npj | computational materials

Perspective

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01538-0

Foundation models for materials discovery – current state and future directions

Check for updates

Edward O. Pyzer-Knapp[1,2] ✉, Matteo Manica[1], Peter Staar[1], Lucas Morin[1], Patrick Ruch[1], Teodoro Laino[1], John R. Smith[3] & Alessandro Curioni[1]

Large language models, commonly known as LLMs, are showing promise in tacking some of the most complex tasks in AI. In this perspective, we review the wider field of foundation models—of which LLMs are a component—and their application to the field of materials discovery. In addition to the current state of the art—including applications to property prediction, synthesis planning and molecular generation—we also take a look to the future, and posit how new methods of data capture, and indeed modalities of data, will influence the direction of this emerging field.

The story of AI is a story of data representations (Fig. 1). Early expert systems relied on hand-crafted symbolic representations which eventually evolved into task specific, hand-crafted representations for early machine learning applications[1]. Since hand-crafting a representation can act as a panacea for a lack of data, and capture a large amount of prior knowledge, this approach persisted for many years. As the availability of data grew, and the amount of compute available to apply to the problem grew with it, thoughts turned to more automated, data-driven ways to learn these representations utilizing the newly popular approach of deep learning[2–4]. This approach, whilst bereft of the injected prior knowledge of their hand-crafted cousins, started to address the related issue of the implicit inclusion of human biases. As this approach gained popularity, driven in part by the advent and enthusiastic uptake of GPUs for model training[5,6], we saw a paradigm shift in the way data was considered, with significant effort being placed in the collection and curation of large data sets for training deep learning models[7,8]. Of course, there is a practical, if not fundamental, limit to the number of clean and large data sets which can be used for such tasks; and fundamental questions about the novelty of scientific discovery which can be brought to bear using models where so much data is already known, which brought into sharp focus the need for more generalizable representations. In 2017, the invention of the transformer architecture[9], which was then developed into the generative pretrained transformer (GPT) models by OpenAI[10–13], demonstrated that there was a route to generalized representations through the mechanism of self-supervised training on large corpora of text. Philosophically, this model can be thought of as harking back to the age of specific feature design, albeit through the lens of an oracle trained through exposure to phenomenal volumes of, often noisy and unlabeled, data. Through this decoupling, the task of representation learning, which is the most data hungry, is performed once—with smaller fine-tuning target-specific tasks now requiring little—or sometimes even no—additional training.

Whilst materials discovery can be a somewhat more nuanced task than language generation[14–16], we have seen that techniques and technology developed in the AI realm for language are often transposed and translated to this important task[14,17,18]. In this perspective, we will chart the current state of foundation models—the general term for the newly evolved class of machine learning models of which large language models (LLMs) are a part—for materials discovery, and reflect on the challenges which should be addressed to maximize the impact of this important development to the scientific community.

Foundation models

The class of AI model commonly referred to as a "foundation model"—of which LLMs are a specific incarnation—is defined as a "model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks"[19]. These models typically exist as a base model, which is often generated through unsupervised pre-training on a large amount of unlabeled data. This base model can then be fine-tuned using (often significantly less) labeled data to perform specific tasks. Optionally, this fine-tuned model can also undergo a process known as *alignment*. In this process, the outputs generated by the model are aligned to the preferences of the end user. In language tasks, this might take the form of reducing harmful outputs not aligned with human values (Fig. 1), whereas for a chemical task this might take the form of generating molecular structures which have improved synthesisability, or chemical correctness. Typically this is achieved through conditioning the exploration of the latent space to particular parts of a desired property distribution. A visual description of how the encoder and decoder tasks

[1]IBM Research Europe, Rüschlikon, Switzerland. [2]Xyme, Oxford, UK. [3]IBM Research—TJ Watson Research Center, Yorktown Heights, NY, USA.
✉e-mail: ed@xyme.ai

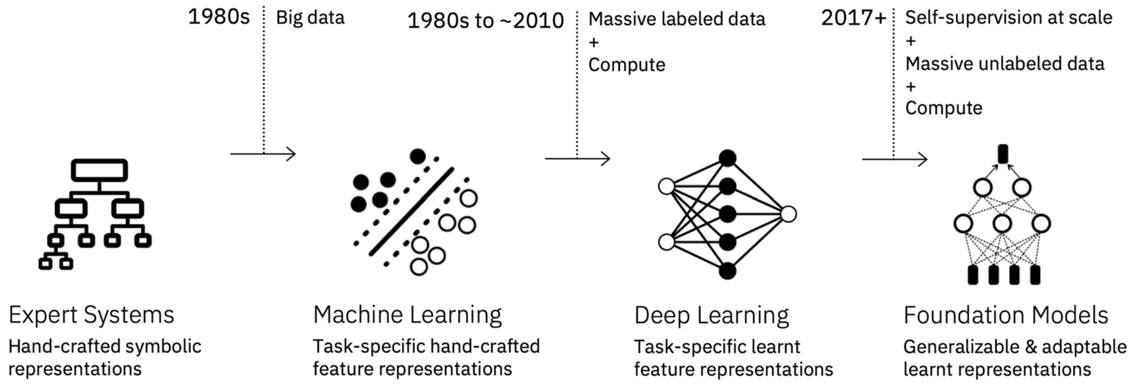
1980s
Big data
1980s to ~2010
Massive labeled data
+
Compute
2017+
Self-supervision at scale
+
Massive unlabeled data
+
Compute
Expert Systems
Hand-crafted symbolic representations
Machine Learning
Task-specific hand-crafted feature representations
Deep Learning
Task-specific learnt feature representations
Foundation Models
Generalizable & adaptable learnt representations

Fig. 1 | Timeline of the development of representations for machine learning. This timeline shows the evolution from hand crafted symbolic representations to today's foundation models through major milestones such as the advent of deep learning.

Fig. 2 | A visual representation of the creation and utilization of a shared latent space, or embedding, via encoder, decoder and predictor models. Molecular representations are transformed into their latent space representation by an encoder model, which can be reversed by a decoder model. Additionally properties can be directly predicted from the encoded representation via the building of an additional property predictor model, which is itself capable of producing a property distribution for the latent space.

