International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 1, Issue 2 December 2024 ISSN: 3048-975X
https://ijrasht.com/

AI for Scientific Discovery: A Survey of Methods, Applications, and Open Challenges

Karan, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822377.cse@pietgroup.co.in

Anjali, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822392.cse@pietgroup.co.in

Ashish, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822403.cse@pietgroup.co.in

Manish, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822448.cse@pietgroup.co.in

Nitika, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India nitika.cse-aiml@piet.co.in

Abstract— Artificial intelligence is increasingly used not merely to automate existing scientific workflows but to actively accelerate the process of scientific discovery itself, from proposing candidate drug molecules and novel materials to inferring governing physical laws directly from experimental data. This paper surveys the growing field of AI for scientific discovery, organizing methods into a taxonomy spanning generative molecular design, graph-based property prediction, structure prediction, symbolic regression, and experiment design through active learning and Bayesian optimization. We review applications across drug discovery, materials science, physics, astronomy, climate science, and chemistry, summarize commonly used benchmarks, and present a worked case study illustrating how a generative model, a property predictor, and an active learning loop can be combined into an end-to-end discovery pipeline. We discuss open challenges including data scarcity in specialized scientific domains, the gap between computational prediction and experimental validation, and the difficulty of ensuring that AI-proposed hypotheses respect known physical constraints. We conclude by outlining future directions connecting AI for science with foundation models, automated laboratories, and human-AI collaborative discovery workflows.

Index Terms—AI for science, scientific discovery, molecular generation, materials informatics, symbolic regression, active learning, protein structure prediction

I. INTRODUCTION

Scientific progress has historically depended on a cycle of hypothesis generation, experimentation, and refinement, a process that is often slow, expensive, and constrained by the limits of human intuition when navigating extremely large and high-dimensional spaces of possible molecules, materials, or physical models. The space of drug-like small molecules, for instance, is estimated to contain on the order of 10^60 candidate compounds, an astronomically large search space that no amount of manual laboratory synthesis could ever exhaustively explore. Artificial intelligence, and deep learning in particular, offers tools well suited to navigating such spaces: generative models can propose novel candidate structures, predictive models can estimate properties of those candidates far faster than physical experimentation or first-principles simulation, and active learning strategies can prioritize which candidates are most valuable to test experimentally.

The term "AI for science" has come to describe a broad and rapidly growing research program applying machine learning across essentially every scientific discipline, including chemistry, biology, materials science, physics, astronomy, and climate science. Unlike many mainstream machine learning applications, AI for science typically must contend with data that is comparatively scarce relative to web-scale corpora, must respect known physical laws and constraints that a purely data-driven model might otherwise violate, and must ultimately connect back to physical experimentation or first-principles theory for validation, since a computational prediction alone rarely constitutes scientific proof.

This paper surveys the state of AI for scientific discovery. Section II reviews the background motivating the use of AI in science and the historical trajectory of the field. Section III introduces a taxonomy of methodological approaches. Section IV describes core algorithmic techniques in technical detail. Section V surveys applications across scientific domains. Section VI summarizes benchmarks. Section VII presents a comparative discussion. Section VIII walks through an illustrative end-to-end case study. Section IX discusses implementation tooling. Section X considers ethical

and societal implications. Section XI discusses open challenges, Section XII outlines future directions, and Section XIII concludes.

A defining feature of this field, distinguishing it from purely predictive machine learning applications, is that its ultimate goal is not simply to fit a function to data but to generate genuinely novel and experimentally validated scientific knowledge — new molecules that are subsequently synthesized and tested, new materials that are subsequently fabricated, or new physical relationships that are subsequently confirmed through independent experimentation. This orientation toward discovery, rather than prediction alone, introduces distinctive methodological demands, including the need for models that can extrapolate meaningfully beyond their training distribution into genuinely novel regions of chemical or physical space, a requirement quite different from the interpolation-focused evaluation common in many standard machine learning benchmarks.

II. BACKGROUND AND MOTIVATION

A. The Traditional Scientific Method and Its Bottlenecks

Classical experimental science proceeds through iterative cycles of hypothesis formation, experimental design, data collection, and analysis, a process that has produced immense scientific progress but that faces well-documented bottlenecks in several domains. Drug discovery, for example, has historically required, on average, well over a decade and enormous financial investment to bring a single new medication from initial discovery to regulatory approval, with the overwhelming majority of candidate compounds failing at some stage of this pipeline, often after substantial resources have already been invested. Materials discovery faces an analogous challenge: identifying a new material with a desired combination of properties, such as high conductivity and mechanical strength, has traditionally relied heavily on chemical intuition and trial-and-error experimentation, since the space of possible material compositions and structures is far too large to explore exhaustively through physical synthesis alone.

