**Pharmacophore**

ISSN-2229-5402

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

# AGENTIC DRUG DISCOVERY REQUIRES A SCIENTIFIC CONSTITUTION FOR AUTONOMOUS PLANNING, TOOL SELECTION, EXPERIMENTATION, AND HUMAN INTERVENTION

**Michael O'Donnell[1]*, Emma Byrne[2], Nora O'Connor[3], Aisling Kelly[4]**

1. ***Department of Agentic Drug Discovery and Scientific Constitution, Faculty of Pharmacy, University of Limerick, Limerick, Ireland.***
2. ***Department of Autonomous Planning and Tool Selection, Faculty of Pharmaceutical Sciences, University College Cork, Cork, Ireland.***
3. ***Department of Experimentation and Human Intervention, School of Pharmacy, University College Dublin, Dublin, Ireland.***
4. ***Department of Agentic AI Governance in Drug Discovery, Faculty of Pharmacy, University of Galway, Galway, Ireland.***

**ARTICLE INFO**

**Received:**
21 June 2025
**Received** in revised form:
14 October 2025
**Accepted:**
15 October 2025
**Available online:**
28 October 2025

***Keywords:*** Agentic artificial intelligence, Autonomous drug discovery, Scientific governance, Tool-using agents, Self-driving laboratories, Human oversight

**ABSTRACT**

Artificial intelligence in drug discovery is moving from bounded prediction toward systems that can formulate plans, select computational and laboratory tools, revise strategies, and initiate experiments. This transition changes the scientific governance problem because an agent may influence not only what is predicted but also which evidence is gathered, which hypotheses are pursued, and which physical actions are undertaken. Conventional evaluation based on predictive accuracy, task completion, or benchmark performance cannot determine whether such agency is scientifically justified, appropriately bounded, reproducible, or accountable. This article develops a proposed scientific constitution for agentic drug discovery. The constitution is conceived as a structured allocation of scientific authority, evidence requirements, permissions, prohibitions, challenge rights, escalation conditions, and human intervention powers. It distinguishes scientific roles from technical capabilities, tool availability from tool suitability, model confidence from calibrated uncertainty, molecular plausibility from experimental confirmation, and successful task execution from pharmaceutical usefulness. The proposed construct connects four governance layers: role-bound authority, evidence warrants, rules for planning and action, and accountability through logging, challenge, override, and explicit reauthorization. It further argues that autonomous experimentation should be governed through consequence-sensitive permissions rather than a binary distinction between autonomous and non-autonomous systems. The contribution is theoretical and has not been empirically validated, standardized, or demonstrated to ensure safe deployment. Its applicability will depend on the task, laboratory environment, evidence maturity, chemical hazards, institutional responsibilities, and effectiveness of human oversight. The scientific constitution is therefore offered as an organizing framework for designing and evaluating bounded agency, not as a regulatory instrument or authorization for operational autonomy.



**To Cite This Article:** O'Donnell M, Byrne E, O'Connor N, Kelly A. Agentic Drug Discovery Requires a Scientific Constitution for Autonomous Planning, Tool Selection, Experimentation, and Human Intervention. Pharmacophore. 2025;16(5):99-108. https://doi.org/10.51847/aoLcKzpo1A

## Introduction

Artificial intelligence has become relevant across molecular representation, property prediction, virtual screening, synthesis planning, experimental prioritization, and development-stage analysis. Yet the significance of an artificial-intelligence output cannot be inferred from predictive performance alone: pharmaceutical impact depends on the suitability of the underlying chemical and biological data, the decision context in which the model is used, and the evidentiary demands of the particular discovery or development stage [1-3]. These distinctions become more consequential when a system is no longer restricted to producing a score or recommendation but can determine what information to retrieve, which model or database to invoke, which experiment to prioritize, and how to revise its own plan.

**Corresponding Author:** Michael O'Donnell; Department of Agentic Drug Discovery and Scientific Constitution, Faculty of Pharmacy, University of Limerick, Limerick, Ireland. E-mail: michael.odonnell@ul.ie.

The growing range of pharmaceutical artificial-intelligence applications should not be interpreted as evidence that all tasks, models, or use contexts have reached the same level of scientific maturity [4]. A system may perform well in retrospective molecular prediction while remaining unreliable for prospective candidate selection, mechanistic interpretation, synthesis decisions, or experimental escalation. It may also be trained on data that are accessible but unsuitable for the intended inference. Consequently, the governance of agentic systems must address not only whether an output is accurate, but whether the system was entitled to pursue the objective, whether its evidence was admissible, whether its tools were appropriate, and whether the resulting action remained within a scientifically defensible boundary.

