Advancing materials science through next-generation machine learning

Rohit Unni, Mingyuan Zhou, Peter Wiecha, Yuebing Zheng

Advancing materials science through next-generation machine learning

Rohit Unni,[1,2] Mingyuan Zhou,[3,4] Peter R. Wiecha[5], and Yuebing Zheng[1,2,*]

[1]*Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, Texas 78712, USA*

[2]*Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78712, USA*

[3]*Department of Statistics and Data Science, The University of Texas at Austin, Austin, Texas 78712, USA*

[4]*McCombs School of Business, The University of Texas at Austin, Austin, Texas 78712, USA*

[5]*LAAS, Université de Toulouse, CNRS, Toulouse, France*

[*]Corresponding Author: zheng@austin.utexas.edu

Abstract: For over a decade, machine learning (ML) models have been making strides in computer vision and natural language processing (NLP), demonstrating high proficiency in specialized tasks. The emergence of large-scale language and generative image models, such as ChatGPT and Stable Diffusion, has significantly broadened the accessibility and application scope of these technologies. Traditional predictive models are typically constrained to mapping input data to numerical values or predefined categories, limiting their usefulness beyond their designated tasks. In contrast, contemporary models employ representation learning and generative modeling, enabling them to extract and encode key insights from a wide variety of data sources and decode them to create novel responses for desired goals. They can interpret queries phrased in natural language to deduce the intended output. In parallel, the application of ML techniques in materials science has advanced considerably, particularly in areas like inverse design, material prediction, and atomic modeling. Despite these advancements, the current models are overly specialized, hindering their potential to supplant established industrial processes. Materials science, therefore, necessitates the creation of a comprehensive, versatile model capable of interpreting human-readable inputs, intuiting a wide range of possible search directions, and delivering precise solutions. To realize such a model, the field must adopt cutting-edge representation, generative, and foundation model techniques tailored to materials science. A pivotal component in this endeavor is the establishment of an extensive, centralized dataset encompassing a broad spectrum of research topics. This dataset could be assembled by crowdsourcing global research contributions and developing models to extract data from existing literature and represent them in a homogenous format. A massive dataset can be used to train a central model that learns the underlying physics of the target areas, which can then be connected to a variety of specialized downstream tasks. Ultimately, the envisioned model would empower users to intuitively pose queries for a wide array

of desired outcomes. It would facilitate the search for existing data that closely matches the sought-after solutions and leverage its understanding of physics and material-behavior relationships to innovate new solutions when pre-existing ones fall short.

Keywords: *deep learning, neural networks, materials science, large language models*

1. Introduction

Machine learning (ML) has experienced remarkable advancements in recent times, solidifying its role as a multifaceted instrument utilized in various fields. It harnesses the power of large datasets and algorithms that are trained to identify complex and non-obvious patterns, enabling the development of sophisticated predictive models. These models often outperform human insight and conventional analytical techniques [1, 2]. The widespread integration of these models has led to significant innovations in areas such as image recognition[3-5], medical research [6-8], natural language processing[9, 10], and robotics[11-13].

Nevertheless, these advanced predictive models are not without their limitations, primarily due to their narrow focus. For example, without an additional computational tuning effort, a model adept at identifying cats will be incapable of processing different types of images[14]. Such models often depend on deep neural networks (DNNs), which process input data through a cascade of intricate, nonlinear operations, tuning their parameters through training, as illustrated in Figure 1. However, even minor alterations in the input data, such as changes in dimensionality or requests for divergent outputs, can compromise the effectiveness of these predictive models. Enhancing a model's capabilities typically requires additional data and extended training periods, yet this often results in only slight usability enhancements[15, 16]. Moreover, the data demands for accurately training DNN-based predictive models are substantial, frequently necessitating tens of thousands to millions of labeled training examples, which makes the training both resource-intensive and time-consuming[2]. As a result of these factors—the high costs associated with data and the models' limited adaptability—predictive ML algorithms have historically been powerful but highly specialized tools.

The plethora of internal parameters that DNNs learn during training can reach into the billions, or even trillions for the largest language models, adding layers of complexity and opacity to the model. Consequently, the decision-making process of these models often resembles a black box, obscuring their predictive reasoning. This opacity can render DNN-based models particularly

inaccessible to individuals without a background in programming or data analysis. This also implies a trust problem, since the predictions of such models can never be totally relied on, without additional verification. In contrast, alternative predictive models like those based on decision trees typically offer greater transparency in their predictions [17]. Efforts are underway to imbue more complex models with this level of interpretability [18]. Data collection and labeling represent significant hurdles in the training of predictive models. Despite the abundance of both labeled and unlabeled data, the challenge lies in harnessing this data effectively, and homogenizing its representation format. This is a prerequisite for being able to train a model. Furthermore, it would be ideal to enhance its adaptability and ease of use. Thus, developing efficient strategies to glean valuable insights from the available data and improve the models' versatility and user-friendliness stands as a pivotal goal for the ongoing evolution of artificial intelligence (AI).

In the past five years, the field of ML has witnessed the emergence of novel applications, particularly in the representation and generation of text and image, heralding promising advancements towards achieving these goals. These innovations have significantly lowered the barriers to entry, enabling wider public access to the capabilities of AI, with notable strides in text and image-based models. Two areas where breakthroughs have been particularly impactful are representation learning and generative modeling. Representation learning shifts the focus from directly categorizing input data to learning a lower-dimensional representation of its essential features. This representation can then be applied to a broader range of downstream tasks [19]. This approach is versatile, accommodating both supervised and unsupervised learning paradigms. For instance, contrastive representation learning trains models to discern distinctions between unlabeled data samples [20]. Generative models, on the other hand, are designed to understand the underlying statistical distributions of data, enabling them to generate new, convincing samples, be it images or text, that closely mimic the characteristics of the data they were trained on [21]. These advancements are not just technical marvels but also pivotal steps towards making AI more versatile and accessible.

Models such as Stable Diffusion and DALL-E 2 are prime examples of generative models, utilizing diffusion-based mechanisms to generate novel images from textual prompts [22, 23]. Similarly, ChatGPT and GPT-4[24], which operate on the generative pre-trained transformer architecture, facilitate a broad spectrum of user interactions through text inputs [25]. These interactions range from generating creative content, browsing the internet and summarizing key

information, to engaging in basic conversations [26]. All these models utilize representation learning to encode the text inputs into a latent dimension that contains all the most important info. This latent representation is then fed into the generative parts of the model where the final output is then constructed. This latent representation translates an input into an encoding, and the generative architectures decode that encoding into the final generated output. It is the parallel development of these two strategies, paired with massive amounts of homogenously formatted data, that have enabled the powerful performance of recent models.

