Artificial Intelligence Review (2025) 58:371
https://doi.org/10.1007/s10462-025-11365-w

Empowering scientific discovery with explainable small domain-specific and large language models

Hengjie Yu[1,2] · Yizhi Wang[1] · Tao Cheng[3] · Yan Yan[4,5] · Kenneth A. Dawson[4] · Sam F. Y. Li[6] · Yefeng Zheng[1,2] · Yaochu Jin[1,2]

Accepted: 14 August 2025

Abstract
As artificial intelligence (AI) increasingly integrates into scientific research, explainability has become a cornerstone for ensuring reliability and innovation in discovery processes. This review offers a forward-looking integration of explainable AI (XAI)-based research paradigms, encompassing small domain-specific models, large language models (LLMs), and agent-based large-small model collaboration. For domain-specific models, we introduce a knowledge-oriented taxonomy categorizing methods into knowledge-agnostic, knowledge-based, knowledge-infused, and knowledge-verified approaches, emphasizing the balance between domain knowledge and innovative insights. For LLMs, we examine three strategies for integrating domain knowledge—prompt engineering, retrieval-augmented generation, and supervised fine-tuning—along with advances in explainability, including local, global, and conversation-based explanations. We also envision future agent-based model collaborations within automated laboratories, stressing the need for context-aware explanations tailored to research goals. Additionally, we discuss the unique characteristics and limitations of both explainable small domain-specific models and LLMs in the realm of scientific discovery. Finally, we highlight methodological challenges, potential pitfalls, and the necessity of rigorous validation to ensure XAI's transformative role in accelerating scientific discovery and reshaping research paradigms.

Keywords AI for science · Explainable AI · Scientific discovery · Domain knowledge · Research paradigm

1 Introduction

Artificial intelligence (AI), including small domain-specific models and large language models (LLMs), has been more commonly adopted across scientific disciplines to collect (Polak and Morgan 2024), augment, and auto-label data (Roh et al. 2021), generate hypotheses and design experiments (Jain et al. 2023), analyze complex data and gain insights (Leist et al. 2022), and enhance predictive modeling and drive new discoveries (Chen et al. 2018). Scientists are excited about the capabilities of AI as it can accomplish many tasks that were

Extended author information available on the last page of the article

Published online: 08 October 2025

not previously feasible (Van Noorden and Perkel 2023). Besides, the unprecedented success of LLMs has revolutionized the approach to scientific problem-solving in many fields, such as life science (Shanker et al. 2024), medicine (Li et al. 2024), drug discovery (Wong et al. 2023; Lu et al. 2025), chemistry (Bran et al. 2024), and materials (Kang and Kim 2024). Collaborations across fields of expertise are increasing, which provides computational scientists with access to novel real-world scientific problems and enhances their research's impact; concurrently, experimentalists benefit from advanced computational analyses that accelerate their research progress and improve the technical correctness (Littmann et al. 2020). However, at the same time, there are widespread concerns regarding AI's potential to increase dependence on uncomprehended patterns, entrench biases, facilitate fraud, and contribute to irreproducible research (Van Noorden and Perkel 2023), especially for those high-stakes research decisions (Eshete 2021). Ignoring these concerns and possible failures in validity, reproducibility, explainability, and generalizability can undermine the credibility of AI in scientific research (Kapoor et al. 2024). Embracing these concerns and actively working to bridge them is essential for unlocking AI's transformative potential in scientific discovery (Yu and Jin 2025).

Recent advances in AI are transforming the scientific research paradigm (Fig. 1), shifting it from traditional hypothesis-driven methods toward explanation-driven approaches guided by AI models. Explainable AI (XAI) is able to make AI more human-understandable and expose its capabilities and limitations by providing explanations for model behaviors, which is essential for users to troubleshoot, manage, and trust AI tools (Gunning et al. 2019). Following the resurgence of AI, particularly deep learning (LeCun et al. 2015), XAI has received increasing attention across numerous fields since 2018. Previous surveys and reviews discussed the current research status of XAI from different perspectives, such as the definition, concepts and taxonomies (Murdoch et al. 2019; Barredo Arrieta et al. 2020; Minh et al. 2022; Górriz et al. 2023), explanation methods (Guidotti et al. 2019), evaluation methods (Vilone and Longo 2021), counterfactuals and causability (Chou et al. 2022), insights from social sciences (Miller 2019a), and stakeholders' desiderata (Langer et al. 2021). By leveraging XAI, researchers can not only enhance the explainability of complex models but also gain deeper insights into the underlying mechanisms driving the observed data. Some surveys and reviews, based on the above concepts, introduced the XAI into scientific community (Roscher et al. 2020) and specific fields, especially for the fields of materials and chemistry (Feng et al. 2020; Oviedo et al. 2022), medicine and healthcare (Tjoa and Guan 2021; Klauschen et al. 2024), and computational biology (Chen et al. 2024). Recently, LLMs have demonstrated remarkable potential in advancing scientific research across diverse domains (Boiko et al. 2023). By integrating agents based on LLMs with a broader set of automated tools in laboratories, researchers can develop autonomous systems capable of designing, optimizing, and executing scientific experiments. These systems leverage LLMs' capabilities in reasoning, problem-solving, and contextual understanding to enhance the efficiency and autonomy of the scientific process. However, the vast number of parameters in LLMs presents significant challenges for explainability, a critical factor in ensuring regulatory compliance and system credibility. Ongoing research seeks to address these challenges by developing techniques to make model outputs and underlying processes more understandable to researchers and stakeholders (Zhao et al. 2024), paving the way for wider adoption of these systems in scientific inquiry.

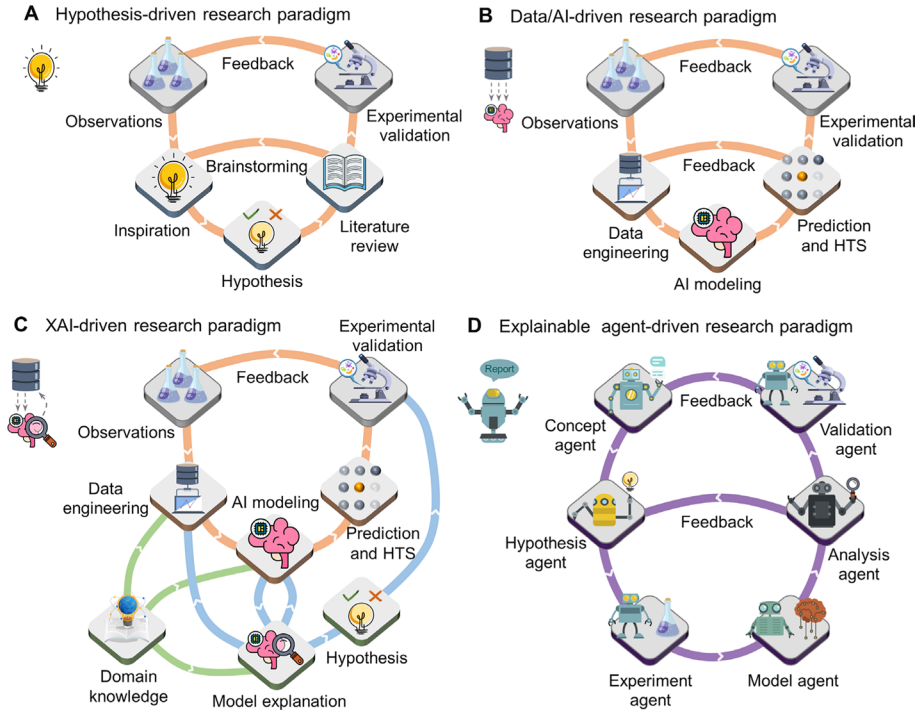


Fig. 1 The evolution of scientific research paradigms: from hypothesis-driven paradigm to XAI-driven and agent-driven paradigms. **A** Developing a sound hypothesis is essential in the classical research workflow, requiring researchers to draw upon their observations, expertise, and inspiration to propose a hypothesis. This hypothesis is then supported by a thorough review of existing literature, ensuring a greater chance of being confirmed. Formulating plausible hypotheses can be time-consuming, but it is essential for ensuring the validity of experimental design and advancing scientific understanding. **B** With the increasing volume of data, AI-driven scientific research significantly accelerates the field by accurately predicting complex systems and enabling high-throughput screening. AI serves as an excellent tool for handling nonlinearities and identifying crucial factors, thereby enhancing research efficiency and precision. **C** Integrating prior knowledge into data engineering and AI modeling enhances their reliability and addresses the limitation of having limited data in certain domains. XAI further promotes scientific research by aiding in model management and debugging, thus enhancing trust in the models. XAI reveals the underlying behavior of AI systems, which can lead to the formulation of new hypotheses. When prior knowledge is incorporated, the explanations provided by XAI align more closely with established prior knowledge, increasing the consistency and validity of the results. **D** The emerging agent-based research paradigm envisions multiple intelligent agents collaborating seamlessly to tackle complex tasks. These agents will integrate domain-specific models tailored for precise tasks with the broader capabilities of LLMs, enabling synergistic interactions with external resources and automated laboratory systems. Crucially, this paradigm must incorporate explainability to meet regulatory, ethical, and scientific demands, ensuring transparency and trustworthiness in both the process and outcomes of research

Instead of the question of whether we trust AI, now the question becomes whether we trust XAI's explanations. Researchers argue that we should stop explaining black-box models for high-stakes decisions because these explanations cannot capture the full behaviors of the model and low-fidelity explanations hurt our trust in model explanations and results (Rudin 2019). However, the alternative—self-explainable models—currently perform poorly when solving complex problems, making them less practical. Another question is whether the model explanation can be mapped into a form familiar to scientists. When XAI

tools provide explanations that are inconsistent with established scientific theories and terminologies, the integration of XAI tools into existing scientific workflows faces challenges. Experimental scientists prefer simpler and more explainable models if gaining scientific insights (Tao et al. 2021) or making trustworthy predictions is their primary interest. Despite the growing interest in XAI for scientific research, the integration of XAI-aided workflows remains relatively rare, highlighting a significant untapped potential for further exploration.

Etymologically, the term "science" comes from the classical Latin word *scientia*, which means knowledge (Oxford English Dictionary 2024). In other words, scientific research serves, to some extent, as a means for humanity to transform reality and transcend boundaries through the application of knowledge and the pursuit of new discoveries. The emergence of vast amounts of data and advancements in AI technologies have created significant opportunities for uncovering hidden patterns and discovering new knowledge through XAI (Krenn et al. 2022). On the other hand, prior knowledge can be a valuable tool for uncovering previously unknown insights (Cornelio et al. 2023) and for mitigating the limitations associated with insufficient training data (Von Rueden et al. 2021). It becomes easier for researchers to understand and trust these models if the model's explanations align with the prior knowledge. This alignment facilitates the seamless integration of XAI tools into existing scientific workflows, ensuring that XAI-driven insights are both reliable and actionable. What differentiates this review from existing works is its distinct perspective rooted in knowledge-driven scientific discovery. Our review emphasizes the role of prior knowledge in XAI, highlighting how knowledge integration enhances model explainability, guides scientific interpretation, and mitigates data limitations. This perspective not only aligns XAI with established scientific methodologies but also facilitates its seamless incorporation into scientific research.

Considering the pivotal role of knowledge in scientific discovery, this review adopts a forward-looking perspective encompassing established (small domain-specific models), emerging (LLMs), and future agent-based (collaboration between large and small models) research paradigms. For small domain-specific models, we present a knowledge-oriented taxonomy derived from recent progress, categorizing methods into knowledge-agnostic, knowledge-based, knowledge-infused, and knowledge-verified approaches. The interplay between leveraging prior knowledge and fostering innovation is explored as a key challenge warranting deeper investigation. For LLMs, we discuss three principal strategies for incorporating knowledge and three analytical approaches to model explainability. Looking ahead, we envision a paradigm shift towards AI scientist agents integrated with automated laboratories, highlighting the critical need for robust explanations and precise identification of explanation targets. Additionally, we highlight the pitfalls and methodological challenges in the thriving XAI-empowered scientific discovery.

2 Examples of explainable AI for scientific research

While the broader landscape of AI applications is vast, the deployment of XAI in real-world scientific research remains relatively nascent. However, a growing body of compelling studies demonstrates that XAI is not only being successfully applied but is also yielding significant and often transformative results across various scientific domains. The following four examples exemplify how XAI, by providing human-understandable insights into

AI's decision-making processes (Fig. 2), is empowering researchers to make more informed decisions, fostering trust, facilitating knowledge generation, and ensuring responsible application of AI.

In dermatology (Fig. 2A), an XAI system was developed to assist in melanoma and nevus diagnosis (Chanda et al. 2024). This system leverages domain knowledge by predicting specific, human-understandable characteristics based on dermatological expertise. The AI then infers a diagnosis from these features, providing clinicians with transparent explanations that align with their own diagnostic reasoning, thereby increasing trust and confidence. Besides, for single-cell data analysis (Fig. 2B), researchers created expiMap (Lotfollahi et al. 2023), a deep learning architecture that integrates domain knowledge from databases, articles, and expert insights to construct a binary matrix of "gene programs". This matrix is used to program a linear decoder within the model, allowing expiMap to learn and represent cell states through biologically understandable gene programs. Furthermore, efforts towards model explainability can be made not only during the model building process but also after it. In the area of antibiotic discovery (Fig. 2C), a graph neural network was initially developed and trained to predict antibiotic activity (Wong et al. 2024), leveraging the power of vast datasets. Critically, after model establishment, researchers shifted their focus to explainability. They applied a subgraph search algorithm, inspired by chemical domain knowledge, to identify the key substructures that influence a compound's activity. This post-prediction analysis provides substructure-based rationales, offering concrete chemical insights that directly guide the discovery of new and effective antibiotic classes. However, understanding and explaining LLMs can be particularly challenging due to their vast parameter counts, often ranging from billions to hundreds of billions. Fortunately, their inherent conversational nature offers a unique avenue toward achieving a certain degree of interpretability. For instance, in the field of disease diagnosis assistance (Fig. 2D), the MedFound LLM was developed for clinical diagnosis (Liu et al. 2025). Crucially, it was fine-tuned using diagnostic rationales and inferential processes derived from physicians' domain knowledge. This involved a self-bootstrapping strategy and a unified preference alignment framework, ensuring the LLM's outputs were not just accurate but also aligned with standard clinical practice and diagnostic hierarchy preferences, making its diagnostic suggestions more interpretable and trustworthy to medical professionals.

In the subsequent sections of this review, we delve into various methodologies for effectively incorporating knowledge into AI systems and subsequently enhancing their explainability, further solidifying XAI's seamless integration into scientific discovery.

3 Small domain-specific models: a knowledge-oriented taxonomy

Based on how and to what extent prior knowledge is incorporated into the domain-specific XAI pipeline, we classify XAI methods into four categories: knowledge-agnostic (data-driven methods that do not rely on domain knowledge), knowledge-based (methods that integrate knowledge into model construction independently), knowledge-infused (methods where model construction model construction is driven by both data and knowledge), and knowledge-verified (methods that integrate knowledge into model explanation rather than model construction).

