JNTM JOURNAL of NEW TECHNOLOGY and MATERIALS

Journal of New Technology and Materials

Vol. 14, No. 1 (2024), 8–20

RESEARCH ARTICLE

Governed Autonomy Readiness for AI-Enabled Materials Research: A Capability-Coupled Decision Model for New-Technology Laboratories

Vincent Marks[1,*] and R. A. Sulaiman[2]

[1]University of Surrey, Guildford GU2 7XH, Surrey, UK

*Correspondence: vincentmarks@btinternet.com

Received date: June 28, 2023; Accepted date: January 30, 2024

Abstract

While there is extensive use of AI algorithms for composition selection, interpretation of measurements, robotic experiments, and organization of materials information, the preparedness of the laboratory setting cannot be determined solely on the basis of AI autonomy. In this paper, we refine the Human-Governed Autonomy Mapping Index (HGAMI) as a materials-informatics decision model for ranking AI-enabled research workflows under concurrent technical and governance criteria. Specifically, five workflow modes are analyzed, namely regression/classification, robotic decision support, generative hypothesis modeling, orchestrated modular autonomy, and general-purpose AI autonomy. All workflow modes have been represented by cognitive inference, physical agency, knowledge creation, module orchestration, and governance assurance criteria. The resulting values of raw autonomy, governance-revised deployability, efficiency of selecting experiments, and balanced capability are computed and explained as laboratory planning metrics. Our research hypothesis concerns the association between the most technically autonomous AI workflow mode and the most responsible one for imminent materials laboratories. The computed metric values reject this hypothesis. Although the general-purpose AI autonomy workflow yields maximal values of raw autonomy ($A_i = 0.938$) and efficiency of experiment selection (84.7%), the value of deployability decreases to $D_i = 0.616$, due to the low level of governance. Meanwhile, the orchestration of modular autonomy leads to the best deployability score ($D_i = 0.640$), because high capability is distributed among modules whose operation can be monitored. We suggest that the approach involving AI autonomy governance should be followed in materials informatics. Machine learning, robotic execution, generative inference, and module orchestration should still be tied to the defined goals, data auditability, validation procedures, and scientific responsibility. Hence, HGAMI can serve as an accessible decision model for intelligent materials laboratory planning.

Keywords: artificial intelligence, materials informatics, autonomous experimentation, materials discovery, governed autonomy, closed-loop optimization, new technology materials

1 Introduction

Optimization and discovery of novel materials now depend heavily on intelligent decisions made within large, uncertain, and experimentally expensive design spaces. Properties of materials utilized in energy conversion, electrochemical storage, catalysis, corrosion protection, optoelectronics, quantum computing, additive manufacturing, polymer engineering, and environmental technology are rarely governed by a single feature. They develop out of complex interactions among compositional, structural, defect-based, thermodynamic, microstructural, surface, interfacial, and aging effects [**?**]. Traditional trial-and-error discovery remains essential for a thorough physical understanding, but is ineffective in covering multi-dimensional design spaces. High-throughput synthesis and characterization can reduce the pace of search, but will still produce slow results if candidate evaluation, experiment interpretation, and decision making are disconnected manual actions.

Materials informatics emerged as a response to this challenge by establishing connections between high-throughput computations, data management, descriptors, machine learning, and feedback. The development of the Materials Project,

AFLOW, OQMD, and NOMAD created solid precedents for database-based exploration of phase diagrams, electronic structure, and structure-function relationships on a large scale [1–3]. Subsequent efforts showed that composition-derived features, structure-derived descriptors, graph representations, and physics-guided variables allow building machine learning-based predictive models accurate enough to assist materials selection and ranking [4–6]. Machine learning-based models change the role of computation in materials discovery. In the past, computation was primarily a tool of explaining experimentally measured material properties. Nowadays, it can also become an element of decision making and help in choosing which materials should be synthesized, what synthesis route should be adopted, and what experiments should be performed.

The current state of the art clearly shows that the most useful AI-enabled materials-discovery pipeline is not simply a better predictor. It is an intelligent decision loop in which data reliability, uncertainty, domain-specific constraints, and experimental feedback determine which next step should be taken. Both active learning and Bayesian optimization provide formal frameworks for decision making under uncertainty. Autonomous material discovery platforms demonstrate how recommendations can be followed by experiments and then incorporated back into models [7–9]. Recent studies applied Bayesian active learning for on-the-fly identification of materials, self-driving approach for thin films and molecules, and autonomous platforms for materials synthesis [10–12]. These achievements are crucial for shifting from post-experimental machine learning to decision-based experimentation. However, they raise an important question: can laboratory-scale materials discovery and optimization fully rely on AI-powered decisions? Can we evaluate AI-based autonomy not only by its performance but also by the quality of its decision-making process?

Recent advances in autonomy-driven materials research show a shift towards more sophisticated interaction with human experts, task-specific decision-making processes, and laboratory environment. Human-in-the-loop Bayesian phase mapping proved that human expertise can be integrated into autonomous discovery so that machine learning can work hand in hand with scientific intuition [13]. Targeted Bayesian algorithm execution emphasizes that many tasks of materials science include finding sub-spaces in a materials design space that meet user-defined requirements and constraints, not maximization of a single function [14]. Autonomous laboratory surveys also show that many researchers appreciate partial automation and prefer modularity and human-in-the-loop approaches because some materials laboratories are not ready for full autonomy [15, 16]. All this leads to a concrete task: materials laboratories need a structured approach to comparing the quality of different kinds of AI-based decisions to evaluate their readiness for materials discovery.

Generation of AI and tool-using language models make this issue even more complicated. Systems based on language models combined with chemistry- and materials-specific tools can perform literature reviews, invoke external programs, generate hypotheses, and even propose procedures for performing certain operations [17]. While such models extend the intellectual capacity of researchers, their outputs need a more thorough validation than regular machine learning outputs. Generated description of a chemical reaction, corrosion mechanism, catalytic pathways, transition states, or proposed procedures for materials synthesis should be experimentally verified by checking it for consistency with laws of physics and chemistry and possible instrumental limitations. In this respect, language fluency is not enough. This difference needs to be clearly understood in any materials research.

