ORIGINAL RESEARCH **Open access**

A Conceptual Theory of Model-Science Interface: Where AI Outputs Become Experimental Inputs

Wei Chen[1*], Li Zhang[1]

Abstract

In the rapidly evolving field of artificial intelligence for materials science, research has overwhelmingly emphasized the development of predictive models, active learning algorithms, and inverse design strategies to accelerate the identification of novel functional materials. Yet, the critical boundary at which these computational outputs become experimental inputs—the model-science interface—remains largely ignored and treated as an unproblematic transmission step. Existing literature on self-driving laboratories and autonomous experimentation systems, while advancing integrated platforms for clean energy discovery and closed-loop workflows, assumes that model predictions, uncertainty estimates, and experimental recommendations flow seamlessly into synthesis protocols, characterization decisions, and iterative loops without significant distortion or loss. This paper proposes the model-science interface as a distinct object of study, worthy of its own conceptual framework rather than being subsumed under broader discussions of automation or machine learning. By formalizing the interface as the active zone of translation between algorithmic intelligence and empirical practice, the framework distinguishes it from upstream modeling or downstream execution phases, thereby enabling systematic analysis of its internal dynamics. The key concepts articulated herein include a typology of interface operation modes differentiated along dimensions of autonomy and stakes, a detailed examination of information transformations that occur when AI outputs cross into experimental inputs—including preservation of core predictions, loss of contextual nuance, addition of laboratory constraints, and potential distortion through interpretation—and the introduction of “interface fidelity” as a conceptual variable that quantifies the quality of this transition across multiple dimensions. These elements, which build directly upon foundational accounts of autonomous chemical experiments and minimal working examples for self-driving laboratories, provide a vocabulary and set of distinctions for diagnosing interface failure modes that can undermine the overall efficacy of materials discovery pipelines. The framework draws upon foundational ideas in autonomous experimentation while elevating the interface itself as the locus of negotiation between computational promise and physical reality. Ultimately, adopting an interface-aware perspective carries profound implications for materials AI practice. It encourages researchers to design interfaces with intentionality, to report interface specifications alongside model performance, and to study information dynamics explicitly, thereby realizing the full potential of self-driving laboratories for accelerating the discovery of materials for clean energy, piezoelectrics, and beyond. This conceptual contribution thus bridges the persistent gap between model sophistication and experimental impact, fostering more accountable, efficient, and robust autonomous materials research ecosystems.

Keywords Active learning, Self-driving laboratories, Model-science interface, Autonomous experimentation, Information transformation, Interface fidelity

*Correspondence:
Wei Chen
wei.chen@outlook.com
[1] Department of Data-Driven Materials Science, School of Materials Engineering, Peking University, Beijing, China

Introduction

The field of artificial intelligence for materials science has witnessed substantial progress through the development of sophisticated computational models capable of predicting material properties, guiding inverse design [1], and supporting active learning strategies [2-7] that iteratively refine hypotheses. Yet, as highlighted in reviews of smart automation for clean energy materials [2], the literature has concentrated primarily on the accuracy and efficiency of these models while paying comparatively little theoretical attention to the subsequent stages of the discovery pipeline. The “model-science interface” emerges as the critical yet undertheorized juncture [5], where AI-generated outputs—such as property predictions, uncertainty quantifications, ranked synthesis targets, or proposed characterization sequences—must be translated into concrete experimental inputs that drive physical laboratory actions. This transition is not a trivial handoff but a complex boundary layer that shapes the ultimate success or failure of autonomous and semi-autonomous research systems [3].

Existing accounts of self-driving laboratories and closed-loop discovery [6-14] often describe integrated workflows in which data flow continuously from experiment to model and back again, yet they leave the precise mechanisms of the return leg—the conversion of model recommendations into executable protocols—implicit and unexamined. For instance, discussions of orchestration platforms [6] emphasize the coordination of hardware and software components but rarely dissect how probabilistic model outputs are rendered into deterministic action commands or how uncertainties influence the selection of next experiments. Similarly, efforts aimed at minimal working examples for self-driving laboratories [9] implicitly rely on well-functioning interfaces without analyzing them as independent objects of inquiry. The result is a persistent black-box quality in current materials AI practice, where assumptions of perfect or near-perfect information transmission persist despite the known challenges of laboratory constraints, equipment variability, and the inherent messiness of real-world experimentation [8, 11].

