**Pharmacophore**

ISSN-2229-5402

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

# WORLD MODELS FOR PHARMACEUTICAL SCIENCE SHOULD SIMULATE BIOLOGICAL CONSEQUENCES RATHER THAN MERELY CONTINUE STATISTICAL PATTERNS

**Min Zhang[1]*, Daiki Takahashi[2], Younes Benali[3], Xiaoli Liu[4]**

1. ***Department of World Models and Biological Consequence Simulation, Faculty of Pharmacy, Jilin Agricultural University, Changchun, China.***
2. ***Department of Beyond-Statistical-Pattern Modeling, Graduate School of Pharmacy, Hokkaido University, Sapporo, Japan.***
3. ***Department of Pharmaceutical World Models and Reasoning, Faculty of Sciences, Mohammed V University, Rabat, Morocco.***
4. ***Department of Biological Consequence Prediction, Faculty of Pharmacy, Nanjing Agricultural University, Nanjing, China.***

**ARTICLE INFO**

**Received:**
23 May 2025
**Received** in revised form:
01 October 2025
**Accepted:**
05 October 2025
**Available online:**
28 October 2025

***Keywords:*** Pharmaceutical world models, Consequence-aware simulation, Biological state, Perturbation prediction, Causal representation, Counterfactual reasoning

**ABSTRACT**

Artificial intelligence in pharmaceutical science increasingly generates molecules, predicts reactions, integrates biological measurements, and coordinates scientific tools. These capabilities are valuable, but they do not by themselves constitute a world model capable of representing what happens to a biological system after a pharmaceutical intervention. The unresolved problem is that models optimized to continue statistically probable sequences may produce chemically or biologically persuasive outputs without encoding intervention-specific state transitions, temporal adaptation, causal structure, uncertainty, or multiscale consequences. This article proposes a consequence-aware pharmaceutical world-model architecture that treats biological simulation as a structured relationship among an initial biological state, an explicitly specified intervention, context-dependent transition dynamics, time, observations, uncertainty, and counterfactual alternatives. The proposed construct distinguishes latent biological state from assay output, molecular action from therapeutic consequence, predictive association from mechanism, and model confidence from calibrated uncertainty. It further links learned cellular representations with mechanistic and quantitative pharmacological constraints, multiscale coupling, provenance, applicability boundaries, and experimental feedback. Validation is framed as a graduated evidentiary problem requiring more than reconstruction or aggregate predictive performance: models should be challenged through held-out perturbations, dose and temporal variation, context transfer, interaction recovery, mechanism-directed experiments, uncertainty assessment, and prospective falsification. The architecture is conceptual rather than empirically validated and cannot establish clinical utility, regulatory acceptability, or universal biological completeness. Its principal implication is that pharmaceutical world models should be judged by whether they support bounded, testable, intervention-conditioned claims about biological consequences—not by whether they generate fluent, realistic, or statistically familiar continuations.



**To Cite This Article:** Zhang M, Takahashi D, Benali Y, Liu X. World Models for Pharmaceutical Science Should Simulate Biological Consequences rather than Merely Continue Statistical Patterns. Pharmacophore. 2025;16(5):55-66. https://doi.org/10.51847/EFVx6KQqNA

## Introduction

Computational pharmaceutical science now includes machine learning for molecular property prediction, target identification, reaction prediction, molecular design, biomarker analysis, pharmacological modeling, and development-stage decision support. These methods can accelerate individual tasks and connect previously separated data types, yet their success remains distributed across heterogeneous objectives, representations, and validation practices rather than organized around a unified account of how an intervention changes a biological system [1-3]. A molecular generator may propose a structure, a reaction model may suggest a transformation, and a cellular model may predict an expression profile, but none of these outputs alone

**Corresponding Author:** Min Zhang; Department of World Models and Biological Consequence Simulation, Faculty of Pharmacy, Jilin Agricultural University, Changchun, China. E-mail: min.zhang@jlau.edu.cn.

specifies the full chain from molecular action to exposure, pathway modulation, cellular adaptation, tissue response, organismal consequence, and clinically relevant outcome.

The conceptual difficulty is not resolved by adding explanations to an otherwise predictive model. Feature attribution, attention visualization, counterfactual examples, and interpretable substructures can help scientists inspect model behavior, but they do not independently demonstrate that the model has recovered a causal biological mechanism [4]. A persuasive explanation may describe correlations encoded by the training data, identify features associated with a prediction, or generate a plausible mechanistic narrative while remaining disconnected from the computations that produced the output. Pharmaceutical consequence modeling therefore requires a stricter distinction among interpretability, mechanistic representation, causal identification, experimental confirmation, and decision usefulness.

A second difficulty arises from the gap between benchmark performance and consequential pharmaceutical validity. Retrospective accuracy may be inflated by dataset similarity, narrow task definitions, favorable data splitting, incomplete comparators, or endpoints that are easier to predict than the decisions they are used to justify. Consequently, strong performance on a molecular or biological benchmark does not establish that a model can guide synthesis, reproduce an intervention response, generalize to a new disease context, anticipate toxicity, or support a development decision [5]. The central problem is therefore architectural and evidentiary: the field lacks a sufficiently explicit definition of what must be represented, connected, and tested before a model can claim to simulate pharmaceutical consequences.

This article proposes a Pharmaceutical Consequence-Simulation World Model as an original conceptual architecture. The construct is defined as a context-bounded model that represents a time-indexed biological state, applies an explicitly specified pharmaceutical intervention, generates a distribution over possible subsequent states, exposes uncertainty and causal assumptions, and produces testable observations across relevant biological scales. Its contribution is not a new trained model, dataset, benchmark, or validated decision system. Rather, it specifies the representational commitments, internal interfaces, evaluation logic, and claim boundaries needed to distinguish consequence-aware simulation from statistical continuation. The architecture is intended to organize future model design and falsification while preventing fluent output, mechanistic vocabulary, or a single performance measure from being treated as sufficient evidence of biological simulation.

