Pharmacophore

ISSN-2229-5402

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

MECHANISTIC PLAUSIBILITY BEFORE PREDICTIVE PERFORMANCE: REORDERING THE STANDARDS BY WHICH COMPUTATIONAL PHARMACOLOGY MODELS ARE JUDGED

Wei Chen[1]*, Yuki Tanaka[2], Mei Lin[3]

1. ***Department of Mechanistic Pharmacology and Model Evaluation, College of Pharmaceutical Sciences, China Agricultural University, Beijing, China.***
2. ***Department of Computational Pharmacology and Biological Reasoning, Graduate School of Pharmacy, University of Tokyo, Tokyo, Japan.***
3. ***Department of Predictive Model Assessment, Faculty of Pharmacy, National University of Singapore, Singapore.***

ARTICLE INFO

Received:
03 June 2024
Received in revised form:
08 October 2024
Accepted:
12 October 2024
Available online:
28 October 2024

Keywords: Computational pharmacology, Mechanistic plausibility, Predictive adequacy, Quantitative systems pharmacology, Model credibility, Causal coherence

ABSTRACT

Computational pharmacology models are increasingly judged through predictive metrics that permit rapid comparison among algorithms but offer only partial evidence about pharmaceutical usefulness. A model may reproduce held-out observations while relying on unstable data associations, biologically implausible relationships, undocumented implementation choices, or causal interpretations that its design cannot support. This creates an unresolved methodological problem: predictive performance is often treated as the principal threshold of model quality even when the intended application requires explanation, mechanistic extrapolation, intervention reasoning, or consequential pharmaceutical decision support. This Original Methodological Position Article proposes a mechanism-first ordering of computational pharmacology evidence. The proposed position distinguishes structural coherence, biological coherence, and causal coherence from predictive adequacy and argues that these dimensions should be examined before performance results are granted substantive pharmaceutical meaning. Predictive evaluation remains necessary, but it is repositioned as a later test of an already scrutinized model rather than a substitute for plausibility, reproducibility, applicability, calibration, uncertainty characterization, or context-of-use specification. The article also introduces a bounded qualification logic in which the permissible influence of a model depends on the strength and relevance of its supporting evidence, the consequences of error, and the reversibility of the decision. The proposal is conceptual rather than empirically validated and does not establish a universal validation sequence, regulatory standard, clinical recommendation, or deployment criterion. Its purpose is to organize more defensible judgments about computational models and to clarify why benchmark success, mechanistic explanation, causal inference, experimental confirmation, and pharmaceutical utility must remain distinct evidentiary claims.

To Cite This Article: Chen W, Tanaka Y, Lin M. Mechanistic Plausibility before Predictive Performance: Reordering the Standards by Which Computational Pharmacology Models Are Judged. Pharmacophore. 2024;15(5):80-9. https://doi.org/10.51847/yTsRGCqU2z

Introduction

Computational models now occupy multiple positions within pharmaceutical research, from molecular-property prediction and target prioritization to systems-level simulation, translational reasoning, and model-informed development. Their increasing influence has intensified a basic evaluative question: what must a model demonstrate before its outputs deserve scientific or decision-making weight? In many contemporary workflows, the immediate answer is predictive performance. Models are compared through error measures, discrimination statistics, rankings, prospective hit rates, or other task-specific indicators. These quantities are often indispensable, but they address only whether an output corresponds sufficiently well to a selected reference under specified evaluation conditions. They do not independently establish that the model represents the relevant pharmacology, that its internal relationships are scientifically defensible, that its behavior will remain reliable under a new intervention or population, or that the modeled output is appropriate for the decision being considered.

Corresponding Author: Wei Chen; Department of Mechanistic Pharmacology and Model Evaluation, College of Pharmaceutical Sciences, China Agricultural University, Beijing, China. E-mail: wei.chen@cau.edu.cn.

Computational models now influence target identification, molecular prioritization, translational analysis, and model-informed development, yet their pharmaceutical value remains conditional on whether their outputs are credible for the decision in which they will be used [1–3]. This distinction is especially important as data-driven methods become integrated with pharmacometrics, bioinformatics, molecular modeling, and quantitative systems pharmacology. A model optimized for retrospective prediction may be useful for ranking compounds within a familiar chemical series, but the same evidence may be insufficient for inferring a biological mechanism, extrapolating beyond the observed domain, comparing interventions, or influencing a development strategy. The scientific meaning of a prediction therefore depends not only on its numerical accuracy but also on the model's construction, the evidence encoded within it, the conditions under which it was evaluated, and the claims attached to its outputs.

For mechanistic in silico models, that credibility must be assessed in relation to the proposed context of use, the consequences of model error, and the degree of influence assigned to the result [4]. Yet current evaluative language often allows conceptually different achievements to collapse into a single vocabulary of validation. Successful fitting may be described as biological validation; explainable outputs may be described as mechanistic insight; associations may be interpreted causally; and favorable benchmark results may be treated as evidence of prospective usefulness. These substitutions are not merely terminological. They alter which uncertainties remain visible, how confidently outputs are communicated, and how strongly a model may influence experimental, clinical, portfolio, or regulatory reasoning. A model can be accurate for the wrong structural reason, biologically plausible but poorly predictive, causally informative only under restrictive assumptions, or useful for exploration while remaining unsuitable for consequential decisions.