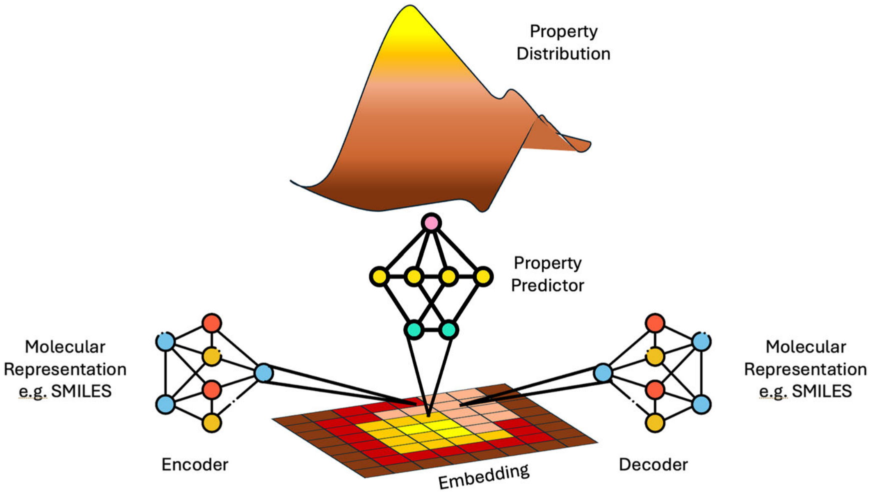
Property Distribution
Property Predictor
Molecular Representation e.g. SMILES
Molecular Representation e.g. SMILES
Encoder
Embedding
Decoder

Table 1 | Examples of tasks in materials discovery which are currently using a foundation model approach

Task	Modalities	Architecture	Example usage
Data extraction	Text, Image	Encoder	Molecular property extraction from patent
Property prediction	Text, Graph, Structural, Spectroscopic,	Encoder, Encoder-Decoder	Downstream prediction of molecular property from molecular graph
Molecular generation	Text, Graph	Decoder	Generate new molecular candidates for a desired task
Synthesis prediction	Text	Encoder-Decoder	Retrosynthetic planning

interact with a latent space, as well as how models trained from that latent space produce a property distribution is shown in Fig. 2, and we also point the interested reader at the following excellent articles which are dedicated to the subject[20–23].

The separation of representation learning from the downstream tasks such as output generation can naturally be crystallized in the model architecture. Whilst the original transformer architecture encompassed both the encoding and decoding tasks, we now frequently see these components as being decoupled models, leading to encoder-only and decoder-only architectures becoming commonplace. Drawing from the success of Bidirectional Encoder Representations from Transformers (BERT)[24], encoder-only models focus solely on understanding and representing input data, generating meaningful representations that can be used for further processing or predictions. Decoder-only models, on the other hand, are designed to generate new outputs by predicting and producing one token at a time based on the given input and previously generated token, making them ideally suited to the task of generating, for example, new chemical entities. Examples of how these types of models could be used in the context of materials discovery are shown in Table 1.

Data extraction

The starting point for successful pretraining and instruction tuning of foundational models is the availability of significant volumes of data, preferably at a high quality. For materials discovery, this principle is even more critical. Materials exhibit intricate dependencies where minute details can significantly influence their properties—a phenomenon known in the cheminformatics community as an "activity cliff". For instance, in the context of high-temperature superconductors like the high-temperature cuprate superconductors, the critical temperature (T_c) can be profoundly affected by subtle variations in hole-doping levels. Models which do not have the richness of information within their training data may miss these effects entirely, potentially leading to non-productive avenues of research inquiry.

Chemical databases provide a wealth of structured information on materials and are therefore a first choice. Indeed resources such as PubChem[21], ZINC[25], and ChEMBL[26] are commonly used to train chemical foundation models[27–29]. However, these sources are often limited in scope and accessibility due to factors such as licensing restrictions (especially for proprietary databases[30]), the relatively small size of datasets, and biased data sourcing. Furthermore, one must ensure the quality and reliability of the extracted data. Source documents often contain noisy, incomplete, or inconsistent information, which can propagate errors into downstream models and analyses. For instance, discrepancies in naming conventions, ambiguous property descriptions, or poor-quality images can hinder accurate extraction and association of materials data. To overcome these limitations, there is an imperative need for robust data-extraction models capable of operating at scale on one of the most common and ubiquitous data-sources, that is, the source documents themselves.

A significant volume of relevant materials information is represented in documents, be that public or proprietary scientific reports, patents or presentations. To extract the relevant materials data from these sources, any AI powered extraction models must efficiently parse and collect the materials information from a variety of habitats. Traditional data-extraction approaches primarily focus on text in the documents[31–33], however, in the realm of materials science, significant information is also embedded in tables, images, and molecular structures. For example, in patent documents, some molecules are selected for their importance and represented by images, while the text can contain any irrelevant structures. Therefore, modern databases aim to extract molecular data not only from text, but from these multiple modalities[34,35]. Additionally, some of the most valuable data arises from the combination of text and images, such as Markush structures in patents, which encapsulate the key patented molecules. Thus, advanced data-extraction models must be adept at handling multimodal data, integrating textual and visual information to construct comprehensive datasets that accurately reflect the complexities of materials science.

Data extraction-based foundation models typically focus on two types of problems—on the one hand identifying the materials themselves and on the other hand identifying and associating described properties with these materials.

For the former, work to date has focused on leveraging the traditional named entity recognition (NER) approaches[36,37] although we note that this is only possible for data encoded within text. Some algorithms have also been developed to identify molecular structures from images in documents, using state-of-the-art computer vision such as Vision Transformers[38,39] and Graph Neural Networks[40]. Recent studies further aim to merge both modalities for extracting general knowledge from chemistry literature[41–43]. For the second type, i.e., property extraction and association, the latest progress in LLMs has allowed such tasks to become much more accurate and leverage schema based extraction[44,45].