Figure 1. Materials research environment suitable for evaluation by the HGAMI framework. Computational analysis, autonomous or semi-autonomous manipulation of samples, molecular modeling, instruments, and conventional characterization are to be evaluated jointly because the usefulness of autonomy relies on close connection of prediction, action, validation, and interpretation of data.

Figure 1 depicts a typical research environment in which materials discoveries with AI assistance will take place. As

shown, the research setup includes computer stations with analysis tools, autonomous or semi-autonomous platforms for sample preparation, instruments, and conventional materials research tools. It highlights the hybrid nature of modern materials research, which includes computational inference, physical actions, reasoning about scientific problems, and experimental validation.

It also demonstrates the reason why AI autonomy cannot be simply treated as a single measure. Machine learning-based prediction can rank candidate materials or structures without involving any hardware resources. An autonomous platform can execute multiple experiments according to an objective provided by human researchers. Language-based models can generate explanations for materials phenomena or suggest candidates without direct experimental support. Orchestration layer can combine multiple independent modules but its outputs can become ambiguous and hard to trace. Full-fledged autonomous agent can perform multiple cognitive functions simultaneously but such breadth makes its decisions difficult to track. When speaking about a laboratory environment, the question becomes more specific: does high autonomy translate into scientific feasibility and deployability in materials research?

The question addressed by this study is thus the following: given a set of materials research workflows assisted by different AI systems, how should we determine which system provides us with the highest level of responsible autonomy? Unlike a generic question about capabilities of various AI systems, it requires to compare the ability of different approaches to make correct decisions and use them in scientific context. The importance of this question is easy to understand. Laboratory facilities may become completely automated, instruments can be remotely operated, and language models can generate hypotheses and experiments. But before investing in all these things, we need a way of evaluating their responsible deployability in a laboratory environment.

This need is addressed in this study by combining autonomous capability descriptors with the corresponding governance modifier. The model allows distinguishing between technical capabilities and scientific autonomy, separating four aspects of AI-based decisions, and accounting for governance and validation. The result is the Human-Governed Autonomy Mapping Index (HGAMI), which gives researchers an opportunity to compare different AI-assisted materials workflows and decide which of them can provide the most responsible autonomy for materials discovery. The method is also supplemented with an experimental-selection efficiency measure and a measure of balance between autonomous capabilities.

The key contribution of this paper is a reproducible classification model for comparing various forms of autonomous materials discovery workflows. Contrary to popular assumptions, the paper will not show that the most autonomous system is always the best solution. Instead, it will demonstrate that responsible materials autonomy is determined by interactions between technical and experimental capabilities of autonomous system, user-defined goals, validation process, traceability, and accountability. These priorities directly reflect new-technology materials research, where any significant result requires a validated decision-making procedure.

2 Materials and Methods

2.1 Study design

The development of a computational methodology was undertaken to investigate if various AI-enabled materials science workflows can be compared using a readiness metric. The following five modes of work were defined, as representatives of possible workflows in materials research laboratories where machines are involved to a greater or lesser extent. Each workflow type corresponds to a specific role in research, although the types are not arranged hierarchically or even sequentially. The types include regression and classification, robotic decision support, generative hypothesis modeling, orchestrated modular autonomy, and general-purpose AI autonomy.

Regression and classification includes standard supervised learning systems designed to predict material properties, rank candidates, or classify materials from descriptors. Robotic decision support includes work in which algorithms drive or communicate with laboratory equipment, including synthesis apparatuses, processing equipment, or measuring apparatuses. Generative hypothesis modeling includes algorithms which propose hypotheses about the underlying phenomena or materials-related mechanisms, candidates based on these hypotheses, or connections to related literature. Orchestrated modular autonomy includes the orchestration of multiple modules in a research pipeline, including databases, predictors, simulators, optimizers, and experimental hardware platforms. General-purpose AI autonomy includes workflows in which a highly-capable AI system can plan experiments, reason about the results, propose a hypothesis, generate predictions from models, decide on actions, and coordinate experiments.

This study utilizes ordinal descriptor coding in lieu of actual experimental measurements. It is appropriate because the primary objective of this paper is not an empirical comparison, and the goal is to develop a laboratory-readiness score. Therefore, each value in the score should be a number normalized between 0 and 1, representing relative capability in each dimension of interest. The descriptor assignment method is designed to allow flexible adjustments by each laboratory based on maturity of their instruments, classes of materials being studied, regulatory requirements, safety considerations, or even validation experiments. In particular, each laboratory may assign new values to some capabilities according to its priorities.

2.2 Definition of capability descriptors

Five normalized variables were specified for each research mode i. Cognitive inference (C_i) measures the capability to extract relationships, predict properties, rank materials candidates, determine uncertainty, or select actions from data. Physical agency (P_i) measures the degree to which the workflow can interact directly with or via robotic control equipment, which in turn controls synthesis, processing, and characterization instruments. Knowledge generation (K_i) measures the capability to generate hypotheses, mechanistic explanations, design rules, or scientific explanations. Module orchestration (O_i) measures the level of interaction between modules in a research workflow. Governance assurance (G_i) measures the ease of interpretability, readiness for validation, human oversight, and clear responsibility attribution.

Technical capabilities are combined using a weighted sum into the raw autonomy score A_i:

$$A_i = w_C C_i + w_P P_i + w_K K_i + w_O O_i, \tag{1}$$

where w_C, w_P, w_K, and w_O are positive numbers that sum to unity. Equation (1) is intended to represent a structured score, rather than a universal principle. An algorithm or hardware can have a high A_i value only if it demonstrates multiple levels of autonomy concurrently. In other words, a purely digital predictor can still be considered very valuable, but not particularly autonomous. As in the present study, the greatest weight is assigned to cognitive inference, because it is necessary for the vast majority of intelligent workflows. The next two weights correspond to experimental action (physical agency) and scientific interpretation (knowledge generation), respectively. The smallest weight is given to orchestration, as it enables but does not constitute full autonomy:

$$w_C = 0.30, \qquad w_P = 0.25, \qquad w_K = 0.25, \qquad w_O = 0.20. \tag{2}$$

These weights assign a materials-research meaning to the autonomy score. While prediction and decision-making are valued above all else, they do not negate the importance of experimental action and knowledge generation. At the same time, the weights allow laboratories to adjust them to reflect the specifics of their workflow: for example, robotic synthesis of air-sensitive powders would warrant higher weight assigned to physical agency w_P, while literature-based materials concept generation would benefit a greater emphasis on knowledge generation.

