Pharmacophore

ISSN-2229-5402

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

WHO IS ACCOUNTABLE WHEN AN ARTIFICIAL INTELLIGENCE AGENT CHOOSES THE NEXT PHARMACEUTICAL EXPERIMENT?

Sven Larsson[1]*, Erik Johansson[2], Anna Nilsson[3], Lars Andersson[4]

1. ***Department of Accountability in AI-Driven Experimentation, Faculty of Pharmacy, Swedish University of Agricultural Sciences, Uppsala, Sweden.***
2. ***Department of Autonomous Agent Decision-Making, Faculty of Pharmacy, KTH Royal Institute of Technology, Stockholm, Sweden.***
3. ***Department of Governance and Responsibility in AI Pharma, Faculty of Pharmaceutical Sciences, Lund University, Lund, Sweden.***
4. ***Department of Ethical Frameworks for Agentic AI, Faculty of Pharmacy, University of Gothenburg, Gothenburg, Sweden.***

ARTICLE INFO

Received:
21 March 2026
Received in revised form:
18 July 2026
Accepted:
20 July 2026
Available online:
28 August 2026

Keywords: Agentic artificial intelligence, Pharmaceutical experimentation, Accountability, Self-driving laboratories, Scientific traceability, Human oversight

ABSTRACT

Artificial intelligence systems are moving from retrospective pharmaceutical prediction toward operational roles in which they formulate plans, invoke computational tools, rank candidate experiments, and coordinate laboratory actions. This transition creates an unresolved accountability problem: an agent may be causally influential in selecting an experiment without possessing the professional authority, institutional duties, or capacity for answerability normally associated with scientific decision-making. This critical review evaluates how accountability should be structured when an artificial intelligence agent participates in choosing the next pharmaceutical experiment. The literature discussed in this critical review was organized around pharmaceutical evidence quality, autonomous chemical experimentation, algorithmic accountability, operational ethics, and human control. The review distinguishes operational agency from decision authority, answerability, and responsibility for remedy. It argues that accountability cannot be assigned solely to the scientist who approves an experiment, the developer who created the model, the organization that deployed the system, or the agent that generated the recommendation. Instead, accountability must be distributed across the full decision lifecycle while preserving identifiable non-delegable duties. The central conceptual contribution is a proposed accountability lens that evaluates agent-selected experiments according to evidence provenance, data suitability, tool validity, uncertainty, feasibility, safety, authorization, traceability, override capacity, and contestability. Existing evidence supports bounded capabilities in planning, robotic execution, and chemistry-tool orchestration, but does not establish routine readiness for autonomous pharmaceutical experiment authorization. Important limitations include heterogeneous validation settings, bespoke laboratory systems, weak prospective comparison, and limited evidence connecting explanations or audit records to improved scientific outcomes. Agentic pharmaceutical science should therefore be governed as a socio-technical experimental system rather than as an autonomous model operating outside established scientific responsibility.

To Cite This Article: Larsson S, Johansson E, Nilsson A, Andersson L. Who Is Accountable When an Artificial Intelligence Agent Chooses the Next Pharmaceutical Experiment? Pharmacophore. 2026;17(4):48-59. https://doi.org/10.51847/VWfLqBkl2y

Introduction

Artificial intelligence already supports target identification, molecular-property prediction, virtual screening, synthesis planning, biomarker analysis, and other activities distributed across drug discovery and development. Nevertheless, the breadth of these applications, the sophistication of their benchmarks, and the plausibility of generated outputs do not by themselves establish prospective pharmaceutical usefulness or routine decision readiness [1-3]. The accountability problem becomes more acute when an artificial intelligence system no longer provides only a prediction for human consideration but participates in selecting the next intervention in an experimental sequence. Choosing an experiment changes the material and informational state of a research program: compounds may be synthesized, biological materials consumed, instruments occupied, personnel exposed to hazards, and subsequent hypotheses conditioned by the resulting observations. Experiment selection is therefore

Corresponding Author: Sven Larsson; Department of Accountability in AI-Driven Experimentation, Faculty of Pharmacy, Swedish University of Agricultural Sciences, Uppsala, Sweden. E-mail: sven.larsson@slu.se

not merely another computational output. It is a consequential scientific act that combines epistemic judgment, resource allocation, risk acceptance, and authority to intervene.

The evidentiary basis for such choices is often less stable than an agent's confident recommendation suggests. Chemical and biological datasets may contain assay-specific artifacts, selective reporting, inconsistent endpoint definitions, historical design biases, and distributions that differ from the compounds, targets, formulations, or experimental conditions under consideration [4]. An agent can therefore process a large quantity of available data while still relying on evidence that is unsuitable for the proposed decision. This distinction between data availability and data suitability is central to accountability. When a recommendation fails because the underlying evidence was biased, incomplete, context-inappropriate, or poorly validated, responsibility cannot be understood solely by inspecting the final model output. It must also encompass who selected the data, defined the endpoint, accepted the validation design, established the operational objective, and permitted the recommendation to influence laboratory action.

The emergence of tool-using language-model systems makes this issue operational rather than hypothetical. Such systems can decompose chemical tasks, retrieve information, invoke external tools, propose synthesis-related actions, and coordinate multi-step workflows, thereby creating a pathway from generated language to chemical action [5]. Their apparent agency arises from the ability to maintain goals, choose among actions, call tools, interpret intermediate outputs, and adapt subsequent steps. Yet these capabilities do not make the system an independent scientific principal. The agent's choices remain conditioned by its prompt, objective function, available tools, training data, retrieved evidence, system permissions, and engineered constraints. What appears to be a unitary autonomous decision is frequently the visible endpoint of a distributed technical and organizational arrangement.