While traditional NER and multimodal approaches have shown promise in extracting materials data from diverse document formats, it is important to recognize that multimodal language models need not independently handle all forms of information. Instead, they can effectively integrate with specialized algorithms that act as intermediary tools to process specific types of content. For instance, Plot2Spectra[46] demonstrates how specialized algorithms can extract data points from spectroscopy plots in scientific literature, enabling large-scale analysis of material properties that would otherwise be inaccessible to text-based models. Similarly, DePlot[47] illustrates the utility of modular approaches by converting visual representations such as plots and charts into structured tabular data, which can then be used as input for reasoning by large language models. These examples emphasize that multimodal models can function as orchestrators, leveraging external tools for domain-specific tasks, thereby enhancing the overall efficiency and accuracy of data extraction pipelines in materials science.

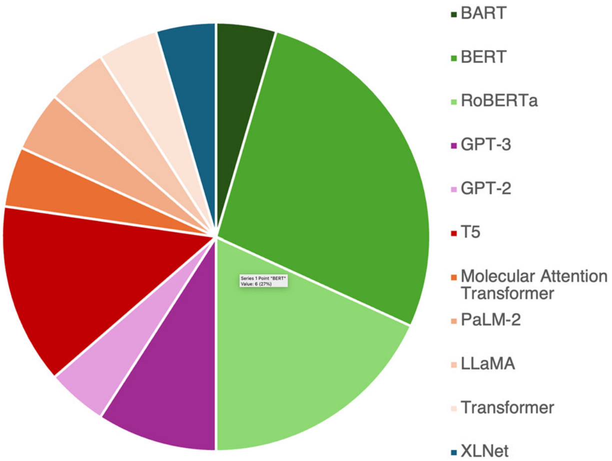

Fig. 3 | Breakdown of common architecture types of models currently reported to perform property prediction. Colors have been determined by the "super type", or originating, architecture. Greens represent models based upon BERT, purples on GPT, oranges refer to architectures based directly on transformers, and blue XLNet. Data based on ref. 155.

Property prediction

The prediction of property from structure is a core component of the value that data-driven approaches can bring to materials discovery. Traditionally, property prediction has either been utilized as a highly approximate initial screen (for example, traditional QSPR methods), or based upon simulation of the fundamental physics of the systems, which can be prohibitively expensive. foundation models offer the opportunity to create powerful predictive capabilities based upon transferrable core components and begin to enable a truly data-driven approach to inverse design.

It is important to note that the current literature is dominated by models which are trained to predict properties from 2D representations of the molecule such as SMILES[48] or SELFIES[49], which can lead to key information such as the 3D conformation of a molecule, being omitted. This is in part due to the significant delta in available datasets for these two respective modalities—with current foundation models being trained on datasets such as ZINC[25] and ChEMBL[26] which both offer datasets ~10^9 molecules—a size not readily available for 3D data. An exception is represented by inorganic solids, such as crystals, where property prediction models usually leverage 3D structures through graph-based or primitive cell feature representations[50,51].

Many of the foundation models that are used for property prediction are encoder-only models based broadly on the BERT[24] architecture[27,52–55], although Fig. 3 shows that other base architectures such as GPT[56–59] are becoming more prevalent. The reuse of both core models, and core architectural components is a strength of the foundation model approach, although there are analogies to be drawn to the limited number of hand-crafted features that ultimately pushed the community into a more data-driven approach.

It is also possible to think of some of the recent class of machine learning based potentials—commonly known as MLIPs (machine learned interatomic potentials)—as foundation models[60], with their core mode of operation being the prediction of energies and forces of a system, based upon a model pre-trained using a large amount of high-quality reference data—typically based on density functional theory (DFT). In much the same way that base representations can be leveraged to build powerful predictors, we are beginning to see pre-trained models such as MEGNET[61], MACE[62], ANI[63] and AIMNET[64,65] being tuned for more specific tasks[66], or for more accurate datasets[67], for which there is a lower available data volume. This is enhanced by efforts such as Optimade[68], which reduces the barrier to bringing together different materials databases, including those built on simulated data.

This offers a fundamentally different route to the use of foundation models for materials property prediction. Whilst many models approach the problem as a direct prediction of a property, this can be hindered by underrepresentation of, for example, rare events. By instead approximating the underlying potential, MLIPs enable traditional simulation techniques at a significantly reduced overhead, and thus the discovery of outcomes which need not be expressed in the original training data.

Of course, for both of these use cases, there is always the problem of understanding transferability and appropriateness of models[69,70]. To this end, we are beginning to see the emergence of techniques to evaluate properties such as model roughness[71,72], which are able to provide quantitative measures which can be related to the likelihood of successful model application and the detection of so-called activity cliffs. We do note, however, that this is still not a solved problem and welcome further research into this area.

Molecular generation

Motivated by the need to overcome the limitations of traditional heuristic and grid-search approaches, AI generative models for material design have gained increasing popularity.

These models are trained to propose novel molecules with desired properties by relying on a variety of molecular structure representations, e.g., text-based SMILES[48] and SELFIES[49] or graph-based approaches[73].

The field of machine learning-based algorithms for designing materials has seen a surge in diversity, with various techniques and applications being proposed. These include VAEs, GANs, GNNs, Transformer-based models, and Diffusion models, each offering unique contributions to the field.

From the end of the last decade, a series of seminal works[74–78] based on textual and graphical representations of molecular entities, have started showing promising application of deep generative models to design materials.

In the way paved by these early efforts, prominent examples have demonstrated how these models can be applied to conditionally design and optimize therapeutics successfully validated in silico or in vitro for a variety of applications: kinase inhibitors[79,80], antivirals[81,82], antimicrobials[83], and disease-specific compounds[84,85].

The usage of such machine learning-based approaches to molecule generation is not limited to the pharmaceutical domain but has soon shown remarkable results in the broader field of material discovery for property-driven design, e.g., sugar/dye molecules via graph generation[86], small molecules, peptides, and polymers generation leveraging language models[87,88], and, semiconductors combining deep learning and DFT[89].

