JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

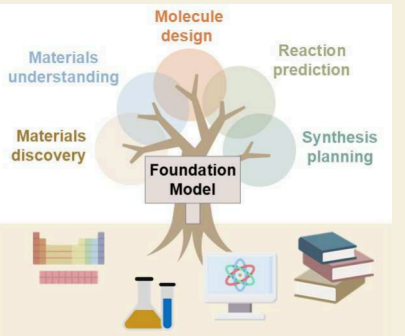
pubs.acs.org/jacsau
Perspective

A Perspective on Foundation Models in Chemistry

Junyoung Choi, Gunwook Nam, Jaesik Choi, and Yousung Jung*

Cite This: JACS Au 2025, 5, 1499−1518
Read Online
ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Foundation models are an emerging paradigm in artificial intelligence (AI), with successful examples like ChatGPT transforming daily workflows. Generally, foundation models are large-scale, pretrained models capable of adapting to various downstream tasks by leveraging extensive data and model scaling. Their success has inspired researchers to develop foundation models for a wide range of chemical challenges, from materials discovery to understanding structure−property relationships, areas where conventional machine learning (ML) models often face limitations. In addition, foundation models hold promise for addressing persistent ML challenges in chemistry, such as data scarcity and poor generalization. In this perspective, we review recent progress in the development of foundation models in chemistry across applications of varying scope. We also discuss emerging trends and provide an outlook on promising approaches for advancing foundation models in chemistry.



KEYWORDS: *foundation model, property prediction, machine learning potentials, inverse design, large-scale, pretraining, downstream tasks*

1. INTRODUCTION

Artificial intelligence (AI) has achieved remarkable success in fields such as computer vision (CV) and natural language processing (NLP), motivating scientists and researchers to explore its potential in chemical and materials science.[1] For instance, machine learning (ML) models are now used to predict molecular and material properties.[2−6] Computational chemists use ML models as surrogate potentials, known as machine learning interatomic potentials (MLIPs), to accelerate computationally demanding *ab initio* molecular dynamics (AIMD) simulations.[7−11] Another growing area is inverse design, where generative models are used to directly generate novel molecules and compounds with desired properties.[12−18] Beyond these, ML has been applied to predict drug interaction,[19,20] synthesis routes,[21−23] reaction products,[24,25] and analyze X-ray diffraction patterns.[26,27]

One of the central challenges for ML in chemistry is the scarcity of large, labeled data.[28] Unlike CV and NLP, where vast amounts of annotated data are readily available, chemical and materials science datasets are often limited and require labor-intensive experiments or enormous computing resources. Moreover, applications like drug discovery or materials design require extrapolation to out-of-domain compounds, another limitation of deep learning models. These challenges limit conventional ML approaches at a practical level. In addition, traditional MLIPs often rely on deliberately curated datasets crafted with domain expertise to avoid extrapolation, limiting their transferability across different systems and thus restricting broader adoption.

Recently, the concept of foundation models has emerged as a new AI paradigm. Foundation models are large-scale, pretrained models that can adapt to a broad range of downstream tasks.[29] By training on vast datasets, foundation models learn general representations that can be shared across different tasks and domains. These models are then adapted to downstream tasks through transfer learning or finetuning. For instance, foundation models in NLP may handle text translation and summarization, while in CV, they are adapted for tasks like image classification and captioning. In the context of chemistry and materials science, a foundation model can be adapted to predict various properties of molecules and crystals,[30] or generate novel compounds with a desired property (Figure 1).[16] Another application is a foundational MLIP, which can be transferred to a wide range of systems and can be further finetuned to a target system when necessary.[31] Recently, large language models (LLMs) such as ChatGPT[32] have been leveraged to solve chemical problems or finetuned for specific tasks.[33,34] Owing to their flexibility and generalizability, foundation models offer not only improved performance over models trained from scratch but also potential solutions to the aforementioned challenges such as data scarcity and robust extrapolation.[35−38] Moreover, multimodal

Received: November 30, 2024
Revised: February 7, 2025
Accepted: February 7, 2025
Published: March 25, 2025

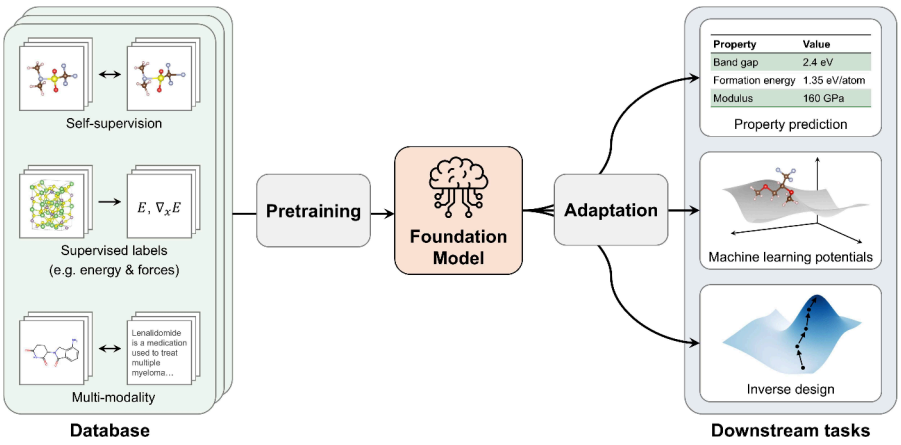


Figure 1. Overview of a foundation model in chemistry for property prediction, machine learning interatomic potentials, and inverse design.

foundation models[39] in chemistry leveraging molecular structures and text would provide a user-friendly platform for text-based inverse design of molecules, which was beyond the capabilities of the conventional ML models.[40,41] However, research related to foundation models is evolving rapidly and extensively, making it challenging to track recent trends and identify promising approaches. Therefore, it is an appropriate time to summarize current efforts and provide an outlook on foundational models in chemistry.

The precise definition of the foundation model in chemistry has not been established, and it may differ depending on the way the scope of downstream tasks is considered. Broadly speaking, a pretrained prediction model that can be adapted to various property prediction tasks (e.g., band gap, formation energy, etc.) can be considered a foundation model for a property prediction domain. In a narrower sense, a foundation model should be able to span multiple applications such as prediction of properties and PES, generation of data, etc. While the latter may be considered closer to an ultimate form of the foundation model, studies on the former focusing on pretraining methods that work well in each domain are also crucial for developing pretraining strategies for the latter. In this perspective, therefore, we consider both definitions, referring to the former as a *small* foundation model and the latter as a *big* foundation model. First, we review the current progress in the *small* foundation model for three respective domains: (1) property prediction, (2) MLIP, and (3) inverse design (Table 1). Then, we look into *big* foundation models adaptable to multiple domains in the (4) multi-domain sections (Table 1). Before this, we provide an overview of common methodologies investigated for foundation models in chemistry. Finally, we outline future directions and opportunities.

2. METHODOLOGY

2.1. Models

2.1.1. Graph Neural Network. Conventional machine learning models often require hand-crafted descriptors to capture input−output relationship.[5,42−46] Designing an effective descriptor requires domain expertise on the target task, and often involves a labor-intensive feature selection process to identify the best-working descriptors. In the case of MLIPs, most conventional approaches are descriptor-based, with the descriptors designed to capture information on the local environment constructed by the elements specific to the target system.[47,48] As a result, this approach limits the transferability of trained MLIPs to other systems.