What makes these models particularly user-friendly is their ability to understand and process requests phrased in everyday language, delivering results that align closely with the user's intentions. This accessibility opens the door to users without specialized machine learning knowledge, democratizing the use of AI. Provided they can form a basic understanding of a problem from a huge amount of data, these generative models also develop efficient learning abilities. An initial pre-training on a large dataset to establish an underlying base model of the target domain is first needed. After this initial pre-training, however, large models require only a limited number of examples to learn to recognize new patterns[27]. Given a database expansive enough to cover a wide array of subjects, these models can attain a more general-purpose utility. The availability of massive amounts of well-formatted data, enabled by the internet, is one of the key factors that has allowed the latest models to advance in size and obtain remarkable results [28] . Users can interact with these systems using natural language, and the models can deduce their intentions and respond appropriately. This intuitive interface has sparked a surge in AI's popularity across various sectors, with its transformative effects already becoming evident in numerous industries.

2. Breakthroughs of deep generative models

ML’s application has been extending beyond its notable success in computer vision and natural language processing (NLP), making considerable inroads into various scientific disciplines. In the realm of materials science, the past decade has witnessed the use of ML algorithms to tackle a diverse array of challenges[29-31]. These applications range from designing energy materials [32, 33] and metamaterials [34-36], for phase and component prediction[37-40], and material behavior prediction[41-44].

Despite these advances, the current research landscape primarily relies on conventional predictive models. Models dedicated to material design and behavior prediction are often constrained to specific paradigms, anchored by assumptions about design factors or environmental conditions. For instance, metamaterials' inverse design models [45-47] have shown proficiency in predicting designs that replicate arbitrary spectra [48-50], and suggesting new designs that rival the best existing ones for certain applications [51, 52]. In atomic modeling, neural networks have been trained to forecast material properties like heat capacity [53-55] as well as interatomic potentials[56-59] and synthesis conditions[60], offering an alternative to density functional theory simulations. Once trained, these models can generate predictions almost instantaneously, a stark contrast to the consistent computational costs of traditional simulations. However, changing any of the parameters that the training sets are based on often necessitates building an entirely new model with new data, which can be extremely time consuming.

Transfer learning (TL) emerges as a promising technique to enhance the training efficiency for related tasks. It involves transferring the knowledge or weights from a pre-trained network to a new network, offering a head start compared to initializing weights randomly (Fig. 2). For example, the insights from a model trained to identify cats can be transferred to a new model aimed at recognizing lions. If the tasks share common predictive features, this can reduce the training time and the number of training samples required. In large language models (LLMs), this is a very common approach for a model to learn specific language skills, after having it pre-trained on a vast corpus of raw text. For example, understanding instructions is a task that the current iteration of ChatGPT was taught via transfer learning of the underlying GPT3.5 model. Also, image generation models can learn secondary tasks from small amounts of additional data in this way, for example drawing in the style of an artist which was not present in the pre-training dataset[61]. While TL can mitigate the data and training demands for closely related tasks, its effectiveness has its limits [62, 63]. In some scenarios, the efficiency gains are minimal, and TL falls short when it comes to predicting tasks that are vastly different, particularly when the source model is trained on a similarly narrow task.

The next frontier for AI applications in materials science lies in expanding their scope, generalizability, and ease-of-use, taking cues from the recent advancements of generative models in other domains. Envisioning a system where more valuable insights can be drawn from a broader array of data inputs and be encoded into a compact representation for diverse downstream tasks is

crucial. Since general material types and their properties comprise a vastly heterogeneous dataset, multi-modal deep learning techniques will be necessary[64]. This evolution would also involve creating intuitive user interfaces that allow queries in standard language, enabling searches for materials, compositions, or designs tailored to specific requirements.

Imagine a scenario where a user could simply articulate a desired material property, such as "a material that reflects all light between 300nm and 700nm," or input a chemical formula to predict its electronic ground state. The AI system would then discern the topic, understand the type of response sought, and either retrieve or generate appropriate information. Similarly, users could be able to input design parameters or compositions of proposed materials and receive accurate predictions of their behaviors and feasibility assessments for fabrication.

Such an AI model would not only bridge the vast array of existing material and atomic data but also integrate data related to the underlying physics. This dual functionality would empower the model to act both as a comprehensive search engine for existing solutions within online databases and as a generative model, capable of generalizing to entirely new inputs, significantly enhancing the field's predictive and exploratory capabilities.

3. All-in-one materials model

In this section we discuss the main challenges towards our vision of a true "foundation model" for materials sciences, that understands and is capable to predict materials in a broad sense, both in terms of the material type (molecule, crystal, meta-material, etc.) and of its properties (optical, thermic, mechanical, etc.). We foresee two large challenges that need to be overcome for a large material model to be feasible. The first and most important challenge is the data, the second challenge is the different, more heterogeneous, and more quantitative nature of the problem, compared with language or vision tasks.

3.1 Data

To realize the vision of a comprehensive, all-encompassing materials model, a significantly expanded data repository is essential. Analogous to models like ChatGPT, which leverage vast quantities of unlabeled text data scraped from social media and other platforms to accurately represent natural language patterns and topics, materials science requires a similar breadth of data. To get an idea of the vast quantities of data used for LLM training, META's latest model "Llama2"

was trained on text comprising around 2 trillion words. Among other sources, the full multi-lingual Wikipedia text data is included, however, Wikipedia text only corresponds to roughly 1% of the full training data[65]. Gathering enough data is therefore a key challenge towards a general materials foundation model. In material science currently, a wealth of research data, both experimental and simulated, is generated, but most of it remains underutilized post-publication. The data employed in existing models represent just a sliver of the global repository, resulting in models with limited predictive breadth, even the most precise ones (Fig. 3). To create a large database of materials data, we suggest two approaches that should ideally be followed simultaneously.

Open data repository

As a first means for gathering worldwide generated materials data we suggest the creation of a centralized database where researchers worldwide can deposit their data. We foresee two main challenges in this endeavor: first, to harness this data effectively, some degree of standardization and manual labeling is imperative to ensure consistency and homogeneity, which would be basic requirements for a high model accuracy. Initiatives to amass such data are in progress[66], but further progress is needed. Also, to efficiently handle these extensive, heterogenous datasets, the development of advanced deep learning architectures is paramount (see below). Second: researchers have little time. Data often remains unpublished not because of confidentiality, but because it costs time to clean the data structure, write meta-information, upload, and label the repository, etc. A system to reward participation will therefore be necessary to motivate researchers to participate in data sharing. National research foundations and other funding agencies today often pay open access publication fees under the condition that the data is shared openly. These programs need to be reinforced and appended on the condition that data is provided in a standardized format on an approved dissemination channel. A data repository could also provide a licensing platform, which could guarantee that contributing researchers are required to be acknowledged by the authors of models that are trained using their data, as well as guarantee open and transparent access to the platform.

