Foundation Model for Material Science

Seiji Takeda[1], **Akihiro Kishimoto**[1], **Lisa Hamada**[1], **Daiju Nakano**[1], **John R. Smith**[2]

[1] IBM Research - Tokyo
[2] IBM Thomas J. Watson Research Center
SEIJITKD@jp.ibm.com, Akihiro.Kishimoto@ibm.com, Lisa.Hamada@ibm.com, dnakano@jp.ibm.com, jsmith@us.ibm.com

Abstract

Foundation models (FMs) are achieving remarkable successes to realize complex downstream tasks in domains including natural language and vision. In this paper, we propose building an FM for material science, which is trained with massive data across a wide variety of material domains and data modalities. Nowadays machine learning models play key roles in material discovery, particularly for property prediction and structure generation. However, those models have been independently developed to address only specific tasks without sharing more global knowledge. Development of an FM for material science will enable overarching modeling across material domains and data modalities by sharing their feature representations. We discuss fundamental challenges and required technologies to build an FM from the aspects of data preparation, model development, and downstream tasks.

Introduction

AI has achieved various milestones such as superhuman performance in playing games (Campbell, Hoane Jr., and Hsu 2002; Silver et al. 2016) and quizzes (Ferrucci et al. 2010) as well as accurate predictions of protein structures in biology (Senior et al. 2020). Essential technologies to achieve such successes include planning and search, machine learning and natural language processing (NLP).

State-of-the-art machine learning methods use large-scale datasets to significantly improve the performance. BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020) learn feature representations for natural language from 3–500 billion tokens. The trend of using large-scale data has also been observed to tackle more complex tasks across multiple domains requiring multimodal data. For example, using 250 million image-text pairs, DALL-E (Ramesh et al. 2021) generates an image that corresponds to a text description.

In general, *foundation models* (FMs) are large models pretrained by broad datasets in self-supervised manners instead of targeting on specific discrete tasks. FMs aim to adapt to a variety of downstream tasks with zero or little additional training. BERT, GPT-3, and DALL-E are well-known examples that attempt to capture generic feature representations specifically on language and/or vision. However, while Bommasani et al. (2021) envision the opportunities and risks of the FMs, they are currently limited to the domains involving NLP and visions.

As a next challenge for AI research, we promote a universal FM pretrained with *multimodal* data for *material science*. Machine learning models have recently been adapted to material science (Gómez-Bombarelli et al. 2016; Pyzer-Knapp et al. 2022). However, these models are usually trained with small, *unimodal* datasets in specific material domains. Existing research uses training data whose size ranges between a few hundred (Takeda et al. 2020) and several hundred thousand (Wu et al. 2019). In addition, each independently learned model addresses only one specific task and does not effectively leverage the feature representations reusable for other models.

On the other hand, there are various types of data representing certain aspects of materials. In principle, training the material FM with such multimodal data should enable more generic, important features to be acquired that are applicable to many downstream tasks within material science. Additionally, as material scientists often come up with new ideas from one discipline to another, the material FM has the potential to reuse its feature representation even across different natural sciences such as chemistry and physics in the long run.

The structure of this paper is as follows: First, we give an overview of material science and AI, followed by their challenges. We then describe necessary steps and AI technologies to address these challenges, finally giving concluding remarks.

Material Science and AI

Material discovery has been key to grow various industries. Depending on the target industrial domain, new materials need to possess specific chemical/physical characteristics. For example, a material for solar cell needs to have a high light absorption efficiency, high mechanical strength durable for inclement weather, and low environmental toxicity. Those characteristics are determined by the internal structures of the materials in both microscopic and macroscopic scales where careful design such as atom configurations plays a crucial role.

In the conventional discovery process, material scientists carry out trial-and-error processes driven by their experience and intuition. A parameter space for material design is infinitely large, where only tiny portions bring desired properties. Therefore developing new materials generally takes 10 to 20 years and costs 10 to 100 million US dollars (Ray 2021).

AI has received increased attention to address obstacles in material discovery over the last several years. Some models are available as open source software to facilitate scientists' tasks, e.g., (Manica et al. 2022; Yang et al. 2017). Recent approaches employ machine learning that performs several tasks summarized here. For those tasks, material structures need to be represented by appropriate representations that include molecular graphs (Weininger 1988), crystal structures (Noh et al. 2020) and polymer strings (Lin et al. 2019), etc.