These bottlenecks share a common structure: an enormous combinatorial search space, an expensive and slow evaluation function corresponding to physical experimentation or high-fidelity simulation, and a scientific community whose collective intuition, while valuable, cannot fully compensate for the scale of the search problem. This structure is precisely the setting in which machine learning methods for generation, prediction, and optimization have proven most valuable, motivating the rapid growth of AI applications across the physical and life sciences over the past decade.

B. Why AI Is Well Suited to Scientific Discovery

Modern deep learning models excel at learning complex, high-dimensional functions from data, which maps naturally onto the problem of predicting a molecule's binding affinity, a material's band gap, or a physical system's future state from a learned representation of its structure. Generative models extend this capability from prediction to proposal, learning to sample novel candidate structures from the same underlying distribution as a training set of known molecules or materials, or, in more advanced formulations, to generate candidates optimized toward a desired property that may not be well represented in the training data at all. Meanwhile, techniques from Bayesian optimization and active learning provide principled strategies for deciding which of many possible experiments to run next, given a limited experimental budget, directly addressing the core bottleneck of expensive physical validation.

AI methods also offer the practical advantage of operating over representations that can incorporate substantial existing scientific knowledge, for example through physics-informed neural network architectures that hard-code known conservation laws, or through graph neural network architectures that encode molecular structure directly as a graph of atoms and bonds, allowing the model to exploit known chemical structure rather than learning it entirely from data, which is particularly valuable in specialized scientific domains where labeled data is far scarcer than in mainstream computer vision or natural language processing applications.

C. Historical Context

Early applications of AI to scientific problems date back to expert systems of the 1970s and 1980s designed to assist with tasks such as organic chemical structure elucidation from spectroscopic data, encoding expert chemical

knowledge as symbolic rules. The subsequent shift toward statistical and, later, deep learning methods enabled data-driven approaches that did not require hand-crafted rules, with quantitative structure-activity relationship models becoming a standard tool in computational chemistry by the 1990s and 2000s. The current wave of AI for science, distinguished by the use of deep generative models, graph neural networks, and large pretrained foundation models, has been substantially accelerated by a small number of high-profile breakthroughs, most notably the substantial improvement in computational protein structure prediction accuracy achieved by deep learning systems in the early 2020s, which demonstrated to the broader scientific community that deep learning could make qualitative, rather than merely incremental, contributions to long-standing scientific problems.

This breakthrough had a catalytic effect on funding and institutional attention across the physical and life sciences, prompting major research funding agencies, pharmaceutical companies, and national laboratories to substantially expand investment in AI-for-science research programs, and prompting a wave of new interdisciplinary research centers explicitly organized around combining machine learning expertise with domain-specific scientific expertise. This institutional shift is itself notable: unlike earlier machine learning application areas that grew primarily from within computer science departments, AI-for-science has increasingly been co-developed by machine learning researchers working in close, sustained collaboration with domain scientists, a collaborative model that has proven important for ensuring that resulting methods respect domain-specific constraints and are validated against scientifically meaningful benchmarks rather than benchmarks that are convenient from a purely machine learning perspective.

III. A TAXONOMY OF AI-FOR-SCIENCE APPROACHES

AI for scientific discovery methods can be organized along two complementary dimensions: the scientific task being addressed, and the underlying machine learning paradigm employed. Along the task dimension, the field encompasses property prediction, in which a model estimates a scientifically relevant quantity, such as a molecule's toxicity or a material's thermal conductivity, from its structure; generative design, in which a model proposes novel candidate structures, optionally optimized toward desired properties; structure and law discovery, in which a model infers an underlying structural configuration, such as a protein's three-dimensional fold, or a governing mathematical law from data; and experiment design, in which a model decides which experiments or simulations to prioritize given a limited budget.

Along the methodological dimension, approaches range from graph-based deep learning, which represents scientific entities such as molecules or crystal structures directly as graphs to exploit their natural relational structure, to physics-informed methods, which embed known physical constraints directly into model architecture or training objective, to foundation-model approaches, which pretrain large models on broad scientific corpora, such as protein sequence databases or chemical reaction datasets, before fine-tuning for specific downstream discovery tasks. Table I summarizes this taxonomy with representative methods for each combination of task and paradigm.

It is useful to note that these two dimensions, task and paradigm, are not independent in practice: certain paradigms have historically proven better suited to certain tasks, and much of the field's methodological innovation has come from successfully applying a paradigm associated with one task to a different task where it was not originally used. For example, techniques originally developed for property prediction, such as equivariant graph neural network architectures, have since been adapted as components within generative design pipelines, where a learned property predictor serves as a differentiable reward signal guiding the generative model toward candidates with desired characteristics, illustrating how the taxonomy's dimensions increasingly interact rather than remaining cleanly separated.