*From Next-Token Prediction to Consequence-Aware Scientific Simulation*

Sequence models demonstrate that statistical continuation can acquire substantial chemical competence. Recurrent neural networks can learn molecular-string distributions and generate focused molecular libraries; transformer models can infer reaction products from learned relationships between reactants and products; and large chemical-language representations can encode structural and property information through pretraining [6-8]. These results show that next-token-like objectives are not scientifically trivial. By learning regularities in chemical notation and training distributions, such systems may support molecular generation, reaction prediction, representation learning, retrieval, and hypothesis formation. The limitation is narrower: success at continuing a chemical sequence does not establish that the model represents the biological system in which the resulting molecule would act.

Tool augmentation can extend a language model beyond its internal statistical representation by allowing it to retrieve chemical information, execute calculations, invoke specialist models, or check molecular properties [9]. Such grounding can improve accuracy and scientific usefulness, particularly where the underlying task has a well-defined external tool. Nevertheless, a collection of tool calls remains different from a world model. The tool-augmented system may calculate descriptors, search literature, predict a target, or estimate a property without maintaining a coherent biological state that changes through time. It may also combine outputs generated under incompatible assumptions, biological contexts, measurement conditions, or uncertainty conventions. Consequence-aware simulation therefore requires more than access to scientific functions; it requires a common representational substrate through which the outputs of those functions can be interpreted as parts of the same evolving system.

Agentic systems further illustrate the distinction between acting in a scientific environment and modeling the environment's consequences. Language-model agents can formulate plans, select tools, execute code, consult external information, and coordinate elements of chemical experimentation [10]. These capabilities may support closed-loop research, but agency and world modeling are separable. An agent may possess an inaccurate model of the system it manipulates, rely on external tools whose outputs are not mutually consistent, or respond to experimental feedback without representing why the result occurred. Conversely, a consequence model might support simulation without autonomously selecting or executing experiments. The proposed architecture therefore treats agents as potential users of a pharmaceutical world model rather than assuming that planning autonomy itself demonstrates biological understanding.

The transition from pattern continuation to consequence-aware simulation requires at least six changes in scientific commitment. First, the model must represent an identifiable biological state rather than only a token history or static molecular input. Second, the intervention must be explicit, including its identity, dose, schedule, route, duration, combination, and context. Third, predicted outputs must arise through temporally ordered state transitions rather than unstructured completion. Fourth, uncertainty must accompany the predicted trajectory and distinguish data noise, model uncertainty, measurement limitations, and unsupported extrapolation. Fifth, counterfactual alternatives must be tied to defensible causal assumptions rather than generated as unconstrained possibilities. Sixth, outputs must be connected to observations that can falsify the simulated consequence. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop from next-token prediction to consequence-aware scientific simulation within the article's central argument.

**Table 1.** Evidence domains, core questions, scientific requirements, and interpretation boundaries for From next-token prediction to consequence-aware scientific simulation

| Construct or evidence domain | Core question | Scientific basis | Role in the article | Failure or overclaiming risk | Boundary statement |
|---|---|---|---|---|---|
| **Statistical sequence continuation** | Can the model reproduce or extend patterns present in chemical or scientific sequences? | Chemical sequence and language models can learn syntax, reaction regularities, molecular structure, and property associations. | Establishes a valuable but insufficient starting point for pharmaceutical world modeling. | Fluency, validity, or benchmark success may be interpreted as evidence that the model understands biological consequences. | Pattern continuation supports representation or prediction claims only within the evaluated task and domain. |
| **Molecular representation** | Does the model encode chemically relevant structure and relationships? | Learned molecular representations can support generation, retrieval, reaction prediction, and property estimation. | Supplies molecular inputs to the proposed architecture. | Molecular embeddings may omit stereochemistry, formulation, exposure, target engagement, context, and biological adaptation. | A chemically informative representation is not a complete biological state. |
| **Tool grounding** | Can external scientific tools supply calculations, retrieval, or specialist predictions? | Tool-augmented systems can combine language-model reasoning with external chemistry functions. | Provides modular capabilities that may inform the world model. | Outputs from different tools may have incompatible assumptions, domains, scales, or uncertainty meanings. | Tool use does not establish an internally coherent consequence model. |
| **Agentic planning** | Can a system select and sequence scientific actions? | Autonomous systems can plan, retrieve information, execute code, and coordinate experimental tasks. | Positions agents as users or controllers of a world model. | Action competence may be mistaken for accurate representation of the manipulated biological system. | Agency and world modeling must be evaluated separately. |
| **Biological state** | What condition of the biological system exists before and after intervention? | Cellular and multiscale models increasingly encode molecular, network, phenotypic, and contextual information. | Defines the object that undergoes change. | A latent vector may be treated as sufficient without showing that it preserves intervention-relevant information. | State adequacy is conditional on the intended consequence and available observations. |
| **Intervention specification** | What is changed, under which dose, schedule, route, duration, and context? | Pharmaceutical effects depend on exposure, target engagement, timing, combinations, and biological context. | Converts passive prediction into intervention-conditioned simulation. | A compound identifier may be treated as a complete intervention description. | An intervention must be represented at the resolution required by the intended claim. |
| **Temporal transition** | How does the state evolve after intervention? | Biological responses include delays, adaptation, feedback, recovery, progression, and repeated exposure. | Provides the core world-model transition function. | Cross-sectional similarity may be misrepresented as temporal dynamics. | Time-ordered prediction is necessary but does not by itself establish causality. |
| **Multiscale consequence** | How are molecular actions linked to cellular, tissue, physiological, and outcome states? | Pharmaceutical consequences emerge through interactions across biological and pharmacological scales. | Prevents direct equivalence between molecular plausibility and therapeutic value. | Missing scale interfaces may be filled by statistically plausible but unsupported interpolation. | Every cross-scale link requires evidence, assumptions, and an applicability boundary. |
| **Uncertainty** | What does the model not know, and how does uncertainty change under distribution shift? | Prediction error can arise from noise, parameter uncertainty, model form, missing mechanisms, | Controls interpretation, abstention, and experimental escalation. | Confidence scores may be presented as calibrated uncertainty. | Uncertainty must be evaluated for the relevant state, intervention, context, and use. |