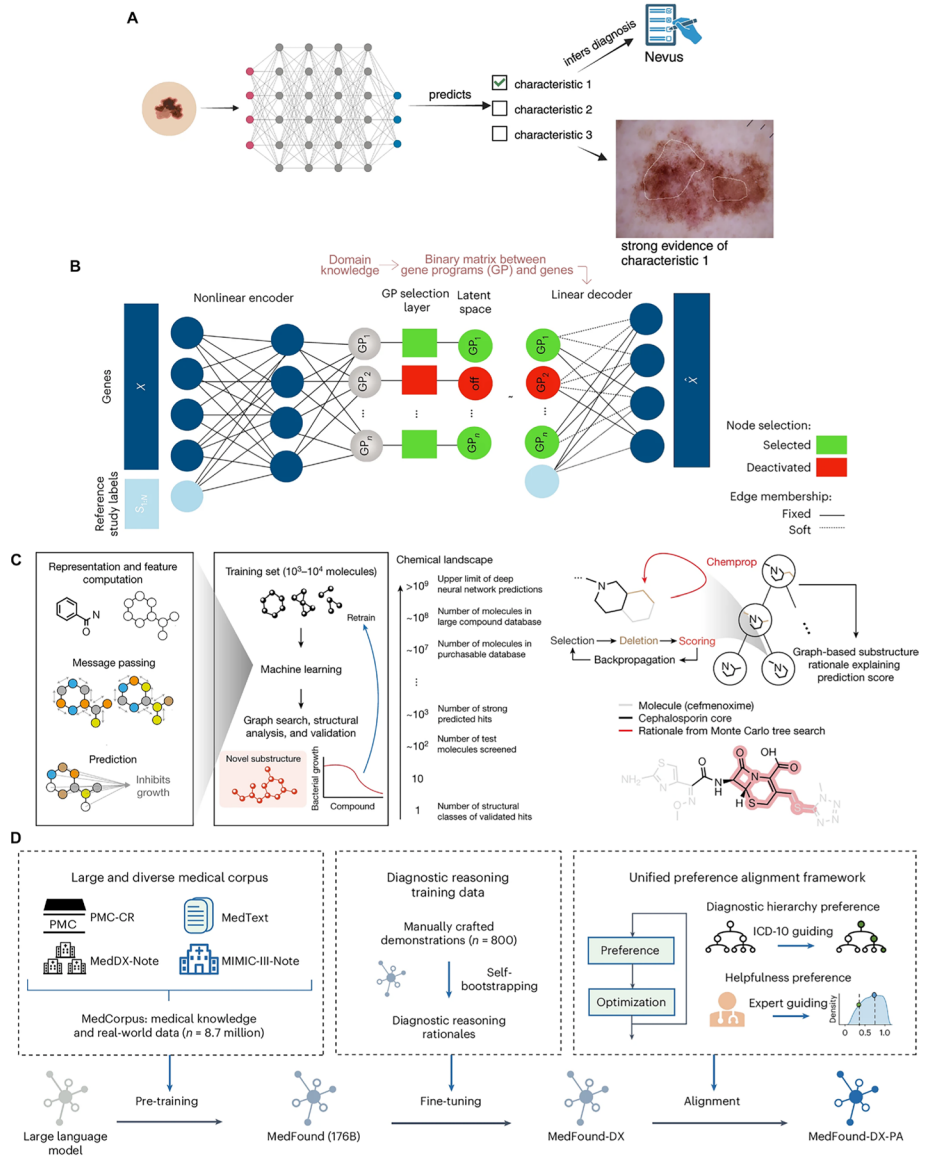
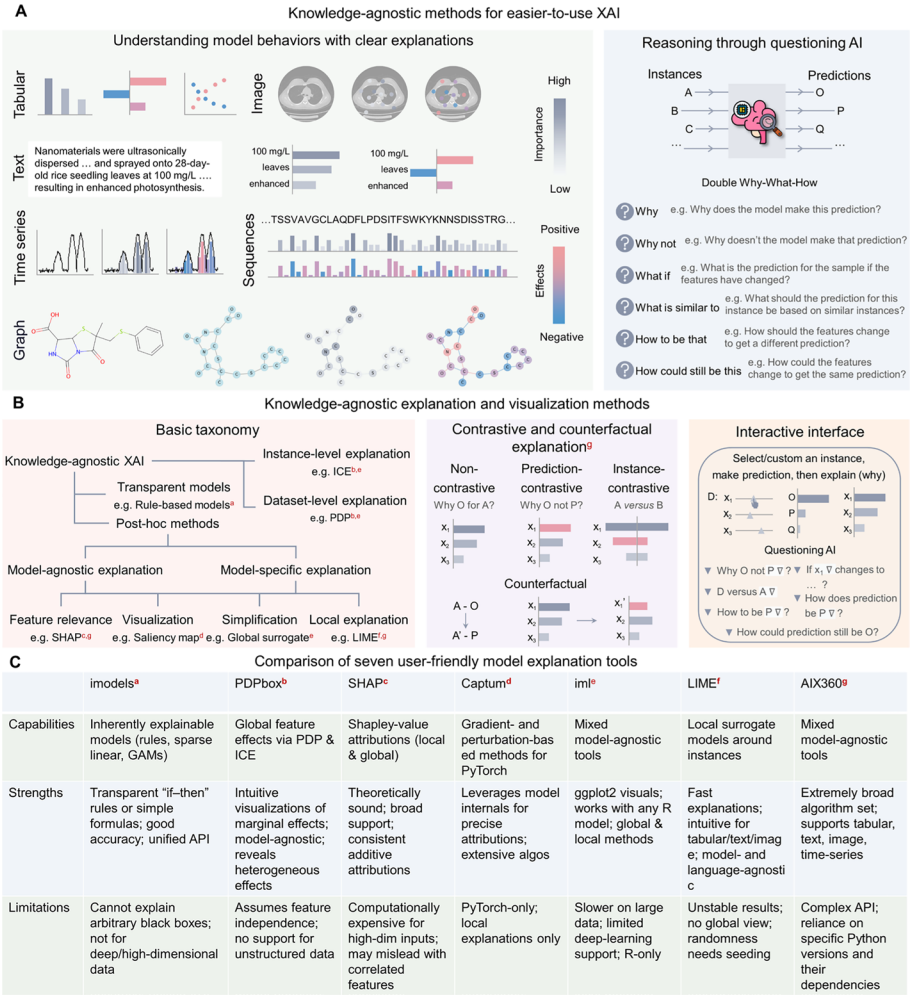


Fig. 2 Four examples of explainable AI for scientific research. **A** Prediction of domain-knowledge-derived features enables explainable medical diagnoses. Reproduced with permission from (Chanda et al. 2024).
Copyright 2024, The Author(s). **B** Domain knowledge guides the construction of a linear decoder for gene program analysis. Adapted with permission from (Lotfollahi et al. 2023). Copyright 2023, The Author(s). **C** Following prediction, a domain knowledge-inspired search method is employed to identify key substructures that influence antibiotic activity. Reproduced with permission from (Wong et al. 2024). Copyright 2023, The Author(s), under exclusive licence to Springer Nature Limited. **D** Diagnostic rationales and domain knowledge refine LLM output and align user preferences. Reproduced with permission from (Liu et al. 2025). Copyright 2025, The Author(s), under exclusive licence to Springer Nature America, Inc

The selection of appropriate categories for different tasks should take into account several factors, including task requirements, data availability, and the accessibility of domain-specific knowledge. Even within a single field, different tasks may require distinct approaches depending on the context and available resources. For example, in the domain of material property prediction, the selection of methods should be guided by data availability and research focus. When sufficient data is available, knowledge-agnostic methods or low-degree knowledge-infused methods may suffice. In scenarios where data is scarce, leveraging knowledge-based methods or high-degree knowledge-infused methods can compensate by incorporating domain expertise. For studies focusing on structure–property relationships, high-degree knowledge-infused methods ensure consistency with scientific theories. Conversely, for exploring unknown structure–property relationships, knowledge-agnostic methods or knowledge-verified methods are more suitable, allowing for unbiased exploration and validation against established knowledge.

3.1 Knowledge-agnostic methods

Knowledge-agnostic explanation methods play a crucial role in the field of AI-aided scientific research (Fig. 3), particularly when dealing with complex and opaque models like deep neural networks. These methods aim to elucidate model decision-making without requiring knowledge of its architecture, parameters, internal workings, or domain-specific expertise, making them highly versatile and applicable across various types of models and domains. Most unified model explanation methods are knowledge-agnostic methods, such as the widely used SHapley Additive exPlanations (SHAP) based on coalitional game theory (Lundberg and Lee 2017) and saliency maps derived from the gradient of predictions (Rebuffi et al. 2020). To provide a concrete example, SHAP offers a unified framework for interpreting machine learning model predictions by attributing contributions to individual features based on Shapley values from cooperative game theory. Shapley values ensure a fair allocation of feature importance by satisfying three key properties: local accuracy, missingness, and consistency. Local accuracy guarantees that the sum of SHAP values across all features, combined with a baseline prediction, equals the model's output for a given instance. Missingness ensures that features absent from the model's input receive zero attribution. Consistency mandates that if a model changes such that a feature's contribution increases, its SHAP value should not decrease. A key factor behind SHAP's widespread adoption in the field of scientific discovery is the availability of user-friendly packages and visualization tools provided by its developers. These resources play a crucial role in enhancing the accessibility and practical applicability of AI research outcomes within the scientific discovery domain. These methods have been used in various scientific research fields with different data types and tasks such as identifying disease-specific metabolite profiles (Buergel et al. 2022), screening risk biomarkers for subgroups of osteoarthritis (Nielsen et al. 2024), investigating single-cell gene regulatory network (Keyl et al. 2023), screening protein–protein binding sites (Hou et al. 2023), disentangling compounding effects in river flood risks under climate change (Jiang et al. 2024), understanding effects of anatomical and pathological markers on Alzheimer's disease (Qiu et al. 2022), achieving explainable diagnosis of ovarian cancer (Xiang et al. 2024), understanding particle-property relationships of nanozymes (Wei et al. 2022) and molecules (Harren et al. 2022), explaining permeability-selectivity trade-off in gas separation membranes based on polymers (Yang et al. 2022), understanding



C Comparison of seven user-friendly model explanation tools

	imodels[a]	PDPbox[b]	SHAP[c]	Captum[d]	iml[e]	LIME[f]	AIX360[g]
Capabilities	Inherently explainable models (rules, sparse linear, GAMs)	Global feature effects via PDP & ICE	Shapley-value attributions (local & global)	Gradient- and perturbation-based methods for PyTorch	Mixed model-agnostic tools	Local surrogate models around instances	Mixed model-agnostic tools
Strengths	Transparent "if–then" rules or simple formulas; good accuracy; unified API	Intuitive visualizations of marginal effects; model-agnostic; reveals heterogeneous effects	Theoretically sound; broad support; consistent additive attributions	Leverages model internals for precise attributions; extensive algos	ggplot2 visuals; works with any R model; global & local methods	Fast explanations; intuitive for tabular/text/image; model- and language-agnostic	Extremely broad algorithm set; supports tabular, text, image, time-series
Limitations	Cannot explain arbitrary black boxes; not for deep/high-dimensional data	Assumes feature independence; no support for unstructured data	Computationally expensive for high-dim inputs; may mislead with correlated features	PyTorch-only; local explanations only	Slower on large data; limited deep-learning support; R-only	Unstable results; no global view; randomness needs seeding	Complex API; reliance on specific Python versions and their dependencies

Fig. 3 Knowledge-agnostic methods for easier-to-use XAI in scientific research. **A** Knowledge-agnostic explanation methods are versatile and applicable across various types of data and models. With the help of visualization methods, researchers can obtain clear explanations of model behavior. **B** Current knowledge-agnostic methods can be categorized along several dimensions: transparent models (recommended for high-risk decisions) versus post hoc methods; model-agnostic explanation versus model-specific explanation; and instance-level explanation versus dataset-level explanation. Besides, contrastive explanations clarify why a model made a specific prediction (O) instead of another plausible outcome (P) from the given input (A). It is also possible to compare the contribution of two samples (A versus B) with respect to different features. Counterfactual explanations, a specific type of contrastive explanation, reveal the minimal changes to the input feature (A) that would cause the model's prediction to switch from the actual outcome (O) to a desired different outcome (P). Additionally, by asking the model questions (such as why, what, and how questions) through user-oriented interactive interfaces, this approach can assist researchers in obtaining a better understanding of relevant scientific contexts. **C** Comparison of seven user-friendly model explanation tools. The superscripts in **(B)** indicate the corresponding tools in **(C)** that support implementation of each method, including [a]imodels (Python: pip install imodels) (Singh et al. 2021), [b]PDPbox (Python: pip install pdpbox), [c]SHAP (Python: pip install shap) (Lundberg and Lee 2017), [d]Captum (Python: pip install captum) (Kokhlikyan et al. 2020), [e]iml (R: install.packages("iml")) (Molnar 2018), [f]LIME (Python: pip install lime) (Ribeiro et al. 2016), and [g]AIX360 (Python: pip install aix360) (Arya et al. 2019)

the benefit-risk trade-off of nanoagrochemicals (Yu et al. 2024), and extracting visual symptoms of plant stresses in machine vision-based phenotyping (Ghosal et al. 2018).

The knowledge-agnostic methods do not imply a lack of value from the knowledge perspective. On the contrary, these methods offer experimental scientists accessible model explanations using existing packages that demands minimal programming effort and limited domain expertise, as shown in Fig. 3C. Therefore, researchers in the fields of materials science and chemistry argue that there is no reason to keep the AI model a black box (Oviedo et al. 2022). Moreover, ensemble strategies are recommended in small-sample experimental studies, as they consider all explanation results rather than relying on a single split or model (Yu et al. 2023). There are many surveys and reviews (Feng et al. 2020; Oviedo et al. 2022; Chen et al. 2024) on knowledge-independent explainable AI, as we mentioned earlier. We provide a basic classification framework in Fig. 3 to facilitate the retrieval of existing methods by researchers. In the context of scientific research, XAI has predominantly been utilized to explain model behavior and extract useful information (evidence-based explanations) (Wu et al. 2023b). However, the potential of XAI extends beyond these applications.

Explanations are contrastive, selected, and socially interactive based on the insights obtained from social science (Miller 2019a). Contrastive and counterfactual explanations are more human-like explanations, receiving AI researchers' attention in recent years (Stepin et al. 2021). When people seek explanations for events or facts, they are often implicitly asking for comparisons to alternative scenarios ("Why did P happen instead of Q?" and "Why could instance A achieve the given target rather than instance B?") (Lipton 1990). Counterfactual explanation was considered a potential way to provide human causally understandable explanations (Chou et al. 2022) byidentifying a new instance that is both close to the original one in feature space and likely to occur in a real-world situation, but yields a different prediction (Mittelstadt et al. 2019). Counterfactual explanations are characterized by desirable properties such as validity (changing the classification outcome), minimality (minimal changes to the input), similarity (closeness to the original instance), plausibility (realistic and coherent with observed data), actionability (focus on mutable features), causality (respecting known causal relationships among features), discriminative power (highlighting reasons for the decision), and diversity (offering varied counterfactuals to enable multiple actionable insights) (Guidotti 2024). These properties collectively ensure that counterfactuals are interpretable, realistic, and practically useful for decision-making. Multiple counterfactual examples enhanced non-expert users' objective understanding and satisfaction scores compared with single example (Bove et al. 2023). Contrastive and counterfactual explanations can be especially valuable in scientific research, where hypotheses often involve exploring differences between experimental conditions, object properties, or theoretical scenarios, enhancing scientific reasoning and discovery.

People are skilled at selecting several causes from a vast array of potential factors to serve as explanations (Miller 2019a). Comparative explanations provide a structured method for focusing on several key factors that significantly influence outcomes rather than all. For example, in the context of bankruptcy prediction, a model-agnostic feature-weighted counterfactual generation method using multi-objective genetic algorithms has been proposed (Cho and Shin 2023). This approach leverages SHAP values to assign feature importance weights, ensuring that counterfactuals prioritize relevant and important features over irrelevant ones. Besides, user preference input can be leveraged to design selective explanation system to be more human-compatible (Lai et al. 2023). Furthermore, human–

computer interaction is a critical topic in human-centered XAI, as human-centric interactive user interfaces could empower user trust, knowledge transfer and model usability (Sokol and Flach 2020; Rong et al. 2024), which receives increasing attention from healthcare (Cálem et al. 2024) to broader fields. For example, a visual case-based reasoning method was proposed for breast cancer management, which could achieve automatic classification and visual explanations (Lamy et al. 2019). Visualization websites or softwares equipped with XAI have been developed to predict chemical toxicity (Togo et al. 2023; Lou et al. 2024). A chemically explainable platform, based on deep graph network, scoring module of atom-level features, and interactive web interface, was presented for the discovery of high-performance semiconductors (Gao et al. 2024a). Most existing work focuses on explaining how models achieve predictions, while asking questions to AI systems will further advance the field.

3.2 Knowledge-based methods

In scenarios with high-stakes decisions such as medicine and healthcare (Albahri et al. 2023), autonomous driving (Atakishiyev et al. 2024), robotics (Christov-Moore et al. 2023), financial services (Sachan et al. 2020), law enforcement (Hall et al. 2022), and policy decision making (Green and Chen 2021), the knowledge-agnostic methods are problematic due to confirmation bias (Ghassemi et al. 2021) and potentially unfaithful explanations (Rudin 2019). These explanations could be insufficiently accurate, confusing, or misleading (Chen et al. 2023) even if we assume that the model does not make the mistake of Clever Hans (i.e., making correct predictions based on wrong evidence). Besides, explanations are inherently context-sensitive, as they arise from specific questions that themselves are grounded in particular contexts (Beckh et al. 2023). However, the knowledge-agnostic method itself is usually not context-aware. Transparent and self-explainable models are therefore recommended for high-stakes decision making (Rudin 2019), but they currently face challenges in predictive accuracy when dealing with complex problems. The use of sophisticated black-box models is sometimes unavoidable and may be more prevalent than realized due to their superior predictive capabilities (London 2019). In these cases, full automation can often be undesirable, not only because of the critical nature of the outcomes but also because human experts can leverage their domain knowledge to complement the model's capabilities, ensuring the success of the task (Zhang et al. 2020). Furthermore, the knowledge foundation provides an effective way to adjust human trust in AI systems according to specific circumstances (discerning when to trust or question the AI). Considering the key roles of knowledge consistency and knowledge-based inspection, we classify individual knowledge-based components with human-understandable results within XAI pipelines as knowledge-based methods (Fig. 4), distinguishing them from knowledge-infused methods (discussed in the next section), which are driven by both data and knowledge.

