Pharmacophore

ISSN-2229-5402

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

From Predicted Binding to Defensible Candidate Advancement: A Conceptual Framework for Integrating Mechanistic Plausibility, Translational Uncertainty, and Decision Thresholds in Artificial Intelligence–Driven Drug Discovery

Elena Petrova[1], Marcus Weber[2]*, Priya Nair[3], Thomas Andersen[4]

1. ***Department of AI-Driven Drug Discovery, Faculty of Pharmacy, Sofia University, Sofia, Bulgaria.***
2. ***Department of Computational Pharmacology and Molecular Informatics, Faculty of Medicine, LMU Munich, Munich, Germany.***
3. ***Department of Pharmaceutical Data Science and Cheminformatics, Indian Institute of Technology Bombay, Mumbai, India.***
4. ***Department of Drug Design and Translational Pharmacology, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.***

ARTICLE INFO

Received:
16 April 2026
Received in revised form:
27 August 2026
Accepted:
02 September 2026
Available online:
28 October 2026

Keywords: AI-driven drug discovery, Candidate advancement, Mechanistic plausibility, Translational evidence, Uncertainty calibration, Predictive validity

ABSTRACT

Artificial intelligence increasingly supports molecular generation, structural prediction, affinity estimation and property prediction in drug discovery, yet a favorable computational score does not itself establish that a candidate should advance. The central problem is therefore not whether AI can predict binding or related molecular properties, but how such predictions should be interpreted alongside mechanistic plausibility, translational evidence, uncertainty and the consequences of an advancement decision. This conceptual article develops an evidence-bounded candidate-advancement account that separates prediction quality from candidate-worthiness. It first identifies the justificatory limits of prediction-centred selection, including state dependence, benchmark leakage, out-of-distribution failure, calibration limitations and the conditional relevance of biological data. It then examines evidence that can strengthen or weaken an advancement case, including target engagement, pharmacological characterization, target assessment, orthogonal experimental confirmation and human translation. The resulting argument treats evidence layers as heterogeneous rather than automatically commensurable and proposes that uncertainty changes the burden of evidence required for progression Favorable AI predictions can therefore remain insufficient when mechanistic incoherence, unresolved distribution shift, weak target biology or adverse translational evidence persists. Conversely, a candidate can become more defensible as distinct uncertainties are retired through evidence appropriate to their level. Candidate advancement should thus be framed as a state-dependent decision under uncertainty, not as the maximization of a single predictive score.

*This is an **open-access** article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, and build upon the work non commercially.*

To Cite This Article: Petrova E, Weber M, Nair P, Andersen T. From Predicted Binding to Defensible Candidate Advancement: A Conceptual Framework for Integrating Mechanistic Plausibility, Translational Uncertainty, and Decision Thresholds in Artificial Intelligence–Driven Drug Discovery. Pharmacophore. 2026;17(5):9-16. https://doi.org/10.51847/rSRfIoYpgX

Introduction

Artificial intelligence now contributes to target identification, molecular generation, structural modelling and property prediction across the drug-discovery workflow. Its growing reach, however, creates a distinction that is easy to obscure: improving the prediction of a molecular attribute is not the same as improving the decision to commit resources to a candidate. Contemporary reviews emphasize both the expanding technical potential of AI and the continuing dependence of drug discovery on experimentally grounded evidence, later-stage validation and expert judgment [1, 2]. The relevant decision problem is therefore not whether an AI system can rank molecules more quickly, but whether the information it produces changes the probability that the compound selected for further investment is genuinely worth advancing.

Corresponding Author: Marcus Weber; Department of Computational Pharmacology and Molecular Informatics, Faculty of Medicine, LMU Munich, Munich, Germany. E-mail: marcus.weber@lmu.de.

That distinction matters because candidate selection occurs in a low-base-rate environment in which most projects fail and decision errors compound over time. Predictive validity provides a useful foundation for this problem because it concerns the relationship between the output of a decision tool and the downstream outcome the tool is intended to inform [3]. A model may perform well against an internal benchmark while contributing little to the clinically meaningful discrimination between candidates if the benchmark is weakly connected to the eventual decision consequence. For candidate advancement, model performance therefore has to be interpreted relative to the decision being made, the uncertainty surrounding that decision and the cost of being wrong.