The development of models has significantly reduced the barriers to accessing generative algorithms for material design. This is largely due to the release of open benchmarks and specialized toolkits for generative molecular design, which encompass a wide range of methods, including evolutionary approaches and generative models, such as GuacaMol[90], Moses[91], TDC (Therapeutics Data Common)[92,93], and GT4SD[94].

Despite the ease of access to these technologies, training generative models at the foundation model scale in material science still presents challenges[21]. Indeed, only recently have we seen attempts to train generative foundation models exhibit promising results in a multi-task setting that leverages extensive pretraining across various chemical tasks[95,96].

Foundation models for materials synthesis

The emergence of foundation models in materials science represents a significant opportunity to revolutionize the synthesis of both inorganic and organic materials. The direct application of foundation models in these domains is still in its early stages, with key developments in the synthesis of both inorganic and organic materials thus far being more closely aligned with traditional machine learning approaches rather than foundation models. Nevertheless, compelling indications in the literature suggest that foundation models will play a pivotal role in the future, making it crucial to carefully analyze these trends to anticipate the forms and characteristics that future foundation models may take as they evolve.

Significant advancements in the synthesis of inorganic materials have been achieved through the application of machine learning and data-driven approaches. Recent studies have highlighted the potential of utilizing novel data sources, such as natural language text from scientific literature[97], to predict synthesis protocols for inorganic materials[98]. For example, word embeddings and variational autoencoders have been used to generate synthesis strategies for perovskite materials[99]. Additionally, natural language processing techniques have been instrumental in designing novel synthesis heuristics derived from text-mined literature data. When combined with active learning, these heuristics optimize the synthesis of novel inorganic materials in powder form[100]. These efforts have led to the development of extraction and analysis pipelines for synthesis information from scientific publications, laying the groundwork for comprehensive, high-quality datasets for inorganic materials[101] and single-atom heterogeneous catalysis[102]. While natural language processing techniques have demonstrated significant potential, it is critical to observe that literature itself often presents inherent limitations that can affect the applicability of extracted synthesis recipes. For example, the quality and consistency of reporting in the scientific literature, as noted by David et al.[103], can pose significant barriers. These challenges include incomplete data, inconsistent terminology, and insufficient experimental details, which can limit the reliability and generalizability of text-derived synthesis strategies. Addressing these limitations will require improved curation of datasets and the development of more robust processing frameworks to ensure accurate and actionable predictions. Supervised machine learning models have been developed to classify and predict suitable synthesis routes and conditions, such as calcination and sintering temperatures, based on target and precursor materials, outperforming traditional heuristics[104]. Reinforcement learning has also been successfully applied to predict optimal synthesis schedules, including time-sequenced reaction conditions for the synthesis of semiconducting monolayer MoS_2 using chemical vapor deposition[75]. Recent advancements have leveraged high-throughput thermochemical data and classical nucleation theory, identifying favorable reactions based on catalytic nucleation barriers and selectivity metrics, such as in Aykol et al.[105]. This method has been validated on well-known compounds such as $LiCoO_2$, $BaTiO_3$, and $YBa_2Cu_3O_7$, showcasing its applicability in identifying both established and unconventional synthetic routes. The integration of such frameworks into foundation models could further enhance their ability to predict and optimize complex synthesis processes by providing structured insights into thermodynamic and kinetic factors, opening up exciting possibilities for leveraging foundation models in both the analysis and prediction of time-series data[106], as well as their application to chemical synthesis. The progress made in machine learning for inorganic synthesis underscores the immense potential of foundation models, particularly in expanding the capabilities of LLMs. Recent studies[107] have demonstrated the effectiveness of LLMs in predicting the synthesizability of inorganic compounds and selecting suitable precursors. In the long term, foundation models are likely to achieve performance levels comparable to specialized ML models, but with significantly reduced development time and costs. This suggests that the future of material synthesis may increasingly be driven by foundation models tailored to multiple tasks within this domain, making them a powerful tool for advancing the field.

The application of foundation models in the synthesis of organic materials follows a similar trajectory. In the realm of organic synthesis, these models hold immense potential for transforming synthetic pathways and optimizing reaction conditions. Early studies have shown that deep learning models can effectively predict reaction outcomes and retrosynthetic pathways, which are critical to the advancement of organic synthesis. A notable example is the Molecular Transformer model[108], the first language model in chemical synthesis, which achieved state-of-the-art accuracy in predicting reaction products by treating reaction prediction as a machine translation problem. Recent advancements in prompt-based inference[109] have extended the capabilities of these models, enabling chemists to steer retrosynthetic predictions, thereby providing more diverse and creative disconnection strategies while overcoming biases in training data. The introduction of domain-specific LLMs[110,111] underscores the importance of fine-tuning with specialized data and integrating advanced chemistry tools[112], significantly enhancing the predictive capabilities for synthetic tasks in organic chemistry. The introduction of the first foundation model[95] capable of addressing multiple tasks across both chemical[17,108,113–115] and natural language domains[116] marks a breakthrough, demonstrating that sharing information across these domains can enhance model performance, particularly in cross-domain tasks. Furthermore, the implementation LLMs in organic chemistry, exemplified by recent advancements in training a foundational large-scale model with 15 billion parameters for retrosynthesis prediction and generative chemistry[117], highlights the potential of foundation models to extend beyond predictive capabilities, and eventually guide experimental efforts in the laboratory. These advancements underscore the potential for further refinement of foundation models, which could lead to enhanced predictive accuracy and, ultimately, more efficient synthesis of complex organic molecules.

The future application of foundation models in the synthesis of both inorganic and organic materials is particularly exciting when considering their ability to learn from diverse data modalities, including spectroscopic data[118], crystallography[119], and atomistic simulations[60]. This capability allows for a more holistic understanding of material behavior, facilitating the design of novel compounds with desired properties. The development of multimodal foundation models, which can simultaneously process and learn from these varied data sources, will represent a transformative approach in the synthesis of both organic and inorganic materials. By capturing complex interactions across different modalities, these models will not only improve predictive accuracy but also enable the generation of new hypotheses for experimental validation. This could lead to the discovery of materials with unprecedented properties, advancing fields such as catalysis, drug design, and energy storage. Moreover, their ability to generalize across different types of data positions them as powerful tools for tackling the intricate challenges in materials science, from understanding reaction mechanisms to optimizing synthesis pathways, all within a unique model.

