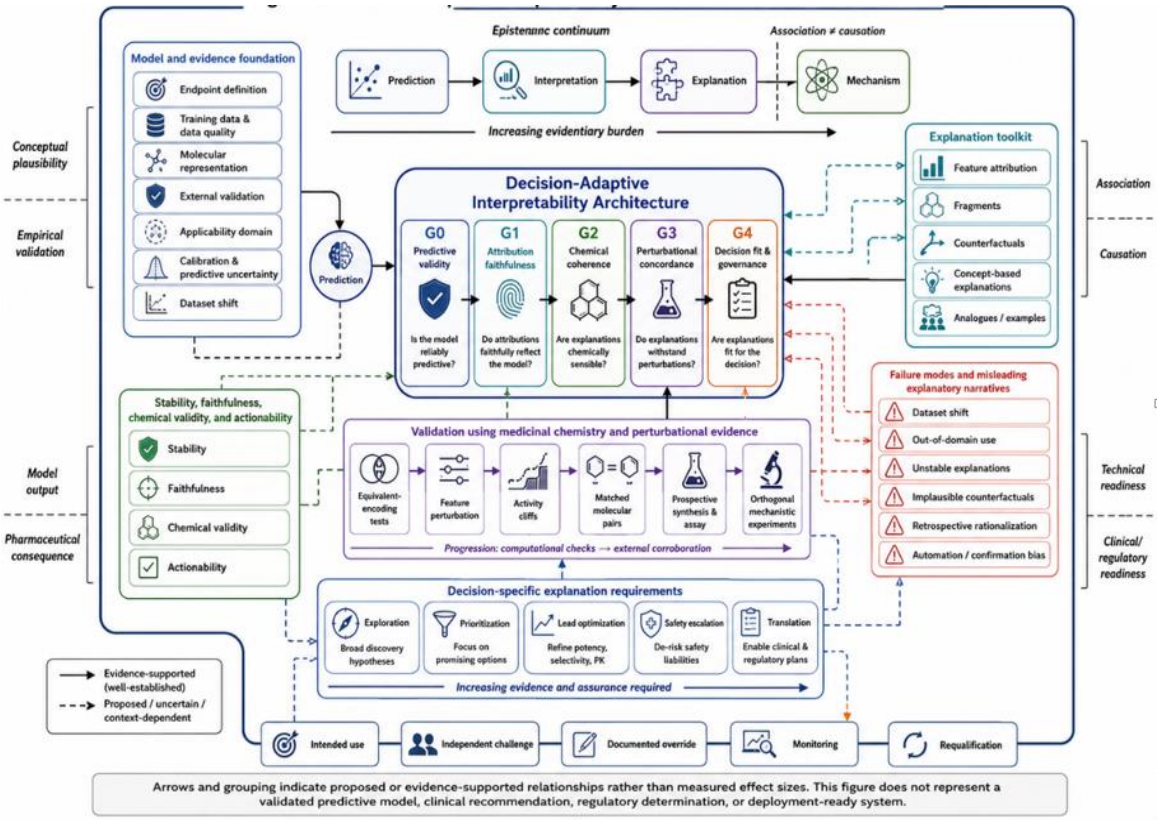

Figure 2. Decision-Adaptive Interpretability Architecture for Pharmaceutical Use.

The figure is an original conceptual synthesis developed for "The Prediction–Explanation Divide in Interpretable Structure–Activity Modelling and How Pharmaceutical Decisions Should Cross It." Arrows and grouping indicate

proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

The architecture has limitations of its own. Its gates are qualitative, their order may require adaptation, and the appropriate evidence burden will vary by endpoint, chemical series, organization, and decision. Prospective studies must determine whether D-PECA improves scientific judgement, identifies misleading explanations, or creates avoidable procedural burden. Until then, it should be treated as a testable conceptual architecture rather than a validated standard

CONCLUSION

The prediction–explanation divide arises when predictive performance, model interpretation, explanatory plausibility, and molecular mechanism are treated as equivalent. D-PECA proposes that pharmaceutical decisions cross this divide only through evidence gates addressing predictive validity, attribution faithfulness, chemical coherence, perturbational concordance, and decision governance. Feature attribution, fragments, counterfactuals, and concepts can contribute to model understanding, but none independently establishes molecular causation. Explanation requirements should increase with decision consequence and remain conditioned on applicability domain, uncertainty, external validation, experimental evidence, and accountable human review. The architecture is an original conceptual synthesis whose scientific value, practical burden, and effects on decision quality require prospective evaluation.

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. 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
3. 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
4. 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
5. 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
6. 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
7. Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. J Cheminform. 2021;13(1):41. doi:10.1186/s13321-021-00519-x
8. 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
9. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369-75. doi:10.1038/s42256-020-0197-y
10. 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

11. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh CY, et al. From black boxes to actionable insights: A perspective on explainable artificial intelligence for scientific discovery. J Chem Inf Model. 2023;63(24):7617-27. doi:10.1021/acs.jcim.3c01642
12. Jiménez-Luna J, Skalic M, Weskamp N, Schneider G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. J Chem Inf Model. 2021;61(3):1083-94. doi:10.1021/acs.jcim.0c01344
13. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J Med Chem. 2020;63(16):8761-77. doi:10.1021/acs.jmedchem.9b01101
14. 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
15. Kong Y, Zhao X, Liu R, Yang Z, Yin H, Zhao B, et al. Integrating concept of pharmacophore with graph neural networks for chemical property prediction and interpretation. J Cheminform. 2022;14(1):52. doi:10.1186/s13321-022-00634-3
16. 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
17. Hartog PBR, Krüger F, Genheden S, Tetko IV. Using test-time augmentation to investigate explainable AI: Inconsistencies between method, model and human intuition. J Cheminform. 2024;16(1):39. doi:10.1186/s13321-024-00824-1
18. 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
19. Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform. 2020;12(1):17. doi:10.1186/s13321-020-00423-w
20. Chuang KV, Gunsalus LM, Keiser MJ. Learning molecular representations for medicinal chemistry. J Med Chem. 2020;63(16):8705-22. doi:10.1021/acs.jmedchem.0c00385
21. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, 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
22. 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
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. Dablander M, Hanser T, Lambiotte R, Morris GM. Exploring QSAR models for activity-cliff prediction. J Cheminform. 2023;15(1):47. doi:10.1186/s13321-023-00708-w
25. 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
26. 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
27. 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
28. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
29. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
30. Liu R, Wallqvist A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. J Chem Inf Model. 2019;59(1):181-9. doi:10.1021/acs.jcim.8b00597
31. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7