Mechanistic information can narrow this gap, but only if it makes the molecular hypothesis more testable rather than merely more interpretable. Physics-informed and explainable AI approaches can connect predictions to interaction geometry, energetic constraints or other domain-informed features, thereby improving the scientific tractability of a model-generated hypothesis [4]. Yet even a coherent mechanistic account at the molecular level does not establish that the intended biology will occur in the relevant system, that exposure will be adequate, or that efficacy will outweigh toxicity. Mechanistic plausibility is therefore an intermediate evidentiary achievement: stronger than an unexplained score, but still distinct from translational adequacy.

This article addresses that distinction as a candidate-advancement problem. Critical analyses of generative AI in medicinal chemistry have already argued that computational novelty requires rigorous real-world evaluation rather than reliance on impressive in silico outputs alone [5]. Building on that premise, the analysis below asks what different evidence layers can legitimately justify, how uncertainty changes the evidence required to progress, and under what conditions a favorable AI prediction should nevertheless result in pause, redirection or termination. The objective is not to replace quantitative modelling with qualitative judgment. It is to specify where model outputs end, where additional evidence becomes decision-relevant, and why those evidentiary transitions matter for defensible candidate selection.

Limitations of Prediction-Centred Candidate Selection

Prediction-centred selection is attractive because it offers a seemingly coherent basis for ranking large chemical spaces. Structural AI has substantially expanded what can be predicted: AlphaFold 3, for example, extends joint structure prediction across proteins, nucleic acids, small molecules, ions and modified residues [6]. Such capability can make a predicted binding configuration more informative than a simple similarity score. It still remains a structural hypothesis. Protein–ligand interactions are state dependent, and state-specific generative approaches such as NeuralPLexer explicitly model ligand-bound conformations rather than assuming that a single protein representation captures all relevant configurations [7]. A predicted complex can therefore support plausibility at one level while leaving functional state, occupancy, exposure and downstream biological consequence unresolved.

Benchmark performance creates a second limitation because the apparent quality of a prediction depends on how the test resembles the eventual use case. Redundancy between training and evaluation sets can inflate performance, while genuinely novel targets, scaffolds or structural contexts impose a harder generalization problem. Recent studies show that stricter external testing, unseen-entity splits and the quality of predicted structural inputs materially influence reported affinity or scoring performance [8-10]. These findings do not invalidate machine-learning-based scoring. They instead define a justificatory ceiling: benchmark success can support confidence that a model has learned something useful about a defined task, but it cannot by itself establish that the same reliability will hold for a candidate that differs materially from the training distribution or is being advanced for a new biological context.

Out-of-distribution uncertainty makes this ceiling operationally important. In deployment-like molecular settings, distribution shift can substantially reduce both predictive performance and calibration [11]. The implication for candidate selection is not that every out-of-distribution molecule should be rejected, but that the evidentiary burden should rise when a candidate lies farther from the region in which the model has been validated. Uncertainty quantification can help, yet it should not be treated as an infallible confidence measure. Comparative work across molecular property tasks found no single neural uncertainty-quantification approach that was consistently superior across all datasets and evaluation metrics [12]. At the same time, individual-level approaches can provide useful additional information: multimodal measures of representation divergence have improved out-of-distribution binding-affinity assessment in polypharmacology settings [13]. These results support a limited but important conclusion: uncertainty estimates can help decide where more evidence is needed, but uncertainty itself requires validation and should not be converted uncritically into a scalar candidate value.

A final limitation is that chemical and biological data are not interchangeable with generic high-volume machine-learning data. Their meaning depends on the hypothesis being tested, assay conditions, target context, measurement quality and the biological relevance of the endpoint. Analyses of AI in drug discovery have emphasized that chemical and especially biological datasets are constrained by scarcity, conditionality and uneven relevance to the decisions for which models are ultimately used [14]. A larger dataset can therefore improve statistical learning without necessarily improving translational relevance. For advancement decisions, the central question is not simply how much evidence exists, but what uncertainty that evidence can actually resolve.

