Pharmacophore

ISSN-2229-5402

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

DO SELF-DRIVING LABORATORIES ACCELERATE DISCOVERY OR AUTOMATE UNCERTAINTY, BIAS, AND IRREPRODUCIBLE DECISIONS?

Chen Hao[1]*, Liu Fang[1], Zhao Lin[2], Wei Zhang[3]

1. ***Department of Self-Driving Labs and Discovery Acceleration, Faculty of Pharmacy, Zhejiang University, Hangzhou, China.***
2. ***Department of Uncertainty and Bias in Automated Labs, Faculty of Pharmaceutical Sciences, Nanjing University, Nanjing, China.***
3. ***Department of Irreproducible Decisions in Autonomous Discovery, Faculty of Pharmacy, University of Melbourne, Melbourne, Australia.***

ARTICLE INFO

Received:
17 April 2026
Received in revised form:
12 August 2026
Accepted:
16 August 2026
Available online:
28 August 2026

Keywords: Self-driving laboratories, Realist review, Closed-loop experimentation, Laboratory automation, Active learning, Uncertainty quantification

ABSTRACT

Self-driving laboratories are increasingly presented as a route to faster, more systematic discovery by coupling algorithmic experiment selection with automated synthesis, testing, and analysis. Yet the same closed-loop structure that can improve experimental efficiency can also repeat hidden biases, amplify measurement error, misinterpret model confidence, and convert poorly specified objectives into reproducible but scientifically weak decisions. This realist review asks not whether laboratory autonomy works in the abstract, but what works, for whom, under which technical and organizational conditions, through which mechanisms, and with what outcomes. The review develops an initial programme theory, searches iteratively for explanatory evidence, and organizes findings as context–mechanism–outcome configurations across design, synthesis, testing, analysis, uncertainty management, and pharmaceutical translation. The synthesis identifies four conditional acceleration mechanisms: faster feedback between prediction and measurement, information-efficient experiment selection, consistent execution of supported procedures, and cumulative learning from structured experimental records. It also identifies four counter-mechanisms: reinforcement of biased or narrow data, exploitation of measurement noise and proxy objectives, propagation of miscalibrated uncertainty, and loss of reproducibility through brittle interfaces or underspecified protocols. The central contribution is a proposed staged-autonomy interpretation in which decision authority expands only when the relevant objective, assay, instrumentation, data lineage, uncertainty estimates, failure recovery, and human escalation pathways have demonstrated adequate reliability. The available evidence remains strongest for bounded chemical and materials tasks and substantially weaker for routine pharmaceutical decision-making, biological complexity, independent replication, and long-duration operation. Self-driving laboratories may accelerate discovery, but acceleration is not an intrinsic property of autonomy; it is an outcome produced only when enabling contexts activate reliable scientific mechanisms while safeguards interrupt error-amplifying feedback.

To Cite This Article: Chen H, Liu F, Zhao L, Wei Z. Do Self-Driving Laboratories Accelerate Discovery or Automate Uncertainty, Bias, and Irreproducible Decisions? Pharmacophore. 2026;17(4):146-56. https://doi.org/10.51847/i178ALlztD

Introduction

Self-driving laboratories combine computational experiment selection with automated execution, measurement, analysis, and iterative updating. Their distinguishing feature is therefore not robotic motion alone but the closure of a decision loop in which observations from one experiment influence what the system does next. Early accounts framed these laboratories as next-generation experimental systems capable of integrating machine learning, robotics, and domain knowledge to explore chemical or materials spaces more efficiently [1]. In pharmaceutical science, this proposition is especially consequential because discovery decisions are distributed across molecular design, synthesis, physicochemical characterization, biochemical testing, cellular evaluation, formulation, and evidence integration. A system that shortens the interval between hypothesis, experiment, and revision could reduce avoidable experimentation and improve the use of scarce material or assay capacity. Conversely, a system that repeatedly acts on an incomplete objective, biased model, drifting instrument, or invalid assay could generate a

Corresponding Author: Chen Hao; Department of Self-Driving Labs and Discovery Acceleration, Faculty of Pharmacy, Zhejiang University, Hangzhou, China. E-mail: chen.hao@zju.edu.cn

large volume of internally consistent but scientifically misleading evidence. The central issue is therefore not whether laboratory automation is technically feasible, but whether closed-loop autonomy produces reliable knowledge under the conditions in which pharmaceutical decisions must be made.

Establishing such autonomy requires more than connecting a predictive model to a robot. Hardware must execute the requested operation within known tolerances; analytical instruments must generate interpretable measurements; software must preserve state across instruments and data systems; decision algorithms must account for constraints and uncertainty; and failure-handling procedures must distinguish recoverable technical events from scientifically informative negative outcomes. These dependencies make the self-driving laboratory a coupled sociotechnical system rather than an isolated algorithmic product. Analyses of autonomous chemical experimentation emphasize that hardware, software, analytical feedback, planning, and error management must operate coherently before a closed loop can be trusted [2]. This coherence is difficult to establish because errors can originate at multiple interfaces and may remain latent. A mislabeled reagent, an unstable baseline, an incorrect unit conversion, or an overconfident prediction may appear locally plausible while changing every subsequent experiment selected by the system. Autonomy can consequently transform small upstream defects into systematic decision trajectories unless the architecture contains mechanisms for detecting disagreement, pausing execution, and requesting human review.