TABLE I. TAXONOMY OF AI-FOR-SCIENCE TASKS AND METHODOLOGICAL PARADIGMS

Task	Representative Paradigm	Example Methods	Typical Domain
Property prediction	Graph neural networks	Message-passing neural networks, SchNet	Chemistry, materials science

Task	Representative Paradigm	Example Methods	Typical Domain
Generative design	Deep generative models	Variational autoencoders, diffusion models, GANs	Drug and materials design
Structure/law discovery	Specialized architectures	AlphaFold-style structure networks, symbolic regression	Structural biology, physics
Experiment design	Bayesian optimization	Gaussian process bandits, active learning loops	Autonomous experimentation

IV. CORE METHODS

A. Deep Generative Models for Molecular and Materials Design

Deep generative models learn a probability distribution over molecular or material structures from a training dataset and allow sampling of novel structures from this learned distribution. Variational autoencoders map discrete or graph-structured molecular representations into a continuous latent space in which optimization or interpolation can be performed more easily than in the original discrete representation, after which a decoder maps optimized latent points back into valid molecular structures. Generative adversarial networks and, more recently, diffusion models have also been adapted to molecular and crystal structure generation, with diffusion-based approaches in particular showing strong results for generating three-dimensional atomic configurations by iteratively refining a noisy initial structure toward a physically plausible one.

A key practical challenge in generative molecular design is ensuring that generated outputs correspond to chemically valid structures, since a naive generative model operating over an unconstrained representation can easily produce syntactically invalid molecules, such as atoms with impossible valence configurations. This has motivated representations and decoding procedures that enforce chemical validity by construction, as well as reinforcement-learning-based fine-tuning procedures that optimize a pretrained generative model toward a desired downstream property, such as predicted binding affinity, while a validity or synthesizability penalty discourages the model from drifting into chemically implausible regions of structure space.

B. Graph Neural Networks for Property Prediction

Molecules and crystal structures are naturally represented as graphs, with atoms as nodes and chemical bonds or spatial proximity as edges, making graph neural networks a natural architectural choice for predicting molecular and material properties. Message-passing neural networks iteratively update each atom's representation by aggregating information from neighbouring atoms and bonds, allowing the network to learn representations that reflect local chemical environment, and after several rounds of message passing, a graph-level readout function produces a prediction for a target property such as solubility, toxicity, or formation energy.

Extensions of this basic architecture incorporate three-dimensional geometric information, such as bond angles and distances, using equivariant neural network architectures that respect the physical symmetries of rotation and translation, ensuring that a model's prediction for a given molecule does not depend on the arbitrary orientation in which that molecule's coordinates happen to be represented. These geometric and equivariant architectures have substantially improved prediction accuracy for properties that depend sensitively on precise three-dimensional structure, such as quantum mechanical energies computed from density functional theory.

C. AI for Protein Structure and Function Prediction

Predicting a protein's three-dimensional structure from its amino acid sequence, long considered one of the most significant unsolved problems in computational biology, has been substantially advanced by deep learning systems that combine attention-based architectures with structural and evolutionary information derived from large databases

of related protein sequences. These systems learn to predict spatial relationships between amino acid residues and iteratively refine a predicted three-dimensional structure to satisfy these predicted relationships along with known geometric constraints of protein backbones, achieving prediction accuracy for many protein families that approaches the accuracy of experimental structure determination methods such as X-ray crystallography.

Beyond structure prediction, related deep learning approaches address protein function prediction, protein-protein interaction prediction, and de novo protein design, in which a model generates an entirely novel amino acid sequence expected to fold into a desired target structure or to perform a desired biochemical function, extending the paradigm from structure prediction to structure-conditioned generative design, analogous to the shift from property prediction to generative design in small-molecule chemistry.

D. Symbolic Regression and Automated Law Discovery

Symbolic regression seeks to discover a closed-form mathematical expression that fits observed experimental data, in contrast to standard regression techniques that fit parameters within a fixed functional form chosen in advance. Genetic programming approaches to symbolic regression evolve a population of candidate mathematical expressions, represented as expression trees, through operations analogous to biological mutation and crossover, selecting expressions that best balance fit to the data against a preference for concise, parsimonious forms. More recent neural approaches to symbolic regression use transformer-based architectures trained on large synthetic datasets of mathematical expressions and corresponding numerical data, learning to translate a table of numerical observations directly into a candidate symbolic expression.

Symbolic regression has been used to rediscover several well-known physical laws directly from raw experimental or simulated data without prior knowledge of the correct functional form, demonstrating the potential for AI systems to contribute not just to prediction but to identifying interpretable, generalizable scientific relationships, a capability of particular interest for scientific domains where an interpretable closed-form law is valued not only for its predictive accuracy but for the physical insight it can offer to human researchers.

