Pharmacophore

ISSN-2229-5402

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

CAN PHARMACEUTICAL WORLD MODELS BECOME SCIENTIFIC INSTRUMENTS? A STATE-OF-THE-ART REVIEW OF SIMULATION, REASONING, AND VALIDATION

Andreas Müller[1]*, Stefan Weber[1], Julia Hoffmann[2], Lukas Schneider[3], Tobias Klein[4]

1. ***Department of Pharmaceutical World Models as Scientific Instruments, Faculty of Pharmacy, Heidelberg University, Heidelberg, Germany.***
2. ***Department of Simulation and Reasoning in World Models, Faculty of Pharmaceutical Sciences, Technical University of Munich, Munich, Germany.***
3. ***Department of Validation and Instrument-Ready Models, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.***
4. ***Department of State-of-the-Art Review of Pharmaceutical Simulation, Faculty of Pharmacy, University of Stuttgart, Stuttgart, Germany.***

ARTICLE INFO

Received:
14 April 2026
Received in revised form:
06 August 2026
Accepted:
09 August 2026
Available online:
28 August 2026

Keywords: Pharmaceutical world models, Scientific instruments, Causal representation, Multimodal artificial intelligence, Molecular simulation, Hypothesis generation

ABSTRACT

Artificial intelligence is increasingly used to represent molecules, predict properties, generate candidates, infer biological states, plan syntheses, and coordinate experimental tools. These capabilities have encouraged the broader idea of a pharmaceutical world model: a computational system intended to represent relevant scientific states, anticipate how interventions alter those states over time, and support reasoning about downstream consequences. However, the evidentiary conditions under which such a system could be treated as a scientific instrument remain unresolved. This state-of-the-art review critically examines the frontier of pharmaceutical world-model architectures, representational commitments, reasoning capabilities, validation practices, failure modes, and readiness boundaries. The review uses a claim-centred analytical framework that distinguishes component capability from integrated simulation, retrospective benchmark performance from prospective experimental confirmation, association from causality, model confidence from calibrated uncertainty, and molecular plausibility from therapeutic relevance. The principal synthesis is a proposed scientific-instrument test requiring explicit representation of state, intervention, temporal transition, causal assumptions, uncertainty, applicability, and measurable biological consequences. Existing foundation, multimodal, generative, perturbational, planning, and agentic systems provide important components of this architecture, and selected prospective studies demonstrate that model outputs can contribute to experimentally testable hypotheses. Nevertheless, the available evidence does not establish a general, intervention-faithful, causally reliable, independently reproducible, and routinely qualified pharmaceutical world model. Important limitations include narrow task definitions, benchmark leakage, limited transportability, incomplete uncertainty evaluation, selective reporting of successful hypotheses, and weak linkage to downstream development decisions. Pharmaceutical world models may become scientific instruments only through intended-use qualification, prospective consequence-based testing, cross-site reproduction, transparent failure reporting, and continuing re-evaluation under distributional and workflow change.

To Cite This Article: Müller A, Weber S, Hoffmann J, Schneider L, Klein T. Can Pharmaceutical World Models Become Scientific Instruments? A State-of-the-Art Review of Simulation, Reasoning, and Validation. Pharmacophore. 2026;17(4):115-24. https://doi.org/10.51847/hn9ejVqJOc

Introduction

Artificial intelligence now contributes to target identification, molecular representation, property prediction, candidate generation, synthesis planning, and other activities distributed across pharmaceutical research. At the same time, contemporary drug-design systems increasingly combine learned representations with medicinal-chemistry knowledge, search procedures, and iterative experimental workflows. This breadth has stimulated expectations that computational models could progress from narrow predictors toward systems that represent scientific environments and anticipate the consequences of proposed

Corresponding Author: Andreas Müller; Department of Pharmaceutical World Models as Scientific Instruments, Faculty of Pharmacy, Heidelberg University, Heidelberg, Germany. E-mail: andreas.mueller@uni-heidelberg.de

interventions. Yet broad application coverage, architectural sophistication, and technical novelty do not by themselves demonstrate that a model functions as a reliable scientific instrument or produces decision-relevant pharmaceutical impact [1-3].

The distinction matters because pharmaceutical data are not neutral observations of a stable world. Chemical and biological datasets are shaped by assay selection, measurement error, historical research priorities, missing negative results, scaffold redundancy, inconsistent endpoints, and experimental context. Large datasets may therefore remain unsuitable for causal inference, temporal simulation, mechanistic explanation, or transport to new chemical and biological settings. A system trained on such evidence can reproduce patterns within the available record without adequately representing the processes that generated those patterns or the consequences of acting beyond the record [4].

The term *world model* is consequently at risk of becoming an expansive label for systems that perform fundamentally different functions. A molecular generator may produce chemically plausible structures without representing biological state transitions. A foundation model may transfer across tasks without encoding interventions or counterfactuals. A tool-using agent may execute searches and calculations without possessing a faithful internal model of pharmaceutical mechanisms. A dynamic cellular model may infer trajectories while remaining unable to establish causal effects. These capabilities are scientifically valuable, but collapsing them into a single category obscures differences among representation, prediction, generation, planning, simulation, explanation, and experimental validation.