This article develops the methodological position that mechanistic plausibility should be evaluated before predictive performance is interpreted as pharmaceutical evidence. The proposed contribution is a reordered hierarchy rather than a new predictive algorithm, empirical benchmark, regulatory instrument, or universal checklist. It separates structural, biological, and causal coherence; explains why each supplies a distinct evidentiary function; and positions calibration, uncertainty, parsimony, reproducibility, applicability, and intended context of use as moderators of the meaning assigned to performance. The central proposition is not that mechanistic models are automatically superior to data-driven models or that every model must reproduce complete biological reality. It is that the strength of a model claim must remain bounded by the weakest scientifically necessary evidentiary layer for its proposed use. Predictive performance can strengthen a credible model, but it should not erase unresolved structural defects, implausible biology, unsupported causal assumptions, or uncertainty that materially changes the relevant decision.

How Predictive Performance Became the Dominant Evaluation Standard

The dominance of predictive performance arose from legitimate methodological needs. Computational pharmaceutical science required common ways to compare representations, algorithms, training procedures, and molecular or biological endpoints. Standardized datasets and quantitative metrics made comparisons more reproducible, allowed researchers to identify incremental improvements, and supported clearer communication across laboratories. Predictive benchmarks also fit the practical organization of machine-learning research: a model receives defined inputs, produces outputs, and is ranked by correspondence with held-out labels. This arrangement creates an apparently neutral basis for progress. It is easier to report a change in error or discrimination than to compare the quality of encoded biological assumptions, the relevance of a data-generating process, or the adequacy of a model for a particular pharmaceutical decision.

Benchmark suites, large-scale method comparisons, and audits of ligand-based classification collectively show how standardized predictive scores became a dominant language of evaluation while also exposing the possibility that apparent gains reflect benchmark design or memorization rather than transferable pharmaceutical knowledge [5–7]. Benchmarking therefore performs two roles that must not be conflated. First, it provides a controlled environment for comparing models under common assumptions. Second, it implicitly defines what the field treats as success. When benchmark success becomes the primary criterion of publication, model selection, or methodological prestige, characteristics that are easy to quantify can displace characteristics that are scientifically harder to establish. A model may consequently be rewarded for exploiting similarities within a dataset even when the intended pharmaceutical task requires extrapolation to unfamiliar chemistry, altered biology, new experimental platforms, or populations not represented in the training evidence.

The problem is not that conventional metrics are intrinsically misleading. The problem is that they are frequently interpreted beyond the conditions under which they were obtained. Random data partitions may test interpolation more than chemical-series transfer. Aggregate results may conceal performance differences across assay types, molecular classes, disease mechanisms, or clinically relevant subgroups. A highly flexible model may achieve favorable error measures while encoding relationships that are unstable under intervention or distribution shift. Conversely, a model with modest average predictive performance may still contribute scientifically by exposing assumptions, organizing mechanistic hypotheses, or identifying conditions under which current knowledge is insufficient. A performance-first culture tends to treat these models as inferior because it lacks an equally developed language for evaluating mechanistic and epistemic contribution.

These limitations are intensified when the underlying chemical or biological observations are treated as interchangeable data points despite differences in assay construction, provenance, relevance, bias, and measurement context [8]. Data availability should therefore be distinguished from data suitability. A large dataset may contain inconsistent endpoints, hidden experimental dependencies, selective reporting, imbalanced chemical space, indirect biological proxies, or measurements collected under conditions that differ from the intended use. Increasing model capacity or optimizing performance cannot

correct a mismatch between the evidence represented in the dataset and the pharmaceutical claim assigned to the output. The proposed methodological response is not to abandon benchmarking but to restrict its interpretation: predictive performance should establish predictive adequacy within an explicitly defined evaluation context, not mechanistic truth, causal validity, experimental confirmation, clinical utility, or general qualification.

Table 1 organizes the evidence, constructs, and interpretation boundaries needed to develop how predictive performance became the dominant evaluation standard within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for How predictive performance became the dominant evaluation standard

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Benchmark standardization	Does the evaluation permit controlled comparison among models?	Common datasets, endpoint definitions, partitions, and metrics	Explains why predictive performance became a practical common language	Treating the benchmark as a complete representation of pharmaceutical reality	Benchmark validity is limited to the task, data, split, and assumptions evaluated
Dataset suitability	Do the data represent the scientific question and intended application?	Provenance, assay context, endpoint relevance, bias, missingness, and domain coverage	Places evidence quality before interpretation of model scores	Equating a large or accessible dataset with a suitable evidentiary base	Data availability does not establish biological or decision relevance
Generalization	Does model behavior transfer beyond familiar observations or chemical structures?	External, temporal, scaffold-aware, platform-aware, or context-aware evaluation	Distinguishes interpolation from transferable predictive knowledge	Reporting memorization or similarity exploitation as broad generalization	Generalization claims must correspond to the actual form of shift tested
Predictive adequacy	Are outputs sufficiently accurate for the defined task?	Context-aligned error, discrimination, ranking, or calibration assessments	Retains predictive performance as necessary evidence	Treating one favorable metric as proof of overall model quality	Predictive adequacy is one evidentiary layer rather than a complete credibility judgment
Structural meaning	Does the model implement the relationships and constraints it claims to represent?	Architecture, equations, topology, code, units, parameterization, and numerical behavior	Introduces evidence that performance measures do not capture	Allowing fit to compensate for implementation or specification defects	Accurate outputs cannot independently verify model structure
Biological meaning	Are learned or encoded relationships compatible with relevant pharmacological knowledge?	Mechanistic evidence, directionality, time scales, perturbations, and multilevel consistency	Connects model outputs to scientifically bounded interpretation	Describing predictive associations as established biological mechanisms	Biological plausibility is graded and remains open to experimental challenge
Causal meaning	Can outputs support intervention or counterfactual claims?	Explicit causal assumptions, temporal ordering, perturbational evidence, and confounding analysis	Separates actionable claims from predictive association	Converting prediction or explanation into unsupported causal inference	Causal conclusions require evidence beyond predictive correspondence
Intended context of use	What decision, user, population, setting, and consequence define adequate evidence?	Development question, available alternatives, reversibility, and cost of error	Determines how much credibility is required	Treating a model as universally valid after success in one application	Credibility and qualification remain use-specific