Given a structure of a material, predicting properties of that material is a major task. For example, electric conductivity is an essential property in developing electronic devices, for example, Organic Light Emitting Diode (OLED). Property prediction is formulated as a regression task to which neural network models have been applied (Gómez-Bombarelli et al. 2016).

Another task is inverse to the material property prediction: given material properties that can be exact values or ranges, new material structures satisfying these properties need to be generated. This task is more challenging because an appropriate material structure needs to be identified in a large search space of molecules estimated to contain at least 10^{60} possible candidates. Deep generative models have been actively researched to address this task (Gómez-Bombarelli et al. 2018). These generative models are specifically for small organic molecules.

Tasks on chemical synthesis are also essential, studied in the context of organic chemistry, including reaction predictions based on neural networks as well as chemical synthesis planning combined with other approaches, e.g., (Kishimoto et al. 2019; Schreck, Coley, and Bishop 2019; Segler, Preuss, and Waller 2018).

Those machine learning models are independently developed, so they are isolated in terms of data modality, material domains, and application tasks, thereby missing links between sharable cross-domain knowledges.

Challenges for Material FM

The material FM will realize overarching modeling across different material domains. Once the material FM is successfully pretrained, its learned feature representation serves as a basic feature set to address most of the downstream tasks in the previous section all together. However, development of the material FM is hindered by the fact that the knowledge representation that can describe a complete picture of materials is not available.

In theory, since materials obey the governing equations such as the Schrödinger equation, machine learning models should effectively share common principles behind the equations across different material domains. However, in practice, it is infeasible to find accurate, numerical solutions for each material. Therefore, knowledge on materials has been represented in various, human-interpretable ways such as molecular graph images also regarded as a sequence of tokens called SMILES (Weininger 1988), three-dimensional conformation of atoms, microscopic images, real-valued physical properties (e.g., orbital energies), and spectroscopic representations. These knowledge representations focus only on partial aspects of materials, without sharing comprehensive knowledge. In learning from richer knowledge from multiple knowledge representations, the material FM raises several fundamental challenges.

First, while most of the standard models have only at most two modalities on natural language and vision (Ramesh et al. 2021; Shridhar et al. 2020), the material FM requires many more modalities ranging from graphs to property values and even images. Relating one modality to others raises an issue caused by the dimensional complexity of multiple modalities.

Second, the fact that only a few material samples are available for some representations poses a challenge for creating multimodal training data. For example, Ramesh et al. (2021) created 250 million image-text pairs for text-to-image generation of DALL-E. On the other hand, while over one billion molecules are represented as SMILES (Irwin et al. 2020), far fewer examples are available for their physical properties that are more valuable than SMILES, e.g., 134K examples in the QM9 database (Ramakrishnan et al. 2014), 91 and 250 samples to design sugar and dye, respectively (Takeda et al. 2020), and at most 33K examples for spectra (NIST 1996). Such sparseness limits the ability to construct a large-scale dataset with no missing modalities.

Third, the material FM is forced to be trained with inaccurate and/or incomplete data. For example, many of the informative data are obtained either by computational simulations or by actual chemical experiments and measurements. These approaches introduce noises to the actual property values. For example, it is difficult to accurately measure or simulate the glass transition temperature, because the transition experimentally occurs over a wide temperature range and because the simulated result depends on the configuration on internal and external conditions (Liu et al. 2017). Additionally, they are time-consuming and have difficulty in scaling up the training data size. Another example is that only small portions of the entire UV/IR spectra are often simulated or experimented on, although the spectroscopic data play a more important role specifically in material science.

Fourth, some knowledge representations are based on images, raising a nontrivial issue on automatically extracting essential information. For example, many of the spectroscopic data illustrated in academic papers are accessible only in an image format without indicating the actual numbers (Probst et al. 2021). Microscopic data are another case where knowledge is represented as images.