Drug design is not reducible to the optimization of a single computational objective. Molecular proposals remain hypotheses that must be interpreted through chemical knowledge, target biology, assay context, exposure, safety, manufacturability, and therapeutic relevance. Artificial intelligence may extend the scale or speed of this reasoning, but it does not remove the need for integrated human and computational scientific judgment [5]. When an agent can select objectives, combine tools, generate intermediate claims, and influence experiments, the central question therefore changes from "How well does the model predict?" to "Under what authority, evidence, constraints, and accountability may the system act?"

Existing discussions of responsible artificial intelligence, autonomous laboratories, reproducibility, and experimental optimization provide important but fragmented elements of an answer. They do not by themselves define a unified scientific governance construct for agentic drug discovery. This article proposes a scientific constitution: a structured and revisable allocation of roles, permissions, evidence warrants, planning duties, action boundaries, escalation triggers, challenge rights, and human intervention powers. The contribution is intended to organize the conditions under which scientific agency may be delegated and evaluated. It does not assume that autonomy is inherently desirable, that human review is invariably sufficient, or that constitutional compliance alone establishes mechanistic truth, pharmaceutical value, safety, or deployment readiness.

*Why Agentic Autonomy Changes the Governance Problem in Drug Discovery*

Agentic autonomy changes the governance problem because the system may become an active participant in the production of evidence rather than a passive generator of predictions. Demonstrated chemical agents can integrate planning, information retrieval, code execution, specialist tools, and experimental interfaces, while autonomous-laboratory research increasingly raises governance questions extending beyond conventional software validation [6-8]. In such systems, an error may alter not only an output but also the sequence of evidence acquisition: an unsuitable database can influence a hypothesis, the hypothesis can determine a tool call, the tool call can shape an experiment, and the experiment can generate data that reinforce the original mistake. Governance must therefore consider trajectories of action and evidence, not isolated model outputs.

The physical reach of an agent further changes the consequence structure. A mobile robotic chemist can navigate a laboratory, conduct experiments, evaluate outcomes, and select subsequent actions within a specified environment [9]. This capability does not establish general scientific reasoning or unrestricted laboratory competence, but it demonstrates that computational choices can be translated into material operations. A false prediction can ordinarily be rejected without physical consequence; an agent-directed experiment may consume scarce material, damage equipment, expose personnel to hazards, generate uninterpretable results, or create misleading evidence that enters later decision processes. The relevant governance question is therefore not merely whether the agent can complete an operation, but whether the operation was scientifically authorized, appropriately controlled, and reversible.

Self-driving laboratories make this distinction especially visible because they connect experiment selection, execution, measurement, and learning within a closed loop [10]. Their value may arise from rapid iteration and systematic exploration, but the same closed-loop structure can amplify objective misspecification, measurement artifacts, instrument drift, inappropriate acquisition functions, or unrecognized domain shifts. If each new result automatically informs the next action, a local scientific error can become a persistent experimental trajectory. A constitution for such agency must consequently distinguish operational continuity from epistemic legitimacy: the ability to continue acting cannot itself justify continued authority.

The proposed theory identifies four properties that separate agentic governance from ordinary model governance. First, temporal extension means that the object of evaluation is a sequence of decisions rather than a single output. Second, tool-mediated reach means that an agent can inherit the capabilities and limitations of databases, software, instruments, and external services. Third, evidence reflexivity means that the agent may generate the data later used to evaluate or reinforce its own hypotheses. Fourth, consequence escalation means that actions can move from reversible computation to costly or hazardous experimentation. These properties require permissions, evidence warrants, logging, challenge, and human intervention to be designed as an integrated scientific system rather than appended as general ethical principles. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why agentic autonomy changes the governance problem in drug discovery within the article's central argument.

**Table 1.** Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why agentic autonomy changes the governance problem in drug discovery