In contrast, deep learning models using neural networks are known to learn the representation of the input data that best describes the underlying relationship between input and output, generally outperforming the descriptor-based models.[49,50] Among these, graph neural networks (GNN) have gained significant attention for their high expressivity and suitability for molecular data,[51] making them promising candidates for foundational models in chemistry. In a GNN, molecules or crystals are represented as graphs $G = (V, E)$, where atoms and bonds are treated as nodes V and edges E, respectively.[52] The graphs are typically featurized with node features v_i^0, such as atom types, and edge features e_{ij}^0, such as bond types and lengths, where i and j denote the node indices of a graph.[51] Node features are then updated via a message-passing framework, where the message is constructed based on the environment of the central node or edge.[53] Specifically, the message m_i^{t+1} for node i at the $(t+1)^{\text{th}}$ message-passing step is obtained in general as

$$m_i^{t+1} = \sum_{j \in N(i)} M_t(v_i^t, v_j^t, e_{ij}^t) \tag{1}$$

where $N(i)$ denotes the number of nodes in the graph, and $M(t)$ is a message function. Then, node features are updated via an update function U_t:

$$v_i^{t+1} = U_t(v_i^t, m_i^{t+1}) \tag{2}$$

After T message-passing steps, a readout function R is applied to the updated node features to obtain a graph-level feature g:

$$g = R(\{v_i^T | i \in G\}) \tag{3}$$

Then g is projected onto multilayer perceptrons (MLP) to output a prediction. It should be noted that the update functions M_t and U_t, and the readout function R can be approximated by learnable neural networks, enabling the effective learning of complex interactions between atoms. Moreover, the recent equivariant GNN enables the use of vectorial features richer in geometric information such as positions, thereby enhancing the expressive power of the GNN.[54−56] As a result, they have been actively explored as MLIPs, where the predictions (e.g., energy and forces) are sensitive to the geometries of molecules and crystals.[57−60]

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

Table 1. Summary of Foundation Models in Chemistry

Domain	Model	Architecture	Pretraining		Downstream task
			Data	Method[a]	
Property prediction	GraphCL[105]	GIN[106]	ZINC15[107] (2M)	CL (aug.)	Molecular property prediction
	MolCLR[108]	GCN,[109] GIN[106]	PubChem[110] (10M)	CL (aug.)	Molecular property prediction
	GraphMVP[111]	GIN,[106] SchNet[112]	GEOM[113] (50K)	CL (2D ↔ 3D)	Molecular property prediction
	Hu et al.[114]	GIN,[106] GCN,[109] GraphSAGE[115]	ZINC15[107] (2M), ChEMBL[116,117] (456K)	PL (node context), GL (node, edge), SL (property)	Molecular property prediction
	GROVER[118]	GTransformer	ZINC15,[107] ChEMBL[116] (total 11M)	PL (motif), GL (node, edge)	Molecular property prediction
	SMILES-BERT[119]	BERT[63]	ZINC[120] (18M)	GL (SMILES)	Molecular property prediction
	ChemBERTa-2[121]	RoBERTa[122]	PubChem[110] (77M)	GL (SMILES), SL (property)	Molecular property prediction
	MoLFormer[123]	Transformer[62]	PubChem[110] (111M), ZINC[124] (1B)	GL (SMILES)	Molecular property prediction
	MOFTransformer[125]	BERT[63]	In-house hMOF (1M)	PL (property)	MOF property prediction
	CT[126]	CGCNN[96]	Matminer[127] (152K), hMOF[128] (275K)	CL (aug.)	Materials property prediction
	CrysGNN[35]	CGCNN,[96] CrysXPP,[129] GATGNN,[130] ALIGNN[97]	OQMD[131] (661K), MP[132] (139K)	CL (crystal system), PL (space group), GL (node, connectivity)	Materials property prediction
	DSSL[133]	DeeperGATGNN[134]	MP[132] (138K)	CL (aug.), PL (micro-property), GL (node)	Materials property prediction
	LLM-Prop[135]	T5[136]	MP[132] (144K)	GL (text[75])	Materials property prediction
Machine learning interatomic potentials	M3GNet[137]	GNN	MPF.2021.2.8[132] (187K)	SL (E, F, S)	MD simulations and structural relaxation
	CHGNet[138]	GNN	MPtrj[132] (1.58M)	SL (E, F, S, M)	MD simulations and structural relaxation
	ALIGNN-FF[139]	ALIGNN[97]	JARVIS-DFT[140] (307K)	SL (E, F, S)	E−V calculation, structural relaxation, structure search
	PFP[141]	TeaNet[142]	PFP molecular dataset (6M), PFP crystal dataset (3M)	SL (E, F, C)	MD simulations, molecular adsorption, order−disorder transition, material discovery for catalysts
	MACE-MP-0[31]	MACE[58]	MPtrj[132] (1.58M)	SL (E, F, S)	35 applications including water, catalyst, MOF, battery cell, etc.
	MACE-OFF23[143]	MACE[58]	OFF23 (1M)	SL (E, F)	Dihedral scans, MD simulations of molecular crystals, organic liquids, etc.
	GNoME potential[144]	NequIP[57]	GNoME (89M)	SL (E, F)	Crystal structure search
	SevenNet[145]	NequIP[57]	MPtrj[132] (1.58M)	SL (E, F, S)	Melt-quench simulation
	MatterSim[36]	M3GNet,[137] Graphormer[65]	In-house data (3M, 17M)	SL (E, F, S)	Calculation of thermodynamics, lattice dynamics, and mechanical properties
	eqV2[146]	EquiformerV2[147]	OMat24 (118M)	SL (E, F, S), Denoising	Structural relaxation
Inverse design	MatterGen[16]	Diffusion	Alex-MP-20[132,148] (607K)	GL (A, X, L)	Generation of crystals with target property
	GP-MoLFormer[149]	MoLFormer[123]	PubChem[150] (111M), ZINC[124] (1B)	GL (SMILES)	Generation of molecules with target property
	CrystalLLM[151]	Pretrained LLaMA-2[152]	–	–	Generation of crystals with target property
Property prediction & MLIP	Zaidi et al.[153]	GNS[154]	PCQM4Mv2[155] (3.4M)	Denoising	Energy, force and molecular property prediction
	GeoSSL-DDM[156]	PaiNN[59]	Molecule3D[157] (1M)	Denoising	Force and molecular property prediction
	ET-OREO[158]	TorchMDNet[66]	MD17[159] (3.5M), ANI1-x[160] (5M), PCQM4Mv2[155] (3M), poly24 (3.5M)	Denoising, SL(F)	MD simulations, molecular property prediction
	3D-EMGP[158]	EGNN[54]	GEOM-QM9[113] (100K)	Denoising	Force and molecular property prediction
	Frad[161]	TorchMDNet[66]	PCQM4Mv2[155] (3.4M)	Denoising	Energy, force and molecular property prediction
Property prediction and inversedesign	KV-PLM[162]	BERT[63]	S2orc[163] (0.3M papers, 1B tokens), PubChem[164]	Multimodal learning (SMILES, text)	Molecular property prediction, reaction classification, SMILES-description retrieval
	MoMu[165]	SciBERT,[166] KV-PLM,[162] GraphCL[105]	PubChem,[164] S2orc[163] (15K graph-text pairs)	Multimodal learning (graph, text)	Graph-description retrieval, molecule captioning, text-to-graph generation, molecular property prediction