Problem definition is crucial for effectively addressing scientific challenges. The process encompasses several key aspects, such as causal relationships within the field, the potential need for problem simplification, the identification of key factors, and the requirement for specific data. Among them, causal relationships are fundamental to ensuring the reproducibility of data-driven scientific research via machine learning (Li et al. 2020). It is crucial to recognize that accurate model predictions do not imply that the variables are the true causes of the observed outcomes (this may be due to correlations). Although there are some

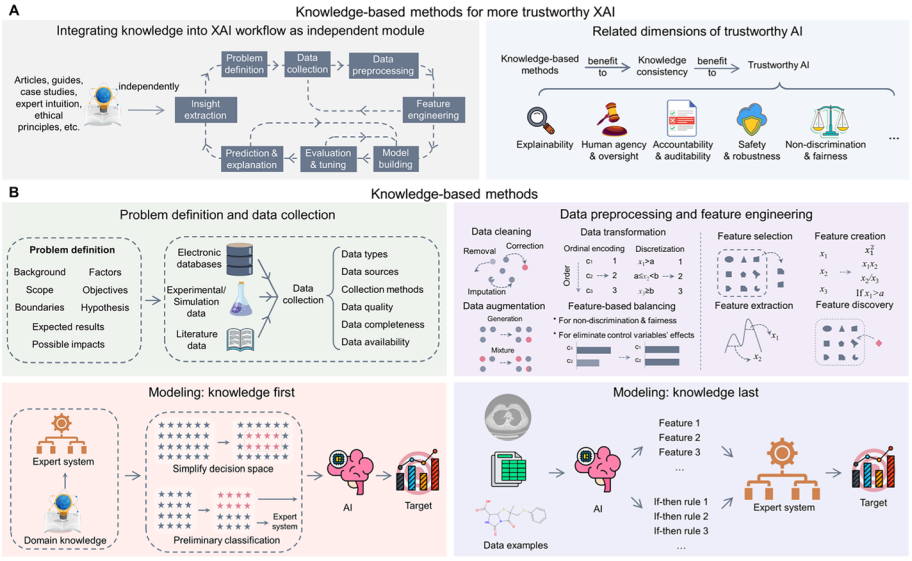


Fig. 4 Knowledge-based methods for more trustworthy XAI in scientific research. **A** The incorporation of domain knowledge can improve the trustworthiness of AI models across different dimensions, such as explainability, human oversight, accountability, and non-discrimination. This is because the integration of prior knowledge can help the model better achieve consistency with domain knowledge, and the independence of prior knowledge can also enable more effective human supervision. **B** Independent prior knowledge modules can be integrated into the steps of problem definition and data collection, data preprocessing and feature engineering, and the initial and final stages of modeling

efforts to integrate causal inference and machine learning (Cui and Athey 2022), such as the example in healthcare (Prosperi et al. 2020), the prevalence of observational $p(y|x)$ modeling underscores the importance of the causal understanding at the outset. Besides, data are the cornerstone of AI for science, and they determine the accuracy and reproducibility of the model from the source. Data preparation may take up 45% or even 80% of the time and effort of the entire project (Whang et al. 2023). With deep domain insights and clear problem definition, the need for extensive data collection can be minimized, and the quality of the collected data can be substantially improved. This alignment between data and problem definition enhances the validity of the research outcomes. For example, to increase the number of negative samples in a study on senolytic discovery using machine learning, the negative samples were selected based on the assumption that compounds from diverse chemical libraries lack senolytic activity (Smer-Barreto et al. 2023). The Materials Project of the Materials Genome Initiative collected computed structural, electronic, and energetic data for uncovering the properties of inorganic materials (Jain et al. 2013). The COVID Moonshot project established an open knowledge base with publicly available compound designs, crystallographic data, assay data, and synthesized molecules to accelerate the identification, synthesis, and test inhibitors against SARS-CoV-2 (Boby et al. 2023).

Data preprocessing and feature engineering are critical components in the development of robust machine learning models, and they stand out as areas where domain knowledge plays a pivotal role, especially for fields with small data (Murdock et al. 2020). Domain knowledge is indispensable during preprocessing, especially when it involves data cleaning,

transformation, augmentation, and balancing. For example, domain knowledge has been used in the material classification based on bandgap energy ranges (Kailkhura et al. 2019), filtering unreasonable samples with positive photovoltaic power at midnight (Luo et al. 2021), data augmentation of XRD powder patterns (Oviedo et al. 2019), and preparation of training data in case of class imbalance (Hirsch et al. 2023). Feature engineering, the process of selecting, modifying, creating or discovering new features, is another area where domain expertise is crucial. Well-engineered features can significantly enhance model performance, sometimes even more so than sophisticated algorithms. For example, structured features were transformed from raw extinction spectra based on plasmonic knowledge, improving the performance of nanoparticle size and distribution prediction (Tan et al. 2022). Fluorescence changes from fluorescamine labeling on a protein were identified as novel features for the prediction of protein corona (Duan et al. 2020). Dimensionality reduction methods (such as principal component analysis) can be used to reduce the dimensionality and improve model performance, but these new features are more difficult to understand than the original features and need to be used with caution in tasks where interpretability is a concern.

The hybrid model, combining individual knowledge-based and data-driven modules, offers ease of use for scientists with limited AI expertise while achieving a balanced performance. These hybrid models can be divided into two main types, named knowledg-first and knowledge-last methods, according to the order of modules. Knowledge-first methods were usually conducted for problem simplification. For instance, using the valence electron concentration criterion (domain knowledge) to narrow down the unexplored space in an active learning-assisted design of a high entropy alloy, six iterative loops led to the synthesis of an alloy with an ultimate strength of 1258 MPa and an elongation of 17.3% (Li et al. 2022a). In a binary classification task with uncertainty, fuzzy logic was first used to classify it into three categories (positive, negative, or boundary), and deep learning was then used to classify the boundary samples (Subhashini et al. 2022). Besides, knowledge-last methods were primarily intended to enhance human supervision. For instance, compared to direct end-to-end melanoma diagnosis, the method of first predicting features and then making a diagnosis based on domain knowledge significantly increased dermatologists' trust in XAI (Chanda et al. 2024).

3.3 Knowledge-infused methods

AI models relying solely on data-driven approaches often lack reasoning capabilities and may generate predictions inconsistent with fundamental laws or amplify pre-existing biases (Ntoutsi et al. 2020). Integrating knowledge into these models introduces appropriate inductive biases, resulting in scientifically valid predictions, simplified model architectures, enhanced generalization on unseen data, and improved explainability (Yu and Wang 2024). There is growing interest in integrating more knowledge into AI models, not only as an individual module, resulting in approaches with various names, such as knowledge-informed AI (Von Rueden et al. 2021) and physics-informed AI (Karniadakis et al. 2021a). This category examines the infusion of domain-knowledge by means of the hypothesis set and learning algorithm (Fig. 5). The hypothesis set refers to the collection of all possible models or functions that a learning algorithm can explore to generate predictions. This set typically includes various model architectures, such as neural networks, decision trees, or support vector machines, and the corresponding learnable parameters, such as weights, biases, or

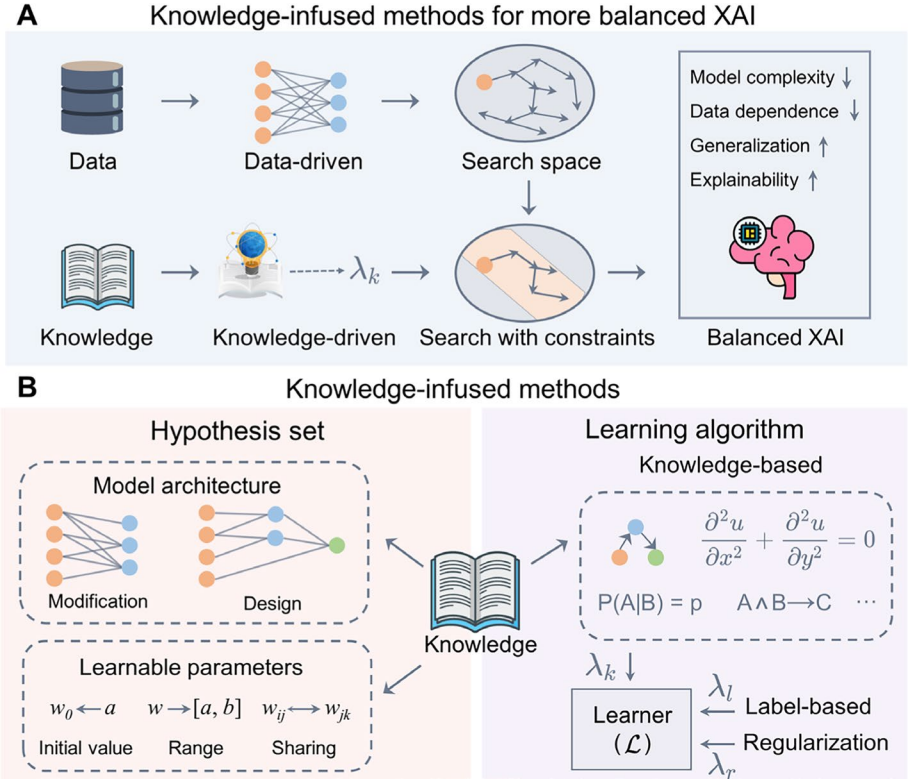


Fig. 5 Knowledge-infused methods for more balanced XAI in scientific research. **A** By infusing knowledge, some constraints are added to the model, effectively narrowing the search space. This helps reduce model complexity, thereby lowering the amount of data required for training while enhancing generalization capabilities. Additionally, knowledge infusion contributes to improving the model's explainability, as it aligns predictions with domain-relevant reasoning. **B** Knowledge can be embedded directly into both the hypothesis set and the learning algorithm. Incorporating knowledge into the hypothesis set helps guide the model toward more plausible solutions, while embedding it in the learning algorithm helps optimize the learning process itself, ensuring that the model prioritizes scientifically valid outcomes

kernel functions. Domain-specific knowledge can be integrated into the hypothesis set by constraining the model architectures or parameter ranges to reflect known principles from the domain. The learning algorithm is responsible for searching through this hypothesis set to find the model that best explains the training data. Domain-specific knowledge can be integrated into the learning algorithm through knowledge-based regularization, parameter priors, or constraint-driven optimization.

To demonstrate how knowledge-infused models can be applied to scientific research, we can examine specific examples of model structures, learning parameters, and loss functions where domain knowledge plays a critical role. Explainable programmable mapper (ExpiMap) was a deep-learning model for single-cell reference that integrated biological knowledge in the forms of gene programs into its structure (Lotfollahi et al. 2023), making data analysis more interpretable and aligned with known biological processes. In scenarios with fewer samples, incorporating prior knowledge made learning models more efficient compared to non-biologically informed models, while expiMap could also handle complex

nonlinear models when more training samples were available. To solve the issue where a model's nodes, edges, and weights lack correspondence with meaningful biological concepts, knowledge-primed neural networks (KPNNs) were designed to allocate interpretable weights across multiple hidden layers (Fortelny and Bock 2020). This enables the model to identify and prioritize key regulatory proteins, facilitating their use in experimental validation and enhancing the biological interpretability of the model's predictions.

Besides, physics-informed neural networks (PINNs) have gained significant attention due to their ability to seamlessly incorporate prior knowledge from physics into the learning process, leading to enhanced accuracy in predictions, reduced data requirements, and the capability to solve complex partial differential equations (PDEs) efficiently. For instance, the incompressibility condition, and Dirichlet and Neumann boundary conditions were formulated into loss terms of the mean squared residuals of the PDE to analyze internal structures and defects in materials (Zhang et al. 2022). The constraints of elastic coupling relation, input features wave speed, and wave amplitude were added to the loss function to predict lab earthquakes (Borate et al. 2023). A physics-informed gated recurrent graph attention unit network integrates prior knowledge as graph regularization to model variable dependencies into a directed graph, ensuring adherence to underlying physical laws for improved unsupervised anomaly detection in industrial cyber-physical systems (Wu et al. 2023a). Besides, the knowledge incorporated into the loss function is not limited to physical knowledge. For instance, by using the knowledge of repetitive streaking and star-shaped patterns caused by metal artifacts in CT scans, an explainable convolutional dictionary network was built with simple and physically meaningful operations (Wang et al. 2022a). This approach not only encodes prior knowledge into a learnable framework but also ensures fine explainability by mapping each network module to specific physical operators, enabling clear analysis of the model's mechanisms and enhancing both performance and understanding in metal artifact reduction. Moreover, the knowledge of mass balance, the range of yield, and key factor responses were infused into loss function to improve carbon cycle quantification in agroecosystems (Liu et al. 2024a). This method was developed by designing a machine learning architecture based on causal relationships from an agricultural process-based model, pre-training it with synthetic data, and fine-tuning it with observed crop yield and carbon flux data. Knowledge-guided loss functions were integrated to constrain variable responses during training, enhancing the accuracy of carbon cycle predictions in agroecosystems.

3.4 Knowledge-verified methods

In the scenarios where large datasets are available, such as drug discovery (Zhao et al. 2020), protein-protein interaction prediction (Bryant et al. 2022), and weather forecasting (Lam et al. 2023), a data-driven modeling approach is particularly advantageous due to the abundance of available data. Relying on purely data-driven modeling can prevent constraints imposed by domain knowledge from limiting the model's capacity to uncover novel insights and discoveries. Researchers can fully harness the potential of large datasets to explore uncharted territories. In this framework, domain knowledge is strategically employed solely in the explanation of the model's outcomes, enhancing interpretability without restricting the model's exploratory capabilities. Therefore, knowledge-verified methods were proposed in these cases (Fig. 6), such as prediction-based methods (finding fundamental rationales that match domain knowledge) and concept-based methods (discovering and validat-

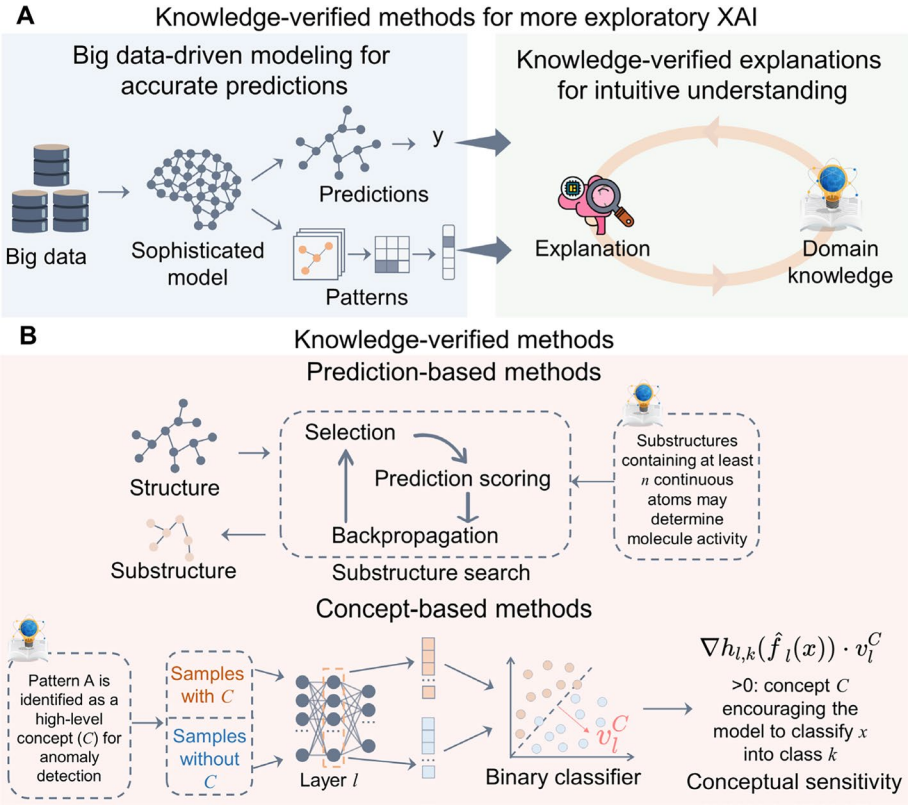


Fig. 6 Knowledge-verified methods for more exploratory XAI in scientific research. **A** In scenarios where abundant data are available, a purely data-driven approach can fully leverage the data's potential to discover new insights without the constraints of existing knowledge. Once the model is developed, knowledge infusion can be used in the model's explanation to enhance understanding of its predictive behavior and uncover meaningful patterns. **B** Two primary methods of knowledge infusion are prediction-based and concept-based approaches. Prediction-based methods focus on identifying the influence of fundamental rationales—such as structures or patterns—on the model's predictions using domain knowledge. Meanwhile, concept-based methods rely on domain-specific concepts, or automatically learned high-level concepts, to assess their impact on model outcomes. These explanation methods ensure that knowledge is applied post-hoc, preserving the model's ability to explore new knowledge while improving the interpretability of its predictions

ing understandable patterns), which not only foster innovation but also allow for a more nuanced understanding of the phenomena involved. Compared to knowledge-agnostic post hoc explanation methods, knowledge-verified methods offer explanations that are more aligned with the intuition of domain experts, focusing on meaningful structures and patterns rather than isolated pixels or individual atoms.

Currently, knowledge-verified methods are less commonly used compared to other explanation techniques, but they are gaining increasing attention from researchers. These methods validate model predictions by comparing them against established domain knowledge, providing an additional layer of reliability and understandability. For example, leveraging domain knowledge—such as the classification of antibiotics based on shared substructures—an explainable deep learning method was employed to identify substructure-based

rationales for compounds with high predicted antibiotic activity and low predicted cytotoxicity (Wong et al. 2024). The substructure identification was guided by three key substructure properties: the maximum size being below a defined atom limit, the substructure being a connected subgraph, and the subgraph's predicted values exceeding an activity threshold. Concept-based methods begin with concepts identified by domain experts. For example, six concepts, i.e., high-grade carcinoma, low-grade carcinoma, invasive lobular carcinoma, ductal carcinoma in-situ, tumor-adjacent desmoplastic stromal changes, and tumor infiltrating lymphocytes, were determined by experienced breast subspecialist pathologists for evaluating the association of specific histomorphological features with the biomarker predictions (Gamble et al. 2021). Additionally, concepts can be automatically discovered (Ghorbani et al. 2019) and evaluated (FEL et al. 2023), reducing the need for manual supervision in labeling them. However, before these concepts are applied as explanations, they can be verified and screened using expert knowledge or experimental data to ensure the reliability and relevance of the automatically identified concepts in scientific contexts.

