**Pharmacophore**

ISSN-2229-5402

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

# SELF-DRIVING PHARMACEUTICAL LABORATORIES NEED EXPLICIT RULES FOR EXPLORATION, CONFIRMATION, FAILURE RECOVERY, ESCALATION, AND STOPPING

**Erik Johansson[1]*, Sofia Lundberg[2], Anders Nilsson[3]**

1. ***Department of Self-Driving Labs and Exploration Rules, Faculty of Pharmacy, KTH Royal Institute of Technology, Stockholm, Sweden.***
2. ***Department of Confirmation and Failure Recovery, Faculty of Pharmaceutical Sciences, Lund University, Lund, Sweden.***
3. ***Department of Escalation and Stopping Rules in Autonomous Labs, Faculty of Pharmacy, Uppsala University, Uppsala, Sweden.***

**ARTICLE INFO**

**Received:**
06 August 2025
**Received** in revised form:
27 November 2025
**Accepted:**
01 December 2025
**Available online:**
28 December 2025

***Keywords:*** Self-driving laboratories, Autonomous experimentation, Pharmaceutical artificial intelligence, Scientific governance, Failure recovery, Human escalation

**ABSTRACT**

Self-driving laboratories are increasingly positioned as closed-loop systems that can select experiments, control instruments, interpret measurements, and update subsequent actions with limited direct intervention. In pharmaceutical science, however, technical autonomy does not itself provide the epistemic governance needed to distinguish an informative experiment from a confirmatory test, a repeat measurement from an independent replication, or a recoverable malfunction from a condition requiring human escalation or campaign termination. Without explicit transition and stopping rules, an autonomous platform may efficiently optimize an invalid objective, reinforce bias, repeat contaminated procedures, overinterpret uncertain measurements, or continue experimentation after the scientific value of additional observations has become negligible. This article develops the Pharmaceutical Autonomous Experimentation Governance State Machine as an original conceptual architecture for organizing autonomous laboratory behavior. The proposed construct separates qualified readiness, exploration, confirmation, replication, failure recovery, human escalation, and stopping into distinct governance states linked by evidence-dependent transition guards. It further treats provenance, auditability, uncertainty assessment, and contamination monitoring as continuous control functions rather than retrospective reporting obligations. The central argument is that autonomous laboratory quality cannot be established through a single optimization score, throughput measure, model-confidence estimate, or experimental success indicator. Evaluation must instead examine whether the system enters the correct scientific state, preserves the distinction between provisional and confirmed evidence, detects failure, escalates unresolved uncertainty, and stops under success, futility, risk, contamination, or resource constraints. The architecture is not presented as an empirically validated controller, regulatory framework, or deployment-ready standard. Its value lies in specifying testable governance relationships and boundary conditions for future pharmaceutical laboratory research.



**To Cite This Article:** Johansson E, Lundberg S, Nilsson A. Self-Driving Pharmaceutical Laboratories Need Explicit Rules for Exploration, Confirmation, Failure Recovery, Escalation, and Stopping. Pharmacophore. 2025;16(6):77-88. https://doi.org/10.51847/LZiC884gRP

## Introduction

Artificial intelligence and machine learning now influence multiple stages of pharmaceutical research, including molecular-property prediction, virtual screening, synthesis planning, experimental design, formulation development, biological assay interpretation, and development decision support. Their scientific value is nevertheless conditional on the suitability of the underlying data, the validity of the selected task, the relationship between computational outputs and experimental evidence, and the extent to which the resulting decision can be reconstructed and challenged. Pharmaceutical discovery is not a single prediction problem: it is a sequence of interdependent scientific judgments involving chemical feasibility, biological relevance, measurement validity, uncertainty, reproducibility, developability, and safety. A system that performs well on one computational objective may therefore remain unsuitable for determining what experiment should be performed, whether an observed result should be believed, or when further experimentation should stop [1].

**Corresponding Author:** Erik Johansson; Department of Self-Driving Labs and Exploration Rules, Faculty of Pharmacy, KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: erik.johansson@kth.se

The transition from isolated prediction models to AI-supported drug design intensifies this problem because generated or prioritized candidates must move through chemical reasoning, synthesis, analytical characterization, biological evaluation, and multidisciplinary interpretation. Computational fluency cannot substitute for evidence that a compound is synthesizable, correctly identified, experimentally active, mechanistically interpretable, reproducible, or relevant to a therapeutic objective. AI-supported pharmaceutical design consequently requires an integrated relationship among machine inference, medicinal chemistry, experimental science, and expert judgment rather than a workflow in which a high model score is treated as a sufficient decision rule [2]. This integration becomes more difficult when models are allowed not only to advise researchers but also to select actions, modify experimental conditions, invoke software tools, or control instruments.

Self-driving laboratories extend closed-loop computation into the physical scientific process. In their general form, they combine experiment planning, robotic or automated execution, analytical measurement, data processing, model updating, and selection of subsequent experiments. Such systems can potentially improve the consistency and information efficiency of experimental campaigns, particularly where experiments are expensive, multidimensional, or difficult to prioritize manually. Yet the closed-loop structure also creates a recursive vulnerability: an error in the objective, measurement process, model, sample identity, or data pipeline can influence the next experiment and then be reinforced through subsequent iterations. The same architecture that accelerates valid learning can accelerate systematic error when scientific-state distinctions are left implicit [3].