Automatic scrapers for research data and articles

The goal is to establish a user-curated database that aggregates datasets and findings from all publications within pertinent fields. However, this ideal solution may be impractical in the near term because it will take time to convince researchers worldwide to standardize and contribute their data. As a secondary measure, leveraging NLP models to mine the vast corpus of peer-reviewed papers presents a more attainable approach for constructing a large-scale database. These models have already proven capable of ingesting thousands of scientific articles to distill key insights[67, 68]. This methodology can be extended to extract material-behavior relationships from experimental and theoretical studies across various material types. While this approach might not offer the comprehensive dataset that would be necessary for general component prediction, material design, or atomistic modeling, it can efficiently facilitate the search for existing solutions. This bypasses the labor-intensive task of manually reviewing the entire body of relevant literature. Additionally, NLP models can be employed to extract other material-related information, such as experimental findings like indices of refraction, mechanical attributes, electronic properties, etc. Last-generation large language models could be specifically fine-tuned on scientific literature for pertinent, quantitative data extraction. This fine-tuning would require a moderate labeling effort, where quantitative data needs to be provided in the desired format, following the specifications for the global dataset. After fine-tuning on this training corpus of publications, the model will become capable to analyze scientific papers and extract the datasets quantitatively and in a standardized way. Note that a vision-capable model would be necessary to extract results also from figures. In the automatic dataset generation, in order to extract useful features from papers that have no “hard” labels, but are described by prose text, strategies of contrastive representation learning can also be used[69].

An essential problem that would need to be addressed is the possibility for models like ChatGPT to generate fake, but convincing sounding information. At the core of the system, the model is attempting to convincingly produce language similar to the language it has been trained on. In a sense it may end up prioritizing producing text that sounds like it is correct, rather than being correct. This would pose a much larger problem in the realm of materials science, as any quantitative falsities would make its predictions for real world applications entirely useless or infeasible to fabricate. Preventing generative models from creating falsities is still an ongoing development in natural language processing[70], particularly relevant in scientific writing[71-73],

so any development of a similar large scale generative model would need to keep up with the ongoing improvements in research.

Potential challenges related with quantitative data

A further potential problem concerns the fact that the most impressive large models today are language and vision models, both being rather qualitative domains. Materials science on the other hand is more strictly quantitative. An LLM-based data extraction procedure may thus face unforeseeable difficulties that are not relevant for common language or vision tasks, but substantial for quantitative data extraction and processing. One potential remedy might involve a strategic shift in the training data utilized. Generative models generally employ a semi-supervised learning approach, blending unsupervised data (abundant but unlabeled) with supervised data (scarce but accurately labeled) to establish robust associations between the data and accurate outcomes. Specifically, ChatGPT relies on unsupervised data for initial pre-training, followed by supervised data for subsequent fine-tuning. Diffusion vision models also use a mix of supervised training (labeled text/image pairs) and unsupervised denoising. To mitigate the generation of inaccuracies, any prospective materials science model may need to emphasize supervised data to a greater extent.

This would necessitate a more stringent selection process for vetting published data and setting elevated standards for user-contributed data to ensure higher fidelity in the model's predictions. This would involve the formatting standards set by a central body for an open data repository. Additionally, analogous to so-called “alignment” efforts, that aim at reducing biases in LLMs, a reinforcement learning-based strategy can be employed to improve its factual accuracy metrics[74, 75]. Human feedback can be used to fine-tune the model. This can apply to using preference modeling to improve the helpfulness of responses the way current LLMs do[76], as well as constructing rewards based on the agreement between proposed quantitative solutions and experimental verification.

A further complication arises from the variable quality and applicability of results, sourced directly from scholarly papers. Simulation outcomes are particularly sensitive to the assumptions made regarding environmental conditions, the dielectric properties of involved materials, or the chosen meshing strategies. Similarly, experimental results can vary significantly depending on the

methodologies and equipment utilized. Compounding this issue is the fact that some findings may be subsequently refuted or deemed non-reproducible.

For a comprehensive model, designed to aggregate data by mining existing literature, a sophisticated NLP component is essential. Contrastive Language-Image Pre-training (CLIP) pairs natural language supervision with image modeling to efficiently encode visual information[69]. A similar strategy may be employed, targeted at language-material pre-training to better target information for improving predictive power. This component must not only contextualize research findings in light of the employed methods, but also develop the capability to assess the relative credibility of different results. Even a centralized database, populated by datasets voluntarily submitted by researchers would necessitate mechanisms to evaluate the outcomes in relation to the adopted methodologies. Ideally, such a system would evolve to discern the reliability of results based on the employed methodologies, thereby enhancing the robustness and utility of its predictions. Just as models like DALL-E and ChatGPT have featured multiple generations that have continually improved results, an all-in-one materials model would be fine-tuned over multiple iterations, punishing and deprioritizing poor results to improve performance, and implementing architectural improvements or novel deep learning concepts.

3.2. Deep Learning techniques

Foundation model architectures

As mentioned already above, bridging the gap between the use of ML in materials science and that in natural language processing and computer vision is not possible alone by training existing models to new data. The models themselves need to be adapted. Computer vision often uses convolutional neural networks (with extensions like the attention mechanism[23]). Further CNNs have been recently proposed, demonstrating that modern network design approaches can lead to parity performance compared with vision transformers[77]. However, convolutional networks are most efficient for array-like, structured data, which an all-encompassing material network won’t be restricted to.

NLP foundation models use different architectures, today mostly transformers[25]. The latter architecture is highly flexible with regards to the data format, but it doesn’t scale ideally with its dimension since self-attention layers are essentially fully-connected. However transformers are extremely strong at generalizing, provided huge amounts of data are available for training [78].

In conclusion, we believe that the heterogeneous character of parametrizations of materials and their properties cannot be ideally solved with a single architecture. An all-in-one scientific materials model will rather be composed of multiple, interconnected models, each relatively specialized. All their respective predictions may then very well be processed by a global multi-modal model (see also below), that may potentially be a transformer.