This neglect of the interface leads to several practical and conceptual limitations. Workflows frequently treat model outputs as directly actionable, an assumption that can break down under conditions of high uncertainty or multi-objective optimization scenarios common in functional materials design [15-18]. Consequently, opportunities to optimize the handoff process are missed, potentially reducing the overall throughput and reliability of discovery campaigns [12, 13]. Furthermore, without a dedicated vocabulary for discussing interface dynamics, it becomes difficult to compare alternative implementations of autonomous systems or to diagnose inefficiencies arising from mismatched levels of model confidence and experimental decisiveness. The present work addresses this shortfall by proposing a conceptual theory of the model-science interface that positions it as the central boundary mediating computational intelligence and empirical practice in materials research.

The importance of theorizing this interface becomes particularly evident when considering recent advances toward fully autonomous materials research [5, 19-23]. Descriptions of hierarchical active learning for nonequilibrium phase diagrams [19] or Bayesian approaches for on-the-fly discovery [7, 24, 25] illustrate the power of iterative loops. Yet, the efficacy of such systems ultimately depends on how effectively the interface translates model-generated recommendations into prioritized experimental queues or adaptive synthesis parameters. In the context of universal self-driving platforms [14], the interface layer determines whether high-dimensional optimization results translate into feasible laboratory instructions or whether critical contextual information is lost along the way. By bringing the model-science interface into explicit focus, this manuscript offers a unifying lens through which to view the diverse contributions to autonomous experimentation systems [3, 21], self-driving laboratories [4, 22], and data-driven experimental design [18].

In this study, the framework moves systematically from problem identification—where the interface is treated as invisible—to concept introduction, typology development, and transformation analysis. Ultimately, an interface-centric perspective promises to enhance the scientific accountability, decision quality, and overall impact of AI-driven materials discovery, transforming implicit assumptions into deliberate design choices [10, 26-28].

Defining the Model-Science Interface

The model-science interface constitutes a foundational yet previously informalized element within contemporary materials discovery ecosystems. To advance theoretical

understanding, it is necessary to delineate this concept with clarity and precision, recognizing that it functions not as a passive conduit but as an active zone of negotiation between algorithmic outputs and empirical actions [8, 26].

The model-science interface is defined as the conceptual and operational boundary in AI-assisted materials research workflows at which outputs generated by computational models—including predictions of material properties, quantified uncertainties, ranked lists of synthesis targets or characterization protocols, and proposed experimental designs—are translated, filtered, prioritized, and augmented to serve as actionable inputs for experimental systems, whether those systems operate autonomously, semi-autonomously, or with human oversight.

This definition deliberately encompasses several interrelated processes. Translation refers to the mapping of abstract, often high-dimensional or probabilistic model outputs onto concrete laboratory parameters such as precursor ratios, temperature profiles, or measurement sequences. Filtering involves the selective retention or discard of recommendations based on practical constraints like equipment availability, safety protocols, or material compatibility. Prioritization ranks competing suggestions according to criteria that may extend beyond pure model confidence, while augmentation incorporates additional domain knowledge or contextual information not captured in the original model. By isolating the interface as a distinct unit of analysis, rather than treating it as an implicit assumption within broader workflow descriptions [3, 21], researchers gain the capacity to examine its internal dynamics independently of upstream model training or downstream experimental execution.

Boundary conditions further clarify the scope. The interface begins at the moment of model output generation and terminates once fully specified directives reach the experimental apparatus. It explicitly excludes the internal mechanics of model construction, training datasets, or algorithmic optimization, as well as the pure physical execution of experiments once initiated. What it does include, however, are the mediating layers—software middleware, control scripts, human judgment points, or robotic interpreters—that bridge the computational and physical domains. In orchestration platforms designed for autonomous experimentation [6, 24], for example, the interface manifests in the conversion of optimization results into executable hardware commands, ensuring that theoretical insights become laboratory reality.