| | | | | | |
|---|---|---|---|---|---|
| | | measurement, and domain shift. | | | |
| **Counterfactual structure** | What would change under an alternative intervention? | Counterfactual claims require causal assumptions and a defined comparison between intervention states. | Supports comparative consequence reasoning. | Generated alternatives may be mislabeled as causal treatment effects. | Counterfactual validity is conditional on identification assumptions and supporting evidence. |
| **Experimental falsification** | Which observation could show that the simulated consequence is wrong? | Perturbation studies can test intervention-specific states, interactions, and mechanisms. | Converts architectural plausibility into an empirical research program. | Validation may be limited to reconstruction or favorable retrospective tasks. | A world-model claim must remain revisable when predicted consequences fail experimentally. |

*What Pharmaceutical World Models Must Represent*

A pharmaceutical world model must first represent the state of the biological system, not merely the content of the data used to describe it. A biological state may include molecular abundance, regulatory activity, pathway configuration, cellular phenotype, tissue composition, disease status, physiological condition, exposure, prior treatment, and environmental context. The required state is necessarily selective rather than exhaustive: no model can encode every molecular event or organismal variable. The relevant criterion is whether the representation preserves the information needed to predict and test the consequences covered by its intended use. Proposals for AI-enabled virtual cells similarly emphasize multimodal state representations, dynamic transitions, perturbations, and relationships across biological scales, while acknowledging that a complete virtual representation remains an aspirational scientific program rather than an achieved general-purpose simulator [11].

Foundation models for single-cell biology show how shared representations may integrate transcriptomic and multi-omic information and transfer across tasks [12]. Such models provide an important component of a pharmaceutical world model because they can encode cellular context more richly than isolated molecular descriptors or manually selected biomarkers. Nevertheless, transferability across analytical tasks is not equivalent to consequence validity. A state representation may support cell annotation, integration, or masked-value prediction while failing to preserve dose sensitivity, treatment history, rare-state behavior, or the variables that determine an adverse response. The proposed architecture therefore requires state representations to be evaluated against the interventions and outcomes they are expected to support, rather than accepted as biologically sufficient because they are large, pretrained, or successful across unrelated downstream tasks.

Biological state must also preserve relational structure. Gene activity, signaling, metabolism, target dependence, compensatory pathways, and disease phenotypes arise from interactions rather than independent feature values. Transfer learning in network biology demonstrates that models can encode context-sensitive relationships among genes and use those relationships for downstream biological predictions [13]. The proposed world model extends this principle by requiring explicit interfaces among molecular, pathway, cellular, tissue, exposure, and physiological representations. These interfaces need not all be implemented through the same modeling formalism: learned representations, mechanistic equations, probabilistic graphical models, pharmacokinetic modules, and quantitative systems-pharmacology components may coexist. What matters is that their exchanged states, assumptions, units, temporal resolutions, and uncertainty meanings remain interpretable and testable. **Figure 1** presents the pharmaceutical consequence-simulation world model, showing how the article's key components and boundaries are connected within what pharmaceutical world models must represent.