This review asks whether pharmaceutical world models can become scientific instruments and, more specifically, what would have to be demonstrated before that description is scientifically defensible. Its contribution is evaluative rather than celebratory. First, it proposes an operational definition that separates world models from large predictors, generators, and agents. Second, it organizes current architectures according to their treatment of state, intervention, time, causality, uncertainty, and consequences. Third, it distinguishes simulation, reasoning, planning, tool use, and hypothesis generation as related but non-equivalent capabilities. Finally, it proposes scientific-instrument criteria and readiness boundaries that connect computational claims to prospective biological evidence, reproducibility, intended use, and continuing qualification. These constructs are review-derived conceptual syntheses and are not presented as empirically validated standards.

State-of-the-Art Review Scope and Analytical Framework

The evidentiary frontier for this review was fixed at 26 July 2026. A state-of-the-art design was used because the central question concerns a rapidly developing technological frontier rather than the pooled effect of a standardized intervention. The review therefore emphasizes current architectures, conceptual commitments, evaluation practices, and readiness boundaries. Its analytical process follows the core logic of state-of-the-art knowledge synthesis: define the purpose and frontier, identify evidence capable of supporting the planned claims, chart the relevant constructs, compare convergent and divergent findings, and produce an explanatory interpretation rather than a publication-by-publication summary [5].

The review scope was organized around seven linked questions: how the technological frontier should be delimited; what qualifies as a scientific world model; how state, intervention, time, causality, and uncertainty are represented; what simulation, reasoning, planning, and hypothesis-generation capabilities have been demonstrated; how models are validated against biological consequences; which failure modes prevent scientific-instrument use; and which research directions could increase readiness. The review design was matched to these purposes by combining architecture classification, evidence-maturity assessment, critical appraisal, gap analysis, and interpretive synthesis. This approach treats methodological transparency as a condition of credibility while recognizing that heterogeneous computational and experimental studies cannot be reduced to a single pooled estimate [6].

Eligible evidence had to contribute directly to a planned claim about pharmaceutical artificial intelligence, world-model components, biological-state modeling, chemical planning, prospective hypothesis testing, uncertainty, reproducibility, predictive validity, or future infrastructure. Evidence was appraised according to claim fit, methodological transparency, validation setting, reproducibility concerns, risk of overstatement, and the distance between the reported task and pharmaceutical decision relevance. Absence of reported information was treated as an evidentiary limitation rather than filled through inference. This proportionate approach preserves the distinction between a transparent state-of-the-art review and a systematic effectiveness review while retaining explicit eligibility, appraisal, extraction, and synthesis procedures [7]. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop state-of-the-art review scope and analytical framework within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for State-of-the-art review scope and analytical framework

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Review-method evidence	State-of-the-art, narrative-synthesis, and transparent-review methodology	A technological frontier can be examined through explicit scope, eligibility, coding, appraisal, and interpretive synthesis	Strong for methodological organization; indirect for pharmaceutical performance	Search terminology, database coverage, and reviewer judgement may affect inclusion and interpretation	No universally accepted reporting standard exists specifically for pharmaceutical world-model reviews

Pharmaceutical AI landscape	Critical reviews of machine learning across discovery and development	Artificial intelligence contributes to multiple pharmaceutical tasks and can reorganize design workflows	Strong for breadth of application; variable for demonstrated downstream impact	Positive-result emphasis, heterogeneous tasks, and inconsistent definitions of success	Limited evidence connecting broad AI adoption to reproducible pharmaceutical decision improvement
Architecture evidence	Foundation models, multimodal systems, generative models, neural–symbolic planning, and agentic systems	Current systems provide reusable representations, integration mechanisms, generation, planning, or tool coordination	Moderate to strong for individual components	Architectural comparisons often use different datasets, tasks, and evaluation criteria	No reviewed architecture demonstrates all required world-model commitments in one qualified system
State and transition evidence	Perturbation models, dynamic cellular models, temporal representation learning, and causal-representation methods	Models can represent biological variation, infer transitions, or predict selected responses to interventions	Moderate within defined datasets and biological contexts	Observational trajectories may be mistaken for causal mechanisms; transport across assays and laboratories is uncertain	Integrated state, dose, time, context, and counterfactual modeling remains incomplete
Capability evidence	Molecular generation, synthesis planning, tool-augmented reasoning, autonomous chemistry, and prospective discovery studies	Systems can generate hypotheses, prioritize candidates, plan actions, and execute bounded workflows	Strong for selected bounded demonstrations; limited for generalization	Successful outputs may depend on retrieval, memorization, filtering, external tools, or unreported human intervention	Faithful simulation and mechanism-sensitive reasoning are rarely tested directly
Validation evidence	Benchmarks, external datasets, prospective synthesis, biological assays, mechanistic follow-up, and independent reproduction	Validation can range from retrospective task performance to experimentally confirmed consequences	Strong evidence that validation setting changes the meaning of performance	Leakage, scaffold similarity, endpoint mismatch, selective reporting, and weak comparator design	Few studies provide preregistered, blinded, cross-site, decision-impact validation
Scientific-instrument readiness	Reproducibility standards, uncertainty evaluation, predictive-validity analysis, and sociotechnical risk assessment	Scientific use requires traceability, calibration, applicability assessment, reproducibility, consequence relevance, and intended-use control	Strong conceptual support; limited direct qualification evidence for integrated world models	Technical repeatability can be confused with predictive validity or operational readiness	No general pharmaceutical world model has established routine, monitored scientific-instrument status

