Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2026, 15(3):1-11
https://doi.org/10.51847/TPKNzXPpkn



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Pharmacoinformatics Should Integrate Uncertainty Before Ranking Virtual Candidates

Laura Keller[1]*, Thomas Lehmann[1], Simon Brunner[2], Christoph Meier[3], David Thompson[4]

[1]*Department of Pharmacoinformatics and Uncertainty Integration, Faculty of Pharmacy, ETH Zurich, Zurich, Switzerland.*
[2]*Department of Virtual Candidate Ranking, Faculty of Pharmacy, EPFL Lausanne, Lausanne, Switzerland.*
[3]*Department of Drug Discovery Informatics, Faculty of Pharmacy, University of Bern, Bern, Switzerland.*
[4]*Department of Uncertainty-Aware Molecular Design, Faculty of Pharmacy, University of Basel, Basel, Switzerland.*

***Email:** laura.keller@pharma.ethz.ch*

ABSTRACT

Virtual screening and molecular-property models increasingly produce ordered candidate lists, yet the numerical precision of a prediction is often treated as if it were equivalent to decision certainty. That interpretation is unsafe because prediction error can arise from noisy reference data, model uncertainty, weak chemical-space support, local structure–activity discontinuities, model disagreement, and miscalibration. This Decision-Science Perspective argues that uncertainty should be represented before compounds are ranked rather than appended as a confidence note after ranking. The central distinction is between a predicted value and the evidential conditions under which that value deserves decision weight. The article develops a proposed risk-adjusted prioritization logic in which chemical-space support, model disagreement, calibration, and decision utility remain analytically visible until the point of candidate selection. No universal uncertainty metric, penalty function, or acceptance threshold is assumed. Instead, uncertainty is treated as a structured decision variable that can justify promotion, down-weighting, deferral, or targeted experimentation. The framework is intended for prospective evaluation against score-only ranking under future-like chemical and data shifts.

Keywords: *Uncertainty quantification, Pharmacoinformatics, Molecular property prediction, Applicability domain, Conformal prediction, Candidate prioritization*

Received: 21 March 2026; **Revised:** 08 July 2026; **Accepted:** 11 Jyly 2026

INTRODUCTION

Machine learning and other computational models now contribute to hit identification, mechanism-oriented analysis, and property optimization in preclinical drug discovery [1]. Their outputs increasingly influence which compounds receive synthesis, assay, profiling, and development resources. Yet ranking workflows still commonly privilege a scalar predicted value—predicted potency, permeability, clearance, toxicity, or another property—while the conditions that determine how much that value should be trusted are handled separately or not at all.

In silico chemistry distinguishes numerical prediction from error and uncertainty, and uncertainty itself can originate at several levels of the modeling process [2]. A value printed to several decimal places can therefore be computationally precise while remaining poorly supported as a decision. This distinction matters most near ranking boundaries, where small numerical differences may reverse the apparent order of candidates despite being smaller than plausible predictive uncertainty.

© 2026 The International Journal of Pharmaceutical Research and Allied Sciences (IJPRAS). **Open Access** - This article is under the **CC BY NC SA** license (https://creativecommons.org/licenses/by-nc-sa/4.0).

Benchmarking work reinforces that uncertainty quantification cannot be reduced to one universally preferred estimator [3]. Performance depends on the dataset, model class, uncertainty method, and downstream use. This Perspective therefore advances a proposed decision principle: before virtual candidates are totally ordered, their predictions should be accompanied by an explicit uncertainty state that preserves data quality, chemical-space support, model disagreement, calibration, and decision context. The purpose is not to replace medicinal chemistry judgment with a new composite score, but to prevent unsupported numerical differences from masquerading as evidence.

Point predictions and false precision

False precision begins when a model score is interpreted without asking whether the evaluation setting resembles the decision setting. Ligand-based classification benchmarks can reward train-test redundancy and memorization rather than genuine generalization, producing apparently strong performance that weakens when similarity structure is controlled [4]. A high retrospective score is therefore not, by itself, evidence that the ordering of new chemical matter is reliable.

The problem is sharper at activity cliffs. Across 30 targets and 24 machine-learning and deep-learning approaches, activity-cliff compounds remained substantially more difficult than smoother regions of structure–activity space [5]. This creates a candidate-level failure mode that aggregate error metrics can conceal: two close analogues may receive confident-looking predictions even when a small structural modification is associated with a large potency change. Specialized representations can improve cliff prediction in defined tasks [6], but improvement of a local model does not eliminate the need to represent residual ranking uncertainty.