Challenges and future look

Tackling the data challenge with multi-modal models

In the wake of the rising popularity of LLMs' applications beyond natural language and the introduction of increasingly powerful foundation models, there has been a surge in interest in multimodal approaches. This has led to the development of vision-language models, such as Flamingo[120], LLaVA[121], and Idefics[122,123] which combine different data types to enhance the perception horizon of foundation models. Alongside modeling approaches, multimodal datasets play a crucial role in the success of such models.

Multiple efforts in building web-crawled datasets combining visuals and textual data have become central in machine learning research, culminating with extensive benchmarks promoting advances in multimodal modeling research, e.g., the Cauldron[124] or MMMU[125]. Besides generic datasets for vision-language modeling, there has been a growing interest in compiling data resources to push the boundaries of foundation model perception. Specifically, in the space of egocentric perception, two notable examples are represented by Ego4D[126] and Ego-Exo4D[127], which provide rich, diverse, and annotated data for research in human activity recognition and exploration.

Despite the growing popularity of multimodal approaches, their application in material discovery remains limited. This limited application is primarily due to the lack of large-scale, high-quality datasets that cover a wide range of materials and their properties. Nevertheless, in recent years, inspired by the successful application of multi-task prompted training and instruction tuning[128], models like Text-chem T5[95] and Regression Transformer (RT)[87,88] have been successfully adopted to solve multimodal tasks to accelerate the design of novel materials exhibiting superior performance to specialized single-task models.

These promising attempts at developing multi-purpose models for materials motivate further and deeper experimentation by combining various data sources in model training. Especially considering the wealth of high-quality data that can be generated by simulations, using the latter to fill gaps in real-world data represents a strategy to build more robust and accurate models.

Capturing new types of data for new types of models

The growing use of generative models for synthetic data generation and data augmentation in the domain of materials discovery remains grounded in data from experimental observations, which is indispensable to support the discovery of novel materials[14,16]. However, fundamental concerns regarding the reproducibility of experimental findings prevail[129], and challenges with respect to effective dissemination of high-quality and findable, accessible, interoperable and reusable (FAIR)[130] experimental data cut across scientific disciplines. The lack of reproducibility, in particular, can usually be traced back to experimental data and metadata that is flawed or missing[131,132]. Additional key challenges are the disparity of data formats and schema that must be addressed in order to achieve data interoperability and standardization[133], as well as the diversity of tools used to help digitize and organize experimental data[134,135].

While the consolidation and organization of already digitized data may be facilitated by the use of LLMs[136], the emergence of multi-modal foundation models capable of processing different data modalities at the same time offers new perspectives for data capture and experimental documentation. For example, the shift from convolutional neural networks (CNNs) to transformer-based foundation models for video and action recognition[10] has spawned a series of foundational vision-language models that exhibit proficiency in diverse video recognition tasks[137–141]. Such models, used to describe and document a wide range of real-world actions, have demonstrated significant gains in performance as exemplified by the increase in top-1 accuracy for action recognition on the Kinetics-400 dataset[142,143] from 73.9% in 2016 to 93.6% in 2024[144].

So far, the training data for large vision-language models has mostly been restricted to common human activities[145], but applications have also been studied in specialized domains such as endoscopic surgical procedures to provide information on surgical gestures, generate feedback for surgeons, and help study the relationship between intraoperative factors and postoperative outcomes[146]. The study and adaptation of multi-modal foundation models to the transcription of specialized procedures has the potential to alleviate data and metadata capture at the roots of the experimentation process, by automatically converting raw data streams of observations and sensor data to a reproducible transcript in the desired target format using multi-modal generative models. Such data capture could be realized with minimal burden on the experimentalist and implemented to generate documentation for any digital system of record, such as electronic laboratory notebooks (ELNs). This type of novel data capture stands to benefit both experimentalists who generate the data as well as data scientists and theoreticians who rely on the data to train and validate scientific models. While this concept is still nascent and lacking open datasets and benchmarks, the automatic transcription of laboratory procedures in real-time using multi-modal foundation models has already been demonstrated in principle[147] and may help standardize the process of documenting manual research, link procedural details with outcomes, facilitate science education and information sharing, and supply more consistent and reproducible data for the next generation of foundation models for materials discovery.

Exploiting the science of approximation through multi-fidelity models

As both our model architectures, and the computational infrastructure on which they run, are able to ingest extreme volumes of data, the risk of data collection biases being transferred into the models increases. Whilst we can observe this to a degree in today's language models—for example through differential capability in English language models compared to other, less common, languages—we believe that this is a particular risk in materials discovery. We hold this view as the very task of discovery requires stepping into unknown chemistries, which may not be well represented in large volume datasets, which can be dominated by a particular part of chemical space (e.g., organic chemistry) or domain (e.g., drug discovery). We view multi-fidelity models as having a significant role to play in mitigating this risk.

A multi-fidelity model is one which can take as input data collected in several different ways, each with their own acquisition cost and accuracy. An example of this might be collecting molecular quantum-chemical data using a range of quantum mechanical approaches (e.g., semi-empirical methods, DFT, MP2). By being able to use data from a variety of sources, in a similar manner to which multi-modal models can draw data from a variety of modalities, multi-fidelity models are able to mitigate data sparsity concerns.

Multi-fidelity methods are a reasonably recent addition to the machine learning for materials discovery toolkit, including recent advances in their use for optimization[148], and property prediction of both molecules[149–151] and materials[152–154], and we believe that their extension to the foundation models arena will continue to drive progress.