E. Active Learning and Bayesian Optimization for Experiment Design

Because physical experiments and high-fidelity simulations are typically far more expensive than evaluating a trained machine learning model, active learning and Bayesian optimization techniques are used to select which candidate experiments to prioritize given a limited budget. Bayesian optimization maintains a probabilistic surrogate model, commonly a Gaussian process, over the relationship between experimental conditions and outcomes, and uses this surrogate model's predicted mean and uncertainty to select the next experiment expected to be most informative, balancing exploitation of currently promising regions of the search space against exploration of regions where the surrogate model remains highly uncertain.

These techniques have been integrated into closed-loop autonomous experimentation platforms, sometimes described as self-driving laboratories, in which a machine learning system proposes an experiment, a robotic platform physically executes it, the resulting measurement is fed back into the surrogate model, and the cycle repeats with minimal human intervention, substantially accelerating the pace at which large experimental parameter spaces, such as those encountered in catalyst or battery electrolyte formulation, can be explored relative to manual experimentation.

F. Foundation Models for Science

Following the success of large pretrained foundation models in natural language processing, the AI for science community has developed analogous foundation models pretrained on large scientific corpora, such as protein sequence databases, chemical reaction datasets, or astronomical survey data, before being fine-tuned or prompted for specific downstream scientific tasks. These models aim to capture broadly useful scientific representations from large volumes of unlabelled or weakly labelled data, which can then be adapted with comparatively little task-specific labelled data, addressing the data scarcity that characterizes many specialized scientific applications relative to mainstream machine learning domains such as web-scale text or image corpora.

V. APPLICATIONS

A. Drug Discovery

AI methods are used throughout the drug discovery pipeline, from virtual screening of large compound libraries to identify candidates likely to bind a target protein, to generative design of novel candidate molecules optimized for predicted binding affinity, synthesizability, and favourable pharmacokinetic properties, to prediction of toxicity and off-target effects earlier in the pipeline than would otherwise be possible through experimental screening alone.

Several AI-designed drug candidates have progressed into clinical trials, and industry adoption of AI-assisted discovery pipelines has grown substantially, with major pharmaceutical companies and dedicated AI-drug-discovery startups reporting meaningful reductions in the time required to identify promising candidate molecules for a given therapeutic target, although the ultimate clinical success rate of AI-discovered candidates relative to traditionally discovered candidates remains an active area of ongoing evaluation given the long timelines of clinical development.

B. Materials Science

In materials science, AI methods support the discovery of new battery electrode and electrolyte materials, catalysts for chemical reactions such as ammonia synthesis or carbon capture, and structural materials with optimized combinations of strength, weight, and durability, typically by combining property-predicting graph neural networks with generative or search-based methods to propose novel candidate compositions and crystal structures.

Large-scale materials discovery efforts have used AI-guided screening to identify thousands of previously unknown, computationally stable candidate materials from a vastly larger space of hypothetical compositions, illustrating the scale advantage that AI-driven virtual screening offers relative to purely experimental materials discovery, though experimental synthesis and validation of computationally proposed materials remains a significant bottleneck that limits how much of this computational output can currently be confirmed and put to practical use.

C. Physics and Astronomy

In physics, machine learning is used to accelerate computationally expensive simulations, such as those used in fluid dynamics, plasma physics, and lattice quantum chromodynamics, by learning fast surrogate models that approximate the outputs of expensive numerical solvers, and to assist in the analysis of experimental data from large physics facilities, such as particle detectors, where AI-based pattern recognition helps identify rare events of scientific interest within enormous volumes of collected data.

In astronomy, AI methods are used for tasks such as automated classification of celestial objects from large-scale sky survey imagery, detection of exoplanets from stellar brightness time series, and the search for anomalous or previously unknown classes of astronomical phenomena that might not conform to existing classification schemes, leveraging the ability of unsupervised and semi-supervised learning methods to flag unusual patterns in data too voluminous for manual human inspection.

Learned simulation surrogates have proven particularly valuable in settings where a single high-fidelity simulation run may require days or weeks of supercomputer time, since a trained neural surrogate can often produce an approximate result in seconds, enabling scientists to explore far larger regions of parameter space, for example when searching for physical parameter values consistent with an observed astronomical signal, than would be feasible using the original expensive simulator alone.

D. Climate and Earth Science

AI methods are increasingly integrated into climate and weather modeling, both to accelerate computationally expensive numerical simulations through learned surrogate models and to directly improve short- and medium-term weather forecasting accuracy using deep learning models trained on historical atmospheric data, with several recent AI-based weather forecasting systems demonstrating forecast accuracy competitive with, and in some metrics exceeding, traditional physics-based numerical weather prediction models at a small fraction of the computational cost.

Beyond weather forecasting, AI is applied to broader earth science problems including satellite-based monitoring of deforestation and land-use change, prediction of extreme weather events, and downscaling of coarse global climate model outputs to finer spatial resolutions relevant for local adaptation planning.