3.5 Trade-off between leveraging prior knowledge and exploring scientific novelty

There is a trade-off between leveraging prior knowledge and exploring scientific novelty (Fig. 7). As mentioned earlier, knowledge-based methods are favored in high-stakes fields due to their transparency and reliability (Batra et al. 2021; Li et al. 2022b). The integration of knowledge reduces reliance on data volume, particularly beneficial when data is scarce or of poor quality. It would accelerate the model's convergence and provide the model explanation with a better alignment with human domain expertise (Singhal et al. 2023). Regardless of the manifestation of knowledge incorporation, the essential substance, however, is to restrict the searching space of models (Willard et al. 2022). Thereby, the valuable rules or implicit patterns that are not yet fully grasped within the domain underlying the data would be neglected, with over-reliance on pre-existing knowledge. Such an overfitting issue can even be exacerbated, leading the model to produce biased predictions if the injected knowledge is overly partial or subjective. On the other hand, AI models have demonstrated a remarkable capability to handle vast amounts of data and uncover complex patterns within

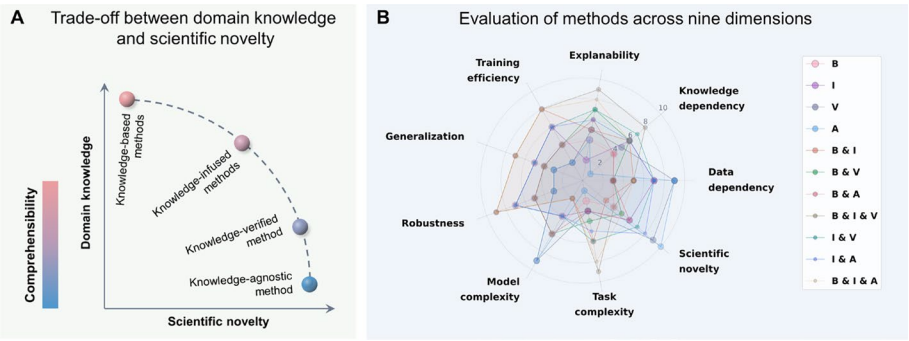


Fig. 7 The trade-off between prior knowledge and novelty, and the positioning of methods on varying dimensions. **A** The trade-off between domain-specific knowledge and scientific novelty in XAI. **B** Nine dimensions representing key considerations for the evaluation and comparison of different methods or method combinations. The methods are abbreviated as knowledge-based (B), knowledge-injected (I), knowledge-verified (V), and knowledge-agnostic (A)

data (Topol 2019). Incorporating explainable methods allows models to generate insights that form the foundation for new scientific discoveries. However, it is noteworthy that the AI's superior performance in some cases may rely on spurious correlations rather than the true causal relationship, driven by the pursuit of the training goal of accuracy, which results in insights provided by the model of less utility (Lapuschkin et al. 2019; Messeri and Crockett 2024). Furthermore, large datasets are frequently unavailable in many domains at their current stage of development (Karniadakis et al. 2021).

Generally speaking, the comprehensibility of the method is proportional to its dependency on prior knowledge. However, such dependency would reduce the model's discovery capabilities beyond the boundary of that knowledge, limiting the potential for new scientific discoveries. Knowledge dependency is highest in knowledge-based methods, where the expertise module serves as the foundation of the entire pipeline. In contrast, the knowledge-infused methods primarily incorporate knowledge during the training process, imposing fewer constraints and allowing the model a broader exploration space. Compared to knowledge-verified methods, knowledge-agnostic approaches tend to focus solely on explaining black-box model behavior within the AI context, and the explanations lacking domain consensus would be innovative but challenging for experimentalists to understand or directly verify, resulting in a gap between the extracted rules or explanations and the expertise of experts. Thereby, as the novelty of a method increases, its comprehensibility tends to decrease, highlighting the trade-off between knowledge integration and scientific innovation.

Thus, we argue that finding the optimal balance between data-driven approaches and knowledge integration requires consideration of the application context and underlying requirements. In our taxonomy, we assess methods and their combinations against key dimensions most considered in the design of AI systems, scoring from 0 to 10, allowing researchers with different backgrounds to select suitable approaches based on their goals. For instance, in domains where the data is abundant but the theoretical understanding is limited, the knowledge-agnostic methods can be an optimal choice, which offers a sufficient level of scientific novelty but produces explanations that are hard to validate and less understandable. Additionally, such methods always require meticulous model architectural design, parameter tuning, and longer convergence times for the model to effectively learn from large amounts of data. On the contrary, in fields where scientific principles are well understood, knowledge-related methods might be prioritized to maintain the integrity and interpretability of the model. This is particularly critical in the development of application-oriented AI, where the primary goal should be to minimize risk and provide reliability (Dong et al. 2023). In these scenarios, models that integrate knowledge throughout the entire process—from data processing to training—and are equipped with knowledge-consistent explanations should be prioritized. These come, however, with greater dependency on domain knowledge limiting the model's ability to explore unknown patterns and increased task complexity. Additionally, factors such as generalization and robustness should also be considered when selecting a suitable method.

Combining human expertise with AI-driven insights remains challenging. Therefore, we recommend verifying and comparing the impact of incorporating specific domain knowledge on model performance and explainability. This requires assessing whether to prioritize knowledge integration or rely more heavily on data based on expert judgment and task requirements. For instance, if domain knowledge suggests a causal relationship that

contradicts empirical data, a critical evaluation is necessary to determine the most appropriate approach to knowledge incorporation. If the domain knowledge is grounded in well-established principles, it may take precedence even if the data does not fully support it. Conversely, if strong data-driven evidence challenges existing knowledge, it may be necessary to reassess and refine the domain understanding. The decision should be guided by task objectives: if the goal is to improve generalization, rigorous validation methods such as cross-validation should be used to ensure that knowledge integration enhances performance without leading to overfitting; if the focus is on model explainability, incorporating domain knowledge may yield smoother and more intuitive factor effects, even if it slightly compromises predictive accuracy.

4 LLMs: knowledge injection and explainability

LLMs have attracted significant attention in scientific research, providing unprecedented capabilities in natural language understanding, generation, reasoning, and interaction with the surrounding environment (Caldas Ramos et al. 2024). LLMs can achieve performance improvements on specific tasks through domain knowledge injection. This approach is particularly valuable for both prediction and generation tasks in domains where sufficient data is lacking. By leveraging the generalization capabilities of LLMs, domain knowledge can be effectively integrated to enhance their predictive accuracy and enable the generation of new, contextually relevant samples. Such knowledge infusion allows the model to compensate for data scarcity, providing more accurate predictions and facilitating the creation of novel content even in low-data scenarios. Moreover, improving the explainability is important for increasing trust in LLMs and future LLM-based scientist agents.

4.1 Knowledge injection into LLMs

The processes of infusing domain knowledge into LLMs typically involve three key approaches (Fig. 8): prompt engineering, retrieval-augmented generation (RAG), and supervised fine-tuning (SFT).

The performance of LLMs is highly sensitive to the quality and design of the prompts provided (Zhou et al. 2022). Prompts can be considered a form of inference code expressed in natural language, guiding the model's reasoning and output. This unique characteristic highlights the dual role of prompts as both an interface for human-model interaction and a mechanism for task-specific optimization, underscoring their critical importance in leveraging the full potential of LLMs. Prompt engineering enables the use of LLMs to generate high-quality textual data for sparse materials science problems, significantly improving classification accuracy (Liu et al. 2024b). By integrating optimized prompts with fine-tuned deep learning models, this approach improves accuracy in material feature-label classification tasks by up to 48% compared to traditional machine learning models. Prompt-based embeddings facilitate advanced molecular optimization by enhancing the model's ability to adjust specific properties, even with limited multi-property data, leading to significant improvements in optimization success rates compared to traditional models (Wu et al. 2024). Few-shot learning enhances the performance of protein language models in predicting protein fitness under extreme data scarcity, requiring only minimal labeled data for

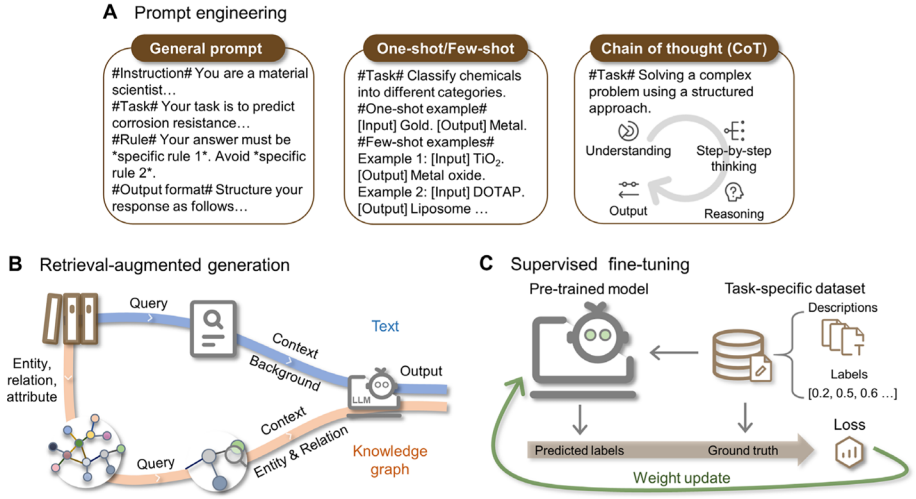


Fig. 8 Three main approaches to infusing domain knowledge into LLMs. **A** Prompt engineering involves tailoring input prompts to guide LLMs with specific domain knowledge, while **B** RAG enhances LLM outputs by integrating external information sources, and **C** SFT improves performance and relevance by fine-tuning LLMs on domain-specific data

optimization, with a 25% improvement in performance (Zhou et al. 2024). Chain-of-thought reasoning, when integrated with tool-augmented frameworks, transforms LLMs from confident information sources into iterative problem-solving engines, enabling precise and dynamic applications in complex domains such as chemistry (Bran et al. 2024). Manually designing prompts requires extensive experience and significant effort. Prompt engineering is evolving towards automation, with several emerging technologies, such as Automatic Prompt Engineering (Google) (Zhou et al. 2022), Active Prompting (Diao et al. 2023), and PromptWizard (Microsoft) (Agarwal et al. 2024), showing great potential. Although still in their early stages, these technologies have not yet been fully leveraged to address specific scientific challenges. However, they hold great potential to revolutionize the field by streamlining and enhancing the prompt design process.

The extent to which next-token prediction, the foundational mechanism of LLMs, can genuinely emulate human intelligence remains a subject of debate (Bachmann and Nagarajan 2024). This approach inherently risks introducing factual inaccuracies, posing challenges to achieving the precision and reliability demanded by knowledge-intensive tasks (Lewis et al. 2020). RAG enhances LLMs by incorporating external data, providing domain-specific context to improve accuracy and reliability beyond pre-training knowledge. For example, RAG enhances BioinspiredLLM by integrating external knowledge sources, enabling precise numerical recall, accurate responses to complex queries, and traceability of information, thus advancing its utility in bio-inspired materials research (Luu and Buehler 2024). RAG and well-designed prompts can achieve numerous powerful functions, such as Nova (Hu et al. 2024), which integrates RAG, prompt engineering, and iterative planning and search methods to enhance the novelty and diversity of scientific ideas generated by LLMs. This approach effectively avoids the issue of simplistic and repetitive suggestions, which can result from limited external innovation knowledge. Graph-based RAG extends traditional text-based RAG by incorporating structured knowledge graphs and hierarchical community

summaries, making it particularly valuable in knowledge-intensive domains where deeper contextual understanding and relationships enhance retrieval and generation accuracy. Graph-based RAG is currently underutilized in addressing specific scientific challenges. However, recent advancements, such as the GraphRAG framework (Microsoft) (Edge et al. 2024) and LightRAG (Guo et al. 2024), are making this approach more accessible to researchers, thereby facilitating broader adoption in scientific problem-solving.

SFT of LLMs tailors their general capabilities to specific scientific domains, enabling enhanced contextual understanding and task-specific performance. SFT has been applied in chemical text mining tasks, enabling automated extraction and transformation tasks such as compound entity recognition and reaction role labeling with high accuracy (Zhang et al. 2024). Besides, fine-tuning GPT-3 enabled it to effectively tackle a range of chemical and materials science tasks, outperforming conventional machine learning models, particularly in low-data scenarios, and offering a powerful tool for both predictive tasks and inverse design (Jablonka et al. 2024). As model parameters continue to scale up, the expenses associated with SFT and the storage of all these parameters escalate significantly, ultimately rendering such processes impractical. Delta-tuning is recognized as a parameter-efficient adaptation strategy that focuses on modifying only a small subset of model parameters during training, optimizing resource use while maintaining performance (Ding et al. 2023).

4.2 Explainability of LLMs

In comparison to smaller, domain-specific models, LLMs possess significantly more parameters, allowing them to generalize across a wide range of tasks and languages. However, this richness in representation comes at the cost of explainability. Existing work has focused on constructing smaller, domain-specific models that are easier to interpret and can help provide explanations for LLMs (Frank et al. 2024). To address this challenge, many approaches have been proposed to uncover the inner working mechanisms of LLMs. In this context, we taxonomize the explanation methods of LLMs that are more likely to be used in scientific discoveries into three main types (Fig. 9): local explanation, global explanation, and conversation-based explanation.

Local explanations involve methods that offer insights into how a model makes predictions for a given input. One common approach is feature attribution, which can be achieved through linear approximations (Yang et al. 2023) or by calculating integrated gradients (Enguehard 2023) to assess the contribution of individual tokens to the model's prediction. Example-based methods explore how the model's behavior changes with the modified input, using techniques like masking, perturbing inputs (Barkan et al. 2024) or the generation of counterfactual samples (Chen et al. 2021). Another important approach is the attention map, a key component of transformer-based architectures, which reveals how the model distributes its focus across different tokens during the forward pass. Visualizing the attention weights assigned to tokens could help with identifying most relevant parts of input for the model's decision-making process (Barkan et al. 2021). Attention maps have also been utilized to investigate how LLMs make predictions in protein phase transition tasks (Frank et al. 2024). In the Lamole approach, the authors integrate gradient-based explanations with attention mechanisms to capture how functional groups and their interactions contribute to molecular property predictions (Wang et al. 2024b).

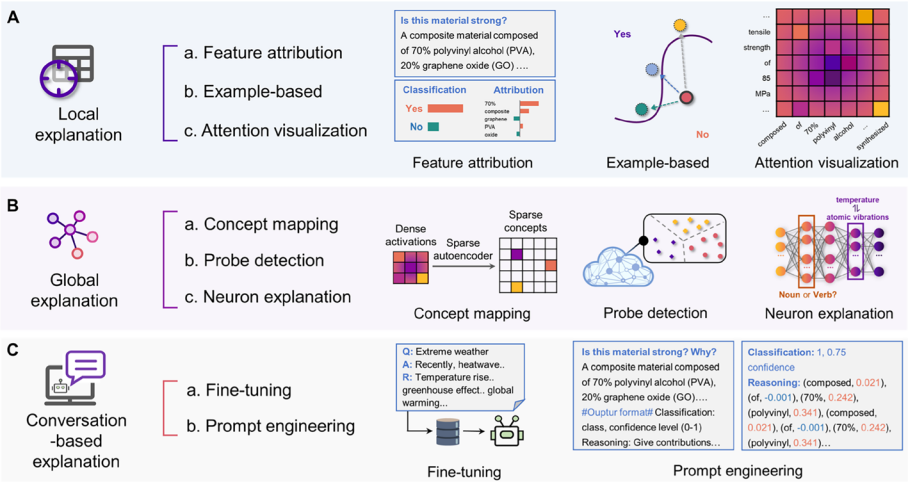


Fig. 9 Explanation approaches for LLMs. **A** Local explanation focuses on specific input–output relationships to reveal how individual predictions are made. **B** Global explanation provides an overarching view of the model's behavior by analyzing general patterns and decision rules. **C** Conversation-based explanation clarifies the model's reasoning and outputs through interactive dialogue with users

Despite the simplicity and intuitiveness of these methods, the relevance of attention weights to the final prediction in LLMs still requires further validation (Bibal et al. 2022). Moreover, feature attribution methods often overlook the complex interactions underlying tokens, which can be particularly evident in scientific research. Additionally, local explanation methods can offer only a fragmented or partial understanding of the model's behavior rather than capturing the full scope of the decision-making process. As a result, these methods limit their usefulness in generating scientifically robust insights.

Global explanations provide an explanation of the overall mechanism encoded in the neurons, hidden layers or modules of a model by consolidating all possible predictions. The concept mapping aims at extracting neural representations into human-understandable concepts. K-sparse autoencoders are introduced to effectively control sparsity and enhance feature extraction from GPT-4 activations, achieving a 16 million latent model trained on 40 billion tokens with improved reconstruction and sparsity (Gao et al. 2024b). A concept bottleneck approach has been proposed for protein generation using LLMs. By analyzing the weights of the linear decoder, researchers have been able to uncover the complex relationships between abstract concepts and amino acids (Ismail et al. 2024). In probe detection approaches, an auxiliary model, referred to as a probe, is trained to classify the representations and model parameters into specific properties (Peng et al. 2022). Neuron explanation methods focus on interpreting the contribution of individual neuron or groups of neurons, as well as their interactions, to the overall language processing of the model. For example, the individual neuron could be activated by different patterns in the input, representing different meanings encoded by that neuron (Mukherjee et al. 2023). One notable example is the identification of a specific circuit of neurons within an LLM responsible for performing the indirect object identification task (Wang et al. 2022b). Existing studies have shown that neurons in shallow layers are more focused on word-level syntax, such as determining the part of speech for each token, while deeper neurons capture higher-level semantic knowledge,

allowing for a more comprehensive understanding of sentence meaning, such as relationships between entities within the sentence (Jawahar et al. 2019).