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

Table 1. continued

Domain	Model	Architecture	Pretraining		Downstream task
			Data	Method[a]	
	MolFM[167]	KV-PLM,[162] GraphMVP,[111] TransE[168]	PubChem,[164] S2orc,[163] DrugBank[169] (15K graph-text pairs), knowledge graphs (E49K, R3.2M)	Multimodal learning (graph, text, knowledge graph)	Graph-description retrieval, molecule captioning, text-to-graph generation, molecular property prediction
	MoleculeSTM[40]	MegaMolBART,[170] GraphMVP,[111] SciBERT	PubChemSTM[164] (281K structure-text pairs)	Multimodal learning (SMILES/graph, text)	Structure-text retrieval, text-based molecule editing, molecular property prediction
	SPMM[41]	BERT[63]	PubChem[171] (50M)	Multimodal learning (SMILES, property)	Property-to-SMILES generation, molecular property prediction, reaction prediction
	ChemDFM[172]	Pretrained LLaMa-13B[173]	Chemical books (1.4K), papers (3.9M), general text	Language modeling	Molecule recognition, text-to-SMILES generation, molecular property prediction, reaction prediction
	nach0[174]	T5[136]	Text from PubMed (13M, 355M tokens), USPTO (119K, 2.9B tokens), ZINC (100M, 4.7B tokens)	Language modeling	14 tasks including molecular property prediction, reaction prediction, text-to-SMILES generation, etc.
	Jablonka et al.[175]	Pretrained GPT-3[32]	–	–	Molecular/materials property prediction, text-to-SMILES generation
	AtomGPT[176]	Pretrained GPT-2[177]	–	–	Materials property prediction
		Pretrained Mistral 7B[178]	–	–	Text-to-materials generation

[a]CL: contrastive learning; aug.: augmentation; PL: predictive learning; GL: generative learning; SL: supervised learning; E: energy; F: forces; S: stress; M: magnetic moments; A: atom types; X: positions; L: lattice.

2.1.2. Language Model. A language model is a key area in NLP that aims to learn and predict the likelihood of word sequences.[61] The development of BERT by Google, built on transformer architecture,[62] marked the beginning of the foundation model era by showcasing that a single model could perform multiple tasks.[29,63] Furthermore, scaling up language models in both model size and data volume has revealed new capabilities, or "emergent abilities", enabling them to solve complex tasks that were previously out of reach.[64] As a result, large language models (LLMs) like ChatGPT[29] have gained widespread popularity in various fields where tasks can be expressed in human languages, such as language translation, research assistance, and more.

An autoregressive language model such as GPT aims to predict the next token y in a sequence given a context X.[61] The context consists of the preceding tokens $x_1, x_2, ..., x_{t-1}$, where t denotes the current position in the sequence. The model is trained by maximizing the conditional probability of the token sequence, expressed as $P(y|X) = P(y|x_1, x_2, ..., x_{t-1})$. Using the chain rule, this objective can be further decomposed into the product of probabilities:[61]

$$\prod_{t=1}^{T} P(y_t|x_1, x_2, ..., x_{t-1}) \tag{4}$$

where T is the sequence length.

One of the most successful language models, a transformer, leveraged self-attention mechanism eliminating the need for recurrent neural networks (RNNs) or convolution.[62] Specifically, the attention is computed by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, and V are the query, key, and value matrices, respectively. To capture rich and intricate information from the input, multi-head attention is used in the transformer, which applies the attention mechanism across h independent subspaces:[62]

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{6}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{7}$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$, d_k and d_v are dimensions of query/key and values, respectively, and $d_{\text{model}}/h = d_k$. In addition, since the transformer does not rely on RNNs or convolution, positional encoding (PE) is used to make aware of the position of a token within the sequence:[62]

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$
$$\text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{8}$$

The transformer is known for its effectiveness in capturing the relative importance and long-range dependencies of each element in a sequence, making it the mainstream model in language models today. In fact, this has also led to the development of GNNs inspired by transformer architecture to effectively learn interactions between nodes.[65−67]

With the success of language models in NLP, efforts to apply these models in chemistry have grown extensively.[68−71] A crucial aspect of this application is identifying effective string-based representations for input. Molecules can be readily represented as a string such as the Simplified Molecular Input Line System (SMILES)[72] and Self-referencing embedded strings (SELFIES),[73] enabling active use of language models in chemistry.[74] For example, transformers have been applied almost directly with minimal modification to predict chemical reactions represented by the rearrangement of SMILES from reactants to products.[25] Conversely, in materials science, the lack of a standardized text-based representation for crystalline structures limits the broader application of language models. However, tools such as Robocrystallographer[75] and ChemNLP[76] have been developed to generate text descriptions for crystals. Additionally, Xiao et al. proposed a Simplified