Deployability D_i is defined by the product of autonomy and governance:

$$D_i = A_i G_i^{\alpha}, \tag{3}$$

where α is the governance sensitivity coefficient. In order to prevent low scores due to a lack of governance despite strong technical capabilities, it was set to 0.40 in this study. This definition reflects a practical problem faced by laboratories where an intelligent workflow is capable of guiding expensive experiments, but its decisions cannot be validated, interpreted, or audited properly. Additionally, α prevents governance from being a formality or a mere decoration. Low governance assurance reduces the total score even when autonomy is relatively high. In fact, the definition of equation (3) reflects the laboratory reality: in order for AI recommendations to lead to actual experimental actions, they must be backed up by data and constrained by the experimenter's preferences and safety requirements.

Efficiency of candidate selection E_i is estimated using an experiment selection efficiency indicator:

$$E_i = 100 \left(1 - e^{-2A_i}\right). \tag{4}$$

This measure is not a throughput, but rather an indication of the potential improvement of candidate selection due to increased autonomy, assuming the workflow is scientifically valid and well-designed. The exponential decay ensures a diminishing effect from increasing autonomy, as early improvements of the algorithm result in a significant gain, but later ones produce a lower increase unless accompanied by governance improvements. Such an interpretation is important for materials research laboratories, since going from manual to uncertainty-aware AI screening can help avoid many inefficient experiments, whereas further increases in autonomy can be counter-productive if the objective function and measurement loop are improperly defined.

Balance of technical capabilities B_i is calculated as follows:

$$B_i = 1 - \left[\max(C_i, P_i, K_i, O_i) - \min(C_i, P_i, K_i, O_i)\right]. \tag{5}$$

In order for an AI-enabled workflow to be useful in materials science, its capabilities cannot be concentrated in a single dimension. A workflow that ranks candidates but makes no use of experimental results would still be scientifically useful as a screening step. Similarly, a workflow that can drive laboratory hardware but generates no scientific insight would still perform well in terms of data acquisition. On the contrary, if the workflow is only good at a single point in the research cycle, it is likely underdeveloped. In practice, materials research usually needs multiple stages of the cycle: from

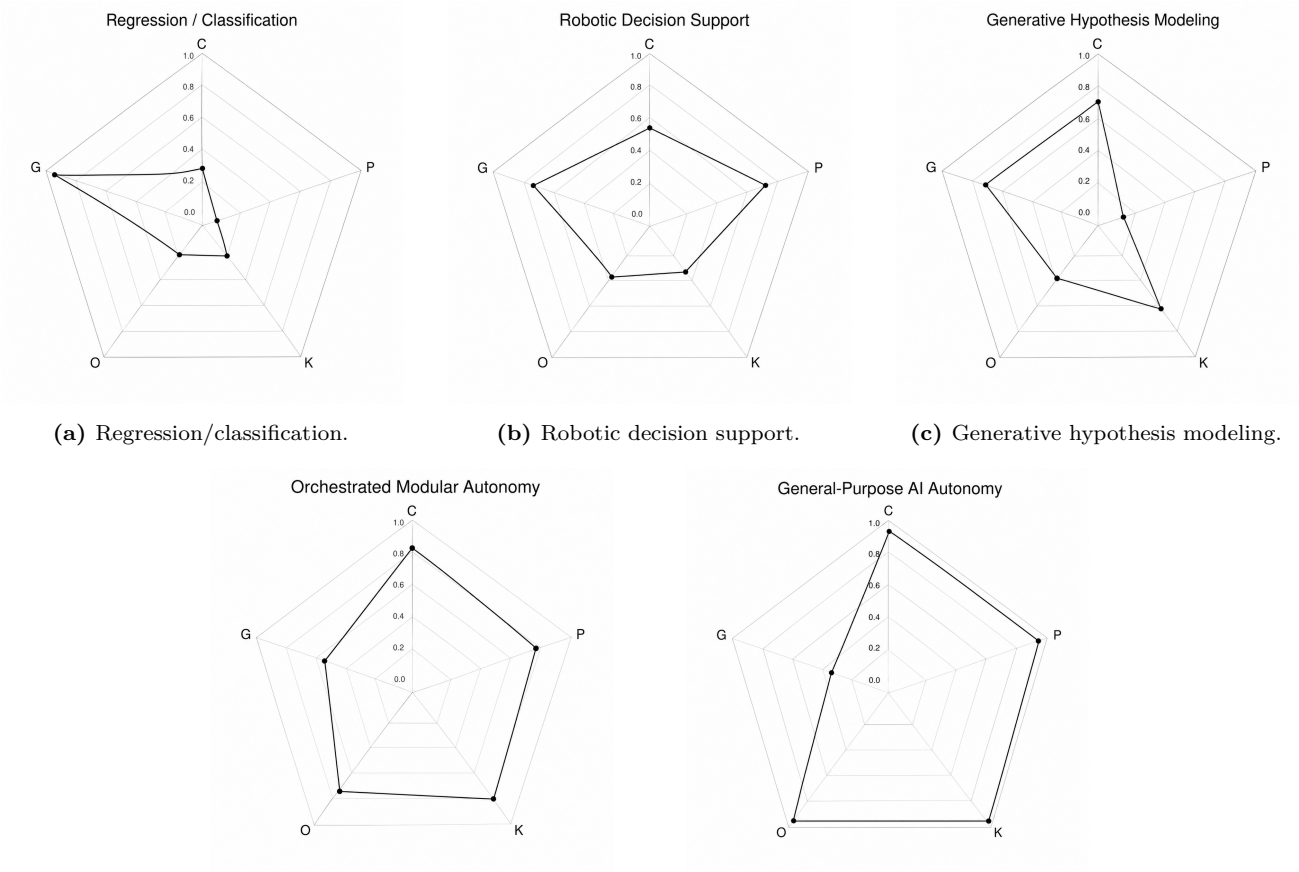
predicting materials to generating hypotheses. This fact is represented by equation (5), adding a diagnostic component to the scoring system.