Existing laboratory orchestration systems demonstrate that algorithms, instruments, measurements, and data services can be coordinated through explicit software-mediated workflows [4]. What remains less developed is a governance layer that determines the scientific meaning of each action and restricts how autonomous systems move from one evidentiary purpose to another. An optimizer may know how to choose the next condition without knowing whether the campaign is still exploratory. A robot may repeat an assay without establishing whether the repetition constitutes technical repeatability or independent replication. A control system may recover from an instrument fault without determining whether the affected data must be quarantined. This article addresses that gap by proposing the Pharmaceutical Autonomous Experimentation Governance State Machine, a conceptual architecture that separates exploration, confirmation, replication, failure recovery, escalation, and stopping and connects them through explicit transition guards, provenance controls, and bounded human authority.

*Why Automation without Stopping Rules Can Amplify Poor Science*

Autonomous experimentation is often evaluated through measures such as throughput, search efficiency, precision, uptime, objective improvement, experimental cost, or degree of human involvement. These measures describe important aspects of system performance, but none independently establishes whether the resulting science is valid. A platform may execute many experiments precisely while pursuing a poorly specified objective. It may identify a numerical optimum that depends on an assay artefact, an unstable reagent, a biased data representation, or a narrow region of experimental space. It may also increase apparent autonomy by reducing human intervention even when intervention would have been scientifically necessary. Multidimensional performance assessment is therefore essential because autonomy, speed, reproducibility, search quality, and scientific reliability are related but non-equivalent properties [5]. The absence of stopping rules is particularly consequential: without explicit termination criteria, the system can continue generating observations after success has been adequately demonstrated, after the hypothesis has become uninformative, after repeated failures indicate futility, or after integrity concerns invalidate further learning.

The danger is magnified when retrospective performance is treated as evidence of prospective pharmaceutical value. A model may appear highly accurate because the evaluation data resemble the training data, because chemical series are distributed across both sets, or because the benchmark contains shortcuts that permit recognition rather than transferable inference. Chemical and biological datasets may also combine inconsistent assay conditions, variable endpoints, publication biases, incomplete negative results, uncertain structures, and unrecorded experimental context. Under these conditions, autonomous action does not remove the underlying weakness; it operationalizes it. A self-driving platform may repeatedly select experiments that reinforce a biased representation of the search space or optimize a signal that is only locally valid. Claims of autonomy can therefore be inflated when retrospective benchmark performance is treated as evidence of prospective pharmaceutical value, particularly where data limitations and train–test similarity permit memorization rather than transferable generalization [6-8].

Bias can also enter through benchmark construction, decoy selection, assay composition, preprocessing, and the definition of positive or negative examples. In structure-based virtual screening, for example, models may exploit differences between ligand and decoy sets that are unrelated to the intended molecular-recognition problem. Comparable problems can occur in experimental automation when plate layout, sample order, instrument drift, solvent system, analytical batch, or reagent source becomes correlated with the target outcome. If the autonomous system is rewarded for improving a measured objective, it may exploit any stable signal available to it, including signals that lack scientific relevance. Bias control must therefore operate before, during, and after autonomous experimentation rather than being treated as a one-time property of a training dataset [9]. A result should not progress toward confirmation merely because the optimizer has converged or the observed response has increased; progression requires evidence that the measurement is valid, the signal is not an artefact, the candidate lies within a defensible applicability region, and the relevant controls have behaved as expected.

The central governance problem is thus not simply that an autonomous laboratory might make an incorrect prediction. It is that automation can couple incorrect assumptions to repeated physical action. A poorly specified objective can drive systematic exploration; a contaminated measurement can update the model; an updated model can prioritize additional contaminated conditions; and a superficially improving trajectory can delay recognition that the campaign is scientifically compromised. Explicit stopping rules interrupt this recursive amplification. They should include successful stopping when prespecified objectives and evidence requirements have been met; futility stopping when additional experimentation is unlikely to resolve the question; integrity stopping when sample, data, protocol, or instrument provenance is compromised; safety stopping when physical or informational risk exceeds delegated authority; and resource stopping when continued experimentation is no longer proportionate to its expected information value. These rules do not guarantee scientific truth, but they prevent optimization logic from being treated as the sole authority over whether experimentation should continue. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why automation without stopping rules can amplify poor science within the article's central argument.

**Table 1.** Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why automation without stopping rules can amplify poor science

| Construct or evidence domain | Core question | Scientific basis | Role in the article | Failure or overclaiming risk | Boundary statement |
|---|---|---|---|---|---|
| **Objective validity** | Is the optimized objective scientifically aligned with the pharmaceutical question? | Autonomous systems act on formalized targets, which may represent only part of the intended scientific value. | Establishes that efficient optimization can still pursue an invalid or incomplete objective. | A convenient surrogate is treated as equivalent to efficacy, mechanism, developability, or therapeutic value. | Objective improvement alone does not establish pharmaceutical usefulness. |
| **Data suitability** | Are the data sufficiently relevant, representative, and contextually documented for the decision? | Chemical and biological datasets may contain assay heterogeneity, missing context, selection bias, and uncertain labels. | Supports entry guards before data are used to direct physical experimentation. | Data availability is mistaken for decision suitability. | A large dataset may remain unsuitable for a specific autonomous decision. |
| **Benchmark generalization** | Does reported performance survive realistic chemical, biological, or experimental separation? | Similarity-rich partitions may reward memorization or narrow interpolation. | Separates retrospective model scores from prospective experimental usefulness. | Benchmark performance is reported as evidence of transferable discovery capacity. | Generalization must be evaluated under task-relevant distribution shifts. |
| **Measurement validity** | Is the observed signal attributable to the intended phenomenon rather than an artefact? | Instrument, assay, batch, environmental, and preprocessing effects can create stable but misleading signals. | Requires controls and orthogonal evidence before state progression. | An optimizer exploits an assay or analytical artefact. | A reproducible measurement can still be scientifically invalid. |
| **Recursive error amplification** | Can an error influence subsequent experiment selection and become self-reinforcing? | Closed-loop updating couples prior observations to future actions. | Explains why autonomous error differs from an isolated incorrect prediction. | Contaminated or biased observations reshape the experimental trajectory. | Faster learning is beneficial only when the information entering the loop is valid. |
| **Multidimensional performance** | Which dimensions of autonomy and scientific quality are being assessed? | Throughput, precision, search efficiency, reliability, reproducibility, and intervention burden are distinct. | Rejects reliance on a single SDL performance indicator. | A high score on one dimension is used to claim overall scientific quality. | No single performance measure establishes valid autonomous science. |
| **Successful stopping** | Has the prespecified scientific purpose been adequately fulfilled? | Continuing after sufficient evidence can waste resources and increase opportunities for contradictory or contaminated data. | Defines stopping as a legitimate positive governance outcome. | The system continues optimizing after the scientific question has been answered. | Stopping after success does not imply universal validity beyond the tested context. |