| Construct or evidence domain | Core question | Scientific basis | Role in the article | Failure or overclaiming risk | Boundary statement |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **Predictive performance** | Does the system produce accurate outputs for a defined task and evaluation setting? | Task-specific comparison against appropriate data, baselines, and error criteria | Establishes a necessary but incomplete measure of technical competence | Treating retrospective performance as proof of prospective discovery value | Predictive performance does not alone establish mechanism, experimental validity, safety, or pharmaceutical usefulness |
| **Agentic planning** | Can the system formulate and revise a sequence of actions toward an approved objective? | Explicit objectives, assumptions, constraints, alternatives, uncertainty, and stopping rules | Makes the agent's intended trajectory inspectable before execution | Confusing a coherent plan with a scientifically correct plan | Planning competence does not create authority to execute every planned action |
| **Tool-mediated agency** | Which databases, models, software systems, instruments, or services may the agent use? | Tool provenance, version, validation state, applicability domain, input compatibility, and access controls | Connects abstract reasoning to computational and physical capabilities | Assuming an available or familiar tool is suitable for the scientific purpose | Tool access is conditional and does not establish tool fitness |
| **Closed-loop experimentation** | How are observations used to select subsequent experiments? | Defined objective function, measurement validity, acquisition rule, operating envelope, and update procedure | Explains how errors or useful evidence can propagate through autonomous cycles | Reinforcement of measurement artifacts, proxy objectives, or false hypotheses | Operational continuation is not evidence that the trajectory remains scientifically justified |
| **Physical action authority** | Which computational recommendations may be translated into laboratory operations? | Approved protocols, hazard assessment, controls, equipment status, and qualified supervision | Marks the transition from reversible analysis to consequential action | Equating technical executability with scientific or safety authorization | Physical experimentation requires a distinct, revocable authorization gate |
| **Evidence reflexivity** | Can the agent generate evidence later used to support its own hypothesis or continued authority? | Independent controls, provenance, predefined interpretation rules, and external challenge | Identifies circularity risks in autonomous evidence production | Self-confirming experimental trajectories and selective evidence generation | Agent-generated evidence requires independent interpretive and quality checks |
| **Consequence-sensitive autonomy** | How should permissions change with novelty, uncertainty, irreversibility, cost, or hazard? | Risk classification, uncertainty assessment, reversibility analysis, and escalation thresholds | Replaces a binary autonomous/non-autonomous distinction with graduated authority | Applying the same permissions to low-risk computation and high-risk experimentation | Greater technical capability does not automatically justify broader scientific authority |
| **Pharmaceutical relevance** | Does the action improve a meaningful discovery decision beyond benchmark completion? | Experimental confirmation, decision-context evaluation, and downstream scientific interpretation | Connects agent evaluation to drug-discovery usefulness | Presenting benchmark gain as evidence of therapeutic or translational value | Pharmaceutical usefulness remains conditional on evidence beyond computational task completion |

*Scientific Roles, Permissions, and Boundaries for Autonomous Agents*

A scientific constitution begins by separating role, capability, and authority. A capability describes what a system can technically perform; a role defines the function assigned to it within a scientific process; authority specifies which actions it is permitted to take and under what conditions. This distinction is necessary because safe self-driving laboratories require safeguards that correspond to both the level of autonomy and the hazards of the environment [11]. An agent capable of generating a synthesis route need not be authorized to order reagents, configure equipment, execute the route, or communicate the result as experimentally established. Authority should therefore be narrower than capability unless additional evidence and supervision justify expansion.

Autonomy should also be decomposed rather than treated as a binary property. The ADePT framework characterizes autonomous laboratory systems through dimensions including adaptability, dexterity, perception, and task complexity [12]. For governance purposes, such decomposition permits permissions to be attached to specific proficiencies. A system may be permitted to perceive instrument state but not modify safety parameters; adapt a computational search strategy but not change the approved biological objective; or execute a validated protocol while being prohibited from designing a materially different experiment. Capability-specific permissions make it possible to expand or restrict authority without inaccurately labeling an entire system as either autonomous or non-autonomous.

Scientific roles can be organized around functions such as objective owner, planner, evidence analyst, tool operator, experimental executor, independent reviewer, and accountable human principal. Human participation may vary across interactive and autonomous laboratory configurations, and the relevant distinction is not simply whether a person appears "in the loop" [13]. A human who receives hundreds of opaque actions after execution may have less meaningful control than a reviewer positioned at a small number of consequential authorization gates. Conversely, requiring manual confirmation for every low-consequence operation may create delay, automation bias, or superficial approvals. The constitution should therefore specify what the human must understand, what evidence must be presented, when intervention is possible, and which decisions cannot be delegated.

General ethical principles are insufficient unless translated into operational duties. Comparative analysis of artificial-intelligence guidelines has shown broad convergence around values such as transparency, fairness, responsibility, privacy, and non-maleficence, alongside substantial divergence in interpretation and implementation [14]. In an agentic drug-discovery system, these principles must become concrete provisions: who may define objectives, which data are admissible, which tools are permitted, what uncertainties must be disclosed, which actions require independent review, and who can suspend authority. Algorithmic systems redistribute decision-making responsibility but do not eliminate the responsibilities of the humans and institutions that design, authorize, supervise, and rely on them [15]. The proposed constitution therefore treats accountability as a traceable chain of scientific responsibility rather than an attribute that can be assigned solely to the autonomous agent.