This conceptualization builds upon yet extends beyond existing portrayals of closed-loop and autonomous systems [14, 27]. Accounts of autonomous experimentation platforms describe integrated data flows but devote less scrutiny to the precise character of the model-to-experiment direction. Likewise, frameworks for next-generation self-driving laboratories [4, 22] highlight hardware-software integration without isolating the translation layer for dedicated study. Treating the interface as a unit of analysis opens new lines of inquiry: how does the representational format of model outputs affect their interpretability in experimental contexts? What mechanisms govern the incorporation of uncertainty information into decision thresholds? How do varying interface architectures influence the adaptability and robustness of discovery campaigns?

Moreover, conceptualizing the interface in this manner facilitates the integration of insights from adjacent fields such as control theory and information science into materials AI. The manner in which uncertainties are communicated across the boundary, for instance, can determine whether a system adopts a conservative or exploratory posture—distinctions that prove crucial in high-throughput environments [7, 25]. In active learning contexts [15, 17], the interface not only governs how experimental data update models but also dictates how updated models shape subsequent experimental choices, thereby closing the loop with greater or lesser fidelity. Thus, the formal definition offered here supplies the necessary foundation for subsequent examination of why the interface matters, how it operates in different modes, and how information is transformed when crossing it. By establishing these conceptual boundaries, the present framework equips the field with a shared language for analyzing and improving the handoff between model and science in pursuit of accelerated materials innovation [2, 20].

Why the Interface Matters

The model-science interface merits dedicated theoretical attention for at least three interrelated reasons that collectively underscore its influence on the success of materials AI workflows. First, the interface is the primary site of information transformation, where the raw outputs of computational models encounter the practical realities of laboratory execution [8, 26]. Because model predictions and uncertainties are expressed in forms that may not align directly with experimental parameters, the interface

becomes the locus at which information is selectively preserved, lost, augmented, or restructured. Without explicit recognition of this transformative role, researchers risk underestimating the cumulative effects of small distortions that propagate through iterative discovery cycles, ultimately diminishing the overall efficiency of autonomous systems [5, 23].

Second, the quality of decisions made within autonomous or semi-autonomous research platforms depends critically on the fidelity of the interface. High-quality model recommendations can lose their value if the interface fails to translate them accurately into actionable protocols. At the same time, modest predictions can gain leverage when the interface intelligently incorporates contextual constraints or human expertise [18]. In closed-loop setups, the interface therefore functions as a decision amplifier or attenuator, modulating how model confidence translates into experimental aggressiveness or caution. This role becomes especially salient in scenarios involving multi-objective optimization or high-dimensional search spaces, where the interface must balance competing priorities that the model itself may not fully resolve [12, 13].

Third, the interface raises important questions of scientific accountability that extend beyond model performance metrics. When an autonomous system proposes an experiment that fails or yields unexpected results, responsibility for the outcome must be traced not only to the model but also to the interface layer that mediated its implementation [11]. By rendering the interface visible as an object of study, the framework promotes greater transparency and traceability, enabling researchers to assign accountability appropriately and to design safeguards that prevent unexamined biases or errors from entering the experimental loop. Collectively, these three arguments—centered on information transformation, decision quality, and scientific accountability—demonstrate that the interface is not a peripheral concern but a core determinant of whether the promise of self-driving laboratories translates into tangible advances in materials discovery [4, 22].

A Typology of Interface Modes

To facilitate systematic analysis, the framework introduces a typology of interface operation modes organized along two orthogonal dimensions: the degree of autonomy granted to the AI system and the level of stakes associated with the experimental decisions being made. This typology can be represented as a 2 × 2 matrix in which the columns correspond to low versus high autonomy and the rows correspond to low versus high stakes. The resulting four modes—advisory, direct execution, human-mediated, and adversarial/competitive—capture distinct ways in which the model-science interface can function, each with characteristic strengths, limitations, and appropriate application contexts within materials research.

In the low-autonomy, low-stakes quadrant resides the advisory mode. Here, the interface operates by presenting model-generated recommendations as suggestions that human researchers may accept, modify, or reject without immediate commitment to automated execution. This mode preserves substantial human oversight while still leveraging computational insights, making it particularly suitable for early-stage exploratory work where uncertainty is high and the cost of individual experiments remains modest. Conceptual examples include scenarios in which active learning algorithms propose candidate compositions for initial screening [16, 17], yet final selection rests with domain experts who incorporate unmodeled factors such as precursor availability or safety considerations.

