Pharmacophore

ISSN-2229-5402

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

CAN FOUNDATION MODELS GENERATE PHARMACOLOGICAL HYPOTHESES THAT ARE MECHANISTICALLY COHERENT, FALSIFIABLE, AND WORTH TESTING?

Philippe Lefevre[1]*, Helga Schmidt[2], Thomas Fischer[3], Anne-Catherine Robert[4]

1. ***Department of Foundation Models and Pharmacological Hypothesis Generation, Faculty of Pharmaceutical Sciences, University of Strasbourg, Strasbourg, France.***
2. ***Department of Mechanistic Coherence in Hypothesis Generation, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.***
3. ***Department of Falsifiability and Testability in AI, Faculty of Pharmacy, University of Zurich, Zurich, Switzerland.***
4. ***Department of Hypothesis Worth Testing and Validation, Faculty of Pharmacy, University of Avignon, Avignon, France.***

ARTICLE INFO

Received:
14 June 2025
Received in revised form:
10 October 2025
Accepted:
14 October 2025
Available online:
28 October 2025

Keywords: Foundation models, Pharmacological hypothesis generation, Mechanistic coherence, Falsifiability, Experimental design, Evidence grounding

ABSTRACT

Foundation models can produce scientifically fluent accounts of targets, pathways, compounds, and disease mechanisms, but linguistic plausibility does not establish that an output constitutes a pharmacological hypothesis. A decision-worthy hypothesis must specify a directional intervention–mechanism–outcome relationship, identify the biological and experimental context in which the claim applies, expose assumptions and contradictory evidence, and state observations that could count against it. This Original Hypothesis-Generation Theory Article develops a conceptual framework for evaluating whether foundation-model outputs are mechanistically coherent, falsifiable, novel, experimentally discriminating, and consequential for pharmaceutical decision-making. The proposed contribution rejects the use of any single performance measure as a sufficient indicator of hypothesis quality. Instead, it treats quality as a structured profile comprising evidentiary grounding, mechanistic specification, falsifiability, scientific novelty, experimental feasibility, uncertainty, expected information gain, and decision consequence. The framework distinguishes prediction from explanation, association from mechanism, model confidence from calibrated uncertainty, and molecular plausibility from synthesizability, developability, safety, and therapeutic value. It further proposes that minimum evidentiary and falsifiability conditions should operate as non-compensatory gates: high fluency, novelty, or predicted activity cannot offset the absence of a traceable mechanism or an experiment capable of challenging the claim. The article outlines how molecular structures, biological networks, literature evidence, alternative hypotheses, and experimental constraints can be integrated into an auditable hypothesis object and evaluated prospectively. The framework remains theoretical and requires domain-specific, leakage-resistant, blinded, and experimentally grounded validation. Its purpose is not to authorize autonomous discovery decisions, but to define the conditions under which machine-generated propositions may become scientifically interpretable candidates for testing.

To Cite This Article: Lefevre P, Schmidt H, Fischer T, Robert AC. Can Foundation Models Generate Pharmacological Hypotheses That Are Mechanistically Coherent, Falsifiable, and Worth Testing? Pharmacophore. 2025;16(5):88-98. https://doi.org/10.51847/ynJWxXwyK3

Introduction

Artificial intelligence now supports prediction, molecular design, prioritization, and workflow automation across pharmaceutical research, but its useful contribution remains conditional on suitable evidence, objectives, experimental integration, and expert practice [1-3]. This conditionality is especially important for foundation models because their apparent breadth can obscure differences among tasks that require fundamentally different kinds of validity. Retrieving a known drug–target association, predicting a molecular property, generating a candidate structure, summarizing a mechanistic paper, and proposing a new pharmacological hypothesis are not interchangeable achievements. Each involves different objects of

Corresponding Author: Philippe Lefevre; Department of Foundation Models and Pharmacological Hypothesis Generation, Faculty of Pharmaceutical Sciences, University of Strasbourg, Strasbourg, France. E-mail: philippe.lefevre@unistra.fr.

inference, different opportunities for error, and different standards of evidence. Hypothesis generation is particularly demanding because it requires a system to move beyond reproducing correlations or familiar scientific language and instead formulate a directional claim whose assumptions, mechanistic dependencies, observable implications, and possible refutations can be examined. A model may therefore perform well on language generation or molecular benchmarks while remaining unable to produce a proposition that deserves experimental attention.

Prospective molecular-generation studies illustrate why computational output cannot be equated with pharmacological knowledge. In one prominent example, generated kinase-inhibitor candidates acquired scientific significance only after synthesis and biological testing connected the computational proposal to observable chemical and pharmacological behavior [4]. Such demonstrations are important because they show that model output can participate in a genuine discovery workflow. They do not, however, establish that a generative system understands the mechanism it describes, that its proposed explanation is causally correct, or that its preferred experiment is the most informative use of limited research resources. A generated molecule can be synthetically inaccessible, unstable, promiscuous, poorly exposed, unsafe, or active for reasons unrelated to the proposed mechanism. Likewise, a plausible textual hypothesis may contain real entities and correctly cited facts while joining them through an unsupported causal transition. The relevant question is therefore not whether a foundation model can produce a hypothesis-shaped sentence, but whether it can produce a structured claim whose evidentiary and experimental commitments are scientifically defensible.