These distinctions change the interpretation of prediction-centred evidence. A structural model can support a binding hypothesis; an affinity model can support comparative ranking; an uncertainty model can indicate where extrapolation is hazardous. None of these layers alone determines whether the candidate is defensible to advance. **Table 1** specifies the evidentiary reach of the principal prediction-centred layers and the additional questions that remain open once each layer has been satisfied.

Table 1. Prediction-centred evidence layers and their justificatory limits for candidate advancement.

Evidence layer	What a favorable result can justify	What remains unresolved	Typical failure mode	Advancement implication
Predicted complex geometry	A structurally plausible interaction hypothesis	Functional engagement, relevant state occupancy, efficacy and exposure	Reliance on one predicted conformation or incorrect local geometry	Treat as a mechanistic hypothesis requiring orthogonal confirmation
Predicted binding or affinity	Relative support for interaction strength within a validated task domain	Causal mechanism, selectivity in the relevant system, tissue exposure and therapeutic effect	Dataset leakage, weak generalization or extrapolation to unseen targets/scaffolds	Raise confidence only within demonstrated applicability
Virtual-screening rank	Prioritization of compounds for follow-up	Whether rank differences are biologically meaningful or robust to assay context	Memorization of benchmark patterns or sensitivity to decoys	Use for experiment ordering, not as a stand-alone advancement criterion
Model-derived uncertainty	An estimate of how reliable a prediction may be under the fitted uncertainty method	Whether the uncertainty model is calibrated for the candidate and decision context	Overconfident out-of-distribution prediction or method-specific miscalibration	Increase confirmatory evidence requirements when uncertainty is high or poorly characterized
Large chemical/biological dataset support	Broader statistical support for learned patterns	Whether the measured endpoint captures the biology relevant to advancement	High-volume but low-relevance measurements	Judge data by decision relevance, not size alone

Note: The advancement implications in the final column are proposed synthesis derived from the analytical distinctions developed in this section; they are not universal quantitative thresholds.

Mechanistic and Translational Evidence for Candidate Advancement

If prediction-centred evidence defines what is not yet known, candidate advancement requires evidence that narrows uncertainty at levels relevant to mechanism and translation. Direct target engagement is one such layer. Target-engagement assays can determine whether a compound physically interacts with its intended target in increasingly realistic experimental contexts, supporting structure–activity relationships and mechanism-of-action claims beyond predicted binding alone [15]. Even this evidence has a boundary: engagement does not establish that the interaction produces the desired biological effect. Deeper pharmacological characterization can therefore alter candidate assessment by revealing efficacy profiles, signaling bias, allosteric behavior or other context-dependent properties that are invisible to a simple affinity estimate [16]. The practical value of mechanistic evidence lies in its ability to retire a specific uncertainty, not in its proximity to the computational prediction.

The same logic applies one level upstream, at target selection. AI can accelerate target identification by integrating genetic, molecular and phenotypic information, but target ranking remains distinct from target validation [17]. A candidate cannot be more defensible than the biological hypothesis it is designed to exploit. Structured target-assessment approaches therefore evaluate dimensions such as causal disease biology, safety, druggability, assayability and differentiation before substantial development resources are committed [18]. These dimensions should not be collapsed into a single score. A target can be compelling genetically yet problematic pharmacologically; it can be druggable but poorly differentiated; it can be measurable in vitro while lacking a credible path to patient benefit. Advancement becomes stronger when these potential contradictions are exposed rather than averaged away.

Orthogonal experimental escalation provides a way to test whether a computationally favorable molecule survives contact with measurements that were not used to generate it. PocketFlow illustrates this progression: generated compounds were taken beyond in silico scoring into biochemical or biophysical evaluation and structural confirmation in specific target programmes [19]. The importance of such examples is not that they establish an overall success rate for AI-designed molecules. Their value is methodological. Each orthogonal test changes the epistemic status of the candidate by addressing a different uncertainty. A prediction that survives synthesis and activity testing is not equivalent to one that exists only computationally; a candidate that also receives structural confirmation occupies a different evidentiary position again.