The low-autonomy, high-stakes quadrant corresponds to the human-mediated mode. In this configuration, the interface deliberately routes critical or high-impact recommendations through human review and approval before any experimental action proceeds. Autonomy is deliberately constrained to ensure that human judgment can intervene when potential consequences—scientific, financial, or safety-related—are substantial. This mode finds conceptual application in workflows involving rare or high-value materials where an erroneous synthesis attempt could waste limited resources or in regulatory contexts where accountability requires documented human authorization [11].

Moving to the high-autonomy, low-stakes quadrant yields the direct execution mode. Here, the interface allows the AI system to translate its recommendations directly into executable commands for automated hardware with minimal or no human intervention. Because the associated stakes remain low—such as routine characterization runs or incremental parameter adjustments—this mode maximizes speed and throughput while still operating within safe boundaries. Conceptual illustrations appear in self-driving laboratory designs [12, 13] where low-risk, repetitive

experiments are delegated entirely to robotic platforms once model confidence exceeds a predefined threshold.

Finally, the high-autonomy, high-stakes quadrant defines the adversarial/competitive mode. In this mode, multiple models or hybrid AI-human agents operate in competition or opposition, each challenging the recommendations of the others before any high-stakes experimental decision is finalized. The interface, therefore, functions as an arena for constructive conflict that refines proposals through iterative critique, thereby elevating overall robustness. This mode is conceptually suited to complex optimization problems where conflicting objectives must be reconciled and where the cost of error is high, such as the design of advanced functional materials requiring simultaneous satisfaction of multiple performance criteria [19].

The value of this 2 × 2 typology lies in its capacity to classify existing and proposed workflows, to diagnose mismatches between chosen mode and problem requirements, and to guide deliberate interface engineering. Rather than assuming a one-size-fits-all approach to automation, researchers can select or hybridize modes according to the specific autonomy and stakes profile of each discovery task [14]. By providing these clear distinctions, the typology transforms the previously implicit handoff between model and experiment into an explicit design variable, thereby advancing more intentional and context-appropriate implementations of autonomous materials research.

Table 1 extends the manuscript's 2 × 2 typology by showing that each interface mode embodies a distinct governance logic, information priority, and implementation challenge rather than merely a different level of automation.

Table 1. A governance-oriented typology of model–science interface modes in AI-driven materials research

Interface mode	Autonomy level	Decision stakes	Pri decisio the in
Advisory mode	Low	Low	AI ge recomm that nonbin interp sele rese
Human-mediated mode	Low	High	recomm are throug appro before
Direct execution mode	High	Low	AI ou convert into n rea instruc minim
Adversarial/competitive mode	High	High	Multipl or hybr ger com propo mu reconci a

Information Transformation across the Interface

When an AI output crosses the model-science interface, information does not pass unchanged; instead, it undergoes a series of transformations that can preserve, lose, add, or distort content in ways that profoundly affect downstream experimental outcomes. Preservation occurs when core elements—such as predicted property values or uncertainty ranges—are faithfully mapped onto experimental parameters without alteration. Loss arises when contextual details, such as the full probabilistic distribution behind a recommendation, are simplified or discarded to meet laboratory constraints. Addition takes place when the interface injects supplementary information unavailable to the model, for example, laboratory-specific equipment calibrations or safety interlocks. Distortion emerges when interpretive layers inadvertently shift meaning, such as when a ranked list of candidates is reordered according to unmodeled feasibility criteria [7, 25].

To capture the overall quality of these transformations, the framework introduces “interface fidelity” as a conceptual variable that quantifies how faithfully model outputs are rendered into experimental inputs. Interface fidelity possesses at least three measurable dimensions. Precision refers to the accuracy with which numerical or categorical outputs are represented in executable form; low precision might manifest as rounding of optimal synthesis temperatures that alters reaction kinetics. Completeness denotes the extent to which all relevant information—predictions, uncertainties, rationales—is carried forward rather than selectively omitted. Timeliness addresses the speed of transmission, particularly critical in closed-loop systems where delays can render recommendations obsolete before execution [6, 24].