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

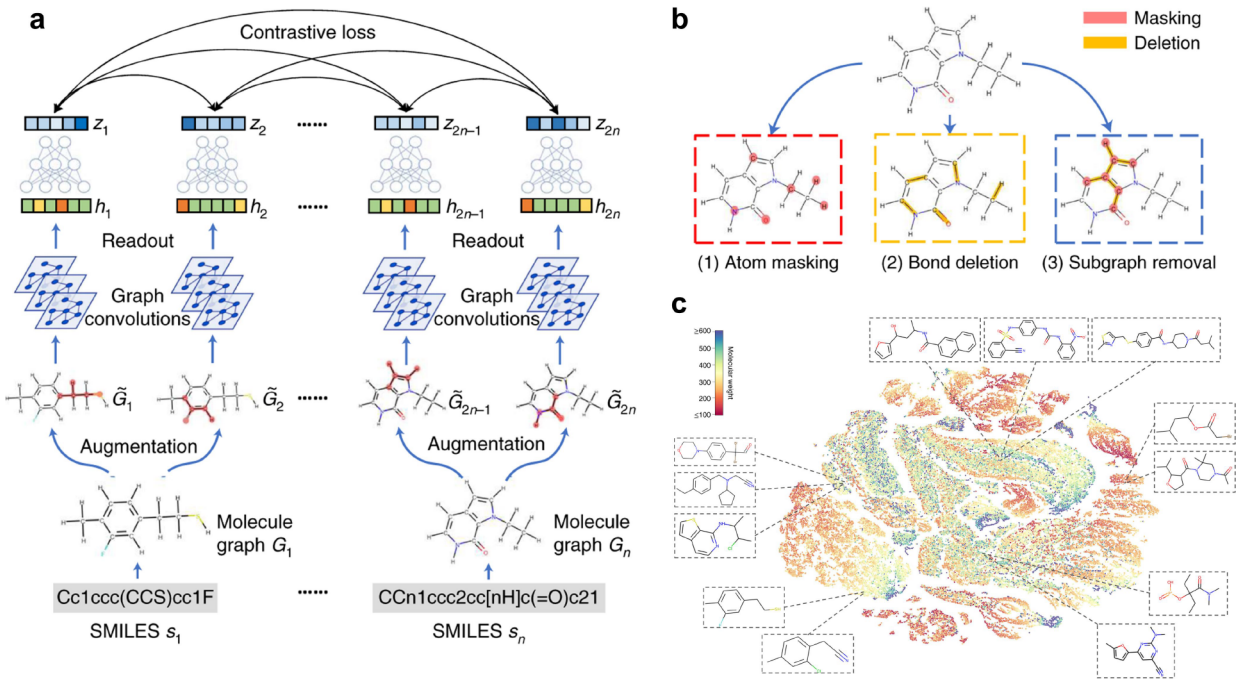


Figure 2. Contrastive learning proposed in MolCLR. Reprinted with permission from ref 108. Copyright 2022 Springer Nature. **a.** Schematic of contrastive learning and **b.** proposed augmentation methods for molecular graphs. **c.** Learned features via contrastive learning visualized by t-SNE.

Line-Input Crystal-Encoding System (SLICES), which offers a more compact and invertible string representation for crystals.[77,78] These advancements are driving increased research into language models for crystalline materials. As a result, large language models have become another key foundation model in the field of chemistry.

2.2. Pretraining Methods

2.2.1. Self-Supervised Learning. A prominent pretraining approach for the foundation model is self-supervised learning (SSL), which leverages the inherent structure of data as a source of supervision without requiring human-labeled data.[79] While supervised learning relies on labeled data which is often costly and time-consuming to generate, SSL enables the use of abundant unlabeled data to learn intrinsic data representations that can be finetuned for downstream tasks with a smaller amount of labeled data. Indeed, most successful foundation models including those based on transformer architectures were pretrained using SSL.

SSL is often categorized into three approaches: contrastive, predictive, and generative learning.[79,80] Contrastive learning has been actively explored in CV and NLP domains.,[81−84] where representations are paired as either positive or negative, with the model learning to maximize similarity between positive pairs while pushing negative pairs apart. The pairing method is crucial and must be designed carefully to ensure the learned representation is relevant to downstream tasks and avoids negative transfer.[80,81] Predictive learning is known as a method that predicts self-generated informative labels from the data, while generative learning refers to reconstructing a graph or text.[79,80] While contrastive learning focuses on inter-data information, these two learning schemes aim to extract intra-data features.[80] In the following sections, we review the approaches to adapting these SSL methods for foundation models in chemistry, and for more general and theoretical discussions and benchmarks on SSL, we refer readers to refs 30, 79, 80.

2.2.2. Multimodal Learning. Multimodal learning aims to capture complementary information from multiple sources or modalities, similar to how humans perceive voice while seeing the speaker.[85,86] This approach has achieved remarkable success in vision-and-language pretraining (VLP).[87−90] For instance, Contrastive Language-Image Pretraining (CLIP) learns a common embedding space for text and images by jointly training respective encoders,[87] such as a transformer for text and a convolutional neural network for images.[91] Specifically, contrastive learning is used to maximize the similarity of embeddings from the real text-image pairs while minimizing it for incorrect pairs, enabling zero-shot image classification by comparing the similarity between the query image and label texts. The success of VLP has encouraged researchers to explore multimodal learning in chemistry, becoming a promising approach to the development of *big* foundation models capable of handling diverse downstream tasks across two modalities (Section 3.5.1). More discussion on multimodal learning for foundation models in materials science can be found in Takeda et al.,[92] to which we refer interested readers.

3. APPLICATIONS

3.1. Property Prediction

Obtaining molecular and material properties through experimental or computational methods like quantum mechanical calculations is resource-intensive, motivating the development of machine learning models to predict various properties by learning structure−property relationships from data.[93−95] However, deep learning requires large datasets, which are often limited in chemistry, especially for experimental data. In addition, machine learning models often struggle with extrapolating to unseen data, a frequent requirement in chemistry such as with newly designed materials. While advancements in architectures and techniques have improved predictive performance in chemistry,[60,96−98] challenges of data

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

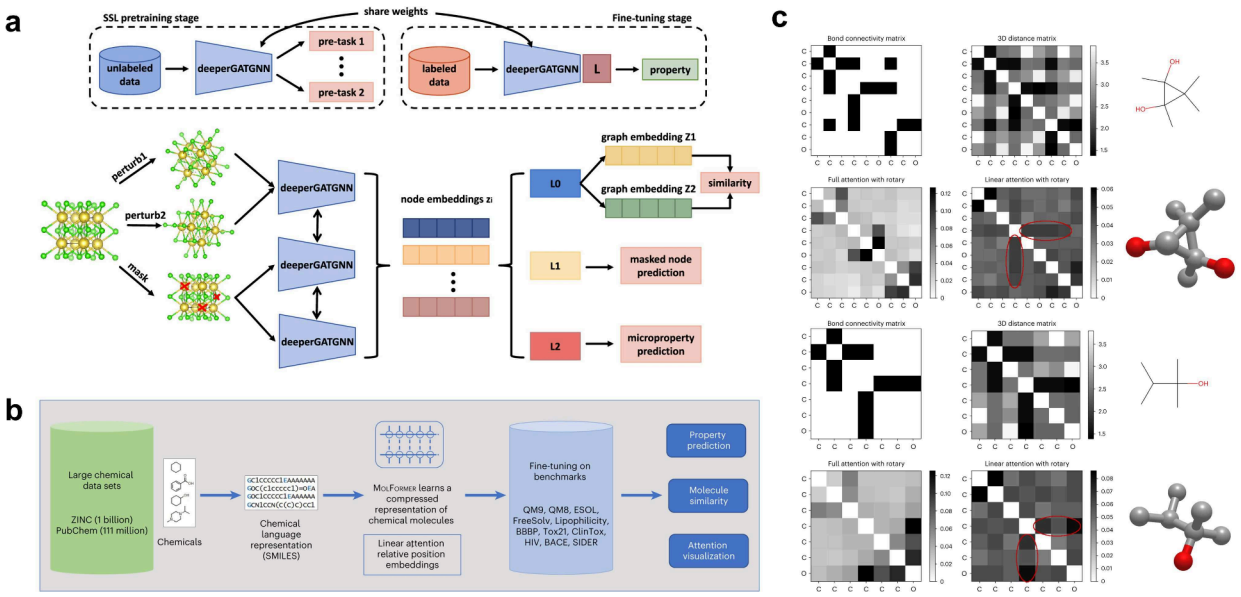


Figure 3. a. Predictive and generative learning on crystal graphs. Reprinted with permission from ref 133. Copyright 2024 American Chemical Society. **b.** Overview of MoLFormer and **c.** its attention maps, with linear attention capturing the 3D distance information from SMILES. Reprinted with permission from ref 123. Copyright 2022 Springer Nature.

scarcity and generalization remain unresolved. Transfer learning partially addresses these issues by reusing representations from large datasets, yet effective transfer depends on close alignment between pretraining and target tasks, requiring domain expertise to avoid negative transfer.[99−101] By contrast, foundation models seek to learn universal representations adaptable across diverse properties, holding the potential to address data limitations and improve generalizability, even in low-data and out-of-distribution (OOD) domains. In this section, we review the progress in *small* foundation models for property prediction, focusing on various self-supervised pretraining strategies.