It is computationally expensive to explain the billions of neurons in LLMs; therefore, a commonly adopted method is to rank neurons and retain only the most important ones. However, this approach may overlook the complex interactions between features or entities, particularly in domains such as molecular structure prediction and gene sequence analysis. Furthermore, current global explanation methods primarily focus on general LLMs, providing linguistic and semantic insights. For non-linguistic tasks, however, translating these interpretations into domain-specific knowledge remains a significant challenge, representing a key difficulty in scientific research.

Fine-tuning and prompt-based approaches involve designing or guiding the model to generate explanations as part of the prediction process, without requiring external interpretability tools. This can be achieved by finetuning LLMs with training data that contains extra causal relationships or contextual explanations between inputs and outputs (Yang et al. 2024). Furthermore, LLMs can be instructed by users to generate reasonings as contextual explanations for their output, making prompt design a powerful tool for guiding the model to produce self-explanations during the prediction process (Huang et al. 2023). Compared to other methods, prompt-based explanations offer particular advantages for users, allowing them to adjust the type and level of detail of the explanation provided based on their specific requirements, without any additional tools or modifications. Moreover, in early 2025, reasoning LLMs began to attract significant attention, with models like DeepSeek-R1 (DeepSeek-AI et al. 2025) leading the way into the open-source stage. At the same time, the reasoning process inherent in the models provides valuable insights into understanding their behavior. DeepSeek-R1-Zero pioneers pure reinforcement learning (RL) training without SFT, leveraging the GRPO algorithm to autonomously develop reasoning capabilities like error self-correction, demonstrating emergent "Aha Moments" during training. In contrast, DeepSeek-R1 adopts a four-stage hybrid framework: starting with SFT cold-start training for output readability, followed by reasoning-oriented RL with automated reward mechanisms (accuracy, formatting, and language consistency), rejection sampling to generate high-quality trajectories, and RL from human feedback for human alignment. Reasoning models hold tremendous potential for applications in scientific discovery, especially when assisted by RAG that incorporate domain-specific knowledge. However, for tasks that require rapid responses, the speed of reasoning may present certain limitations.

However, concerns arise regarding the potential for LLMs to generate "hallucinated" content—outputs that may seem plausible but are factually incorrect (Augenstein et al. 2024). Additionally, even when the model's output is reliable, the self-explanations it generates may not align with its actual reasoning process, raising issues about the LLM's faithfulness (Madsen et al. 2024). Fortunately, the above-mentioned method of injecting knowledge into the model will help alleviate these problems. The integration of domain knowledge injection with interpretability analysis of large models offers transformative opportunities for scientific discovery, particularly in addressing complex challenges that demand extensive knowledge bases. These approaches not only enhance the reliability and explainability of AI-driven methods but also lower the barriers for practitioners without extensive AI expertise. Techniques such as retrieval-augmented generation and prompt engineering provide accessible and effective tools, enabling domain experts to leverage advanced AI systems for accelerating breakthroughs in their fields.

5 Agent-based large-small model collaboration: automated laboratories for scientific discovery

Recent advancements in scientist agents and laboratory robotics have catalyzed the emergence of a transformative agent-based scientific research paradigm, particularly in fields such as biomedicine (Gao et al. 2024c) and chemistry (Bran et al. 2024). In research domains characterized by standardized experimental procedures and a high frequency of experiments, agent-based collaborations can significantly reduce setup complexity and automate routine tasks, freeing scientists to focus on higher-level analysis and innovation while ensuring transparency and oversight through explainability. Within the workflow of agent-based automated laboratories (Fig. 10A), agents powered by LLMs seamlessly integrate with diverse external tools to execute complex sequences of scientific tasks. For example, Coscientist, driven by GPT-4, autonomously designs, plans, and executes scientific experiments by leveraging tools such as internet and documentation search, code execution, and experimental automation, demonstrating its versatility and efficacy across six diverse tasks, including the optimization of palladium-catalyzed cross-couplings (Boiko et al. 2023). However, agent-based automated laboratories are still in their early stages of development and face numerous challenges. LLM-based autonomous agents face signifi-

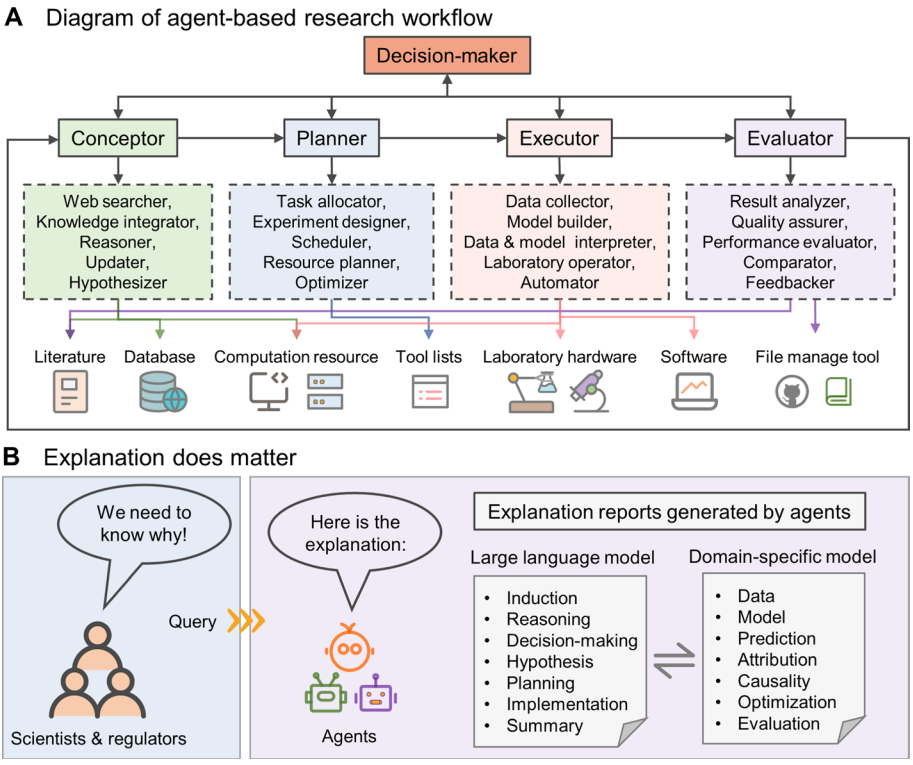


Fig. 10 Diagram of agent-based research workflow and the need for explainability. **A** Diagram of the division of labor among agents, along with the external resources and tools they utilize. **B** Emphasis on the need for agent-based systems to provide explanations for their decision-making processes, as well as for the models and data they employ

cant challenges including role-playing limitations, generalized human alignment, prompt robustness, hallucination, knowledge boundary management, and efficiency (Wang et al. 2024a). The challenges in robotic automation of laboratories lay in developing versatile, safe, and interoperable systems that can perform a broad spectrum of tasks with high precision and autonomy (Angelopoulos et al. 2024).

Ensuring explainability (Fig. 10B) across decision-making processes in intelligent agent-based scientists and LLM-powered robots, as well as throughout data analysis and modeling using both domain-specific models and LLMs, is essential for maintaining scientific rigor and building trust in automated results. Practitioners often face challenges in selecting and interpreting appropriate explainability methods, but LLMs offer more accessible, user-friendly explanations through interactive natural language dialogue systems (Slack et al. 2023). This explanation not only fosters confidence among researchers and stakeholders but also facilitates the validation and reproducibility of scientific findings. Furthermore, the explainability of AI is crucial for harnessing human intellectual engagement and deriving scientific insights that can inform new design principles (Su et al. 2024). As researchers increasingly rely on autonomous experimentation, their role may shift toward translating the results into scientific understanding, a process that is significantly enhanced by advances in XAI (Tom et al. 2024).

6 Comparison between explainable small domain-specific models and LLMs

Understanding the practical distinctions between explainable small domain-specific models and explainable LLMs is essential for their effective application in scientific research. While small models built from scratch offer transparency, precision, and deep integration of expert knowledge within narrowly defined domains, LLMs, leveraging prompt engineering, RAG, and fine-tuning, provide scalable generalization and rapid adaptability across diverse scientific tasks. Figure 11 below systematically compares these two model classes across five key dimensions relevant to scientific use, including explainability, knowledge injection, generalizability, domain-specific performance, and data requirements.

7 Challenges, pitfalls, and future directions

While the development of XAI has provided valuable insights into the mechanics of AI decisions and enhanced trust in AI tools, many expectations have been proposed in existing research, including human-centered design (Kong et al. 2024), robust benchmark (Kenny et al. 2021), and human-machine collaboration system (Saeed and Omlin 2023). These expectations have already been addressed or can be effectively mitigated through the combinations of methods we suggested above. Moreover, several challenges have been pointed out to be encountered within the XAI pipeline, such as knowledge identification (Xie et al. 2021; Tiddi and Schlobach 2022), knowledge representation (Peng et al. 2023), knowledge conflict (Xu et al. 2024), explanation evaluation (Longo et al. 2024), and accuracy trade-off (Crook et al. 2023). These issues can either be technically resolved or become apparent once explanations are provided. In this review, we aim to outline unintended pitfalls related

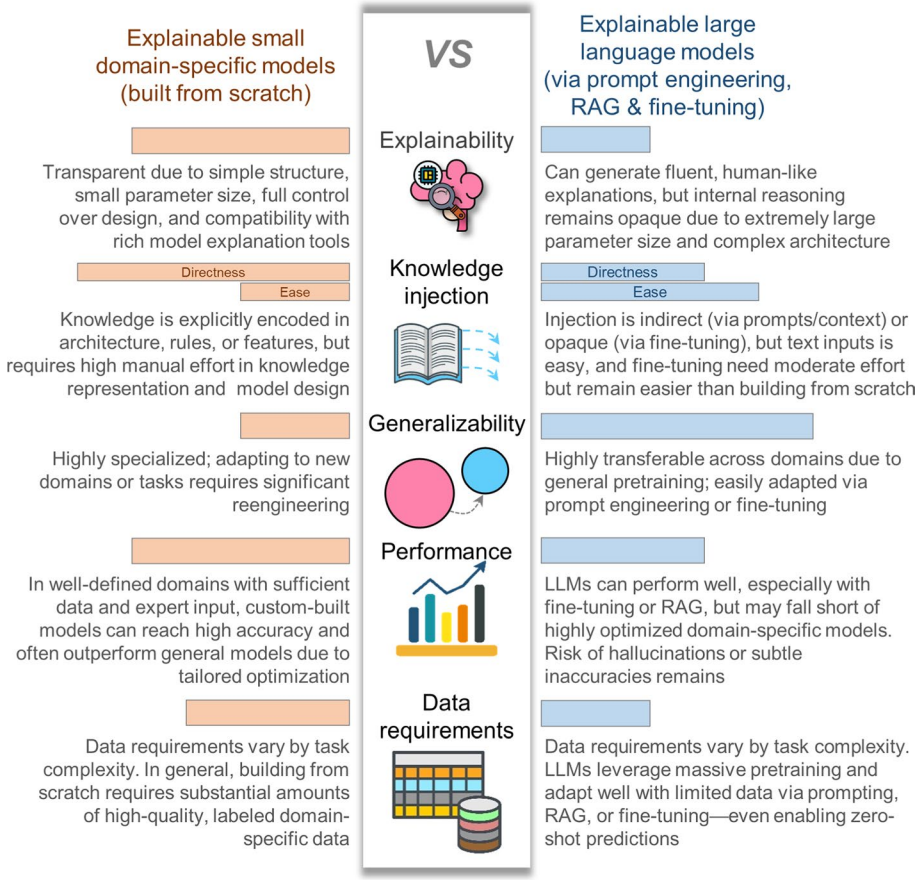


Fig. 11 Comparative assessment of explainable small domain-specific models and LLMs for scientific applications. Small domain-specific models are constructed from the ground up, allowing full architectural control and direct integration of expert knowledge. In contrast, explainable LLMs build upon pretrained foundation models, and are adapted to specific tasks through prompt engineering, RAG, or fine-tuning

to scientific research that imperceptibly make errors, particularly from the perspective of experimental scientists who may integrate XAI tools into their work, discussed as follows:

Illusion of explanation: Explanations yielded by XAI methods sometimes appear reasonable but actually reflect spurious correlations that the model has learned due to biases or errors, even though the model has demonstrated high accuracy. These explanations may fail to uncover the true causal mechanisms within the data generation process, providing only superficial explanations that lack the depth required for a comprehensive understanding of complex scientific phenomena.

Over-reliance on model explanations: Experimental scientists may risk placing excessive trust in the explanations produced by XAI methods (Ehsan and Riedl 2024), potentially leading them to mistakenly accept these explanations as the definitive rationale for model decisions. Such over-reliance is particularly concerning when the model is based on spurious correlations or flawed assumptions (Lapuschkin et al. 2019; Messeri and Crock-

ett 2024), which can result in erroneous research findings or conclusions. Therefore, it is crucial to approach these explanations with a critical mindset and rigorously validate them through experimental verification.

Oversimplified explanations: Explanations produced by XAI methods may be inherently partial. For example, local interpretable model-agnostic explanations (LIME) and SHAP use linear approximations to represent a model's decision rules (Renftle et al. 2024; Salih et al. 2025), potentially overlooking higher-order nonlinear relationships in complex models. It would lead researchers to unintentionally focus on the 'easy-to-interpret' parts of the model while overlooking potentially more critical but less explainable decision processes (Miller 2019b), thereby hindering a comprehensive understanding of the scientific phenomena.

Bias toward hypotheses: Different XAI methods can produce vastly different explanations, and even variations of the parameters can do so (Weber et al. 2023). Researchers might unconsciously favor methods or emphasize a subset of explanations that align with their expectations or hypotheses to support their assumptions (Holman et al. 2015). Such a cognitive bias can lead to the neglect of rigorous comparison and validation of the chosen methods.

High-dimensional explanation: In scientific research, data is often high-dimensional, as seen in fields like genomics, which imposes additional challenges for explanation (Pahud de Mortanges et al. 2024). Summarizing explanations from complex models can be difficult, as these models are more prone to overfitting to noise and irrelevant information. The interaction effects between features for the model can be intricate and multifaceted. Interpreting the model requires nuanced consideration, making it challenging to derive clear and actionable insights from the model.

Explanation of generative AI: Generative AI has garnered significant attention from the research community, particularly in drug (Gangwal and Lavecchia 2024; Farhadi et al. 2025), material (Liu et al. 2023), and protein (Ingraham et al. 2023) design. In these fields, XAI is critical as it empowers scientists to understand why a model produces a specific design, enabling them to adjust, verify, and optimize outputs for safe, effective, and ethical high-impact applications (Schneider 2024). However, achieving true XAI remains challenging due to the opacity of commercial Generative AI models, the inherent complexity of their systems and outputs, and the difficulty in evaluating explanations themselves. Future directions should focus on enhancing interactivity, allowing deeper engagement with explanation systems, and broadening the scope of explanations to encompass not only inputs but also various facets of the generated outputs, including specific attributes and holistic characteristics.

A persistent issue in XAI is balancing explainability and predictive performance, as simpler models offer transparency but often lack the accuracy of deep learning models. Future research should focus on enhancing the explainability of high-performing models without compromising their effectiveness. Another critical challenge is the integration of domain knowledge into XAI models; while prior knowledge can improve reliability and explainability, scalable and systematic methods for its incorporation remain an open research area. Additionally, current evaluation metrics for explainability are often designed for general AI applications rather than scientific discovery, necessitating new domain-specific metrics to ensure AI-generated insights are both valid and practically useful. As LLMs and autonomous systems begin to play a role in hypothesis generation and experimentation, ensuring their outputs are explainable, trustworthy, and actionable will be critical, particularly in

fields where validation and reproducibility are paramount. Ultimately, fostering effective human-AI collaboration, where researchers can intuitively interact with and trust AI explanations, will be essential for embedding XAI into scientific workflows and maximizing its transformative potential.

8 Conclusion

The integration of XAI into scientific research is set to revolutionize discovery, innovation, and collaboration. This review has highlighted three key paradigms: small domain-specific models, LLMs, and agent-based large-small model collaboration. For small domain-specific models, we introduced a knowledge-oriented taxonomy that categorizes methods into knowledge-agnostic, knowledge-based, knowledge-infused, and knowledge-verified approaches. This framework helps researchers balance established knowledge with novel insights, ensuring that models are both reliable and innovative. LLMs have emerged as powerful tools in scientific research, offering advanced capabilities in natural language understanding, generation, reasoning, and interaction. The review discussed strategies for integrating domain knowledge into LLMs and enhancing their explainability, which is crucial for building trust and effective application in scientific discovery. The development of LLM-based scientist agents highlights the need for robust, context-aware explanations tailored to specific research goals. Looking ahead, agent-based automated laboratories represent a transformative shift. These laboratories, powered by explainable LLMs and advanced robotics, will enable the execution of complex scientific tasks with unprecedented precision and efficiency.

However, methodological challenges and usage pitfalls persist. Issues such as spurious correlations, over-reliance on model explanations, and cognitive biases highlight the need for rigorous validation and critical interpretation. Additionally, high-dimensional data and oversimplified explanations demand more nuanced approaches to ensure scientific rigor. Emerging technologies like LLM-based agents and automated laboratories show promise but face hurdles in robustness, explanation, and transparency. Ensuring explainability across all stages of AI integration is essential to maintain trust, foster human-AI collaboration, and drive meaningful scientific insights.