| | | | | | |
|---|---|---|---|---|---|
| **Futility stopping** | Is further experimentation likely to provide decision-relevant information? | Repeated non-convergence or uninformative measurements may indicate diminishing value. | Prevents indefinite experimentation driven by algorithmic persistence. | Resource consumption is mistaken for scientific thoroughness. | Futility depends on the stated question, uncertainty, and available alternatives. |
| **Integrity stopping** | Has contamination, leakage, provenance loss, or protocol deviation compromised the evidence chain? | Scientific claims require traceable materials, data, code, models, and decisions. | Connects stopping to reproducibility and contamination governance. | Compromised observations continue updating the model. | Halting protects the evidence base but does not identify the root cause by itself. |
| **Safety and authority stopping** | Does the proposed action exceed physical, chemical, computational, or delegated authority limits? | Autonomous platforms may encounter unanticipated hazards, novel actions, or permission conflicts. | Establishes a boundary between autonomous execution and human accountability. | Technical capability is mistaken for authorized scientific action. | A system must stop or escalate when authority, risk, or scope limits are exceeded. |

*Exploration, Confirmation, Replication, and Failure Recovery*

Exploration and confirmation are often combined within a single optimization loop, but they serve different scientific purposes. Exploration seeks information: it searches uncertain space, compares alternatives, updates beliefs, and may tolerate adaptive objectives or acquisition policies within defined bounds. Bayesian optimization and robotic experimentation provide practical mechanisms for selecting experiments when evaluations are costly and the response surface is not known in advance. They can balance exploitation of promising conditions against exploration of uncertain regions and can update subsequent actions from measured outcomes. Bayesian optimization and closed-loop robotic exploration show how experiment selection can balance information acquisition and objective improvement, but they do not by themselves define confirmatory, replication, recovery, or stopping obligations [10-12]. An exploratory result remains provisional because the candidate, criterion, model, or experimental conditions may have been selected adaptively from the same observations used to judge success.

Confirmation begins only when the claim under examination has been sufficiently specified to resist post hoc reinterpretation. The candidate or hypothesis should be frozen, the primary acceptance criteria defined before the confirmatory run, relevant positive and negative controls specified, and the analytical method qualified for the intended inference. Confirmation may also require an orthogonal measurement that relies on a materially different detection principle. Failure during this stage cannot automatically be interpreted as disproof of the underlying hypothesis because the failure may arise from instrumentation, materials, execution, analysis, or model assumptions. Automated multistep synthesis illustrates this distinction: a platform may optimize process variables successfully yet still encounter catalyst deactivation or route-specific complications that the optimizer does not represent, requiring human diagnosis and corrective intervention [13]. The confirmatory state must therefore preserve the scientific claim while allowing the experiment to be questioned.

Replication addresses a further question: whether a confirmed result persists when relevant sources of variation are deliberately reintroduced. Repeating the same measurement from the same prepared sample can estimate technical repeatability, but it does not provide the same evidence as using a new batch, a separately prepared sample, another instrument, an independent analytical method, a different operator, or another laboratory. The replication design should reflect the anticipated use of the result. A reaction condition intended for transfer requires evidence across batches and equipment states; a biological finding may require different plates, cell preparations, or assay days; a formulation result may require independently manufactured lots and stability-relevant conditions. Machine-learning-enhanced high-throughput experimentation can improve the integration of design, execution, and analytical processing, but the evidentiary meaning of a repeated experiment still depends on how independence and variation are constructed [14]. Autonomous repetition must not be labelled replication merely because the software has executed the same protocol more than once.

Failure recovery is a separate governance state because its purpose is neither discovery nor confirmation. It begins when a control fails, an instrument becomes unreliable, an observation is physically implausible, modalities disagree, model behaviour changes unexpectedly, contamination is suspected, or repeated experiments do not converge. Entry into recovery should freeze affected outputs, quarantine relevant materials and data, preserve raw records, and identify the earliest point at which the evidence chain may have been compromised. Corrective actions should be bounded and diagnostic: recalibrating an instrument, replacing a reagent, rerunning a control, restoring a verified software version, testing an alternative analytical method, or repeating the last valid transition. Recovery must not silently redefine the objective, relax the acceptance criterion, discard inconvenient observations, or repeatedly rerun the experiment until a favourable result appears. Return to exploration, confirmation, or replication is permissible only after that state's entry conditions have been re-established; otherwise, the laboratory must escalate the unresolved condition or stop the campaign.