Figure 1 conceptualizes the model–science interface as the central boundary through which AI outputs are transformed into experimental inputs, showing how operation modes, information transformation, interface fidelity, and failure modes jointly shape downstream experimental impact.

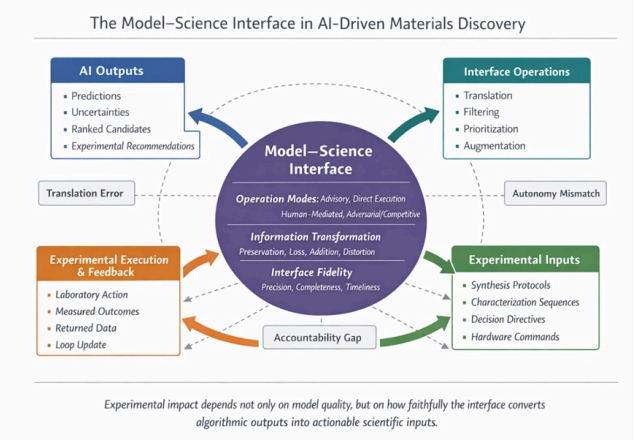

Figure 1. The model science interface in AI-driven materials discovery

By attending explicitly to these transformations and dimensions of fidelity, the framework equips researchers to evaluate and improve interface designs systematically. For instance, in advisory or human-mediated modes, completeness may be prioritized to support informed human judgment, whereas direct execution modes may emphasize precision and timeliness to maintain loop speed [9, 27]. Across all modes, recognizing the potential for distortion encourages the incorporation of transparency mechanisms that expose the reasoning behind interface-level decisions. Ultimately, analyzing information transformation in this manner shifts attention from isolated model performance to the holistic integrity of the discovery pipeline, offering a conceptual foundation for more reliable autonomous materials research [10, 28].

Interface Failure Modes

Even when the model-science interface is recognized as a distinct layer, its operation can still break down in systematic ways that undermine the entire materials discovery pipeline. The framework proposed here distinguishes three primary conceptual failure modes—translation error, autonomy mismatch, and accountability gap—each arising from different aspects of how information crosses the boundary between model outputs and experimental inputs. These failure modes are not mere implementation bugs but fundamental risks inherent to any interface that negotiates between probabilistic computational representations and deterministic laboratory actions. By naming and conceptualizing them explicitly, the framework equips researchers to anticipate, diagnose, and mitigate interface-level breakdowns rather than attributing all shortcomings to upstream modeling deficiencies or downstream experimental execution.

The first failure mode, translation error, occurs when model outputs are misinterpreted or incorrectly mapped onto experimental parameters, resulting in a disconnect between the intended recommendation and actual implementation. As articulated in discussions of Bayesian active learning for closed-loop discovery [7, 25], model outputs often take the form of high-dimensional probability distributions or ranked lists that must be compressed or discretized to fit laboratory hardware constraints. In translation error, this compression introduces distortions: a predicted optimal composition might be rounded to the nearest available precursor ratio, or an uncertainty range might be collapsed into a single point estimate that misguides synthesis temperature selection. Such errors propagate through iterative loops, causing the system to explore suboptimal regions of the design space. What distinguishes translation error from simple model inaccuracy is that the original prediction may have been correct within its own representational frame. Yet, the interface fails to preserve semantic fidelity during the conversion process. Conceptual examples appear frequently in self-driving laboratory literature where robotic platforms receive discretized instructions that inadvertently shift reaction conditions away from the model’s intended optimum [6, 24], thereby reducing the effectiveness of

autonomous experimentation systems [3, 21]. Without deliberate interface design that includes validation steps or reversible mappings, translation errors can silently erode discovery efficiency even when model performance metrics appear strong.

The next failure mode, autonomy mismatch, arises when the level of decisional autonomy granted at the interface fails to align with the confidence expressed in the model output. This mismatch is particularly acute in high-stakes or high-uncertainty scenarios common to functional materials optimization. For instance, a model may output a recommendation accompanied by wide uncertainty bounds. Yet, the interface proceeds with direct execution as though the prediction were definitive [14, 27], leading to premature commitment of resources to unpromising experiments. Conversely, an interface operating in overly conservative human-mediated mode may require human approval for every low-stakes suggestion, thereby negating the speed advantages promised by self-driving platforms [4, 22]. Autonomy mismatch, therefore, represents a calibration failure at the boundary layer: the interface does not dynamically adapt its autonomy level to the epistemic status of the incoming recommendation. The consequences extend beyond wasted experimental cycles; repeated mismatches can erode trust in the overall system and discourage adoption of more advanced autonomous workflows [11]. By treating autonomy mismatch as a distinct interface failure rather than a general automation shortcoming, the framework highlights the need for interfaces that incorporate confidence-aware gating mechanisms capable of modulating autonomy on the fly.