Tokenizer

The "tokenizer" in LLMs is an independent model that performs a kind of translation pre-processing step, converting the natural language input into a compressed, learned representation. In materials, the tokenizer would be a model that converts the bare material and property inputs into a latent representation. This can then be easier processed by further deep learning models. The choice of tokenizer model can have a severe impact on the full model's performance. For instance, in multi-lingual LLMs, an English-only tokenizer does in fact work, but the full model then performs significantly worse compared to using a dedicated, multi-lingual tokenizer [79]. The use of latent representations to compress input data has been explored in some works in this field[80], [62] but further advancement in this area is necessary. Advancement will involve understanding how to encode most efficiently information specific to materials tasks, such as atomic and structural properties.

Multimodality

As stated before, even if high-quality materials data were available at a similar scale, due to their heterogeneity, the approaches that lead to the tremendous success of LLMs and computer vision foundation models cannot be directly applied. On the one hand, various very different types of material families with totally different parametrization format must be dealt with, on the other hand a plethora of physical and chemical properties needs to be described, again requiring also different description formats. The problem is thus comparable to recent attempts of combining language with vision or audio models, which is generally described as "multimodal deep learning" [64]. In fact, our perspective for a universal materials tokenizer is by itself a multimodal representation model [81] Multimodality concepts will need to be applied throughout the entire model design, to allow for a universal deep learning neural network to treat all possible types of

different materials and their properties. Since it will be impossible to design from the start an all-encompassing architecture that comprises all potential possibilities, focusing the development also on extensibility will as well be an important criterion.

Applicable techniques that go beyond language and vision foundation models include also incorporating physics-based knowledge into modeling, which can yield more meaningful features. Physics informed neural networks (PINNs) or neural operators may for instance be used (for adequate materials families or property categories), either for regularization or for full downstream predictions in corresponding branches of the global model[82-85]. The path to such a proposed all-in-one model would require many steps (Fig. 4). A massive training dataset would be built, using the methods discussed in the previous sections. Data would be fed in, comprised of materials descriptions such as molecular compositions, lattice information, geometric parameters, paired with their corresponding physical or chemical behavior. One could even imagine including info about experimental methods or simulation conditions in the dataset, such that the global model can learn to propose experiments or simulations methods for the characterization of a material. This totality of data could then be parametrized in a homogenous format.

On the DL side, different models are needed for different parts of the process. First, a unique tokenizing model to convert qualitative and quantitative data about materials, their properties and their behavior to numerical values and compress them to a latent representation. This strategy is already employed with models such as LLMs, but adapting the architectures to account for differences in the target domains can greatly improve their performance. Next, a central foundation model can be trained. This will require the largest amount of data. If the dataset is massive enough, a suitably large model can develop a global understanding of the problem, such that it becomes able to learn efficiently from few-shot data, like large language models, and use results from similar classes of tasks to assist in training.

Finally, numerous submodule models can be connected to the central pre-trained model that are specified for narrower tasks. A statistical model will always have some amount of error, and deep learning furthermore is a black-box method, so predicted results always need to be verified. It may therefore be interesting to train a second input pre-processor model on the task to search for potentially available existing data. This would prevent the chance of generating an approximate solution when an exact one is already known. The portion of data corresponding to existing solutions would comprise a portion of the total predictive power of the model (Fig. 5), that

would be, on average, more accurate. Technically, such a model could be granted access to a database, looking for similarities in the tokenized material descriptions. Self-prompting techniques, in which a model refines its output by iteratively adding externally acquired information to its input, could further enhance performance for queries on totally new inputs[86]. In the future, it may also be interesting to incorporate NLP models to parse a user's human readable input to understand the intended request in natural language and convert it into quantitative values as input for the rest of the model. Likewise, a separate module for converting the final predicted suggestion back into easily human-readable language would also be helpful to make the model accessible for a large audience. By adopting these advancements in ease-of-use and scope, AI in materials science can unearth its full potential, and enable innovative transformations across many applications, revolutionizing how we design, discover, and use materials.

4. Conclusions

In conclusion, while not yet there, we foresee that machine learning in materials research will find itself taking a similar trajectory as research in natural language processing and computer vision, two fields that have garnered significant interest since the inception of the field. A few years ago, the most powerful state-of-the-art models in NLP and vision were still limited to highly specialized predictions, given a well curated set of labeled data. In remarkably short time, however, researchers developed models that could efficiently encode insights from significantly larger datasets and mostly unlabeled, raw samples, spanning a far larger scope of potential inputs and outputs. Developments in representation learning have allowed for useful intermediary representations of data to be learned for both supervised and unsupervised tasks, which can be mapped to a wide range of tasks. This has helped address data concerns as well as increase the generalizability of models, two key drawbacks of standard predictive modeling.

Likewise, breakthroughs in generative modeling have allowed for unparalleled potential in the creation of new data aimed at specific tasks. These large-scale generative models have begun a revolution of large language and image generation models. They can accomplish a wide range of tasks with input in everyday language. This has also allowed non-technical users to leverage its power for their own needs. We claim that, as materials research is approaching a similar crossroads, models start to become sufficiently complex and well-designed to allow for high performance on a wide range of, for now, well specified tasks. At the same time, there is a large

wealth of high-quality data, both labeled and unlabeled, being generated constantly through worldwide research efforts, albeit not as easily accessible as text and images are on the internet. Establishing a centralized repository would enable researchers around the world to contribute their data, fostering a collaborative environment for the advancement of materials science. Utilizing this data to train an all-in-one model tailored for domain specific needs would allow for revolutionary advancement in many applications. Collective efforts by the community. careful design of the models, as well as an intelligent combination of techniques of NLP and computer-vision foundation models, can allow this knowledge to be properly harnessed and pave the way for transformative effects in research and application.

Acknowledgements

R.U and Y.Z acknowledge the financial support of the National Institute of General Medical Sciences of the National Institutes of Health (1R01GM146962-01).

P.R.W acknowledges funding from the French Agence Nationale de la Recherche (ANR) under the grant ANR-22-CE24-0002 (project NAINOS).