Mechanistic Plausibility in Systems Pharmacology and Model-Informed Development

Mechanistic plausibility refers here to a bounded judgment that the relationships encoded or learned by a computational model are sufficiently coherent with its stated scientific interpretation and intended use. It is not equivalent to proving that a mechanism is uniquely true. Nor does it require every useful model to represent all biological processes explicitly. Instead, mechanistic plausibility asks whether the scientific content necessary for the proposed claim has been identified, represented, and challenged. For a phenomenological exposure model, the relevant mechanistic burden may be limited. For a systems model used to compare interventions, extrapolate across biological scales, or justify a counterfactual claim, the burden is greater. Plausibility is therefore conditional on what the model is expected to explain or influence.

QSP practice has consequently emphasized transparent model construction, context-dependent assessment, and early alignment with the pharmaceutical question rather than reliance on a single universal validation recipe [9–11]. These practices reflect the distinctive purpose of systems pharmacology. A QSP model commonly integrates evidence from several biological levels, formalizes hypothesized relationships, represents feedback or adaptation, and explores scenarios that cannot all be observed directly. Its value can lie partly in making assumptions explicit and testing whether disparate observations can coexist within one quantitative account. Predictive fit remains relevant, but it is not the only output of the modeling process. The model also establishes a structured argument about which entities matter, how they interact, which time scales are relevant, and why a simulated intervention produces a particular response.

The mechanism-first position proposed here organizes this argument through three distinguishable forms of coherence. Structural coherence concerns whether the equations, architecture, topology, parameters, units, numerical implementation, and

constraints correspond to the model's stated design. Biological coherence concerns whether modeled entities, directions, temporal patterns, cross-scale relationships, and emergent behaviors remain defensible against the relevant pharmacological evidence. Causal coherence becomes necessary when the model is used to support claims about interventions, counterfactuals, treatment effects, or controllable mechanisms. These forms of coherence are related but non-substitutable. A correctly coded model may represent implausible biology; a biologically plausible network may not identify a causal effect; and a predictive system may perform well without satisfying either requirement. **Figure 1** presents the mechanism-first model evaluation balance, showing how the article's key components and boundaries are connected within mechanistic plausibility in systems pharmacology and model-informed development.

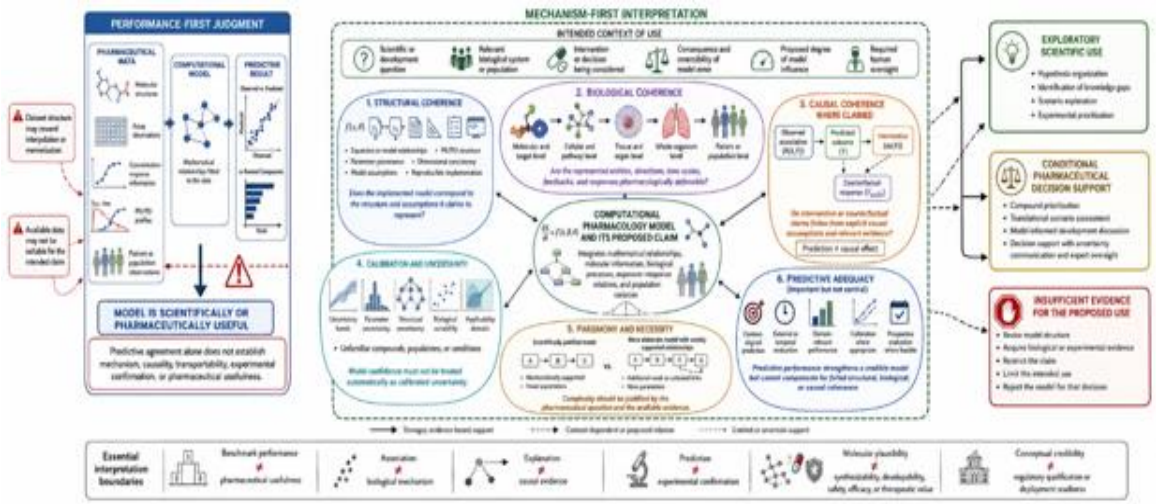

Figure 1. Mechanism-First Model Evaluation Balance.