The published evidence also covers markedly different meanings of a self-driving laboratory. Some platforms automate a fixed procedure while leaving scientific objectives and experiment selection to a person. Others use optimization algorithms to choose among predefined conditions, and a smaller group integrates design, execution, analysis, and iterative selection. Reviews of the field show substantial variation in application domain, robotic architecture, analytical modality, search strategy, objective complexity, and degree of demonstrated autonomy [3]. Such heterogeneity complicates claims that self-driving laboratories accelerate "discovery" as a general category. Faster optimization of a bounded reaction condition, autonomous identification of a material with a specified property, and selection of a therapeutically credible molecule are not equivalent outcomes. They differ in the number of causal layers separating experimental measurement from the intended decision, the consequences of error, and the amount of external confirmation required. Benchmark success, prospective experimental confirmation, mechanistic relevance, developability, and routine pharmaceutical readiness must therefore be treated as distinct evidence states.

The unresolved gap is explanatory. Existing accounts commonly organize systems by instruments, algorithms, or application areas, but these classifications do not by themselves explain why similar closed-loop arrangements succeed in one setting and fail or mislead in another. Physical configurations—including batch, flow, mobile, centralized, and distributed laboratories—also allocate flexibility, authority, and failure risk differently [4]. This review therefore uses realist reasoning to examine the interaction between technical configuration and scientific context. Its objective is to identify the contexts that enable or inhibit reliable autonomy, the mechanisms triggered by those contexts, and the outcomes that follow. The review does not estimate a universal acceleration effect and does not treat autonomy as a single intervention. Instead, it develops an explanatory synthesis across design, synthesis, testing, and analysis; distinguishes mechanisms of acceleration from mechanisms of uncertainty amplification; and proposes evidence-informed principles for staged autonomy. These principles are intended to organize evaluation and implementation. They have not been empirically validated as a universal maturity standard or regulatory framework.

Realist-Review Questions and Context–Mechanism–Outcome Approach

A realist review begins from the proposition that an intervention does not produce an outcome independently of its setting. Outcomes arise when resources or opportunities introduced by an intervention activate particular mechanisms among participants or within a system. Realist synthesis is therefore suited to complex programmes whose effects depend on implementation conditions, interpretation, and interaction among components [5]. Applied to self-driving laboratories, the "intervention" is not merely a robot or optimization algorithm. It is an arrangement of objectives, models, data, instruments, protocols, analytical methods, authority rules, and human responsibilities. The initial programme theory proposed here is that autonomy may accelerate reliable discovery when the search space is scientifically meaningful, experimental measurements are valid, uncertainty is represented appropriately, execution is traceable, and decision authority is bounded by evidence. Under these conditions, closed-loop feedback may activate information-efficient exploration, consistent execution, and cumulative learning. Where objectives are weak proxies, data are biased, measurements drift, or oversight is ineffective, the same feedback structure may activate error reinforcement, premature convergence, and opaque propagation of unreliable conclusions.

Context is defined as the set of conditions that changes whether and how a mechanism operates, rather than as a descriptive list of laboratory features. Realist-methodological work cautions that context should be linked explicitly to causal explanation [6]. For this review, contextual domains include the scientific maturity of the objective; dimensionality and coverage of the search space; assay validity and biological relevance; instrument calibration and maintenance; data provenance; availability of negative and failed experiments; model applicability; uncertainty calibration; protocol completeness; interoperability; human expertise; escalation rights; and the consequences of an incorrect decision. Mechanisms are the processes through which these conditions shape system behaviour. Examples include reduction of experiment-selection latency, prioritization of information gain, standardization of routine operations, reinforcement of sampling bias, exploitation of noisy measurements, automation complacency, and suppression of alternatives outside the encoded search space. Outcomes include not only speed

or target optimization but also validity, novelty, reproducibility, resource use, failure recovery, transferability, and the evidentiary strength of downstream pharmaceutical decisions.

The review questions follow this explanatory structure. First, how should realist questions and context–mechanism–outcome configurations be operationalized transparently for self-driving laboratories? Second, how do configurations of automation across design, synthesis, testing, and analysis allocate scientific authority and failure risk? Third, through which mechanisms may autonomy accelerate discovery? Fourth, through which mechanisms may it amplify uncertainty, bias, and irreproducibility? Fifth, which contextual conditions govern whether beneficial or harmful mechanisms dominate? Sixth, what evidence-informed principles can support staged autonomy? Seventh, which implementation and evaluation priorities are required before autonomous outputs can influence consequential pharmaceutical decisions? Evidence searching is iterative because explanatory relevance may become apparent only after an initial configuration is proposed. Comparative research on realist-review searching supports combining structured retrieval with citation-based and supplementary approaches to locate evidence that can refine programme theories [7]. The synthesis therefore privileges explanatory contribution over superficial topical similarity and permits evidence from adjacent experimental sciences only when its transfer limitations are made explicit.