This distinction matters because common computational evaluations often substitute accessible proxy tasks for the harder question of pharmaceutical usefulness. High predictive accuracy, favorable molecular-generation scores, successful retrieval, or expert-rated fluency may indicate competence within a bounded task, yet none alone demonstrates that a generated hypothesis will improve target selection, assay design, compound prioritization, translational reasoning, or termination of an unproductive program. Critical analyses of artificial intelligence in drug discovery have emphasized that technical performance on available endpoints can remain disconnected from meaningful impact when the endpoint, data-generating process, validation design, or decision context is poorly aligned with the intended use [5]. The same problem becomes more acute in hypothesis generation because the system is not merely ranking a predefined set of candidates. It is constructing the explanatory object to be judged, selecting which evidence to foreground, determining which uncertainties to omit, and implicitly proposing what should be tested next. Without a theory of hypothesis quality, evaluation risks rewarding persuasive narratives that cannot be discriminated from informed speculation.

This article develops such a theory for pharmaceutical foundation systems. Its original contribution is a proposed multidimensional account of a decision-worthy pharmacological hypothesis, organized around the distinction between fluent scientific text and an auditable hypothesis object. The argument proceeds in four steps. First, it defines the structural and epistemic features that distinguish a hypothesis from a plausible explanation. Second, it separates mechanistic coherence, falsifiability, novelty, and experimental value as related but non-equivalent dimensions. Third, it proposes a hypothesis-quality model in which evidence provenance, mechanistic links, explicit falsifiers, feasible experiments, uncertainty, and decision consequences are represented separately and subjected to non-compensatory minimum conditions. Fourth, it outlines how experiments and evaluation protocols should test not only whether a generated claim appears reasonable, but whether it survives leakage-resistant scrutiny, distinguishes among alternatives, and produces information capable of changing a pharmaceutical decision. The contribution is conceptual rather than empirically validated; it is intended to define a research object and evaluation agenda, not to establish clinical, regulatory, or deployment readiness.

What Distinguishes a Hypothesis from Fluent Scientific Text

Scientific and chemical AI systems can generate candidate designs, plan multistep actions, and invoke specialist tools, yet these capabilities remain dependent on task definition, external grounding, and the surrounding experimental infrastructure [6-8]. Their outputs may resemble expert reasoning because they use appropriate terminology, cite recognizable pathways, and organize claims into familiar explanatory forms. However, fluency is a property of expression, whereas a hypothesis is an epistemic commitment. A fluent passage can remain compatible with several mutually inconsistent mechanisms, omit the conditions under which its claim should hold, or avoid specifying any result that would weaken it. By contrast, a pharmacological hypothesis should identify at least an intervention or perturbation, a molecular target or process, a relevant biological context, a directional mechanistic chain, a predicted observation, and a possible observation that would count against the proposed relationship. These components convert an attractive narrative into an object that can be inspected, compared with alternatives, and exposed to evidence.

The difference becomes clearer when a generated proposition is carried into the laboratory. A recent study demonstrated that a large language model could participate in generating a breast-cancer treatment hypothesis that was subsequently subjected to experimental investigation [9]. The significance of such work lies not only in whether the initial proposition receives support, but in the conversion of language into a testable structure. That conversion requires choices about the biological model, intervention, comparator, dose or exposure context, temporal sequence, readout, controls, and interpretation of positive, negative, or ambiguous findings. A proposition such as “inhibiting pathway X may sensitize cells to drug Y” is not yet sufficiently specified merely because the target, pathway, and drug are real. It becomes a stronger hypothesis when it states where the relation is expected to occur, which mechanistic intermediate should change, how the combined intervention differs from plausible alternatives, and what outcome would challenge the proposed sensitization mechanism. Laboratory exposure

therefore does not simply verify generated prose; it reveals whether the prose contained enough structure to guide discriminating inquiry.

A useful distinction can be made among four increasingly demanding output classes. The first is scientific text, which may be grammatical and topically appropriate without making a testable commitment. The second is a scientific proposition, which asserts a relation but may not specify directionality, context, or evidentiary basis. The third is a testable hypothesis, which links a defined perturbation to a predicted observation under specified conditions and identifies a potential falsifier. The fourth is a decision-worthy pharmacological hypothesis, which additionally contains a plausible and traceable mechanistic chain, distinguishes itself from prior or competing explanations, can be investigated through feasible experiments, and could alter a defined research decision. These classes should not be treated as a single continuum of writing quality. A passage can be highly fluent yet fail at the proposition level, while a concise and stylistically plain statement may constitute a strong testable hypothesis. The distinction also prevents expert preference for polished language from being mistaken for evidence of scientific validity.