This review addresses the question of who is accountable when an artificial intelligence agent chooses the next pharmaceutical experiment. Its objective is not to determine legal liability or propose that contemporary agents possess moral responsibility. Instead, it develops a critical framework for separating operational agency, decision authority, answerability, and remedial responsibility across the experimental lifecycle. The review evaluates evidence concerning planning, tool use, experiment selection, physical execution, interpretation, traceability, override, and human control. It also distinguishes demonstrated component capability from benchmark performance, plausible utility, experimental confirmation, translational value, and routine pharmaceutical readiness. The article's original contribution is a proposed model of distributed but non-diffuse accountability: responsibility may be shared across actors and stages, but each consequential decision must remain connected to identifiable evidence obligations, authorization rights, monitoring duties, escalation routes, and mechanisms for correction.

Critical Review Method

This article used a purposive and iterative critical-review approach rather than a systematic-review protocol. The literature was assembled to examine the accountability implications of artificial intelligence systems that participate in planning, selecting, coordinating, or interpreting pharmaceutical experiments. The objective was not to identify every publication concerning artificial intelligence in pharmaceutical research, but to select sources capable of clarifying the article's central distinctions among operational agency, decision authority, answerability, and remedial responsibility.

The 33 references were organized across five intersecting evidence domains: pharmaceutical artificial intelligence and data quality; autonomous chemical experimentation and self-driving laboratories; tool-using artificial intelligence and experiment-selection systems; algorithmic accountability, ethics, and governance; and traceability, interpretability, human control, and contestability. Foundational reviews and widely cited conceptual papers were used to define established concerns, while primary demonstrations were included when they provided direct evidence of planning, tool use, robotic execution, closed-loop experimentation, or chemistry-workflow orchestration. Additional sources were retained when they clarified important limitations involving dataset bias, information leakage, synthesizability, objective misspecification, explanation fidelity, reproducibility, or organizational responsibility.

Sources were selected according to their direct relevance to one or more decision stages in the agentic experimental lifecycle: objective definition, evidence acquisition, planning, tool invocation, experiment ranking, authorization, physical execution, measurement appraisal, interpretation, monitoring, and response to failure. Preference was given to peer-reviewed studies, major critical reviews, and influential governance analyses that supported specific claims or exposed identifiable interpretation boundaries. Publications were not included solely because they discussed artificial intelligence ethics or pharmaceutical machine learning in general. Sources were excluded from the core synthesis when they did not address experiment-affecting decisions, accountability mechanisms, evidence quality, autonomous scientific workflows, or human and organizational control.

The selected literature was critically compared according to the type of evidence provided, the setting in which the reported capability was evaluated, the extent of prospective or physical validation, the reproducibility of the workflow, and the limits of the conclusions that could reasonably be drawn. Demonstrations of component capability were distinguished from evidence of routine pharmaceutical usefulness, autonomous authorization, regulatory acceptability, or improved scientific outcomes. The review should therefore be interpreted as a structured critical synthesis and conceptual framework rather than as a systematic, scoping, or exhaustive review.

Critical-Review Scope and Accountability Lens

The scope of this critical review is defined by the transition from computational recommendation to experiment-affecting action. It includes systems that generate or rank experiments, invoke chemical or biological tools, coordinate automated equipment, interpret resulting measurements, or update subsequent experimental choices. It excludes the broader question of whether all pharmaceutical artificial intelligence is ethically acceptable and does not treat ordinary predictive modeling as equivalent to agentic experimentation. The accountability lens begins from the principle that responsibility should connect the anticipated benefits and harms of artificial intelligence to identifiable practices of design, authorization, oversight, and response [6]. For this review, operational agency means that a system causally participates in selecting or coordinating actions; decision authority means the recognized power to permit an experiment; answerability means the obligation and capacity to explain and justify the decision; and remedial responsibility means the duty to investigate failure, correct systems, compensate for harm where appropriate, and prevent recurrence. These constructs are analytically separated because an agent may possess operational agency without possessing the other three attributes.

The review does not treat the recurrence of accountability, transparency, fairness, safety, or explicability in ethics guidance as evidence that these principles have been implemented effectively. Comparative analysis of artificial intelligence ethics guidelines shows substantial convergence around high-level values while also revealing variation in their interpretation and operationalization [7]. A pharmaceutical laboratory can therefore claim adherence to responsible-artificial-intelligence principles while leaving unresolved who is permitted to approve an agent-selected experiment, what information that person must receive, how uncertainty is represented, when execution must be stopped, and who investigates an irreproducible or unsafe outcome. The critical question is not whether accountability appears in a governance statement, but whether the experimental system creates a traceable chain from evidence and model design to recommendation, authorization, execution, interpretation, and corrective action.