Finally, since large-scale SMILES data are mostly for small organic molecules, specifically for drugs (Irwin et al. 2020), they do not best represent materials. While the number of molecular structures can be arbitrarily increased by creating artificial molecules on the basis of sampling, such an augmentation does not always lead to covering more practical material structures. For example, SMILES has dif-

Name	Size	Property examples	Note
ZINC20 (Irwin et al. 2020)	>1B	SMILES, 3-D atom conformation	Bioactive, bionegic, and drug-like molecules
PubChem (Kim et al. 2021)	112M	SMILES, molecular weight, TPSA, XLogP	Generic organic molecule database. Incomplete property sets sometimes found as text (examples omitted)
PubChemQC (Nakata and Shimazaki 2017)	3.2M	SMILES, HOMO-LUMO energy gap, dipole moment, 3-D atom conformation	DFT calculation for molecules selected from PubChem.
ChEMBLE (Gaulton et al. 2017)	2M	SMILES, polar surface area, bioactivities, molecular weight	Bioactive molecules
OQMD (Kirklin et al. 2015)	1M	Composition ratio, band gap, stability, formation energy	Inorganic crystal structures. DFT calculation.
QM9 (Ramakrishnan et al. 2014)	134K	SMILES, heat capacity, HOMO-LUMO energy gap	DFT calculation with up to 9 atoms. 20 physical properties.
Chemistry Webbook (NIST 1996)	$<$ 51K	Mass spectra, IR spectra, UV/Vis spectra	IR spectra for 16K molecules. UV/Vis spectra for 1.6K molecules.

Table 1: Examples of available datasets

ficulty in accurately describing materials based on macromolecules (e.g., polymers) due to their stochastic properties (Lin et al. 2019).

AI Technologies for Material FM

We discuss three necessary steps and their technical obstacles, while referring to AI technologies to realize the material FM.

Data Preparation

Data preparation is the first essential step for the success of the material FM. In general, the importance of large-scale datasets has been recognized in the material informatics community, e.g., (Dima et al. 2016; Jha et al. 2021). However, the material FM needs a large-scale *multimodal* dataset that describes various aspects for each material. The data should be collected by various methods to cover broad material domains across different modalities. As discussed in the previous section, the type, usefulness, quantity, and quality of the data depend on the modality. We discuss several approaches to collect large-scale data.

Integration of public datasets One straightforward approach is to integrate several public, structured datasets that represent basic properties as a uniform dataset. Table 1 summarizes examples of the datasets (e.g., ZINC20 and ChEMBLE). However, the uniform dataset cannot always fill in the data points of all properties, since those datasets do not possess property values of all materials. Considering only a set of materials whose physical properties are all available results in discarding the majority of the material data. Additionally, a large number of physical property values stored there are calculated either by simulations with a very small number of atoms due to high computational overhead (e.g., up to 9 atoms for QM9) or less informative heuristic values (XLogP and TPSA) calculated instantly, on the basis of counting material substructures.

Information extraction from text and images Despite conveying rich information, Table 1 shows that spectra data are very sparse. Although extracting spectra from academic papers is one way to increase the data size, they are provided only as images in those papers. This is also the case for papers with microscopic data. NLP and image processing techniques are needed for extracting useful information (e.g., actual numbers for spectra) from an image and associating them with its corresponding material, as is successfully done for chemical reactions (Lowe 2012). For example, the work of Lowe (2012) needs to be combined with automated data extraction algorithms for plots in a chart, e.g., (Cliche et al. 2017).

Simulations and experiments supported by AI Despite time-consuming steps, performing computational simulations or actual experiments is necessary to increase the dataset size. Such cases include both sparse spectroscopic data and more basic physical properties available in QM9 and PubChemQC. Active learning (Settles 2009) is one way for effective data collection, which needs to be applied here. For example, given an available budget such as time and money, this task is regarded as a problem to choose the next material to simulate and determine a set of experimental conditions that maximize the chance of obtaining desired results. Related work attempts to improve experimental design (Eyke, Green, and Jensen 2020) and automated data generation in chemical space (Smith et al. 2018; Park et al. 2020). Acceleration of physical simulation by surrogate model is another promising solution (Toledo-Marin et al. 2021).