*The Proposed Scientific Constitution: Authority, Evidence, and Accountability*

The proposed scientific constitution is a theory of conditional delegation. It specifies which scientific roles may be assigned to an autonomous agent, what actions accompany each role, what evidence must exist before those actions are permitted, and who remains accountable for their consequences. Responsible machine-learning practice already requires duties to be distributed across design, validation, implementation, and monitoring rather than concentrated at the point of model deployment [16]. Agentic drug discovery extends this lifecycle logic because the system may alter its plan, invoke additional tools, generate evidence, and initiate physical operations. The constitution therefore treats authority as temporary, task-specific, consequence-sensitive, and revocable rather than as a permanent property conferred by technical capability.

The first constitutional pillar is authority. An authority ledger should define what an agent may observe, recommend, select, execute, modify, communicate, and approve. The second pillar is evidence. Every consequential action should require an evidence warrant documenting the objective, relevant data, provenance, model or tool version, applicability limits, uncertainty, alternatives considered, and foreseeable consequences. Minimum-information initiatives demonstrate why scientific claims cannot be interpreted reliably when data handling, modeling decisions, evaluation procedures, and limitations are insufficiently reported [17]. Within the proposed constitution, an evidence warrant is not merely a retrospective reporting device; it is a prospective condition for delegated action.

The third pillar is accountability, operationalized through traceable responsibility and reproducible records. Reproducibility standards in life-science machine learning emphasize the importance of documented data, code, models, workflows, and computational environments [18]. Agentic systems require an expanded record that also preserves retrieved sources, prompts or instructions, intermediate plans, rejected alternatives, tool calls, parameter changes, experimental states, uncertainty declarations, human approvals, and the reasons for escalation or non-escalation. Accountability is therefore not satisfied by identifying the final user of an agent. It requires an attributable chain connecting the human or institutional objective owner to system designers, tool providers, experiment supervisors, reviewers, and final decision authorities.

A fourth constitutional element is the right to challenge authority and evidence. Data leakage and related methodological failures can produce apparently strong scientific conclusions that do not survive independent examination [19]. The constitution should permit a scientist, reviewer, safety mechanism, independent agent, or failed reproduction attempt to contest a plan, action, or claim without requiring proof that harm has already occurred. Regulatory authorization in another artificial-intelligence context has also shown that authorization status and transparent scientific characterization are not interchangeable [20]. Accordingly, constitutional compliance cannot be interpreted as regulatory acceptance or universal scientific validity.

**Figure 1** presents the agentic drug-discovery scientific constitution, showing how the article's key components and boundaries are connected within the proposed scientific constitution: authority, evidence, and accountability.


**Figure 1.** Agentic Drug-Discovery Scientific Constitution.

The figure is an original conceptual synthesis that organizes the article's central contribution across agentic autonomy changes the governance problem in drug discovery, scientific roles, permissions, and boundaries for autonomous agents, scientific roles, boundaries for autonomous agents, proposed scientific constitution: authority, evidence, and accountability, proposed scientific constitution. 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.

*Rules for Planning, Tool Selection, Experimentation, and Escalation*

Constitutional planning begins with an approved scientific objective and a bounded action space. Autonomous investigation of chemical reactivity demonstrates that machine-directed search can select experiments and update its strategy within a defined experimental domain [21]. The scientific validity of such activity depends on what has been fixed before the search begins: the objective, available actions, prohibited regions, measurement procedure, controls, stopping criteria, and interpretation rules. An agent should not be permitted to silently redefine a target property, replace a mechanistic question with an easier proxy, or expand an experiment beyond the approved chemical and safety envelope. Every material plan revision should therefore produce a new evidence warrant or trigger escalation.

Tool selection should be governed by scientific fitness rather than availability or computational convenience. For each database, predictive model, simulation package, search service, laboratory instrument, or robotic interface, the agent should verify identity, version, intended use, input compatibility, validation state, known limitations, and applicability to the current problem. Bayesian optimization in chemical synthesis illustrates that experiment selection depends on the objective, the available observations, the surrogate representation, and the acquisition strategy [22]. A tool can be technically functional while scientifically unsuitable. The proposed rule is therefore that no consequential tool call should inherit authority merely because it is accessible to the agent.