Definitions and Architectures of Scientific World Models

Current architecture families provide complementary but incomplete foundations for pharmaceutical world models. Generalist foundation models contribute transferable pretrained representations; multimodal systems provide mechanisms for aligning heterogeneous biomedical observations; and hybrid neural–symbolic systems connect learned knowledge to explicit search or planning procedures [8-10]. These elements matter because a scientific world model must do more than compress data. It must maintain a representation of the relevant environment, distinguish observations from interventions, update state through time, and expose consequences in forms that can be compared with experiments. The proposed definition in this review is therefore narrower than the common use of *foundation model*, *generative model*, or *agent*: a pharmaceutical world model is an intervention-conditioned computational representation of scientifically relevant states and transitions that produces testable consequence distributions together with explicit uncertainty and applicability boundaries.

Generative biological models illustrate both the promise and incompleteness of this definition. Protein language models can learn statistical and functional constraints sufficiently well to propose sequences that retain experimentally measurable activity across multiple protein families [11]. Such results show that generative pretraining can capture biologically meaningful structure rather than merely reproduce literal training examples. However, sequence-level functionality does not establish therapeutic suitability, organism-level effects, manufacturability, safety, exposure, or comparative clinical value. A protein

generator may therefore supply hypotheses or components to a pharmaceutical world model without itself representing the broader world in which those proteins act.

Architecturally, the proposed scientific-instrument test contains six connected commitments. First, the system must define an explicit state comprising the chemical, biological, experimental, and contextual variables relevant to its intended use. Second, it must encode interventions separately from observations so that a compound, dose, schedule, combination, or experimental manipulation is not treated as another passive feature. Third, it must contain a transition model that relates current state and intervention to future state over an appropriate time scale. Fourth, any causal or counterfactual claim must be linked to stated assumptions rather than inferred from predictive performance alone. Fifth, outputs must include calibrated uncertainty, applicability assessment, and the option to abstain. Sixth, predicted consequences must be mapped to measurable experimental outcomes capable of updating, challenging, or invalidating the model. **Figure 1** presents the scientific-instrument test for pharmaceutical world models, showing how the article's key components and boundaries are connected within definitions and architectures of scientific world models.