Reference data add a second source of false precision. ADMET measurements can contain experimental noise large enough to distort model fitting and candidate comparison; denoising work shows that identifying unreliable labels can improve downstream prediction in studied settings [7]. The appropriate inference is not that noise can always be removed, but that a candidate score inherits uncertainty from the data used to train and test the model. Ranking should therefore distinguish numerical separation from evidential separation: two candidates are meaningfully ordered only when the predictive difference remains credible under the uncertainty sources relevant to that comparison.

Aleatoric and epistemic uncertainty

A useful decision architecture separates uncertainty associated with data variability from uncertainty associated with what the model does not know. In practice, these components are not perfectly identifiable, but the distinction prevents experimental noise, sparse sampling, model misspecification, and parameter uncertainty from being treated as interchangeable. Comparative molecular-property studies show that scalable uncertainty estimators differ materially in performance; ensembles and bootstrap approaches were often stronger than Monte Carlo dropout in the evaluated graph-convolutional settings [8]. Other neural-network benchmarks similarly show that molecule-specific uncertainty estimates can be informative while remaining method- and task-dependent [9].

The decision value of uncertainty emerges when it changes what is done with a prediction. Evidential deep learning has produced uncertainty estimates correlated with prediction error and, in retrospective molecular screening, confidence filtering altered which candidates were favored [10]. That result supports uncertainty-aware selection as a feasible computational operation, but not the stronger claim that an evidential model is prospectively superior for medicinal chemistry. Likewise, in elementary-reaction property prediction, deep ensembles were the most reliable among the compared methods and could separate useful uncertainty components [11]. The chemical setting is different from drug-candidate ranking, yet the methodological lesson transfers cautiously: disagreement among uncertainty estimators is itself information about model dependence.

For pharmacoinformatics, the practical consequence is to retain at least two questions before ranking. First, how much irreducible or measurement-associated variability surrounds the target quantity? Second, how much uncertainty reflects incomplete model knowledge, representation, or training support? A single variance term may combine both, but the decision response need not be the same. Noisy endpoints may call for replicate measurement or assay qualification, whereas epistemic uncertainty may call for new chemical examples, alternative models, or deliberate exploration of unsupported regions.

chemical-space uncertainty

A candidate can be attractive according to a model while lying outside the region in which that model has been meaningfully tested. Real-world-oriented molecular OOD work shows that deployment-like covariate shifts can

degrade both predictive performance and uncertainty calibration [12]. Chemical novelty should therefore not be represented merely by a scaffold label. It is better treated as a model-specific support problem defined by the relationship among training chemistry, representation, endpoint, and intended deployment population.

This matters because common split strategies do not create equivalent notions of novelty. A broad comparison across molecular models, datasets, and ten split strategies found that scaffold splitting did not always produce a markedly harder OOD problem than random splitting [13]. The result cautions against using "scaffold split" as an automatic synonym for prospective generalization. Applicability-domain analysis leads to the same conclusion from another direction: model support is conditional on the particular model and dataset, and no single applicability-domain strategy is uniformly best [14].

Support boundaries also have decision consequences. Generative work can deliberately place molecules near applicability-domain boundaries [15]. illustrating that predicted desirability and support are separable dimensions. External physicochemical validation further shows that nominal prediction intervals may require rescaling and that uncertainty behavior differs by property [16]. Neither result can be transferred directly into a universal drug-ranking cutoff. Instead, they support a narrower rule for this Perspective: a virtual candidate should carry an explicit chemical-space support status alongside its predicted value, and movement toward unsupported chemistry should increase the burden of experimental confirmation rather than automatically lower or raise the candidate's rank.

Chemical-space uncertainty is therefore not a veto on innovation. A compound outside dense training support may be exactly the molecule worth testing. The decision error occurs when extrapolation is hidden. In the proposed architecture, unsupported but valuable candidates remain eligible, but they enter a different evidential state from equally valued candidates surrounded by close, well-characterized neighbors.

Model disagreement

Model disagreement provides another window onto epistemic risk. Neural architecture search studies show that molecular predictions and their associated uncertainty can change with architecture choice [17], while pharmaceutical activity benchmarks demonstrate that common uncertainty methods may still miss difficult errors in regions of steep structure–activity relationships [18]. Agreement is therefore not proof of correctness, and disagreement is not proof that a compound is poor. Both are diagnostic states whose meaning depends on chemical support and the failure modes shared by the models.