*Governance States and Transition Rules for Self-Driving Laboratories*

Technical autonomy should be distinguished from scientific authority. A laboratory may possess automated planning, robotic execution, measurement, and model-updating capabilities while remaining unable to determine whether a result has changed evidentiary status. Existing self-driving laboratory architectures show that autonomy depends on coordinated computational, analytical, and physical capabilities, but they do not by themselves provide a complete governance ontology for deciding when exploration must end, when confirmation may begin, or when unresolved failure requires escalation [15]. The proposed Pharmaceutical Autonomous Experimentation Governance State Machine therefore treats scientific purpose as an explicit system state rather than an informal description added after an experiment has been completed.

The first state is qualified readiness, in which instruments, materials, analytical procedures, software, models, permissions, and provenance services are checked before autonomous activity begins. This requirement follows from the fact that automated execution depends on procedures being translated into explicit and sufficiently unambiguous operations [16]. A machine-readable protocol should identify required inputs, permissible parameter ranges, safety limits, expected controls, analytical dependencies, and completion conditions. Where the protocol depends on tacit knowledge, unresolved natural-language ambiguity, undocumented instrument behaviour, or unavailable controls, the system should not infer that it is ready. It should instead request clarification, enter recovery, or escalate the unresolved precondition.

Following readiness, the system may enter exploration, confirmation, replication, or a bounded analytical state only through defined transition guards. Integrated synthesis platforms already separate planning, physical execution, in-line analysis, and intervention as distinguishable workflow stages [17]. PAEGS extends this separation by assigning each stage a scientific purpose and by prohibiting transitions that would change the meaning of evidence without satisfying new obligations. An exploratory optimum cannot become a confirmed finding solely because a model has converged. A confirmed result cannot become a replicated result through repeated measurements of the same sample. A recovered instrument cannot resume operation merely because it produces data again; the affected state must first be requalified.

Governance becomes more important as autonomous systems acquire broader tool-use capabilities. Language-model agents can decompose scientific goals, generate code, search information sources, and invoke experimental tools, thereby expanding both the useful action space and the opportunity for unsupported or unauthorized action [18]. Mobile robotic systems likewise demonstrate that orthogonal analytical measurements and predetermined pass-or-fail criteria can structure progression and replication within exploratory chemistry [19]. These examples support the feasibility of explicit state transitions, but they do not establish a universal governance standard. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop governance states and transition rules for self-driving laboratories within the article's central argument.

**Table 2.** Components, relationships, uncertainties, and validation needs within Governance states and transition rules for self-driving laboratories

| Component or layer | Inputs | Core function | Expected output | Uncertainty or limitation | Validation requirement |
|---|---|---|---|---|---|
| **Governance envelope** | Scientific objective, risk classification, delegated authority, resource limits, permitted tools | Defines the campaign's operational and scientific boundaries | Explicit permissions, prohibitions, objectives, and accountable roles | Objectives may be incomplete, conflicting, or poorly translated into machine-readable form | Expert review and prospective testing across representative pharmaceutical tasks |
| **Qualified-readiness state** | Instrument qualification, material identity, protocol version, model version, controls, contamination status | Determines whether the system may begin autonomous action | Readiness decision with traceable supporting evidence | Latent faults and undocumented dependencies may remain | Independent readiness audits and fault-injection testing |
| **Exploration state** | Search space, priors, acquisition policy, constraints, resource budget | Acquires information and proposes promising candidates or conditions | Provisional observations and ranked hypotheses | Adaptive selection can bias interpretation and exaggerate apparent certainty | Prospective assessment of search behaviour under realistic distribution shift |
| **Confirmation state** | Frozen claim, predefined criteria, controls, orthogonal method, qualified protocol | Tests whether an exploratory signal survives controlled evaluation | Confirmed, rejected, indeterminate, or escalated status | A valid protocol may still test an invalid scientific assumption | Comparison with blinded expert interpretation and independent testing |
| **Replication state** | Confirmed result, replication plan, independent materials or equipment, equivalence criteria | Tests persistence under prespecified sources of variation | Replicated, non-replicated, context-dependent, or indeterminate status | Apparent independence may be compromised by shared data, materials, or preprocessing | Multibatch, multi-instrument, and where feasible multisite evaluation |

| | | | | | |
|---|---|---|---|---|---|
| **Failure-recovery state** | Failure event, raw data, control results, instrument diagnostics, recent changes | Quarantines affected evidence and conducts bounded root-cause tests | Restored qualification, unresolved failure, escalation, or stopping | Root causes may be multifactorial or unobservable | Controlled evaluation using known and unknown failure scenarios |
| **Human-escalation state** | Trigger reason, uncertainty, failed guards, affected lineage, permissible options | Transfers decision authority to a qualified human | Authorized bounded action, additional evidence request, scope revision, or stop | Human judgments may be inconsistent, biased, or delayed | Human-factors studies and inter-rater evaluation of escalation decisions |
| **Stopping-and-containment state** | Success, futility, risk, contamination, repeated failure, resource exhaustion | Halts experiment selection and safely preserves materials, data, and equipment state | Terminal or suspended campaign with complete rationale | Premature and delayed stopping involve task-specific trade-offs | Prospective comparison of alternative stopping policies |
| **Transition-guard layer** | Acceptance criteria, uncertainty limits, provenance completeness, safety constraints, resource conditions | Prevents unauthorized or scientifically invalid state changes | Passed, failed, or escalated transition decision | Poorly selected thresholds can formalize weak science | Domain-specific calibration and external audit |
| **Audit-and-contamination control plane** | Event logs, sample lineage, code, data, model and instrument versions, access records | Maintains traceability and identifies affected evidence chains | Reconstructable campaign history and contamination alerts | Complete logging does not guarantee semantic correctness | Independent reconstruction and contamination-challenge exercises |