FM Creation

Once we have a large-scale multimodal dataset, machine learning methods based on deep neural networks are reasonable approaches to create the material FM. Particularly, neural networks based on successful encoder-decoder architectures in natural language (Vaswani et al. 2017) and images (Kingma and Welling 2014) are promising in chemistry and material science e.g., (Gómez-Bombarelli et al. 2018).

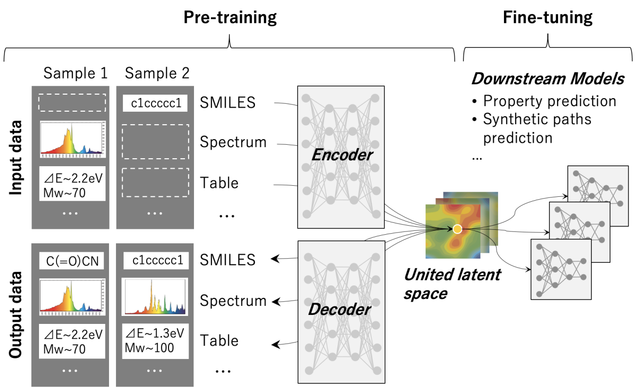

Figure 1: Overview of FM architecture for material science

Given a training dataset, an encoder-decoder-based neural network first encodes its input to learn an effective representation in a feature space (or latent space) with reduced dimensions. It then decodes the encoded input, returning the decoded result as output that is compared against a target for training. One form of training is to attempt to generate the output identical to the input. For language translation, the input and output are a description in a source language and its translation into a target language, respectively.

Creating the material FM on the basis of an encoder-decoder-based neural network is a rational choice, since it can reuse the feature representation of the encoder across different downstream tasks. We discuss a few approaches for the material FM.

One straightforward approach is to train a neural network that encodes all elements of each material to a latent space that is decoded to the identical elements. Obviously, this approach has a high dimension of the input space that is difficult to learn as well as a serious issue caused by a large number of missing data points across different modalities.

A more reasonable approach is first to prepare several sets of neural networks (called the *unimodal neural networks* in this paper) each of which learns for a feature representation of its corresponding, specific modality. Another neural network (called the *unification neural network*) then attempts to learn commonalities among these feature representations of the unimodal neural networks as a *united latent space*.

A schematic diagram of the FM architecture is exhibited in Figure 1. Here, we propose development of the material FM that can encode input data including several missing modalities to feature representations united together, and decode them to a completed set of modalities. The downstream models can also receive those representations on the united latent space as their input to address their downstream tasks.

In training unimodal neural networks, there are various knowledge representations, including a text-based graph topological representation (i.e., SMILES), a tabular representation (e.g., physical properties such as HOMO-LUMO and images (e.g., microscopic data). There are many approaches to learn unimodality, which can be further improved by follow-up research. For example, a graph convolution neural network is one way to learn feature representation on graph structures of molecules (Altae-Tran et al. 2017), while deep generative models are another way by regarding SMILES as text in language (Gómez-Bombarelli et al. 2018). In general, variants of Transformer (Vaswani et al. 2017) are strong candidates because of Transformer's promise in many domains including language (Devlin et al. 2019; Vaswani et al. 2017), vision (Khan et al. 2022), graphs (Yun et al. 2019), and chemical reaction predictions (Schwaller et al. 2018).

Although we envision that dealing with more than two modalities is the key to realize the material FM, starting with two modalities is a reasonable choice because of the literature available in the AI research community.

Given a training example represented as feature vectors in two feature spaces S_1 and S_2, contrastive learning (Jaiswal et al. 2020) attempts to group feature vectors in S_1 and S_2 closer if those vectors represent similar examples, and vice versa. CLIP (Radford et al. 2021) is based on contrastive learning to contrast natural language text and images. Other related approaches include contrastive-loss-based alignment of different feature vectors in one latent space (Nakayama and Nishida 2017). This work was recently expanded to support three modalities of text, audio and video (Alayrac et al. 2020) and implemented to Transformer (Akbari et al. 2021).

Contrastive learning has just started being applied to the tasks in chemistry, such as relating molecule names in IUPAC to their SMILES representations (Guo et al. 2022) and SMILES to a three-dimensional representation (Liu et al. 2022). Investing in ideas behind contrastive learning is one way to embody the unification neural network. A recent algorithm to predict masked latent representations is another promising approach (Baevski et al. 2022).