2.3 Encoded workflow modes

Descriptive values are listed below (Table 1) and illustrated graphically (Figure 2). In particular, each workflow type is separated into technical capabilities and governance assurance in the scoring system. Regression and classification workflows are characterized by a high cognitive inference, but lack physical agency and orchestration capabilities. Robotic decision support algorithms are characterized by a high physical agency, but low knowledge generation capabilities. Hypothesis modeling workflows have a high knowledge generation capability, but a low physical agency value. Orchestration workflows include high values in almost all of the technical categories. General-purpose autonomy workflows have extremely high values for technical capabilities, but a slightly lower governance assurance.

Table 1. Encoded AI-enabled materials-research modes used for HGAMI computation.

Research mode	C_i	P_i	K_i	O_i	G_i
Regression and classification	0.30	0.05	0.20	0.10	0.95
Robotic decision support	0.55	0.75	0.35	0.40	0.75
Generative hypothesis modeling	0.70	0.10	0.70	0.35	0.65
Orchestrated modular autonomy	0.85	0.75	0.80	0.85	0.55
General-purpose AI autonomy	0.95	0.90	0.95	0.95	0.35



(a) Regression/classification. **(b)** Robotic decision support. **(c)** Generative hypothesis modeling.

(d) Orchestrated modular autonomy. **(e)** General-purpose AI autonomy.

Figure 2. Capability and governance profiles for the five AI-enabled materials-research modes. The axes correspond to cognitive inference (C_i), physical agency (P_i), knowledge generation (K_i), module orchestration (O_i), and governance assurance (G_i). The profiles clarify why high raw autonomy must be interpreted together with governance strength and capability balance.

Table 1 plays an important role in the calculations since it reveals the modeling assumptions used. The values are not intended to represent a universal indicator of every imaginable laboratory readiness; rather, they indicate laboratory readiness according to five well-known workflow modes. In this case, regression/classification is assigned a high level of

governance and a low level of physical agency, while the ready-for-use general-purpose AI autonomy has a high technical breadth but a weak governance assurance. Thus, this distinction helps avoid the confusion between sophisticated and ready modes.

Figure 2 presents radar profiles based on the descriptor values in Table 1. This graphical representation of the laboratory workflows helps understand the distribution of capabilities. Regression/classification is largely limited to digital capabilities, robotic decision support involves mostly physical agency, generative hypothesis modeling is strong at cognitive and knowledge generation tasks, orchestrated modular autonomy exhibits balanced capability, and general-purpose AI autonomy has high technical breadth but a low governance assurance.

As seen from Fig. 2, different workflow modes present various types of laboratory capabilities. As such, they help illustrate the methodological argument that different laboratory research problems involve different capabilities and thus need different workflow modes. For instance, the problem associated with catalyst screening requires uncertainty-aware inference and validation, whereas the automated synthesis of powders involves more physical and material agency. Thus, this graphical presentation allows readers to better understand the relation between numeric descriptors and laboratory decisions.

2.4 Computational procedure

Four consecutive stages were used to perform the calculations. In the first stage, descriptor values were assigned for each workflow mode. In the second step, raw autonomy was calculated using Eq. (1). Then, the responsible deployability was evaluated using Eq. (3). Finally, experiment-selection efficiency and capability balance were assessed using Eqs. (4) and (5). The results were compared to see if technical autonomy implies responsible readiness.

3 Results

3.1 Raw autonomy and deployability outputs

The outcomes of HGAMI analysis are given in Table 2 and Figure 3. Indeed, there is no complete correspondence between the two indicators, as the general-purpose AI workflow mode has the highest value of raw autonomy, $A_i = 0.938$. This happens due to strong cognitive inference, physical agency, knowledge generation, and orchestration. At the same time, the responsible deployability is $D_i = 0.616$ for the general-purpose AI, which is lower than that of orchestrated modular autonomy.

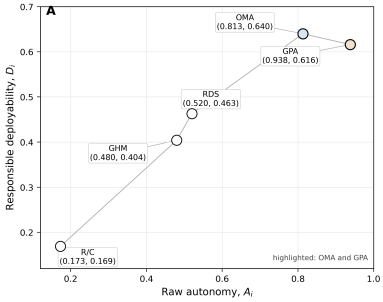
Table 2. HGAMI outputs for five AI-enabled materials-research modes.

Research mode	**Raw autonomy** A_i	**Deployability** D_i	**Efficiency** E_i **(%)**
Regression and classification	0.173	0.169	29.2
Robotic decision support	0.520	0.463	64.6
Generative hypothesis modeling	0.480	0.404	61.7
Orchestrated modular autonomy	0.813	0.640	80.3
General-purpose AI autonomy	0.938	0.616	84.7

Table 2 responds directly to the primary computation-based aspect of the research question, showing the maximum values for A_i and E_i associated with general-purpose AI autonomy, while D_i is the greatest for orchestrated modular autonomy. This table disproves the notion that the most technologically advanced form of AI automatically translates into the most deployable form of AI, for a materials lab, which suggests that a versatile workflow does not necessarily equate to a more deployable system if its decisions are not easily reviewed by examining data, model, optimizer, instruments, and validation rule.

In Figure 3, a graphical representation of the same outputs is provided. Panel (a) displays the relationship between the raw autonomy and responsible deployability scores, which reveals the impact of the governance penalty. The comparison indicates that general-purpose AI autonomy has the highest raw autonomy, while orchestrated modular autonomy has the highest responsible deployability due to its greater technical proficiency and control. Panel (b) presents the values used for the above comparison.

Regression and classification yields the lowest score on raw autonomy, $A_i = 0.173$. Robotic decision support has much better raw autonomy score at 0.520 and an efficiency indicator at 64.6%. On raw autonomy score, generative hypothesis modeling is at 0.480, and deployability score is 0.404. On the other hand, orchestrated modular autonomy attains the greatest responsible deployability at 0.640, along with a very good raw autonomy at 0.813 and an efficiency indicator at 80.3%. General-purpose AI autonomy has the highest raw technical and efficiency indicators.