A particularly important recent result comes from ADME modeling under realistic data shifts: chemical-distance metrics and model-based variance captured complementary information, and learned error models that combined signals were generally more robust for ranking prediction error across several endpoints and shifts [19]. Representation-space distance has independently been shown to correlate with prediction error in neural chemical models, allowing confidence to be tightened by restricting predictions to better-supported regions [20]. These findings justify preserving data-support and model-disagreement channels separately until they are reconciled for a specific decision.

The manuscript therefore distinguishes transient from persistent disagreement. A new analogue may initially produce divergent predictions because it occupies sparse local chemistry; retraining with informative measurements may resolve that divergence. Persistent disagreement after relevant data are added suggests a different problem, such as model-form dependence, unresolved endpoint noise, or a local discontinuity. Conversely, stable agreement under a shared biased representation can create a false consensus. The decision consequence depends not simply on the magnitude of disagreement, but on when it arises, whether it persists, and what evidence could resolve it.

Before candidate ranking proceeds, the temporal logic of these disagreement states is easier to inspect comparatively than in prose alone.

Table 1. Temporal states of model disagreement and their consequences for virtual-candidate prioritization

State and entry conditions	Evidence through time	Transition and reassessment	
Experimental counterpart	**Entry condition / State-specific evidence / Persistence or transition**	**Decision state / Failure signature**	**Transfer boundary / Prospective test / Decision consequence**
Confirmatory assay on	**Entry condition:** Similar models agree within a well-sampled chemical neighborhood	**Decision state:** Supported concordance	**Transfer boundary:** Agreement does not establish external validity

representative candidates	**State-specific evidence:** Prediction order is stable across model choices and support diagnostics **Persistence or transition:** Persists unless new chemistry or endpoint data shift the domain	**Failure signature:** Hidden shared bias remains possible	**Prospective test:** Time-forward test on related chemistry **Decision consequence:** Rank can carry greater provisional weight
Measure the candidate and informative near neighbors	**Entry condition:** A candidate extends beyond dense local support and plausible models separate **State-specific evidence:** Distance and model variance rise together or differently **Persistence or transition:** May resolve after local measurements are added	**Decision state:** Sparse-neighborhood divergence **Failure signature:** Extrapolation appears as unstable ordering	**Transfer boundary:** Local support does not imply program-wide transfer **Prospective test:** Refit and test whether rank stabilizes **Decision consequence:** Defer forced ordering; prioritize information gain
Orthogonal assay or representation-sensitive experiment	**Entry condition:** Disagreement remains after relevant local data are incorporated **State-specific evidence:** Competing representations or architectures retain different orderings **Persistence or transition:** Persists across retraining or resampling	**Decision state:** Persistent model-form conflict **Failure signature:** Rank depends on model family	**Transfer boundary:** No model consensus can be treated as mechanistic truth **Prospective test:** Prespecified comparison across model families **Decision consequence:** Preserve competing ranks; avoid single-score commitment
Test chemically distinct or time-forward compounds	**Entry condition:** Models agree because they share data, descriptors, or inductive biases **State-specific evidence:** Low apparent disagreement despite weak external support **Persistence or transition:** Can persist until genuinely shifted data arrive	**Decision state:** False consensus risk **Failure signature:** Uniform confidence with later error	**Transfer boundary:** Internal agreement does not establish deployment calibration **Prospective test:** Challenge with future-like chemistry **Decision consequence:** Down-weight consensus when support evidence is weak
Repeat or orthogonal measurement where noise is plausible	**Entry condition:** New measurements discriminate among prior models **State-specific evidence:** One or more uncertainty channels contract after informative data **Persistence or transition:** Moves to supported concordance or exposes persistent conflict	**Decision state:** Evidence-resolved transition **Failure signature:** Expected reduction fails to occur	**Transfer boundary:** Resolution is endpoint- and domain-specific **Prospective test:** Confirm stability on subsequent candidates **Decision consequence:** Update rank only after evidence changes the uncertainty state

Note. The decision-state labels and decision consequences are proposed synthesis; they are qualitative and do not encode a numerical confidence scale.

The same logic is summarized visually as a chemical-space support field in which disagreement is interpreted relative to support boundaries rather than as a free-standing confidence score.