The figure is an original conceptual synthesis that organizes the article's central contribution across predictive performance became the dominant evaluation standard, mechanistic plausibility in systems pharmacology and model-informed development, a reordered hierarchy of computational pharmacology evidence, testing structural, biological, and causal coherence before prediction, testing structural, causal coherence before prediction. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Show directional relations only where the manuscript supports them; use dashed connectors for uncertain, context-dependent, or proposed relations. Use a clean white background, precise alignment, professional sans-serif typography, thin uniform connectors, generous white space, and a restrained color-blind-safe palette. Do not make the figure resemble a generic flowchart, dashboard, pipeline of rectangular boxes, circular wheel, marketing infographic, or decorative molecular collage. Do not use journal, publisher, corporate, software, database, regulatory, or product logos; copied scientific figures; interface screenshots; stock laboratory photography; proprietary molecular renderings; or invented numerical outputs. Ensure all labels remain legible at journal-column width and the design is suitable for high-resolution export.

Mechanistic plausibility also clarifies the relationship between systems pharmacology and model-informed development. A model may help organize evidence, compare scenarios, expose knowledge gaps, or support the design of future experiments before it is qualified to influence a consequential development choice. The same principle applies when systems models are extended toward clinical development, where mechanistic extrapolation may inform a question but does not establish clinical utility independently of the relevant evidence and assumptions [12]. The proposed mechanism-first balance therefore does not award automatic priority to complex mechanistic models. Complexity can increase biological expressiveness, but it can also introduce weakly identified parameters, untestable assumptions, compensating errors, and opaque dependencies. Mechanistic detail is valuable only when it is justified by the intended question and supported strongly enough to constrain interpretation. The appropriate aim is not maximal mechanism but sufficient, inspectable, and context-relevant mechanism before predictive results are used to enlarge the model's scientific or pharmaceutical claim.

A Reordered Hierarchy of Computational Pharmacology Evidence

The proposed hierarchy begins by separating the order in which evidence is interpreted from the order in which a model may be technically developed. In practice, data preparation, mechanistic reasoning, parameter estimation, predictive testing, and uncertainty analysis frequently proceed iteratively. A mechanism-first standard does not prohibit such iteration. It instead specifies an order of adjudication: before favorable performance is used to justify a scientific or pharmaceutical claim, the model must satisfy the forms of coherence required by that claim. The hierarchy therefore moves from intended context of use to structural coherence, biological coherence, causal coherence where applicable, reproducibility, uncertainty and applicability, predictive adequacy, and finally a bounded qualification statement. Each level restricts the interpretation available at the next level. Failure at an earlier level cannot be neutralized merely by stronger performance at a later one.

The evidentiary intensity applied at each level should be proportionate rather than uniform. A model used to organize hypotheses may tolerate unresolved parameter uncertainty that would be unacceptable in a model intended to influence a dose-development strategy. Similarly, a screening model used within a well-represented chemical domain may require strong evidence of transportability and calibration but little claim to biological mechanism, whereas a systems model used to compare interventions must justify why its structure represents the relevant causal and pharmacological relationships. The evidentiary burden for a QSP model should therefore be proportional to its scope, uncertainty, assessment cost, and potential decision consequence [13]. This proportionality is essential because unlimited validation is impossible, yet minimal testing can be inadequate when model error has consequential or difficult-to-reverse effects.

A multidimensional evaluation is required because conceptual structure, implementation correctness, empirical correspondence, and intended use answer different questions about model credibility [14]. The first proposed level, intended context of use, defines the model question, relevant population or biological system, intervention, endpoint, user, and expected degree of influence. Structural coherence then asks whether the mathematical, computational, and data-processing implementation corresponds to the stated model. Biological coherence asks whether the encoded entities and relationships are compatible with the available pharmaceutical evidence. Causal coherence is evaluated only when the model supports intervention or counterfactual claims. Reproducibility and traceability determine whether these claims can be inspected independently. Calibration, uncertainty, and applicability constrain reliance on model outputs, while predictive adequacy tests whether the model performs sufficiently under evaluation conditions aligned with the intended use.

Qualification occupies the final level because it is a conclusion about evidence and use, not an intrinsic property acquired by a model once and retained permanently. Established PBPK practice demonstrates that qualification depends on the specified application, available evidence, assumptions, limitations, and reporting record [15]. A model qualified for one interaction pathway, population, compound class, or development question may require substantial reassessment before being used elsewhere. The required credibility should also increase with the model's proposed influence and the consequences of an erroneous conclusion [16]. The proposed hierarchy therefore ends in a qualification claim ceiling: the strongest statement that may defensibly be made for the defined use. Possible outcomes range from exploratory scientific use, through conditional decision support with explicit oversight, to a determination that the model is insufficiently credible for the proposed purpose.

Testing Structural, Biological, and Causal Coherence before Prediction

Structural coherence concerns the correspondence between the model as described and the model as implemented. Its assessment includes the correctness of equations or computational architecture, dimensional consistency, direction and connectivity of represented relationships, parameter provenance, initial and boundary conditions, numerical stability, data transformations, and version-controlled implementation. For learned models, structural review also includes the relationship between the target variable, representation, training objective, and claimed interpretation. A pre-existing mechanistic model transferred to a new research context must be reassessed for scope, assumptions, represented biology, parameter provenance, technical implementation, and relevant uncertainty [17]. Adaptation is not a clerical change of inputs: a new disease stage, intervention, population, endpoint, or time scale may alter which mechanisms and parameters are scientifically necessary.