Human testing forms a qualitatively different translational boundary because it exposes a candidate to biological complexity that cannot be reproduced by molecular prediction alone. Rentosertib, an AI-discovered and AI-generated TNIK inhibitor, progressed to randomized phase 2a evaluation in idiopathic pulmonary fibrosis [20]. This is important evidence that AI-originated assets can move through the discovery stack into controlled clinical testing. It is not evidence that AI-originated drugs as a class are clinically superior, nor that the upstream AI methods caused any observed therapeutic effect. For the present framework, the trial is instead a boundary marker: human evidence is qualitatively different from molecular and preclinical evidence, and its interpretation remains candidate-, indication- and design-specific.

Candidate advancement is therefore better understood as staged derisking than as the accumulation of interchangeable positive signals. Academic drug-discovery models emphasize targeted investments that progressively reveal translational potential,

while industrial frameworks and candidate-target-profile approaches use explicit multidimensional criteria to guide nomination and progression [21-23]. These approaches differ in setting and cannot be treated as universal decision rules, but they converge on an important principle: advancement decisions improve when the evidence required for the next commitment is specified in relation to the risks that remain unresolved.

That principle also explains why a candidate can justifiably pause or redirect despite favorable computational scores. If target engagement is absent, the predicted binding hypothesis is weakened. If target biology remains uncertain, stronger chemistry may not solve the programme-level problem. If pharmacology reveals a context-dependent liability, further optimization may need to change mechanism rather than simply improve potency. If human or translational evidence conflicts with preclinical expectations, the appropriate response may be to stop rather than to search for another model that reproduces the desired ranking. **Table 2** therefore treats candidate advancement as a set of decision states defined by evidence sufficiency, unresolved uncertainty and counterevidence rather than by a single composite score.

Table 2. Candidate-advancement decision conditions under heterogeneous evidence and uncertainty.

Decision state	Evidence configuration	Uncertainty condition	Counterevidence pattern	Required next evidence
Progress	Prediction is supported by coherent mechanism and relevant translational evidence	Remaining uncertainty is bounded and proportionate to the next development commitment	No major contradiction at target, pharmacology, exposure or safety level	Evidence appropriate to the next development stage
Pause	Core hypothesis remains plausible, but one or more decision-critical uncertainties are unresolved	Uncertainty is material and potentially reducible	Evidence is incomplete rather than clearly adverse	Targeted experiment designed to retire the specific unresolved uncertainty
Redirect	The candidate retains useful value, but the original mechanism, indication, exposure strategy or optimization objective is no longer adequately supported	Uncertainty points to a change in the path rather than simple confirmation	Local contradiction can potentially be addressed by reframing or redesign	Evidence testing the revised mechanistic or translational hypothesis
Stop	Prediction is outweighed by non-resolvable or high-consequence counterevidence	Remaining uncertainty cannot be reduced at acceptable scientific or resource cost, or the expected value of doing so is low	Mechanistic incoherence, unacceptable liability, weak target rationale or translational failure dominates	No routine confirmatory evidence; termination or major programme redefinition

Note: These decision states are proposed analytical categories, not empirically validated universal gates. Their use depends on programme context, disease area, modality, available alternatives and the consequences of incorrect advancement.

Proposed Candidate-Advancement Framework

Candidate advancement should be treated as an inference across heterogeneous evidence layers rather than as the accumulation of positive scores. Predictive validity is relevant because a model must discriminate outcomes that matter to the eventual decision [3]. Target engagement adds evidence that an intended molecular interaction occurs [15], while deeper pharmacological characterization asks what that interaction does in a biological system [16]. Structured target assessment adds still another level by testing disease relevance, safety, druggability, assayability and differentiation [18]. These forms of evidence are connected but not interchangeable. Stronger support at one level cannot automatically erase a serious contradiction at another. The proposed framework therefore assigns each layer a justificatory reach: evidence should influence advancement only to the extent that it resolves uncertainty relevant to the next commitment.

Figure 1 illustrates the distinct justificatory roles of prediction-centred, mechanistic, and translational evidence, highlighting how unresolved uncertainty limits the extent to which favourable computational findings can support candidate advancement.