Appraisal is also claim-specific. A source may provide rigorous evidence that a robotic platform executed a defined workflow but offer limited support for claims about general scientific autonomy or pharmaceutical translation. Conversely, a methodological study of uncertainty or dataset bias may not evaluate a physical laboratory but may explain a mechanism through which autonomous selection becomes unreliable. Realist appraisal distinguishes whether evidence is relevant to a theory, sufficiently rich to explain a configuration, and methodologically credible for the inference being made [8]. This review accordingly codes demonstrated capability, benchmark performance, plausible utility, prospective experimental confirmation, translational relevance, and routine readiness as separate evidence categories. It uses “not clearly reported” where implementation or validation information is absent and does not convert missing information into negative evidence automatically. Original configurations and maturity interpretations are labelled as proposed syntheses rather than validated models. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop realist-review questions and context-mechanism-outcome approach within the article’s central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Realist-review questions and context-mechanism-outcome approach

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Programme-theory development	Realist reviews, methodological analyses, explanatory technical studies, and prospective laboratory demonstrations	Closed-loop autonomy may produce different outcomes through context-dependent mechanisms	Strong methodological basis for explanatory synthesis; application to self-driving laboratories remains proposed	Programme theories may reflect reviewer assumptions unless competing explanations are tested	Prospective studies designed explicitly to compare alternative mechanisms
Context specification	Evidence describing assays, instruments, search spaces, data, organizations, interfaces, and decision consequences	Technical and organizational conditions govern whether autonomy is reliable	Strong conceptual support; empirical reporting is inconsistent across platforms	Context may be reduced to a checklist without explaining its causal role	Minimum contextual reporting standard for autonomous experimentation
Mechanism identification	Process evidence connecting system resources to changes in selection, execution, interpretation, or oversight	Faster feedback, information-efficient sampling, standardization, error reinforcement, and automation complacency may mediate outcomes	Direct support for some acceleration processes; indirect support for several failure mechanisms	Mechanisms may be inferred from outcomes without direct observation	Fault-injection, mediation, and process-tracing studies
Outcome classification	Prospective experiments, replication studies, benchmark evaluations, and implementation reports	Autonomy may affect speed, resource use, validity, novelty, reproducibility, and transferability	Evidence is strongest for bounded experimental tasks	Optimization success may be presented as discovery without external confirmation	Shared outcome definitions extending beyond speed and target attainment
Iterative evidence searching	Database searching, citation tracking, author searching,	Different search methods retrieve	Methodologically supported for realist reviews	Selective supplementary searching may	Transparent documentation of search decisions

	and theory-directed supplementary retrieval	different types of explanatory evidence		privilege confirming evidence	and theory changes
Evidence appraisal	Relevance, explanatory richness, methodological rigour, and transfer-risk assessment	Evidence can be useful for one inference without supporting broader claims	Strong methodological support	Journal prestige or technical sophistication may substitute incorrectly for claim-level appraisal	Empirical evaluation of appraisal consistency in technical realist reviews
Bias and uncertainty analysis	Dataset audits, calibration studies, split comparisons, drift analysis, and prospective monitoring	Bias and miscalibration can alter autonomous experiment selection and subsequent data generation	Strong evidence for underlying computational risks; limited direct SDL amplification studies	Feedback-generated data can conceal the origin of bias	Longitudinal monitoring of uncertainty and bias within closed loops
Pharmaceutical interpretation	Molecular, biological, developability, translational, and decision-impact evidence	Experimental optimization does not alone establish therapeutic usefulness	Strong conceptual agreement; sparse end-to-end autonomous pharmaceutical evidence	Surrogate objectives may be mistaken for therapeutic value	Prospective, independently reproduced chemical–biological campaigns

Configurations of Automation across Design, Synthesis, Testing, and Analysis

Automation configurations differ according to which functions are connected and which decisions remain external to the system. At one end, a digital workflow may record experiments while humans choose candidates, specify protocols, interpret results, and decide when to stop. At another, software may coordinate experiment selection, equipment instructions, data capture, and model updating. ChemOS illustrates the orchestration layer required to connect an experiment-planning process with physical execution and return observations to the planner [9]. In realist terms, orchestration is a resource that can activate faster feedback only when instruments expose reliable controls, data formats remain interoperable, and the system state is preserved correctly. It does not itself establish that an objective is scientifically valid or that a selected experiment is informative. The same orchestration layer can accelerate an invalid assay or repeatedly execute an incorrectly encoded procedure. Configuration analysis must therefore distinguish connectivity from epistemic authority: a platform may coordinate many instruments yet remain scientifically dependent on fixed human assumptions embedded in the objective, protocol, representation, and stopping rule.

Design-to-synthesis configurations connect computational proposal or planning to physical production. A robotic flow platform informed by artificial-intelligence planning demonstrated that retrosynthetic analysis and supported synthesis operations can be integrated prospectively [10]. The evidentiary significance is substantial but bounded. The platform establishes that a computationally proposed route can be translated into executable operations within a defined hardware and reaction repertoire; it does not establish that any synthetically plausible molecular proposal can be produced autonomously. In pharmaceutical use, this distinction matters because route feasibility depends on reagent availability, chemoselectivity, purification, solid-state behaviour, scale, safety, and undocumented practical judgement. The mechanism of acceleration is the reduction of handoffs between planning and execution, but it is activated only where route descriptions map reliably onto available modules. Outside that supported domain, automation may create brittle execution, repeated failure, or false negative conclusions about synthesizability. Human review therefore remains important at boundaries where the system encounters unfamiliar transformations, hazardous conditions, analytical ambiguity, or conflicting route constraints.