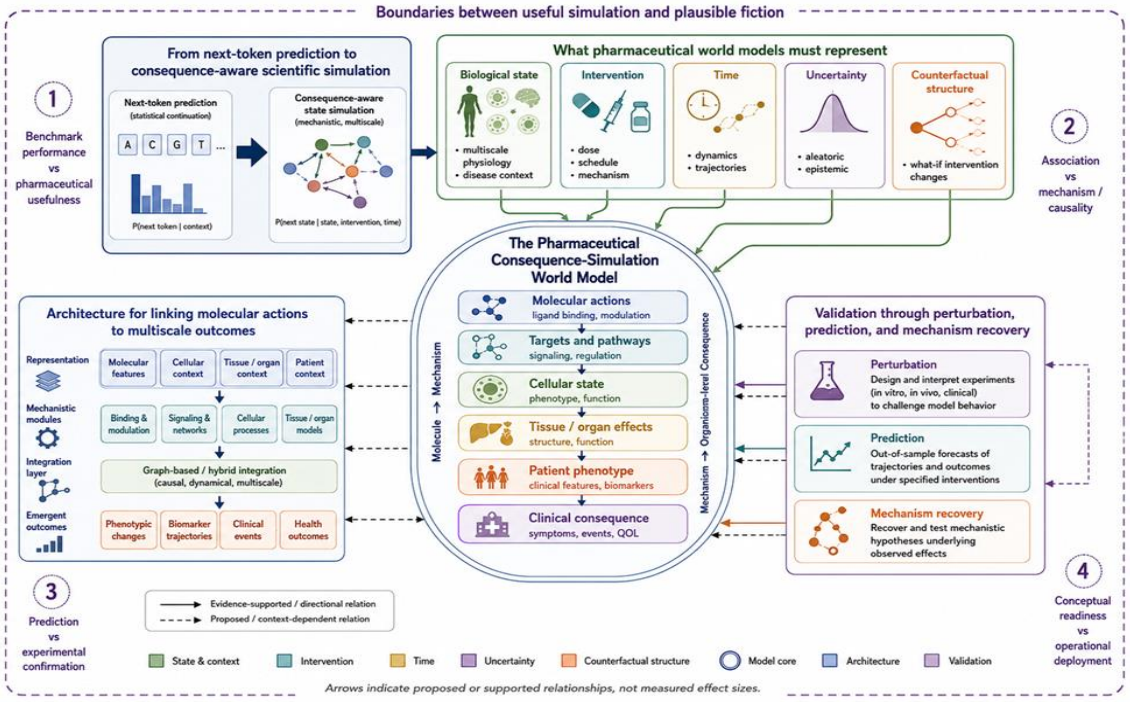


**Figure 1.** Pharmaceutical Consequence-Simulation World Model.

The figure is an original conceptual synthesis that organizes the article's central contribution across next-token prediction to consequence-aware scientific simulation, pharmaceutical world models must represent, biological state, intervention, time, uncertainty, and counterfactual structure, biological state, counterfactual structure, architecture for linking molecular actions to multiscale outcomes. 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.

A world model must finally encode the transformation between states under a defined intervention. Perturbation-response models demonstrate that learned latent spaces can approximate changes between cellular distributions after treatment, providing an important precursor to consequence simulation [14]. The proposed architecture, however, requires the intervention to contain more than a perturbation label. It should represent the molecular or biological action, dose, route, schedule, duration, combination, target-engagement assumptions, prior treatment, and relevant context. The transition should produce a probability distribution over future states rather than a single deterministic answer, because biological responses are heterogeneous and incompletely observed. It should also distinguish the latent state from the assays used to measure it, allowing apparent prediction errors to be traced to transition dynamics, observation noise, missing context, or model-form failure. A pharmaceutical world model is therefore defined not by scale alone, but by whether its representations support explicit, uncertain, intervention-conditioned, temporally ordered, and experimentally contestable claims about biological consequences.

*Biological State, Intervention, Time, Uncertainty, and Counterfactual Structure*

The proposed architecture begins with a formal distinction between biological state and the data through which that state is observed. A biological state is the condition of the modeled system at a specified time and level of resolution; it may include molecular concentrations, regulatory activity, pathway configuration, cell composition, tissue organization, physiological variables, disease features, exposure, prior treatment, and environmental context. Measurements such as transcriptomic profiles, imaging features, biomarkers, or assay readouts are observations generated from that state rather than the state itself. This distinction is essential because different assays may provide incomplete, noisy, delayed, or technically distorted views of the same underlying biology. A world model that equates an assay vector with biological reality may reconstruct measured data accurately while failing to represent the variables that determine response to intervention.

An intervention must be represented as an operation on the biological state rather than as an additional descriptive token. Causal representation learning provides the theoretical motivation for separating variables that predict one another from variables and mechanisms that remain meaningful under intervention [15]. In the proposed architecture, an intervention specification includes the manipulated entity, hypothesized target or mechanism, dose, route, schedule, duration, treatment sequence, combination status, and biological context. The required resolution depends on the intended claim. A compound identifier may be sufficient for retrieval of an observed assay profile but inadequate for predicting tissue exposure, delayed pathway adaptation, or combination toxicity. The intervention operator must therefore expose what is manipulated, what is assumed about its action, and which parts of the resulting transition are learned from data, mechanistically constrained, or presently unknown.

Counterfactual structure adds a more demanding requirement. Predicting an observed outcome for a treated sample is not identical to estimating what would have happened to the same biological system under a different intervention or no intervention. Causal treatment-outcome modeling requires a clearly specified estimand and assumptions concerning treatment assignment, confounding, consistency, positivity, and transportability [16]. Accordingly, the proposed world model should not label every alternative generated trajectory as a counterfactual. Counterfactual claims should be reserved for comparisons supported by an explicit causal structure and an identified relationship between interventions and potential outcomes. Where these conditions are not met, the architecture should describe outputs as conditional simulations, scenario analyses, or hypothesis-generating alternatives rather than causal treatment effects.

Time and uncertainty complete the core state-transition structure. Biological consequences unfold through exposure, delay, feedback, adaptation, compensation, recovery, progression, and repeated intervention. Distributional trajectory methods illustrate how time-separated cell populations may be linked to infer probable state movement, although inferred trajectories are not direct observations of individual histories or proof of mechanism [17]. Every predicted transition should therefore be indexed by time and accompanied by uncertainty reflecting biological variability, measurement error, parameter uncertainty, model-form uncertainty, missing mechanisms, and domain shift. Neural uncertainty methods can behave differently under changing molecular prediction conditions, demonstrating that confidence estimates cannot be assumed to be interchangeable or calibrated [18]. The proposed architecture consequently treats uncertainty as a structured output attached to states, transitions, and counterfactual comparisons rather than as a single confidence score appended after prediction.

*Architecture for Linking Molecular Actions to Multiscale Outcomes*

The Pharmaceutical Consequence-Simulation World Model is proposed as a modular architecture rather than a single monolithic neural network. Its central object is a transition distribution that maps an initial biological state, intervention, context, and time specification to possible future states. Surrounding this transition are specialized components for molecular representation, target and pathway reasoning, cellular-state modeling, tissue and physiological coupling, exposure, observation generation, uncertainty, counterfactual comparison, provenance, and experimental revision. Modularity is scientifically important because each scale may require a different representational language, temporal resolution, or validation strategy. It also permits a component to be revised or rejected without treating the entire architecture as a single indivisible source of truth.