Given source and target sequences of tokens, the attention mechanism in neural networks, which is a successful factor for Transformer (Vaswani et al. 2017), attempts to identify relations between source tokens and target tokens. Representing the unification neural network as an attention mechanism for feature vectors of two different modalities is another approach, since attention can calculate how elements in a feature vector in S_1 contribute to represent those in S_2. This approach has been studied in the context of cross-modal representation alignments to jointly model language and vision. Tan and Bansal (2019) and Lu, D. Batra, and Lee (2019) report to compute those cross-modal attention by key-query dot product between different modalities, while Ye et al. (2019) and Kamath et al. (2019) report to process concatenated multi modal representations by a single Transformer. See (Khan et al. 2022) for a comprehensive survey.

Another approach other than contrastive learning to deal with multi missing modal data will be an extension of masked language model (Devlin et al. 2019) to multi modalities. Masking partial or full tokens of an input modality (e.g. SMILES, properties table, etc.), the FM should be trained so to predict the masked tokens. This approach is simple but careful masking strategy by for example curriculum training will be needed.

Since creating an FM falls into a task of discovering effective neural network architectures, other AI-based approaches that achieve this goal are useful in general. For example, Neural Architecture Search is an important subject to automate some of the architecture creation tasks (Benmeziane et al. 2021). Improving the performance of the decoder is formulated as a heuristic search problem addressed by new heuristic search that is more adaptive than standard beam search, e.g., (Freitag and Al-Onaizan 2017) .

One final note is that the dataset contains noisy data points whose distributions are difficult to estimate due to various approaches for constructing the data. This gives the AI research community opportunities to develop algorithms robust to such unpredictable noises.

Downstream Tasks

Once the material FM is made, it can deal with several essential downstream tasks in a more uniform way.

One attractive capability is to model the missing modalities of a material at a time. This enables material scientists to obtain comprehensive information on a material of interest. This task is addressed by encoding represented knowledge of an input material with limited available modalities to a feature vector incorporating information about the modalities of all the prepared material samples.

One technical challenge is how to effectively leverage the unification neural network. This is related to the approach on autoregressive and diffusion models that leverage the latent spaces trained by contrastive learning (Ramesh et al. 2022). For transformer-based models, word masking (Devlin et al. 2019) is related.

The material FM can support a downstream task to complete missing modalities, which eventually equals to modality conversion; predictive or generative modeling task. For example, given an IR spectrum visually drawn by a material scientist, the FM can generate candidates of molecular structures that are represented in SMILES as well as meet that IR spectrum. After the IR spectrum is transformed into numbers (see the discussion in the Data Preparation subsection), this task is addressed by performing encoding and decoding steps in the direction from that IR spectrum to its corresponding molecular structure. In a similar manner, other variety of modality conversions; property to molecule, text to molecule, molecule to property are realized.

In practice, molecular generation algorithms need to account for structural constraints of molecules in each material domain (e.g. structural symmetry, inclusion of a backbone structure, tuning of functional groups, etc.), as well as experimental conditions (e.g. temperature profile of polymerization, etc.). The material FM needs to provide an effective framework combined with other approaches that handle constraints, e.g., (Lim et al. 2020; Takeda et al. 2020).

Modeling across different material domains is another notable capability ensured for downstream tasks. For example, accurately modeling electronic conjugated systems of polymers is key to successfully design new conductive polymers with high electric conductivity. Even if no training example is available for the polymers, the material FM pretrained with the electronic conjugated systems of other non-polymer materials (e.g., small organic molecule and semiconductor) can leverage generic features necessary to predict the electronic conjugated systems for a polymer of interest. With a smaller number of available training examples, which is often the case in the material industry in practice, the model can be tuned further.

Conclusion

In this paper, we proposed building a Foundation Model (FM) for material science, which is trained with massive data acquired across a broad range of data modalities and material domains. We identified existing issues and argued on required technologies from the viewpoints of data preparation, model development, and downstream tasks. In material science, incorporation of multiple representations of material is the key for accurate modeling. The FM for material science will integrate those representations on a united latent space, so that overarching modeling across different disciplines the same as human scientists do will be achieved.