*Escalation Thresholds and Human Decision Points*

Escalation is not a residual category for events that automation cannot process. It is a planned governance transition that recognizes where scientific, ethical, safety, or interpretive authority should return to a qualified person. Escalation thresholds should be defined before the campaign and should identify the conditions under which the system must pause rather than improvise. Such conditions include actions outside the validated domain, proposals involving unfamiliar hazards, conflicting analytical modalities, failed controls, unresolved contamination, repeated recovery attempts, unstable uncertainty estimates, or requests to alter the scientific objective. In consequential settings, inspectable decision logic is preferable where feasible because post hoc explanations may not reveal why a transition rule acted as it did [20].

Model confidence should not be used as a substitute for calibrated uncertainty. A high softmax probability, narrow ensemble spread, or stable acquisition score may reflect the internal behaviour of the model without accurately representing the probability that the proposed experiment will be valid or informative. Uncertainty assessment should therefore distinguish uncertainty arising from noisy measurements, limited data, model structure, distribution shift, and disagreement among analytical sources. Machine-assisted decision research has emphasized that uncertainty quantification is necessary when automated recommendations influence consequential actions [21]. In a self-driving pharmaceutical laboratory, uncertainty should govern permissions: sufficiently low and validated uncertainty may permit bounded autonomous action, whereas poorly calibrated or structurally unresolved uncertainty should trigger confirmation, additional measurement, escalation, or stopping. An escalation packet should be designed for decision use rather than merely exposing raw system output. It should state which transition was attempted, which guard failed, what evidence is affected, how uncertainty was estimated, whether the event has occurred before, which samples or data descendants may be contaminated, and what actions remain permissible. Communicating uncertainty in a form that supports review is essential because an uninterpretable alert can create either blind acceptance or indiscriminate rejection [22]. **Figure 1** presents the autonomous laboratory governance state machine, showing how the article's key components and boundaries are connected within escalation thresholds and human decision points.

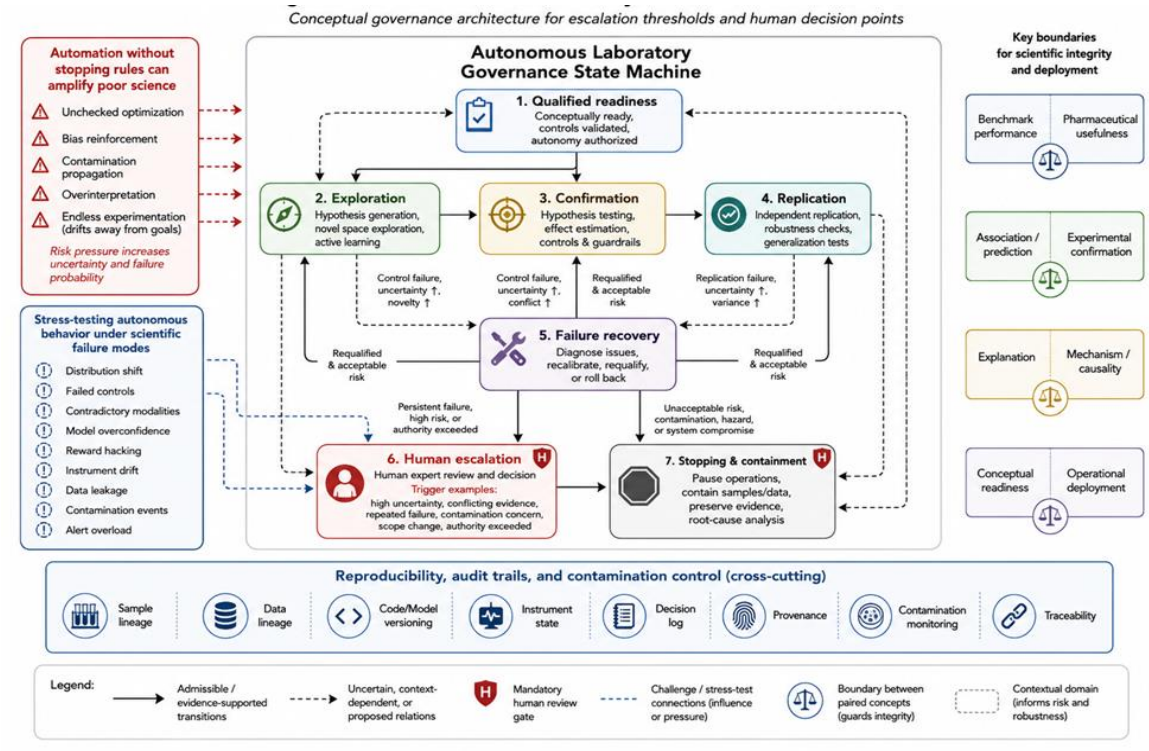


**Figure 1.** Autonomous Laboratory Governance State Machine.

The figure is an original conceptual synthesis that organizes the article's central contribution across automation without stopping rules can amplify poor science, exploration, confirmation, replication, and failure recovery, failure recovery, governance states and transition rules for self-driving laboratories, escalation thresholds and human decision points, reproducibility, audit trails, and contamination control. 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.