The proposed theory therefore treats a hypothesis as a structured package rather than as a sentence. Its elements include a claim, a context, a mechanism, an evidence ledger, competing explanations, a predicted observation, an explicit falsifier, and an experimental route capable of distinguishing among alternatives. Each element has a different interpretation boundary. Citations can ground an entity or relation but cannot establish an entire causal chain merely through accumulation. Mechanistic terminology can organize a claim but does not prove causal understanding. Novel wording can indicate linguistic variation without constituting scientific novelty. An executable assay can produce data without being capable of discriminating the proposed mechanism from a simpler explanation. Finally, expert agreement that a hypothesis is "plausible" may reflect conformity with current knowledge rather than the hypothesis's capacity to generate informative tests. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop what distinguishes a hypothesis from fluent scientific text within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for What distinguishes a hypothesis from fluent scientific text

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Linguistic fluency	Is the output coherent, grammatical, and appropriately expressed?	Language-model generation and scientific discourse conventions	Establishes whether the proposition can be read and interpreted	Fluent wording may conceal unsupported inference, contradiction, or absence of a testable claim	Fluency is necessary for communication but insufficient for hypothesis quality
Hypothesis structure	Does the output identify a perturbation, context, mechanism, prediction, and possible falsifier?	Scientific reasoning and experimental logic	Converts prose into an inspectable hypothesis object	A vague association may be presented as a directional hypothesis	Structural completeness does not establish truth
Directionality	Does the hypothesis specify what is expected to cause or modify what?	Pharmacological intervention logic and causal ordering	Distinguishes a mechanism-bearing claim from co-occurrence	Correlation may be rewritten as causal influence	Directionality is a claim requiring evidence, not a stylistic feature
Biological and pharmacological context	In which system, state, exposure range, and temporal setting should the claim hold?	Context dependence of targets, pathways, compounds, and phenotypes	Defines the domain in which testing and interpretation are meaningful	Context-free claims may imply unjustified generality	A hypothesis may be valid only under narrowly specified conditions
Mechanistic chain	Are intermediate links between perturbation and outcome explicitly represented?	Molecular pharmacology, pathway biology, systems pharmacology, and causal reasoning	Makes assumptions and missing transitions visible	Familiar mechanistic terms may be connected without evidence	Internal coherence does not alone prove the mechanism
Evidence provenance	Which evidence supports, contradicts, or fails to address each component?	Literature, molecular, biological, and experimental records	Enables claim-level audit and source tracing	Citation quantity may be mistaken for support quality	Evidence must be matched to the precise claim and context
Falsifier	What observation would reduce confidence in the hypothesis?	Falsifiability and discriminating experimental design	Prevents confirmation-only narratives	The hypothesis may be insulated from negative	A falsifier must be operationally interpretable, not merely stated

				results through post hoc explanation	
Competing explanations	Which alternative mechanisms could produce the same observation?	Differential reasoning and causal identification	Determines whether an experiment can distinguish the proposed mechanism	A positive result may be attributed exclusively to one mechanism	Support for one prediction does not eliminate alternatives
Scientific novelty	Is the proposed relation mechanistically or experimentally distinct from prior work?	Prior-art assessment, literature mapping, and knowledge representation	Separates rediscovery from a new testable contribution	Paraphrased prior knowledge may be labelled novel	Textual difference is not scientific novelty
Experimental discriminability	Can a feasible experiment produce meaningfully different expectations under competing hypotheses?	Assay design, controls, perturbation logic, and measurement theory	Links falsifiability to an actionable test	An experiment may be feasible but non-informative	Experimental execution does not guarantee mechanistic discrimination
Decision consequence	Could possible results change a defined pharmaceutical decision?	Decision theory, portfolio reasoning, and translational context	Distinguishes interesting speculation from decision-relevant inquiry	Hypotheses may be ranked by novelty or ease rather than consequence	Scientific interest and pharmaceutical value are related but not equivalent
Uncertainty	Which claims, evidence links, model outputs, and experimental interpretations remain uncertain?	Predictive uncertainty, evidentiary conflict, and model limitations	Prevents apparent precision from replacing calibrated qualification	A single confidence score may hide heterogeneous uncertainty	Uncertainty must remain attached to the component from which it arises

Table note: This table is an original conceptual synthesis. It does not define validated universal thresholds, clinical criteria, regulatory standards, or autonomous decision rules.

Mechanistic Coherence, Falsifiability, Novelty, and Experimental Value

Mechanistic coherence concerns whether the proposed links between a pharmacological perturbation and an expected outcome can coexist within the relevant chemical, biological, and temporal context. It is not established simply by naming a target, pathway, and phenotype that each appear in the literature. Chemical and biological data may contain assay-specific labels, indirect endpoints, heterogeneous conditions, missing negative results, and correlations that do not support the mechanistic inference imposed on them [10]. A coherent hypothesis should therefore expose its intermediate links: target engagement, molecular or cellular consequence, pathway-level response, compensatory process, and observable phenotype. Each link should be classified according to whether it is directly supported, indirectly supported, contradicted, or presently ungrounded. This structure helps identify mechanistic gaps, but it does not turn coherence into causal proof. Several incorrect mechanisms can be internally coherent, and a true mechanism can appear incomplete because relevant biology has not yet been characterized. Mechanistic coherence should consequently be evaluated as compatibility and evidentiary traceability, not as a substitute for experimental confirmation.

Falsifiability adds a different requirement. A pharmacological hypothesis should expose itself to a result that would reduce confidence in its defining relation rather than merely list experiments likely to produce confirmatory observations. Reliable machine-learning practice in chemistry already requires careful data partitioning, baseline selection, transparent reporting, and reproducibility because seemingly strong results can arise from weaknesses in evaluation design [11]. The analogous requirement for hypothesis generation is to specify the conditions under which a proposed mechanism would fail. This may involve absence of target engagement despite adequate exposure, persistence of the phenotype after blocking the proposed intermediate, reproduction of the effect in a target-null system, or observation of the same outcome through a competing mechanism. Falsifiability is thus stronger than testability. Any claim can be paired with an experiment, but only some experiments generate outcomes that meaningfully discriminate the claim from alternatives. The system should state in advance which result would support the mechanism, which would challenge it, which would remain ambiguous, and which technical failures would prevent interpretation.