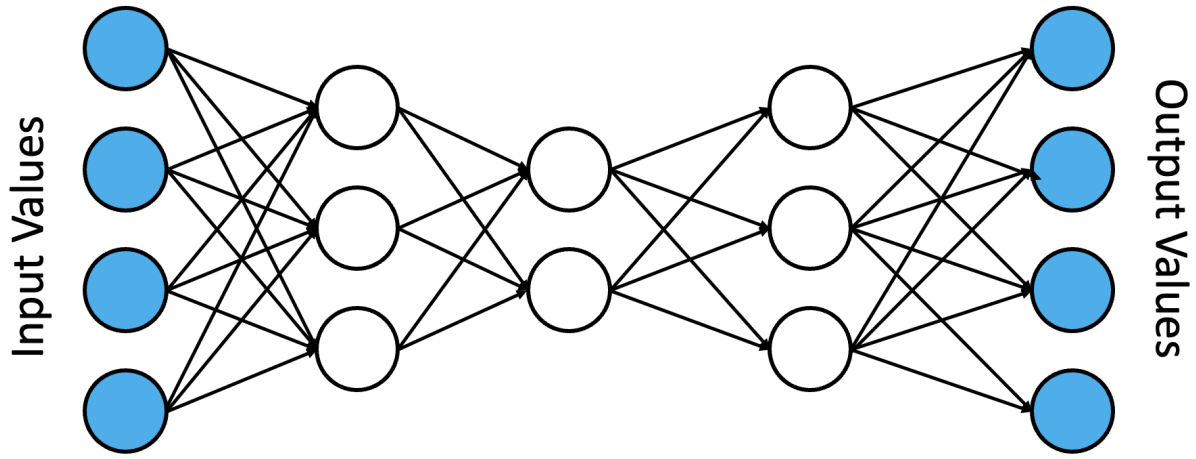

Figure 1: Basic architecture of a standard feedforward neural network. Values in the neurons in one layer are transformed to the values in the next layer based on the weight connections between each node. The overall transformation is dependent on the dimension of the input and output values.

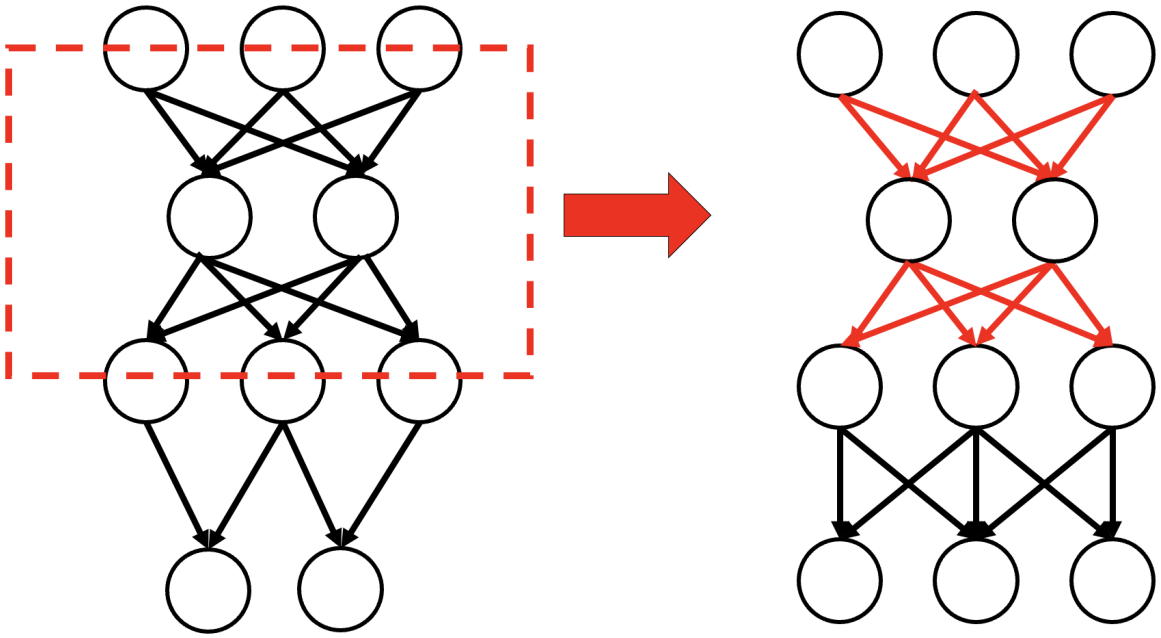

Figure 2: Diagram of basic transfer learning operation between models with similar input and slightly different outputs. The first model is trained (left) and the weight connections after training are transferred to a new model (right) rather than initializing the values randomly.

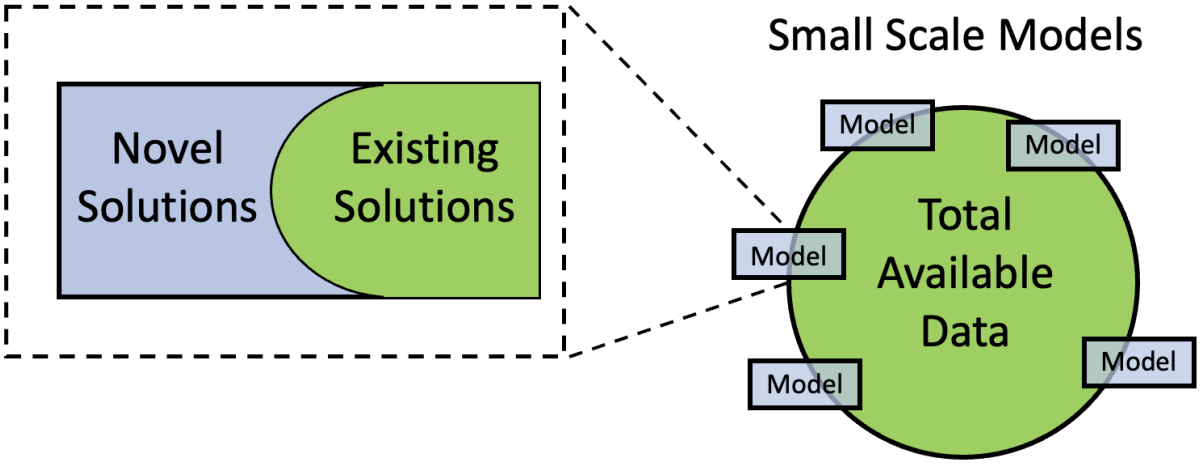

Figure 3: Diagram representing the difference in scope of potential materials models. Currently, models with a singular focus are trained on a small portion of all available global data. These models can both accurately retrieve solutions based on existing data (green), as well as offer suggestions for novel solutions (blue) within limited constraints.