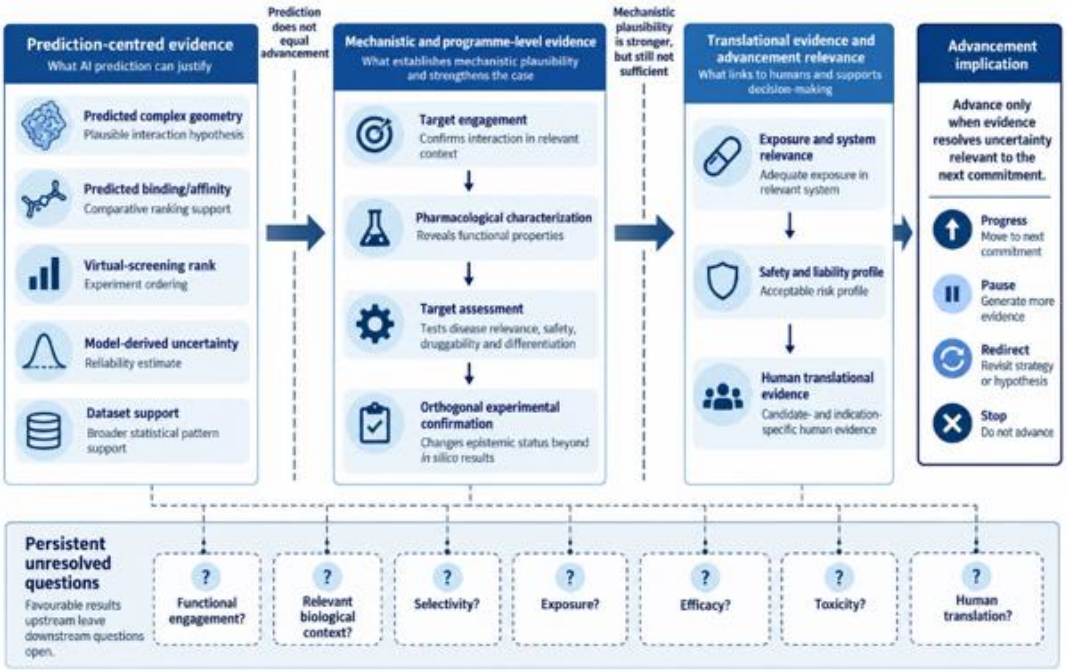


Figure 1. Evidence-layer boundaries in AI-supported candidate advancement.

Uncertainty changes that evidentiary requirement. Translational AI frameworks increasingly emphasize biological grounding, validation, uncertainty and deployment conditions rather than model accuracy in isolation [24]. Calibrated prediction intervals can also make uncertainty more explicit for molecular ADMET tasks [25]. Neither approach supplies a universal advancement threshold. Instead, uncertainty should alter the burden of evidence. A candidate near the validated domain of a well-calibrated model may justify focused confirmation, whereas an out-of-distribution candidate should require stronger orthogonal support because deployment-like shifts can impair both performance and calibration [11]. Candidate-level uncertainty methods can help identify where that burden is greatest [13]. The framework therefore treats uncertainty as a modifier of what evidence is needed next, not as an intrinsic penalty assigned to the molecule.

A practical consequence is that evidence sufficiency should be defined relative to the decision being contemplated. The evidence needed to synthesize another analogue is not necessarily the evidence needed to nominate a development candidate, initiate toxicology studies or enter a clinical programme. As commitment rises, unresolved uncertainties that were previously tolerable can become decision limiting. This makes advancement thresholds conditional on both scientific state and decision consequence. The framework therefore distinguishes evidence that changes confidence from evidence that changes action. A new assay result may increase confidence in binding without altering the current decision, whereas evidence of poor target engagement, incompatible exposure or a mechanism-linked safety liability may immediately change the appropriate state. This distinction prevents an ever-growing evidence dossier from being mistaken for a stronger advancement case simply because more positive observations have accumulated.

The decision itself is not binary. Staged discovery practice supports progressive derisking and explicit progression criteria [21]. Multidimensional industrial experience likewise shows the value of attending to target, tissue, safety and patient-related considerations rather than optimizing a single property [22]. Candidate-target-profile approaches further demonstrate how explicit criteria can organize what must be learned before nomination [23]. Building on these principles, progress is appropriate when the evidence configuration is sufficiently coherent for the next commitment; pause applies when a material uncertainty is unresolved but testable; redirect applies when the underlying opportunity remains credible but the mechanism, indication or optimization strategy requires revision; and stop applies when high-consequence counterevidence overwhelms the plausible value of additional confirmation. These states are contextual policies, not universal gates.