Conclusion

Foundation models are already showing promise in tackling some of the hardest problems in materials discovery. In this perspective we outline some of the fundamentals of this new area of study, and their applicability to the task of materials discovery. We highlight some areas in which foundation models are already being used to significant impact—namely property prediction, retrosynthesis, and molecular generation—and also look to the future to outline areas which we believe are key to continuing to unlock value. These areas hinge on exploiting the natural multi-modality and multi-fidelity characteristics of materials data through increasingly powerful and elegant modeling approaches. We believe that building from the current state of the art into these areas will unlock a significant value stream, potentially enabling substantial acceleration of materials discovery by leveraging the ever-growing troves of data produced by the research community.

Received: 18 September 2024; Accepted: 25 January 2025;
Published online: 06 March 2025

References

1. Zhong, G., Wang, L.-N., Ling, X. & Dong, J. An overview on data representation learning: from traditional feature learning to recent deep learning. *J. Finance Data Sci* **2**, 265–278 (2016).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
3. Ma, J., Jiang, X., Fan, A., Jiang, J. & Yan, J. Image matching from handcrafted to deep features: a survey. *Int. J. Comput. Vis.* **129**, 23–79 (2021).
4. Snyder, S. H. et al. The Goldilocks paradigm: comparing classical machine learning, large language models, and few-shot learning for drug discovery applications. *Commun. Chem.* **7**, 1–11 (2024).
5. Pandey, M. et al. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **4**, 211–221 (2022).
6. Wang, Y. E., Wei, G.-Y. & Brooks, D. Benchmarking TPU, GPU, and CPU platforms for deep learning. Preprint at https://doi.org/10.48550/arXiv.1907.10701 (2019).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems*, Vol. 25 (Curran Associates, Inc., 2012).
8. LeCun, Y. & Cortes, C. MNIST handwritten digit database. (2010).
9. Vaswani, A. et al. Attention is all you need. in *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017).
10. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *OpenAI blog* (2018).
11. Radford, A. et al. Language models are unsupervised multitask learners. Preprint at Semantic Scholar https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (2018).
12. Brown, T. et al. "Language models are few-shot learners." *Adv. neural inf. process. syst.* **33**, 1877–1901 (2020).
13. OpenAI et al. GPT-4 technical report. *Preprint* https://doi.org/10.48550/arXiv.2303.08774 (2024).
14. Pyzer-Knapp, E. O. et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput. Mater.* **8**, 1–9 (2022).
15. Hautier, G. Finding the needle in the haystack: materials discovery and design through computational ab initio high-throughput screening. *Comput. Mater. Sci.* **163**, 108–116 (2019).
16. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
17. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
18. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
19. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at http://arxiv.org/abs/2108.07258 (2022).
20. Ramos, M. C., Collison, C. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci*. https://doi.org/10.1039/D4SC03921A (2024).
21. Takeda, S., Kishimoto, A., Hamada, L., Nakano, D. & Smith, J. R. Foundation model for material science. *Proc. AAAI Conf. Artif. Intell.* **37**, 15376–15383 (2023).
22. Deng, L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Trans. Signal Inf. Process.* **3**, e2 (2014).
23. Ivanenkov, Y. et al. The Hitchhiker's guide to deep learning driven generative chemistry. *ACS Med. Chem. Lett.* **14**, 901–915 (2023).
24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* vol. 1 (long and short papers), pp. 4171–4186 (2019).
25. Irwin, J. J. & Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
26. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100–D1107 (2012).
27. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at https://doi.org/10.48550/arXiv.2010.09885 (2020).
28. Ross, J. et al. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **4**, 1256–1264 (2022).
29. Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **62**, 2064–2076 (2022).
30. Lawson, A. J., Swienty-Busch, J., Géoui, T. & Evans, D. The making of reaxys—towards unobstructed access to relevant chemistry information. in *The Future of the History of Chemical Information* (eds McEwen, L. R. & Buntrock, R. E.) Vol. 1164, 127–148 (American Chemical Society, 2014).
31. Akhondi, S. A. et al. Automatic identification of relevant chemical compounds from patents. *Database* **2019**, baz001 (2019).
32. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).