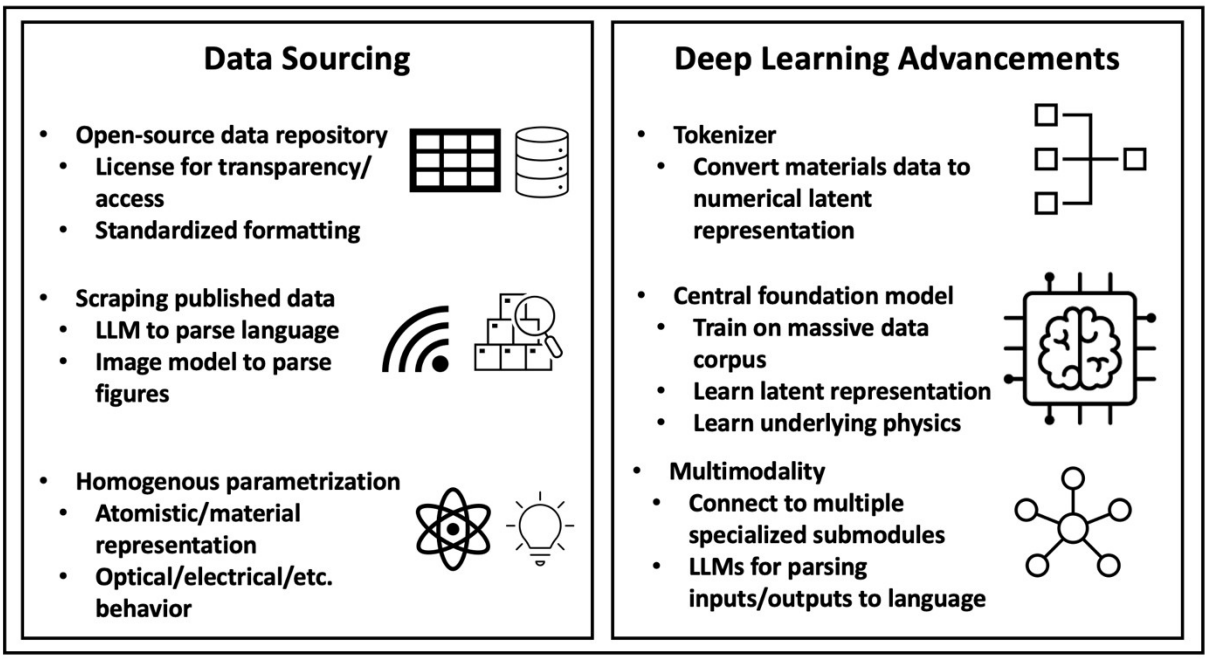

Figure 4: Challenges to be solved for a hypothetical all-in-one model. A massive central homogenously formatted dataset is needed. This can be sourced from an open-access repository as well as developing advanced models to scrape published results. On the modeling side, advancements are needed to efficiently tokenize and represent key information from materials data. A central foundation model needs to be built that can learn the underlying working principles and physics. This can then be connected in a modular fashion to numerous sub-models trained for more specific tasks.

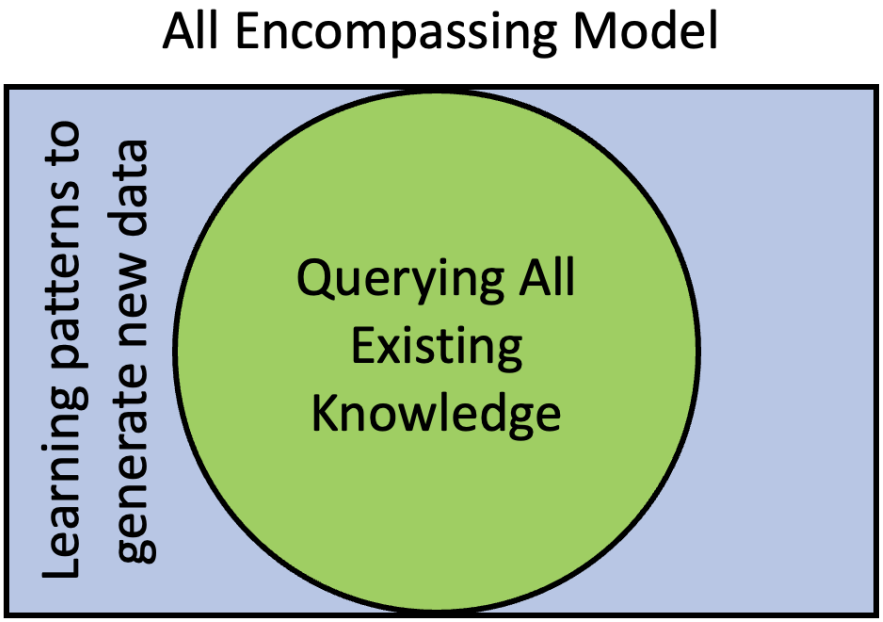

Figure 5: Diagram representing the scope of an idealized all-in-one model. Such a hypothetical model would be able to search a large range of existing data and learn underlying patterns to generate new solutions for a wider variety of tasks.