Machine-executable synthesis creates a related but distinct configuration. A modular robotic system driven by a chemical programming language showed that synthetic operations can be formalized and executed through reusable instructions [11]. Formalization can improve consistency by making procedural steps explicit, reducing informal transcription, and enabling repeated execution of supported operations. It can also improve auditability when the instruction set, hardware state, reagent identity, and analytical output are retained together. However, abstraction can conceal material differences between instruments, vessels, transfer lines, mixing regimes, or environmental conditions. A protocol may be syntactically portable while remaining chemically non-equivalent across laboratories. Testing and analysis introduce further dependencies: the platform must know whether a measurement is technically valid, whether a signal corresponds to the intended species, and whether the result should trigger repetition, adaptation, or termination. Consequently, machine executability should be evaluated separately from chemical fidelity, analytical validity, and cross-platform reproducibility.

Literature-to-execution configurations attempt to extend automation by converting reported procedures into standardized instructions. Digitization and automatic execution of synthetic literature demonstrate the potential to reduce the distance between published methods and laboratory action [12]. Yet published procedures may omit tacit knowledge, instrument-specific adjustments, visual cues, negative trials, or the reasoning used to interpret deviations. When these omissions are carried into a closed loop, the system may treat incomplete reporting as complete procedural knowledge. The realist synthesis therefore represents the self-driving laboratory as a context-dependent engine rather than a uniform pipeline: design, synthesis, testing,

and analysis are coupled, but each interface can activate either reliable learning or compounded error. **Figure 1** presents the self-driving laboratory context–mechanism–outcome engine, showing how the article's key components and boundaries are connected within configurations of automation across design, synthesis, testing, and analysis.

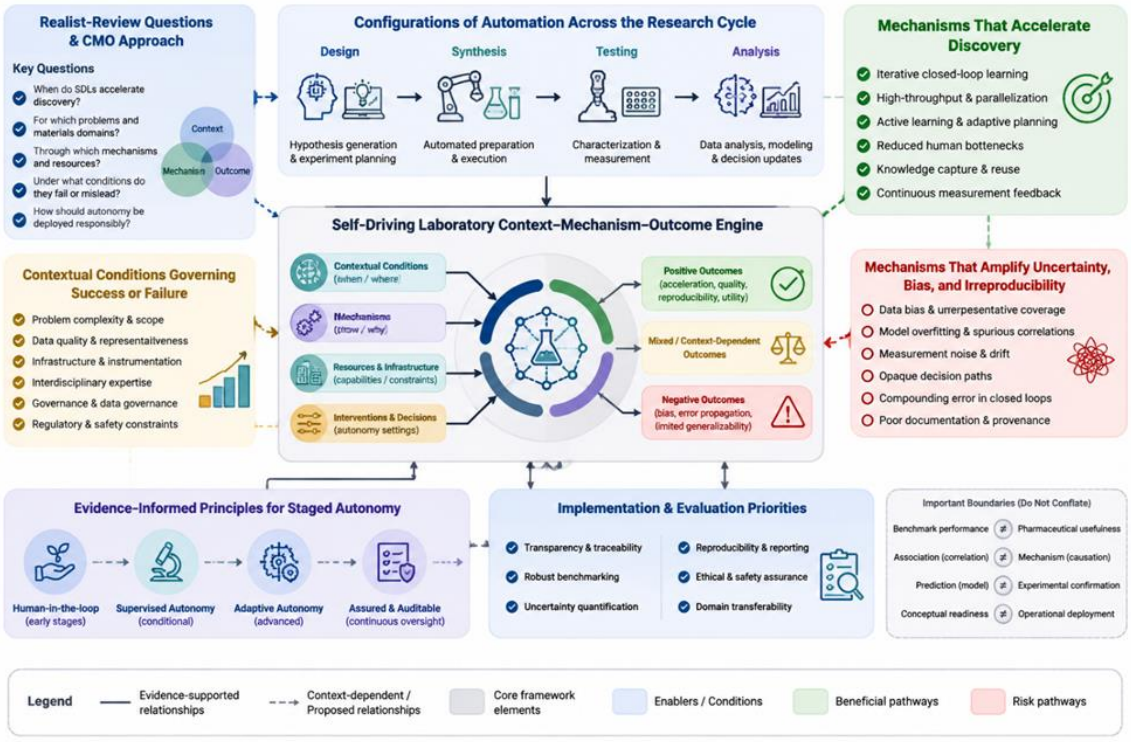

Figure 1. Self-Driving Laboratory Context–Mechanism–Outcome Engine.

The figure is an original conceptual synthesis that organizes the article's central contribution across realist-review questions and context-mechanism-outcome approach, configurations of automation across design, synthesis, testing, and analysis, configurations of automation across design, mechanisms by which autonomy may accelerate discovery, mechanisms that amplify uncertainty, bias, and irreproducibility, mechanisms that amplify uncertainty. 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.

Mechanisms by Which Autonomy May Accelerate Discovery

The first acceleration mechanism is information-directed exploration. Conventional experimental campaigns may prioritize familiar conditions, convenient compounds, or locally promising regions even when those choices add little new knowledge. A self-driving system can instead select experiments according to an explicit acquisition rule intended to reduce uncertainty, test competing hypotheses, or explore under-sampled regions. Robotic reactivity search has shown that machine learning can guide physical experimentation toward unfamiliar chemical behaviour within a defined reagent and analytical space [13]. The resulting advantage is not that the system "understands" chemistry independently, but that it can apply a consistent selection policy across more alternatives than a person could inspect manually. This mechanism remains conditional on whether the representation, candidate pool, and measurement system contain the distinctions needed to recognize genuinely informative outcomes.