The literature was interpreted according to its relevance to these decision functions rather than according to artificial intelligence terminology alone. The reviewed sources included critical pharmaceutical reviews, bounded demonstrations of autonomous chemistry, analyses of algorithmic accountability, and studies examining mechanisms for translating governance principles into practical controls. Publicly available ethics tools show that operationalization may involve impact assessment, documentation, model evaluation, stakeholder review, oversight procedures, and mechanisms for remediation, although their coverage and maturity remain uneven [8]. The present review uses these categories as appraisal dimensions rather than assuming that the tools themselves are validated solutions. It asks whether each source supports a specific claim, what type of evidence is provided, where validation occurred, what bias or reproducibility concerns remain, and what cannot legitimately be concluded. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop critical-review scope and accountability lens within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Critical-review scope and accountability lens

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Pharmaceutical artificial intelligence capability	Critical reviews of machine learning across drug discovery and development	Computational methods can assist multiple pharmaceutical tasks	Strong support for breadth of application; variable support for prospective impact	Selective benchmarks, retrospective evaluation, heterogeneous datasets, and inconsistent comparators	Evidence that computational assistance improves prospective experimental decisions across realistic programs
Pharmaceutical data suitability	Analyses of chemical and biological data generation, labeling, and validation	Large datasets can support modeling but may encode assay, selection, and historical biases	Strong conceptual and methodological support	Dataset shift, endpoint ambiguity, duplicated chemistry, missing negative results, and context-dependent labels	Decision-specific standards for determining whether available data are suitable for an agent-selected experiment
Operational agency	Demonstrations in which artificial intelligence plans tasks, calls tools, or coordinates actions	Agents can participate causally in multi-step scientific workflows	Direct support in bounded demonstrations	Bespoke tools, narrow tasks, prompt sensitivity, incomplete fault characterization, and limited external replication	Prospective evidence that operational agency remains reliable under changing pharmaceutical conditions
Accountability principles	Normative frameworks and comparative analyses of artificial intelligence ethics guidance	Accountability, explicability, safety, justice, and human autonomy are recurring governance requirements	Strong support for principle-level convergence	Principles are interpreted differently and may be adopted symbolically	Role-specific duties, decision rights, and enforcement mechanisms for pharmaceutical experimentation

Translation from principles to practice	Reviews of ethics tools, impact assessments, documentation methods, and governance procedures	Practical instruments may connect ethical commitments to design and oversight activities	Moderate support for availability of tools; limited support for effectiveness	Fragmented methods, self-assessment bias, inconsistent adoption, and weak outcome evaluation	Evidence that governance instruments prevent unsafe, wasteful, or irreproducible experimental decisions
Scientific authority	Professional judgment, organizational governance, and research-quality requirements	Experiment authorization involves expertise, institutional authority, and acceptance of risk	Strong conceptual support	Authority may be nominal when reviewers lack information, time, competence, or genuine power to refuse	Minimum conditions for meaningful rather than ceremonial human authorization
Accountability boundaries	Critical comparison of prediction, explanation, experimental confirmation, and deployment	Operational capability does not establish moral agency, legal liability, scientific validity, or routine readiness	Strong support for maintaining conceptual distinctions	Anthropomorphic language can obscure the actual distribution of control and responsibility	Empirical testing of proposed accountability structures across real pharmaceutical workflows

Agent Roles in Planning, Tool Use, Experiment Selection, and Interpretation

Across self-driving laboratory architectures, robotic reactivity search, and artificial-intelligence-informed flow synthesis, the central operational pattern is a closed loop linking objective definition, experiment selection, physical execution, measurement, and model updating [9-11]. Within this loop, "the agent" should not be treated as a single undifferentiated actor. Planning involves decomposing a scientific objective into candidate actions and dependencies. Tool use involves choosing external databases, predictors, synthesis planners, analytical routines, or laboratory interfaces and specifying how they are invoked. Experiment selection involves ranking or choosing an intervention under an explicit or implicit objective. Execution coordination translates a proposed action into instrument instructions, materials, timing, and procedural constraints. Interpretation assigns meaning to the resulting measurements and determines whether the evidence changes the next decision. These roles may be implemented by one software architecture, several interacting models, or a broader automation stack, but their accountability implications differ because each role introduces distinct opportunities for error and control.

Embodied autonomy adds dependencies that are absent from purely computational recommendation. A mobile robotic chemist, for example, must navigate a physical environment, access instruments, handle samples, schedule operations, and adapt to laboratory conditions while remaining coupled to an optimization procedure [12]. Such a platform demonstrates that an algorithmic recommendation can be converted into sustained experimental action, but it also shows why model-level assessment is insufficient. An experiment may fail because of navigation error, sample contamination, incorrect instrument configuration, unstable environmental conditions, unrecognized equipment drift, or a mismatch between software representation and physical reality. Accountability must therefore follow the complete causal chain. The developer of an optimization model is not automatically responsible for a mechanical handling fault, yet the organization cannot treat hardware, software, and procedure as unrelated when their integration creates the autonomous capability. Before execution, the system must establish which component generated the proposed action, which constraints were checked, what physical assumptions were made, and which person or function had authority to halt the sequence.

Tool use creates a second layer of distributed dependence. Chemistry-tool augmentation can improve language-model performance by enabling access to calculations, databases, structure-processing functions, and domain-specific software, but the reliability of the resulting recommendation depends on tool selection, invocation parameters, versioning, output validity, and interpretation [13]. An agent may choose an appropriate tool but use an invalid parameter; retrieve an accurate value but attach it to the wrong compound; obtain a valid retrosynthetic route but overlook unavailable reagents or incompatible conditions; or combine outputs from individually credible tools in a scientifically invalid way. Consequently, a tool call should be treated as a scientific operation requiring provenance rather than as an invisible extension of model reasoning. A defensible record should identify the tool, version, input, parameters, returned output, error state, transformation steps, and the basis on which the agent used the result. Without this record, later explanation may reproduce a persuasive narrative while failing to reconstruct what actually caused the experimental recommendation.