Figure 2 presents the proposed state-dependent decision framework, demonstrating how evidence sufficiency, unresolved uncertainty, and counterevidence jointly inform whether a candidate should progress, pause, redirect, or stop.

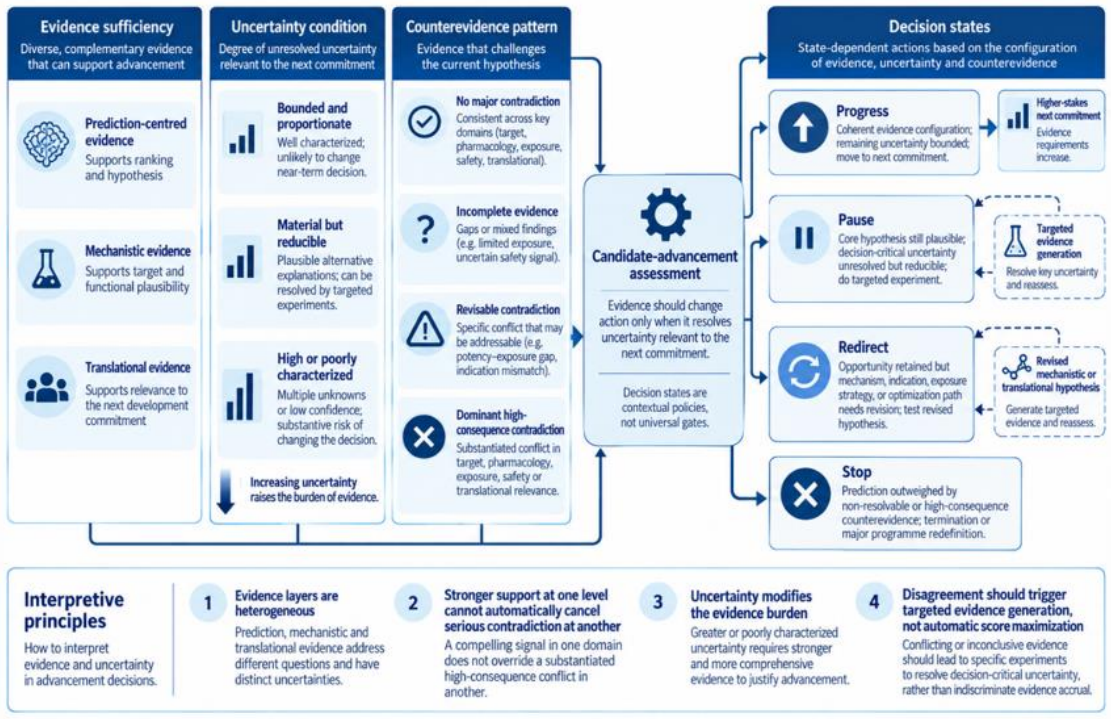


Figure 2. State-dependent candidate-advancement decisions under heterogeneous evidence and uncertainty.

The strongest test of the framework occurs when prediction and downstream evidence disagree. Mechanistically grounded AI can make hypotheses more testable [4], but an apparently strong prediction should not dominate when the candidate lies outside the model's reliable domain [11], when target assessment exposes a major biological or safety weakness [18], or when translational evidence contradicts the preclinical account. Human testing illustrates why the highest evidence layer remains candidate and indication specific rather than a general validation of AI-enabled discovery [20]. The framework is therefore non-compensatory only where counterevidence is decision-critical: a severe liability cannot be neutralized merely by adding more favorable predictions. In less decisive conflicts, disagreement should trigger targeted evidence generation rather than automatic rejection.

The same logic also prevents negative evidence from being treated uniformly. Some contradictions are diagnostic and redirectable: a compound may fail because the exposure profile is inadequate while the target hypothesis remains strong, or a phenotype may suggest that the original indication is poorly matched to the mechanism. Other contradictions undermine the core proposition, such as reproducible absence of target engagement under relevant conditions or a mechanism-linked liability that cannot be separated from efficacy. The framework therefore asks whether counterevidence attacks the candidate, the mechanism, the indication, or the broader target hypothesis. Redirect is justified only when the remaining evidence supports a coherent alternative path; otherwise, relabeling failure as redirection would merely postpone termination.