(a) Raw autonomy–deployability frontier.

B

Code	Workflow mode	A_i	D_i	E_i (%)	G_i
R/C	Regression/classification	0.173	0.169	29.2	0.95
RDS	Robotic decision support	0.520	0.463	64.6	0.75
GHM	Generative hypothesis modeling	0.480	0.404	61.7	0.65
OMA	Orchestrated modular autonomy	0.813	0.640	80.3	0.55
GPA	General-purpose AI autonomy	0.938	0.616	84.7	0.35

(b) Numerical HGAMI output values.

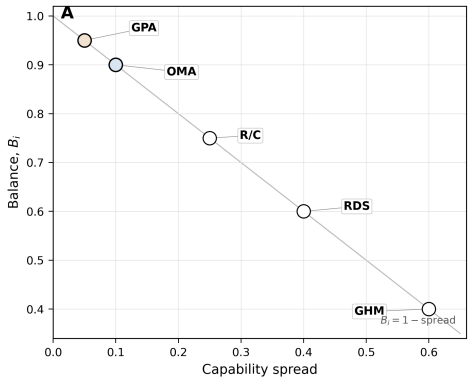
Figure 3. Autonomy and responsible-deployability outputs for the five AI-enabled materials-research modes. The frontier plot separates technical breadth from deployable readiness, while the numerical panel reports the values of A_i, D_i, E_i, and G_i used in the interpretation.

3.2 Capability spread and balance outputs

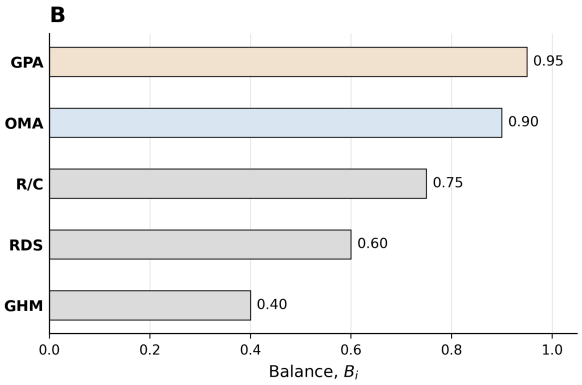
Table 3 presents capability spread and balance score, and Figure 4 visualizes the same relationship. The balance score identifies whether a workflow is specialized or broadly capable. General-purpose AI autonomy has the highest balance score of 0.95, followed by orchestrated modular autonomy with 0.90. Regression and classification has a moderate balance score of 0.75 because all its capability values are low to moderate, producing a relatively small spread. Robotic decision support and generative hypothesis modeling show lower balance because their capability profiles are uneven.

Table 3. Capability spread and balance scores for AI-enabled materials-research modes.

Research mode	Capability spread	Balance score B_i
Regression and classification	0.25	0.75
Robotic decision support	0.40	0.60
Generative hypothesis modeling	0.60	0.40
Orchestrated modular autonomy	0.10	0.90
General-purpose AI autonomy	0.05	0.95



(a) Capability spread–balance map.



(b) Balance ranking.

C

Code	Workflow mode	Spread	B_i	D_i
R/C	Regression/classification	0.25	0.75	0.169
RDS	Robotic decision support	0.40	0.60	0.463
GHM	Generative modeling	0.60	0.40	0.404
OMA	Modular autonomy	0.10	0.90	0.640
GPA	General-purpose AI	0.05	0.95	0.616

(c) Balance-value table.

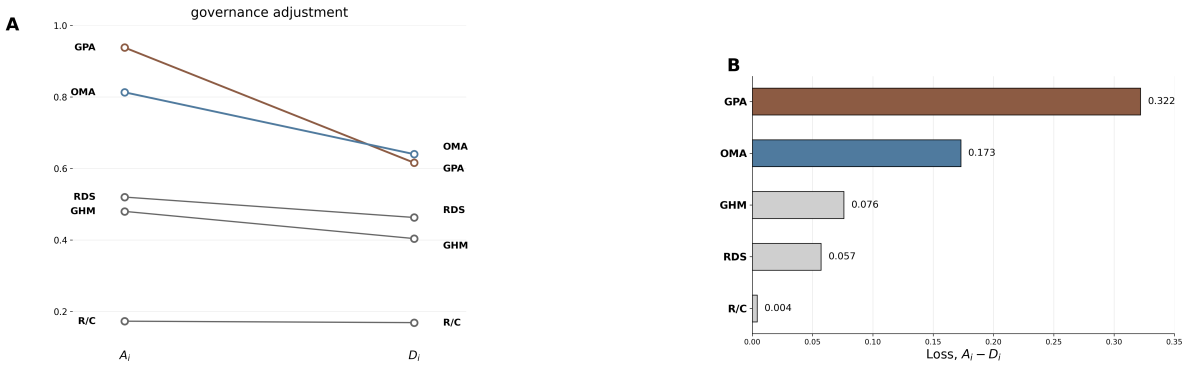
Figure 4. Capability balance analysis for the five AI-enabled materials-research modes. Broadly distributed capability is strongest for general-purpose AI autonomy and orchestrated modular autonomy, but responsible workflow selection also requires interpretation of governance strength and task-specific laboratory constraints.

Table 3 gives a second diagnostic output of HGAMI on top of the other two. It indicates how distributed or concentrated the capability is. High balance for AI autonomy indicates a wide range of technologies but should be combined with low governance assurance. On the other hand, regression/classification is rather balanced since all four technical aspects still have quite low values; hence, balance alone does not guarantee maturity. Table 3 becomes relevant only if used in combination with Tables 1 and 2.

Figure 4 presents the balancing results visually in three ways. Panel (a) compares capability distribution with B_i, panel (b) ranks the modes according to their balance scores, and panel (c) shows the numbers for distribution and balance metrics. The figure illustrates that balance alone cannot be equated with deployability; it needs to be analyzed together with raw capability and governance-adjusted deployability.