Interpretation is the most easily overlooked role because it often occurs after measurements have been generated and may be described as summarization rather than decision-making. In a closed experimental loop, however, interpretation determines whether a result is treated as confirmation, contradiction, anomaly, technical failure, or evidence for a revised hypothesis. This classification directly shapes the next experiment. The review therefore proposes a role decomposition in which planning, tool use, selection, execution coordination, measurement appraisal, and interpretation are evaluated separately even when they are implemented within one agentic interface. This decomposition is conceptual and has not been validated as a universal architecture. Its purpose is to prevent the language of autonomous agency from concealing the actual locations of epistemic dependence and human control. Under this model, an agent may generate and prioritize experimental options, but progression

to execution requires an independently identifiable authorization act based on evidence appropriate to the experiment's scientific value, feasibility, uncertainty, and risk.

Responsibility Gaps across Developers, Scientists, Organizations, and Agents

An artificial intelligence agent can contribute causally to an experimental decision without becoming an independent bearer of scientific, professional, or institutional accountability. Algorithms are embedded in organizational arrangements that determine their objectives, permissible actions, evidence sources, interfaces, and consequences. Treating the agent as the accountable actor would therefore obscure the prior human decisions that made its conduct possible. Algorithmic accountability instead requires examination of who designed the decision process, who selected the criteria that the system optimizes, who approved its use, and who benefits from or bears the consequences of its recommendations [14]. In pharmaceutical experimentation, the agent should be represented as an operational component whose actions are attributable and auditable, but not as a substitute for the developers, scientists, managers, and institutions that possess authority and continuing duties.

Scientists retain responsibility when they rely on agent-generated recommendations, although that responsibility must be interpreted realistically. A scientist cannot meaningfully supervise a system when its evidence is inaccessible, its tool calls cannot be reconstructed, or organizational pressures make refusal impractical. Professional oversight therefore requires more than a final approval click. The reviewer must understand the scientific objective, relevant uncertainties, experimental hazards, plausible alternatives, and limits of the evidence supporting the recommendation. Ethical analysis of algorithmic decision-making in healthcare similarly indicates that professional duties persist when computational systems contribute to high-stakes judgments [15]. For pharmaceutical science, this principle implies that scientists remain responsible for evaluating whether an experiment is scientifically warranted and appropriately controlled, while organizations remain responsible for ensuring that reviewers have adequate competence, information, time, and authority.

Organizations possess additional non-delegable responsibilities because they define the environment in which agentic decisions are produced and acted upon. These duties include determining acceptable use cases, validating data and models, controlling tool and instrument access, establishing authorization thresholds, monitoring performance, documenting deviations, and responding to harmful or irreproducible outcomes. Lifecycle governance models developed for healthcare artificial intelligence similarly allocate oversight across data acquisition, development, validation, implementation, monitoring, and response [16]. Such models are not validated pharmaceutical governance standards, but they reveal why accountability cannot be concentrated only at the moment of experiment approval. A defective recommendation may originate in outdated data, inappropriate objectives, untested software updates, poorly integrated equipment, insufficient training, or an institutional decision to tolerate uncertainty without appropriate safeguards.

Responsibility gaps commonly emerge at interfaces. Developers may assume that scientists will recognize model limitations; scientists may assume that validation teams have tested relevant failure modes; organizations may assume that an audit trail establishes accountability; and oversight functions may intervene only after harm has occurred. Mapping reviews of artificial intelligence ethics in healthcare demonstrate that relevant concerns are distributed across data practices, technical development, professional use, organizational implementation, and wider social effects [17]. The corresponding pharmaceutical interpretation is that accountability must be distributed without becoming diffuse. Every lifecycle stage should have a named decision owner, a defined evidence obligation, an escalation route, and a record of what was accepted or rejected. The agent may be described as selecting an experiment operationally, but responsibility for permitting that selection to influence laboratory action remains attached to human and institutional roles.

Evidence, Traceability, Override, and Contestability Requirements

Evidence requirements should be proportional to the consequence of the proposed experiment. An agent that suggests a literature hypothesis may require less control than one that authorizes synthesis, biological testing, scale-up, or work involving hazardous materials. Explainable artificial intelligence may help researchers identify molecular features, relationships, or model considerations relevant to a recommendation, but an explanation is not equivalent to evidence that the model is correct, causal, or experimentally useful [18]. A scientifically defensible recommendation should therefore be accompanied by the data sources used, their relevance to the current context, the model and tool versions involved, uncertainty information, known applicability limits, feasible alternatives, and the reason the proposed experiment is expected to reduce an important uncertainty or advance a defined objective.

Interpretability and explanation must also be distinguished. A post-hoc explanation can provide a simplified account of an opaque system while failing to represent the mechanism that actually determined its output. In high-stakes settings, reliance on such explanations may create misplaced confidence, particularly when users cannot determine whether the explanation is faithful, stable, or sensitive to small changes in the input [19]. Where feasible, experiment-selection systems should use transparent objectives, explicit constraints, interpretable decision stages, and directly inspectable evidence. Where opaque components remain necessary, the burden of validation should increase rather than be displaced onto a persuasive narrative. The appropriate question is not whether the agent can produce a reason, but whether an independent reviewer can reconstruct and challenge the evidentiary path from inputs to recommended action.

Traceability requires more than preserving the final prompt and response. Data leakage, inappropriate partitions, repeated entities across evaluation sets, preprocessing performed before data separation, and other workflow errors can produce

apparently strong but irreproducible scientific claims [20]. An end-to-end record should therefore include data provenance, exclusions, transformations, model selection, parameter settings, tool calls, intermediate outputs, constraint checks, uncertainty estimates, human interventions, execution conditions, measurements, and subsequent interpretation. The record should be sufficiently detailed to reproduce not merely the text of the recommendation but the process that generated it. Version control is essential because an identical prompt can produce a different decision when a model, database, tool, or laboratory interface has changed. Traceability is consequently both a scientific reproducibility requirement and an accountability mechanism.