The molecular-to-cellular interface must connect a drug or biological intervention to target engagement, pathway modulation, compensatory signaling, and phenotypic response. Mechanistic pathway models demonstrate that parameterized biological relationships can link molecular interventions with predicted response states, while also exposing assumptions and parameter uncertainty [19]. Within the proposed architecture, mechanistic modules may constrain learned state transitions, provide interpretable intermediate variables, or identify violations of known biological relationships. They should not be treated as automatically correct merely because they use equations or named pathways. Mechanistic models may omit relevant feedback, represent only selected cell types, inherit uncertain parameters, or fail when transported to a new disease context. Their role is therefore to make assumptions explicit and testable rather than to guarantee mechanistic validity.

The cellular-to-organism interface requires additional layers for tissue context, distribution, metabolism, exposure, physiological regulation, disease progression, and population variability. Quantitative systems pharmacology offers a relevant foundation because it integrates pharmacology, systems biology, disease mechanisms, and quantitative simulation across multiple levels of organization [20]. The proposed architecture extends this logic by allowing learned representations and mechanistic modules to exchange bounded state variables. For example, a cellular model may predict a pathway or phenotypic response conditioned on target engagement, while a pharmacokinetic component constrains whether the required exposure is attainable and a tissue-level component determines whether the relevant cell state is present. These interfaces must preserve units, time scales, assumptions, and uncertainty so that apparently coherent outputs do not conceal contradictions between modules.

Validation of such an architecture cannot be reduced to fitting all available data simultaneously. Quantitative systems-pharmacology validation is necessarily tied to context of use, model purpose, assumptions, and the evidence required for a particular decision [21]. Population simulation and in silico trial concepts may extend mechanistic models toward heterogeneous outcome scenarios, but simulated populations do not substitute for clinical observations or establish clinical utility [22]. The architecture should therefore maintain a provenance and applicability ledger recording data sources, assay conditions, parameter origins, model versions, calibration domains, failed tests, and permitted claim levels. Its expected output is not a single definitive forecast but a structured set of possible consequences, their supporting mechanisms or associations, uncertainty, predicted observations, and explicit conditions under which the result should be rejected, qualified, or tested experimentally.

*Validation through Perturbation, Prediction, and Mechanism Recovery*

A consequence-aware world model should be validated by asking whether it predicts what changes after a defined intervention, not merely whether it reconstructs the dominant structure of an observational dataset. Genetic interaction screens provide rich single-cell phenotypes for testing whether a model recovers nonlinear interaction structure; multiplex chemical transcriptomics measures heterogeneous cellular responses to pharmacological perturbations; and prediction of unseen multigene interventions tests whether a model generalizes beyond conditions explicitly represented during training [23-25]. Together, these evidence types support a validation strategy centered on held-out interventions, matched controls, context variation, and response-

specific structure. They do not establish that transcriptomic prediction alone is equivalent to tissue, organismal, safety, or therapeutic consequence.

Aggregate similarity metrics can obscure biologically weak models. A prediction may correlate highly with the observed post-treatment state because it reproduces cell identity, baseline expression, batch structure, or a response shared across many perturbations. Systema was developed to evaluate perturbation-response prediction beyond such systematic variation and to examine whether models discriminate individual perturbations and recover the organization of a perturbation landscape [26]. This principle is central to the proposed validation framework. Evaluation should subtract or control for readily predictable background structure and determine whether the model captures the direction, specificity, magnitude ordering, and context dependence of the intervention effect. Simple baselines, including control-state, mean-response, nearest-neighbor, and linear models, should remain visible so that complexity is not mistaken for biological information.

Mechanism recovery requires stronger evidence than accurate endpoint prediction. A model claiming to represent mechanism should predict the effects of intervening at multiple points in the proposed pathway, including target inhibition, genetic disruption, mediator alteration, rescue, or compensatory activation. The expected evidence may include correct response direction, interaction structure, mediation patterns, temporal ordering, and agreement across orthogonal assays. Post hoc pathway enrichment, attention maps, or generated explanations can support hypothesis formation but cannot alone establish that the named mechanism produced the prediction. Prospective experiments should be selected partly for their capacity to falsify the model, distinguish competing mechanisms, and reveal where apparently correct predictions arise for the wrong biological reason. **Figure 2** presents the evidence, validation, and translation boundary observatory for world models for pharmaceutical science should simulate, showing how the article's key components and boundaries are connected within validation through perturbation, prediction, and mechanism recovery.