33. Zhang, Y. et al. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database* **2016**, baw049 (2016).
34. Morin, L., Weber, V., Meijer, G. I., Yu, F. & Staar, P. W. J. PatCID: an open-access dataset of chemical structures in patent documents. *Nat. Commun.* **15**, 1–11 (2024).
35. Papadatos, G. et al. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res* **44**, D1220–D1228 (2016).
36. Weston, L. et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **59**, 3692–3702 (2019).
37. Gupta, T., Zaki, M. & Krishnan, N. M. A. & Mausam MatSciBERT: a materials domain language model for text mining and information extraction. *npj Comput. Mater.* **8**, 1–11 (2022).
38. Rajan, K., Brinkhaus, H. O., Agea, M. I., Zielesny, A. & Steinbeck, C. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nat. Commun.* **14**, 5045 (2023).
39. Qian, Y. et al. MolScribe: robust molecular structure recognition with image-to-graph generation. *J. Chem. Inf. Model.* **63**, 1925–1934 (2023).
40. Morin, L. et al. MolGrapher: Graph-based Visual Recognition of Chemical Structures. in 2023 IEEE/CVF InternationalConference on Computer Vision (ICCV) 19495–19504 https://doi.org/10.1109/ICCV51070.2023.01791 (IEEE, Paris, France, 2023).
41. Fan, V. et al. OpenChemIE: an information extraction toolkit for chemistry literature. *J. Chem. Inf. Model.* **64**, 5521–5534 (2024).
42. Cai, H. et al. Uni-SMART: universal science multimodal analysis and research transformer. Preprint at https://doi.org/10.48550/arXiv.2403.10301 (2024).
43. Wang, J. et al. Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space. *Brief. Bioinforma.* **23**, bbac461 (2022).
44. Shetty, P. et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *NPJ Comput. Mater.* **9**, 52 (2023).
45. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
46. Jiang, W. et al. Plot2Spectra: an automatic spectra extraction tool. *Digital Discov* **1**, 719–731 (2022).
47. Liu, F. et al. DePlot: one-shot visual language reasoning by plot-to-table translation. in *Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N) 10381–10399 (Association for Computational Linguistics, 2023).
48. Weininger, D. SMILES. a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
49. Krenn, M. et al. SELFIES and the future of molecular string representations. *Patterns* **3**, 100588 (2022).
50. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
51. Siriwardane, E. M. D., Zhao, Y., Perera, I. & Hu, J. Generative design of stable semiconductor materials using deep learning and density functional theory. *npj Comput. Mater.* **8**, 1–12 (2022).
52. Ock, J., Guntuboina, C. & Barati Farimani, A. Catalyst energy prediction with CatBERTa: unveiling feature exploration strategies through large language models. *ACS Catal* **13**, 16032–16044 (2023).
53. Yüksel, A., Erva, U., Atabey, Ü. & Tunca, D. SELFormer: molecular representation learning via SELFIES language models. *Mach. Learn. - Sci. Tech. 4* **no. 2**, 025035 (2023).
54. Yu, J. et al. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discov.* **2**, 409–421 (2023).
55. Li, J. & Jiang, X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. *Proc. Int. Wirel. Commun. Mob. Comput. Conf.* **2021**, 7181815 (2021).
56. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
57. Winter, B., Winter, C., Schilling, J. & Bardow, A. A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing. *Digital Discov* **1**, 859–869 (2022).
58. Adilov, S. Generative pre-training from molecules. *ChemRxiv* https://doi.org/10.26434/chemrxiv-2021-5fwjd (2021).
59. Liu, Z. et al. MolXPT: wrapping molecules with text for generative pre-training. in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 2:* Short Papers*)* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) 1606–1616 (Association for Computational Linguistics, 2023).
60. Batatia, I. et al. A foundation model for atomistic materials chemistry. Preprint at http://arxiv.org/abs/2401.00096 (2024).
61. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
62. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
63. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
64. Anstine, D., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. Preprint at https://doi.org/10.26434/chemrxiv-2023-296ch-v2 (2024).
65. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).
66. Focassio, B., Freitas, L. P. M. & Schleder, G. R. Performance assessment of universal machine learning interatomic potentials: challenges and directions for materials' surfaces. Preprint at https://doi.org/10.48550/arXiv.2403.04217 (2024).
67. Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
68. Andersen, C. W. et al. OPTIMADE, an API for exchanging materials data. *Sci. Data* **8**, 217 (2021).
69. Speckhard, D. et al. How big is big data? *Faraday Discuss*. https://doi.org/10.1039/D4FD00102H (2024).
70. Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput. Mater.* **9**, 1–9 (2023).
71. Dicks, L., Graff, D., Jordan, K., Coley, C. & Pyzer-Knapp, E. A physics-inspired approach to the understanding of molecular representations and models. *Mol. Syst. Des. Eng.* **9**, 449–455 (2024).
72. Graff, E. et al. Evaluating the roughness of structure–property relationships using pretrained molecular representations. *Digital Discov* **2**, 1452–1460 (2023).
73. Cayley, O. the mathematical theory of isomers. *Philos. Mag.* **47**, 444–446 (1874).
74. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).

75. Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
76. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. in *International Conference on Machine Learning*, PMLR 2323–PMLR 2332 (2018).
77. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Adv. Neural Inf. Process. Syst.* **31**, 6410–6421 (2018).
78. Prykhodko, O. et al. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminform.* **11**, 1–13 (2019).
79. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
80. Born, J. et al. Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3D effects in a 1D model. *J. Chem. Inform.* **62**, 240–257 (2022).
81. Chenthamarakshan, V. et al. Cogmol: target-specific and selective drug design for covid-19 using deep generative models. *Adv. Neural Inf. Process. Syst.* **33**, 4320–4332 (2020).
82. Born, J. et al. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-COV-2. *Mach. Learn. Sci. Technol.* **2**, 025024 (2021).
83. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
84. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **11**, 1–10 (2020).
85. Born, J. et al. PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **24**, 102269 (2021).
86. Takeda, S. et al. Molecular inverse-design platform for material industries. in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2961–2969 (2020).
87. Born, J. & Manica, M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mach. Intell.* **5**, 432–444 (2023).
88. Park, N. H. et al. Artificial intelligence driven design of catalysts and materials for ring opening polymerization using a domain-specific language. *Nat. Commun.* **14**, 3686 (2023).
89. Siriwardane, E. M. D., Zhao, Y., Perera, I. & Hu, J. Generative design of stable semiconductor materials using deep learning and density functional theory. *npj Comput. Mater.* **8**, 164 (2022).
90. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
91. Polykovskiy, D. et al. Molecular sets (Moses): a benchmarking platform for molecular generation models. *Front. Pharm.* **11**, 1931 (2020).
92. Huang, K. et al. Therapeutics data commons: machine learning datasets and tasks for drug discovery and development. *Adv. Neural Inf. Process. Syst*. **35** (2021).
93. Huang, K. et al. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **11**, 191–200 (2022).
94. Manica, M. et al. Accelerating material design with the generative toolkit for scientific discovery. *npj Comput. Mater.* **9**, 1–6 (2023).
95. Christofidellis, D. et al. Unifying molecular and textual representations via multi-task language modelling. in *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202, 6140–6157 (JMLR.org, 2023).
96. Chang, J. & Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nat. Commun.* **15**, 2323 (2024).
97. Wang, Z. et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Sci. Data* **9**, 231 (2022).
98. He, T. et al. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Sci. Adv.* **9**, eadg8180 (2023).
99. Kim, E. et al. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
100. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
101. Wang, Z. et al. ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discov* **1**, 313–324 (2022).
102. Suvarna, M., Vaucher, A. C., Mitchell, S., Laino, T. & Pérez-Ramírez, J. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nat. Commun.* **14**, 7964 (2023).
103. Sun, W. & David, N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discuss*. https://doi.org/10.1039/D4FD00112E (2024).
104. Karpovich, C., Pan, E., Jensen, Z. & Olivetti, E. Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction. *Chem. Mater.* **35**, 1062–1079 (2023).
105. Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
106. Liang, Y. et al. Foundation models for time series analysis: a tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 6555–6565 (2024).
107. Chen, Z. et al. MatChat: a large language model and application service platform for materials science. *Chin. Phys. B* **32**, 118104 (2023).
108. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
109. Thakkar, A. et al. Unbiasing retrosynthesis language models with disconnection prompts. *ACS Cent. Sci.* **9**, 1488–1498 (2023).
110. Vaucher, A. C. et al. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **12**, 2573 (2021).
111. Zhang, C. et al. SynAsk: unleashing the power of large language models in organic synthesis. Preprint at https://doi.org/10.48550/arXiv.2406.04593 (2024).
112. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
113. Cretu, M. T. et al. Standardizing chemical compounds with language models. *Mach. Learn. Sci. Technol.* **4**, 035014 (2023).
114. Zipoli, F., Baldassari, C., Manica, M., Born, J. & Laino, T. Growing strings in a chemical reaction space for searching retrosynthesis pathways. *npj Comput. Mater.* **10**, 1–14 (2024).
115. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 015016 (2021).
116. Vaucher, A. C. et al. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).
117. Yang, Y. et al. BatGPT-Chem: a foundation large model for chemical engineering. Preprint at https://doi.org/10.26434/chemrxiv-2024-1p4xt (2024).