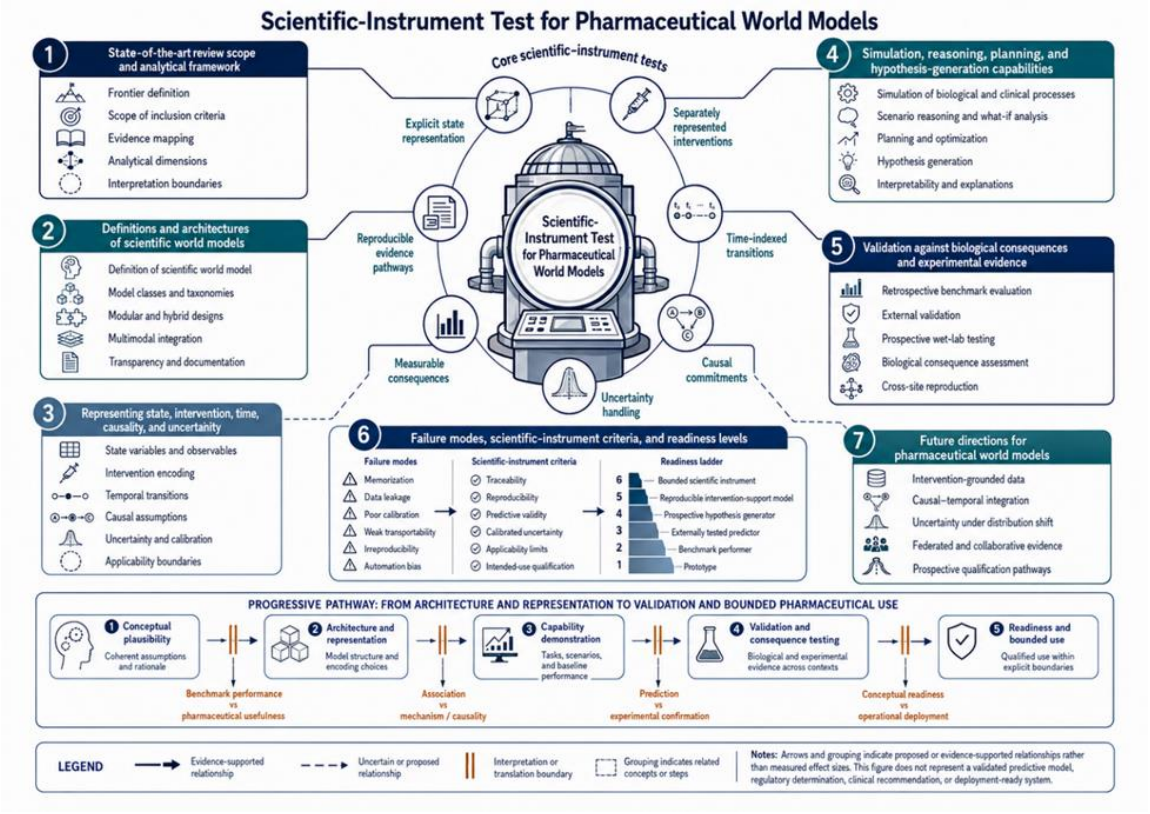

Figure 1. Scientific-Instrument Test for Pharmaceutical World Models.

The figure is an original conceptual synthesis that organizes the article's central contribution across state-of-the-art review scope and analytical framework, definitions and architectures of scientific world models, representing state, intervention, time, causality, and uncertainty, representing state, simulation, reasoning, planning, and hypothesis-generation capabilities, hypothesis-generation capabilities. 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 test deliberately separates internal architecture from evidentiary qualification. A foundation model may supply a reusable state encoder, but transfer across tasks does not establish intervention fidelity [8]. A multimodal system may align structures, images, omics, and text, but alignment does not guarantee that the fused representation corresponds to a causally coherent biological state [9]. A neural–symbolic planner may conduct explicit search through possible chemical actions, but planning success does not establish that biological outcomes have been simulated [10]. Likewise, a generative protein model may produce functional sequences without representing the downstream pharmaceutical environment in which function becomes benefit, toxicity, resistance, or failure [11]. Pharmaceutical world models should therefore be classified by the commitments they demonstrably implement, not by scale, interface fluency, or branding. Under the proposed taxonomy, most current architectures are best interpreted as partial world-model components whose scientific value depends on how they are integrated, validated, and bounded for a specified use.

Representing State, Intervention, Time, Causality, and Uncertainty

A pharmaceutical world model requires an explicit account of *state*: the variables considered sufficient to describe the scientific situation relevant to an intended decision. Depending on the task, state may include molecular structure, target conformation, gene expression, cellular phenotype, tissue context, exposure, dose, prior treatment, and experimental conditions. Perturbation-response models demonstrate that latent cellular representations can be conditioned on interventions and used to predict changes across selected cellular contexts [12]. This capability is closer to world modeling than ordinary classification because it relates a represented state to a proposed action and an expected response. Nevertheless, a transcriptomic response remains only one projection of biological state. Unmeasured signaling, metabolism, spatial organization, assay conditions, and cellular history can alter the consequences of the same intervention. State adequacy must therefore be evaluated against the intended scientific use rather than inferred from the dimensionality or apparent richness of the representation.

Time is equally fundamental because pharmaceutical interventions act through processes with different delays, durations, and reversibility. Dynamical single-cell models can infer transient trajectories and directionality from molecular observations, extending analysis beyond static clustering and steady-state assumptions [13]. Such models provide useful representations of evolving biological systems, but an inferred trajectory is not automatically a measured transition law. Temporal predictions may depend on sampling density, model assumptions, unobserved variables, and the correspondence between molecular change and function. A pharmaceutical world model should therefore specify its temporal resolution, transition horizon, memory assumptions, and treatment of path dependence. It should also distinguish immediate molecular engagement, delayed cellular adaptation, tissue-level response, organismal exposure, and longer-term resistance or toxicity, because success at one time scale does not validate predictions at another.

Causality determines whether a model can support questions about what would happen under an intervention rather than merely what tends to co-occur in observed data. Causal representation learning provides a conceptual basis for separating stable generative factors, interventions, and counterfactual structure from correlations that may change across environments [14]. In pharmaceutical science, this distinction is critical because the same molecular signature may arise from different mechanisms, and the same intervention may produce different effects across disease states or treatment histories. Causal claims require explicit assumptions about confounding, measurement, intervention assignment, and the stability of relationships across contexts. A model may predict a response accurately without identifying the responsible mechanism, and it may describe an association without supporting a counterfactual comparison. Accordingly, causal interpretation should be limited to the intervention and context for which assumptions and experimental evidence are defensible.

Uncertainty completes the minimum representational framework. Molecular prediction studies show that uncertainty-estimation methods differ materially in how well they identify likely errors and that apparent confidence cannot be assumed to be calibrated [15]. For a pharmaceutical world model, uncertainty should be attached not only to final predictions but also to state estimation, intervention representation, transition dynamics, causal assumptions, and applicability. Aleatoric variation, model uncertainty, measurement error, missing modalities, and distribution shift may require different treatments. The system should communicate when evidence is insufficient, when a prediction depends strongly on assumptions, and when abstention or additional experimentation is preferable to action. Calibration must also be decision-specific: an uncertainty estimate adequate for prioritizing inexpensive assays may be inadequate for eliminating a candidate, changing a development program, or extrapolating to a new patient population.