3.1.1. Contrastive Learning. One of the early studies on contrastive learning for graph data was Deep Graph Infomax (DGI),[102] inspired by Deep InfoMax[83] from CV. DGI generated pairs by comparing local and global views of the graph. Specifically, it learns node embedding in such a way that the embedding of the node and its parent graph (positive pair) becomes similar while the node and the other graphs (negative pair) are dissimilar. InfoGraph[103] adopted a similar approach, with a more focus on graph-level embeddings. When combined with supervised learning, InfoGraph outperformed purely supervised learning on all 12 targets of the QM9[104] dataset and surpassed semi-supervised baselines on 11 targets.

While these methods focused on contrastive learning of local and global features within graphs,[80] more recent works have explored contrastive learning of graph-level embeddings via data augmentation (Figure 2).[105,108,111,179−181] For instance, GraphCL[105] applied data augmentation to graph data for contrastive learning of graph-level embeddings. Specifically, GraphCL introduced four data augmentation techniques: node dropping, edge perturbation, attribute masking, and subgraph sampling. During training, one augmentation method was randomly applied, treating the augmented and original graphs as a positive pair, while pairing the augmented graph with other graphs as negatives. After transfer learning, GraphCL achieved state-of-the-art (SOTA) performance on 5 of 9 tasks, including MoleculeNet[182] classification task. Given the importance of choosing suitable augmentation techniques for downstream performance, the same group further proposed an automated framework to optimize augmentation selection for pretraining.[179]

More recently, Wang et al. introduced MolCLR,[108] a contrastive learning framework for molecular graphs trained on an enlarged dataset of 10 million molecular graphs from PubChem.[110] Benchmarking on MoleculeNet classification and regression tasks showed that MolCLR achieved significant improvements compared to the supervised counterparts. Additionally, t-SNE visualization of the learned features showed that molecules with similar topologies and functional groups were clustered together (Figure 2c), demonstrating MolCLR's effectiveness in capturing meaningful molecular representations.

Recognizing that 3-dimensional (3D) molecular geometry provides richer information than 2-dimensional (2D) molecular topology, Liu et al. proposed GraphMVP, a framework that utilizes both 3D geometry and contrastive learning to enhance representation learning.[111] In GraphMVP's pretraining stage, each molecule's 2D topology and 3D geometry were paired as positives, with negatives defined as pairs of different molecules. In the downstream task, only 2D molecular graphs were used to predict properties, with the test dataset differing from the pretraining dataset. When combined with conventional 2D SSL, GraphMVP outperformed prior SSL approaches across all 8 classification and 6 regression tasks. Moreover, the authors suggested that 3D geometry served as privileged information, improving separability among molecules and accelerating convergence as explained by VC theory.[183]

Contrastive learning has also shown promise for crystalline materials. Crystals differ from molecules by combining a broader range of elements with diverse symmetries, which may necessitate different approaches. One of the earliest studies by Magar et al. proposed three augmentation methods based on random perturbations, atom masking, and edge masking.[126] Using CGCNN[96] as a baseline model, this approach showed

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

improvements in 7 out of 9 Matbench[94] benchmarks. CrysGNN, proposed by Das et al., took a different approach by defining crystals of the same (different) crystal system as positive (negative) pairs.[35] The rationale behind this lies in the fact that some electronic and optical properties such as band gap and dielectric constant depend on the spacegroup and crystal structures.

3.1.2. Predictive and Generative Learning. In chemistry, the labels for predictive learning can be contextual information of a node,[114] a motif within a molecule such as a functional group,[118] chemical properties of atoms such as electronegativity, number of valence electrons, and covalent radius,[35] or the space group.[133] For generative learning, a typical approach is to mask nodes or/and edges and reconstruct them[114,133] or to reconstruct the entire molecule or crystal,[184] which is essentially generative modeling.

Predictive and generative learning are often applied in a multi-task fashion and are sometimes integrated with contrastive or supervised learning.[35,114,118,133] For example, Hu et al. proposed a general framework to learn representations from both local and global aspects of the molecular graph.[114] The local features were learned through predictive learning on the node context and generative learning by reconstructing masked nodes and edges, while the global features were learned via supervised learning of molecular properties. Rong et al. took a similar approach but differed at the graph-level representation learning, where they suggested predicting the graph-level motif such as functional groups, rather than supervised learning.[118]

For crystalline materials, CrysGNN, mentioned earlier in Section 3.1.1, combines generative and predictive learning with contrastive learning.[35] This model reconstructs node features and connectivity at the node level and uses space group prediction to learn graph-level embeddings. In downstream tasks on the Materials Project (MP)[132] and JARVIS-DFT[140] datasets, CrysGNN improved the performance of all four models tested across various properties owing to the synergy of these SSL methods. CrysGNN also demonstrated improvements even on small experimental datasets, implying its practical utility. More recently, Fu et al. proposed using atomic-level microproperties as predictive labels in addition to contrastive learning as shown in Figure 3a.[133] Their approach is motivated by the fact that macro-properties such as elastic properties or band gaps depend on the ensemble of microproperties of local atoms, such as atomic stiffness or valence electrons. However, selecting relevant pretraining labels for downstream tasks requires domain knowledge and must be done carefully.

Language models also have been employed in terms of predictive and generative learning. For example, Wang et al. pretrained BERT through a Masked SMILES Recovery task, training the model to recover randomly masked SMILES.[119] After finetuning, this model outperformed the SOTA on tested three datasets, demonstrating the effectiveness of their pretraining approach. Similarly, Zhang et al. used BERT with SMILES augmentation by varying starting atoms and traversal orders, which were then masked for pretraining.[185] The finetuned model achieved SOTA performance on most of the 60 molecular prediction tasks, including ADMETlab[186] and MoleculeNet.

Chithrananda et al. proposed ChemBERTa, and first systematically demonstrated the potential of transformers in molecular representation learning by exploring data size, tokenizer methods, and string representations.[187] The following work by Ahmad et al. developed ChemBERTa-2,[121] investigating pretraining strategies such as masked language modeling and multi-task regression with labels computed using RDKit,[188] without the need for extra experiments. ChemBERTa-2 pretrained with multi-task regression outperformed the original model and SOTA models on some MoleculeNet tasks.

Ross et al. introduced MoLFormer, a transformer-based model pretrained on the PubChem[110] and ZINC[124] datasets, encompassing 1.1 billion SMILES using masked language modeling (Figure 3b).[123] MoLFormer was benchmarked on various downstream tasks, including 2D and 3D molecular datasets like MoleculeNet and QM9, outperforming baseline and other language models and even some GNN models. Analysis of MoLFormer's attention maps revealed that the model effectively captured molecular structural features from SMILES (Figure 3c). Kang et al. proposed MOFTransformer, a multimodal transformer for metal−organic frameworks (MOFs), combining local features derived from CGCNN with global energy grid features.[125] The pretraining task was predictive learning to predict MOF properties such as topology and void fraction and to classify metal cluster-organic linkers. MOFTransformer achieved SOTA across diverse properties, including gas adsorption, diffusion, and electronic properties, surpassing baselines including CGCNN.