Another failure mode, accountability gap, emerges when no clear actor—human or algorithmic—assumes responsibility for decisions made at the interface itself. In many current descriptions of orchestration platforms and universal self-driving laboratories [6, 14], the interface layer is rendered invisible, so that when an experiment fails, blame is diffusely assigned either to the model or to the laboratory hardware without tracing the precise translation or filtering choices that occurred in between. This gap becomes especially problematic in hybrid systems where multiple models contribute recommendations or where human oversight is intermittent [8, 26]. Accountability gap undermines scientific traceability and ethical oversight because the rationale for why a particular synthesis parameter was selected—whether through filtering, augmentation, or prioritization—remains unlogged and unreviewable. The framework, therefore, proposes that interface designs must embed explicit logging and attribution protocols so that every transformation step carries a traceable provenance. Without such mechanisms, even well-intentioned autonomous experimentation systems risk operating in a regulatory and intellectual vacuum where lessons learned from interface-level mistakes cannot be systematically incorporated into future designs [5, 23].

Collectively, these failure modes illustrate that the model-science interface is not a neutral transmission channel but a potential source of systemic fragility. Translation Error compromises fidelity of representation, autonomy mismatch disrupts appropriate calibration of action to confidence, and accountability gap severs the link between decision and responsibility. Addressing them requires moving beyond ad-hoc implementations toward deliberate interface engineering that incorporates the conceptual vocabulary developed here. Only by recognizing these failure modes as interface-specific phenomena—rather than collateral damage of broader automation efforts—can the field achieve the robust, accountable materials discovery pipelines envisioned in recent reviews of smart automation and self-driving laboratories [2, 9, 20, 27].

Table 2 consolidates the manuscript's theoretical core by showing that interface failures do not arise abstractly, but from identifiable transformations that selectively threaten particular dimensions of interface fidelity.

Table 2. Diagnostic matrix linking information transformation, interface fidelity, and interface failure modes

Transformation process at the interface	What changes as AI output becomes experimental input	Fidelity dimension is most at risk	ass failu is p ma
Preservation	Core predictions, confidence estimates, or ranked options are carried forward with minimal alteration	Precision	Tran Erro pre valu nume sem misre exe f

Loss	Probabilistic detail, rationale, uncertainty structure, or contextual qualifiers are dropped to simplify action	Completeness	Aut Mis when infor de trea suffi de exe
Addition	Laboratory constraints, calibration rules, safety interlocks, or tacit human knowledge are inserted	Completeness and sometimes precision	Acco Gap a constr conse undoc or una
Distortion	Ranking order, meaning, or relative importance of outputs shifts during filtering or reconciliation	Precision + completeness	Trar Err Acco C esp w distor invi u
Temporal delay	Output reaches execution too late relative to system dynamics, queue conditions, or experimental state	Timeliness	Aut mis whe outp acte thoug
Conflict reconciliation	Multiple models or agents are merged into a single actionable recommendation	Consensus robustness plus traceability	Acco Gap actor recor rule; if one l don

Illustrative Applications

To demonstrate the framework's utility, conceptual cases—closed-loop synthesis, human-in-the-loop characterization, and multi-model recommendation—illustrate how an interface-centric perspective clarifies challenges in materials AI workflows. These thought experiments draw on the typology of modes, information transformation, and failure modes to reveal dynamics at the model–science interface.

A closed-loop synthesis platform operates in direct execution mode, translating active-learning outputs into robotic commands [2, 20]. The framework shows how interface transformations add laboratory constraints (e.g., reagent purity, reactor limits) while risking loss of model uncertainty [7, 25]. Translation error—such as discretization of composition space—may drive convergence to local optima, a risk invisible at the level of model accuracy. Monitoring fidelity (precision, completeness, timeliness) can help detect when constraints dominate recommendations, reframing automation as a tunable design space.