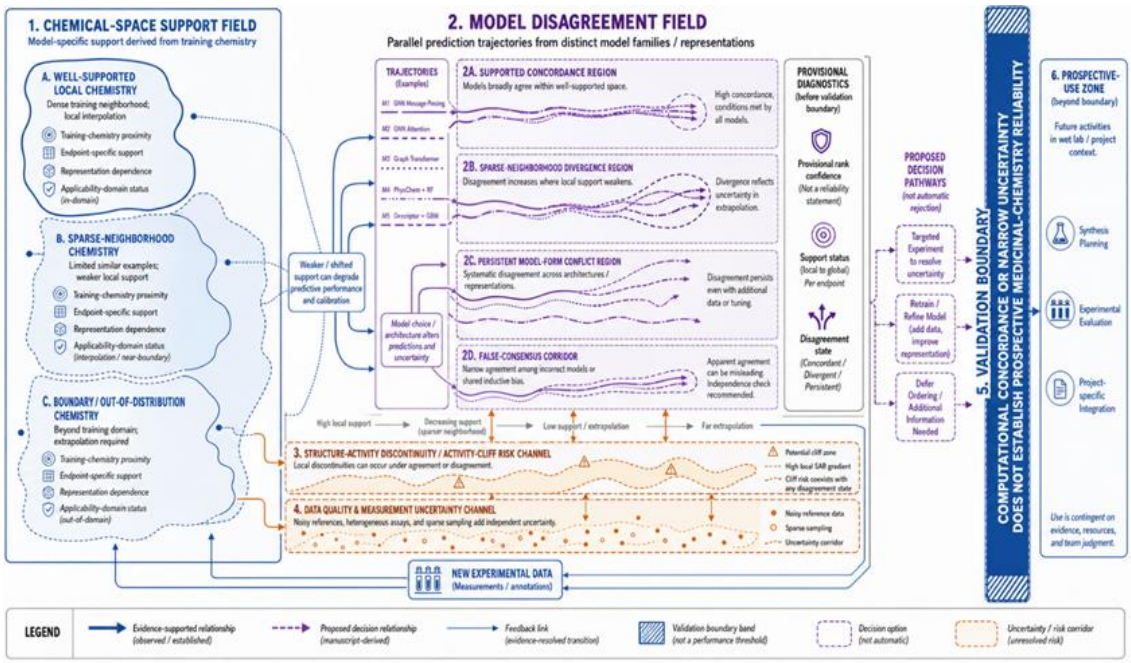


Figure 1. Chemical-Space Support, Model Discordance, and the Limits of Rank Confidence

Calibration

Uncertainty becomes decision-grade only when its numerical expression is tested against observed outcomes. Conformalized graph learning for ADMET prediction demonstrates that prediction intervals can be constructed and evaluated for empirical coverage rather than assumed from model confidence alone [21]. Calibration, however, is not equivalent to usefulness. A method may achieve nominal coverage with intervals so broad that the candidate ranking remains uninformative.

Conformal prediction software for cheminformatics makes this distinction operational by separating validity-oriented coverage from efficiency and by exposing the dependence of predictions on calibration data and methodological assumptions [22]. Large-scale comparison of conventional QSAR and conformal approaches across hundreds of targets using later published data further shows that uncertainty-aware methods must be judged in the downstream prediction setting rather than only on the data from which they were built [23].

For candidate prioritization, calibration should therefore be treated as a transfer gate. A calibrated interval in one chemical series, assay regime, or historical dataset is evidence about that setting; it is not a transferable guarantee. The relevant question is whether uncertainty retains its advertised behavior for the chemistry and decision horizon being ranked now. This makes calibration a cross-scale problem linking observed residual behavior, molecular support, model output, and the final decision state.

Table 2 makes those transfer conditions explicit without treating calibration at one level as automatic validation at the next.

Table 2. Calibration transfer conditions from molecular evidence to candidate-ranking decisions

Source state and scale		Transfer conditions	Qualification endpoint
Analytical scale / Governing condition / Molecular or chemical state	**Observable or model output / Sampling or data support / Experimental counterpart**	**Failure signature**	**Prospective test / Framework endpoint**
Analytical scale: Reference-data level **Governing condition: Residual behavior is estimable from relevant measurements** **Molecular or chemical state: Compound measurements reflect the endpoint being modeled**	**Observable or model output:** Observed error or residual distribution **Sampling or data support:** Replicates or sufficiently informative reference observations **Experimental counterpart:** Repeat measurement or assay-quality review	Apparent precision driven by noisy or inconsistent labels	**Prospective test:** Test error behavior in newly generated data **Framework endpoint:** Retain or expand data-related uncertainty