Reproducibility forms part of structural coherence because a model that cannot be reconstructed or interrogated cannot be evaluated adequately. Executable code alone is insufficient when preprocessing steps, parameter transformations, solver configurations, stochastic procedures, calibration choices, or model versions remain undocumented. Likewise, a textual model description is insufficient if it omits implementation decisions that materially affect outputs. Reproducible implementation is necessary for structural scrutiny [18]. Nevertheless, reproduction of the same result does not establish validity. It confirms that another investigator can obtain and inspect the reported behavior. The reproduced model may still encode incorrect assumptions, fit an unsuitable target, or perform reliably only because the same hidden dependency has been preserved.

Biological coherence concerns whether the model's represented or learned relationships remain compatible with the evidence relevant to its intended claim. This assessment may involve the identity and direction of pathways, relative time scales, dose and concentration dependence, feedback behavior, compensatory processes, cross-level consistency, and responses to known perturbations. It should also include contradictory evidence and conditions under which the proposed mechanism is expected to fail. Interpretability methods may assist this evaluation by identifying features, substructures, examples, or input changes associated with model outputs. However, an attribution, saliency map, counterfactual example, or generated explanation may illuminate model behavior without demonstrating that the learned relationship corresponds to a pharmacological mechanism [19]. Explanation is evidence about model reasoning only to the extent that the explanatory method is itself appropriate and reliable; it is not experimental confirmation of the represented biology.

Causal coherence becomes necessary when model outputs are used to answer "what if" questions involving intervention, modification, or counterfactual comparison. A model that predicts an outcome from correlated observations does not necessarily estimate how that outcome would change if a target, treatment, pathway, exposure, or patient characteristic were altered. Claims about intervention, treatment response, or counterfactual outcomes require causal assumptions that are not supplied by predictive association alone [20]. Relevant scrutiny may include temporal ordering, confounding structures, intervention consistency, selection mechanisms, transport assumptions, and comparison with perturbational evidence. The mechanism-first position does not require a fully identified causal model for every pharmaceutical prediction. It requires causal coherence only when causal meaning is attached to the result. When the necessary assumptions or evidence are unavailable,

the appropriate response is to restrict the claim to association or prediction rather than describing the model as mechanistically actionable.

Balancing Calibration, Parsimony, Uncertainty, and Intended Context of Use

After the required coherence conditions have been addressed, predictive performance must be interpreted together with calibration and uncertainty. Accuracy describes correspondence between outputs and selected observations, whereas calibration concerns whether expressed probabilities or intervals correspond appropriately to observed frequencies or errors. Uncertainty extends further by addressing imperfect knowledge of parameters, structures, measurements, biological variability, data-generating conditions, and applicability. Once mechanistic plausibility has been examined, prediction must still be interpreted together with the uncertainties that determine whether a result can be relied upon in drug design [21]. A model may be accurate on average but unreliable for precisely those compounds, pathways, populations, or intervention conditions that matter most to a decision.

Uncertainty methods should be judged according to the uncertainty source and intended use rather than treated as interchangeable confidence generators. Ensembles, Bayesian approximations, evidential methods, conformal approaches, prediction intervals, sensitivity analyses, and scenario analyses answer different questions and rely on different assumptions. Available uncertainty estimators capture different properties and may behave differently when molecular inputs depart from the training domain [22]. An estimate that correlates with interpolation error may fail to identify biological extrapolation, assay shift, population shift, or structural misspecification. For this reason, uncertainty evaluation should include not only average calibration but also performance under relevant forms of unfamiliarity, subgroup variation, alternative model structures, and assumptions that cannot be resolved from the available data.

No single uncertainty representation should be presumed adequate across computational pharmacology applications. Comparative evidence indicates that no examined neural uncertainty method is uniformly reliable across datasets, tasks, and evaluation criteria [23]. The proposed hierarchy therefore treats uncertainty adequacy as a context-dependent judgment. For exploratory analysis, qualitative sensitivity and identification of unsupported regions may be sufficient. For prospective experimental prioritization, the relationship between uncertainty and error may need to be demonstrated within the relevant chemical or biological domain. For consequential development decisions, unresolved structural and translational uncertainties may need to be represented through alternative scenarios rather than compressed into one numerical interval. Model confidence should not be described as calibrated uncertainty unless the relevant calibration claim has been evaluated.

Parsimony moderates mechanistic plausibility by requiring model complexity to be justified by the scientific question and available evidence. A model that omits a necessary mechanism may be too simple, but a model containing weakly supported pathways, parameters, or learned components may create an appearance of mechanistic richness without greater credibility. Parsimony therefore means justified complexity, not automatic selection of the smallest model. Component necessity can be examined through reduced alternatives, sensitivity analysis, identifiability assessment, robustness testing, and comparison of whether competing structures change the relevant decision. The relevant performance standard must ultimately reflect the losses and consequences of the pharmaceutical decision rather than a generic prediction error chosen for convenience [24]. Intended context of use connects these considerations: acceptable complexity, uncertainty, calibration, and predictive adequacy depend on what the model will influence, who will use it, what alternatives exist, and how reversible an erroneous decision would be.