118. Alberts, M., Laino, T. & Vaucher, A. C. Leveraging infrared spectroscopy for automated structure elucidation. Preprint at https://doi.org/10.26434/chemrxiv-2023-5v27f (2023).
119. Ozawa, K., Suzuki, T., Tonogai, S. & Itakura, T. Graph-text contrastive learning of inorganic crystal structure toward a foundation model of inorganic materials. Preprint at https://doi.org/10.26434/chemrxiv-2024-mpl8l (2024).
120. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
121. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. in *Thirty-seventh Conference on Neural Information Processing Systems* (2023).
122. Laurençon, H. et al. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. Preprint at https://doi.org/10.48550/arXiv.2306.16527 (2023).
123. Liu, H., Li, C., Li, Y. & Lee, Y. J. Improved baselines with visual instruction tuning. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 26296–26306 (2024).
124. Laurençon, H., Tronchon, L., Cord, M. & Sanh, V. What matters when building vision-language models? *Adv. Neural Inf. Process. Syst.* **37**, 87874–87907 (2025).
125. Yue, X. et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9556–9567 (2024).
126. Ego4D: around the world in 3,000 h of egocentric video. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 18995–19012 (2022).
127. Grauman, K. et al. Ego-Exo4D: understanding skilled human activity from first- and third-person perspectives. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19383–19400 (2024).
128. Sanh, V. et al. Multitask prompted training enables zero-shot task generalization. in *International Conference on Learning Representations* (2022).
129. Baker, M. 1500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
130. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
131. Gonçalves, R. S. & Musen, M. A. The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data* **6**, 190021 (2019).
132. Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24 (2020).
133. Jablonka, K. M., Patiny, L. & Smit, B. Making the collective knowledge of chemistry open and machine actionable. *Nat. Chem.* **14**, 365–376 (2022).
134. Higgins, S. G., Nogiwa-Valdez, A. A. & Stevens, M. M. Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nat. Protoc.* **17**, 179–189 (2022).
135. Kanza, S. et al. Digital research environments: a requirements analysis. *Digital Discov.* **2**, 602–617 (2023).
136. Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discov.* **2**, 1233–1250 (2023).
137. Ni, B. et al. Expanding language-image pretrained models for general video recognition. Preprint at https://doi.org/10.48550/arXiv.2208.02816 (2022).
138. Lin, B. et al. Video-LLaVA: learning united visual representation by alignment before projection. Preprint at https://doi.org/10.48550/arXiv.2311.10122 (2023).
139. Zhao, L. et al. VideoPrism: a foundational visual encoder for video understanding. *Preprint* https://doi.org/10.48550/arXiv.2402.13217 (2024).
140. Chen, Z. et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.* **67**, 220101 (2024).
141. Wang, Y. et al. InternVideo2: scaling foundation models for multimodal video understanding. Preprint at https://doi.org/10.48550/arXiv.2403.15377 (2024).
142. Kay, W. et al. The kinetics human action video dataset. Preprint at https://doi.org/10.48550/arXiv.1705.06950 (2017).
143. Sasaki, R., Fujinami, M. & Nakai, H. Application of object detection and action recognition toward automated recognition of chemical experiments. *Digital Discov.* **3**, 2458–2464 (2024).
144. Action Classification on Kinetics-400. https://paperswithcode.com/sota/action-classification-on-kinetics-400
145. Gupta, N. et al. Human activity recognition in artificial intelligence framework: a narrative review. *Artif. Intell. Rev.* **55**, 4755–4808 (2022).
146. Kiyasseh, D. et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng.* **7**, 780–796 (2023).
147. Thakkar, A. et al. Using foundation models to promote digitization and reproducibility in scientific experimentation. *in NeurIPS 2023 AI for Science Workshop* (2023).
148. Fare, C., Fenner, P., Benatan, M., Varsi, A. & Pyzer-Knapp, E. O. A multi-fidelity machine learning approach to high throughput materials screening. *npj Comput. Mater.* **8**, 1–9 (2022).
149. Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D. & Lió, P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat. Commun.* **15**, 1517 (2024).
150. Greenman, K. P., Green, W. H. & Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. Sci.* **13**, 1152–1162 (2022).
151. Yang, C.-H. et al. Multi-fidelity machine learning models for structure–property mapping of organic electronics. *Comput. Mater. Sci.* **213**, 111599 (2022).
152. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput Sci.* **1**, 46–53 (2021).
153. Patra, A. et al. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **172**, 109286 (2020).
154. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
155. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* (2025).

Acknowledgements

The authors would like to acknowledge their IBM colleagues for their continued contributions to this exciting area of research.

Author contributions

All authors contributed to the conception, structuring, and writing of this perspective.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Edward O. Pyzer-Knapp.

Reprints and permissions information is available at http://www.nature.com/reprints