Analytical scale: Local chemical-support level **Governing condition: Calibration set represents the candidate neighborhood adequately** **Molecular or chemical state: Candidate lies within or near modeled support**	**Observable or model output:** Distance/support diagnostic plus prediction **Sampling or data support:** Chemically relevant neighbors **Experimental counterpart:** Add near-neighbor compounds or orthogonal chemistry	Nominal interval remains narrow despite weak support	**Prospective test:** Evaluate coverage as chemistry moves outward **Framework endpoint:** Mark support-qualified or support-limited prediction
Analytical scale: Model level **Governing condition: Predictive uncertainty tracks error for the fitted model family** **Molecular or chemical state: Multiple plausible representations remain compatible with data**	**Observable or model output:** Ensemble, evidential, conformal, or error-model output **Sampling or data support:** Training and calibration samples match intended use **Experimental counterpart:** Compare model families and retraining responses	Low uncertainty on wrong predictions or unstable uncertainty across models	**Prospective test:** Time-forward error-ranking and calibration test **Framework endpoint:** Preserve model-dependent uncertainty state
Analytical scale: Prediction-interval level **Governing condition: Empirical coverage and interval efficiency are both acceptable for the task** **Molecular or chemical state: Candidate-specific interval is interpretable in the endpoint scale**	**Observable or model output:** Prediction set or interval **Sampling or data support:** Valid calibration sample under stated assumptions **Experimental counterpart:** Obtain discriminating assay result	Correct nominal coverage with unusably broad intervals, or undercoverage	**Prospective test:** Prespecified coverage and efficiency on future-like compounds **Framework endpoint:** Decide whether prediction can influence rank
Analytical scale: Candidate-decision level **Governing condition: Calibration remains credible under the actual ranking context** **Molecular or chemical state: Candidate competes with alternatives with different support states**	**Observable or model output:** Provisional rank plus uncertainty descriptors **Sampling or data support:** Evidence sufficient for the intended decision **Experimental counterpart:** Targeted experiment chosen for expected decision value	Rank reversal under plausible calibration or support changes	**Prospective test:** Compare uncertainty-aware and score-only prioritization prospectively **Framework endpoint:** Promote, down-weight, defer, or test; no automatic rejection

Note. The framework endpoint is proposed decision logic. Calibration at one analytical scale does not establish transferability to the next.

Decision thresholds and utility

A decision threshold converts predictive evidence into action, but no universal threshold follows from model accuracy alone. Distance-aware molecular-property modeling shows that lack of support in nonlinear structure–property space can be associated with higher prediction difficulty [24]. Such evidence justifies using support as a decision variable; it does not justify importing a published numerical distance cutoff into an unrelated medicinal-chemistry program.

The proposed framework therefore treats thresholds as context-dependent commitments rather than fixed properties of a model. A small uncertainty penalty may be appropriate when synthesis and testing are inexpensive and false rejection is costly. A much stronger evidential requirement may be appropriate when advancing a candidate triggers expensive profiling or when the modeled endpoint carries substantial downstream consequence. Because those utilities are program-specific, this Perspective does not assign numerical weights.

Crucially, an uncertain prediction need not trigger rank reversal. The appropriate decision may be abstention, deferred ordering, or a targeted experiment selected to resolve the uncertainty that matters most. High predicted value with weak chemical-space support is therefore different from modest predicted value with strong support, and both are different from model disagreement driven by unresolved representation choice. Decision thresholds should operate on these states rather than on predicted value alone. The next section will use this logic to define how rankings can be constructed without collapsing heterogeneous uncertainty into a single, falsely precise leaderboard.

Ranking under uncertain predictions

Ranking under uncertainty should begin by refusing to compress unlike uncertainties prematurely. A candidate with a high predicted value and weak chemical-space support differs from one with moderate predicted value,

strong support, and narrow calibrated uncertainty. Likewise, two candidates with similar mean predictions may differ because one is stable across model families while the other changes order under plausible representations. Comparative UQ studies show that no single uncertainty measure performs uniformly across settings [3], and pharmaceutical activity benchmarks show that uncertainty estimators may fail precisely in steep structure–activity regions [18]. The practical implication is that a ranking should preserve the dimensions responsible for uncertainty until their decision relevance is explicit.

A robust ranking is therefore not necessarily a different total order. It may contain tied, deferred, or experimentally contingent states. Chemical-distance and model-derived uncertainty can provide complementary information [19], while architecture choice and latent-space support can change apparent confidence [17]. Candidate order should consequently be stress-tested across plausible model and support assumptions. If a nominal top candidate falls sharply when one reasonable representation or support definition changes, its rank is fragile rather than simply "high."

This Perspective proposes that uncertainty modifies the burden of evidence rather than mechanically subtracting a penalty from every score. Confidence filtering has already been shown to alter retrospective molecular selection [10], but that does not establish a universal penalty function. The relevant output is instead a rank state: provisionally supported, rank-fragile, support-limited, disagreement-limited, or evidence-seeking. These states are proposed labels, not empirically validated classes.