Consequences for Model Qualification and Pharmaceutical Decision-Making

A mechanism-first standard changes model qualification from a declaration of general validity into a bounded argument about permissible use. Qualification should state what question the model can inform, the evidence supporting that use, the assumptions that remain necessary, the conditions under which performance was evaluated, and the circumstances that require reassessment. Emerging efforts toward harmonized MIDD principles similarly frame model evidence through risk, credibility, and context-specific regulatory use [25]. Harmonization need not imply one universal evidentiary threshold. Its more defensible purpose is to establish common expectations for transparency, reasoning, uncertainty communication, and the connection between model influence and demonstrated credibility.

The proposed hierarchy also affects how machine learning is integrated with pharmacometrics and systems pharmacology. A learned component may improve parameter estimation, identify patterns, construct surrogate models, approximate computationally intensive processes, or support molecular and clinical prediction. These functions do not remove the need to specify what the learned component contributes and which scientific claim depends on it. For machine learning within pharmacometrics and systems pharmacology, best practice therefore requires explicit use cases, interpretable limitations, and integration with established quantitative reasoning [26]. Hybrid models should be examined for failure modes at their interfaces: a mechanistic shell does not automatically make a learned component biologically valid, and a highly predictive learned component does not establish that the surrounding mechanistic interpretation is correct.

For pharmaceutical decision-making, the proposed order encourages earlier separation of model development from model authority. A model may be useful before it is qualified for high-influence use because it can reveal inconsistencies, organize knowledge, prioritize measurements, compare scenarios, or expose uncertainty. Conversely, a technically mature model may remain unsuitable for a proposed decision when the supporting evidence does not match the relevant population, mechanism, endpoint, or consequence. Early experience with model-informed regulatory discussions illustrates that the central issue is

how a particular model may inform a particular development problem, not whether a modeling approach is acceptable in the abstract [27]. A mechanism-first dossier would therefore begin with the question and proposed influence, then document coherence, reproducibility, uncertainty, predictive adequacy, alternative explanations, and the intended qualification boundary. The practical consequence is not automatic rejection of models with incomplete mechanisms. It is explicit limitation of their claims and influence. A model may support exploratory compound prioritization while remaining unsuitable for causal target claims. A QSP model may organize translational hypotheses while requiring further evidence before it informs regimen selection. A PBPK model may support a qualified extrapolation within one application but not another population or pathway. Industry experience suggests that early model-focused dialogue may improve methodological clarity and strategic learning, although participation or favorable discussion cannot be interpreted as universal qualification or acceptance [28]. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop consequences for model qualification and pharmaceutical decision-making within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Mechanistic Plausibility before Predictive Performance Reordering the Standards by Which Computational Pharmacology

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Intended context of use	Scientific question, model users, population or system, intervention, endpoint, decision alternatives, consequence of error	Defines what the model must support and how strongly it may influence action	Explicit use statement and proposed claim ceiling	The context may change during development or differ among stakeholders	Prospective confirmation that evaluation conditions match the actual use
Structural coherence	Equations, architecture, topology, parameters, code, units, preprocessing, solver and version records	Tests whether the implemented model corresponds to its stated construction	Traceable and technically inspectable model specification	Correct implementation can still represent incorrect science	Independent verification, reproduction, and challenge testing
Biological coherence	Pharmacology, disease biology, pathway evidence, time scales, known perturbations, translational observations	Tests compatibility between modeled relationships and relevant biological evidence	Bounded mechanistic interpretation and identified contradictions	Available biological knowledge may be incomplete or context dependent	Predefined behavior tests, perturbational comparison, and negative controls
Causal coherence where claimed	Temporal structure, causal assumptions, intervention evidence, confounder assessment, counterfactual definitions	Determines whether intervention or counterfactual conclusions are defensible	Restricted causal claim or explicit reclassification as association or prediction	Causal effects may be unidentifiable from available evidence	Perturbational studies, causal sensitivity analyses, and assumption testing
Reproducibility and traceability	Executable model, data transformations, parameter records, software environment, documentation	Enables independent reconstruction and interrogation	Reproduced implementation and auditable evidence chain	Reproduction does not establish biological or predictive validity	Cross-team or independent reproduction
Parsimony and component necessity	Alternative structures, reduced models, sensitivity, identifiability, parameter support	Determines whether complexity is justified by the scientific question	Defensible level of mechanistic or computational complexity	Simplification may omit important mechanisms; complexity may conceal weak support	Comparative model analysis and component-necessity testing
Calibration, uncertainty, and applicability	Prediction distributions, error behavior, domain information, parameter and structural uncertainty, shift scenarios	Limits reliance on outputs and identifies unsupported conditions	Calibrated or otherwise qualified uncertainty statement and applicability boundary	No single method captures every relevant uncertainty source	Evaluation under context-relevant domain, platform, temporal, biological, or population shifts
Predictive adequacy	Context-aligned observations, prespecified outcomes, appropriate evaluation partitions	Tests whether an already scrutinized model performs sufficiently for its defined task	Predictive result bounded to the evaluated conditions	Favorable aggregate metrics may conceal subgroup or domain failure	External, temporal, mechanistically informative, or prospective evaluation as appropriate
Human oversight and decision integration	Model evidence, alternatives, uncertainties, decision costs, reversibility, domain expertise	Connects model outputs to accountable pharmaceutical reasoning	Documented decision role and permissible model influence	Human review can introduce inconsistency or automation bias	Prospective evaluation of decision processes and disagreement resolution
Qualification and reassessment	Integrated evidence record, unresolved limitations, change history, proposed degree of influence	Assigns the strongest claim supported for a specific use	Exploratory use, conditional decision support, or insufficient credibility	Qualification does not automatically transfer across uses or model versions	Reassessment after material changes in context, structure, data, evidence, or influence