Experiment-selection permissions should also reflect cost, novelty, and information value. Pool-based active learning can prioritize molecular candidates under constrained screening resources, but the selected sequence remains conditional on the candidate pool, predictive model, acquisition policy, and available labels [23]. The constitution should require the agent to disclose whether a proposed experiment is intended to exploit an apparently favorable region, explore uncertainty, discriminate among mechanisms, test a failure hypothesis, or improve representational coverage. These purposes are not interchangeable. A strategy optimized only for immediate predicted gain may repeatedly sample a narrow chemical neighborhood while producing little evidence about generalization or mechanism.

Physical experimentation requires a distinct authorization gate. Closed-loop reaction optimization shows how experimental results can be used iteratively to select subsequent conditions [24]. Before execution, however, the agent should demonstrate that the protocol is within its permission envelope, hazards are represented, equipment and materials are compatible, controls are adequate, measurements can answer the stated question, and failure or interruption is manageable. Escalation should be

mandatory when novelty, uncertainty, irreversibility, cost, hazard, evidence conflict, protocol deviation, or repeated failure exceeds predefined limits. Escalation is not itself a scientific verdict; it transfers authority to a qualified human or independent review process capable of narrowing, revising, postponing, or rejecting the proposed action.

*Logging, Reproducibility, Challenge, and Human Override*

Agentic logging must preserve the decision process, not only the final output. Reproducibility problems in health-oriented machine learning show that insufficient methodological disclosure can obstruct independent evaluation even when a model or result is publicly described [25]. An agentic record should therefore include the initiating objective, authority state, source provenance, retrieved evidence, plan versions, tool configurations, parameter choices, generated code, intermediate outputs, uncertainty estimates, experimental conditions, observations, failed actions, human interactions, and final interpretation. Logs should be time-stamped, version-linked, tamper-evident, and interpretable by an independent reviewer.

Reproducibility should be assessed at several levels. Computational replayability asks whether analyses can be rerun with the same versions and inputs. Decision reproducibility asks whether another qualified investigator can reconstruct why the agent selected a particular action. Experimental reproducibility asks whether the protocol and context are sufficient for independent repetition. Persistent limitations in the reproducibility of machine-learning health research demonstrate that inaccessible data, code, implementation details, and environments can prevent meaningful reconstruction [26]. Exact numerical duplication may remain impossible for stochastic systems or physical experiments, but unexplained divergence should become visible rather than disappearing behind an aggregate task-success claim.

The proposed constitution grants a formal challenge right. Biological machine-learning validation recommendations organize scrutiny across data, optimization, model, and evaluation domains [27]. Agentic drug-discovery challenges should additionally examine objective legitimacy, authority, tool suitability, experiment controls, uncertainty handling, safety, provenance, and interpretation. A challenge may request additional evidence, a counter-analysis, an alternative tool, an independent reproduction, or suspension of the action. The challenged agent should not be the sole judge of whether the objection is valid. Independent adjudication is especially important when the disputed evidence was generated or selected by the same system whose authority is under review.

Human override should be technically effective, scientifically informed, and proportionate to consequence. Evaluations comparing artificial intelligence with clinicians have shown how design and reporting limitations can coexist with strong claims of superiority [28]. Analogously, nominal human supervision cannot compensate for opaque evidence, delayed warnings, excessive action volume, or interfaces that do not permit meaningful interruption. Override mechanisms should allow authorized humans to pause execution, revoke a tool, narrow the objective, require additional controls, roll back a reversible state, terminate a workflow, or retire a capability. Resumption should require explicit reauthorization supported by a documented root-cause assessment and revised permissions; authority should never be restored merely because the system has restarted.

*Evaluation Scenarios for Safe and Scientifically Valid Agency*

Evaluation must distinguish task completion from constitutionally valid scientific conduct. Comparative molecular studies show that conclusions about model superiority depend on the task, dataset, comparator, and evaluation design; benchmark structure may reward memorization rather than transferable prediction; and uncertainty quality must be evaluated separately from prediction error [29-31]. An agent can therefore complete a benchmark while violating its authority, selecting an unsuitable tool, concealing uncertainty, exploiting chemical similarity, or generating an experimentally meaningless result. Evaluation should examine both what the system achieved and whether the process by which it acted remained scientifically justified.