A risk-adjusted candidate-prioritization framework

The proposed framework reconciles five elements before a candidate is advanced: predicted value, data-related uncertainty, chemical-space support, model disagreement, and calibration. These elements remain separate until the decision boundary because their failure modes differ. Large ADME benchmarks support combining nonredundant uncertainty signals for error ranking [19], but they do not establish the weighting system proposed here. Utility enters only at the final decision layer, where the cost of a false advance, false rejection, or additional experiment is program-specific.

The framework therefore uses evidence reconciliation rather than an opaque composite score. Concordant high-value predictions with strong support and acceptable calibration can receive greater provisional weight. High-value predictions outside support remain candidates, but carry an experimental-resolution requirement. Persistent model disagreement prevents confident total ordering, while apparently strong consensus with poor external support remains vulnerable to shared bias. Activity-cliff risk is retained as a separate failure signature because low ensemble disagreement does not guarantee smooth local structure–activity behavior [18].

The framework endpoint is not a probability of development success. It is a proposed decision state indicating whether current computational evidence justifies promotion, down-weighting, deferral, or a discriminating experiment. **Table 3** reconciles the principal evidence patterns without imposing numerical scores.

Table 3. Evidence reconciliation before committing to a virtual-candidate rank

Evidence configuration		Interpretive alternatives	Adjudication requirement
Evidence pattern / Molecular or chemical state	**Justified inference / Experimental counterpart**	**Transfer boundary / Failure signature / Inference not justified**	**Prospective test / Candidate-prioritization endpoint**
Evidence pattern: Concordant and well supported Molecular or chemical state: Candidate lies within relevant local support and plausible models retain similar order	**Justified inference:** Rank is comparatively stable under tested assumptions **Experimental counterpart:** Confirmatory assay	**Transfer boundary:** Internal support does not guarantee future chemistry **Failure signature:** Shared model bias may remain hidden **Inference not justified:** Candidate is proven superior	**Prospective test:** Time-forward stability test **Candidate-prioritization endpoint:** Provisional promotion
Evidence pattern: High value with weak support Molecular or chemical state: Candidate lies near or beyond the model's supported region	**Justified inference:** Candidate is promising but evidence-limited **Experimental counterpart:** Measure	**Transfer boundary:** Extrapolation limits transfer from training chemistry **Failure signature:** Attractive score depends on unsupported chemistry	**Prospective test:** Coverage and rank test after local expansion **Candidate-prioritization endpoint:** Evidence-seeking priority

	candidate and informative neighbors	**Inference not justified:** Candidate should automatically be rejected or promoted	
Evidence pattern: Persistent model conflict Molecular or chemical state: Plausible models retain different orderings after relevant refitting	**Justified inference:** Current evidence does not determine a stable total order **Experimental counterpart:** Orthogonal assay or representation-sensitive experiment	**Transfer boundary:** Model family constrains inference **Failure signature:** Rank changes with representation or architecture **Inference not justified:** Averaged prediction represents consensus truth	**Prospective test:** Prespecified cross-model comparison **Candidate-prioritization endpoint:** Deferred ordering
Evidence pattern: Low disagreement with local cliff risk Molecular or chemical state: Models agree but nearby structural changes show unstable activity relationships	**Justified inference:** Agreement alone is insufficient for confidence **Experimental counterpart:** Matched-analogue or orthogonal activity testing	**Transfer boundary:** Agreement does not remove local SAR discontinuity **Failure signature:** False consensus around a cliff **Inference not justified:** Low variance guarantees reliability	**Prospective test:** Prospective cliff-sensitive challenge **Candidate-prioritization endpoint:** Down-weight apparent consensus
Evidence pattern: Calibrated but decision-ambiguous Molecular or chemical state: Interval behavior is acceptable but candidates remain insufficiently separated	**Justified inference:** Prediction uncertainty is honestly represented **Experimental counterpart:** Experiment chosen for expected discrimination	**Transfer boundary:** Calibration does not determine utility **Failure signature:** Valid uncertainty remains too broad for ordering **Inference not justified:** Calibration selects the best candidate automatically	**Prospective test:** Compare decision yield with score-only ranking **Candidate-prioritization endpoint:** Targeted evidence acquisition

Note. The row states and prioritization endpoints are proposed synthesis. They do not constitute validated confidence grades or numerical decision thresholds.

The framework's central decision boundary is summarized in **Figure 2**, which separates predictive trajectories from justified candidate action.