Contestability extends traceability by ensuring that a recommendation can be challenged before and after execution. Current approaches to explainable artificial intelligence may create reassurance without delivering fidelity, causal understanding, or actionable safety information [21]. A contestable system must therefore permit qualified individuals to question the objective, inspect supporting evidence, request alternative analyses, identify excluded constraints, pause execution, escalate disagreement, and initiate independent review. Override must remain technically available and organizationally legitimate. An operator who can press a stop button but is discouraged from doing so does not possess meaningful control. Similarly, a post-experiment appeal mechanism is inadequate when an unsafe action cannot be interrupted. Contestability should be designed as a sequence of review, challenge, pause, escalation, and remedy functions rather than as a single explanatory interface.

Failure Cases Involving Unsafe, Wasteful, or Irreproducible Experiments

A first failure pathway begins with benchmark performance that does not represent the intended experimental setting. Ligand-based classification benchmarks may reward recognition of closely related molecules or dataset-specific patterns rather than generalization to genuinely different chemistry [22]. If such a model is incorporated into an agent, the system may confidently prioritize compounds that resemble known examples while failing on novel targets, scaffolds, assay conditions, or chemical spaces. The immediate consequence may be wasted synthesis and testing, but the scientific consequence can be broader: repeated low-value experiments may reinforce a false hypothesis, distort portfolio priorities, and generate additional training data from a biased selection process. Accountability therefore requires reviewers to ask what the benchmark measures, what chemical distinctions exist between training and proposed experiments, and whether performance remains credible under decision-relevant separation.

A second failure pathway arises when models exploit biases that are easier to learn than the chemical or biological relationships they are intended to represent. Structure-based virtual-screening datasets can contain systematic differences between active and inactive examples that enable models to achieve strong apparent performance without learning transferable interaction patterns [23]. An agent built on such models may select experiments for reasons unrelated to the proposed mechanism, even when its predictions appear technically well calibrated within the original dataset. Bias control should consequently precede experimental escalation. Relevant checks include alternative data splits, matched controls, artifact testing, applicability analysis, comparison with simple baselines, and inspection of whether the model remains informative when obvious shortcuts are removed. Failure to conduct these checks transfers hidden benchmark assumptions into physical experimentation.

A third failure pathway concerns objective gaming in generative molecular design. Benchmark frameworks demonstrate that model behavior and rankings vary substantially across distribution-learning and goal-directed tasks, indicating that apparent success is dependent on the selected objective and evaluation structure [24]. An agent may optimize a numerical proxy while generating compounds that are chemically repetitive, unstable, reactive, toxicologically concerning, or irrelevant to the therapeutic problem. Multiobjective optimization reduces but does not eliminate this risk because the included objectives may remain incomplete, uncertain, or mutually incomparable. Pharmaceutical usefulness cannot be inferred from novelty, predicted activity, or a composite score alone. Before an agent-generated proposal advances, the objective should be examined for missing dimensions, exploitable shortcuts, and conflicts with medicinal chemistry, pharmacology, formulation, safety, and development requirements.

A fourth failure pathway occurs when computational plausibility is mistaken for experimental feasibility. Molecules proposed by generative models can satisfy design objectives while remaining difficult to synthesize according to computational assessments and retrosynthetic analysis [25]. Even an apparently feasible route may rely on unavailable starting materials, hazardous transformations, fragile intermediates, unsuitable selectivity, impractical purification, or conditions incompatible with the available laboratory. Synthesizability is also distinct from developability: a compound that can be prepared may still be unsuitable because of stability, exposure, formulation, toxicity, or manufacturability constraints. The accountability failure arises when these distinctions are collapsed into one automated score. Agentic systems should instead pass through staged feasibility and safety reviews, with explicit reasons recorded whenever a proposal is rejected, modified, or allowed to progress.

Models of Distributed Accountability and Human Control

The most defensible model of human control does not position the scientist as a passive monitor of an otherwise autonomous process. High-stakes biomedical artificial intelligence is commonly framed as most useful when human and machine capabilities are complementary, when responsible practices extend across the system lifecycle, and when high levels of automation coexist with strong human and organizational control [26-28]. Applied to pharmaceutical experimentation, this means that agents may search large decision spaces, compare candidate actions, coordinate tools, and detect patterns, while humans retain responsibility for defining scientifically meaningful objectives, evaluating contextual evidence, authorizing consequential interventions, interpreting unexpected findings, and responding to failure. Control should be allocated according to capability, but accountability should remain attached to actors capable of justification, judgment, and remedy.

Hybrid intelligence provides a useful model for task allocation because it treats human and machine contributions as components of a coordinated problem-solving system rather than as competitors for total autonomy [29]. However, complementarity does not arise automatically. It requires explicit interfaces for information exchange, uncertainty communication, disagreement, feedback, and learning. The scientist must be able to identify what the agent considered, what it ignored, and how alternative assumptions would change the recommendation. The agent's role should also vary with evidence maturity. It may autonomously perform reversible, low-risk computational operations while requiring progressively stronger review for resource-intensive, hazardous, biologically consequential, or strategically irreversible experiments. Such graduated authority is proposed here as a governance principle rather than as an empirically validated universal scale.

A distributed accountability chain should connect each agentic function to an identifiable human or institutional duty. Developers and data stewards are responsible for documenting design assumptions, data provenance, known limitations, and foreseeable misuse. Model validators are responsible for testing decision-relevant performance, calibration, bias, and failure modes. Tool and pipeline integrators are responsible for interface integrity, version control, permissions, and error handling. Scientists are responsible for evaluating scientific relevance and experimental justification. Project and organizational decision owners are responsible for resource commitments and accepted risk. Independent quality, safety, ethics, or governance functions are responsible for challenge, escalation, incident review, and corrective action. **Figure 1** presents the distributed accountability chain for agentic experiments, showing how the article's key components and boundaries are connected within models of distributed accountability and human control.