Rubungo et al. developed LLM-Prop, a language model pretrained on text descriptions of crystal structures, then finetuned to predict crystal properties.[135] Using Robocrystallographer[75] to convert crystal structures into text, they pretrained a T5 model[136] with span-masking.[189] LLM-Prop outperformed GNN baselines on MP dataset benchmarks while using 35k fewer training points out of 125k. These studies underscore language models' capabilities in molecular and crystal representation learning, providing competitive results comparable to those of geometrical GNNs.

3.2. Machine Learning Interatomic Potential

In computational chemistry, the potential energy surface (PES) of a system is of significant importance as it provides valuable information about the system, such as local minima and transition states. In particular, one can sample the PES to collect meaningful points through molecular dynamics (MD) or Monte Carlo simulations. To obtain the PES, one can resort to first-principle methods such as DFT or empirical force fields, such as the Lennard-Jones potential. However, there is a trade-off between accuracy and computational cost, where the former is accurate but slow, while the latter is generally faster but less accurate. Because of this, tasks such as MD simulations, which require numerous energy and force evaluations on large cells, mostly employ empirical force fields at the expense of accuracy. In contrast, AIMD has been performed in limited applications using small cells and short time scales. This trade-off between accuracy and speed poses a challenge to the reliability of MD simulations regarding the accuracy of the potential and the rigor of the simulation setup.

Machine learning interatomic potentials (MLIPs) have emerged as an alternative that offers a compromise between these trade-offs.[8] MLIPs are predictive models trained on DFT-level total energy and its derivatives, namely forces and stress. Once properly designed and trained, MLIPs serve as both accurate and efficient surrogate potentials, thereby enhancing the reliability of MD simulations. Conventional

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

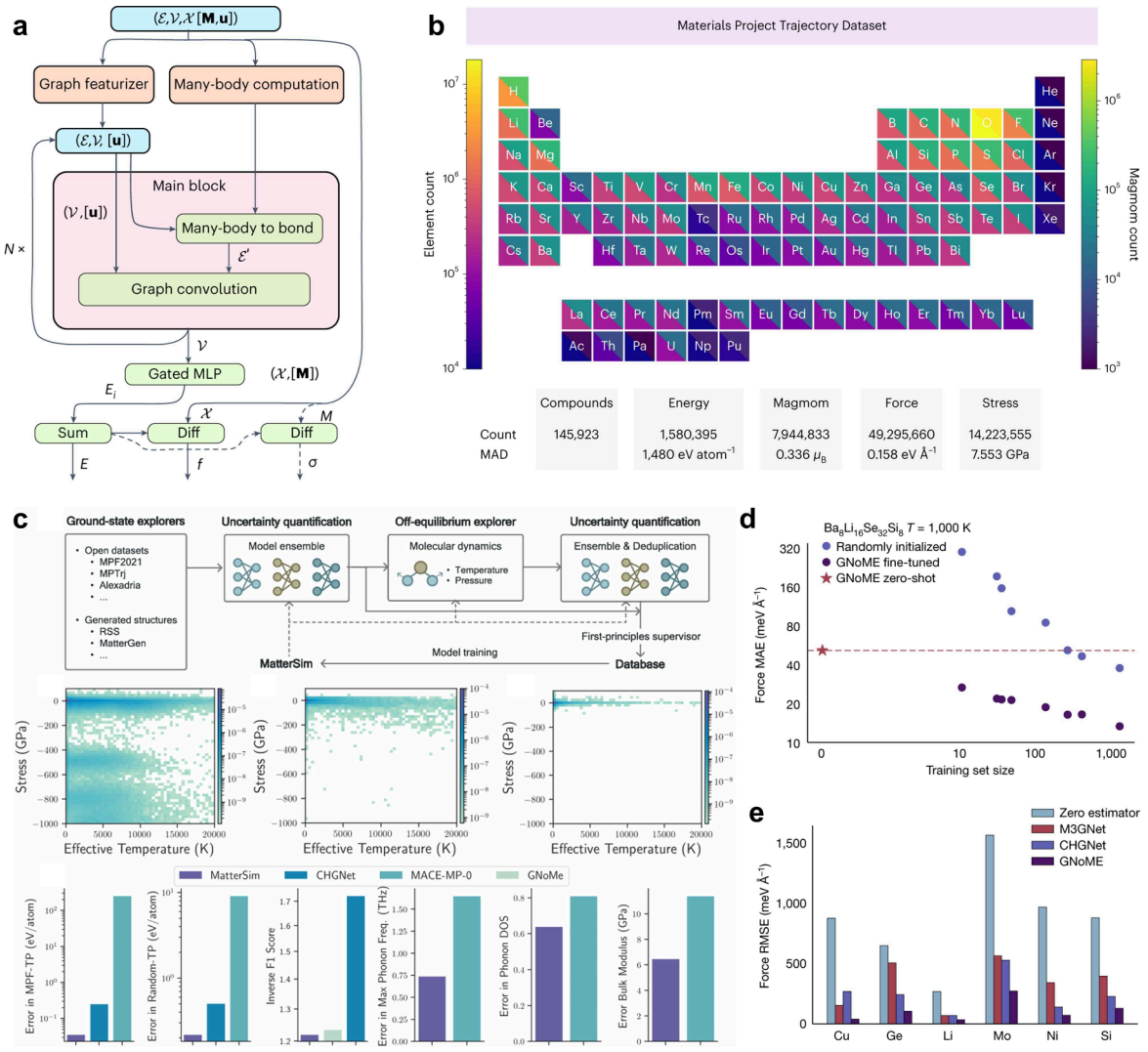


Figure 4. a. Architecture of GNN-based M3GNet. Reprinted with permission from ref 137. Copyright 2022 Springer Nature. **b.** Statistics of the MP trajectory data used to train CHGNet. Reprinted from ref 138 under the terms of the CC BY license. **c.** Data generation pipeline and statistics proposed in MatterSim and **e.** its zero-shot performance in predicting properties via simulations. Reprinted from ref 36 with permission from the authors. Copyright 2024, the authors. **d.** Finetuned and zero-shot performances of GNoME in comparison to NequIP[57] trained from scratch, and **f.** comparison of the zero-shot performances of universal MLIPs on force prediction of elemental systems. Reprinted from ref 144 under the terms of the CC BY license.

MLIPs were descriptor-based models, where the descriptors are designed to contain information about the local environment constructed from the combinations of elements in the target system. An intrinsic challenge of this approach is that the trained MLIP cannot be easily transferred to other systems. Moreover, a typical bottleneck in training MLIPs is the generation of data, which often requires an active learning scheme to ensure that the MLIP has seen all relevant configurations likely to be encountered during simulations.