A distinguishing characteristic of AI applications in climate science, relative to many other AI-for-science domains, is the exceptionally high stakes attached to forecast reliability under conditions that may differ substantially from the historical training data, since climate change is expected to produce increasingly frequent extreme events outside the range of historical experience, placing particular importance on rigorous out-of-distribution evaluation and close collaboration with domain climate scientists when interpreting model outputs intended to inform adaptation and mitigation policy decisions.

E. Chemistry and Reaction Prediction

AI models are used to predict the outcome of chemical reactions, to plan multi-step synthetic routes for producing a target molecule from available starting materials, and to identify plausible reaction mechanisms consistent with experimental observations, tasks that traditionally relied heavily on the accumulated experience of expert synthetic chemists and that can now be substantially accelerated using models trained on large databases of published chemical reactions.

VI. BENCHMARKS AND EVALUATION

Evaluating AI-for-science methods requires benchmarks that reflect the specific scientific task being addressed, and the field has developed a range of standardized datasets spanning molecular property prediction, materials property prediction, and structure prediction. Table II summarizes several widely used benchmarks. A recurring theme across these benchmarks is the tension between benchmark performance and genuine scientific utility: strong performance on a held-out test set drawn from the same distribution as the training data does not guarantee that a model will extrapolate reliably to the genuinely novel molecules or materials that a real discovery application requires it to propose or evaluate, since scientific discovery is, almost by definition, concerned with venturing beyond the previously characterized distribution.

The MoleculeNet benchmark suite aggregates multiple molecular property prediction datasets spanning quantum mechanical, physical chemistry, biophysical, and physiological properties, providing a standardized basis for comparing molecular representation learning methods. The Materials Project and Open Catalyst Project datasets provide large collections of computed material and catalytic surface properties derived from density functional theory calculations, used to train and evaluate property-predicting graph neural networks at scale. For structure prediction, the Critical Assessment of Protein Structure Prediction competition has served for decades as the primary independent benchmark against which computational protein structure prediction methods are evaluated against blind, previously unpublished experimental structures.

TABLE II. REPRESENTATIVE BENCHMARKS IN AI FOR SCIENTIFIC DISCOVERY

Benchmark	Domain	Task
MoleculeNet	Chemistry	Molecular property prediction
Materials Project	Materials science	Formation energy and stability prediction
Open Catalyst Project	Catalysis	Catalytic surface reaction energy prediction
CASP	Structural biology	Blind protein structure prediction
USPTO reaction datasets	Chemistry	Reaction outcome and retrosynthesis prediction

VII. COMPARATIVE ANALYSIS AND DISCUSSION

Generative and predictive AI-for-science methods involve a characteristic trade-off between the breadth of structures they can consider and the reliability of their predictions in genuinely novel regions of structure space. Purely data-driven graph neural network predictors tend to achieve high accuracy on molecules or materials similar to their training distribution but can degrade substantially, sometimes without clear warning, when applied to structurally novel candidates, precisely the regime most relevant to genuine discovery. Physics-informed approaches, which embed known physical constraints directly into the model, tend to generalize more reliably to novel structures because they are not relying solely on statistical similarity to training examples, but they require that the relevant physical constraints be known and correctly specified in advance, which is not always the case for the more empirically characterized properties, such as biological toxicity, that are often of greatest practical interest.

A related trade-off concerns computational cost versus fidelity. Machine learning surrogate models can evaluate candidate structures orders of magnitude faster than first-principles simulation or physical experimentation, enabling screening of vastly larger candidate sets, but this speed advantage is only useful to the extent that the surrogate model's predictions are sufficiently accurate to correctly prioritize genuinely promising candidates, an accuracy that is difficult to guarantee without ongoing experimental or high-fidelity-simulation validation, particularly as a discovery campaign moves into regions of structure space increasingly distant from previously characterized examples.

A third dimension of comparison concerns interpretability. Symbolic regression and physics-informed methods tend to produce results that are directly interpretable by domain scientists, since a discovered closed-form expression or a constraint-respecting prediction can be inspected and reasoned about using established scientific vocabulary, whereas purely data-driven deep learning predictors, including most graph neural network property predictors, generally produce accurate but comparatively opaque predictions whose internal reasoning is difficult for a domain scientist to directly inspect or validate against existing theoretical understanding. This interpretability gap is not merely an academic concern: domain scientists evaluating whether to trust and act upon an AI-generated prediction, particularly one that contradicts prior expectations, are often more willing to do so when they can trace at least a partial theoretical or mechanistic justification for the prediction, which purely black-box predictors are less able to provide.