Next, it considers human-in-the-loop characterization of piezoelectrics or ferroelectric perovskites [16, 17]. In this human-mediated mode, expert input adds tacit knowledge (e.g., calibration nuances), potentially improving decisions despite reduced uncertainty completeness [18]. However, autonomy mismatch may arise if humans systematically override or rubber-stamp recommendations. The framework suggests refining handoffs—for example, through confidence-calibrated prompts—so human and model strengths are aligned.

The third case examines multi-model recommendation in a self-driving laboratory [14]. Operating in Adversarial/Competitive Mode, the interface reconciles conflicting outputs, introducing meta-reasoning but also potential bias [19]. Here, fidelity must include consensus robustness alongside precision, completeness, and timeliness. Without clear responsibility, an Accountability Gap can undermine traceability. The framework thus recasts the interface as a deliberate integration layer rather than an ad hoc aggregator.

Across cases, the framework shifts focus from model performance to boundary dynamics that determine experimental outcomes. It offers a shared vocabulary—modes, fidelity, failure types—to systematically analyze and improve interface design, bridging visions of self-driving labs [4, 22] with real information flows [8, 26].

Implications for Materials AI Practice

The framework suggests several practical shifts. First, researchers should report interface design choices alongside model metrics. Rather than treating the interface as invisible, studies of autonomous experimentation systems [3, 12, 13, 21] should document operation modes, translation mechanisms, and fidelity dimensions to enable comparison and best-practice development, and the field should systematically study information loss and distortion at the interface. Complementing focus on accuracy and throughput [1, 15], researchers can analyze preservation of uncertainty, constraint addition, and fidelity. Using concepts like translation error, autonomy mismatch, and accountability gap [5, 23] helps diagnose whether issues stem from models, interfaces, or execution, enabling more targeted experimentation. We would also suggest that interface transparency should become a core design principle [6, 9, 24, 27]. Logging transformations, incorporating confidence-aware autonomy, and ensuring traceable reconciliation can reduce accountability gaps and improve fidelity. This may involve standardized interface description languages or middleware exposing transformation logic.

Overall, the framework shifts materials AI from a model-centric paradigm toward balanced attention to boundary management. Advances in inverse design and active learning [2, 4, 20, 22] remain essential, but their impact depends on treating the interface as a first-class object of design and analysis.

Conclusion

This study advances a conceptual theory of the model–science interface as the boundary where AI outputs become experimental inputs. Defining the interface explicitly introduces a typology of modes, an account of information transformation, the concept of interface fidelity, and a vocabulary of failure modes for systematic analysis.

An interface-aware perspective enhances robustness, transparency, and accountability in self-driving laboratories and autonomous experimentation systems. It shifts attention from assumed seamless data flow to deliberate boundary design. The broader call is for the materials AI community to adopt and extend this lens: making the interface visible is essential to fully realizing the promise of autonomous materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial Support

None

Ethics statement

None

Received: 27 Jun 2021 Revised: 15 Aug 2021 Accepted: 14 Sep 2021
Published online: 18 January 2022

References

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2:0121. https://doi.org/10.1038/s41570-018-0121.

Tabor DP, Roch LM, Saikin SK, Kreisbeck C, Sheberla D, Montoya JH, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. Nat Rev Mater. 2018;3(5):5-20. https://doi.org/10.1038/s41578-018-0005-z.

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Persson KA. Autonomous experimentation systems for materials development. Matter. 2021;4(9):2702-26. https://doi.org/10.1016/j.matt.2021.06.036.

Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282-91. https://doi.org/10.1016/j.trechm.2019.02.007.

Montoya JH, Aykol M, Anapolsky A, Gopal CB, Herring PK, Hummelshøj JS, et al. Toward autonomous materials research: Recent progress and future challenges. Appl Phys Rev. https://doi.org/10.1063/5.0071507.

Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, et al. ChemOS: Orchestrating autonomous experimentation. Sci Robot. 2018;3(19):eaat5559. https://doi.org/10.1126/scirobotics.aat5559.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11:5966. https://doi.org/10.1038/s41467-020-19597-w.