Simulation, Reasoning, Planning, and Hypothesis-Generation Capabilities

Hypothesis generation is one of the clearest demonstrated capabilities at the current frontier. Generative modeling has contributed to the identification of kinase inhibitors that were synthesized and experimentally tested, providing evidence beyond virtual novelty or internal scoring [16]. The scientific value of this result lies in producing a proposition that could be challenged by experiment. However, successful generation does not show that the model simulated the full biological system. Candidate selection may reflect learned chemical patterns, target-specific predictors, optimization objectives, human filtering, or external calculations. A world-model claim would require the system to represent why the intervention should alter a relevant state, which downstream consequences are expected, how uncertainty propagates, and under which conditions the prediction should fail. Prospective molecular generation is therefore best treated as experimentally anchored hypothesis formation rather than automatic evidence of mechanistic simulation.

Prediction-led discovery similarly shows how computational systems can search large chemical spaces and prioritize hypotheses that differ from familiar compounds. Deep learning has been used to identify antibacterial candidates that proceeded to experimental evaluation, demonstrating that model-guided prioritization can produce biologically consequential findings [17]. This evidence is stronger than retrospective classification because the selected candidates encounter physical and biological tests not contained in a benchmark score. Yet the resulting capability remains conditional on the task definition, training labels, screening library, assay design, and follow-up strategy. Prediction can direct attention efficiently without representing the complete causal structure of bacterial response, resistance, host effects, or clinical exposure. The distinction between prioritization and simulation should therefore remain explicit even when prospective experiments succeed.

Reasoning claims require particular caution because fluent explanations and successful task completion can arise from several mechanisms. Chemistry language models augmented with calculators, databases, search systems, code execution, and specialist tools can solve bounded tasks more effectively than unaided language generation [18]. This architecture is scientifically useful because it externalizes operations that require precision and allows intermediate outputs to be inspected. Nevertheless, tool use does not establish that the language model possesses a faithful internal representation of chemistry. An

answer may result from retrieval, pattern matching, decomposition, or correctly invoking an external program. Evaluation should therefore identify which component produced each result, whether intermediate claims are traceable, how errors propagate across tools, and whether the system can detect contradictions or invalid assumptions. Tool-augmented performance is evidence of orchestrated problem solving, not by itself evidence of mechanistic reasoning.

Planning extends this orchestration across multiple actions. Autonomous chemical systems have integrated literature retrieval, code, experimental planning, and laboratory interfaces to perform bounded research tasks [19]. Such demonstrations show that an agent can connect computational and physical operations, adapt a plan to intermediate information, and produce experimentally testable outputs. They do not yet establish a general pharmaceutical world model. The represented environment may be limited to available tools and encoded procedures; biological consequences may remain outside the planning loop; and human intervention may be difficult to separate from autonomous action. A rigorous capability taxonomy should therefore distinguish simulation of state transitions, reasoning over explicit scientific claims, planning of actions, execution through tools, and generation of hypotheses. These capabilities may be combined, but evidence for one should not be used as proof of the others.

Validation against Biological Consequences and Experimental Evidence

Validation begins with clarity about what a test can establish. Molecular benchmark suites standardize datasets, endpoints, and splits, enabling comparison among algorithms on defined prediction tasks [20]. They are valuable for detecting implementation errors, estimating relative performance, and studying representation choices. Their evidentiary reach is nevertheless limited. A benchmark evaluates behavior under a specified data-generating and partitioning scheme; it does not automatically test whether the model will support a future pharmaceutical decision. Benchmarks may omit intervention timing, assay variability, negative results, development constraints, and the downstream consequences that matter to drug discovery. A pharmaceutical world model should therefore not be considered validated merely because its components perform well on established datasets. Benchmark competence is an early validation layer, not a substitute for consequence-based evaluation.

Representation studies further demonstrate that apparent superiority depends on the task, training regime, data volume, and evaluation split. Comparative analysis of learned molecular representations shows that no single architecture dominates all property-prediction conditions and that conclusions can change with dataset and validation design [21]. This context dependence is important for world models because a representation useful for one endpoint may discard information essential to another. Internal validity requires careful split selection, appropriate baselines, control of leakage, and analysis of performance across chemical neighborhoods. External validity requires evidence that relevant relationships persist across new compounds, assays, targets, laboratories, and biological contexts. Model comparison should consequently be framed around intended use and failure conditions rather than a single aggregate ranking.

Prospective target-based studies provide a more demanding form of evidence. An AI-supported workflow using predicted structural information contributed to the discovery and experimental testing of a small-molecule inhibitor for a defined target [22]. This type of result demonstrates that computational representations and prioritization can be connected to laboratory evidence in a real discovery program. However, one successful target-specific application does not establish broad predictive validity. The result may depend on target tractability, structural quality, compound-selection choices, assay configuration, and expert intervention. Prospective validation should therefore report the complete decision pathway, including rejected hypotheses, failed compounds, model revisions, and the contribution of conventional methods. Without these denominators, the incremental value and reliability of the computational system remain difficult to estimate.