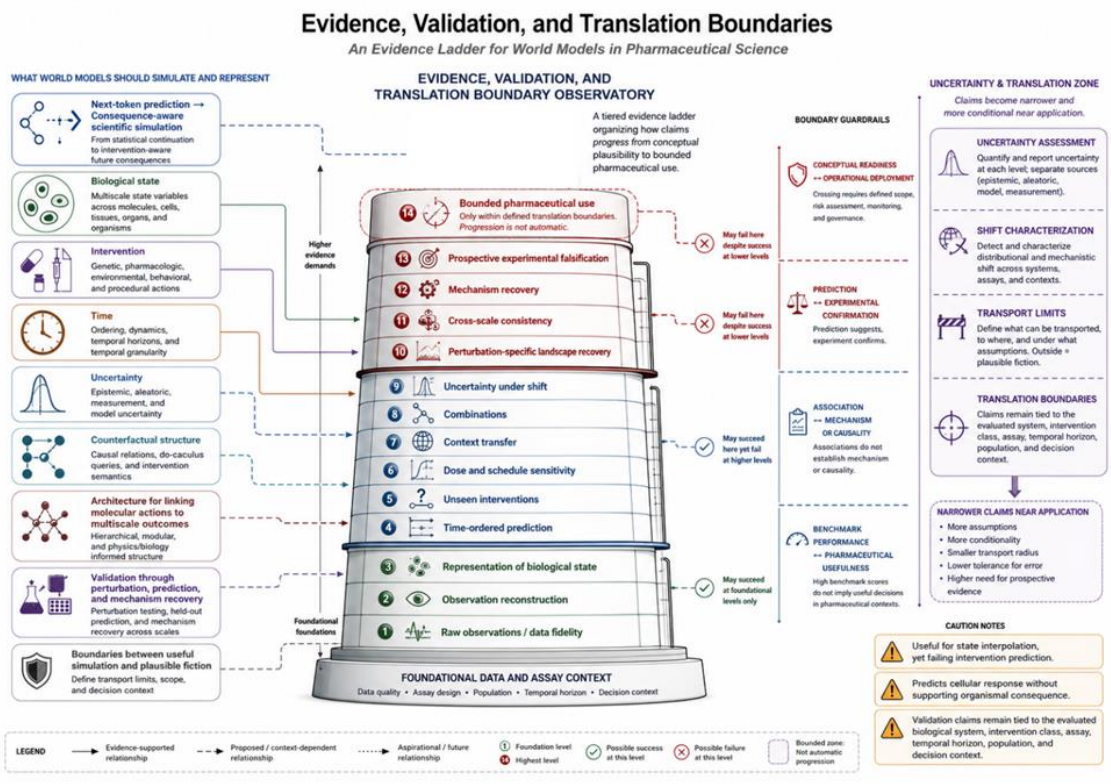


**Figure 2.** Evidence, Validation, and Translation Boundaries.

The figure is an original conceptual synthesis that shows how claims move from conceptual plausibility through evaluation, uncertainty assessment, and bounded pharmaceutical use. 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.

The complete validation strategy is therefore an evidence ladder rather than a single score. Lower levels test whether the model reconstructs observations and represents biological state. Intermediate levels examine time-ordered prediction, unseen interventions, dose and schedule sensitivity, context transfer, combinations, and uncertainty under shift. Higher levels test perturbation-specific landscape recovery, cross-scale consistency, mechanism recovery, and prospective experimental falsification. Progression is not automatic: a model may be useful for state interpolation while failing intervention prediction, or predict a cellular response without supporting an organismal consequence. Validation claims must remain attached to the specific biological system, intervention class, assay, temporal horizon, population, and decision context that were evaluated.

*Boundaries between Useful Simulation and Plausible Fiction*

The first boundary concerns memorization and generalization. Ligand-based benchmarks can reward models for recognizing compounds or neighborhoods similar to those represented in training rather than learning relationships that generalize to unfamiliar chemical space [27]. A world model may exhibit the same problem at multiple levels: it may retrieve familiar molecular responses, reproduce common cell-state shifts, or continue a frequently observed disease trajectory without modeling why the transition occurs. Realistic evaluation therefore requires intervention-level, scaffold-level, context-level, temporal, and where possible mechanism-level holdouts. Randomly separating observations while leaving nearly identical interventions or biological contexts on both sides of the split may test interpolation but cannot support a broad consequence-simulation claim.

The second boundary concerns local discontinuity. Activity cliffs show that small structural differences can correspond to large changes in biological activity and that strong average performance can conceal serious errors in chemically local neighborhoods [28]. Analogous cliffs may occur at higher levels when modest changes in dose, timing, genotype, tissue state, co-medication, or pathway activation produce qualitatively different outcomes. A model trained to favor smooth statistical continuation may suppress these transitions because they are uncommon or difficult to infer from sparse data. Consequence-aware evaluation must therefore include local stress tests, boundary cases, rare biological states, and interventions expected to expose nonlinearities. Failure at such boundaries does not render the model universally useless, but it sharply restricts the domain in which its outputs should be interpreted.

The third boundary concerns evaluation design and claim discipline. Chemical machine-learning guidance emphasizes that data construction, splitting strategy, baselines, applicability analysis, uncertainty, and intended use jointly determine the credibility of a reported result [29]. For pharmaceutical world models, evaluation must similarly align with the consequence being claimed. Reconstruction supports a representation claim; held-out perturbation prediction supports a bounded predictive claim; mechanism-directed intervention may support a mechanism-recovery claim; and prospective replication may support usefulness in a defined research context. None of these automatically establishes clinical benefit, safety, regulatory acceptability, or broad deployment readiness. The model's interface should communicate these claim ceilings rather than presenting all outputs with the same visual or linguistic authority.

The final boundary concerns data suitability. Pharmaceutical data are generated through assays, selection processes, measurement technologies, historical research priorities, and biological contexts that influence what is observed and labeled [30]. More data do not necessarily provide the variables, interventions, temporal coverage, or negative evidence needed to identify consequence dynamics. A model may produce a plausible trajectory because it has learned the language and typical correlations of biomedical data while lacking exposure measurements, target-engagement evidence, rare toxicity states, failed experiments, or relevant population diversity. In this article, plausible fiction therefore denotes an output that appears scientifically coherent but exceeds the model's evidentiary support, causal assumptions, applicability domain, or experimental confirmation. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop boundaries between useful simulation and plausible fiction within the article's central argument.