Recently, MLIPs based on GNN architecture have been increasingly reported. The superior performance and descriptor-free nature of GNNs which encode the local environment of any combination of elements into a fixed-length feature vector suggest the potential for a foundational MLIP, which has indeed been reported as a "universal MLIP" (Figure 4).[31,36,137−139,141,143−145] These universal MLIPs have been trained on large databases encompassing diverse elements, making them transferable across various systems.[190,191] Chen and Ong reported the first universal MLIP, M3GNet, which can perform relaxation or MD simulations of various crystal structures.[137] They exploited relaxation trajectory data from the MP database, consisting of over 60,000 compounds and 89 elements. Deng et al. developed CHGNet, a universal MLIP capable of predicting the magnetic moment, which is essential in materials such as transition metal oxides.[138] MACE,[58] a recent equivariant MLIP, has also been employed to develop universal MLIPs for crystals (MACE-MP-0)[31] and organic molecules (MACE-OFF23).[143] In particular, MACE-MP-0 extensively investigated its transferability to a wide variety of systems including solids, liquids, surfaces, porous materials, and even battery cells. Although not always quantitatively accurate, it demonstrated qualitative universality across a wide range at unprecedented levels.[31]

These universal MLIPs not only serve as effective surrogates for tasks such as relaxation, but they also act as foundation models that facilitate the development of MLIPs tailored to specific systems through finetuning when more accurate potentials are required, such as for MD simulations. For

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

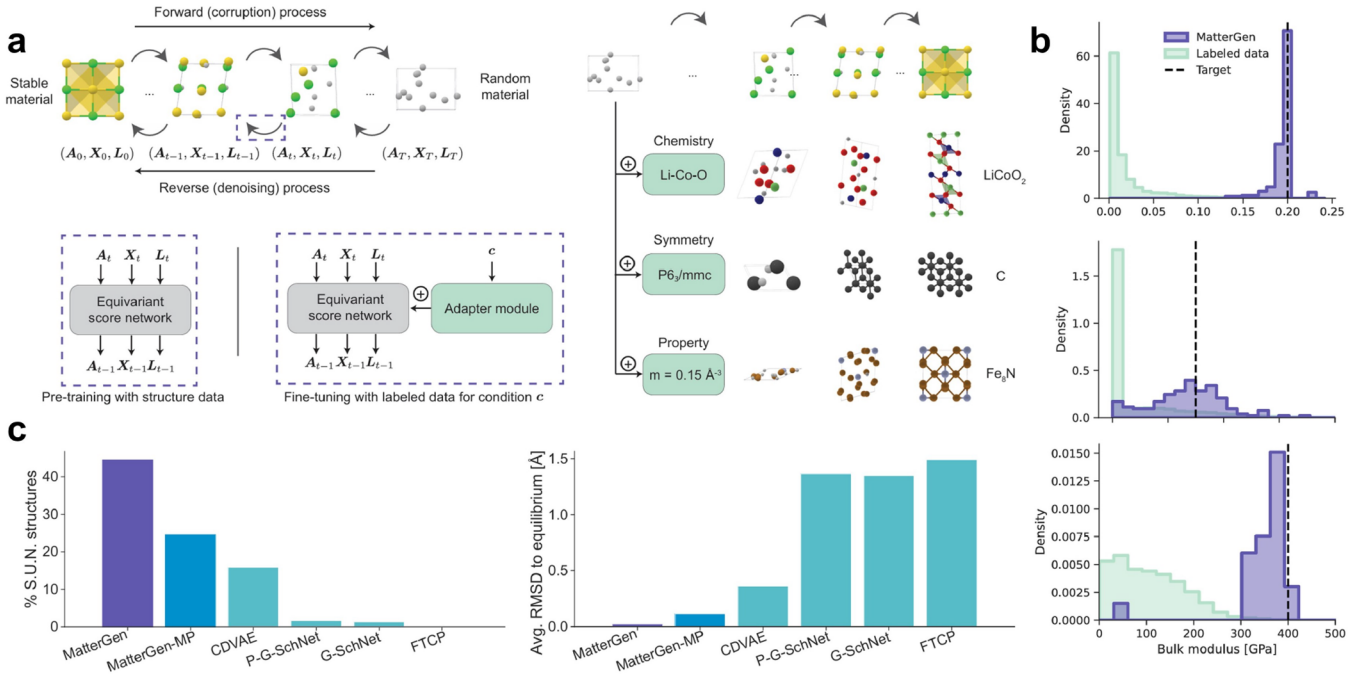


Figure 5. a. Pipeline of diffusion-based MatterGen, **b.** its performance in inverse design, and **c.** generating stable, unique, and novel (S.U.N.) materials. Reproduced from ref 16 with permission from the authors. Copyright 2023, the authors.

example, Jun et al. finetuned CHGNet to study ionic conduction in nitride Li-ion conductors, observing that the anchoring of dopants and Li vacancies lowered the ionic conductivity.[192] Lu et al. finetuned M3GNet to extract Li atomic energies to understand Li-ion migration across interfaces between the cathode and solid electrolytes, where the estimated migration barriers were consistent with the experiments.[193]

Recently, Google reported a universal MLIP called GNoME potential, as part of their work on crystal structure exploration.[144] It was trained on a huge amount of relaxation trajectory data ($\sim 10^8$) obtained during the process. Remarkably, through case studies, it was shown that the zero-shot performance of the GNoME potential was comparable to that of SOTA MLIPs trained from scratch (Figure 4d, e). Furthermore, Microsoft reported MatterSim, a universal MLIP trained on a massive dataset of more than 17 million entries, including public databases and in-house datasets calculated over a wide range of conditions covering 0−5000 K and 0−1000 GPa as shown in the top and middle of Figure 4c.[36] MatterSim achieved SOTA performance on Matbench Discovery,[38] demonstrating its capability of predicting the stability of new, unknown compounds. In addition, it successfully calculated lattice dynamics and thermodynamics, such as phonon dispersion and phase diagrams, respectively, far outperforming the previous universal MLIPs trained solely on relaxation trajectory data (bottom of Figure 4c). These works illustrate the feasibility of a universal foundational MLIP, even with zero-shot or, if necessary, few-shots.

3.3. Inverse Design

Inverse design is a paradigm where desired properties are specified first, guiding the design of new molecules and materials to meet those criteria. High-throughput screening, a common inverse design strategy, generates a large pool of candidates and filters them step by step to identify those with the target properties. However, since initial candidates are often created by fragmenting known molecules and enumerating or substituting them, the success rate tends to be low, and the process relies heavily on the experimenter's intuition. Generative models, on the other hand, can learn the underlying data structure from large datasets and utilize data distributions to generate novel candidates. Inspired by the success of generative models in producing images[194,195] and text,[32] this approach is increasingly being explored in chemistry.