The second mechanism is reduction of operational latency between experiment selection and execution. A mobile robotic chemist demonstrated that an autonomous agent could navigate among existing instruments, conduct repeated experiments, analyse outputs, and select subsequent conditions within a bounded campaign [14]. Mobility can reduce the need to construct a fully integrated fixed platform and can increase access to heterogeneous instruments. Nevertheless, physical flexibility does not guarantee scientific flexibility. The robot remains constrained by its manipulators, navigation accuracy, supported vessels, instrument interfaces, and programmed recovery behaviours. Acceleration emerges when routine transfers and measurements are dependable enough that the system can continue without repeated human intervention. If those conditions are absent, mobility may merely relocate bottlenecks or introduce additional positioning and coordination failures.

A third mechanism is rapid conversion of measurements into revised experimental priorities. Reactivity-seeking autonomous platforms illustrate how online analytical signals can be interpreted by a learned model and used to redirect subsequent experimentation [15]. This compression of the hypothesis–experiment–analysis cycle may help laboratories investigate uncertain regions while the relevant materials, instruments, and contextual information remain available. It can also make negative or unexpected outcomes operationally useful by feeding them directly into the next decision. Yet the mechanism

works only when the analytical signal is a valid indicator of the scientific phenomenon of interest. If peak detection, classification, or novelty scoring is unreliable, the system may optimize responsiveness to an artefact rather than learning chemistry.

The most pharmaceutically relevant acceleration mechanism is prospective learning across multiple molecular objectives. A closed loop linking predictions, physical measurements, and model updating has been demonstrated for multiproperty molecular discovery [16]. Such systems may help expose trade-offs that are obscured when properties are optimized sequentially or in separate teams. However, molecular-property improvement remains an intermediate outcome and does not establish biological efficacy, safety, developability, or therapeutic value. End-to-end autonomous synthesis and characterization have also produced experimentally confirmed materials within a constrained domain [17]. Together, these demonstrations support the feasibility of integrated learning loops, but they do not show that autonomy transfers without loss across different chemical classes, assay systems, organizations, or pharmaceutical decision stages. The realist conclusion is therefore conditional: autonomy accelerates discovery when rapid feedback produces scientifically valid information that can change the next decision, not merely when robots execute experiments faster.

Mechanisms that Amplify Uncertainty, Bias, and Irreproducibility

The principal counter-mechanism is feedback reinforcement of biased evidence. When an experiment planner is trained or evaluated on chemically redundant data, apparently strong predictions may reflect memorization of familiar structural patterns rather than generalization to meaningfully new candidates. Ligand-based benchmark analyses have shown that redundancy between training and validation data can reward memorization [18]. In a passive model, such bias affects a prediction. In a self-driving laboratory, it can also affect which experiment is performed next and therefore which new data become available. The system may repeatedly sample familiar regions, obtain predictable results, and interpret this self-generated confirmation as evidence that its model is improving. Closed-loop operation can thus convert static dataset bias into an adaptive sampling trajectory.

A related mechanism is exploitation of unintended dataset cues. Chemical datasets may contain differences in composition, preparation, decoy construction, or annotation that allow a model to distinguish classes without learning the intended molecular determinants. Bias-controlled evaluation in structure-based virtual screening has demonstrated that models can benefit from such unintended structure [19]. When these models guide experiments, the resulting selection policy may systematically exclude unfamiliar chemical regions or prioritize candidates for reasons unrelated to the desired mechanism. More automation does not correct this problem automatically. Unless the laboratory tests whether model decisions remain valid under alternative splits, external compounds, changed assay conditions, and prospective challenge sets, closed-loop efficiency can make a biased selection process operate more consistently rather than more truthfully.

A third counter-mechanism is misinterpretation of predictive uncertainty. Comparative studies of scalable uncertainty-estimation methods show that their calibration and out-of-domain behaviour vary across methods and tasks [20]. An autonomous system that treats an uncalibrated score as reliable uncertainty may select experiments too aggressively, avoid unfamiliar candidates unjustifiably, or stop before competing explanations have been tested. The problem is compounded when uncertainty estimates are derived from the same modelling assumptions that generated the prediction. Apparent agreement among related models may then be mistaken for genuine epistemic security. Reliable autonomy requires prospective monitoring of whether uncertainty predicts actual error, especially as the experiment-selection policy changes the distribution of newly generated data.

Uncertainty also has multiple sources that demand different responses. Explainable approaches distinguish uncertainty associated with noisy observations from uncertainty arising from limited knowledge or insufficient domain coverage [21]. Repeating an experiment may help estimate measurement variability, whereas epistemic uncertainty may require new chemical regions, alternative representations, or orthogonal assays. Treating both as a single confidence value prevents the system from choosing an appropriate corrective action. Moreover, systematic evaluations of molecular prediction show that representations, architectures, tasks, and data-splitting strategies can materially alter reported performance [22]. If these design choices are hidden within an automated platform, users may receive a seemingly precise recommendation without understanding the methodological decisions that shaped it. Autonomy amplifies uncertainty when it accelerates action while reducing opportunities to examine how uncertainty was produced.