**Table 2.** Evaluation, implementation, and research priorities arising from World Models for Pharmaceutical Science Should Simulate Biological Consequences rather than Merely

| Component or layer | Inputs | Core function | Expected output | Uncertainty or limitation | Validation requirement |
|---|---|---|---|---|---|
| **Biological-state encoder** | Multimodal molecular, cellular, tissue, physiological, disease, exposure, and treatment-history data | Represent the intervention-relevant condition of the system | Time-indexed latent state with interpretable links to measured variables | State incompleteness, missing modalities, batch effects, representation bias | Demonstrate preservation of information needed for the intended perturbation and consequence tasks |
| **Intervention operator** | Molecular identity, target hypothesis, dose, route, schedule, duration, combination, and context | Apply a defined manipulation to the current state | Structured intervention-conditioned transition request | Incomplete target engagement, off-target effects, ambiguous exposure, adherence or delivery uncertainty | Test across held-out interventions, doses, schedules, and combinations |
| **Temporal transition model** | Initial state, intervention, context, time horizon, and prior state history | Generate distributions over possible future states | Time-resolved biological trajectories | Sparse longitudinal observations, unobserved intermediate states, model-form uncertainty | Evaluate prospective time points, onset, adaptation, recovery, and repeated-treatment behavior |
| **Multiscale coupling layer** | Molecular, pathway, cellular, tissue, pharmacokinetic, | Link consequences across | Cross-scale state updates and constraints | Incompatible units, time scales, aggregation, missing | Test linked predictions using measurements |

| | | | | | |
|---|---|---|---|---|---|
| | physiological, and disease-state variables | organizational scales | | feedback, uncertain parameters | obtained across multiple scales |
| **Observation and assay model** | Latent states and assay-specific measurement processes | Convert modeled states into testable observations | Predicted omics, imaging, biochemical, phenotypic, exposure, safety, or outcome measurements | Measurement error, censoring, platform effects, label noise | Validate separately from the latent transition model using matched assay conditions |
| **Causal and counterfactual layer** | Intervention structure, causal assumptions, covariates, and alternative actions | Compare possible consequences under defined alternatives | Conditional or causally identified intervention contrasts | Confounding, non-identifiability, positivity failure, unsupported causal assumptions | Use randomized interventions, negative controls, sensitivity analyses, or defensible identification strategies |
| **Mechanism and provenance ledger** | Biological relations, parameter sources, dataset origins, model versions, and assumptions | Make predictions traceable and contestable | Evidence-linked mechanism hypotheses and audit trail | Incomplete provenance, conflicting mechanisms, non-identifiable parameters | Reconstruct selected predictions and test proposed mediators through targeted perturbation |
| **Uncertainty and applicability controller** | Predictive distributions, domain distance, model disagreement, measurement limitations, and validation history | Qualify, restrict, or abstain from unsupported claims | Calibrated uncertainty and explicit claim boundary | Unknown unknowns, subgroup miscalibration, incomplete shift detection | Assess calibration, coverage, error detection, and abstention under relevant distribution shifts |
| **Perturbation-validation layer** | Genetic and chemical interventions, matched controls, combinations, and context variation | Determine whether the model recovers intervention-specific biology | Perturbation discrimination, landscape recovery, and context-specific response predictions | Shared background variation may dominate aggregate metrics | Compare with simple baselines and evaluate perturbation-specific effects rather than average reconstruction alone |
| **Mechanism-recovery layer** | Targeted knockout, inhibition, rescue, mediator, temporal, and orthogonal assay experiments | Distinguish predictive association from experimentally supported mechanism | Bounded mechanism-recovery claim | Post hoc explanations, incomplete pathways, alternative mechanisms | Require directional, mediation, rescue, or falsification evidence across multiple intervention points |
| **Prospective experimental feedback** | Frozen predictions, preregistered criteria, new perturbations, and measured outcomes | Falsify, update, or restrict the model | Revised modules, uncertainty, or narrowed applicability domain | Selective testing, feedback bias, failed experiments, retrospective reinterpretation | Replicate prospective tests across laboratories, platforms, contexts, and intervention classes |
| **Decision-use boundary** | Intended scientific decision, consequence of error, validation evidence, and oversight plan | Limit use to claims supported by available evidence | Research recommendation, abstention, escalation, or restricted scenario analysis | Automation bias, interface-driven trust, transfer of validation between uses | Qualify the model separately for each context of use and retain accountable human review |

*Limitations and Boundary Conditions*

The principal limitation of the proposed architecture is that a biologically complete state is not observable. Pharmaceutical systems span interacting molecular, cellular, tissue, physiological, behavioral, and environmental processes, many of which are measured sparsely or not at all. Any world model will therefore encode a selective and purpose-dependent state rather than the biological world in its entirety. The architecture can make this incompleteness visible through observation models, uncertainty, provenance, and applicability boundaries, but it cannot eliminate the underlying limitation. A state representation suitable for predicting an acute transcriptional response may be inadequate for chronic adaptation, immune effects, tissue remodeling, toxicity, formulation performance, or patient-level benefit.

A second limitation is that causal and mechanistic structure cannot generally be recovered from observational scale alone. Large multimodal datasets may improve representation learning while preserving confounding, selection bias, missing interventions, and weak temporal identification. Mechanistic modules may add scientific structure but can also create false reassurance when pathways are incomplete, parameters are poorly identified, or model equations reflect assumptions rather than established biology. Counterfactual trajectories remain conditional on causal assumptions, and experimental validation of one mechanism does not establish that the architecture captures all relevant mechanisms. The proposed construct should therefore be treated as a framework for organizing falsifiable claims, not as a method for converting correlation automatically into causality.