Human involvement remains necessary even when the laboratory's immediate decision rule is computationally explicit. Distribution shift can invalidate apparently general rules when assays, instruments, materials, populations, or laboratory conditions differ from those represented during model development [23]. Automation can also create unintended consequences, including overreliance on machine outputs, loss of tacit expertise, propagation of upstream errors, and diffusion of accountability [24]. Escalation should therefore be based not only on predicted risk but also on novelty, disagreement, provenance failure, and the consequences of an incorrect action. The human reviewer should have authority to authorize a bounded continuation, demand additional evidence, revise the campaign scope, quarantine affected outputs, or stop. Human review, however, cannot convert inadequate evidence into valid evidence; it can only determine the defensible next action.

*Reproducibility, Audit Trails, and Contamination Control*

A governed autonomous laboratory requires more than a chronological list of instrument commands. It needs an evidence-centred audit trail capable of reconstructing why each experiment was selected, which state the campaign occupied, which transition guard was applied, what data and materials were used, and how the resulting observation influenced subsequent action. A defensible autonomous laboratory requires versioned workflows, automated reproducibility checks, and explicit protection against information leakage because traceability without integrity control cannot secure a scientific claim [25-27]. The audit record should include raw and processed data, code and model versions, acquisition functions, protocol revisions, instrument configurations, calibration status, sample lineage, reagent lots, environmental conditions, human overrides, failed controls, and stopping decisions.

Reproducibility should be evaluated at multiple levels. Computational reproducibility asks whether the same code, model, data, and environment produce the same output. Analytical repeatability asks whether the same sample measured under nominally identical conditions yields a consistent result. Experimental reproducibility asks whether independently prepared materials and repeated procedures produce sufficiently concordant evidence. Scientific reproducibility asks whether the interpretation persists when relevant assumptions, methods, and contexts are challenged. Guidance for machine learning in biological research emphasizes that splitting strategies, data preprocessing, evaluation design, and interpretation must remain aligned with the scientific question [28]. These distinctions are essential because a perfectly reproducible computation may repeatedly generate an invalid prediction, whereas a scientifically robust conclusion may tolerate bounded experimental variability.

Contamination control must cover both physical and informational pathways. Physical contamination includes sample carryover, reagent degradation, plate effects, environmental exposure, instrument residue, sample swaps, and mislabeled materials. Informational contamination includes leakage between training and evaluation sets, use of confirmatory outcomes during model selection, reuse of supposedly independent controls, untracked preprocessing changes, and propagation of compromised observations into subsequent model updates. Reproducibility standards for life-science machine learning emphasize transparent reporting and accessible artifacts, but autonomous experimentation additionally requires lineage-aware containment because one contaminated observation can recursively influence later experiment selection [29]. The system should therefore identify every descendant decision affected by a compromised sample, dataset, model, or protocol version.

Auditability is not equivalent to validity. A laboratory may record every event while still pursuing an invalid objective, using a biased assay, or interpreting an association as a mechanism. Logs must therefore preserve scientific rationale as well as operational history. Each state transition should record the claim being tested, the evidence required, the uncertainty remaining, the rejected alternatives, and the authority under which the transition occurred. When contamination is detected, the system should freeze affected outputs, identify their dependencies, prevent further model updates from using them, and document whether the campaign can be reconstructed from the last qualified state. These controls allow independent scrutiny, but they do not establish that the underlying model, assay, or pharmaceutical interpretation is correct.

*Stress-Testing Autonomous Behavior under Scientific Failure Modes*

Stress testing should evaluate how an autonomous laboratory behaves when its assumptions fail, not merely how accurately its models perform under familiar benchmark conditions. Molecular benchmark suites can support standardized comparison across tasks, representations, and prediction endpoints, but their results do not establish that a laboratory will detect invalid measurements, recover from equipment failure, or stop after contamination [30]. Governance testing should therefore examine complete campaign behaviour: state recognition, guard enforcement, uncertainty response, evidence quarantine, escalation, and termination. A system that predicts molecular properties accurately but continues operating after failed controls should not be considered scientifically robust.

Generative molecular-design benchmarks further illustrate why objective-specific evaluation is necessary. Distribution-learning tasks and goal-directed optimization tasks examine different model behaviours, and success on one does not imply balanced performance on the other [31]. In an autonomous laboratory, goal-directed optimization creates opportunities for reward hacking, where the system exploits weaknesses in the objective, assay, or scoring model. Stress tests should introduce candidates that satisfy the numerical objective while violating synthesizability, chemical stability, structural plausibility, assay validity, or broader pharmaceutical constraints. The expected response is not necessarily rejection; it is recognition that the evidence is insufficient for automatic progression.

Uncertainty systems should be challenged under data scarcity, noisy measurements, out-of-domain chemistry, model disagreement, and abrupt instrument or assay drift. Comparative work in molecular property prediction shows that uncertainty-estimation methods differ in calibration and their ability to rank likely errors [32]. Stress testing should therefore avoid assuming that any uncertainty score is inherently meaningful. The system should be evaluated on whether uncertainty increases under known shift, whether high uncertainty changes the permitted action, and whether false reassurance leads to an inappropriate transition. A laboratory that reports uncertainty but ignores it in decision logic has not operationalized uncertainty governance.