3.3 Governance-adjustment outputs

Governance assurance is the metric that most effectively differentiates between mere technical capability and mature responsibility of deploying an AI. Governance adjustment in the current analysis is represented in Figure 5. Regression/classification has the highest governance score since all three aspects associated with its operation – inputs, outputs, and validation – are generally well-defined. For example, a model predicting band gaps, formation energy, corrosion rate, or adsorption energy can be tested on the held-out data, compared to other models, and inspected through feature-importance or explainability [18–20].



(a) Raw-to-deployable transformation.

(b) Governance-loss magnitude.

C

Code	Mode	A_i	D_i	G_i	$A_i - D_i$
R/C	Regression/classification	0.173	0.169	0.95	0.004
RDS	Robotic decision support	0.520	0.463	0.75	0.057
GHM	Generative modeling	0.480	0.404	0.65	0.076
OMA	Modular autonomy	0.813	0.640	0.55	0.173
GPA	General-purpose AI	0.938	0.616	0.35	0.322

(c) Governance-adjustment values.

Figure 5. Governance adjustment of raw autonomy into responsible deployability. The figure demonstrates that the largest technical capability does not necessarily produce the strongest responsible deployment score when traceability, validation readiness, and human oversight are weaker.

Figure 5 isolates the governance-adjustment behavior behind the deployability score. Panel (a) shows the transformation from raw autonomy to deployability, panel (b) ranks the resulting autonomy loss $A_i - D_i$, and panel (c) reports the corresponding values. The general-purpose mode loses the largest amount because its raw capability is high while governance assurance is low. Regression/classification loses almost no autonomy because its technical scope is narrower and its validation pathway is more straightforward.

3.4 HGAMI-guided workflow architecture

Figure 6 summarizes the recommended architecture. The workflow begins with a human-defined materials objective because the scientific problem must be formulated before autonomy has value. The objective may be higher ionic con-

ductivity, improved corrosion resistance, reduced catalyst overpotential, enhanced thermal stability, increased mechanical toughness, or improved optoelectronic response. The human researcher then defines descriptors, design boundaries, measurable objectives, and safety constraints. HGAMI scoring is applied before deployment to determine which AI mode is appropriate. The selected workflow is then executed under validation rules, and the experimental outcome returns to the human researcher for interpretation and redesign.

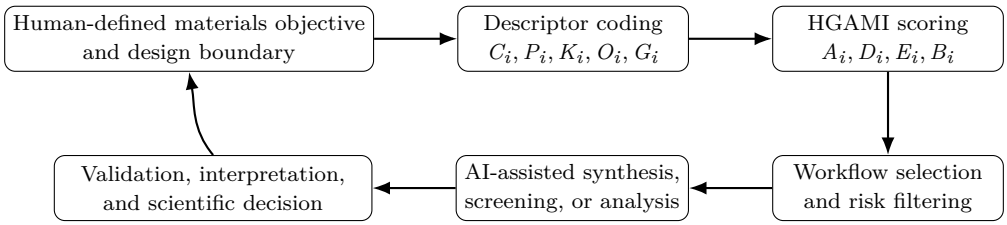


Figure 6. HGAMI-guided architecture for selecting AI-enabled materials-research workflows. The researcher defines the materials objective and boundary conditions, the workflow is encoded using capability and governance descriptors, and AI-assisted execution remains connected to validation and scientific interpretation.

In other words, Fig. 6 converts the numerical result into a laboratory workflow loop. As mentioned above, the loop starts and ends with the materials scientist not because of the marginality of AI but due to the necessity to guarantee scientific relevance. The validity of the materials result is conditional upon the objective, boundary conditions, sample history, measurement precision, and interpretation. In this loop, HGAMI is used pre-execution to choose a responsible workflow and post-execution to guide laboratory strategy refinement.

4 Discussion

4.1 Interpretation of raw autonomy and deployability

The results show that raw autonomy and responsible deployability should not be considered interchangeable terms. In Fig. 3, the important feature of the plot is not the magnitude of each coordinate but the separation between raw-autonomy and deployability axes. On the one hand, general-purpose AI autonomy offers the largest distance to the origin on the raw-autonomy axis. On the other hand, this approach fails to dominate the deployability frontier. Or, to put it differently, the scientific value of orchestrated modular autonomy can be decomposed into accountable modules.

Regression and classification produce the smallest raw autonomy score because the technique is limited primarily to digital inference. However, this does not mean that it is scientific insignificant. In a wide range of materials laboratories, regression and classification models provide the most reliable initial step towards AI integration due to their ease of training, systematic comparability, and ability to compare with measured outcomes [4, 6, 21]. Deployability value is high because governance assurance is also high.

The robotic decision support provides scientific value due to the possibility of building an autonomously executable loop. Namely, the system can suggest the next experimental condition, prepare a candidate for synthesis, run a series of measurements, and update the model. While such a process can reduce labor costs dramatically, deployability shows that there is an increased risk associated with autonomous physical action. Robotic decision support relies on accurate instrument calibration, experimental constraints, measurement controls, sample tracking, and explicit stopping criteria. Without these, the system can quickly collect data without creating any scientific value.

Generative hypothesis modeling provides value for the organization of literature knowledge base, identification of materials analogues, generation of explanatory models, and derivation of design rules. However, lower deployability value reflects the risk of persuasive but non-verifiable hypotheses. Materials science faces fundamental limitations due to physics of materials and processing. Any generated explanation requires physical verification based on existing literature, experimental testing, and materials theory. Thus, the present result suggests a controlled place for generative AI within ideation process.

Finally, the greatest scientific value for new-technology materials discovery can be provided by orchestrated modular autonomy. Indeed, such an approach allows combining several specialized modules without full reliance on a black-box system. Specifically, an autonomous materials workflow can incorporate a database query module, descriptor generator, predictive model, uncertainty estimator, a Bayesian optimizer, robotic execution controller, and validation module. Given such modularity, errors in each component can be separately identified and corrected. This way, modular autonomy increases scientific value compared to full-blown general-purpose AI.