A third limitation concerns use beyond the validated context. Performance may change across laboratories, assay platforms, chemical series, doses, tissues, disease stages, demographic groups, co-medications, or clinical workflows. Human oversight does not by itself correct these shifts, particularly when users cannot inspect the relevant evidence or uncertainty. The architecture cannot establish clinical benefit, safety, regulatory acceptability, or operational readiness without application-specific evidence outside the scope of this conceptual article. Its implementation should remain bounded to clearly stated research uses, preserve the possibility of abstention, and prevent outputs generated for hypothesis formation from being silently repurposed as treatment recommendations or authoritative development decisions.

*Research Agenda and Implementation Priorities*

The first research priority is to develop interoperable representations of state, intervention, context, time, and observation. This requires shared schemas that distinguish latent biology from measurement, specify dose and schedule, preserve assay provenance, and connect learned representations with mechanistic and pharmacological variables. New datasets should emphasize longitudinal, factorial, and multimodal perturbations rather than merely increasing the number of passively observed samples. Particularly valuable designs would combine chemical and genetic interventions, multiple doses and times, treatment combinations, exposure measurements, matched controls, and orthogonal cellular or physiological outcomes. These resources should include failed and null experiments because absence of response is essential for learning applicability boundaries and avoiding selective success narratives.

The second priority is evaluation infrastructure that rewards intervention-specific and mechanistically informative performance. Benchmark design should use realistic holdouts, including new scaffolds, targets, perturbations, combinations, contexts, time points, laboratories, and populations. Results should be compared with transparent simple baselines and reported separately for state reconstruction, temporal forecasting, perturbation discrimination, context transfer, uncertainty, local discontinuity, cross-scale consistency, and mechanism recovery. Prospective challenges should freeze models and claims before experimental outcomes are known. Failed predictions should be retained and classified according to whether they arose from state representation, intervention specification, transition dynamics, observation error, missing mechanism, domain shift, or unsupported causal assumptions.

The third priority is staged implementation with explicit claim ceilings. Early systems should be used for bounded tasks such as generating discriminating experiments, comparing mechanistic hypotheses, or prioritizing perturbations for laboratory testing. Advancement to higher-consequence uses should require independent validation, reproducible provenance, calibrated uncertainty, defined abstention behavior, and evidence that users understand the model's permitted claims. Version changes should trigger re-evaluation rather than inherit validation automatically. Governance should assign responsibility for data suitability, model construction, experimental testing, interpretation, and final decisions. This staged approach does not assume that a universal pharmaceutical world model is achievable; it creates a practical route for evaluating whether narrower consequence models are scientifically informative, falsifiable, and appropriately bounded.

**Conclusion**

World models for pharmaceutical science should be defined by their ability to represent how explicit interventions alter biological systems through time, context, and scale—not by their capacity to continue statistically likely chemical or biomedical patterns. The proposed Pharmaceutical Consequence-Simulation World Model organizes this requirement around biological state, intervention operators, temporal transitions, multiscale coupling, observations, uncertainty, counterfactual structure, provenance, applicability boundaries, and experimental feedback. Its central contribution is an architectural and evidentiary distinction between fluent plausibility and bounded consequence simulation. The construct remains conceptual and does not establish that present models can reproduce complete biological systems, recover causal mechanisms universally, predict clinical outcomes, or support regulatory or operational use. Its value lies instead in making stronger claims conditional on stronger tests: held-out perturbations, dose and temporal variation, context transfer, mechanism-directed experiments, uncertainty under shift, cross-scale consistency, and prospective falsification. Pharmaceutical world models will become scientifically meaningful only when their failures are as visible and informative as their successful predictions.

**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. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673-85. doi:10.1038/s41586-023-05905-z
4. 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
5. 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
6. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci. 2018;4(1):120-31. doi:10.1021/acscentsci.7b00512
7. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83. doi:10.1021/acscentsci.9b00576
8. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
9. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8
10. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0
11. Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, Roed M, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell. 2024;187(25):7045-63. doi:10.1016/j.cell.2024.11.015
12. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470-80. doi:10.1038/s41592-024-02201-0
13. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616-24. doi:10.1038/s41586-023-06139-9
14. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715-21. doi:10.1038/s41592-019-0494-8
15. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proc IEEE. 2021;109(5):612-34. doi:10.1109/JPROC.2021.3058954
16. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nat Med. 2024;30(4):958-68. doi:10.1038/s41591-024-02902-1
17. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell. 2019;176(4):928-43. doi:10.1016/j.cell.2019.01.006
18. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
19. Fröhlich F, Kessler T, Weindl D, Shadrin A, Schmiester L, Hache H, et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. Cell Syst. 2018;7(6):567-79. doi:10.1016/j.cels.2018.10.013
20. Azer K, Kaddi CD, Barrett JS, Bai JPF, McQuade ST, Merrill NJ, et al. History and future perspectives on the discipline of quantitative systems pharmacology modeling and its applications. Front Physiol. 2021;12:637999. doi:10.3389/fphys.2021.637999
21. Kirouac DC. How do we "validate" a QSP model? CPT Pharmacometrics Syst Pharmacol. 2018;7(9):547-8. doi:10.1002/psp4.12310
22. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: Concepts and early adoptions. Brief Bioinform. 2019;20(5):1699-708. doi:10.1093/bib/bby043
23. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science. 2019;365(6455):786-93. doi:10.1126/science.aax4438
24. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders LM, Cao J, Packer JS, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020;367(6473):45-51. doi:10.1126/science.aax6234
25. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. 2024;42(6):927-35. doi:10.1038/s41587-023-01905-6

26. Viñas Torné R, Wiatrak M, Piran Z, Fan S, Jiang L, Teichmann SA, et al. Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. Nat Biotechnol. 2026;44(6):1050-9. doi:10.1038/s41587-025-02777-8
27. 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
28. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
29. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
30. 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