Representation and partition choices can materially alter apparent molecular-model performance [33], while explainability methods may produce chemically persuasive attributions that do not establish mechanism or causality [34]. Stress tests should consequently include scaffold shifts, assay changes, corrupted labels, spurious structural correlates, contradictory analytical modalities, persuasive but invalid explanations, sample swaps, control failure, repeated non-convergence, and alert overload. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop stress-testing autonomous behavior under scientific failure modes within the article's central argument.

**Table 3.** Evaluation, implementation, and research priorities arising from Self-Driving Pharmaceutical Laboratories Need Explicit Rules for Exploration, Confirmation, Failure Recovery, Escalation

| Evaluation dimension | What must be tested | Suitable evidence or assessment | Failure signal | Interpretive limitation | Decision relevance |
|---|---|---|---|---|---|
| **Scientific-state recognition** | Whether the system distinguishes exploration, confirmation, replication, recovery, escalation, and stopping | Expert-labelled campaign traces; blinded state classification; transition review | Exploratory output is treated as confirmed or repeated measurement as replication | Expert agreement does not prove that the state ontology is complete | Determines whether the correct evidentiary obligations are applied |
| **Transition-guard enforcement** | Whether mandatory criteria are checked before state changes | Immutable logs; attempted bypass scenarios; protocol audits | Transition occurs despite failed control, missing provenance, or | Formal compliance can preserve poorly designed rules | Determines whether optimization can override governance |

| | | | | | |
|---|---|---|---|---|---|
| | | | excessive uncertainty | | |
| **Objective robustness** | Whether the system detects invalid, incomplete, or exploitable objectives | Adversarial objectives; competing pharmaceutical constraints; surrogate-objective perturbations | Score improves while scientific or pharmaceutical suitability degrades | Test objectives cannot represent all real campaign goals | Determines whether autonomous optimization should continue |
| **Distribution-shift response** | Whether model behaviour changes appropriately under unfamiliar chemistry, assays, instruments, or materials | Scaffold split, assay transfer, temporal split, new-laboratory or instrument evaluation | Confidence remains high while error or inconsistency increases | Simulated shift may not reproduce prospective deployment conditions | Determines whether the system may act, confirm, or escalate |
| **Uncertainty calibration** | Whether uncertainty corresponds to observed error and governs action | Calibration analysis, selective prediction, coverage assessment, out-of-domain challenge | High-confidence failure or uncertainty without behavioural consequence | Calibration is local to the tested distribution | Determines whether autonomous authority should be reduced |
| **Measurement-integrity response** | Whether failed controls, drift, artefacts, and contradictory modalities are detected | Fault injection, altered controls, instrument drift, orthogonal measurement disagreement | Invalid measurement updates the model or promotes a candidate | Known fault libraries may omit novel failure mechanisms | Determines whether evidence must be quarantined or remeasured |
| **Physical contamination control** | Whether sample, reagent, plate, and instrument contamination is contained | Sample swaps, carryover tests, altered reagent lots, blinded contamination events | Contaminated lineage propagates into later experiments | Deliberate contamination may be easier to detect than natural contamination | Determines whether the campaign can return to a qualified state |
| **Informational contamination control** | Whether leakage and compromised data descendants are identified | Train–test leakage, confirmatory-data reuse, preprocessing contamination, model-version conflicts | Evaluation data influence model selection or future experiment choice | Some dependencies may be undocumented or indirect | Determines which results and models must be invalidated |
| **Failure-recovery behaviour** | Whether recovery remains bounded, diagnostic, and evidence preserving | Repeated-failure scenarios; known instrument faults; ambiguous root causes | Endless retries, criterion relaxation, or selective data deletion | Successful rerun does not prove root-cause identification | Determines whether the system may resume or must escalate |
| **Human-escalation quality** | Whether appropriate events reach qualified reviewers with usable information | Human-factors studies; packet-completeness ratings; missed-event review | Alert fatigue, silent continuation, or evidence-poor escalation | Reviewer expertise and institutional practice vary | Determines when autonomous authority transfers |
| **Stopping behaviour** | Whether the laboratory stops for success, futility, risk, contamination, and resource limits | Counterfactual campaign analysis; prospective stopping-policy comparison | Continued experimentation after trigger or premature termination | No universal optimal stopping rule exists | Determines scientific and resource proportionality |
| **Explanation validity** | Whether explanations remain stable, chemically meaningful, and distinguishable from causal evidence | Attribution perturbation, expert chemical review, counterfactual testing | Persuasive explanation persists after the underlying relation disappears | Domain plausibility does not establish causal truth | Determines whether an explanation may support confirmation or only further inquiry |

*Limitations and Boundary Conditions*

PAEGS is a conceptual governance architecture rather than a validated laboratory-control system. The selected literature supports its constituent concerns—closed-loop experimentation, benchmark limitations, uncertainty, reproducibility, autonomous tool use, and failure detection—but does not empirically validate the full state machine. The proposed states and transitions may require substantial modification across medicinal chemistry, biochemical screening, formulation science,

process development, analytical laboratories, and drug-delivery research. Thresholds appropriate for one assay, instrument, or hazard class should not be transferred automatically to another.

The architecture also depends on the quality of the scientific objective, analytical methods, controls, and human expertise surrounding it. Explicit rules can make decisions more traceable without making weak assumptions correct. A well-governed system may still produce an inconclusive result, fail to detect an unknown contamination pathway, or escalate to a reviewer who reaches an incorrect judgment. Governance should therefore be understood as a means of structuring scientific responsibility and reducing preventable failure, not as a guarantee of truth, safety, reproducibility, or therapeutic relevance.