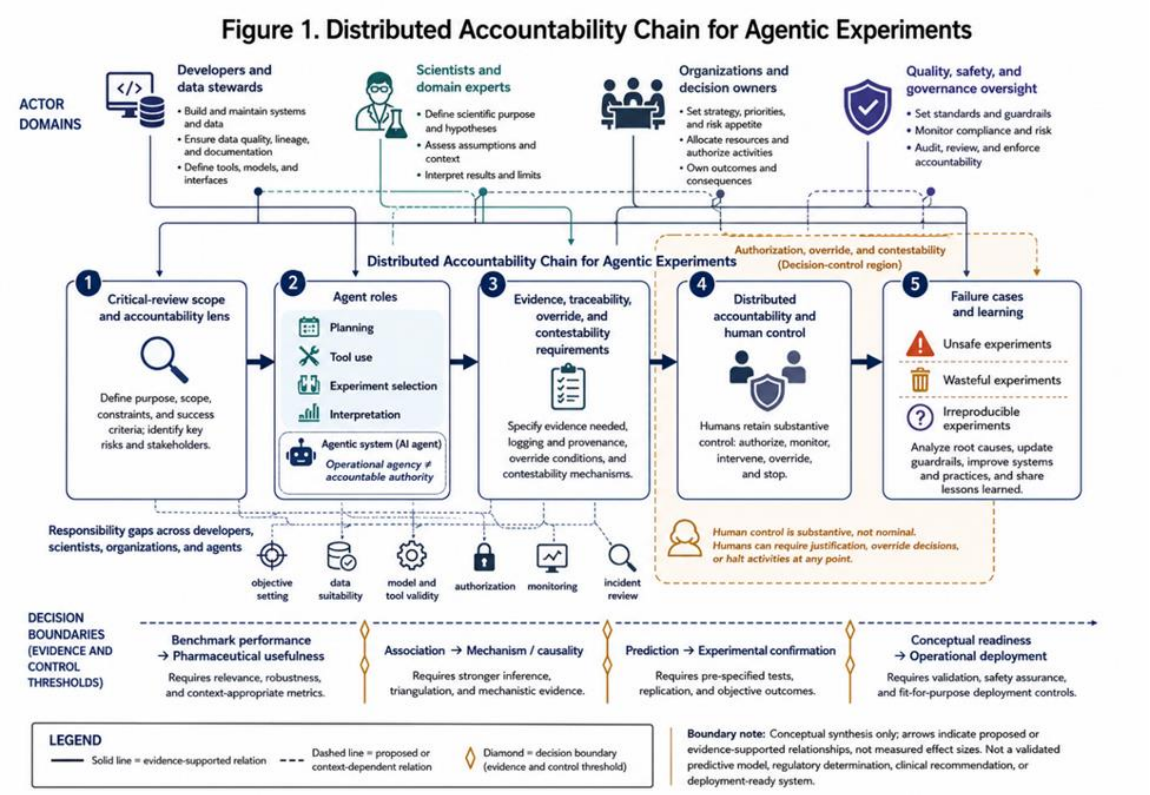

Figure 1. Distributed Accountability Chain for Agentic Experiments

The figure is an original conceptual synthesis that organizes the article's central contribution across critical-review scope and accountability lens, agent roles in planning, tool use, experiment selection, and interpretation, agent roles in planning, tool use, experiment selection, responsibility gaps across developers, scientists, organizations, and agents. 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.

Distributed accountability must not become accountability dilution. Assigning responsibilities to several actors can create the appearance that everyone is involved while leaving no one answerable for the final decision. The proposed chain therefore separates contribution from ownership. Multiple actors may contribute evidence, models, tools, and review, but each experiment should have one identifiable authorization owner and one organizational function responsible for ensuring that the authorization process is adequate. Incident review should examine both local error and system design: why the recommendation was produced, why controls failed to stop it, how the outcome was interpreted, and whether incentives or workload weakened oversight. This model does not determine legal liability and should not be treated as a validated governance architecture. It is an explanatory framework for preventing operational agency from being mistaken for accountable authority.

Research and Governance Priorities for Agentic Pharmaceutical Science

The first research priority is to evaluate complete agentic systems rather than isolated model components. Reviews of autonomous chemical discovery show substantial progress in planning, optimization, robotics, measurement, and analysis, but also reveal heterogeneous capabilities and reliance on bespoke integration [30]. Future studies should evaluate whether an agent improves the quality of experimental decisions relative to expert practice and established computational workflows. Relevant outcomes include whether selected experiments are informative, feasible, reproducible, safe, and appropriately prioritized. Prospective comparisons should record rejected recommendations, human modifications, failures, stopping decisions, and negative results rather than reporting only successful demonstrations. Evaluation should also test distribution shift, incomplete data, conflicting tools, instrument faults, and changes in objectives, because these conditions are intrinsic to pharmaceutical research rather than exceptional disturbances.

The second priority is interoperable infrastructure for traceability and control. Autonomous-discovery research identifies standardized representations, compatible hardware and software, robust decision algorithms, and complete workflow evaluation as continuing needs [31]. Pharmaceutical systems additionally require a shared experiment-decision record linking objectives, evidence, models, tool calls, permissions, authorization, execution conditions, analytical quality, interpretation, and outcome. Such records should be portable across laboratory information systems and should preserve failed and interrupted workflows. Interoperability is not merely an efficiency goal. Without it, organizations cannot reconstruct why an agent selected an experiment, compare performance across sites, detect drift after software changes, or determine whether a failure originated in data, modeling, orchestration, equipment, or human review.