Implications and Research Priorities

The first implication is methodological: AI models intended to influence candidate advancement should be evaluated against the decision they are meant to improve. Predictive validity links a tool to downstream utility [3], but prospective studies should go further by asking whether model-assisted selection changes which candidates enter costly experiments, which uncertainties are retired, and which weak candidates are stopped earlier. Model discrimination and calibration remain necessary outcomes, yet they are incomplete when the claimed value of the system is better portfolio choice. Evaluation should therefore distinguish technical performance from incremental decision value.

Second, uncertainty quantification should be tested as an experimental-allocation tool. Deployment-like distribution shifts can expose serious reliability losses [11], while comparative UQ research shows that no single method should be presumed universally adequate [12]. Candidate-level OOD methods provide a promising route for identifying unreliable predictions [13], and conformal approaches can provide calibrated intervals in specific ADMET tasks [25]. The unresolved question is whether such information leads teams to choose better confirmatory experiments. Prospective comparisons could test whether uncertainty-guided evidence generation reduces avoidable advancement without simply increasing experimental burden.

Third, mechanistic evidence needs an incremental-value test of its own. Target engagement can strengthen an interaction hypothesis [15], pharmacological characterization can expose context-dependent properties [16], and structured target assessment can reveal programme-level weaknesses [18]. Their value should therefore be studied by asking when these evidence layers change a candidate decision that would otherwise have been made from prediction alone. This would distinguish mechanistic information that merely explains an existing ranking from evidence that genuinely redirects development.

A related research priority is to establish explicit records of why advancement decisions were made. Many AI studies report model metrics, and many discovery programmes document experimental milestones, but neither alone reconstructs the causal logic of the decision. Prospective decision logs could record the evidence available at each commitment point, the uncertainties regarded as material, the counterevidence considered decisive, the experiment selected to resolve uncertainty and whether subsequent evidence changed the state. Such records would enable retrospective calibration of advancement policies without pretending that heterogeneous evidence can be reduced to a universal score. They would also make it possible to compare whether AI-supported programmes fail because predictions were inaccurate, because relevant uncertainty was ignored, or because sound predictions were translated into poor decisions.

Because these states are decision constructs rather than natural categories, their reliability also depends on governance. Teams should specify who can override an AI-supported recommendation, what evidence can trigger reassessment, and whether stopping rules protect against sunk-cost continuation. These governance questions are especially important when model confidence and domain expertise disagree. The relevant research outcome is not conformity to either source of judgment, but whether the process makes the reason for advancement, delay or termination inspectable and scientifically contestable.

Finally, advancement thresholds should be investigated as context-sensitive policies rather than universal cutoffs. Clinical translation demonstrates that evidence requirements change as commitment and consequence increase [20]. Staged academic prioritisation [21], multidimensional industrial decision frameworks [22], explicit candidate profiles [23] and emerging AI implementation standards [24] all point toward conditional rather than fixed progression logic. The proposed framework would be weakened if a well-calibrated integrated model consistently produced equally defensible decisions without explicit evidence-layer distinctions. It would be strengthened if separating justificatory reach, uncertainty and counterevidence improved prospective candidate selection across programmes.