References

[1] E. Alpaydin, Introduction to Machine Learning, 3 ed., MIT Press, Cambridge, MA, 2014.
[2] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data 8(1) (2021) 53.
[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Curran Associates Inc., Lake Tahoe, Nevada, 2012, pp. 1097–1105.
[4] J. Chai, H. Zeng, A. Li, E.W.T. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Machine Learning with Applications 6 (2021) 100134.
[5] S.V. Mahadevkar, B. Khemani, S. Patil, K. Kotecha, D.R. Vora, A. Abraham, L.A. Gabralla, A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions, IEEE Access 10 (2022) 107293-107329.
[6] J.C. Caicedo, J. Roth, A. Goodman, T. Becker, K.W. Karhohs, M. Broisin, C. Molnar, C. McQuin, S. Singh, F.J. Theis, A.E. Carpenter, Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images, Cytometry. Part A : the journal of the International Society for Analytical Cytology 95(9) (2019) 952-965.
[7] J.C. Caicedo, S. Cooper, F. Heigwer, S. Warchal, P. Qiu, C. Molnar, A.S. Vasilevich, J.D. Barry, H.S. Bansal, O. Kraus, M. Wawer, L. Paavolainen, M.D. Herrmann, M. Rohban, J. Hung, H. Hennig, J. Concannon, I. Smith, P.A. Clemons, S. Singh, P. Rees, P. Horvath, R.G. Linington, A.E. Carpenter, Data-analysis strategies for image-based cell profiling, Nature Methods 14(9) (2017) 849-863.
[8] M.M. Ahsan, S.A. Luna, Z. Siddique, Machine-Learning-Based Disease Diagnosis: A Comprehensive Review, Healthcare (Basel) 10(3) (2022).
[9] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent Trends in Deep Learning Based Natural Language Processing [Review Article], IEEE Computational Intelligence Magazine 13(3) (2018) 55-75.
[10] E. Cambria, B. White, Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article], IEEE Computational Intelligence Magazine 9(2) (2014) 48-57.
[11] T. Zhang, H. Mo, Reinforcement learning for robot research: A comprehensive review and open issues, International Journal of Advanced Robotic Systems 18(3) (2021) 17298814211007305.
[12] D. Kim, S.-H. Kim, T. Kim, B.B. Kang, M. Lee, W. Park, S. Ku, D. Kim, J. Kwon, H. Lee, J. Bae, Y.-L. Park, K.-J. Cho, S. Jo, Review of machine learning methods in soft robotics, PLOS ONE 16(2) (2021) e0246102.
[13] H. Nguyen, H. La, Review of Deep Reinforcement Learning for Robot Manipulation, 2019 Third IEEE International Conference on Robotic Computing (IRC), 2019, pp. 590-595.
[14] T. Hastie, The elements of statistical learning : data mining, inference, and prediction, Second edition. ed., Springer, New York, NY, 2009.
[15] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3(1) (2016) 1-40.
[16] M. Kaya, S. Hajimirza, Using a Novel Transfer Learning Method for Designing Thin Film Solar Cells with Enhanced Quantum Efficiencies, Scientific Reports 9(1) (2019) 5034.
[17] L. Breiman, Random Forests, Machine Learning 45(1) (2001) 5-32.
[18] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy (Basel) 23(1) (2020).
[19] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013).
[20] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-Supervised Learning: Generative or Contrastive, IEEE Transactions on Knowledge and Data Engineering 35(1) (2023) 857-876.
[21] S. Bond-Taylor, A. Leach, Y. Long, C.G. Willcocks, Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models, IEEE Transactions on Pattern Analysis & Machine Intelligence 44(11) (2022) 7327-7347.
[22] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, in: M. Marina, Z. Tong (Eds.) Proceedings of the 38th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, 2021, pp. 8821--8831.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 10674-10685.
[24] OpenAi, GPT-4 Technical Report, arXiv e-prints (2023) arXiv:2303.08774.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 6000–6010.
[26] P.P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121-154.
[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.) Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 1877-1901.
[28] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349(6245) (2015) 255-260.
[29] J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat 1(3) (2019) 338-358.
[30] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J.L. Billinge, E. Holm, S.P. Ong, C. Wolverton, Recent advances and applications of deep learning methods in materials science, npj Computational Materials 8(1) (2022) 59.
[31] K. Hippalgaonkar, Q. Li, X. Wang, J.W. Fisher, J. Kirkpatrick, T. Buonassisi, Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics, Nature Reviews Materials 8(4) (2023) 241-260.
[32] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, A Critical Review of Machine Learning of Energy Materials, Advanced Energy Materials 10(8) (2020) 1903242.
[33] A. Chen, X. Zhang, Z. Zhou, Machine learning: Accelerating materials development for energy storage and conversion, InfoMat 2(3) (2020) 553-576.
[34] P.R. Wiecha, A. Arbouet, C. Girard, O.L. Muskens, Deep learning in nano-photonics: inverse design and beyond, Photonics Research 9(5) (2021) B182-B200.
[35] M.A. Bessa, P. Glowacki, M. Houlder, Bayesian Machine Learning in Metamaterial Design: Fragile Becomes Supercompressible, Advanced Materials 31(48) (2019) 1904845.
[36] K. Yao, Y. Zheng, Deep-Learning-Assisted Inverse Design in Nanophotonics, in: K. Yao, Y. Zheng (Eds.), Nanophotonics and Machine Learning: Concepts, Fundamentals, and Applications, Springer International Publishing, Cham, 2023, pp. 113-140.
[37] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Mater. 3 (2017).
[38] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Physical Review B 89(9) (2014) 094104.
[39] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Data Mined Ionic Substitutions for the Discovery of New Compounds, Inorganic Chemistry 50(2) (2011) 656-663.
[40] C. Miles, R. Samajdar, S. Ebadi, T.T. Wang, H. Pichler, S. Sachdev, M.D. Lukin, M. Greiner, K.Q. Weinberger, E.-A. Kim, Machine learning discovery of new phases in programmable quantum simulator snapshots, Physical Review Research 5(1) (2023) 013026.
[41] R.S. Hegde, Deep learning: a new tool for photonic nanostructure design, Nanoscale Advances 2(3) (2020) 1007-1023.
[42] Y. Liu, K.P. Kelley, R.K. Vasudevan, H. Funakubo, M.A. Ziatdinov, S.V. Kalinin, Experimental discovery of structure–property relationships in ferroelectric materials via active learning, Nature Machine Intelligence 4(4) (2022) 341-350.
[43] X.-L. Yu, B. Yi, X.-Y. Wang, Prediction of the Glass Transition Temperatures for Polymers with Artificial Neural Network, Journal of Theoretical and Computational Chemistry 07(05) (2008) 953-963.
[44] D.R. Cassar, A.C. Carvalho, E.D. Zanotto, Predicting glass transition temperatures using neural networks, Acta Mater. 159 (2018).
[45] R. Unni, K. Yao, X. Han, M. Zhou, Y. Zheng, A mixture-density-based tandem optimization network for on-demand inverse design of thin-film high reflectors, Nanophotonics 10(16) (2021) 4057-4065.