Limitations and Boundary Conditions

The proposed hierarchy is a methodological position rather than an empirically validated demonstration that mechanism-first assessment produces better pharmaceutical outcomes than performance-first evaluation. Its ordering is supported by distinctions among credibility, context of use, reproducibility, uncertainty, prediction, and causality, but the practical benefits, resource requirements, and inter-reviewer consistency of the hierarchy remain to be tested. Different model classes will require different evidence. A molecular ranking model, population pharmacokinetic model, PBPK model, QSP model, causal model,

and hybrid neural–mechanistic system cannot be assessed through identical technical procedures. The proposal should therefore be understood as an order of questions and claims rather than a fixed checklist or universal scoring instrument.

Mechanistic plausibility is itself limited by the incompleteness and revisability of biological knowledge. A model can be consistent with current evidence while representing only one of several plausible mechanisms. Conversely, a model may appear inconsistent because the evidence used to assess it is incomplete, indirect, or collected under a different biological context. Structural and biological coherence should therefore be graded rather than treated as binary certification. Parsimony also presents an unresolved tension: simplifying a model may improve identifiability and inspectability while removing mechanisms necessary for extrapolation. Increasing detail may improve biological expressiveness while introducing uncertain parameters and weakly testable assumptions. The proposed hierarchy does not supply a universal solution to this trade-off.

The hierarchy cannot establish clinical benefit, regulatory acceptance, experimental confirmation, synthesizability, developability, safety, efficacy, therapeutic value, or deployment readiness. It also cannot convert observational associations into causal effects or guarantee transportability to new compounds, populations, laboratories, or development stages. Risks of misuse include describing the hierarchy as a certification pathway, treating passage through a conceptual gate as proof of model truth, or assigning numerical readiness scores unsupported by evidence. Its proper use is narrower: to make the reasoning behind model claims visible, identify where evidence is insufficient, and prevent predictive performance from being used to obscure unresolved scientific or decision-relevant limitations.

Research Agenda and Implementation Priorities

The first research priority is operational definition. Structural, biological, and causal coherence require model-class-specific criteria that can be assessed consistently without reducing complex scientific judgment to superficial checklists. Comparative studies should examine how experts evaluate the same model–context pair, where their judgments differ, and whether structured evidence templates improve agreement. Research should also investigate adversarial test cases in which models achieve similar predictive performance for different structural reasons. Such studies could determine whether mechanism-first assessment identifies hidden implementation errors, compensating parameter configurations, implausible pathways, or unsupported causal interpretations that conventional benchmarking misses.

A second priority is prospective evaluation under scientifically relevant shifts. Uncertainty and calibration methods should be tested across chemical-series changes, assay-platform changes, temporal changes, new biological settings, population differences, and altered intervention conditions. Evaluation should ask not only whether uncertainty correlates with prediction error but also whether it identifies circumstances in which mechanistic assumptions cease to be defensible. Hybrid mechanistic and machine-learning models require particular attention because uncertainty may arise from data, learned representations, parameter estimation, model structure, biological abstraction, and interactions among these layers. Shared reporting infrastructures should preserve model versions, assumptions, parameter provenance, transformations, uncertainty analyses, evaluation contexts, and reasons for subsequent claim restriction.

Implementation should proceed in stages. Early-stage adoption could require explicit context-of-use statements and separate reporting of predictive, mechanistic, causal, and uncertainty claims. A second stage could introduce independent structural review, predefined biological challenge tests, and documented applicability boundaries for models expected to influence prospective experimentation or development strategy. Higher-influence use would require stronger evidence, cross-functional review, explicit human accountability, and predefined triggers for reassessment. Prospective decision-impact studies should then test whether this process changes experimental priorities, prevents inappropriate model reuse, improves uncertainty communication, or alters the quality and reversibility of pharmaceutical decisions. Until such evidence is available, implementation should remain cautious, transparent, and proportional to consequence.