TABLE III. COMPARATIVE SUMMARY OF AI-FOR-SCIENCE METHOD FAMILIES

Method Family	Novel-Structure Reliability	Computational Cost	Data Requirement
Purely data-driven GNN predictors	Moderate to low outside training distribution	Low at inference	High labelled data
Physics-informed models	Higher, constraint-consistent	Low to moderate	Lower, constraints substitute for data
Generative design models	Depends on decoder validity checks	Low per sample	Moderate to high
Foundation models (pretrained)	Improving with scale	High pretraining, low fine-tuning	Large unlabelled + modest labelled

VIII. ILLUSTRATIVE CASE STUDY

Consider a representative end-to-end AI-driven materials discovery workflow aimed at identifying a new solid-state electrolyte material with high ionic conductivity for use in a next-generation battery. The workflow begins with a generative model, trained on a database of known crystal structures, proposing a large set of candidate compositions and structures that share statistical and structural similarity with known ionic conductors while introducing novel variations in composition or atomic arrangement not present in the training database.

Each proposed candidate is then evaluated by a graph neural network property predictor, trained on a database of density-functional-theory-computed ionic conductivity values, which rapidly assigns an estimated conductivity score to each of the many thousands of generated candidates, a task that would be computationally prohibitive if every candidate required a full first-principles simulation. Candidates with the highest predicted conductivity, together with

a subset of candidates selected specifically because the property predictor's uncertainty estimate is high in their vicinity of structure space, are then prioritized for follow-up evaluation using higher-fidelity density functional theory simulation, an application of the active learning principle described in Section IV that balances exploiting currently promising predictions against exploring regions where the surrogate model's reliability is least certain.

Only after this multi-stage computational funnel narrows many thousands of initial candidates down to a small handful of the most promising structures does the workflow proceed to physical laboratory synthesis and experimental measurement of ionic conductivity, at which point the experimental outcome is fed back into the training data for the property predictor, incrementally improving its accuracy for future discovery campaigns. This case study illustrates a general pattern common to many AI-for-science applications: generative proposal, fast computational filtering, uncertainty-guided prioritization, and a final experimental validation step that remains, at least currently, an irreducible bottleneck that AI accelerates but does not eliminate.

IX. IMPLEMENTATION CONSIDERATIONS AND TOOLING

A. Software Frameworks and Databases

The AI-for-science ecosystem has benefited from the growth of specialized open-source software libraries for representing and manipulating molecular and materials structures, computing standard molecular descriptors, and constructing graph neural network architectures tailored to chemical and physical data, alongside large public databases of experimentally or computationally characterized molecules, materials, and protein structures that serve as training and benchmark data for the field.

B. Integration with Experimental Workflows

Realizing the full potential of AI for scientific discovery requires tight integration between computational prediction pipelines and physical experimental infrastructure, an integration that remains organizationally and technically nontrivial, since experimental laboratories have traditionally been organized around manual or semi-automated workflows not originally designed to interface with a rapid, iterative computational discovery loop. Investment in laboratory automation and standardized data formats for experimental results is an important, if less visible, complement to algorithmic research progress in enabling AI-for-science methods to translate into accelerated real-world discovery.

C. Data Curation and Quality Control

Because scientific training datasets are frequently aggregated from multiple experimental sources, published literature, and computational databases that vary in measurement protocol, precision, and quality control standard, careful data curation is often as important to model performance as algorithmic innovation itself. Inconsistent units, measurement conditions, or annotation conventions across data sources can introduce noise that a model may struggle to distinguish from genuine scientific signal, and several research groups within the AI-for-science community have specifically focused on developing curated, harmonized benchmark datasets and data quality standards intended to reduce this often underappreciated source of model error, an effort that, while less prominent than headline-grabbing new model architectures, is widely regarded within the field as foundational to reliable progress.

X. ETHICAL AND SOCIETAL CONSIDERATIONS

AI for scientific discovery raises dual-use concerns distinct from many other machine learning applications, since the same generative and predictive methods used to discover beneficial new medicines or materials could, in principle, be redirected toward identifying harmful compounds, including toxic chemicals or biological agents. Responsible development of AI-for-science tools increasingly involves deliberate consideration of dual-use risk during model design and release, including decisions about which model weights, training data, or generative capabilities to make openly available versus restricting to vetted research collaborations.

A further consideration concerns equitable access to the substantial computational resources required to train and deploy large AI-for-science models, which risks concentrating the benefits of AI-accelerated discovery among well-

resourced institutions and companies unless deliberate efforts are made to broaden access, for example through shared computational infrastructure, open datasets, and open-source pretrained models that lower the barrier to entry for researchers at less well-resourced institutions, including those in low- and middle-income countries that may bear a disproportionate burden of diseases or environmental challenges that AI-accelerated discovery could help address.