Overall, general-purpose autonomy offers the highest raw autonomy score and highest efficiency indicator but not the highest deployability value. This is the key methodological finding of this paper. Autonomous AI for materials discovery must not only be evaluated for a wide range of tasks accomplished but for an appropriate balance between technical

power and scientific responsibility. The result of this paper suggests that current AI autonomy development focus is wrong. Instead of trying to maximize machine autonomy, researchers should aim at maximizing scientifically informed autonomy.

From the materials discovery perspective, general-purpose autonomy can cause serious damage. Indeed, the use of expensive precursors, hazardous processing techniques, extended aging periods, and irreversible sample synthesis can make an erroneous autonomous decision dangerous. Thus, even the most efficient general-purpose autonomy cannot be used irresponsibly. By contrast, deployability ranking can help in choosing a workflow with detailed explanations, uncertainties accounting, application of constraints, and validation mechanisms. These elements guarantee that laboratory decisions remain scientifically valid despite physical autonomy.

4.2 Capability balance and workflow maturity

Figure 4 explains why HGAMI should not be interpreted as a single ranking. Low value of spread may mean that a workflow is broadly capable or narrowly limited. The difference between the two cases is crucial for materials discovery. Thus, Figure 4 encourages a diagnostic interpretation of the results: pick the workflow which capability profile matches materials bottleneck, and check if the governance is acceptable.

The capability analysis shows that the workflow should not be judged by its average capability. A workflow may be balanced because of its wide capabilities, e.g., orchestration autonomy, or uniform capability limitation, e.g., regression and classification. In the latter case, capability balance should not be confused with workflow maturity. In other words, HGAMI results should be interpreted not independently but in combination with deployability and raw autonomy measures.

Namely, the orchestration-based workflow is balanced in capability but weak in governance. Robotic decision support lacks capability balance because it is focused on physical experimentation. On the other hand, it may be the best option if the materials task requires repetitive synthesis or parameter tuning. Finally, generative hypothesis modeling lacks capability balance but can be useful at an early stage of ideation.

Scientifically speaking, workflow maturity should be determined by the workflow-fit to materials discovery problem. In battery research, physical agency and long-duration validation will be the bottlenecks due to cycling tests, stability issues, and safety considerations. In catalyst discovery, uncertainty-driven experiments and mechanistic explanation will play key roles due to complex interactions and specificity of reaction pathways. In quantum materials, descriptor generation and physical interpretation of generated patterns will dominate. All these scenarios can be reflected in HGAMI calibration with different weights.

4.3 Governance assurance and responsible deployment

Figure 5 presents the most obvious answer to the question why HGAMI is needed. Without a governance adjustment, the general-purpose autonomy would seem to be the most promising option. With a governance adjustment, the reader will notice that not all technical advantage can be readily used because of lack of experimental governance. In other words, HGAMI converts an abstract governance principle into a laboratory-readiness metric.

Robotic systems require a different governance mechanism than a purely digital model. Indeed, while prediction error is an issue, experimental control is also required to validate autonomous decision-making. Any closed-loop robotic optimizer may be rigorous mathematically. Still, the recommendations will be no better than measurement feedback used to update the model. Thus, robotics-assisted laboratories should always use additional validation samples, instrumental calibration, and human supervision in critical decision moments.

Generative models pose a unique governance challenge since the recommendation will usually be linguistically or conceptually expressed. Generated hypothesis can be formally correct but scientifically unsound. In materials discovery, incorrectness of a generated hypothesis can drive the researcher towards a wrong idea based on wrong reasoning. Thus, the generative system requires a special governance structure, which includes literature tracing, contradiction checking, domain-specific constraints, and separation between suggestion and validation. The model must not generate experimental conclusion out of linguistic hypothesis.

By contrast, orchestrating autonomy can take advantage of the governance structure inherent to modularity. Since all workflow modules are separated, their auditability can be ensured independently. This is a considerable practical advantage over general-purpose autonomy because the failure of candidate material requires inspection of each workflow module independently to find the root of the failure. Auditability contributes to the workflow's reproducibility and aligns with high standards of materials discovery.

4.4 Workflow architecture for materials laboratories

The workflow architecture demonstrates that human participation is not a flaw in autonomous laboratory work. Rather, it is a scientific layer that ensures accountability of the research. Humans are the ones who determine the meaningful objective, exclude impossible regions, recognize artifact measurements, judge the reasonableness of proposed explanations,

and assess the usefulness of the outcome. At the same time, humans need assistance with searching, data collection, and pattern recognition.

Materials discovery requires such participation because experimental measurement can be influenced by impurity content, distribution of particle sizes, processing atmosphere, possible surface reconstruction, aging, and measurement orientation. Such factors cannot be ignored in materials discovery because their presence leads to unreliable results and false conclusions.

4.5 Materials-journal relevance and methodological boundaries

In terms of materials-journal relevance, this paper clearly falls into the category of informatics-related materials science. More specifically, HGAMI can be classified as an autonomous-materials decision-modeling method. The reasons are twoffold: (i) Descriptors and weights used in this method can be adjusted easily by anyone familiar with materials informatics; (ii) the method deals with a practical problem of evaluating AI-enabled materials workflows and selecting the most promising options.

The problem of autonomous-materials workflow evaluation is quite relevant for new-technology materials laboratories. Indeed, when deciding on the introduction of autonomous materials, researchers have to consider various approaches available and their relative benefits. This problem is closely related to materials science and engineering because of involvement of materials informatics, database technologies, and experiment management.

The calculated HGAMI values can be interpreted differently. In this work, they serve as structured ordinal descriptors rather than empirical values derived from laboratory measurements. This limitation is important since HGAMI is supposed to offer a tool for pre-laboratory decision-making based on the workflow characteristics. Future calibration of the method may involve experts, retrospective analyses of finished autonomous laboratory work, or cross-domain comparisons.

Machine learning, autonomous experiment, Bayesian optimization, and materials databases are established research directions [1, 5, 8]. The unique contribution of HGAMI is the coupling between capability score and governance-assurance-adjusted deployability. This idea is kept throughout the paper: raw autonomy is the measure of technical breadth, while deployability is the measure of experimental readiness.