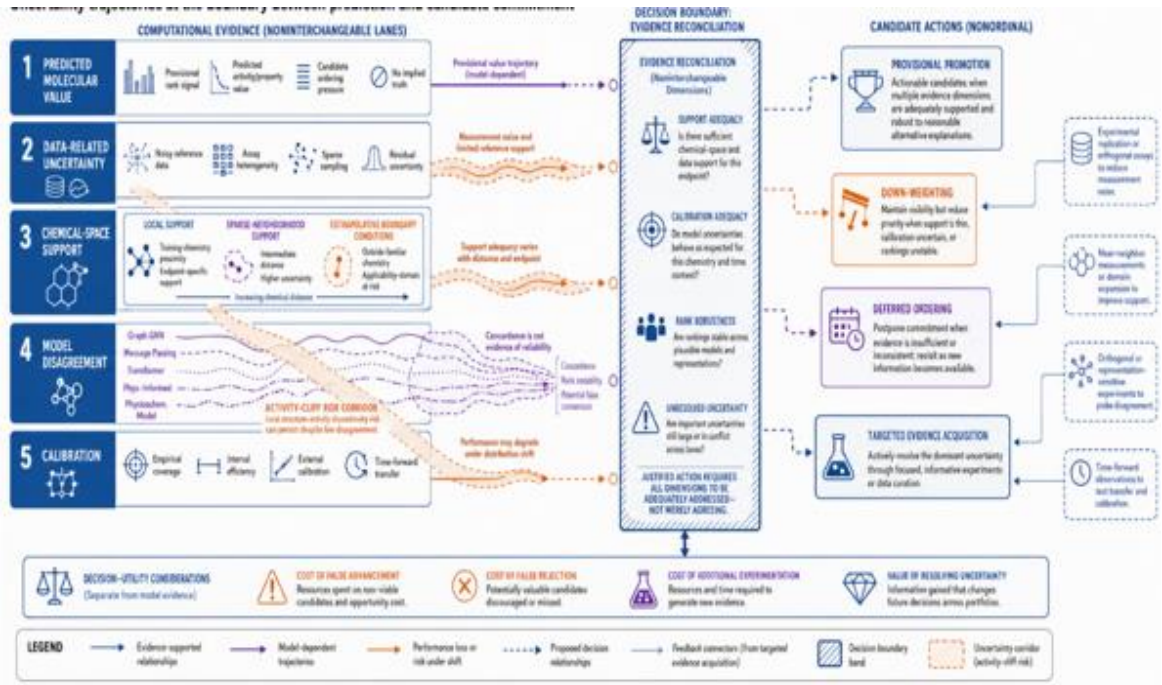


Figure 2. Uncertainty Trajectories at the Boundary Between Prediction and Candidate Commitment

Prospective evaluation

The proposed architecture should be tested prospectively rather than defended by retrospective plausibility. Molecular-property benchmarking shows that random splitting can overstate future-like performance and that

chronological evaluation is preferable when available [25]. A prospective test should therefore freeze the model, uncertainty procedures, candidate-ranking rule, and comparator before new compounds are evaluated.

Evaluation should compare uncertainty-aware prioritization with a prespecified score-only or existing workflow. Outcomes should include prediction error, calibration, rank stability, frequency of deferred decisions, and whether targeted experiments resolve uncertainty efficiently. Distribution shift should be characterized explicitly rather than assumed from a scaffold label. Because the framework is non-empirical, no performance advantage is claimed here.

A particularly informative design would stratify analysis by chemical support and disagreement state before outcomes are known. That prevents a favorable average result from hiding failure in the candidates for which uncertainty awareness is most needed. Recalibration events, rank reversals after new measurements, and cases in which models agree yet fail should be retained as primary diagnostic outcomes rather than discarded as exceptions. The relevant comparison is therefore not merely whether one ranking achieves a lower global error, but whether uncertainty-aware prioritization identifies when computational evidence is insufficient and routes those cases to useful evidence generation.

Reporting standards

Reports should disclose the data split, chemical-space definition, model family, uncertainty estimator, calibration procedure, applicability-domain method, and intended decision use. Evaluation guidance for chemical machine learning emphasizes that claims must match the experimental or computational design used to test them [26]. Accordingly, an uncertainty-aware ranking paper should state whether validation is random, scaffold-based, chronological, external, or genuinely prospective and should describe the shift being approximated.

Reproducibility also requires transparent reporting of data provenance, preprocessing, representations, model selection, hyperparameter procedures, and code or model availability where feasible [27]. Uncertainty-specific reporting should additionally define what each reported interval, variance, distance, or disagreement measure represents. Calibration and predictive accuracy should be reported separately, as should cases in which uncertainty fails to identify large errors. Proprietary-data restrictions should be stated rather than obscured, because reproducibility and external validity are related but distinct properties.