Contextual Conditions Governing Success or Failure

The first governing context is the reliability of the physical and analytical loop. An integrated self-driving laboratory for thin-film materials demonstrated how synthesis, characterization, and experiment planning can be linked within a defined platform [23]. Its significance lies not only in adaptive selection but also in the stability of the operations that generate measurements for the planner. In pharmaceutical laboratories, analogous reliability would require verified liquid handling, compound identity, concentration control, contamination monitoring, instrument calibration, sample traceability, and assay quality. Where these conditions are strong, rapid iteration can accumulate interpretable evidence. Where they are weak, the system may repeatedly learn from technical variation while representing it as molecular or biological information.

The second context is data suitability at the moment of decision. Automated experimentation can increase the speed and regularity of chemical data generation, but its value depends on experimental design, analytical validity, metadata completeness, and the ability to interpret failed runs [24]. Data volume is therefore not a sufficient contextual condition. A

laboratory may produce many measurements while omitting preparation details, environmental conditions, instrument state, unsuccessful procedures, or reasons for manual intervention. Such omissions reduce the ability to reproduce the campaign or determine whether later model changes reflect chemistry, instrumentation, or procedural drift. Successful autonomy requires data structures that preserve the experimental circumstances needed to interpret each result rather than only the final target value.

The third context is fit between the optimization method and the scientific problem. Bayesian reaction optimization can improve experiment selection in real chemical synthesis, but its performance depends on the objective, parameter types, prior assumptions, noise, constraints, and cost of evaluation [25]. An algorithm that performs well for smooth condition optimization may be poorly suited to discontinuous failure regions, sparse categorical decisions, changing objectives, or biological assays with delayed outcomes. Context also determines whether exploration is affordable. A highly informative experiment may consume scarce material, create safety risks, or delay an entire campaign. Autonomous selection should therefore incorporate feasibility and consequence rather than treating every candidate as an interchangeable query to an objective function.

The fourth context is domain complexity and representational completeness. Synthetic-biology self-driving laboratories face dynamic cellular states, biological variability, delayed readouts, and context-dependent responses that differ from many chemical optimization tasks [26]. These features make it harder to assume that the measured outcome depends only on the selected design variables. Reproducibility also depends on whether automated procedures preserve the implementation details needed to recreate the experiment. Computer-science abstractions can improve portability, but poorly specified abstractions may hide consequential platform differences and create reproducibility debt [27]. Success therefore requires an appropriate level of procedural abstraction: sufficiently general to support transfer, yet sufficiently explicit to preserve chemically and biologically meaningful conditions. Across these contexts, autonomy succeeds when the laboratory can detect when its own assumptions no longer hold.

Evidence-Informed Principles for Staged Autonomy

The first principle is evidence before authority. An automated function should receive scientific decision rights only after its behaviour has been evaluated under conditions resembling its intended use. Benchmarking frameworks for noisy experimental optimization can support comparison of planners across varied response surfaces and noise conditions [28]. Such benchmarks are useful for detecting obvious weaknesses and testing sensitivity, but they cannot reproduce all forms of laboratory failure. Progression to physical control should therefore require prospective evidence showing that the planner can handle realistic constraints, missing measurements, failed experiments, and distributional change. Authority should remain narrow when validation is narrow.

The second principle is context-specific qualification rather than universal algorithm ranking. Comparative Bayesian-optimization research across experimental materials domains shows that relative performance changes with the problem and search space [29]. A planner should consequently be qualified for a defined class of tasks rather than labelled generally superior. Qualification should specify supported variable types, objective structures, constraint handling, expected noise, cost assumptions, and conditions requiring escalation. Requalification is needed when the assay, instrument, chemical domain, objective, or consequence of error changes materially. This approach treats an optimizer as one component of an evidence-bearing laboratory configuration rather than as a transferable source of autonomous intelligence.

The third principle is incremental integration of autonomy. Analyses of automated research platforms distinguish physical automation from autonomous scientific decision-making and support adding decision functions progressively [30]. A staged system may begin with fixed-protocol execution, advance to monitored analytical adaptation, and only later receive bounded authority to select experiments or recommend stopping. Each stage should retain explicit human rights to challenge, pause, redirect, and inspect the basis of a recommendation. Advancement should depend not only on successful runs but also on demonstrated failure recognition, safe recovery, provenance completeness, and performance under deliberately difficult cases. Staging is thus a mechanism for matching authority to evidence maturity, not a claim that every laboratory should progress toward unrestricted autonomy.

The fourth principle is modularity with multidimensional evaluation. Atlas demonstrates how a modular experiment-planning architecture can support constrained, mixed-variable, multiobjective, asynchronous, and molecular optimization modes [31]. Modularity can help laboratories replace or validate components independently, but broad software functionality should not be mistaken for equal experimental readiness in every mode. Evaluation frameworks for self-driving laboratories accordingly argue that performance should extend beyond discovery speed to include reliability, efficiency, reproducibility, autonomy, and resource use [32]. A staged-autonomy assessment should also consider calibration, assay validity, recovery from failure, human workload, and downstream decision impact. The proposed stages are explanatory and evaluative constructs derived from the reviewed evidence; they are not validated certification levels.