The third priority is investment in the socio-technical environment surrounding autonomy. Community analyses of autonomous experimentation emphasize that progress depends on coordinated development of hardware, software, data infrastructure, algorithms, standards, and workforce capabilities [32]. Pharmaceutical institutions should add governance competence to this list. Scientists need training in uncertainty, data provenance, model limitations, and automation bias; developers need exposure to laboratory hazards, assay variability, medicinal chemistry, and development constraints; and oversight personnel need access to technically interpretable records. Organizations should test escalation and incident-response procedures through simulation before granting systems broader permissions. Independent evaluation, adversarial testing, and cross-site replication should be incorporated progressively as agents move from computational recommendation toward physical execution.

The fourth priority is to validate agentic decision-making against the realities of pharmaceutical experimentation. High-throughput experimentation in pharmaceutical synthesis demonstrates that automation can address difficult chemical problems while remaining dependent on expert experimental design, analytical quality, contextual interpretation, and iterative judgment [33]. Agentic systems should therefore be integrated with, rather than presumed to supersede, established experimental disciplines. Readiness should be assessed through staged evidence: conceptual plausibility, computational evaluation, prospective recommendation, bounded execution, independent replication, workflow integration, and continued monitoring. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop research and governance priorities for agentic pharmaceutical science within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Who Is Accountable When an Artificial Intelligence Agent Chooses the Next Pharmaceutical

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Agent-supported hypothesis and experiment generation	Literature, internal project data, molecular representations, assay histories, and explicit research objectives	Expert comparison, provenance review, alternative-hypothesis testing, and documentation of rejected options	Emerging computational capability	Generated rationales may be fluent but evidentially incomplete or dependent on hidden assumptions	Appropriate for structured recommendation under review; not sufficient for autonomous authorization
Tool-augmented chemical agents	Databases, calculators, predictive models, synthesis planners, and laboratory software	Tool-level validation, parameter logging, version control, error injection, and independent result verification	Demonstrated in bounded chemistry tasks	Reliability depends on external tools, orchestration, permissions, and result interpretation	Treat every tool call as a traceable scientific operation
Algorithmic experiment selection	Candidate experiments defined by optimization, active-learning, or decision objectives	Prospective comparison with experts and standard workflows; evaluation under distribution shift	Demonstrated in selected closed-loop systems	Objectives may omit scientific value, safety, feasibility, or strategic constraints	Selection authority should expand only with decision-relevant prospective evidence

Robotic or automated execution	Physical instruments, samples, environmental conditions, procedures, and scheduling constraints	Fault testing, recovery assessment, cross-site replication, contamination control, and execution-quality monitoring	Bounded laboratory demonstrations	Bespoke hardware and local tacit knowledge limit portability	Physical execution requires additional safety and accountability gates beyond model validation
Explainability and scientific interpretation	Model outputs, chemical features, biological evidence, and experimental measurements	Fidelity, stability, user relevance, causal caution, and comparison with transparent alternatives	Active methodological development	Explanations may be persuasive without being faithful, causal, or decision-improving	Use explanations as review artifacts, not proof of validity
End-to-end traceability	Data, prompts, models, tools, permissions, human actions, experimental conditions, and outcomes	Reconstruction testing, audit completeness, reproducibility exercises, and version-difference analysis	Recognized requirement with uneven implementation	Records may be incomplete, fragmented, or too complex for timely review	Traceability is necessary for accountability but does not alone establish adequate oversight
Human authorization and override	Experiments with varying scientific, resource, safety, and strategic consequences	Evaluation of reviewer competence, information access, workload, refusal authority, and response time	Strong conceptual support; limited direct outcome evidence	Human involvement may be ceremonial or vulnerable to automation bias	Define substantive authorization rights and escalation triggers rather than relying on a generic human-in-the-loop label
Distributed organizational governance	Development, validation, deployment, monitoring, incident review, and corrective action	Role-assignment exercises, governance audits, simulated failures, and prospective process evaluation	Proposed and partially operationalized across adjacent high-stakes domains	Responsibilities may be fragmented or jurisdiction-dependent	Allocate lifecycle duties while preserving one identifiable authorization owner for each experiment
Pharmaceutical translation	Medicinal chemistry, synthesis, assays, formulation, safety, developability, and portfolio decisions	Progressive validation from computational testing to bounded execution and independent replication	Component-level evidence; limited routine agentic readiness	Pharmaceutical usefulness requires more than predicted activity or synthetic feasibility	Routine autonomous experiment authorization is not established

Figure 2 presents the evidence, validation, and translation boundary observatory for who is accountable when an artificial intelligence, showing how the article's key components and boundaries are connected within research and governance priorities for agentic pharmaceutical science.