Together, the integration of XAI into scientific research holds immense promise. By leveraging the strengths of explainable small domain-specific models and LLMs, and by addressing methodological challenges and pitfalls, we can usher in a new era of scientific advancement. This review serves as a foundation for future research, encouraging the scientific community to embrace the opportunities and challenges presented by XAI to drive innovation and collaboration in scientific discovery.

Acknowledgements This work is supported by International Collaboration Fund for Creative Research of National Science Foundation of China (NSFC ICFCRT) under the Grant No. W2441019, Westlake Education Foundation under the Grant No. 103110846022301, and China Postdoctoral Science Foundation under the Grant No. 2024M762941.

Author contributions Conceptualization: HY, YJ; Methodology: HY, YW, YJ; Visualization: HY, YW; Supervision: YJ; Funding acquisition: YJ, HY; Writing—original draft: HY, YW; Writing—review & editing: YJ, HY, YZ, YW, TC, SFYL, YY, KD.

Funding Thios study was funded by China Postdoctoral Science Foundation, 2024M762941, International Collaboration Fund for Creative Research Teams of National Natural Science Foundation of China, W2441019, Westlake Education Foundation, 103110846022301.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Agarwal E, Singh J, Dani V, Magazine R, Ganu T, Nambi A (2024) PromptWizard: task-aware prompt optimization framework. https://doi.org/10.48550/ARXIV.2405.18369

Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J, Ouyang C, Gupta A, Gu Y, Deveci M (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 96:156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Angelopoulos A, Cahoon JF, Alterovitz R (2024) Transforming science labs into automated factories of discovery. Sci Robot 9(95):eadm6991. https://doi.org/10.1126/scirobotics.adm6991

Arya V, Bellamy RKE, Chen P-Y, Dhurandhar A, Hind M, Hoffman SC, Houde S, Liao QV, Luss R, Mojsilović A, Mourad S, Pedemonte P, Raghavendra R, Richards J, Sattigeri P, Shanmugam K, Singh M, Varshney KR, Wei D, Zhang Y (2019) One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. https://doi.org/10.48550/ARXIV.1909.03012

Atakishiyev S, Salameh M, Yao H, Goebel R (2024) Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. IEEE Access 12:101603–101625. https://doi.org/10.1109/ACCESS.2024.3431437

Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, DiResta R, Ferrara E, Hale S, Halevy A, Hovy E, Ji H, Menczer F, Miguez R, Nakov P, Scheufele D, Sharma S, Zagni G (2024) Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell 6(8):852–863. https://doi.org/10.1038/s42256-024-00881-z

Bachmann G, Nagarajan V (2024) The pitfalls of next-token prediction. https://doi.org/10.48550/ARXIV.2403.06963

Barkan O, Hauon E, Caciularu A, Katz O, Malkiel I, Armstrong O, Koenigstein N (2021) Grad-SAM: explaining transformers via gradient self-attention maps. In: Proceedings of the 30th ACM international conference on information & knowledge management. Association for computing machinery, New York, NY, USA, pp 2882–2887

Barkan O, Toib Y, Elisha Y, Weill J, Koenigstein N (2024) LLM explainability via attributive masking learning. In: Findings of the association for computational linguistics: EMNLP 2024. association for computational linguistics, Miami, Florida, USA, pp 9522–9537

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Batra R, Song L, Ramprasad R (2021) Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater 6(8):655–678. https://doi.org/10.1038/s41578-020-00255-y

Beckh K, Müller S, Jakobs M, Toborek V, Tan H, Fischer R, Welke P, Houben S, Von Rueden L (2023) Harnessing prior knowledge for explainable machine learning: an overview. 2023 IEEE conference on secure and trustworthy machine learning (SaTML). IEEE, Raleigh, pp 450–463

Bibal A, Cardon R, Alfter D, Wilkens R, Wang X, François T, Watrin P (2022) Is Attention Explanation? An Introduction to the Debate. In: Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for computational linguistics, Dublin, Ireland, pp 3889–3900

Boby ML, Fearon D, Ferla M, Filep M, Koekemoer L, Robinson MC, The COVID Moonshot Consortium‡, Chodera JD, Lee AA, London N, Von Delft A, Von Delft F, Achdout H, Aimon A, Alonzi DS, Arbon R, Aschenbrenner JC, Balcomb BH, Bar-David E, Barr H, Ben-Shmuel A, Bennett J, Bilenko VA, Borden B, Boulet P, Bowman GR, Brewitz L, Brun J, Bvnbs S, Calmiano M, Carbery A, Carney DW, Cattermole E, Chang E, Chernyshenko E, Clyde A, Coffland JE, Cohen G, Cole JC, Contini A, Cox L, Croll TI, Cvitkovic M, De Jonghe S, Dias A, Donckers K, Dotson DL, Douangamath A, Duberstein S, Dudgeon T, Dunnett LE, Eastman P, Erez N, Eyermann CJ, Fairhead M, Fate G, Fedorov O, Fernandes RS, Ferrins L, Foster R, Foster H, Fraisse L, Gabizon R, García-Sastre A, Gawriljuk VO, Gehrtz P, Gileadi C, Giroud C, Glass WG, Glen RC, Glinert I, Godoy AS, Gorichko M, Gorrie-Stone T, Griffen EJ, Haneef A, Hassell Hart S, Heer J, Henry M, Hill M, Horrell S, Huang QYJ, Huliak VD, Hurley MFD, Israely T, Jajack A, Jansen J, Jnoff E, Jochmans D, John T, Kaminow B, Kang L, Kantsadi AL, Kenny PW, Kiappes JL, Kinakh SO, Kovar B, Krojer T, La VNT, Laghnimi-Hahn S, Lefker BA, Levy H, Lithgo RM, Logvinenko IG, Lukacik P, Macdonald HB, MacLean EM, Makower LL, Malla TR, Marples PG, Matviiuk T, McCorkindale W, McGovern BL, Melamed S, Melnykov KP, Michurin O, Miesen P, Mikolajek H, Milne BF, Minh D, Morris A, Morris GM, Morwitzer MJ, Moustakas D, Mowbray CE, Nakamura AM, Neto JB, Neyts J, Nguyen L, Noske GD, Oleinikovas V, Oliva G, Overheul GJ, Owen CD, Pai R, Pan J, Paran N, Payne AM, Perry B, Pingle M, Pinjari J, Politi B, Powell A, Pšenák V, Pulido I, Puni R, Rangel VL, Reddi RN, Rees P, Reid SP, Reid L, Resnick E, Ripka EG, Robinson RP, Rodriguez-Guerra J, Rosales R, Rufa DA, Saar K, Saikatendu KS, Salah E, Schaller D, Scheen J, Schiffer CA, Schofield CJ, Shafeev M, Shaikh A, Shaqra AM, Shi J, Shurrush K, Singh S, Sittner A, Sjö P, Skyner R, Smalley A, Smeets B, Smilova MD, Solmesky LJ, Spencer J, Strain-Damerell C, Swamy V, Tamir H, Taylor JC, Tennant RE, Thompson W, Thompson A, Tomásio S, Tomlinson CWE, Tsurupa IS, Tumber A, Vakonakis I, Van Rij RP, Vangeel L, Varghese FS, Vaschetto M, Vitner EB, Voelz V, Volkamer A, Walsh MA, Ward W, Weatherall C, Weiss S, White KM, Wild CF, Witt KD, Wittmann M, Wright N, Yahalom-Ronen Y, Yilmaz NK, Zaidmann D, Zhang I, Zidane H, Zitzmann N, Zvornicanin SN (2023) Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. Science 382(6671):7201. https://doi.org/10.1126/science.abo7201

Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. Nature 624(7992):570–578. https://doi.org/10.1038/s41586-023-06792-0

Borate P, Rivière J, Marone C, Mali A, Kifer D, Shokouhi P (2023) Using a physics-informed neural network and fault zone acoustic monitoring to predict lab earthquakes. Nat Commun 14(1):3693. https://doi.org/10.1038/s41467-023-39377-6

Bove C, Lesot M-J, Tijus CA, Detyniecki M (2023) Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In: Proceedings of the 28th international conference on intelligent user interfaces. ACM, Sydney NSW Australia, pp 188–203

Bran M, Cox S, Schilter O, Baldassari C, White AD, Schwaller P (2024) Augmenting large language models with chemistry tools. Nat Mach Intell 6(5):525–535. https://doi.org/10.1038/s42256-024-00832-8

Bryant P, Pozzati G, Elofsson A (2022) Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13(1):1265. https://doi.org/10.1038/s41467-022-28865-w

Buergel T, Steinfeldt J, Ruyoga G, Pietzner M, Bizzarri D, Vojinovic D, Upmeier Zu Belzen J, Loock L, Kittner P, Christmann L, Hollmann N, Strangalies H, Braunger JM, Wild B, Chiesa ST, Spranger J, Klostermann F, Van Den Akker EB, Trompet S, Mooijaart SP, Sattar N, Jukema JW, Lavrijssen B, Kavousi M, Ghanbari M, Ikram MA, Slagboom E, Kivimaki M, Langenberg C, Deanfield J, Eils R, Landmesser U (2022) Metabolomic profiles predict individual multidisease outcomes. Nat Med 28(11):2309–2320. https://doi.org/10.1038/s41591-022-01980-3

Caldas Ramos M, Collison C, White AD (2024) A review of large language models and autonomous agents in chemistry. Chem Sci. https://doi.org/10.1039/D4SC03921A

Cálem J, Moreira C, Jorge J (2024) Intelligent systems in healthcare: a systematic survey of explainable user interfaces. Comput Biol Med 180:108908. https://doi.org/10.1016/j.compbiomed.2024.108908

Chanda T, Hauser K, Hobelsberger S, Bucher T-C, Garcia CN, Wies C, Kittler H, Tschandl P, Navarrete-Dechent C, Podlipnik S, Chousakos E, Crnaric I, Majstorovic J, Alhajwan L, Foreman T, Peternel S, Sarap S, Özdemir İ, Barnhill RL, Llamas-Velasco M, Poch G, Korsing S, Sondermann W, Gellrich FF, Heppt MV, Erdmann M, Haferkamp S, Drexler K, Goebeler M, Schilling B, Utikal JS, Ghoreschi K, Fröhling S, Krieghoff-Henning E, Reader Study Consortium, Salava A, Thiem A, Dimitrios A, Ammar AM, Vučemilović AS, Yoshimura AM, Ilieva A, Gesierich A, Reimer-Taschenbrecker A, Kolios AGA, Kalva A, Ferhatosmanoğlu A, Beyens A, Pföhler C, Erdil DI, Jovanovic D, Racz E, Bechara FG, Vaccaro F, Dimitriou F, Rasulova G, Cenk H, Yanatma I, Kolm I, Hoorens I, Sheshova IP, Jocic I, Knuever J, Fleißner J, Thamm JR, Dahlberg J, Lluch-Galcerá JJ, Figueroa JSA, Holzgruber J, Welzel J, Damevska K, Mayer KE, Maul LV, Garzona-Navas L, Bley LI, Schmitt L, Reipen L, Shafik L, Petrovska L, Golle L, Jopen L, Gogilidze M, Burg MR, Morales-Sánchez MA, Sławińska M, Mengoni M, Dragolov M, Iglesias-Pena N, Booken N, Enechukwu NA, Persa O-D, Oninla OA, Theofilogiannakou P, Kage P, Neto RRO, Peralta R, Afiouni R, Schuh S, Schnabl-Scheu S, Vural S, Hudson S, Saa SR, Hartmann S, Damevska S, Finck S, Braun SA, Hartmann T, Welponer T, Sotirovski T, Bondare-Ansberga V, Ahlgrimm-Siess V, Frings VG, Simeonovski V, Zafirovik Z, Maul J-T, Lehr S, Wobser M, Debus D, Riad H, Pereira MP, Lengyel Z, Balcere A, Tsakiri A, Braun RP, Brinker TJ (2024) Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nat Commun 15(1):524. https://doi.org/10.1038/s41467-023-43095-4

Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039

Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, Sahai S, Mahmood F (2023) Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 7(6):719–742. https://doi.org/10.1038/s41551-023-01056-8

Chen V, Yang M, Cui W, Kim JS, Talwalkar A, Ma J (2024) Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. Nat Methods 21(8):1454–1461. https://doi.org/10.1038/s41592-024-02359-7

Chen Q, Ji F, Zeng X, Li F-L, Zhang J, Chen H, Zhang Y (2021) KACE: Generating knowledge aware contrastive explanations for natural language inference. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 2516–2527

Cho SH, Shin K (2023) Feature-weighted counterfactual-based explanation for bankruptcy prediction. Expert Syst Appl 216:119390. https://doi.org/10.1016/j.eswa.2022.119390

Chou Y-L, Moreira C, Bruza P, Ouyang C, Jorge J (2022) Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf Fusion 81:59–83. https://doi.org/10.1016/j.inffus.2021.11.003

Christov-Moore L, Reggente N, Vaccaro A, Schoeller F, Pluimer B, Douglas PK, Iacoboni M, Man K, Damasio A, Kaplan JT (2023) Preventing antisocial robots: a pathway to artificial empathy. Sci Robot 8(80):eabq3658. https://doi.org/10.1126/scirobotics.abq3658

Cornelio C, Dash S, Austel V, Josephson TR, Goncalves J, Clarkson KL, Megiddo N, El Khadir B, Horesh L (2023) Combining data and theory for derivable scientific discovery with AI-Descartes. Nat Commun 14(1):1777. https://doi.org/10.1038/s41467-023-37236-y

Crook B, Schlüter M, Speith T (2023) Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). https://doi.org/10.48550/ARXIV.2307.14239

Cui P, Athey S (2022) Stable learning establishes some common ground between causal inference and machine learning. Nat Mach Intell 4(2):110–115. https://doi.org/10.1038/s42256-022-00445-z

DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, Zhu Q, Ma S, Wang P, Bi X, Zhang X, Yu X, Wu Y, Wu ZF, Gou Z, Shao Z, Li Z, Gao Z, Liu A, Xue B, Wang B, Wu B, Feng B, Lu C, Zhao C, Deng C, Zhang C, Ruan C, Dai D, Chen D, Ji D, Li E, Lin F, Dai F, Luo F, Hao G, Chen G, Li G, Zhang H, Bao H, Xu H, Wang H, Ding H, Xin H, Gao H, Qu H, Li H, Guo J, Li J, Wang J, Chen J, Yuan J, Qiu J, Li J, Cai JL, Ni J, Liang J, Chen J, Dong K, Hu K, Gao K, Guan K, Huang K, Yu K, Wang L, Zhang L, Zhao L, Wang L, Zhang L, Xu L, Xia L, Zhang M, Zhang M, Tang M, Li M, Wang M, Li M, Tian N, Huang P, Zhang P, Wang Q, Chen Q, Du Q, Ge R, Zhang R, Pan R, Wang R, Chen RJ, Jin RL, Chen R, Lu S, Zhou S, Chen S, Ye S, Wang S, Yu S, Zhou S, Pan S, Li SS, Zhou S, Wu S, Yun T, Pei T, Sun T, Wang T, Zeng W, Zhao W, Liu W, Liang W, Gao W, Yu W, Zhang W, Xiao WL, An W, Liu X, Wang X, Chen X, Nie X, Cheng X, Liu X, Xie X, Liu X, Yang X, Li X, Su X, Lin X, Li XQ, Jin X, Shen X, Chen X, Sun X, Wang X, Song X, Zhou X, Wang X, Shan X, Li YK, Wang YQ, Wei YX, Zhang Y, Xu Y, Li Y, Zhao Y, Sun Y, Wang Y, Yu Y, Zhang Y, Shi Y, Xiong Y, He Y, Piao Y, Wang Y, Tan Y, Ma Y, Liu Y, Guo Y, Ou Y, Wang Y, Gong Y, Zou Y, He Y, Xiong Y, Luo Y, You Y, Liu Y, Zhou Y, Zhu YX, Huang Y, Li Y, Zheng Y, Zhu Y, Ma Y, Tang Y, Zha Y, Yan Y, Ren ZZ, Ren Z, Sha Z, Fu Z, Xu Z, Xie Z, Zhang Z, Hao Z, Ma Z, Yan Z, Wu Z, Gu Z, Zhu Z, Liu Z, Li Z, Xie Z, Song Z, Pan Z, Huang Z, Xu Z, Zhang Z, Zhang Z (2025) DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. https://doi.org/10.48550/ARXIV.2501.12948

Diao S, Wang P, Lin Y, Pan R, Liu X, Zhang T (2023) Active prompting with chain-of-thought for large language models. https://doi.org/10.48550/ARXIV.2302.12246

Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Hu S, Chen Y, Chan C-M, Chen W, Yi J, Zhao W, Wang X, Liu Z, Zheng H-T, Chen J, Liu Y, Tang J, Li J, Sun M (2023) Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 5(3):220–235. https://doi.org/10.1038/s42256-023-00626-4

Dong Z, Wang J, Li Y, Deng Y, Zhou W, Zeng X, Gong D, Liu J, Pan J, Shang R, Xu Y, Xu M, Zhang L, Zhang M, Tao X, Zhu Y, Du H, Lu Z, Yao L, Wu L, Yu H (2023) Explainable artificial intelligence incorporated with domain knowledge diagnosing early gastric neoplasms under white light endoscopy. Npj Digit Med 6(1):1–9. https://doi.org/10.1038/s41746-023-00813-y