The strongest available validation examples connect model-guided prioritization to multiple layers of experimental consequence. Explainable deep learning has supported the discovery of an antibiotic structural class followed by biological testing and mechanistic investigation [23]. Such studies move beyond an isolated activity result by examining whether the predicted candidates produce coherent downstream effects. Even so, scientific-instrument qualification requires further evidence: independent reproduction, predeclared endpoints, comparison with expert and conventional baselines, calibration of prospective failure probabilities, and testing across settings that differ from model development. Validation should thus be organized as a ladder from retrospective performance to external prediction, prospective wet-laboratory testing, mechanism-sensitive intervention studies, cross-site reproduction, and decision impact. Advancement along this ladder is conditional, reversible, and specific to the intended pharmaceutical use.

Failure Modes, Scientific-Instrument Criteria, and Readiness Levels

A primary failure mode is mistaking chemical similarity recognition for general scientific prediction. Analysis of ligand-based benchmarks has shown that common evaluation designs can reward memorization or analogue recognition rather than extrapolation to genuinely new chemical space [24]. This problem is especially consequential for world-model claims because high performance may arise without learning transferable transition rules. Leakage can occur through duplicated compounds, shared scaffolds, correlated assay series, target-family overlap, or preprocessing decisions that preserve hidden information. Scientific-instrument evaluation should therefore include chemically and biologically meaningful holdouts, temporal separation where appropriate, adversarial checks, and explicit analysis of performance as distance from the training domain increases. A model that cannot identify its own extrapolations should not be trusted to simulate unfamiliar interventions.

Reproducibility is a second criterion, but it must be understood broadly. Life-science machine-learning standards emphasize transparent reporting of data provenance, preprocessing, code, model configuration, computational environment, and

evaluation procedures [25]. For a pharmaceutical world model, reproducibility also requires versioned state definitions, intervention encodings, tool dependencies, prompts or policies, stochastic settings, and experimental interfaces. Independent investigators should be able to reconstruct not only the final output but the evidence path by which it was generated. Reproducibility does not mean that every biological experiment yields an identical value; it means that variation is characterized, procedures are traceable, and conclusions remain stable within prespecified tolerances. A result that cannot be independently reconstructed cannot support strong scientific-instrument claims.

A third failure mode arises from the sociotechnical setting in which models are interpreted and acted upon. Machine learning can produce unintended consequences even when technical performance appears acceptable, particularly when users overtrust outputs, workflow incentives reward speed over challenge, or responsibility becomes distributed across people and systems [26]. In pharmaceutical research, an apparently precise recommendation may suppress alternative hypotheses, narrow chemical exploration, or obscure uncertainty introduced by upstream data and tools. Instrument qualification must therefore include human responsibilities, escalation rules, challenge pathways, documentation of overrides, and safeguards against automation bias. The relevant unit of evaluation is not only the model but the model–data–tool–experiment–human system through which scientific decisions are made.

The proposed readiness framework treats predictive validity as the central bridge between technical performance and scientific use. Predictive validity concerns whether a model anticipates the downstream consequences relevant to an actual discovery or development decision, rather than merely fitting existing measurements [27]. Under this synthesis, readiness progresses qualitatively from a representational prototype, to a retrospective benchmark performer, an externally tested predictor, a prospective experimental hypothesis generator, a reproducible intervention-support model, a translational decision-support candidate, and finally a routinely monitored scientific instrument. These levels are proposed analytical categories, not validated regulatory classifications. Advancement requires evidence appropriate to the intended use, and failure at calibration, transportability, reproducibility, consequence relevance, or workflow integration can halt or reverse progression. **Figure 2** presents the evidence, validation, and translation boundary observatory for can pharmaceutical world models become scientific instruments, showing how the article's key components and boundaries are connected within failure modes, scientific-instrument criteria, and readiness levels.