Conclusion

Computational pharmacology models should not be judged as though predictive performance were a complete account of scientific credibility. The proposed mechanism-first position reorders evaluation by requiring the intended context of use, structural coherence, biological coherence, causal coherence where claimed, reproducibility, parsimony, applicability, calibration, and uncertainty to constrain the meaning assigned to prediction. Predictive adequacy remains essential, but it becomes a later evidentiary test rather than a mechanism for overriding unresolved deficiencies. The final product of evaluation is therefore not universal validation but a bounded statement about what the model may reasonably support and how strongly it may influence a pharmaceutical decision. This contribution is conceptual and requires empirical assessment across model classes, organizations, and decision settings. Its principal value lies in making evidentiary distinctions explicit: association is not causality, explanation is not mechanism, prediction is not experimental confirmation, confidence is not calibrated uncertainty, and benchmark success is not pharmaceutical usefulness.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. 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. Marshall S, Madabushi R, Manolis E, Krudys K, Staab A, Dykstra K, et al. Model-informed drug discovery and development: Current industry good practice and regulatory expectations and future perspectives. CPT Pharmacometrics Syst Pharmacol. 2019;8(2):87-96. doi:10.1002/psp4.12372
4. Musuamba FT, Skottheim Rusten I, Lesage R, Russo G, Bursi R, Emili L, et al. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT Pharmacometrics Syst Pharmacol. 2021;10(8):804-25. doi:10.1002/psp4.12669
5. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
6. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
7. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
8. 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
9. Cucurull-Sanchez L, Chappell MJ, Chelliah V, Cheung SYA, Derks G, Penney M, et al. Best practices to maximize the use and reuse of quantitative and systems pharmacology models: Recommendations from the United Kingdom Quantitative and Systems Pharmacology Network. CPT Pharmacometrics Syst Pharmacol. 2019;8(5):259-72. doi:10.1002/psp4.12381
10. Ramanujan S, Chan JR, Friedrich CM, Thalhauser CJ. A flexible approach for context-dependent assessment of quantitative systems pharmacology models. CPT Pharmacometrics Syst Pharmacol. 2019;8(6):340-3. doi:10.1002/psp4.12409
11. Nijsen MJMA, Wu F, Bansal L, Bradshaw-Pierce E, Chan JR, Liederer BM, et al. Preclinical QSP modeling in the pharmaceutical industry: An IQ Consortium survey examining the current landscape. CPT Pharmacometrics Syst Pharmacol. 2018;7(3):135-46. doi:10.1002/psp4.12282
12. Bai JPF, Earp JC, Strauss DG, Zhu H. A perspective on quantitative systems pharmacology applications to clinical drug development. CPT Pharmacometrics Syst Pharmacol. 2020;9(12):675-7. doi:10.1002/psp4.12567
13. Lemaire V, Hu C, van der Graaf PH, Chang S, Wang W. No recipe for quantitative systems pharmacology model validation, but a balancing act between risk and cost. Clin Pharmacol Ther. 2024;115(1):25-8. doi:10.1002/cpt.3082
14. Braakman S, Pathmanathan P, Moore H. Evaluation framework for systems models. CPT Pharmacometrics Syst Pharmacol. 2022;11(3):264-89. doi:10.1002/psp4.12755
15. Shebley M, Sandhu P, Emami Riedmaier A, Jamei M, Narayanan R, Patel A, et al. Physiologically based pharmacokinetic model qualification and reporting procedures for regulatory submissions: A consortium perspective. Clin Pharmacol Ther. 2018;104(1):88-110. doi:10.1002/cpt.1013
16. Kuemmel C, Yang Y, Zhang X, Florian J, Zhu H, Tegenge M, et al. Consideration of a credibility assessment framework in model-informed drug development: Potential application to physiologically-based pharmacokinetic modeling and simulation. CPT Pharmacometrics Syst Pharmacol. 2020;9(1):21-8. doi:10.1002/psp4.12479
17. Weis M, Baillie R, Friedrich C. Considerations for adapting pre-existing mechanistic quantitative systems pharmacology models for new research contexts. Front Pharmacol. 2019;10:416. doi:10.3389/fphar.2019.00416
18. Kirouac DC, Cicali B, Schmidt S. Reproducibility of quantitative systems pharmacology models: Current challenges and future opportunities. CPT Pharmacometrics Syst Pharmacol. 2019;8(4):205-10. doi:10.1002/psp4.12390
19. 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
20. 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
21. Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today. 2021;26(2):474-89. doi:10.1016/j.drudis.2020.11.027
22. 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
23. 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

24. Watson OP, Cortés-Ciriano I, Taylor AR, Watson JA. A decision-theoretic approach to the evaluation of machine learning algorithms in computational drug discovery. Bioinformatics. 2019;35(22):4656-63. doi:10.1093/bioinformatics/btz293
25. Marshall S, Ahamadi M, Chien J, Iwata D, Farkas P, Filipe A, et al. Model-informed drug development: Steps toward harmonized guidance. Clin Pharmacol Ther. 2023;114(5):954-9. doi:10.1002/cpt.3006
26. Terranova N, Renard D, Shahin MH, Menon S, Cao Y, Hop CECA, et al. Artificial intelligence for quantitative modeling in drug discovery and development: An Innovation and Quality Consortium perspective on use cases and best practices. Clin Pharmacol Ther. 2024;115(4):658-72. doi:10.1002/cpt.3053
27. Madabushi R, Benjamin JM, Grewal R, Pacanowski MA, Strauss DG, Wang Y, et al. The US Food and Drug Administration's model-informed drug development paired meeting pilot program: Early experience and impact. Clin Pharmacol Ther. 2019;106(1):74-8. doi:10.1002/cpt.1457
28. Galluppi GR, Brar S, Caro L, Chen Y, Frey N, Grimm HP, et al. Industrial perspective on the benefits realized from the FDA's model-informed drug development paired meeting pilot program. Clin Pharmacol Ther. 2021;110(5):1172-5. doi:10.1002/cpt.2265