Duan Y, Coreas R, Liu Y, Bitounis D, Zhang Z, Parviz D, Strano M, Demokritou P, Zhong W (2020) Prediction of protein corona on nanomaterials by machine learning using novel descriptors. NanoImpact 17:100207. https://doi.org/10.1016/j.impact.2020.100207

Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Larson J (2024) From local to global: a graph RAG approach to query-focused summarization. https://doi.org/10.48550/ARXIV.2404.16130

Ehsan U, Riedl MO (2024) Explainability pitfalls: beyond dark patterns in explainable AI. Patterns. https://doi.org/10.1016/j.patter.2024.100971

Enguehard J (2023) Sequential integrated gradients: a simple but effective method for explaining language models. https://doi.org/10.48550/ARXIV.2305.15853

Eshete B (2021) Making machine learning trustworthy. Science 373(6556):743–744. https://doi.org/10.1126/science.abi5052

Farhadi A, Zamanifar A, Faezipour M (2025) Application of generative AI in drug discovery. In: Zamanifar A, Faezipour M (eds) Application of generative AI in healthcare systems. Springer, Cham, pp 155–174

Fel T, Boutin V, Béthune L, Cadene R, Moayeri M, Andéol L, Chalvidal M, Serre T (2023) A holistic approach to unifying automatic concept extraction and concept importance estimation. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds) Advances in neural information processing systems. Curran Associates, Inc, pp 54805–54818

Feng J, Lansford JL, Katsoulakis MA, Vlachos DG (2020) Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. Sci Adv 6(42):eabc3204. https://doi.org/10.1126/sciadv.abc3204

Fortelny N, Bock C (2020) Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol 21(1):190. https://doi.org/10.1186/s13059-020-02100-5

Frank M, Ni P, Jensen M, Gerstein MB (2024) Leveraging a large language model to predict protein phase transition: a physical, multiscale, and interpretable approach. Proc Natl Acad Sci U S A 121(33):e2320510121. https://doi.org/10.1073/pnas.2320510121

Gamble P, Jaroensri R, Wang H, Tan F, Moran M, Brown T, Flament-Auvigne I, Rakha EA, Toss M, Dabbs DJ, Regitnig P, Olson N, Wren JH, Robinson C, Corrado GS, Peng LH, Liu Y, Mermel CH, Steiner DF, Chen P-HC (2021) Determining breast cancer biomarker status and associated morphological features using deep learning. Commun Med 1(1):14. https://doi.org/10.1038/s43856-021-00013-3

Gangwal A, Lavecchia A (2024) Unleashing the power of generative AI in drug discovery. Drug Discov Today 29(6):103992. https://doi.org/10.1016/j.drudis.2024.103992

Gao J, Wang Z, Han Y, Gao M, Li J (2024a) CEEM: a chemically explainable deep learning platform for identifying compounds with low effective mass. Small 20(4):2305918. https://doi.org/10.1002/smll.202305918

Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, Ektefaie Y, Kondic J, Zitnik M (2024c) Empowering biomedical discovery with AI agents. Cell 187(22):6125–6151. https://doi.org/10.1016/j.cell.2024.09.022

Gao L, la Tour TD, Tillman H, Goh G, Troll R, Radford A, Sutskever I, Leike J, Wu J (2024b) Scaling and evaluating sparse autoencoders. https://doi.org/10.48550/ARXIV.2406.04093

Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3(11):e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Ghorbani A, Wexler J, Zou JY, Kim B (2019) Towards automatic concept-based explanations. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems. Curran Associates Inc

Ghosal S, Blystone D, Singh AK, Ganapathysubramanian B, Singh A, Sarkar S (2018) An explainable deep machine vision framework for plant stress phenotyping. Proc Natl Acad Sci U S A 115(18):4613–4618. https://doi.org/10.1073/pnas.1716999115

Górriz JM, Álvarez-Illán I, Álvarez-Marquina A, Arco JE, Atzmueller M, Ballarini F, Barakova E, Bologna G, Bonomini P, Castellanos-Dominguez G, Castillo-Barnes D, Cho SB, Contreras R, Cuadra JM, Domínguez E, Domínguez-Mateos F, Duro RJ, Elizondo D, Fernández-Caballero A, Fernandez-Jover E, Formoso MA, Gallego-Molina NJ, Gamazo J, González JG, Garcia-Rodriguez J, Garre C, Garrigós J, Gómez-Rodellar A, Gómez-Vilda P, Graña M, Guerrero-Rodriguez B, Hendrikse SCF, Jimenez-Mesa C, Jodra-Chuan M, Julian V, Kotz G, Kutt K, Leming M, De Lope J, Macas B, Marrero-Aguiar V, Martinez JJ, Martinez-Murcia FJ, Martínez-Tomás R, Mekyska J, Nalepa GJ, Novais P, Orellana D, Ortiz A, Palacios-Alonso D, Palma J, Pereira A, Pinacho-Davidson P, Pinninghoff MA, Ponticorvo M, Psarrou A, Ramírez J, Rincón M, Rodellar-Biarge V, Rodríguez-Rodríguez I, Roelofsma PHMP, Santos J, Salas-Gonzalez D, Salcedo-Lagos P, Segovia F, Shoeibi A, Silva M, Simic D, Suckling J, Treur J, Tsanas A, Varela R, Wang SH, Wang W, Zhang YD, Zhu H, Zhu Z, Ferrández-Vicente JM (2023) Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. Inf Fusion 100:101945. https://doi.org/10.1016/j.inffus.2023.101945

Green B, Chen Y (2021) Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. Proc ACM Hum-Comput Interact 5(CSCW2):1–33. https://doi.org/10.1145/3479562

Guidotti R (2024) Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Discov 38(5):2770–2824. https://doi.org/10.1007/s10618-022-00831-6

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42. https://doi.org/10.1145/3236009

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI—Explainable artificial intelligence. Sci Robot 4(37):eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Guo Z, Xia L, Yu Y, Ao T, Huang C (2024) LightRAG: simple and fast retrieval-augmented generation. https://doi.org/10.48550/ARXIV.2410.05779

Hall SW, Sakzad A, Choo KR (2022) Explainable artificial intelligence for digital forensics. Wires Forensic Sci 4(2):e1434. https://doi.org/10.1002/wfs2.1434

Harren T, Matter H, Hessler G, Rarey M, Grebner C (2022) Interpretation of structure-activity relationships in real-world drug design data sets using explainable artificial intelligence. J Chem Inf Model 62(3):447–462. https://doi.org/10.1021/acs.jcim.1c01263

Hirsch V, Reimann P, Treder-Tschechlov D, Schwarz H, Mitschang B (2023) Exploiting domain knowledge to address class imbalance and a heterogeneous feature space in multi-class classification. VLDB J 32(5):1037–1064. https://doi.org/10.1007/s00778-023-00780-6

Holman L, Head ML, Lanfear R, Jennions MD (2015) Evidence of experimental bias in the life sciences: why we need blind data recording. PLoS Biol 13(7):e1002190. https://doi.org/10.1371/journal.pbio.1002190

Hou Z, Yang Y, Ma Z, Wong K, Li X (2023) Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning. Commun Biol 6(1):73. https://doi.org/10.1038/s42003-023-04462-5

Hu X, Fu H, Wang J, Wang Y, Li Z, Xu R, Lu Y, Jin Y, Pan L, Lan Z (2024) Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. https://doi.org/10.48550/ARXIV.2410.14255

Huang S, Mamidanna S, Jangam S, Zhou Y, Gilpin LH (2023) Can large language models explain themselves? A Study of LLM-generated self-explanations. https://doi.org/10.48550/arXiv.2310.11207

Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, Tie S, Xue V, Cowles SC, Leung A, Rodrigues JV, Morales-Perez CL, Ayoub AM, Green R, Puentes K, Oplinger F, Panwar NV, Obermeyer F, Root AR, Beam AL, Poelwijk FJ, Grigoryan G (2023) Illuminating protein space with a programmable generative model. Nature 623(7989):1070–1078. https://doi.org/10.1038/s41586-023-06728-8

Ismail AA, Oikarinen T, Wang A, Adebayo J, Stanton S, Joren T, Kleinhenz J, Goodman A, Bravo HC, Cho K, Frey NC (2024) Concept bottleneck language models for protein design. https://doi.org/10.48550/ARXIV.2411.06090

Jablonka KM, Schwaller P, Ortega-Guerrero A, Smit B (2024) Leveraging large language models for predictive chemistry. Nat Mach Intell 6(2):161–169. https://doi.org/10.1038/s42256-023-00788-1

Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 1(1):011002. https://doi.org/10.1063/1.4812323

Jain M, Deleu T, Hartford J, Liu C-H, Hernandez-Garcia A, Bengio Y (2023) GFlownets for AI-driven scientific discovery. Digit Discov 2(3):557–577. https://doi.org/10.1039/D3DD00002H

Jawahar G, Sagot B, Seddah D (2019) What Does BERT learn about the structure of language? In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for computational linguistics, Florence, Italy, pp 3651–3657

Jiang S, Tarasova L, Yu G, Zscheischler J (2024) Compounding effects in flood drivers challenge estimates of extreme river floods. Sci Adv 10(13):eadl4005. https://doi.org/10.1126/sciadv.adl4005

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY-J (2019) Reliable and explainable machine-learning methods for accelerated material discovery. NPJ Comput Mater 5(1):108. https://doi.org/10.1038/s41524-019-0248-2

Kang Y, Kim J (2024) ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. Nat Commun 15(1):4705. https://doi.org/10.1038/s41467-024-48998-4

Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, Gundersen OE, Hofman JM, Hullman J, Lones MA, Malik MM, Nanayakkara P, Poldrack RA, Raji ID, Roberts M, Salganik MJ, Serra-Garcia M, Stewart BM, Vandewiele G, Narayanan A (2024) Reforms: consensus-based recommendations for machine-learning-based science. Sci Adv 10(18):eadk3452. https://doi.org/10.1126/sciadv.adk3452

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L (2021) Physics-informed machine learning. Nat Rev Phys 3(6):422–440. https://doi.org/10.1038/s42254-021-00314-5

Kenny EM, Ford C, Quinn M, Keane MT (2021) Explaining black-box classifiers using *post-hoc* explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif Intell 294:103459. https://doi.org/10.1016/j.artint.2021.103459

Keyl P, Bischoff P, Dernbach G, Bockmayr M, Fritz R, Horst D, Blüthgen N, Montavon G, Müller K-R, Klauschen F (2023) Single-cell gene regulatory network prediction by explainable AI. Nucleic Acids Res 51(4):e20–e20. https://doi.org/10.1093/nar/gkac1212

Klauschen F, Dippel J, Keyl P, Jurmeister P, Bockmayr M, Mock A, Buchstab O, Alber M, Ruff L, Montavon G, Müller K-R (2024) Toward explainable artificial intelligence for precision pathology. Annu Rev Pathol Mech Dis 19(1):541–570. https://doi.org/10.1146/annurev-pathmechdis-051222-113147

Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O (2020) Captum: A unified and generic model interpretability library for PyTorch. https://doi.org/10.48550/ARXIV.2009.07896

Kong X, Liu S, Zhu L (2024) Toward human-centered XAI in practice: a survey. Mach Intell Res 21(4):740–770. https://doi.org/10.1007/s11633-022-1407-3

Krenn M, Pollice R, Guo SY, Aldeghi M, Cervera-Lierta A, Friederich P, dos Passos GG, Häse F, Jinich A, Nigam A, Yao Z, Aspuru-Guzik A (2022) On scientific understanding with artificial intelligence. Nat Rev Phys 4(12):761–769. https://doi.org/10.1038/s42254-022-00518-3

Lai V, Zhang Y, Chen C, Liao QV, Tan C (2023) Selective explanations: leveraging human input to align explainable AI. Proc ACM Hum-Comput Interact 7(CSCW2):1–35. https://doi.org/10.1145/3610206

Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, Hu W, Merose A, Hoyer S, Holland G, Vinyals O, Stott J, Pritzel A, Mohamed S, Battaglia P (2023) Learning skillful medium-range global weather forecasting. Science 382(6677):1416–1421. https://doi.org/10.1126/science.adi2336

Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B (2019) Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med 94:42–53. https://doi.org/10.1016/j.artmed.2019.01.001

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell 296:103473. https://doi.org/10.1016/j.artint.2021.103473

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R (2019) Unmasking clever Hans predictors and assessing what machines really learn. Nat Commun 10(1):1096. https://doi.org/10.1038/s41467-019-08987-4

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Leist AK, Klee M, Kim JH, Rehkopf DH, Bordas SPA, Muniz-Terrera G, Wade S (2022) Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Sci Adv 8(42):eabk1942. https://doi.org/10.1126/sciadv.abk1942

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc, pp 9459–9474

Li J, Liu L, Le TD, Liu J (2020) Accurate data-driven prediction does not mean high reproducibility. Nat Mach Intell 2(1):13–15. https://doi.org/10.1038/s42256-019-0140-2

Li H, Yuan R, Liang H, Wang WY, Li J, Wang J (2022a) Towards high entropy alloy with enhanced strength and ductility using domain knowledge constrained active learning. Mater des 223:111186. https://doi.org/10.1016/j.matdes.2022.111186

Li X-H, Cao CC, Shi Y, Bai W, Gao H, Qiu L, Wang C, Gao Y, Zhang S, Xue X, Chen L (2022b) A survey of data-driven and knowledge-aware eXplainable AI. IEEE Trans Knowl Data Eng 34(1):29–49. https://doi.org/10.1109/TKDE.2020.2983930

Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim L-L, Lim CC, Ruamviboonsuk P, Raman R, Corsino L, Echouffo-Tcheugui JB, Luk AOY, Chen LJ, Sun X, Hamzah H, Wu Q, Wang X, Liu R, Wang YX, Chen T, Zhang X, Yang X, Yin J, Wan J, Du W, Quek TC, Goh JHL, Yang D, Hu X, Nguyen TX, Szeto SKH, Chotcomwongse P, Malek R, Normatova N, Ibragimova N, Srinivasan R, Zhong P, Huang W, Deng C, Ruan L, Zhang C, Zhang C, Zhou Y, Wu C, Dai R, Koh SWC, Abdullah A, Hee NKY, Tan HC, Liew ZH, Tien CS-Y, Kao SL, Lim AYL, Mok SF, Sun L, Gu J, Wu L, Li T, Cheng D, Wang Z, Qin Y, Dai L, Meng Z, Shu J, Lu Y, Jiang N, Hu T, Huang S, Huang G, Yu S, Liu D, Ma W, Guo M, Guan X, Yang X, Bascaran C, Cleland CR, Bao Y, Ekinci EI, Jenkins A, Chan JCN, Bee YM, Sivaprasad S, Shaw JE, Simó R, Keane PA, Cheng C-Y, Tan GSW, Jia W, Tham Y-C, Li H, Sheng B, Wong TY (2024) Integrated image-based deep learning and language models for primary diabetes care. Nat Med 30(10):2886–2896. https://doi.org/10.1038/s41591-024-03139-8

Lipton P (1990) Contrastive explanation. Roy Inst Philos Suppl 27:247–266. https://doi.org/10.1017/S1358246100005130

Littmann M, Selig K, Cohen-Lavi L, Frank Y, Hönigschmid P, Kataka E, Mösch A, Qian K, Ron A, Schmid S, Sorbie A, Szlak L, Dagan-Wiener A, Ben-Tal N, Niv MY, Razansky D, Schuller BW, Ankerst D, Hertz T, Rost B (2020) Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nat Mach Intell 2(1):18–24. https://doi.org/10.1038/s42256-019-0139-8

Liu Y, Yang Z, Yu Z, Liu Z, Liu D, Lin H, Li M, Ma S, Avdeev M, Shi S (2023) Generative artificial intelligence and its applications in materials science: current situation and future perspectives. J Materiomics 9(4):798–816. https://doi.org/10.1016/j.jmat.2023.05.001

Liu L, Zhou W, Guan K, Peng B, Xu S, Tang J, Zhu Q, Till J, Jia X, Jiang C, Wang S, Qin Z, Kong H, Grant R, Mezbahuddin S, Kumar V, Jin Z (2024a) Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. Nat Commun 15(1):357. https://doi.org/10.1038/s41467-023-43860-5

Liu S, Wen T, Pattamatta ASLS, Srolovitz DJ (2024b) A prompt-engineered large language model, deep learning workflow for materials classification. Mater Today 80:240–249. https://doi.org/10.1016/j.mattod.2024.08.028

Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, Wang H, Tao L, Sun Y, Song Z, Hong T, Yang J, Gao T, Zhang J, Li X, Zhang J, Sang Y, Yang Z, Xue K, Wu S, Zhang P, Yang J, Song C, Wang G (2025) A generalist medical language model for disease diagnosis assistance. Nat Med 31(3):932–942. https://doi.org/10.1038/s41591-024-03416-6

London AJ (2019) Artificial intelligence and black-box medical decisions: *accuracy versus explainability*. Hastings Cent Rep 49(1):15–21. https://doi.org/10.1002/hast.973

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S (2024) Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf Fusion 106:102301. https://doi.org/10.1016/j.inffus.2024.102301

Lotfollahi M, Rybakov S, Hrovatin K, Hediyeh-zadeh S, Talavera-López C, Misharin AV, Theis FJ (2023) Biologically informed deep learning to query gene programs in single-cell atlases. Nat Cell Biol. https://doi.org/10.1038/s41556-022-01072-x