Finally, integrating materials domains and modalities does not necessarily indicate that only one absolute FM should exist; as is seen in the NLP domain, the emergence and culling of various models will occur in developing the material FM, contributing to accelerate material science.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *NeurIPS*.

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; Fauw, J. D.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 25–37.

Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; and Pande, V. 2017. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science*, 3(4): 283–293.

Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *PMLR*, volume 162, 1298–1312.

Benmeziane, H.; Maghraoui, K. E.; Ouarnoughi, H.; Niar, S.; Wistuba, M.; and Wang, N. 2021. A Comprehensive Survey on Hardware-Aware Neural Architecture Search. Arxiv preprint: arXiv:2101.09336.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; and *et al.*, S. A. 2021. On the Opportunities and Risks of Foundation Models. ArXiv preprint arXiv:2108.07258.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. ArXiv preprint arXiv:2005.14165.

Campbell, M.; Hoane Jr., A. J.; and Hsu, F. 2002. Deep Blue. *Artificial Intelligence*, 134(1–2): 57–83.

Cliche, M.; Rosenberg, D. S.; Madeka, D.; and Yee, C. 2017. Scatteract: Automated Extraction of Data from Scatter Plots. In *ECML/PKDD*, volume 1, 135–150.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 4171–4186.

Dima, A.; Bhaskarla, S.; Becker, C.; Brady, M.; Campbell, C.; Dessauw, P.; Hanisch, R.; Kattner, U.; Kroenlein, K.; Newrock, M.; Peskin, A.; Plante, R.; Li, S.-Y.; Rigodiat, P.-F.; Amaral, G. S.; Trautt, Z.; Schmitt, X.; Warren, J.; and Youssef, S. 2016. Informatics Infrastructure for the Materials Genome Initiative. *The Journal of The Minerals, Metals and Materials Society*, 68: 2053–2064.

Eyke, N. S.; Green, W. H.; and Jensen, K. F. 2020. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry and Engineering*, 5: 1963–1972.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3): 59–79.

Freitag, M.; and Al-Onaizan, Y. 2017. Beam Search Strategies for Neural Machine Translation. In *Proceedings of the 1st Workshop on Neural Machine Translation*, 56–60.

Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; ns, M. P. M.; Overington, J. P.; Papadatos, G.; Smit, I.; and Leach, A. R. 2017. The ChEMBL Database in 2017. *Nucleic Acids Research*, 45(D1): D945–D954.

Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; and Aspuru-Guzik, A. 2016. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15: 1120–1127.

Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2): 268–276.

Guo, Z.; Sharma, P.; Martinez, A.; Du, L.; and Abraham, R. 2022. Multilingual Molecular Representation Learning via Contrastive Pre-training. In *ACL*, 3441–3453.

Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; and Sayle, R. A. 2020. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 60(12): 6065–6073.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2020. A Survey on Contrastive Self-Supervised Learning. arxiv Preprent arXiv:2011.00362.

Jha, D.; Gupta, V.; Ward, L.; Yang, Z.; Wolverton, C.; Foster, I.; k. Liao, W.; Choudhary, A.; and Agrawal, A. 2021. Enabling deeper learning on big data for materials informatics applications. *Scientific Reports*, 11(4244).

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2019. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*, 1780–1790.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in Vision: A Survey. *ACM Computing Surveys*, 1–38.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; and Bolton, E. E. 2021. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1): D1388–D1395.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; and Wolverton, C. 2015. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(15010).

Kishimoto, A.; Buesser, B.; Chen, B.; and Botea, A. 2019. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning. In *NeurIPS*, 7224–7234.

Lim, J.; Hwang, S.-Y.; Moon, S.; Kimb, S.; and Kim, W. Y. 2020. Scaffold-based molecular design with a graph generative model. *Chemical Science*, 11: 1153–1164.

Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; and Olsen, B. D. 2019. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Central Science*, 5(9): 1523–1531.

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*.

Liu, Y.; Zhao, T.; Ju, W.; and Shi, S. 2017. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3): 159–177.

Lowe, D. M. 2012. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. thesis, University of Cambridge.