A related concern involves the allocation of scientific credit and intellectual property in an increasingly AI-assisted discovery process. When a generative model proposes a novel candidate molecule or material that is subsequently synthesized and validated, questions arise regarding appropriate attribution among the human researchers who curated training data, developed the model, and interpreted its outputs, and the AI system itself, an issue that intersects with ongoing broader debates about AI-assisted authorship and invention that current scientific publication norms and intellectual property law were not originally designed to address, and that professional societies and journals in several scientific disciplines have only recently begun to formally consider.

XI. CHALLENGES AND OPEN PROBLEMS

A central challenge across nearly all AI-for-science applications is data scarcity relative to mainstream machine learning domains: many scientifically important properties are expensive and slow to measure experimentally, resulting in labelled datasets that are orders of magnitude smaller than those used to train state-of-the-art models in computer vision or natural language processing, which constrains model capacity and increases the risk of overfitting to idiosyncrasies of a small training set rather than learning genuinely generalizable scientific relationships.

A second challenge is out-of-distribution generalization to genuinely novel structures, which, as discussed in Section VII, is precisely the regime in which purely data-driven models are least reliable but that genuine scientific discovery inherently requires.

A third challenge concerns physical plausibility: purely data-driven generative models can propose structures that are statistically consistent with training data patterns but that violate basic physical or chemical constraints, such as impossible bond geometries or thermodynamically unstable configurations, requiring careful validation or physics-informed constraints to filter such implausible outputs before they are considered for expensive experimental follow-up.

A fourth, often underappreciated challenge concerns the gap between computational prediction and experimental reality. Many AI-for-science methods are validated primarily against other computational benchmarks, such as density-functional-theory-computed properties, which themselves carry approximation errors relative to true physical measurements, meaning strong benchmark performance does not automatically guarantee accurate real-world experimental outcomes, and closing this simulation-to-experiment gap remains an active and important area of ongoing research across nearly every scientific domain to which AI has been applied.

A fifth challenge concerns reproducibility and independent verification of AI-driven scientific claims. Because many AI-for-science models involve complex, large-scale training pipelines and proprietary or partially disclosed training datasets, independently reproducing a reported result can be considerably more difficult than reproducing a traditional computational chemistry or physics calculation performed using openly documented methods, raising concerns about the broader scientific community's ability to verify and build upon AI-generated scientific claims with the same rigor traditionally expected of published computational or experimental results.

XII. FUTURE RESEARCH DIRECTIONS

One promising direction involves deeper integration of AI-for-science methods with autonomous, closed-loop laboratory platforms, extending the self-driving laboratory concept introduced in Section IV from individual research demonstrations toward broadly deployed infrastructure that continuously generates new experimental data to refine predictive and generative models with minimal manual intervention, potentially compressing discovery timelines substantially relative to conventional manually operated laboratory workflows.

A second direction explores scientific foundation models trained jointly across multiple modalities relevant to a given scientific domain, such as combining molecular structure, textual scientific literature, and experimental measurement

data within a single pretrained model, with the goal of enabling more flexible, natural-language-guided scientific discovery workflows in which a researcher can query a model for candidate hypotheses or experimental designs using ordinary scientific language rather than specialized computational tooling.

A further direction concerns human-AI collaborative discovery workflows that explicitly preserve a central role for human scientific judgment, using AI systems to rapidly generate and filter candidate hypotheses while relying on human researchers to apply domain expertise, contextual judgment, and creative scientific insight that current AI systems do not reliably replicate, a collaborative framing that several researchers in the field view as more realistic and more likely to be broadly adopted by the scientific community than fully autonomous AI-driven discovery in the near term. Finally, continued progress on uncertainty quantification for scientific machine learning models, allowing a model to reliably flag when a given prediction falls outside its region of trustworthy competence, is likely to remain a high-priority research direction given the central importance of extrapolation reliability to genuine scientific discovery applications.

A final direction worth highlighting concerns the standardization of evaluation practice across the AI-for-science community. Because different subfields have historically developed benchmarks and evaluation conventions independently, often prioritizing metrics convenient for machine learning comparison over metrics that best reflect downstream scientific utility, there is growing interest in developing cross-domain evaluation frameworks that explicitly measure a method's ability to generalize to novel structures, its calibration under distributional shift, and its ultimate rate of successful experimental validation rather than benchmark accuracy alone, which would provide a more scientifically meaningful basis for comparing methods and would help the field avoid over-indexing on benchmark improvements that do not translate into genuine acceleration of real-world scientific discovery.

XIII. CONCLUSION

AI for scientific discovery represents one of the most consequential applications of modern machine learning, extending the field's impact from prediction and pattern recognition toward the active generation of new scientific knowledge. This survey reviewed the taxonomy of methodological approaches spanning generative molecular and materials design, graph-based property prediction, protein structure prediction, symbolic regression, and active-learning-guided experiment design, examined applications across drug discovery, materials science, physics, astronomy, climate science, and chemistry, and presented a worked case study illustrating how these components combine into an end-to-end discovery pipeline. While the field faces significant challenges related to data scarcity, out-of-distribution generalization, physical plausibility, and the persistent gap between computational prediction and experimental validation, its continued integration with foundation models, autonomous laboratories, and human-AI collaborative workflows suggests that AI will play an increasingly central and durable role in accelerating the pace of scientific discovery across essentially every quantitative scientific discipline.