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. https://doi.org/10.1021/acs.accounts.2c00220.

Baird SG, Sparks TD. What is a minimal working example for a self-driving laboratory? Matter. 2022;5(6):1747-50. https://doi.org/10.1016/j.matt.2022.11.007.

Soldatov MA, Butova VV, Pashkov D, Butakova MA. Self-driving laboratories for development of new functional materials and optimizing known reactions. Nanomaterials. 2021;11(3):619. https://doi.org/10.3390/nano11030619.

Bennett JA, Abolhasani M. Autonomous chemical science and engineering enabled by self-driving laboratories. Curr Opin Chem Eng. 2022;36:100831. https://doi.org/10.1016/j.coche.2022.100831.

Rooney MB, MacLeod BP, Oldford R, Thompson ZJ. A self-driving laboratory designed to accelerate the discovery of adhesive materials. Digit Discov. 2022;1(5):1026-37. https://doi.org/10.1039/D2DD00029F.

Tamasi MJ, Gormley AJ. Biologic formulation in a self-driving biomaterials lab. Cell Rep Phys Sci. 2022;3(7):100974. https://doi.org/10.1016/j.xcrp.2022.100974.

Epps RW, Volk AA, Ibrahim MYS, Abolhasani M. Universal self-driving laboratory for accelerated discovery of materials and molecules. Chem. 2021;7(10):2697-719. https://doi.org/10.1016/j.chempr.2021.09.004.

Guo Z, Yamaguchi R. Machine learning methods for protein-protein binding affinity prediction in protein design. Front Bioinform. 2022;2:1065703. https://doi.org/10.3389/fbinf.2022.1065703.

Balachandran PV, Kowalski B, Sehirlioglu A, Lookman T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. Nat Commun. 2018;9:1668. https://doi.org/10.1038/s41467-018-03821-9.

Nicol AA, Owens SM, Le Coze SS, MacIntyre A, Eastwood C. Comparison of high-technology active learning and low-technology active learning classrooms. Act Learn High Educ. 2018;19(3):253-65. https://doi.org/10.1177/1469787417731176.

Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design. Integr Mater Manuf Innov. 2017;6(3):207-17. https://doi.org/10.1007/s40192-017-0098-z.

Ament S, Yu H, Wu C, Hattrick-Simpers J, DeCost B, Sarker S, et al. Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. Sci Adv. 2021;7(51):eabg4930. https://doi.org/10.1126/sciadv.abg4930.

Turner JM. The matter of a clean energy future. Science. 2022;376(6600):1361.

Flores-Leonar MM, Mejía-Mendoza LM, Aguilar-Granda A, Sanchez-Lengeling B, Tribukait H, Amador-Bedolla C, et al. Materials acceleration platforms: On the way to autonomous experimentation. Curr Opin Green Sustain Chem. 2020;25:100370.

Serrano M, Boniface M, Calisti M, Schaffers H, Domingue J, Willner A, et al. Next generation internet research and experimentation. In: Serrano M, et al. Building the future internet through FIRE. Aalborg: River Publishers; 2022. p. 43-84.

Zhang R, Mozaffari A, de Pablo JJ. Autonomous materials systems from active liquid crystals. Nat Rev Mater. 2021;6(5):437-53.

Gongora AE, Snapp KL, Whiting E, Riley P, Reyes KG, Morgan EF, et al. Using simulation to accelerate autonomous experimentation: A case study using mechanics. iScience. 2021;24(4).

Hagnell MK, Åkermo M. The economic and mechanical potential of closed loop material usage and recycling of fibre-reinforced composite materials. J Clean Prod. 2019;223:957-68.

Strieth-Kalthoff F, Sandfort F, Kühnemund M, Schäfer FR, Kuchen H, Glorius F. Machine learning for chemical reactivity: The importance of failed experiments. Angew Chem Int Ed. 2022;61(29):e202204647.

Petrino R, Castrillo LG, Yilmaz B, Dodt C, Tuunainen E, Khoury A, et al. Policy statement on minimal standards for safe working conditions in Emergency Medicine. Eur J Emerg Med. 2022;29(6):389-90.

National Academies of Sciences, Engineering, and Medicine. New directions for chemical engineering. Washington (DC): National Academies Press; 2022.