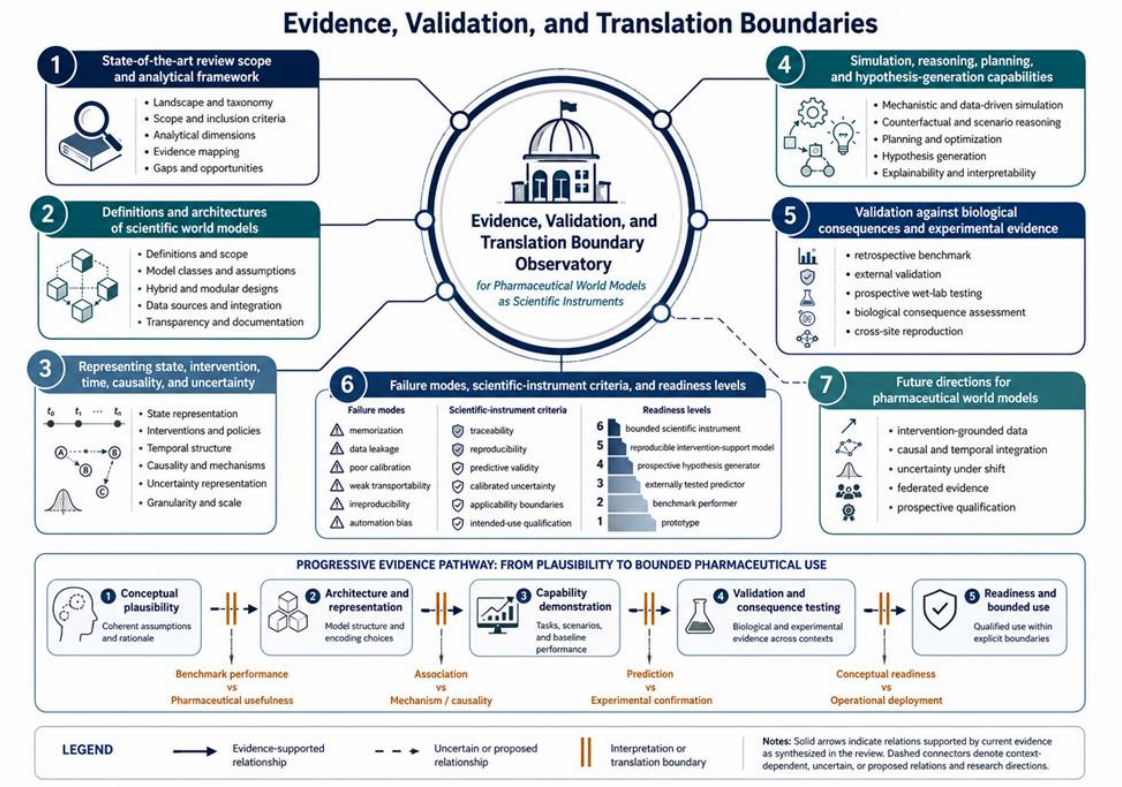

Figure 2. Evidence, Validation, and Translation Boundaries

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.

Future Directions for Pharmaceutical World Models

Future progress will depend partly on more transferable representations of biological state. Foundation models have begun to support predictions in network biology, multi-task analysis of single-cell multi-omics, and transcriptional representation across human cell types [28-30]. These systems may provide reusable encoders for cellular identity, regulatory context, and disease-associated variation. Their contribution to world modeling will remain limited, however, unless they are connected to explicit interventions, temporal transitions, causal assumptions, and consequence-based validation. Future architecture research should therefore prioritize modular integration rather than treating scale alone as the route to generality. A useful platform may combine specialized state encoders, mechanistic or learned transition components, uncertainty modules, and experimental interfaces while preserving traceability across modules.

Evidence diversity is another priority. Federated learning may enable institutions to develop or test molecular models across distributed proprietary datasets without centralizing all underlying records [31]. This could widen chemical, assay, target, and organizational coverage while preserving local control. Federated scale does not automatically improve validity: sites may use incompatible endpoints, assays, preprocessing rules, or patient and compound populations, and model updates may still reveal sensitive information. Collaborative world-model development should therefore include shared ontologies, harmonized endpoint definitions, site-level diagnostics, leakage assessment, privacy safeguards, and common external validation tasks. The objective should be distributed scientific challenge rather than merely distributed optimization.

Prospective evaluation should also move from selected demonstrations to structured research programs. Models should be locked before testing, their intended use and failure criteria specified, and their predictions compared with expert practice and conventional computational baselines. Full denominators are needed: proposed interventions, rejected candidates, failed experiments, abstentions, overrides, and model revisions should be recorded alongside successes. Uncertainty should be tested under chemical, biological, temporal, and laboratory shifts rather than only within familiar benchmarks [15]. Reproducibility packages should preserve model versions, data lineage, tool calls, experimental protocols, and human decisions so that independent groups can reconstruct the complete evidence path [25]. These practices would not guarantee translation, but they would make reliability and incremental scientific contribution assessable.

Finally, qualification should be organized around bounded intended uses. A system might be fit to prioritize compounds for a low-cost assay while remaining unfit to terminate a program, infer a mechanism, recommend a dose, or generalize to a new disease context. Each use should specify acceptable errors, required calibration, applicability boundaries, escalation rules, monitoring indicators, and requalification triggers. Predictive validity should be judged against the downstream consequence that matters for that use rather than against technical performance in isolation [27]. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop future directions for pharmaceutical world models within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Can Pharmaceutical World Models Become Scientific Instruments? A State-of-the-Art Review of Simulation

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Foundation biological representations	Network biology, single-cell multi-omics, transcriptional regulation, and protein sequences	Transfer across held-out tasks, contexts, or biological labels	Externally tested representational capability	Interventions, longitudinal transitions, causal transport, and uncertainty under shift remain incomplete	Promising state encoders, but not complete pharmaceutical world models
Multimodal state construction	Molecular structures, omics, imaging, phenotypes, assays, and scientific text	Modality-specific and fused-task evaluation; missing-modality testing	Emerging component-level evidence	Alignment may combine incompatible measurements or conceal missing causal variables	Multimodality increases observational coverage but does not guarantee coherent scientific state
Perturbation-response modeling	Cellular states under compounds, genetic changes, or environmental interventions	Held-out perturbations, cell contexts, doses, and time points	Moderate predictive maturity in bounded datasets	Transport across laboratories, disease states, and intervention regimes is uncertain	One of the most direct routes toward intervention-conditioned state prediction
Temporal and causal transition modeling	Longitudinal, dynamic, or perturbational biological data	Trajectory prediction, intervention tests, counterfactual checks, and mechanism-sensitive validation	Early and heterogeneous	Observational directionality may be mistaken for causal mechanism	Necessary for world-model status but requires stronger intervention-grounded evidence
Generative molecular and protein design	Candidate structures or sequences	Prospective synthesis, activity assays, functional	Prospective hypothesis-	Molecular novelty and activity do not establish	Demonstrates testable hypothesis generation rather