Novelty is likewise irreducible to textual dissimilarity or molecular unfamiliarity. Learned molecular representations behave differently across endpoints, datasets, and chemical domains, demonstrating that apparent generalization depends on how the problem and representation are constructed [12]. A comparable problem arises when foundation models generate hypotheses. A proposition may appear novel because it combines rarely co-occurring terms, proposes an uncommon molecule, or paraphrases a known mechanism in unfamiliar language. Scientific novelty instead requires comparison with prior mechanistic claims, tested interventions, known target–disease relations, failed programs, and established experimental designs. It may

arise from a new mechanistic link, a new context in which an existing mechanism should operate, a discriminating experiment that resolves an old contradiction, or a new consequence derived from existing evidence. Novelty must also be evaluated together with plausibility and testability. A highly unusual claim may be novel because it violates well-supported constraints, whereas a modest reformulation may be valuable because it identifies a decisive experiment overlooked by prior work.

Experimental value integrates but does not collapse the preceding dimensions. Explainable artificial intelligence can expose features or relations that influence a model's output, yet an explanation of model behavior should not automatically be treated as a biological mechanism [13]. In the same way, a hypothesis should not be judged valuable merely because it includes a mechanistic narrative or an interpretable rationale. Its experimental value depends on whether a feasible test can distinguish the proposed mechanism from credible alternatives, reduce uncertainty that matters to the project, and change a defined decision. Mechanistic coherence without falsifiability produces a persuasive story. Falsifiability without feasibility produces an inaccessible test. Novelty without grounding produces speculation. Feasibility without discriminability produces data that may not resolve the claim. The proposed theory therefore treats mechanistic coherence, falsifiability, novelty, and experimental value as interacting but non-substitutable dimensions. Their relationship motivates a hypothesis-quality model in which no aggregate score can compensate for the absence of a traceable mechanism, an explicit falsifier, or an interpretable experimental route.

A Hypothesis-Quality Model for Pharmaceutical Foundation Systems

A pharmaceutical foundation system should not assign hypothesis quality through a single score derived from fluency, predicted confidence, novelty, or benchmark performance. Molecular systems therefore require separate assessment of calibration, error ranking, distribution learning, and goal-directed generation because performance on one of these dimensions does not establish reliability on the others [14-16]. The same principle applies to hypothesis generation. A model may express high confidence in a poorly grounded mechanistic link, produce a novel proposition that cannot be tested, or generate a feasible experiment that does not distinguish the proposed mechanism from alternatives. The proposed model therefore represents quality as a Hypothesis Quality Vector comprising evidentiary grounding, mechanistic coherence, falsifiability, novelty, experimental discriminability, feasibility, uncertainty, and decision consequence. These dimensions are related but non-equivalent. Their separation is intended to preserve the scientific meaning of each judgment and to prevent strong performance in one dimension from concealing failure in another.

The model begins with a hypothesis object rather than unconstrained prose. This object contains a defined perturbation, molecular target or process, biological and pharmacological context, mechanistic chain, expected observation, competing explanations, and explicit falsifier. It is linked to an evidence ledger that records which components are directly supported, indirectly supported, contradicted, or unsupported. The model then applies non-compensatory gates. At minimum, the proposition must contain traceable evidence, an explicit mechanism, a result capable of challenging the claim, and a feasible path to discriminating experimentation. This design is necessary because generative optimization may produce outputs that satisfy a computational objective while remaining experimentally impractical; assessments of generated molecules, for example, show that optimization does not guarantee synthesizability [17]. In the proposed theory, a hypothesis that fails a minimum gate is revised, deferred, or rejected rather than rescued by a favorable aggregate score.

The quality vector is followed by two linked judgments: expected information gain and decision consequence. Expected information gain concerns how strongly an experiment could distinguish among plausible mechanisms or reduce uncertainty attached to a consequential claim. Decision consequence concerns whether the possible results could alter target prioritization, candidate selection, assay strategy, resource allocation, or termination of an unproductive line of investigation. These judgments should remain separate because an experiment can be scientifically informative without affecting a current pharmaceutical decision, while a high-stakes decision may depend on an experiment that produces only limited mechanistic learning. The system should therefore expose the assumptions used to rank experiments, identify which uncertainties are expected to change, and show how each possible result would affect the decision. **Figure 1** presents the pharmacological hypothesis quality crucible, showing how the article's key components and boundaries are connected within a hypothesis-quality model for pharmaceutical foundation systems.