Scenario suites should test difficult cases rather than only nominal workflows. Relevant scenarios include a confident prediction beyond the applicability domain, contradictory database records, an obsolete but accessible tool, a protocol lacking controls, optimization toward an inadequate proxy, repeated negative experiments, a hazardous indirect tool call, a request to communicate an unconfirmed result as established, and a human instruction that conflicts with a safety boundary. Molecular benchmark suites demonstrate the value of using heterogeneous tasks, datasets, splits, and performance measures [32]. For agentic systems, this heterogeneity should extend to roles, action consequences, evidence conflicts, intervention timing, and the reversibility of failures.

No single aggregate score should determine acceptance. Evaluation should separately assess constitutional fidelity, goal integrity, evidence admissibility, tool suitability, planning quality, experimental validity, uncertainty calibration, reproducibility, challenge responsiveness, override effectiveness, generalization, pharmaceutical relevance, and accountability. A system may be strong in one dimension and unacceptable in another. For example, a technically successful experiment may remain scientifically invalid because its controls cannot distinguish the proposed mechanism, while a cautious refusal may represent appropriate constitutional behavior rather than task failure. Metrics and expert assessments must therefore be interpreted in relation to the permitted role and the consequence of error.

Evaluation should proceed through staged environments: static case analysis, simulated tool use, sandboxed computational action, instrument emulation, bounded laboratory execution, independent reproduction, and only then prospective use in a real discovery workflow. Advancement should require evidence appropriate to the next level of consequence, and permissions should be narrowed when performance is unstable or challenge mechanisms fail. **Table 2** organizes the evidence, constructs,

and interpretation boundaries needed to develop evaluation scenarios for safe and scientifically valid agency within the article's central argument. **Figure 2** presents the evidence, validation, and translation boundary observatory for agentic drug discovery requires a scientific constitution, showing how the article's key components and boundaries are connected within evaluation scenarios for safe and scientifically valid agency.

**Table 2.** Evaluation, implementation, and research priorities arising from Agentic Drug Discovery Requires a Scientific Constitution for Autonomous Planning, Tool Selection

| Component or layer | Inputs | Core function | Expected output | Uncertainty or limitation | Validation requirement |
|---|---|---|---|---|---|
| **Authority and role evaluation** | Role definition, permission ledger, prohibited actions, consequence class | Determine whether the agent acts only within explicitly delegated authority | Traceable record of permitted, denied, and escalated actions | Role definitions may omit emergent capabilities or indirect actions | Prospective tests using hidden permission conflicts and attempted scope expansion |
| **Evidence-warrant evaluation** | Data provenance, model validity, applicability domain, uncertainty, alternatives | Determine whether the evidence required for an action is present and relevant | Action-specific evidence record with stated limitations | Completeness does not establish correctness or causal validity | Independent scientific review and counter-analysis |
| **Planning evaluation** | Objective, assumptions, constraints, alternatives, stopping rules | Assess whether plans are bounded, interpretable, and aligned with the approved purpose | Versioned plan with explicit decision rationale | Detailed plans may still pursue an invalid hypothesis | Comparison with expert plans and controlled goal-drift scenarios |
| **Tool-admissibility evaluation** | Tool identity, version, validation state, inputs, dependencies | Verify that selected tools are fit for the assigned scientific function | Approved tool call, substitution, or escalation | Tool behavior may change across versions or contexts | Cross-tool comparison, version-change testing, and misleading-tool challenges |
| **Experimental authorization** | Protocol, hazards, controls, instrument state, reversibility | Prevent unsupported movement from prediction to physical action | Execute, revise, postpone, or reject decision | Hazard models and protocol checks cannot capture all laboratory contingencies | Staged simulation, instrument emulation, and bounded laboratory assessment |
| **Logging and reproducibility** | Complete action trace, data, code, versions, prompts, experimental metadata | Enable reconstruction, replay, attribution, and audit | Independent replay package and decision history | Stochastic and physical processes may prevent exact replication | Multi-site computational and experimental reproduction |
| **Challenge and override** | Objection source, disputed evidence, action state, intervention authority | Suspend, contest, revise, or terminate agent activity | Adjudicated challenge and documented corrective action | Human reviewers may be delayed, biased, or overloaded | Blind scientific challenges and human-factors studies |
| **Generalization and uncertainty** | External tasks, temporal or scaffold shifts, novelty, calibration data | Determine whether competence and uncertainty remain reliable outside familiar conditions | Failure profile and bounded applicability statement | No finite scenario set establishes universal robustness | Prospective out-of-distribution evaluation |
| **Pharmaceutical decision relevance** | Experimental confirmation, downstream evidence, decision consequence | Distinguish benchmark completion from useful discovery support | Bounded statement of decision value | Downstream value may be delayed, context-dependent, or confounded | Prospective comparison within real discovery decisions |
| **Reauthorization and retirement** | Root-cause analysis, corrected controls, revised permissions, repeat tests | Decide whether suspended capability should be restored, narrowed, or retired | Explicit new authority state | Corrective changes may address symptoms rather than underlying causes | Repeated adversarial evaluation before restoration |