REFERENCES

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, 2021.
[2] R. Gomez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernandez-Lobato, B. Sanchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," ACS Central Science, vol. 4, 2018.
[3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in Proc. ICML, 2017.
[4] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Muller, "SchNet: A deep learning architecture for molecules and materials," Journal of Chemical Physics, vol. 148, 2018.
[5] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," Chemical Science, vol. 9, 2018.
[6] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Materials, vol. 1, 2013.
[7] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, et al., "Open Catalyst 2020 (OC20) dataset and community challenges," ACS Catalysis, vol. 11, 2021.
[8] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) — Round XIV," Proteins, vol. 89, 2021.
[9] M. Schwaller, T. Laino, T. Gaudin, D. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," ACS Central Science, vol. 5, 2019.
[10] C. W. Coley, W. H. Green, and K. F. Jensen, "Machine learning in computer-aided synthesis planning," Accounts of Chemical Research, vol. 51, 2018.
[11] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science, vol. 324, 2009.

[12] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," Science Advances, vol. 6, 2020.
[13] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, "Discovering symbolic models from deep learning with inductive biases," in Proc. NeurIPS, 2020.
[14] B. Kim, S. Azizi, C. Lee, and W. Chueh, "Bayesian optimization for materials science," in Materials Discovery and Design, Springer, 2018.
[15] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, and A. I. Cooper, "A mobile robotic chemist," Nature, vol. 583, 2020.
[16] R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, F. Yao, and A. Aspuru-Guzik, "Data-driven strategies for accelerated materials design," Accounts of Chemical Research, vol. 54, 2021.
[17] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3D neural networks," Nature, vol. 619, 2023.
[18] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al., "Learning skillful medium-range global weather forecasting," Science, vol. 382, 2023.
[19] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven Earth system science," Nature, vol. 566, 2019.
[20] R. J. Murphy, K. T. Butler, and A. Walsh, "Machine learning for materials discovery: A review of state of the art and challenges," APL Materials, vol. 8, 2020.
[21] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature, vol. 559, 2018.
[22] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," Journal of Computational Chemistry, vol. 38, 2017.
[23] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," Drug Discovery Today, vol. 23, 2018.
[24] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," Nature Reviews Drug Discovery, vol. 18, 2019.
[25] T. Blaschke, J. Arus-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist, K. Papadopoulos, and A. Patronov, "REINVENT 2.0: An AI tool for de novo drug design," Journal of Chemical Information and Modeling, vol. 60, 2020.
[26] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," Science, vol. 361, 2018.
[27] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, "Scaling deep learning for materials discovery," Nature, vol. 624, 2023.
[28] C. J. Bartel, "Review of computational approaches to predict the thermodynamic stability of inorganic solids," Journal of Materials Science, vol. 57, 2022.
[29] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds," arXiv preprint, 2018.
[30] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) equivariant graph neural networks," in Proc. ICML, 2021.
[31] J. Hoffmann, L. Bruno, and P. Battaglia, "Learning to simulate complex physics with graph networks," in Proc. ICML, 2020.
[32] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Scientific Data, vol. 1, 2014.
[33] K. Rajan, "Materials informatics: The materials "gene" and big data," Annual Review of Materials Research, vol. 45, 2015.
[34] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Computational Materials, vol. 2, 2016.
[35] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for accurate and interpretable prediction of material properties," Physical Review Letters, vol. 120, 2018.
[36] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gomez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in Proc. NeurIPS, 2015.
[37] E. J. Bjerrum and R. Threlfall, "Molecular generation with recurrent neural networks (RNNs)," arXiv preprint, 2017.
[38] N. De Cao and T. Kipf, "MolGAN: An implicit generative model for small molecular graphs," arXiv preprint, 2018.
[39] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3D," in Proc. ICML, 2022.
[40] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," Science, vol. 373, 2021.
[41] J. Watson, D. Juergens, N. Bennett, B. Trippe, J. Yim, H. Eisenach, W. Ahern, A. Borst, R. Ragotte, L. Milles, et al., "De novo design of protein structure and function with RFdiffusion," Nature, vol. 620, 2023.
[42] A. Radford, R. Józefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," arXiv preprint, 2017.
[43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020.
[44] R. J. Hickman, M. Aldeghi, F. Hase, and A. Aspuru-Guzik, "Bayesian optimization with known experimental and design constraints for chemistry applications," Digital Discovery, vol. 1, 2022.