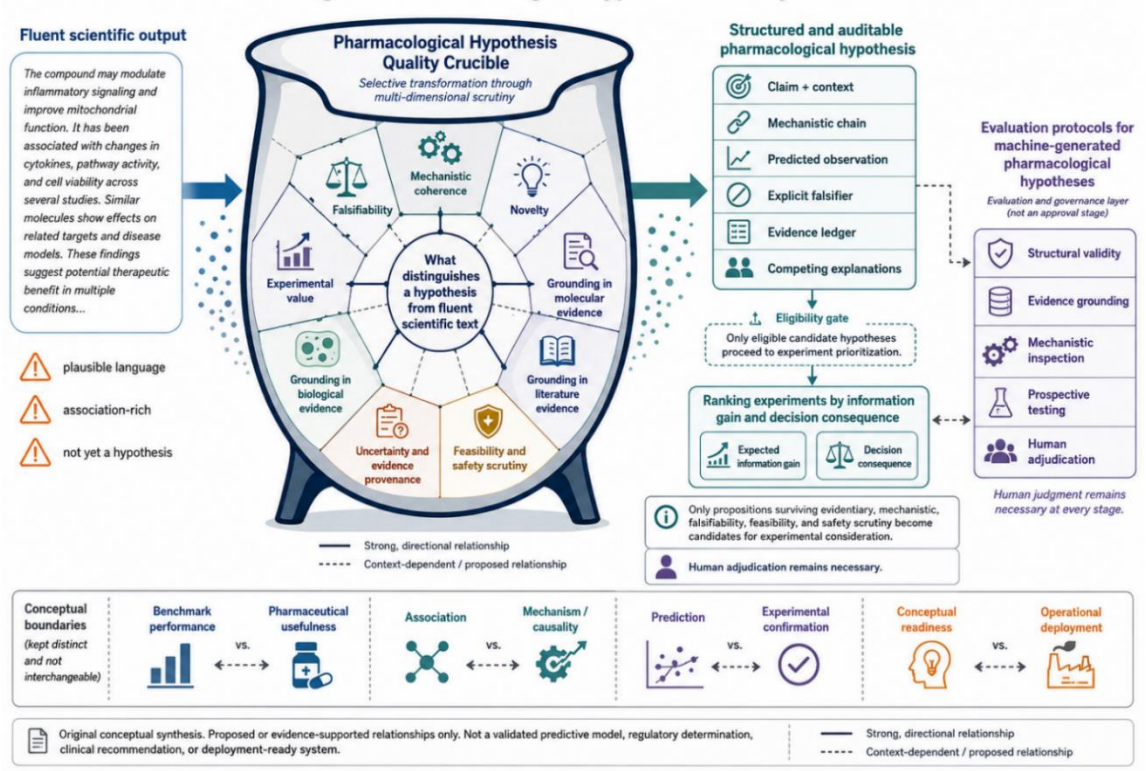

Figure 1. Pharmacological Hypothesis Quality Crucible.

The figure is an original conceptual synthesis that organizes the article's central contribution across distinguishes a hypothesis from fluent scientific text, mechanistic coherence, falsifiability, novelty, and experimental value, mechanistic coherence, experimental value, a hypothesis-quality model for pharmaceutical foundation systems, grounding in molecular, biological, and literature evidence. 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 crucible metaphor emphasizes selective transformation rather than automatic progression. Fluent output enters as unstructured material, but only propositions that survive evidentiary, mechanistic, falsifiability, feasibility, and safety scrutiny emerge as candidates for experimental consideration. Even then, the output is not labelled true, clinically useful, or development-ready. Its status is limited to a structured and auditable candidate hypothesis. Human adjudication remains necessary because the relative importance of mechanistic uncertainty, novelty, feasibility, and decision consequence depends on disease context, development stage, available assays, ethical constraints, and organizational priorities. The proposed model is therefore not an autonomous ranking engine. It is a representation and governance framework intended to make the reasoning behind hypothesis selection visible, contestable, and revisable.

Grounding in Molecular, Biological, and Literature Evidence

Molecular grounding concerns whether the entities and interactions in a hypothesis are compatible with available structural, physicochemical, and pharmacological knowledge. Protein-structure prediction can provide valuable constraints on binding-site geometry, domain organization, accessibility, and possible interaction patterns, but structural prediction alone does not establish target engagement, functional modulation, causal mechanism, exposure, or therapeutic effect [18]. A foundation model should therefore distinguish structural plausibility from pharmacological confirmation. It should also record whether a proposed ligand, target state, binding mode, or molecular consequence is observed, computationally predicted, inferred by analogy, or presently unsupported. This distinction is important because a structurally plausible interaction can fail through conformational dynamics, competition, cellular localization, metabolism, protein abundance, compensatory signaling, or insufficient free exposure.

Literature grounding requires more than attaching citations to generated text. Biomedical language models can improve entity recognition, relation extraction, question answering, and other text-mining functions that help identify relevant evidence [19]. Yet a retrieved paper may support only one component of a generated claim, and its findings may apply to a different compound, disease state, model system, dose, or biological context. The proposed evidence ledger therefore operates at claim level. Each mechanistic edge should link to the precise passage or result that supports it, together with information about experimental system, intervention, comparator, outcome, and limitation. Contradictory and null findings should remain visible rather than being omitted to preserve narrative coherence. Grounding quality should consequently be assessed through citation entailment, contextual correspondence, source provenance, and coverage of counterevidence, not citation count.

Generative biomedical language models can produce domain-appropriate text and assist with knowledge mining, but generation remains distinct from experimental validation [20]. This distinction becomes especially important when a system synthesizes evidence across papers. A model may combine individually accurate statements into a novel but invalid causal sequence, conflate an observed biomarker with a therapeutic mechanism, or infer clinical relevance from preclinical findings without sufficient translational support. Literature evidence should therefore be assigned explicit roles: entity confirmation, association support, mechanistic support, boundary evidence, contradiction, or prior-art evidence. The model should also disclose when support is indirect, such as when evidence comes from a related target, species, tissue, or compound class. Such role labelling reduces the risk that a long reference list will be interpreted as cumulative proof of a claim that no individual source directly establishes.