	optimized for selected objectives	characterization, and comparator analysis	generation evidence exists	developability, safety, exposure, or therapeutic value	than complete consequence simulation
Tool-augmented and agentic research systems	Literature, databases, code, synthesis planning, instruments, and laboratory procedures	Task completion, traceability, error analysis, human-intervention accounting, and cross-site repetition	Bounded workflow demonstrations	Tool dependence, hidden errors, security, reproducibility, and weak biological consequence modeling	Useful orchestration layer whose reasoning and autonomy claims require component-level testing
Federated pharmaceutical learning	Distributed institutional, proprietary, or sensitive datasets	Cross-site performance, site-level calibration, privacy analysis, and common external tasks	Early collaborative maturity	Heterogeneous endpoints, governance, privacy leakage, and incentive alignment	May increase evidence diversity if coupled to harmonized validation and transparent governance
Intended-use scientific-instrument qualification	Specific discovery or development decisions with defined consequences	Prospective decision-impact testing, calibration, abstention, monitoring, and requalification	Not established as a general routine capability	No single model is suitable for all pharmaceutical uses or consequence levels	Qualification should be use-specific, reversible, and continuously monitored rather than claimed from model scale

Conclusion

Pharmaceutical world models can plausibly become scientific instruments, but only if the term is reserved for systems that do more than predict, generate, retrieve, or automate. The central contribution of this review is a proposed scientific-instrument test that requires an explicit state, a separately represented intervention, a time-indexed transition model, stated causal assumptions, calibrated uncertainty, applicability boundaries, measurable consequences, and a reproducible path from prediction to experimental challenge. Current foundation, multimodal, perturbational, generative, planning, and agentic systems provide important elements of this architecture, and prospective studies show that model-guided hypotheses can contribute to successful experiments. The evidence nevertheless remains fragmented, task-specific, and insufficient to establish a general, causally reliable, independently reproducible, and routinely qualified pharmaceutical world model. Benchmark performance must remain separate from usefulness, association from mechanism, prediction from experimental confirmation, and conceptual readiness from operational deployment. Progress will depend on intervention-grounded longitudinal data, transparent negative-result reporting, uncertainty testing under meaningful distribution shifts, preregistered prospective comparisons, independent cross-site reproduction, and qualification for bounded intended uses. The proposed taxonomy and readiness framework are interpretive tools for organizing this transition, not validated standards or claims of current deployment readiness.

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. 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
4. 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

5. Barry ES, Merkebu J, Varpio L. State-of-the-art literature review methodology: A six-step approach for knowledge synthesis. Perspect Med Educ. 2022;11(5):281-8. doi:10.1007/s40037-022-00725-9
6. Snyder H. Literature review as a research methodology: An overview and guidelines. J Bus Res. 2019;104:333-9. doi:10.1016/j.jbusres.2019.07.039
7. Siddaway AP, Wood AM, Hedges LV. How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annu Rev Psychol. 2019;70:747-70. doi:10.1146/annurev-psych-010418-102803
8. Moor M, Banerjee O, Shakeri Hossein Abad Z, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. doi:10.1038/s41586-023-05881-4
9. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773-84. doi:10.1038/s41591-022-01981-2
10. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018;555(7698):604-10. doi:10.1038/nature25978
11. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099-106. doi:10.1038/s41587-022-01618-2
12. 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
13. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020;38(12):1408-14. doi:10.1038/s41587-020-0591-3
14. 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
15. 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
16. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
17. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702. doi:10.1016/j.cell.2020.01.021
18. 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
19. 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
20. 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
21. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
22. Ren F, Ding X, Zheng M, Korzinkin M, Cai X, Zhu W, et al. AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel CDK20 small molecule inhibitor. Chem Sci. 2023;14(6):1443-52. doi:10.1039/D2SC05709C
23. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2024;626(7997):177-85. doi:10.1038/s41586-023-06887-8
24. 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
25. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7
26. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797
27. Scannell JW, Bosley J, Hickman JA, Dawson GR. Predictive validity in drug discovery: What it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022;21(12):915-31. doi:10.1038/s41573-022-00552-x
28. 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
29. 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
30. Fu X, Mo S, Buendia A, Laurent AP, Shao A, Alvarez-Torres MDM, et al. A foundation model of transcription across human cell types. Nature. 2025;637(8047):965-73. doi:10.1038/s41586-024-08391-z
31. Hanser T. Federated learning for molecular discovery. Curr Opin Struct Biol. 2023;79:102545. doi:10.1016/j.sbi.2023.102545