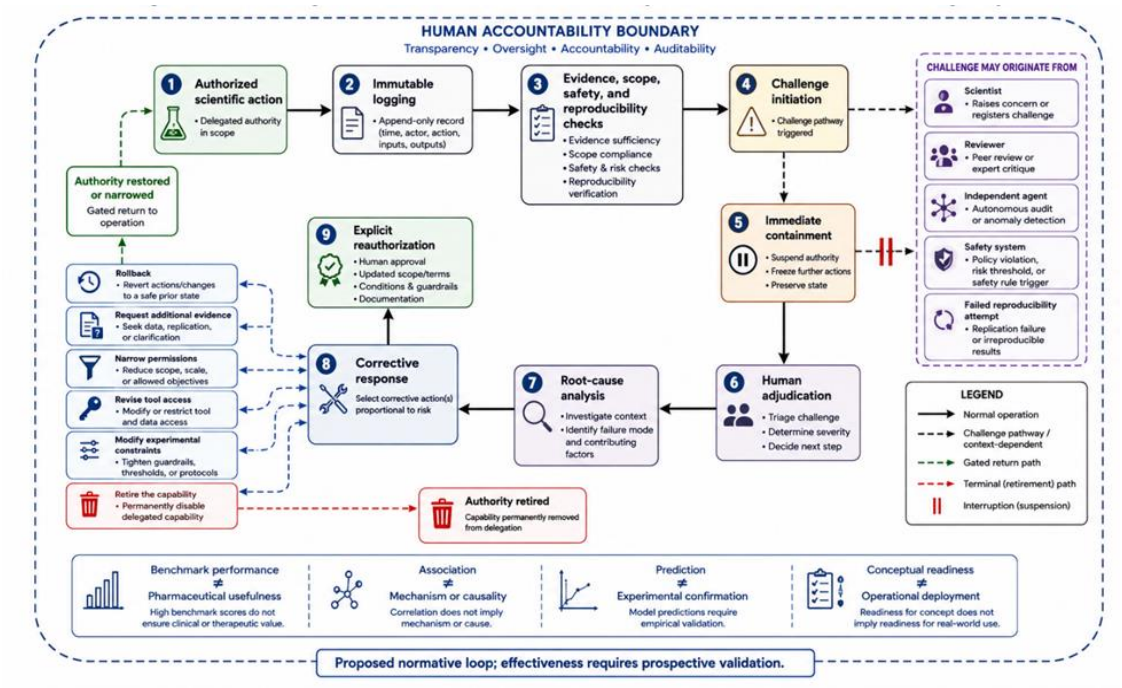


**Figure 2.** Evidence, Validation, and Translation Boundaries.

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.

*Limitations and Boundary Conditions*

The proposed scientific constitution is a conceptual governance theory rather than an empirically validated control architecture. The selected literature supports its components—autonomous planning, tool use, laboratory action, reproducibility, uncertainty assessment, challenge, and oversight—but does not establish that their integration will reliably prevent scientific error or physical harm. The theory also does not specify universal quantitative thresholds for evidence sufficiency, uncertainty, intervention, or acceptable residual risk. These thresholds will vary across computational screening, synthesis planning, biological experimentation, preclinical interpretation, and other pharmaceutical contexts.

The evidence base is heterogeneous. Some supporting studies concern chemistry agents or autonomous laboratories, whereas others address machine-learning validation, clinical artificial intelligence, reporting, or reproducibility. Their transfer to agentic drug discovery is conceptually informative but conditional. Laboratory hazards, biological complexity, intellectual-property restrictions, proprietary platforms, data-access limitations, institutional structures, and jurisdictional requirements may alter how constitutional provisions can be implemented. Human oversight may also fail through automation bias, limited expertise, excessive workload, delayed intervention, or organizational pressure.

Constitutional compliance cannot establish that a molecular hypothesis is causal, that a generated structure is synthesizable or developable, that an experiment is externally reproducible, or that a candidate has therapeutic value. Nor does the framework determine legal liability, regulatory acceptability, cybersecurity sufficiency, biosafety compliance, or ethical permissibility in every context. It may be misused as a checklist that legitimizes premature autonomy. Its provisions must instead remain contestable, task-specific, auditable, and subordinate to applicable scientific, institutional, safety, and legal requirements.