Implementation and Evaluation Priorities

Implementation should begin with interoperable infrastructure and equitable access, while recognizing that centralization and distribution generate different risks. Shared or distributed self-driving laboratories may broaden access to expensive instruments and specialized expertise, but they also introduce coordination, fidelity, cybersecurity, scheduling, and provenance challenges [33]. A protocol executed at a remote facility may differ because of instrument configuration, reagent source, environmental conditions, or local maintenance. Networked autonomy therefore requires machine-readable procedures

accompanied by platform identity, calibration state, software version, deviations, and raw analytical evidence. Accessibility should be assessed by whether users can understand, challenge, and reproduce the generated results, not merely whether they can submit an experimental request.

Evaluation must also use pharmaceutically meaningful endpoints. Critical analyses of artificial intelligence in drug discovery caution that computational performance is frequently disconnected from prospective impact [34]. The same warning applies to autonomous experimentation. A laboratory may optimize yield, solubility, signal intensity, or predicted activity without improving the probability of identifying a safe and effective medicine. Evaluation should therefore state what decision the autonomous output is intended to support and what additional evidence remains necessary. Molecular plausibility, successful synthesis, assay activity, mechanism confirmation, cellular relevance, exposure, safety, manufacturability, and clinical utility are successive but non-equivalent evidence layers.

A third priority is prospective validation across the development pathway. Machine-learning applications in drug discovery and development span target identification, molecular design, biomarker analysis, clinical development, and manufacturing, each with distinct data and validation requirements [35]. Self-driving laboratories should not be assigned authority across these stages on the basis of a single chemical optimization demonstration. Evaluation should include prospective challenge sets, independent replication, negative-result capture, orthogonal measurements, time-ordered assessment, and comparison with an appropriate human-directed or conventional workflow. Long-duration campaigns are particularly important because they expose instrument drift, reagent degradation, software updates, and maintenance burden that short demonstrations may not reveal.

Finally, implementation requires capability-specific robotics assessment and experimental–computational co-design. The proposed ADePT framework separates adaptability, dexterity, perception, and task complexity when assessing autonomous laboratory robotics [36]. This decomposition is useful because a platform may manipulate standard microplates reliably while lacking the perception or adaptability needed for unfamiliar samples and failure states. Broader recommendations for data science in experimental chemistry emphasize collaboration among experimentalists, data scientists, software developers, and instrument specialists [37]. Such co-design should define objectives, admissible actions, data standards, uncertainty responses, stopping rules, and escalation pathways before autonomous operation begins. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop implementation and evaluation priorities within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Do Self-Driving Laboratories Accelerate Discovery or Automate Uncertainty, Bias, and Irreproducible Decisions

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Fixed-protocol robotic execution	Stable, well-characterized procedures with limited adaptation	Execution fidelity, traceability, calibration checks, recovery from routine failures	Demonstrated in multiple chemical configurations	Does not establish autonomous scientific reasoning	Appropriate entry stage when procedural boundaries are explicit
Closed-loop condition optimization	Bounded variables, measurable objectives, manageable experimental noise	Prospective comparison, repeated campaigns, constraint and failure reporting	Repeated domain-specific demonstrations	Performance may depend strongly on objective and search-space structure	Supports conditional information efficiency, not universal acceleration
Reactivity or novelty-seeking experimentation	Defined chemical spaces with online analytical feedback	Independent confirmation, alternative novelty definitions, out-of-domain challenge	Early prospective evidence	Novelty may reflect representation or analytical artefacts	Promising mechanism requiring stronger replication and falsification
Multiproperty molecular selection	Molecular candidates with experimentally measurable property trade-offs	Time-ordered evaluation, external compounds, synthesis success, physical confirmation	Early direct molecular demonstration	Does not establish biological, safety, or therapeutic value	Relevant to early discovery only when downstream evidence boundaries are explicit
Biological self-driving laboratories	Stateful, variable systems with delayed or context-dependent readouts	Replicated perturbations, controls, orthogonal assays, temporal and batch robustness	Emerging and heterogeneous	Biological causality and transfer are difficult to infer from optimization alone	Requires narrower authority and stronger escalation mechanisms
Uncertainty-aware experiment planning	Changing data distributions and unfamiliar candidates	Calibration, error–uncertainty correspondence, drift	Methodologically developed but unevenly implemented	Confidence estimates may fail outside the training domain	Uncertainty must influence action and escalation, not

		monitoring, response-specific actions			appear only as a reported score
Distributed or remotely accessible SDLs	Multiple sites, instruments, users, and software environments	Cross-platform transfer, provenance completeness, security, protocol-equivalence testing	Early implementation	Platform heterogeneity may undermine reproducibility	Access gains are conditional on transparent infrastructure
Staged pharmaceutical autonomy	Consequential decisions spanning chemical and biological evidence	Prospective evidence gates, independent replication, human override, downstream decision impact	Proposed review synthesis	No validated universal maturity or regulatory standard exists	Authority should expand only with task-specific evidence and demonstrated failure control