Lou S, Yu Z, Huang Z, Wang H, Pan F, Li W, Liu G, Tang Y (2024) In silico prediction of chemical acute dermal toxicity using explainable machine learning methods. Chem Res Toxicol 37(3):513–524. https://doi.org/10.1021/acs.chemrestox.4c00012

Lu J, Choi K, Eremeev M, Gobburu J, Goswami S, Liu Q, Mo G, Musante CJ, Shahin MH (2025) Large language models and their applications in drug discovery and development: a primer. Clin Transl Sci 18(4):e70205. https://doi.org/10.1111/cts.70205
Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. https://doi.org/10.48550/ARXIV.1705.07874
Luo X, Zhang D, Zhu X (2021) Deep learning based forecasting of photovoltaic power generation by incorporating domain knowledge. Energy 225:120240. https://doi.org/10.1016/j.energy.2021.120240
Luu RK, Buehler MJ (2024) BioinspiredLLM: conversational large language model for the mechanics of biological and bio-inspired materials. Adv Sci 11(10):2306724. https://doi.org/10.1002/advs.202306724
Madsen A, Chandar S, Reddy S (2024) Are self-explanations from large language models faithful? In: Findings of the association for computational linguistics ACL 2024. Association for computational linguistics, Bangkok, Thailand and virtual meeting, pp 295–337
Messeri L, Crockett MJ (2024) Artificial intelligence and illusions of understanding in scientific research. Nature 627(8002):49–58. https://doi.org/10.1038/s41586-024-07146-0
Miller T (2019a) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
Minh D, Wang HX, Li YF, Nguyen TN (2022) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55(5):3503–3568. https://doi.org/10.1007/s10462-021-10088-y
Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency. ACM, Atlanta GA USA, pp 279–288
Molnar C (2018) iml: An R package for interpretable machine learning. JOSS 3(26):786. https://doi.org/10.21105/joss.00786
Mukherjee S, Mitra A, Jawahar G, Agarwal S, Palangi H, Awadallah A (2023) Orca: progressive learning from complex explanation traces of GPT-4. https://doi.org/10.48550/arXiv.2306.02707
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci USA 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116
Murdock RJ, Kauwe SK, Wang AY-T, Sparks TD (2020) Is domain knowledge necessary for machine learning materials properties? Integr Mater Manuf Innov 9(3):221–227. https://doi.org/10.1007/s40192-020-00179-z
Nielsen RL, Monfeuga T, Kitchen RR, Egerod L, Leal LG, Schreyer ATH, Gade FS, Sun C, Helenius M, Simonsen L, Willert M, Tahrani AA, McVey Z, Gupta R (2024) Data-driven identification of predictive risk biomarkers for subgroups of osteoarthritis using interpretable machine learning. Nat Commun 15(1):2817. https://doi.org/10.1038/s41467-024-46663-4
Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M, Alani H, Berendt B, Kruegel T, Heinze C, Broelemann K, Kasneci G, Tiropanis T, Staab S (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wires Data Min & Knowl 10(3):e1356. https://doi.org/10.1002/widm.1356
Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono NTP, Ramasamy S, DeCost BL, Tian SIP, Romano G, Gilad Kusne A, Buonassisi T (2019) Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. NPJ Comput Mater 5(1):60. https://doi.org/10.1038/s41524-019-0196-x
Oviedo F, Ferres JL, Buonassisi T, Butler KT (2022) Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res 3(6):597–607. https://doi.org/10.1021/accountsmr.1c00244
Oxford English Dictionary (2024) s.v. "science." https://doi.org/10.1093/OED/9588368826
Pahud de Mortanges A, Luo H, Shu SZ, Kamath A, Suter Y, Shelan M, Pöllinger A, Reyes M (2024) Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. NPJ Digit Med 7(1):1–10. https://doi.org/10.1038/s41746-024-01190-w
Peng C, Xia F, Naseriparsa M, Osborne F (2023) Knowledge graphs: opportunities and challenges. Artif Intell Rev 56(11):13071–13102. https://doi.org/10.1007/s10462-023-10465-9
Peng H, Wang X, Hu S, Jin H, Hou L, Li J, Liu Z, Liu Q (2022) COPEN: Probing conceptual knowledge in pre-trained language models. https://doi.org/10.48550/arXiv.2211.04079
Polak MP, Morgan D (2024) Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 15(1):1569. https://doi.org/10.1038/s41467-024-45914-8
Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, Rich S, Wang M, Buchan IE, Bian J (2020) Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2(7):369–375. https://doi.org/10.1038/s42256-020-0197-y

Qiu S, Miller MI, Joshi PS, Lee JC, Xue C, Ni Y, Wang Y, De Anda-Duran I, Hwang PH, Cramer JA, Dwyer BC, Hao H, Kaku MC, Kedar S, Lee PH, Mian AZ, Murman DL, O'Shea S, Paul AB, Saint-Hilaire M-H, Alton Sartor E, Saxena AR, Shih LC, Small JE, Smith MJ, Swaminathan A, Takahashi CE, Taraschenko O, You H, Yuan J, Zhou Y, Zhu S, Alosco ML, Mez J, Stein TD, Poston KL, Au R, Kolachalama VB (2022) Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun 13(1):3404. https://doi.org/10.1038/s41467-022-31037-5

Rebuffi S-A, Fong R, Ji X, Vedaldi A (2020) There and back again: revisiting backpropagation saliency methods. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Renftle M, Trittenbach H, Poznic M, Heil R (2024) What do algorithms explain? The issue of the goals and capabilities of explainable artificial intelligence (XAI). Humanit Soc Sci Commun 11(1):1–10. https://doi.org/10.1057/s41599-024-03277-x

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, San Francisco California USA, pp 1135–1144

Roh Y, Heo G, Whang SE (2021) A survey on data collection for machine learning: a big data - AI integration perspective. IEEE Trans Knowl Data Eng 33(4):1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

Rong Y, Leemann T, Nguyen T-T, Fiedler L, Qian P, Unhelkar V, Seidel T, Kasneci G, Kasneci E (2024) Towards human-centered explainable AI: a survey of user studies for model explanations. IEEE Trans Pattern Anal Mach Intell 46(4):2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. IEEE Access 8:42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Sachan S, Yang J-B, Xu D-L, Benavides DE, Li Y (2020) An explainable AI decision-support-system to automate loan underwriting. Expert Syst Appl 144:113100. https://doi.org/10.1016/j.eswa.2019.113100

Saeed W, Omlin C (2023) Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst 263:110273. https://doi.org/10.1016/j.knosys.2023.110273

Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, Menegaz G (2025) A perspective on explainable artificial intelligence methods: SHAP and LIME. Adv Intell Syst n/a(n/a):2400304. https://doi.org/10.1002/aisy.202400304

Schneider J (2024) Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda. Artif Intell Rev 57(11):289. https://doi.org/10.1007/s10462-024-10916-x

Shanker VR, Bruun TUJ, Hie BL, Kim PS (2024) Unsupervised evolution of protein and antibody complexes with a structure-informed language model. Science 385(6704):46–53. https://doi.org/10.1126/science.adk8946

Singh C, Nasseri K, Tan Y, Tang T, Yu B (2021) Imodels: a python package for fitting interpretable models. JOSS 6(61):3192. https://doi.org/10.21105/joss.03192

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2

Slack D, Krishna S, Lakkaraju H, Singh S (2023) Explaining machine learning models with interactive natural language conversations using TalkToModel. Nat Mach Intell 5(8):873–883. https://doi.org/10.1038/s42256-023-00692-8

Smer-Barreto V, Quintanilla A, Elliott RJR, Dawson JC, Sun J, Campa VM, Lorente-Macías Á, Unciti-Broceta A, Carragher NO, Acosta JC, Oyarzún DA (2023) Discovery of senolytics using machine learning. Nat Commun 14(1):3445. https://doi.org/10.1038/s41467-023-39120-1

Sokol K, Flach P (2020) One explanation does not fit all: the promise of interactive explanations for machine learning transparency. Kunstl Intell 34(2):235–250. https://doi.org/10.1007/s13218-020-00637-y

Stepin I, Alonso JM, Catala A, Pereira-Farina M (2021) A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9:11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Su Y, Wang X, Ye Y, Xie Y, Xu Y, Jiang Y, Wang C (2024) Automation and machine learning augmented by large language models in a catalysis study. Chem Sci 15(31):12200–12233. https://doi.org/10.1039/D3SC07012C

Subhashini LDCS, Li Y, Zhang J, Atukorale AS (2022) Integration of fuzzy logic and a convolutional neural network in three-way decision-making. Expert Syst Appl 202:117103. https://doi.org/10.1016/j.eswa.2022.117103

Tan EX, Chen Y, Lee YH, Leong YX, Leong SX, Stanley CV, Pun CS, Ling XY (2022) Incorporating plasmonic featurization with machine learning to achieve accurate and bidirectional prediction of nanoparticle size and size distribution. Nanoscale Horiz 7(6):626–633. https://doi.org/10.1039/D2NH00146B
Tao H, Wu T, Aldeghi M, Wu TC, Aspuru-Guzik A, Kumacheva E (2021) Nanoparticle synthesis assisted by machine learning. Nat Rev Mater 6(8):701–716. https://doi.org/10.1038/s41578-021-00337-5
Tiddi I, Schlobach S (2022) Knowledge graphs as tools for explainable machine learning: a survey. Artif Intell 302:103627. https://doi.org/10.1016/j.artint.2021.103627
Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learning Syst 32(11):4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
Togo MV, Mastrolorito F, Ciriaco F, Trisciuzzi D, Tondo AR, Gambacorta N, Bellantuono L, Monaco A, Leonetti F, Bellotti R, Altomare CD, Amoroso N, Nicolotti O (2023) Tiresia: an eXplainable artificial intelligence platform for predicting developmental toxicity. J Chem Inf Model 63(1):56–66. https://doi.org/10.1021/acs.jcim.2c01126
Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, Lo S, Pablo-García S, Rajaonson EM, Skreta M, Yoshikawa N, Corapi S, Akkoc GD, Strieth-Kalthoff F, Seifrid M, Aspuru-Guzik A (2024) Self-driving laboratories for chemistry and materials science. Chem Rev 124(16):9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055
Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
Van Noorden R, Perkel JM (2023) AI and science: what 1,600 researchers think. Nature 621(7980):672–675. https://doi.org/10.1038/d41586-023-02980-0
Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 76:89–106. https://doi.org/10.1016/j.inffus.2021.05.009
Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J (2021) Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2021.3079836
Wang H, Li Y, He N, Ma K, Meng D, Zheng Y (2022a) Dicdnet: deep interpretable convolutional dictionary network for metal artifact reduction in CT images. IEEE Trans Med Imaging 41(4):869–880. https://doi.org/10.1109/TMI.2021.3127074
Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J (2024a) A survey on large language model based autonomous agents. Front Comput Sci 18(6):186345. https://doi.org/10.1007/s11704-024-40231-1
Wang K, Variengien A, Conmy A, Shlegeris B, Steinhardt J (2022b) Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. https://doi.org/10.48550/arXiv.2211.00593
Wang Z, Lin Z, Lin W, Yang M, Zeng M, Tan KC (2024b) Explainable molecular property prediction: aligning chemical concepts with predictions via language models. https://doi.org/10.48550/arXiv.2405.16041
Weber L, Lapuschkin S, Binder A, Samek W (2023) Beyond explaining: Opportunities and challenges of XAI-based model improvement. Inf Fusion 92:154–176. https://doi.org/10.1016/j.inffus.2022.11.013
Wei Y, Wu J, Wu Y, Liu H, Meng F, Liu Q, Midgley AC, Zhang X, Qi T, Kang H, Chen R, Kong D, Zhuang J, Yan X, Huang X (2022) Prediction and design of nanozymes using explainable machine learning. Adv Mater 34(27):2201736. https://doi.org/10.1002/adma.202201736
Whang SE, Roh Y, Song H, Lee J-G (2023) Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J 32(4):791–813. https://doi.org/10.1007/s00778-022-00775-9
Willard J, Jia X, Xu S, Steinbach M, Kumar V (2022) Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Comput Surv 55(4):66:1–66:37. https://doi.org/10.1145/3514228
Wong F, De La Fuente-Nunez C, Collins JJ (2023) Leveraging artificial intelligence in the fight against infectious diseases. Science 381(6654):164–170. https://doi.org/10.1126/science.adh1114
Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, Li A, Cubillos-Ruiz A, Krishnan A, Jin W, Manson AL, Friedrichs J, Helbig R, Hajian B, Fiejtek DK, Wagner FF, Soutter HH, Earl AM, Stokes JM, Renner LD, Collins JJ (2024) Discovery of a structural class of antibiotics with explainable deep learning. Nature 626(7997):177–185. https://doi.org/10.1038/s41586-023-06887-8
Wu W, Song C, Zhao J, Xu Z (2023a) Physics-informed gated recurrent graph attention unit network for anomaly detection in industrial cyber-physical systems. Inf Sci 629:618–633. https://doi.org/10.1016/j.ins.2023.01.136
Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh C-Y, Hou T (2023b) From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. J Chem Inf Model 63(24):7617–7627. https://doi.org/10.1021/acs.jcim.3c01642

Wu Z, Zhang O, Wang X, Fu L, Zhao H, Wang J, Du H, Jiang D, Deng Y, Cao D, Hsieh C-Y, Hou T (2024) Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. Nat Mach Intell 6(11):1359–1369. https://doi.org/10.1038/s42256-024-00916-5

Xiang H, Xiao Y, Li F, Li C, Liu L, Deng T, Yan C, Zhou F, Wang X, Ou J, Lin Q, Hong R, Huang L, Luo L, Lin H, Lin X, Chen H (2024) Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nat Commun 15(1):2681. https://doi.org/10.1038/s41467-024-46700-2

Xie X, Niu J, Liu X, Chen Z, Tang S, Yu S (2021) A survey on incorporating domain knowledge into deep learning for medical image analysis. Med Image Anal 69:101985. https://doi.org/10.1016/j.media.2021.101985

Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, Xu W (2024) Knowledge conflicts for LLMs: a survey. https://doi.org/10.48550/ARXIV.2403.08319

Yang J, Tao L, He J, McCutcheon JR, Li Y (2022) Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. Sci Adv 8(29):eabn9545. https://doi.org/10.1126/sciadv.abn9545

Yang M, Zhu M, Wang Y, Chen L, Zhao Y, Wang X, Han B, Zheng X, Yin J (2024) Fine-tuning large language model based explainable recommendation with explainable quality reward. AAAI 38(8):9250–9259. https://doi.org/10.1609/aaai.v38i8.28777

Yang S, Huang S, Zou W, Zhang J, Dai X, Chen J (2023) Local interpretation of transformer based on linear decomposition. In: Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for computational linguistics, Toronto, Canada, pp 10270–10287

Yu R, Wang R (2024) Learning dynamical systems from data: an introduction to physics-guided deep learning. Proc Natl Acad Sci U S A 121(27):e2311808121. https://doi.org/10.1073/pnas.2311808121

Yu H, Tang S, Li SFY, Cheng F (2023) Averaging strategy for interpretable machine learning on small datasets to understand element uptake after seed nanotreatment. Environ Sci Technol 57(34):12760–12770. https://doi.org/10.1021/acs.est.3c01878

Yu H, Tang S, Hamed EM, Li SFY, Jin Y, Cheng F (2024) Optimizing the benefit–risk trade-off in nano-agrochemicals through explainable machine learning: beyond concentration. Environ Sci: Nano. https://doi.org/10.1039/D4EN00213J

Yu H, Jin Y (2025) Unlocking the potential of AI researchers in scientific discovery: What Is Missing? https://doi.org/10.48550/ARXIV.2503.05822

Zhang E, Dao M, Karniadakis GE, Suresh S (2022) Analyses of internal structures and defects in materials using physics-informed neural networks. Sci Adv 8(7):eabk0644. https://doi.org/10.1126/sciadv.abk0644

Zhang W, Wang Q, Kong X, Xiong J, Ni S, Cao D, Niu B, Chen M, Li Y, Zhang R, Wang Y, Zhang L, Li X, Xiong Z, Shi Q, Huang Z, Fu Z, Zheng M (2024) Fine-tuning large language models for chemical text mining. Chem Sci 15(27):10600–10611. https://doi.org/10.1039/D4SC00924J

Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM, Barcelona Spain, pp 295–305

Zhao L, Ciallella HL, Aleksunes LM, Zhu H (2020) Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug Discov Today 25(9):1624–1638. https://doi.org/10.1016/j.drudis.2020.07.005

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2024) Explainability for large language models: a survey. ACM Trans Intell Syst Technol 15(2):1–38. https://doi.org/10.1145/3639372

Zhou Z, Zhang L, Yu Y, Wu B, Li M, Hong L, Tan P (2024) Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. Nat Commun 15(1):5566. https://doi.org/10.1038/s41467-024-49798-6

Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, Ba J (2022) Large Language models are human-level prompt engineers. https://doi.org/10.48550/ARXIV.2211.01910

Authors and Affiliations

Hengjie Yu[1,2] · Yizhi Wang[1] · Tao Cheng[3] · Yan Yan[4,5] · Kenneth A. Dawson[4] · Sam F. Y. Li[6] · Yefeng Zheng[1,2] · Yaochu Jin[1,2]

✉ Yaochu Jin
jinyaochu@westlake.edu.cn

[1] School of Engineering, Westlake University, Hangzhou 310030, Zhejiang, China

[2] Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou 310024, Zhejiang, China

[3] SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK

[4] Centre for BioNano Interactions, School of Chemistry, University College Dublin, Belfield, Dublin 4, Ireland

[5] School of Biomolecular and Biomedical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

[6] Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Singapore