The proposal does not establish clinical utility, regulatory acceptability, manufacturing readiness, or universal deployment requirements. It does not determine whether a molecular candidate is safe, effective, synthesizable at scale, developable, or therapeutically valuable. It also does not prescribe numerical escalation or stopping thresholds. Such values must be justified prospectively for the intended pharmaceutical context, including the consequences of false continuation, false stopping, missed failure, and excessive human intervention.

*Research Agenda and Implementation Priorities*

The first research priority is to develop a shared ontology and machine-readable specification for autonomous scientific states. This work should test whether independent experts can consistently distinguish exploration, confirmation, replication, recovery, escalation, and stopping from complete campaign records. Formal methods could then encode state preconditions, invariants, prohibited transitions, and evidence requirements. Evaluation should focus not only on agreement with expert labels but also on whether the ontology prevents scientifically invalid shortcuts under realistic experimental conditions.

The second priority is prospective governance evaluation. Controlled studies should compare autonomous campaigns operating with explicit state and stopping rules against campaigns governed primarily by optimizer convergence or local performance measures. Relevant outcomes include reproducibility, number and severity of invalid transitions, contamination propagation, missed escalation events, unnecessary human interventions, resource use, and independently confirmed scientific findings. Fault-injection studies should test instrument failures, sample swaps, leakage, model drift, contradictory modalities, ambiguous root causes, and repeated non-convergence.

The third priority is staged implementation. Initial systems should operate in advisory or shadow mode, generating state and transition recommendations without controlling consequential actions. After independent audit, limited authority could be granted for low-risk, well-characterized tasks with validated instruments, stable protocols, and clear rollback procedures. Broader authority should depend on demonstrated transition compliance, uncertainty calibration, contamination containment, successful recovery, and manageable escalation burden. Shared reporting standards, interoperable provenance schemas, and cross-platform challenge suites will be necessary before governance performance can be compared across laboratories.

## Conclusion

Self-driving pharmaceutical laboratories require governance rules that define not only which experiment should occur next but also what scientific state the experiment occupies, what evidence permits progression, when failure requires containment, when uncertainty exceeds delegated authority, and when experimentation must stop. The Pharmaceutical Autonomous Experimentation Governance State Machine organizes qualified readiness, exploration, confirmation, replication, failure recovery, human escalation, and stopping as distinct but connected states supported by transition guards and a continuous audit-and-contamination control plane. Its central contribution is to separate optimization performance from scientific validity and to make the evidentiary meaning of autonomous action explicit. The construct is proposed rather than validated and must be adapted and prospectively evaluated across pharmaceutical contexts. It does not replace expert judgment, experimental confirmation, independent replication, clinical evaluation, or regulatory assessment; it specifies where those forms of authority and evidence must constrain autonomous behaviour.

**Acknowledgments:** None

**Conflict of interest:** None

**Financial support:** None

**Ethics statement:** None

## References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3

3. 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
4. Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, et al. ChemOS: Orchestrating autonomous experimentation. Sci Robot. 2018;3(19). doi:10.1126/scirobotics.aat5559
5. Volk AA, Abolhasani M. Performance metrics to unleash the power of self-driving labs in chemistry and materials science. Nat Commun. 2024;15(1):1378. doi:10.1038/s41467-024-45569-5
6. 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
7. 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
8. 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
9. Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
10. Häse F, Roch LM, Aspuru-Guzik A. Phoenics: A Bayesian optimizer for chemistry. ACS Cent Sci. 2018;4(9):1134-45. doi:10.1021/acscentsci.8b00307
11. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature. 2021;590(7844):89-96. doi:10.1038/s41586-021-03213-y
12. 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
13. Nambiar AMK, Breen CP, Hart T, Kulesza T, Jamison TF, Jensen KF. Bayesian optimization of computer-proposed multistep synthetic routes on an automated robotic flow platform. ACS Cent Sci. 2022;8(6):825-36. doi:10.1021/acscentsci.2c00207
14. Eyke NS, Koscher BA, Jensen KF. Toward machine learning-enhanced high-throughput experimentation. Trends Chem. 2021;3(2):120-32. doi:10.1016/j.trechm.2020.12.001
15. Seifrid M, Pollice R, Aguilar-Granda A, Chan ZM, Hotta K, Ser CT, et al. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. Acc Chem Res. 2022;55(17):2454-66. doi:10.1021/acs.accounts.2c00220
16. Mehr SHM, Craven M, Leonov AI, Keenan G, Cronin L. A universal system for digitization and automatic execution of the chemical synthesis literature. Science. 2020;370(6512):101-8. doi:10.1126/science.abc2986
17. Coley CW, Thomas DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science. 2019;365(6453). doi:10.1126/science.aax1566
18. Boiko DA, MacKnight R, Kline B, Gomes GDP. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0
19. Dai T, Vijayakrishnan S, Szczypiński FT, Ayme JF, Simaei E, Fellowes T, et al. Autonomous mobile robots for exploratory synthetic chemistry. Nature. 2024;635(8040):890-7. doi:10.1038/s41586-024-08173-7
20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
21. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. 2019;1(1):20-3. doi:10.1038/s42256-018-0004-1
22. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3
23. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9). doi:10.1016/S2589-7500(20)30186-2
24. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797
25. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. PLoS Comput Biol. 2017;13(6). doi:10.1371/journal.pcbi.1005510
26. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol. 2017;35(4):342-6. doi:10.1038/nbt.3780
27. 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
28. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0
29. 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
30. 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

31. 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
32. 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
33. 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
34. 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