CONCLUSION

Pharmacoinformatics should not rank virtual candidates as though a predicted value were self-validating. Numerical order becomes decision-relevant only when the supporting data, chemical neighborhood, model dependence, calibration, and intended utility are considered together. The proposed framework therefore preserves heterogeneous uncertainty until a transparent decision boundary and allows promotion, down-weighting, deferral, or targeted experimentation instead of forcing every compound into a falsely precise leaderboard. Its value remains a hypothesis for prospective testing, not an established improvement in drug-development success. The central requirement is simpler: uncertainty should enter the ranking decision before computational precision is mistaken for evidence.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Catacutan DB, Alexander J, Arnold A, Stokes JM. Machine learning in preclinical drug discovery. Nat Chem Biol. 2024;20(8):960-73. doi:10.1038/s41589-024-01679-1
2. Frömbgen T, Surzhikova E, Dölz J, Proppe J, Kirchner B, Jacob CR. Uncertainty quantification for in silico chemistry. Chem Rev. 2026;126(7):4189-236. doi:10.1021/acs.chemrev.5c00931
3. Lanini J, Huynh MTD, Scebba G, Schneider N, Rodríguez-Pérez R. UNIQUE: A framework for uncertainty quantification benchmarking. J Chem Inf Model. 2024;64(22):8379-86. doi:10.1021/acs.jcim.4c01578

4. 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
5. 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
6. Niu D, Deng Z, Wang X, Li Z. Task-specific activity cliff prediction method based on transfer learning and a hyper connection graph model. J Chem Inf Model. 2025;65(15):8345-59. doi:10.1021/acs.jcim.5c00949
7. Adrian M, Chung Y, Cheng AC. Denoising drug discovery data for improved absorption, distribution, metabolism, excretion, and toxicity property prediction. J Chem Inf Model. 2024;64(16):6324-37. doi:10.1021/acs.jcim.4c00639
8. 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
9. 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
10. 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
11. Liu Y, Mo Y, Cheng Y. Uncertainty qualification for deep learning-based elementary reaction property prediction. J Chem Inf Model. 2024;64(21):8131-41. doi:10.1021/acs.jcim.4c01358
12. Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. J Chem Inf Model. 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
13. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. J Chem Inf Model. 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
14. Piazza L, Poles C, Bononi G, Granchi C, Di Stefano M, Poli G, et al. ADvisor: An open-source tool for applicability domain definition and optimization in molecular predictive modeling. J Chem Inf Model. 2025;65(21):11583-600. doi:10.1021/acs.jcim.5c01672
15. Uchibori Y, Kaneko H. Generation of molecules near the applicability domain boundaries of property prediction models. J Chem Inf Model. 2026;66(12):6866-79. doi:10.1021/acs.jcim.5c03220
16. Brown TN, Sangion A, Arnot JA. Identifying uncertainty in physical-chemical property estimation with IFSQSAR. J Cheminform. 2024;16(1):65. doi:10.1186/s13321-024-00853-w
17. Jiang S, Qin S, Van Lehn RC, Balaprakash P, Zavala VM. Uncertainty quantification for molecular property predictions with graph neural architecture search. Digit Discov. 2024;3(8):1534-53. doi:10.1039/D4DD00088A
18. Kötter A, Singh K, Matter H, Hessler G, Grebner C. Upgrading reliability in molecular property prediction by robust quantification of uncertainty from machine learning models. J Chem Inf Model. 2025;65(20):10819-31. doi:10.1021/acs.jcim.5c00464
19. Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. J Chem Inf Model. 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
20. Janet JP, Duan C, Yang T, Nandy A, Kulik HJ. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. Chem Sci. 2019;10(34):7913-22. doi:10.1039/C9SC02298H
21. Li P, Hua L, Ma Z, Hu W, Liu Y, Zhu J. Conformalized graph learning for molecular ADMET property prediction and reliable uncertainty quantification. J Chem Inf Model. 2024;64(23):8705-17. doi:10.1021/acs.jcim.4c01139
22. Arvidsson McShane S, Norinder U, Alvarsson J, Ahlberg E, Carlsson L, Spjuth O. CPSign: Conformal prediction for cheminformatics modeling. J Cheminform. 2024;16(1):75. doi:10.1186/s13321-024-00870-9
23. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11(1):4. doi:10.1186/s13321-018-0325-4
24. Kim JY, Vlachos DG. Distance-aware molecular property prediction in nonlinear structure-property space. J Chem Inf Model. 2025;65(13):6744-56. doi:10.1021/acs.jcim.5c01037
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. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
27. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z