Biological grounding requires representation of relationships that extend beyond isolated molecules. Graph-based models can encode drug, protein, pathway, and adverse-effect relations, allowing hypotheses to be situated within broader pharmacological networks [21]. However, an edge in a biological graph may represent co-occurrence, statistical association, curated interaction, predicted relation, or experimentally supported mechanism. These meanings should not be treated as equivalent. The proposed model therefore requires typed edges, evidence provenance, directionality, and context. It should also represent alternative pathways, feedback loops, compensatory mechanisms, and conflicting observations. Biological graphs can organize relational evidence and expose missing links, but they do not by themselves establish causality. Their principal role in hypothesis generation is to constrain claims, identify contradictions, and support comparison among mechanistic alternatives.

Ranking Experiments by Information Gain and Decision Consequence

Foundation models may generate candidate molecules or mechanistic propositions by navigating learned representations of chemical space. Continuous molecular representations demonstrate how computational systems can search for candidates that optimize specified objectives [22]. For pharmacological hypothesis generation, however, the objective should not be limited to predicted potency, novelty, or similarity. A candidate experiment should be valued according to what can be learned from its possible outcomes. This requires explicit competing hypotheses and an account of how observations would update confidence in each alternative. An experiment that is likely to produce a positive result may have low information value when all plausible mechanisms predict the same outcome. Conversely, a negative or ambiguous result can be valuable when it excludes a central mechanistic link, reveals a hidden dependency, or prevents further investment in an unsupported direction.

Bayesian optimization provides a useful conceptual basis for sequential experiment selection because it formalizes how observations can update a model and influence the next choice [23]. Within the proposed framework, expected information gain should be estimated over mechanistic uncertainty rather than only over an objective function. A candidate experiment might be prioritized because it distinguishes target-mediated activity from a phenotypic off-target effect, separates pathway dependence from general cytotoxicity, or determines whether a proposed combination acts through interaction rather than independent activity. The ranking should account for assay sensitivity, anticipated noise, model-system relevance, cost, time, material availability, and the possibility of uninterpretable outcomes. Because prior assumptions can dominate information calculations, the system should disclose the alternatives considered, the expected observations under each alternative, and the uncertainty attached to those expectations.

Prospective chemical optimization illustrates how iterative observations can improve subsequent experimental choices [24]. Nevertheless, pharmacological hypothesis testing differs from optimizing reaction conditions or a single measurable objective. Biological outcomes may be delayed, heterogeneous, context-dependent, or jointly produced by several mechanisms. The most informative experiment is therefore not necessarily the experiment predicted to maximize an endpoint. It may instead be a perturbation designed to break a proposed causal chain, a rescue experiment that tests the specificity of an effect, a temporal measurement that distinguishes primary from secondary responses, or a model-system comparison that reveals context dependence. The proposed ranking process should reward experiments capable of revising the mechanism, including experiments expected to produce informative negative findings.

Active learning can help prioritize candidates from a large screening pool when experimental resources are constrained [25]. In hypothesis generation, the pool should include not only compounds but also assays, perturbations, model systems, readouts, and mechanistic alternatives. Selection should then be conditioned on decision consequence. Before an experiment is recommended, the system should specify which pharmaceutical decision it could change and how each possible result would alter that decision. A result may justify advancing a mechanism, redesigning a compound, changing an assay, collecting additional evidence, pausing a program, or rejecting the hypothesis. This formulation prevents information gain from becoming an abstract mathematical objective detached from research priorities. It also prevents organizational urgency from replacing scientific discriminability. A decision-worthy experiment must be both informative and relevant to an explicitly defined action.

Evaluation Protocols for Machine-Generated Pharmacological Hypotheses

Evaluation should begin before laboratory testing by determining whether the generated object is structurally valid, evidence-grounded, mechanistically inspectable, and capable of being challenged. Guidance for machine learning in the chemical sciences emphasizes appropriate controls, data partitions, baselines, leakage checks, transparent reporting, and external testing [26]. A hypothesis-generation protocol should extend these principles to the level of scientific claims. Systems should be tested on evidence that was unavailable during model training or withheld through a documented temporal cutoff. Evaluators should compare unaided foundation models, retrieval-augmented systems, tool-using agents, domain specialists, and hybrid human–

AI teams under matched information and time conditions. Assessment should include claim parsing, citation entailment, mechanistic-edge review, identification of missing and contradictory evidence, quality of falsifiers, novelty against prior art, experimental feasibility, and calibration of uncertainty.

Benchmark design must also address hidden substitution. Ligand-based classification studies have shown that common benchmarks can reward memorization rather than transferable generalization [27]. An analogous risk arises when hypothesis-generation systems reproduce familiar claims, retrieve near-duplicate hypotheses, or exploit templates associated with well-studied targets. Evaluations should therefore include paraphrase-resistant prior-art checks, temporally held-out literature, unfamiliar target–disease combinations, explicit negative evidence, and cases in which a plausible narrative conflicts with experimental constraints. Reviewers should be blinded to whether a hypothesis was produced by a human or a model when assessing structural quality and scientific plausibility. Separate evaluators should then inspect provenance, because a hypothesis that appears reasonable may still rely on unsupported or misassigned evidence.