Conclusion

Self-driving laboratories can accelerate discovery, but autonomy is neither a uniform intervention nor an independent cause of scientific progress. Reliable acceleration emerges when bounded objectives, valid measurements, appropriate experiment planners, dependable execution, explicit uncertainty, complete provenance, reproducible protocols, and effective human escalation activate mechanisms of rapid feedback and information-efficient learning. The same closed-loop architecture can amplify uncertainty when biased data, proxy objectives, analytical drift, brittle abstractions, or overconfident models determine both what is measured and what is selected next. The original contribution of this realist review is a context–mechanism–outcome interpretation that separates configurations of automation from the authority assigned to them and frames staged autonomy as an evidence-dependent allocation of decision rights. Current evidence supports technically credible, domain-specific autonomous experimentation, but it does not establish routine pharmaceutical readiness, universal reproducibility, or improved therapeutic outcomes. Future evaluation should therefore compare autonomous campaigns prospectively, expose systems to realistic failures, examine long-duration stability, verify uncertainty under adaptive sampling, and trace whether experimental gains improve consequential downstream decisions. The appropriate question is not whether laboratories should become fully autonomous, but which decisions a particular system can justify under specified conditions, what evidence supports that authority, and how reliably the system recognizes when it should stop and defer.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. 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
3. Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-driving laboratories for chemistry and materials science. Chem Rev. 2024;124(16):9633-732. doi:10.1021/acs.chemrev.4c00055
4. Bayley O, Savino E, Slattery A, Noël T. Autonomous chemistry: Navigating self-driving labs in chemical and material sciences. Matter. 2024;7(7):2382-98. doi:10.1016/j.matt.2024.06.003
5. Jagosh J. Realist synthesis for public health: Building an ontologically deep understanding of how programs work, for whom, and in which contexts. Annu Rev Public Health. 2019;40:361-72. doi:10.1146/annurev-publhealth-031816-044451
6. Greenhalgh J, Manzano A. Understanding 'context' in realist evaluation and synthesis. Int J Soc Res Methodol. 2022;25(5):583-95. doi:10.1080/13645579.2021.1918484
7. Duddy C, Roberts N. Identifying evidence for five realist reviews in primary health care: A comparison of search methods. Res Synth Methods. 2022;13(2):190-203. doi:10.1002/jrsm.1523
8. Dada S, Dalkin S, Gilmore B, Hunter R, Mukumbang FC. Applying and reporting relevance, richness and rigour in realist evidence appraisals: Advancing key concepts in realist reviews. Res Synth Methods. 2023;14(3):504-14. doi:10.1002/jrsm.1630

9. 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
10. Coley CW, Thomas DA 3rd, 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
11. Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. Science. 2019;363(6423). doi:10.1126/science.aav2211
12. 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
13. 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
14. 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
15. Caramelli D, Granda JM, Mehr SHM, Cambié D, Henson AB, Cronin L. Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network. ACS Cent Sci. 2021;7(11):1821-30. doi:10.1021/acscentsci.1c00435
16. Koscher BA, Canty RB, McDonald MA, Greenman KP, McGill CJ, Bilodeau CL, et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. Science. 2023;382(6677). doi:10.1126/science.adi1407
17. Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature. 2023;624(7990):86-91. doi:10.1038/s41586-023-06734-w
18. 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
19. 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
20. 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
21. Yang CI, Li YP. Explainable uncertainty quantifications for deep learning-based molecular property prediction. J Cheminform. 2023;15(1):13. doi:10.1186/s13321-023-00682-3
22. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
23. MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(20). doi:10.1126/sciadv.aaz8867
24. Shi Y, Prieto PL, Zepel T, Grunert S, Hein JE. Automated experimentation powers data science in chemistry. Acc Chem Res. 2021;54(3):546-55. doi:10.1021/acs.accounts.0c00736
25. 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
26. Martin HG, Radivojevic T, Zucker J, Bouchard K, Sustarich J, Peisert S, et al. Perspectives for self-driving labs in synthetic biology. Curr Opin Biotechnol. 2023;79:102881. doi:10.1016/j.copbio.2022.102881
27. Canty RB, Abolhasani M. Reproducibility in automated chemistry laboratories using computer science abstractions. Nat Synth. 2024;3(11):1327-39. doi:10.1038/s44160-024-00649-8
28. Häse F, Aldeghi M, Hickman RJ, Roch LM, Christensen M, Liles E, et al. Olympus: A benchmarking framework for noisy optimization and experiment planning. Mach Learn Sci Technol. 2021;2(3):035021. doi:10.1088/2632-2153/abedc8
29. Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. NPJ Comput Mater. 2021;7(1):188. doi:10.1038/s41524-021-00656-9
30. Canty RB, Koscher BA, McDonald MA, Jensen KF. Integrating autonomy into automated research platforms. Digit Discov. 2023;2(5):1259-68. doi:10.1039/D3DD00135K
31. Hickman RJ, Sim M, Pablo-García S, Tom G, Woolhouse I, Hao H, et al. Atlas: A brain for self-driving laboratories. Digit Discov. 2025;4(4):1006-29. doi:10.1039/D4DD00115J
32. 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
33. Canty RB, Bennett JA, Brown KA, Buonassisi T, Kalinin SV, Kitchin JR, et al. Science acceleration and accessibility with self-driving labs. Nat Commun. 2025;16(1):3856. doi:10.1038/s41467-025-59231-1
34. 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
35. 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

36. 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
37. Yano J, Gaffney KJ, Gregoire JM, Hung L, Ourmazd A, Schrier J, et al. The case for data science in experimental chemistry: Examples and recommendations. Nat Rev Chem. 2022;6(5):357-70. doi:10.1038/s41570-022-00382-w