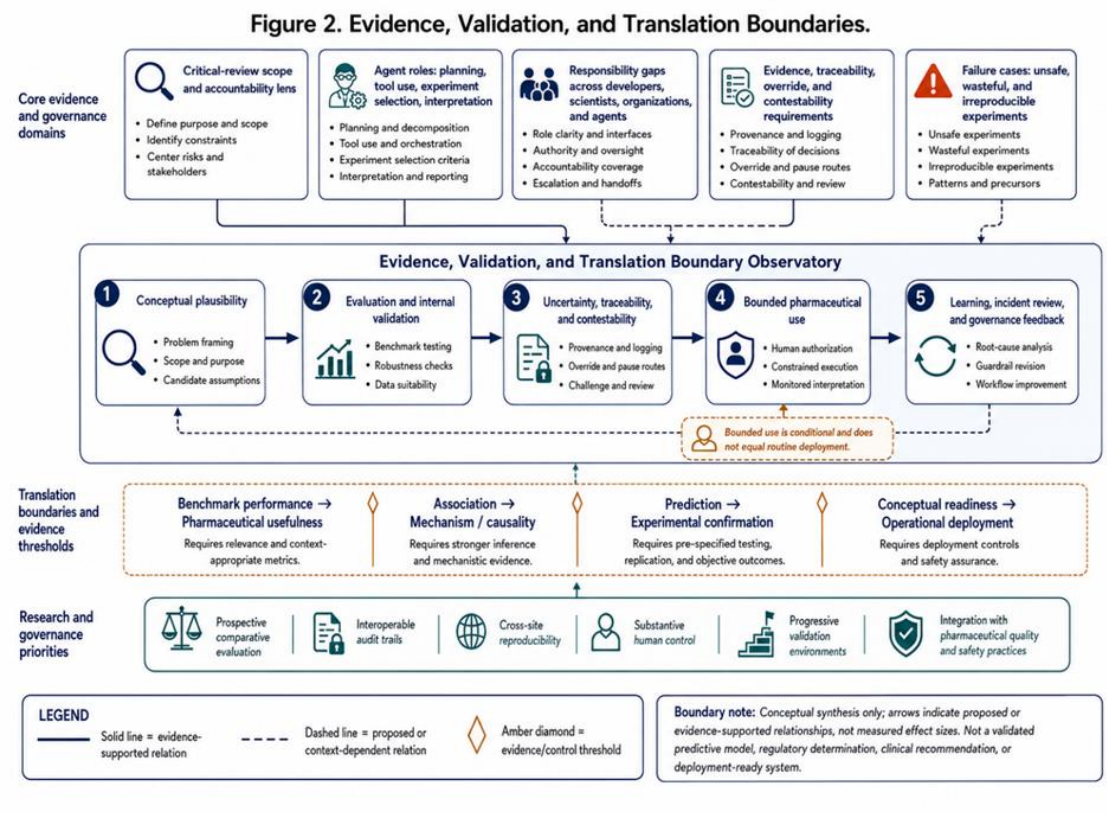

Figure 2. Evidence, Validation, and Translation Boundaries

The figure is an original conceptual synthesis that shows how claims move from conceptual plausibility through evaluation, uncertainty assessment, and bounded pharmaceutical use. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Conclusion

When an artificial intelligence agent chooses the next pharmaceutical experiment, accountability does not transfer to the agent merely because its output is operationally consequential. The accountable system is the broader socio-technical arrangement that defines the objective, supplies the evidence, validates the models and tools, sets permissions, reviews the recommendation, authorizes execution, interprets the result, and responds to failure. This review's central contribution is a proposed separation of operational agency from decision authority, answerability, and remedial responsibility. It further proposes that accountability should be distributed across the experimental lifecycle without becoming diffuse: each consequential experiment should remain connected to identifiable evidence obligations, control rights, authorization ownership, escalation procedures, and corrective duties. Current evidence demonstrates bounded capabilities in planning, tool use, experiment selection, robotic execution, and interpretation, but it does not establish routine readiness for autonomous pharmaceutical experiment authorization. Explanations, confidence scores, audit records, and human presence are each insufficient when treated in isolation. Responsible agentic pharmaceutical science requires their integration with data-suitability assessment, prospective validation, substantive override, contestability, cross-site reproducibility, and existing experimental governance. The proposed tables and figures organize these requirements conceptually; they do not constitute validated legal, regulatory, clinical, or deployment frameworks.

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, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
4. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
5. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0
6. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach (Dordr). 2018;28(4):689-707. doi:10.1007/s11023-018-9482-5
7. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-99. doi:10.1038/s42256-019-0088-2
8. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics. 2020;26(4):2141-68. doi:10.1007/s11948-019-00165-5
9. 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
10. 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
11. 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
12. 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
13. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8
14. Martin K. Ethical implications and accountability of algorithms. J Bus Ethics. 2019;160(4):835-50. doi:10.1007/s10551-018-3921-3
15. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586
16. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27(3):491-7. doi:10.1093/jamia/ocz192
17. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med. 2020;260:113172. doi:10.1016/j.socscimed.2020.113172
18. 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
19. 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
20. 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
21. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
22. 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
23. 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
24. 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
25. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23. doi:10.1021/acs.jcim.0c00174
26. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
27. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6

28. Shneiderman B. Human-centered artificial intelligence: Reliable, safe and trustworthy. Int J Hum Comput Interact. 2020;36(6):495-504. doi:10.1080/10447318.2020.1741118
29. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid intelligence. Bus Inf Syst Eng. 2019;61(5):637-43. doi:10.1007/s12599-019-00595-2
30. Coley CW, Eyke NS, Jensen KF. Autonomous discovery in the chemical sciences. Part I: Progress. Angew Chem Int Ed Engl. 2020;59(52):22858-93. doi:10.1002/anie.201909987
31. Coley CW, Eyke NS, Jensen KF. Autonomous discovery in the chemical sciences. Part II: Outlook. Angew Chem Int Ed Engl. 2020;59(52):23414-36. doi:10.1002/anie.201909989
32. Stach EA, DeCost BL, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26. doi:10.1016/j.matt.2021.06.036
33. Krska SW, DiRocco DA, Dreher SD, Shevlin M. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. Acc Chem Res. 2017;50(12):2976-85. doi:10.1021/acs.accounts.7b00428