[46] Z.A. Kudyshev, A.V. Kildishev, V.M. Shalaev, A. Boltasseva, Machine-learning-assisted metasurface design for high-efficiency thermal emitter optimization, Applied Physics Reviews 7(2) (2020) 021407.
[47] J.A. Fan, Freeform metasurface design based on topology optimization, MRS Bulletin 45(3) (2020) 196-201.
[48] J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria, B.G. DeLacy, J.D. Joannopoulos, M. Tegmark, M. Soljačić, Nanophotonic particle simulation and inverse design using artificial neural networks, Science Advances 4(6) (2018).
[49] C.C. Nadell, B. Huang, J.M. Malof, W.J. Padilla, Deep learning for accelerated all-dielectric metasurface design, Optics Express 27(20) (2019) 27523-27535.
[50] I. Malkiel, I. Mrejen, A. Nagler, U. Arieli, L. Wolf, H. Suchowski, Plasmonic nanostructure design and characterization via Deep Learning, Light: Science & Applications 7(1) (2018) 1-8.
[51] P.R. Wiecha, O.L. Muskens, Deep Learning Meets Nanophotonics: A Generalized Accurate Predictor for Near Fields and Far Fields of Arbitrary 3D Nanostructures, Nano Letters 20(1) (2020) 329-338.
[52] W. Ma, Z. Liu, Z.A. Kudyshev, A. Boltasseva, W. Cai, Y. Liu, Deep learning for the design of photonic structures, Nature Photonics 15(2) (2021) 77-90.
[53] R.Q. Snurr, Machine learning heat capacities, Nature Materials 21(12) (2022) 1342-1343.
[54] S.M. Moosavi, B.Á. Novotny, D. Ongari, E. Moubarak, M. Asgari, Ö. Kadioglu, C. Charalambous, A. Ortega-Guerrero, A.H. Farmahini, L. Sarkisov, S. Garcia, F. Noé, B. Smit, A data-science approach to predict the heat capacity of nanoporous materials, Nature Materials 21(12) (2022) 1419-1425.
[55] M.N. Aldosari, K.K. Yalamanchi, X. Gao, S.M. Sarathy, Predicting entropy and heat capacity of hydrocarbons using machine learning, Energy and AI 4 (2021) 100054.
[56] V.L. Deringer, M.A. Caro, G. Csányi, Machine Learning Interatomic Potentials as Emerging Tools for Materials Science, Advanced Materials 31(46) (2019) 1902765.
[57] F.A. Faber, L. Hutchison, B. Huang, J. Gilmer, S.S. Schoenholz, G.E. Dahl, O. Vinyals, S. Kearnes, P.F. Riley, O.A. von Lilienfeld, Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error, Journal of Chemical Theory and Computation 13(11) (2017) 5255-5264.
[58] R. Pederson, B. Kalita, K. Burke, Machine learning and density functional theory, Nature Reviews Physics 4(6) (2022) 357-358.
[59] P. Cats, S. Kuipers, S. de Wind, R. van Damme, G.M. Coli, M. Dijkstra, R. van Roij, Machine-learning free-energy functionals using density profiles from simulations, APL Materials 9(3) (2021) 031109.
[60] H. Huo, C.J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain, G. Ceder, Machine-Learning Rationalization and Prediction of Solid-State Synthesis Conditions, Chemistry of Materials 34(16) (2022) 7323-7336.
[61] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, arXiv [cs.CV] (2023).
[62] Y. Qu, L. Jing, Y. Shen, M. Qiu, M. Soljačić, Migrating Knowledge between Physical Scenarios Based on Artificial Neural Networks, ACS Photonics 6(5) (2019) 1168-1174.
[63] Z. Fan, C. Qian, Y. Jia, M. Chen, J. Zhang, X. Cui, E.-P. Li, B. Zheng, T. Cai, H. Chen, Transfer-Learning-Assisted Inverse Metasurface Design for 30% Data Savings, Physical Review Applied 18(2) (2022) 024022.
[64] C. Akkus, L. Chu, V. Djakovic, S. Jauch-Walser, P. Koch, G. Loss, C. Marquardt, M. Moldovan, N. Sauter, M. Schneider, R. Schulte, K. Urbanczyk, J. Goschenhofer, C. Heumann, R. Hvingelby, D. Schalk, M. Aßenmacher, Multimodal Deep Learning, arXiv [cs.CL] (2023).
[65] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, arXiv [cs.CL] (2023).
[66] K.M. Jablonka, L. Patiny, B. Smit, Making the collective knowledge of chemistry open and machine actionable, Nature Chemistry 14(4) (2022) 365-376.
[67] R. Yan, X. Jiang, W. Wang, D. Dang, Y. Su, Materials information extraction via automatically generated corpus, Scientific Data 9(1) (2022) 401.
[68] T. Gupta, M. Zaki, N.M.A. Krishnan, Mausam, MatSciBERT: A materials domain language model for text mining and information extraction, npj Computational Materials 8(1) (2022) 102.
[69] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv [cs.CV] (2021).
[70] M. Abdullah, A. Madain, Y. Jararweh, ChatGPT: Fundamentals, Applications and Social Impacts, 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022, pp. 1-8.
[71] H.H. Thorp, ChatGPT is fun, but not an author, Science 379(6630) (2023) 313-313.

[72] G. Grimaldi, B. Ehrler, AI et al.: Machines Are About to Change Scientific Publishing Forever, ACS Energy Letters 8(1) (2023) 878-880.
[73] A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models, Nature Reviews Physics 5(5) (2023) 277-280.
[74] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring How Models Mimic Human Falsehoods, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3214-3252.
[75] P. Christiano, J. Leike, T.B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, arXiv [stat.ML] (2023).
[76] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, J. Kaplan, Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, arXiv [cs.CL] (2022).
[77] Z. Liu, H. Mao, C.Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966-11976.
[78] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, J. Carreira, Perceiver: General Perception with Iterative Attention, in: M. Marina, Z. Tong (Eds.) Proceedings of the 38th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, 2021, pp. 4651--4664.
[79] M. Ali, M. Fromm, K. Thellmann, R. Rutmann, M. Lübbering, J. Leveling, K. Klug, J. Ebert, N. Doll, J.S. Buschhoff, C. Jain, A.A. Weber, L. Jurkschat, H. Abdelwahab, C. John, P.O. Suarez, M. Ostendorff, S. Weinbach, R. Sifa, S. Kesselheim, N. Flores-Herr, Tokenizer Choice For LLM Training: Negligible or Crucial?, arXiv [cs.LG] (2023).
[80] Y. Kiarashinejad, S. Abdollahramezani, A. Adibi, Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures, npj Computational Materials 6(1) (2020) 12.
[81] W. Guo, J. Wang, S. Wang, Deep Multimodal Representation Learning: A Survey, IEEE Access 7 (2019) 63373-63394.
[82] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686-707.
[83] C. White, J. Berner, J. Kossaifi, M. Elleithy, D. Pitt, D. Leibovici, Z. Li, K. Azizzadenesheli, A. Anandkumar, Physics-Informed Neural Operators with Exact Differentiation on Arbitrary Geometries, The Symbiosis of Deep Learning and Differential Equations III, 2023.
[84] M. Chen, R. Lupoiu, C. Mao, D.-H. Huang, J. Jiang, P. Lalanne, J.A. Fan, High Speed Simulation and Freeform Optimization of Nanophotonic Devices with Physics-Augmented Deep Learning, ACS Photonics 9(9) (2022) 3110-3123.
[85] Y. Augenstein, T. Repän, C. Rockstuhl, Neural Operator-Based Surrogate Solver for Free-Form Electromagnetic Inverse Design, ACS Photonics 10(5) (2023) 1547-1557.
[86] J. Li, Z. Zhang, H. Zhao, Self-Prompting Large Language Models for Zero-Shot Open-Domain QA, arXiv [cs.CL] (2023).