This study is directly linked to materials science and engineering because of a tight connection between each workflow type and materials discovery function: property prediction, synthesis planning, characterization, closed-loop optimization, experimental validation, and new-materials discovery. The method can be calibrated for batteries, catalysts, quantum materials, alloys, polymers, and corrosion-resistant materials.

4.6 Practical implications for AI-enabled materials discovery

The following implications may be drawn from the paper:

First, laboratories lacking automation tools should realize that it is premature to give up AI-assisted materials discovery. Regression, classification, uncertainty-guided screening, and interpretable feature analysis can deliver significant scientific value even with minimal automation.

Second, laboratories that have developed robotic or HTS platform should focus on the establishment of experimental boundaries before introducing new autonomous techniques. Robot can improve experiment throughput. Still, throughput does not equal discovery.

Third, generative AI must be used carefully in materials discovery due to the lack of constraints. Its scientific value lies in acceleration of ideation and literature organization, not generation of thermodynamically impossible or energetically unreasonable materials.

Fourth, laboratory workflow architecture can benefit significantly from modular autonomy, in which separate modules can be independently validated and combined in a holistic workflow.

Thus, for new-technology materials laboratories, the ideal AI system seems to be hybrid, i.e., governed by human expertise and powered by machine search.

5 Limitations and future work

Three limitations of this work can be noted:

First, all HGAMI values are ordinal, so they must be interpreted as a methodologically transparent test rather than empirical values.

Second, governance adjustment cannot be universal since it depends on domain-specific constraints, literature tracing, and validation criteria.

Third, efficiency indicator is a purely computational quantity, thus, it cannot be directly measured.

Future work should be aimed at HGAMI calibration with real autonomous or semi-autonomous materials laboratories. The calibration can involve expert elicitation for estimation of HGAMI parameters. It can also involve a retrospective validation in which HGAMI scores are compared to actual experimental performance.

6 Conclusion

In this paper, the Human-Governed Autonomy Mapping Index (HGAMI) was proposed as a decision model for evaluating the relative autonomy of AI-enabled workflows in materials laboratories in terms of technical capability and governance. The findings indicate that the most technically autonomous AI approach does not have to be the best possible choice in terms of being responsible for materials research in the short term. While general-purpose AI autonomy received the highest score in autonomy ($A_i = 0.938$) and efficiency (84.7%), the lower governance score ($G_i = 0.529$) led to worse deployability compared to orchestrated modular autonomy. In this case, the highest deployability score ($D_i = 0.640$) was attained by combining high technical capability with more auditable and controllable nature of a workflow. Therefore, one can conclude that in pursuit of autonomy in the lab, materials researchers need to make sure that they are able to ensure traceability, reproducibility, interpretability, validation, and human accountability. Accordingly, the governed autonomy approach is suggested by HGAMI where humans will formulate the materials objectives, design space, measurement criteria, stopping conditions, and validation strategy and where the AI system is responsible for helping in candidate prioritization, uncertainty quantification, experiment planning, resource management, and pattern detection. Overall, HGAMI enables materials scientists to make a well-reasoned choice between alternative modes of AI-assisted material discovery, synthesis optimization, characterization, and closed-loop experimentation. Further research can focus on benchmarking HGAMI against actual materials labs' performance (hit rates, time to validation, costs per candidate validated, etc.).

Acknowledgements

The authors acknowledge the materials-informatics, autonomous-experimentation, and machine-learning research communities whose contributions support the development of intelligent materials workflows.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All descriptor values and computational outputs used in this study are reported in the tables. No external numerical dataset is required to reproduce the calculations.

References

[1] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Materials, 1* (2013) 011002.

[2] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, AFLOW for high-throughput materials discovery, *Computational Materials Science, 58* (2012) 218–226.

[3] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database, *JOM, 65* (2013) 1501–1509.

[4] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Computational Materials, 2* (2016) 16028.

[5] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: Recent applications and prospects, *npj Computational Materials, 3* (2017) 54.

[6] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, *Nature, 559* (2018) 547–555.

[7] J. Mockus, *Bayesian Approach to Global Optimization: Theory and Applications*, Kluwer Academic Publishers, Dordrecht, 1989.

[8] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, Taking the human out of the loop: A review of Bayesian optimization, *Proceedings of the IEEE, 104* (2016) 148–175.

[9] P. I. Frazier, A tutorial on Bayesian optimization, *arXiv preprint* arXiv:1807.02811 (2018).

[10] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, I. Takeuchi, On-the-fly closed-loop materials discovery via Bayesian active learning, *Nature Communications, 11* (2020) 5966.

[11] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Hase, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, Self-driving laboratory for accelerated discovery of thin-film materials, *Science Advances, 6* (2020) eaaz8867.

[12] N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, Y. Zeng, et al., An autonomous laboratory for the accelerated synthesis of inorganic materials, *Nature, 624* (2023) 86–91.

[13] F. Adams, A. McDannald, I. Takeuchi, A. G. Kusne, Human-in-the-loop for Bayesian autonomous materials phase mapping, *Matter, 7* (2024) 697–709.

[14] S. R. Chitturi, A. Ramdas, Y. Wu, B. Rohr, S. Ermon, J. Dionne, F. H. da Jornada, M. Dunne, C. Tassone, W. Neiswanger, D. Ratner, Targeted materials discovery using Bayesian algorithm execution, *npj Computational Materials, 10* (2024) 156.

[15] E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. Saikin, S. Smullin, V. Stanev, B. Maruyama, Autonomous experimentation systems for materials development: A community perspective, *Matter, 4* (2021) 2702–2726.

[16] L. Hung, J. A. Yager, D. Monteverde, D. Baiocchi, H. Kwon, S. Sun, S. K. Suram, Autonomous laboratories for accelerated materials discovery: A community survey and practical insights, *Digital Discovery, 3* (2024) 1273–1279.

[17] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, P. Schwaller, Augmenting large language models with chemistry tools, *Nature Machine Intelligence, 6* (2024) 525–535.

[18] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems, 30* (2017).

[19] M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) 1135–1144.

[20] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, *arXiv preprint* arXiv:1702.08608 (2017).

[21] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science, *APL Materials, 4* (2016) 053208.