Lu, J.; D. Batra, D. P.; and Lee, S. 2019. Vilbert: Pretraining taskagnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 13–23.

Manica, M.; Cadow, J.; Christofidellis, D.; Dave, A.; Born, J.; Clarke, D.; Gaetan, Y.; Teukam, N.; Hoffman, S. C.; Buchan, M.; Chenthamarakshan, V.; Donovan, T.; Hsu, H. H.; Zipoli, F.; Schilter, O.; Giannone, G.; Kishimoto, A.; Hamada, L.; Padhi, I.; Wehden, K.; McHugh, L.; Khrabrov, A.; Das, P.; Takeda, S.; and Smith, J. R. 2022. GT4SD: Generative Toolkit for Scientific Discovery. arxiv Preprint arXiv:2207.03928.

Nakata, M.; and Shimazaki, T. 2017. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling*, 57(6): 1300–1308.

Nakayama, H.; and Nishida, N. 2017. ero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation Journal*, 31: 49–64.

NIST 1996. 1996. Chemistry WebBook. https://webbook.nist.gov/chemistry/. Accessed: 2022-08-30.

Noh, J.; Gu, G. H.; Kim, S.; and Jung, Y. 2020. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 11: 4871–4881.

Park, N. H.; Zubarev, D. Y.; Hedrick, J. L.; Kiyek, V.; Corbet, C.; and Lottier, S. 2020. A Recommender System for Inverse Design of Polycarbonates and Polyesters. *Macromolecules*, 53(24): 10847–10854.

Probst, J.; Howes, P.; Arosio, P.; Stavros, S.; and DeMello, A. 2021. Broad-Band Spectrum, High-Sensitivity Absorbance Spectroscopy in Picoliter Volumes. *Analytical Chemistry*, 93(21): 7673–7681.

Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J.; and Curioni, A. 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(84).

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 8748–8763.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1: 140022.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In *ICML*, 8821–8831.

Ray, A. 2021. Autonomous Materials Discovery Is Changing How We Know Material Sciences As A Field. *Analytics India Magazine*. Accessed: 2022-8-30. Available at https://analyticsindiamag.com/autonomous-materials-discovery-is-changing-how-we-know-material-sciences-as-a-field/.

Schreck, J. S.; Coley, C. W.; and Bishop, K. J. M. 2019. Learning Retrosynthetic Planning through Simulated Experience. *ACS Central Science*, 5(6): 970–981.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2018. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9): 1572–1583.

Segler, M. H. S.; Preuss, M.; and Waller, M. P. 2018. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature*, 555: 604–610.

Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; T.Green; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; and Hassab, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature*, 577: 706–710.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin–Madison.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–489.

Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; and Roitberg, A. E. 2018. Less is more: Sampling chemical space with active learning. *Journal of Chemical Physics*, 148(24): 241733.

Takeda, S.; Hama, T.; Hsu, H.-H.; Piunova, V. A.; Zubarev, D.; Sanders, D. P.; Pitera, J. W.; Kogoh, M.; Hongo, T.; Cheng, Y.; Bocanett, W.; Nakashika, H.; Fujita, A.; Tsuchiya, Y.; Hino, K.; Yano, K.; Hirose, S.; Toda, H.; Orii, Y.; and Nakano:, D. 2020. Molecular Inverse-Design Platform for Material Industries. In *KDD*, 2961–2969.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 5100–5111.

Toledo-Marin, J. Q.; Fox, G.; Sluka, J. P.; and Glazier, J. A. 2021. Deep Learning Approaches to Surrogates for Solving the Diffusion Equation for Mechanistic Real-World Simulations. *Frontiers in Physiology*, 12(667828).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1): 31–36.

Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; and Yoshida, R. 2019. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Computational Materials*, 5(66).

Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; and Tsuda., K. 2017. ChemTS: an efficient python library for de novo molecular generation. *Science and Technology of Advanced Materials*, 18(1): 972–976.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. In *CVPR*, 10502–10511.

Yun, S.; Jeong, M.; Kim, R.; Kang, J.; and Kim, H. J. 2019. Graph Transformer Networks. In *NeurIPS*, 11960–11970.