*Research Agenda and Implementation Priorities*

The first research priority is to formalize the constitution in a machine-readable ontology of roles, permissions, evidence warrants, prohibited actions, escalation triggers, challenge rights, and authority states. This work should determine whether different laboratories and software platforms can represent the same constitutional rule without losing its scientific meaning. Evaluation datasets should include ambiguous authority, conflicting evidence, unsuitable tools, indirect action chains, objective drift, absent controls, and attempted circumvention. Red-team studies should examine whether agents can exploit gaps between formally permitted actions to produce a prohibited outcome.

The second priority is prospective evaluation of the relationships among uncertainty, consequence, and intervention. Studies should test whether calibrated uncertainty can meaningfully restrict action, whether human reviewers recognize when escalation is necessary, and whether override remains effective under time pressure and high action volume. Multi-laboratory reproduction should assess portability across instruments, data systems, protocols, and organizational settings. Failure reporting should document near misses, abandoned plans, rejected tool calls, unsuccessful reproductions, and human corrections rather than only completed tasks.

Implementation should proceed incrementally. Initial use should emphasize read-only evidence organization and bounded computational assistance, followed by sandboxed tool use, supervised experiment recommendation, instrument emulation, and tightly constrained physical execution. Each increase in authority should require independent evidence that logging, challenge, and override remain effective at the new consequence level. Institutions will also need interoperable provenance infrastructure, versioned tool registries, qualified oversight roles, incident-review procedures, and explicit reauthorization processes. These priorities are research and governance requirements, not evidence that broad agentic deployment is currently justified.

**Conclusion**

Agentic drug discovery changes the scientific governance problem because artificial intelligence may determine not only what is predicted but also which evidence is sought, which tools are invoked, which experiments are conducted, and how subsequent decisions are shaped. This article proposes a scientific constitution that links role-bound authority, evidence warrants, planning and action rules, reproducible records, challenge rights, human override, and explicit reauthorization. Its central claim is that safe and scientifically valid agency cannot be established through a single performance measure or successful task demonstration. Authority must remain conditional on evidence, context, consequence, and continuing accountability. The constitution is an original theoretical synthesis rather than a validated standard, regulatory instrument, or deployment authorization. Its value will depend on whether future work can translate its provisions into testable rules, effective intervention mechanisms, interoperable records, and prospective evidence of bounded pharmaceutical usefulness.

**Acknowledgments:** None

**Conflict of interest:** None

**Financial support:** None

**Ethics statement:** None

**References**

1. 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
2. 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 used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
3. 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
4. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010
5. 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
6. 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
7. 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
8. Tobias AV, Wahab A. Autonomous 'self-driving' laboratories: A review of technology and policy implications. R Soc Open Sci. 2025;12(7):250646. doi:10.1098/rsos.250646
9. Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature. 2020;583(7815):237-41. doi:10.1038/s41586-020-2442-2
10. Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282-91. doi:10.1016/j.trechm.2019.02.007
11. Leong SX, Griesbach CE, Zhang R, Darvish K, Zhao Y, Mandal A, et al. Steering towards safe self-driving laboratories. Nat Rev Chem. 2025;9(10):707-22. doi:10.1038/s41570-025-00747-x
12. Salazar-Villacis P, Benyahia B. The ADePT framework for assessing autonomous laboratory robotics. Commun Chem. 2026;9(1):99. doi:10.1038/s42004-026-01932-9
13. Hysmith H, Foadian E, Padhy SP, Kalinin SV, Moore RG, Ovchinnikova OS, et al. The future of self-driving laboratories: From human in the loop interactive AI to gamification. Digit Discov. 2024;3(4):621-36. doi:10.1039/D4DD00040D
14. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-99. doi:10.1038/s42256-019-0088-2

15. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586
16. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
17. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
18. 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
19. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
20. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digit Med. 2020;3:118. doi:10.1038/s41746-020-00324-0
21. Granda JM, Donina L, Dragone V, Long DL, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature. 2018;559(7714):377-81. doi:10.1038/s41586-018-0307-8
22. 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
23. 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
24. Angello NH, Rathore V, Beker W, Wołos A, Jira ER, Roszak R, et al. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. Science. 2022;378(6618):399-405. doi:10.1126/science.adc8743
25. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA. 2020;323(4):305-6. doi:10.1001/jama.2019.20866
26. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med. 2021;13(586). doi:10.1126/scitranslmed.abb1655
27. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4
28. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368. doi:10.1136/bmj.m689
29. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
30. 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
31. 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
32. 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