Aggregate performance can conceal failures in difficult regions of chemical and biological space. Activity-cliff analyses demonstrate that models may perform adequately overall while failing where small structural changes produce large differences in activity [28]. Hypothesis protocols should therefore include challenge sets constructed around mechanistic discontinuities, context shifts, contradictory studies, uncertain target engagement, failed prior programs, and competing causal explanations. Prospective validation should assess whether selected hypotheses lead to discriminating experiments, whether predicted uncertainty corresponds to observed error or disagreement, and whether experimental results change decisions in the manner anticipated. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop evaluation protocols for machine-generated pharmacological hypotheses within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Can Foundation Models Generate Pharmacological Hypotheses That Are Mechanistically Coherent, Falsifiable, and Worth Testing?

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Hypothesis parsing	Generated text, entities, relations, conditions, predictions	Convert prose into a structured hypothesis object	Intervention, context, mechanism, predicted observation, falsifier	Ambiguous language and unresolved entities	Expert-annotated, multi-domain hypothesis corpora
Evidence retrieval and provenance	Literature, databases, molecular records, assay reports	Link each claim component to supporting, contradicting, or missing evidence	Claim-level evidence ledger	Retrieval omissions, outdated evidence, publication bias	Blinded citation-entailment and provenance audits
Mechanistic assessment	Typed molecular and biological relations	Test compatibility, directionality, and completeness of mechanistic chains	Supported, indirect, contradicted, and missing links	Coherent but incorrect mechanisms may pass	Independent expert review and prospective perturbation studies
Falsifiability assessment	Proposed mechanism, alternatives, assays, controls	Determine whether an outcome could count against the claim	Explicit counterprediction and interpretable failure condition	Technical failure may resemble mechanistic refutation	Protocol-level adjudication before experimentation
Novelty assessment	Time-stamped prior literature, knowledge graphs, failed programs	Distinguish scientific novelty from textual difference	Mechanistic, contextual, or experimental novelty profile	Incomplete prior-art coverage	Temporal holdout and independent novelty review
Experimental discriminability	Alternative hypotheses, expected observations, assay characteristics	Identify experiments that separate competing explanations	Ranked discriminating experiments	Assumptions about noise and biological response may be wrong	Comparison of predicted and realized discrimination
Feasibility and safety gate	Synthesis, exposure, assay, toxicity, material, and operational constraints	Exclude hypotheses that cannot be responsibly tested	Testable, revisable, deferred, or rejected status	Feasibility estimates remain context-specific	Medicinal-chemistry, pharmacology, and safety review
Uncertainty representation	Model uncertainty, evidence conflict, domain shift, assay variability	Retain uncertainty at claim, edge, and experiment levels	Component-specific uncertainty profile	Calibration may not transfer across domains	Temporal, external, and prospective calibration studies
Information-gain ranking	Competing mechanisms, prior beliefs, candidate experiments	Prioritize tests expected to reduce consequential uncertainty	Experiment portfolio ranked by learning value	Rankings depend on priors and model assumptions	Prospective comparison with realized belief updating

Decision-consequence analysis	Project objectives, alternatives, costs, time, safety, development stage	Connect possible experimental results to defined actions	Advance, revise, pause, stop, or gather evidence pathways	Utilities vary across projects and organizations	Embedded prospective decision studies
Human adjudication	Complete hypothesis package and audit trail	Review contested evidence, mechanisms, feasibility, and safety	Test, revise, defer, or reject decision	Inter-reviewer disagreement and institutional bias	Recorded multidisciplinary review and disagreement analysis
Prospective evaluation	Unseen evidence, independent laboratories, preregistered protocols	Test real-world generalization and usefulness	Observed experimental and decision consequences	A successful case cannot establish broad capability	Multisite, domain-specific, leakage-resistant validation
Reporting and governance	Model version, prompts, tools, sources, exclusions, revisions	Make generation and evaluation reproducible and auditable	Complete hypothesis-generation record	Proprietary systems may limit transparency	Versioned reporting standards and external audit

Table note: This table is an original theoretical synthesis for research evaluation. It is not a validated benchmark, laboratory authorization process, clinical decision instrument, regulatory qualification pathway, or operational deployment standard.

Limitations and Boundary Conditions

The proposed framework is limited first by the evidence from which it is constructed. Most available studies evaluate molecular prediction, generative design, information retrieval, autonomous chemistry, or isolated prospective demonstrations rather than repeated generation of pharmacological hypotheses across independent domains. A framework assembled from these neighboring evidence types can define distinctions and evaluation requirements, but it cannot establish how often foundation models will generate high-quality hypotheses or whether such systems will improve pharmaceutical productivity. Even direct laboratory testing of a machine-generated proposition provides only context-specific evidence and should not be generalized into a claim of broad autonomous scientific competence [9]. The framework therefore remains a proposed theory of representation and evaluation.

A second limitation is that mechanistic truth is not fully observable at the time hypotheses are generated. Molecular and biological evidence can be incomplete, contradictory, or dependent on assay conditions, species, tissue, exposure, disease state, and temporal scale [10]. Expert reviewers may also disagree about whether a mechanistic link is sufficiently grounded or whether an experiment genuinely falsifies the claim. The Hypothesis Quality Vector does not eliminate these uncertainties; it makes them explicit. Its components should not be converted into universal thresholds without empirical study, and non-compensatory gates will require domain-specific interpretation. What counts as sufficient evidence for an exploratory cellular hypothesis may be inadequate for a translational or clinically oriented claim.