Conclusion

A favorable AI prediction is evidence about a candidate, not a decision to advance it. Defensible candidate advancement depends on whether the prediction remains credible when confronted with mechanism, target biology, pharmacology, translational context and uncertainty appropriate to the next development commitment. These evidence layers should not be treated as automatically commensurable because each resolves a different class of uncertainty and each has limits to what it can justify. The resulting decision is therefore state dependent: progress when the evidence is sufficiently coherent, pause when a material uncertainty is reducible, redirect when the opportunity remains credible but the path does not, and stop when consequential counterevidence or disproportionate unresolved uncertainty dominates. This account does not require a universal score or fixed threshold. Its central claim is conditional: AI can strengthen candidate selection when its predictions are embedded within an evidence process that makes uncertainty visible, demands mechanism-appropriate confirmation and allows downstream evidence to override upstream computational confidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Hasselgren C, Oprea TI. Artificial intelligence for drug discovery: Are we there yet? Annu Rev Pharmacol Toxicol. 2024;64:527-50. doi:10.1146/annurev-pharmtox-040323-040828
2. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
3. Scannell JW, Bosley J, Hickman JA, Dawson GR, Truebel H, Ferreira GS, et al. Predictive validity in drug discovery: What it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022;21(12):915-31. doi:10.1038/s41573-022-00552-x
4. Lavecchia A. Physics-inspired explainable AI for mechanistic and decision support in drug discovery. Drug Discov Today. 2026;31(3):104679. doi:10.1016/j.drudis.2026.104679
5. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2
6. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493-500. doi:10.1038/s41586-024-07487-w
7. Qiao Z, Nie W, Vahdat A, Miller TF 3rd, Anandkumar A. State-specific protein-ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell. 2024;6(2):195-208. doi:10.1038/s42256-024-00792-z

8. Cao D, Chen G, Jiang J, Yu J, Zhang R, Chen M, et al. Generic protein-ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. Nat Mach Intell. 2024;6(6):688-700. doi:10.1038/s42256-024-00849-z
9. Gorantla R, Gema AP, Yang IX, Serrano-Morrás Á, Suutari B, Juárez-Jiménez J, et al. Learning binding affinities via fine-tuning of protein and ligand language models. J Chem Inf Model. 2025;65(22):12279-91. doi:10.1021/acs.jcim.5c02063
10. Hsu WT, Grevtsev S, Herz AM, Douglas T, Magarkar A, Biggin PC. Can AI-predicted complexes teach machine learning to compute drug binding affinity? J Chem Inf Model. 2025;65(24):13051-6. doi:10.1021/acs.jcim.5c01848
11. 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
12. 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
13. Badkul A, Xie L, Zhang S, Xie L. Multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology. Nat Mach Intell. 2025;7(12):1985-95. doi:10.1038/s42256-025-01151-2
14. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
15. St John-Campbell S, Bhalay G. Target engagement assays in early drug discovery. J Med Chem. 2025;68(12):12331-68. doi:10.1021/acs.jmedchem.4c03115
16. Kenakin T. Know your molecule: Pharmacological characterization of drug candidates to enhance efficacy and reduce late-stage attrition. Nat Rev Drug Discov. 2024;23(8):626-44. doi:10.1038/s41573-024-00958-9
17. Pun FW, Podolskiy D, Izumchenko E, Mortlock A, Oprea TI, Scheibye-Knudsen M, et al. Target identification and assessment in the era of AI. Nat Rev Drug Discov. 2026;25(7):534-52. doi:10.1038/s41573-026-01412-8
18. Emmerich CH, Gamboa LM, Hofmann MCJ, Bonin-Andresen M, Arbach O, Schendel P, et al. Improving target assessment in biomedical research: The GOT-IT recommendations. Nat Rev Drug Discov. 2021;20(1):64-81. doi:10.1038/s41573-020-0087-3
19. Jiang Y, Zhang G, You J, Zhang H, Yao R, Xie H, et al. PocketFlow is a data-and-knowledge-driven structure-based molecular generative model. Nat Mach Intell. 2024;6(3):326-37. doi:10.1038/s42256-024-00808-8
20. Xu Z, Ren F, Wang P, Cao J, Tan C, Ma D, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. Nat Med. 2025;31(8):2602-10. doi:10.1038/s41591-025-03743-2
21. Murray AJ, Cox LR, Adcock HV, Roberts RA. Academic drug discovery: Challenges and opportunities. Drug Discov Today. 2024;29(4):103918. doi:10.1016/j.drudis.2024.103918
22. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018;17(3):167-81. doi:10.1038/nrd.2017.244
23. Bergström F, Lindmark B. Accelerated drug discovery by rapid candidate drug identification. Drug Discov Today. 2019;24(6):1237-41. doi:10.1016/j.drudis.2019.03.026
24. Gangwal A, Lavecchia A. IMPACT framework: Establishing global standards for artificial intelligence implementation, methodology, and translation in drug discovery. Wiley Interdiscip Rev Comput Mol Sci. 2026;16(2). doi:10.1002/wcms.70072
25. 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