A third limitation concerns misuse. A structured and well-grounded hypothesis can still be unsafe, ethically inappropriate, or strategically irrelevant. Molecular plausibility does not establish synthesizability, developability, tolerability, therapeutic exposure, clinical benefit, or regulatory acceptability. Likewise, ranking an experiment highly does not authorize its execution. The framework should not be used to generate treatment recommendations, determine patient eligibility, bypass biosafety review, or automate portfolio decisions. Its proposed value is epistemic and organizational: it can help expose assumptions, evidence gaps, falsifiers, and decision dependencies. Human scientific, ethical, safety, and governance oversight remains indispensable at every stage.

Research Agenda and Implementation Priorities

The first research priority is to create domain-specific datasets in which pharmacological hypotheses are represented as structured objects rather than undifferentiated passages. Such datasets should contain expert annotations of perturbations, contexts, mechanistic edges, predicted observations, falsifiers, alternatives, evidence provenance, and uncertainty. They should include strong hypotheses, superficially plausible but unsupported claims, known failed mechanisms, contradictory evidence, and hypotheses that are scientifically valid but experimentally inaccessible. Evaluation should use temporal and source-level separation to reduce memorization and evidence leakage, consistent with established concerns about misleading chemical benchmarks [27]. Inter-annotator disagreement should be retained as data because uncertainty about mechanism and interpretation is itself part of the evaluation problem.

The second priority is prospective comparison of generation and adjudication strategies. Studies should compare unaided language models, retrieval-augmented models, chemistry- and biology-tool systems, domain experts, and hybrid teams under matched conditions. Hypotheses should be registered before experimental results are known, together with their mechanism, expected observations, falsifiers, uncertainty, and ranked experiments. Independent evaluators should assess whether the proposed experiments distinguish among alternatives and whether observed findings update beliefs and decisions as predicted. Sequential experimental-design methods offer useful foundations for this work, but realized information gain and decision

consequences must be measured rather than assumed [24]. Multisite replication will be necessary because a system that performs well in one target class or disease area may fail under different evidence structures and experimental constraints.
The third priority is staged implementation with explicit stopping rules. Early use should be limited to evidence organization, hypothesis structuring, contradiction detection, and generation of candidate falsifiers under expert supervision. A subsequent stage could examine experiment-ranking support in low-risk preclinical settings, provided that the model's assumptions and uncertainty remain inspectable. Broader integration into project governance should occur only after prospective evidence shows reproducibility, calibration, and added value beyond expert or conventional computational workflows. Reporting standards should document model and retrieval versions, prompts, tools, excluded evidence, human revisions, competing hypotheses, and reasons for accepting or rejecting recommendations. Implementation should advance according to validated capability, not according to model scale, fluency, or commercial availability.

Conclusion

Foundation models can produce pharmacological propositions that are fluent, plausible, and scientifically styled, but these features do not establish that the propositions are mechanistically coherent, falsifiable, novel, or worth testing. This article proposes that machine-generated pharmacological hypotheses should be represented as auditable objects containing explicit contexts, mechanistic chains, evidence provenance, competing explanations, predicted observations, falsifiers, feasible experiments, uncertainty, and decision consequences. The central contribution is the Pharmacological Hypothesis Quality Crucible, which treats quality as a multidimensional profile governed by non-compensatory evidentiary and experimental conditions rather than by a universal score. The framework distinguishes benchmark performance from pharmaceutical usefulness, explanation from mechanism, association from causality, confidence from calibrated uncertainty, and test execution from experimental discrimination. Its purpose is not to claim that foundation models already possess general hypothesis-generation competence, but to define what such competence would require and how it could be evaluated. Progress will depend on prospective, leakage-resistant, domain-specific studies in which hypotheses are assessed not only for plausibility, but for their capacity to survive criticism, generate informative experiments, and change decisions without exceeding scientific, ethical, or operational boundaries.

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 G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97-113. doi:10.1038/nrd.2017.232
3. 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
4. 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
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. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620(7972):47-60. doi:10.1038/s41586-023-06221-2
7. 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
8. 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
9. Abdel-Rehim A, Zenil H, Orhobor OI, Fisher M, Collins RJ, Bourne E, et al. Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. J R Soc Interface. 2025;22(227):20240674. doi:10.1098/rsif.2024.0674
10. 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

11. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
12. 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
13. 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
14. 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
15. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
16. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839
17. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23. doi:10.1021/acs.jcim.0c00174
18. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. doi:10.1038/s41586-021-03819-2
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi:10.1093/bioinformatics/btz682
20. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac409
21. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13). doi:10.1093/bioinformatics/bty294
22. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76. doi:10.1021/acscentsci.7b00572
23. Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. Phoenics: A Bayesian optimizer for chemistry. ACS Cent Sci. 2018;4(9):1134-45. doi:10.1021/acscentsci.8b00307
24. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Martinez Alvarado JI, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature. 2021;590(7844):89-96. doi:10.1038/s41586-021-03213-y
25. Graff DE, Shakhnovich EI, Coley CW. Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem Sci. 2021;12(22):7866-81. doi:10.1039/D0SC06805E
26. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
27. 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