We briefly introduce the concept of the generative model. The goal of the generative model is to learn the probability distribution of data $p(x)$, from which realistic data can be sampled.[196] Training of generative models involves the reconstruction of the training data, which is essentially SSL in nature as aforementioned in Section 3.1.2. For example, a variational autoencoder (VAE) maps by encoder training data x to a feature in latent space enforcing Gaussian distribution, from which latent feature z is sampled and decoded by decoder to the original input x.[197] A more recent approach, the diffusion model, gradually adds Gaussian noise to x until the input data becomes pure Gaussian noise, then removes the noise to reconstruct the x.[198] For the inverse design on a target property, a property predictor can be used in VAE to regularize the latent space by the property,[199] or the denoising process in the diffusion model is conditioned on the property.[16] Another type of generative model, the language model, reconstructs text by either predicting masked words based on the context of a text[63] or predicting the next word based on a sequence of previous words.[177]

A *small* foundational generative model for inverse design can be seen as a pretrained generative model which can be adapted to various inverse design tasks. For instance, MatterGen, a diffusion model developed by Microsoft, was first unconditionally pretrained on a large database such as Alexandria[148] and MP,[132] then finetuned by conditioning on the property such as band gap, elastic modulus, spacegroup, and

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

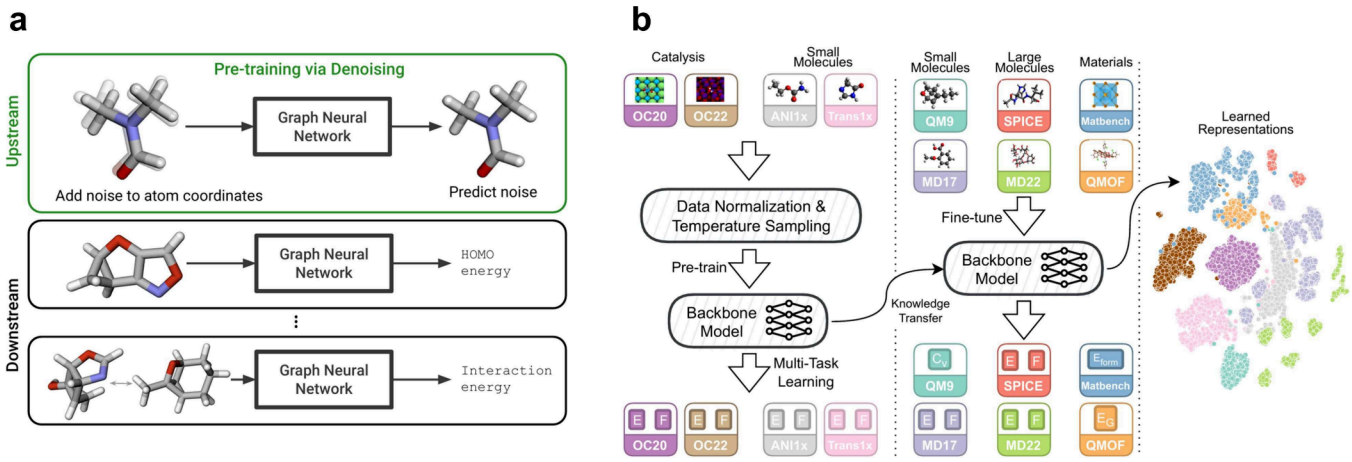


Figure 6. a. Illustration of denoising as a pretraining for properties and force field prediction. Reprinted from ref 158 with permission from the authors. Copyright 2023, the authors. **b.** Supervised pretraining on force fields dataset via joint multi-domain pretraining (JMP) proposed by Shoghi et al. Reprinted from ref 200 with permission from the authors. Copyright 2024, the authors.

composition to enable inverse design for each property (Figure 5a, b).[16] Notably, MatterGen significantly outperformed the previous SOTA model by more than a factor of 2 in generating stable, unique, and novel (S.U.N.) materials, which were more than 10 times closer to the DFT local energy minimum (Figure 5c). IBM reported GP-MoLFormer, a transformer pretrained on 1.1 billion SMILES. This was finetuned via a parameter-efficient novel approach called pair-tuning for property-guided optimization, performing comparably to or better than baselines on inverse design for logP, QED (drug-likeness), and DRD2 activity.[149] Gruver et al. proposed finetuning the pretrained LLaMA-2 70B[152] for generation of crystals with desired spacegroup, composition, and E_{hull} using string representations of crystals and prompting, demonstrating the adaptability of the pretrained LLM to atomistic data.[151]

However, generative models have generally not been explored through pretraining and finetuning schemes; they are most often trained directly from scratch for a target property. As a result, few *small* foundational generative models exist in this context. Alternatively, some generative models suggest the possibility of a *big* foundation model, which can be adapted for both property prediction and inverse design. For example, Gómez-Bombarelli et al. first proposed using a VAE in molecular inverse design by integrating it with a property regressor.[199] The VAE successfully optimized molecules for the target property (5 × QED − SAS), while the performance of the regressor was comparable to that of purely supervised GNN. This implies that representations can be learned such that they can be transferred across both property prediction and inverse design. Following this, we will focus on *big* foundation models that are adaptable to these two domains further in Section 3.5, and we refer the readers interested in generative models in chemistry to refs 18, 196 for more detail.

3.4. Multi-domain: Property Prediction and MLIP

Since MLIPs are predictive models by nature, those that perform well at property prediction are also expected to excel in predicting force fields. Additionally, total energy and forces are related to the global and local features of molecules or crystals, respectively, suggesting that learning force fields itself can serve as an effective pretraining task for property prediction.[200] In this section, we review two approaches to utilizing force fields: denoising by learning fictitious force fields, and supervised learning on energy and forces.

3.4.1. Denoising. Denoising was originally developed for generative modeling and aims to learn the derivative of the log probability distribution of data, $\nabla_x \log p(x)$, rather than directly modeling $p(x)$.[201] For configurations of materials, this probability distribution can be expressed by the Boltzmann distribution, $p(x) \propto \exp(-E(x))$, where $E(x)$ represents the potential energy of configuration x.[153,158] Then, $\nabla_x \log p(x) \propto \nabla_x E(x)$, which is equivalent to the force. In practice, denoising is implemented by predicting noise added to material configurations, often assumed to be Gaussian since equilibrium configurations are located at local minima (Figure 6a). Consequently, denoising can be considered akin to learning the virtual forces that guide noisy coordinates back to their original equilibrium positions.[158]

This denoising approach may benefit downstream tasks such as the prediction of forces and other quantum mechanical properties that are highly sensitive to 3D coordinates.[37,153,156,158,161,184,202−204] For instance, Jiao et al. achieved a notable improvement in force prediction on the MD17 dataset, reducing the average error from 0.2086 to 0.0968 kcal/ mol·Å after pretraining, and outperformed other SSL methods across 11 out of 12 properties in the QM9 dataset.[158] Denoising has also been explored for predicting properties of crystalline materials.[184,203,204] Notably, Song et al. applied denoising to fractional coordinates and lattice parameters using a denoising diffusion probabilistic model (DDPM)[198] framework.[184] This method achieved SOTA performance on the JARVIS-DFT benchmark and showed better results in data-limited scenarios, such as with experimental data.

3.4.2. Supervised Pretraining. Thanks to advances in computational resources and speed, high-quality force field databases are expanding and becoming readily available. As a result, energy and force data can serve as effective pretraining labels, both in terms of abundance and relevance to downstream tasks.[146,200,205] Gao et al. leveraged 86 million DFT-calculated energies from the PubChem PM6[206] database for pretraining to adapt for force prediction and molecular property tasks.[205] Importantly, they applied zero-force regularization, given that the molecular structures in the training data were optimized. Feng et al. introduced a force-

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

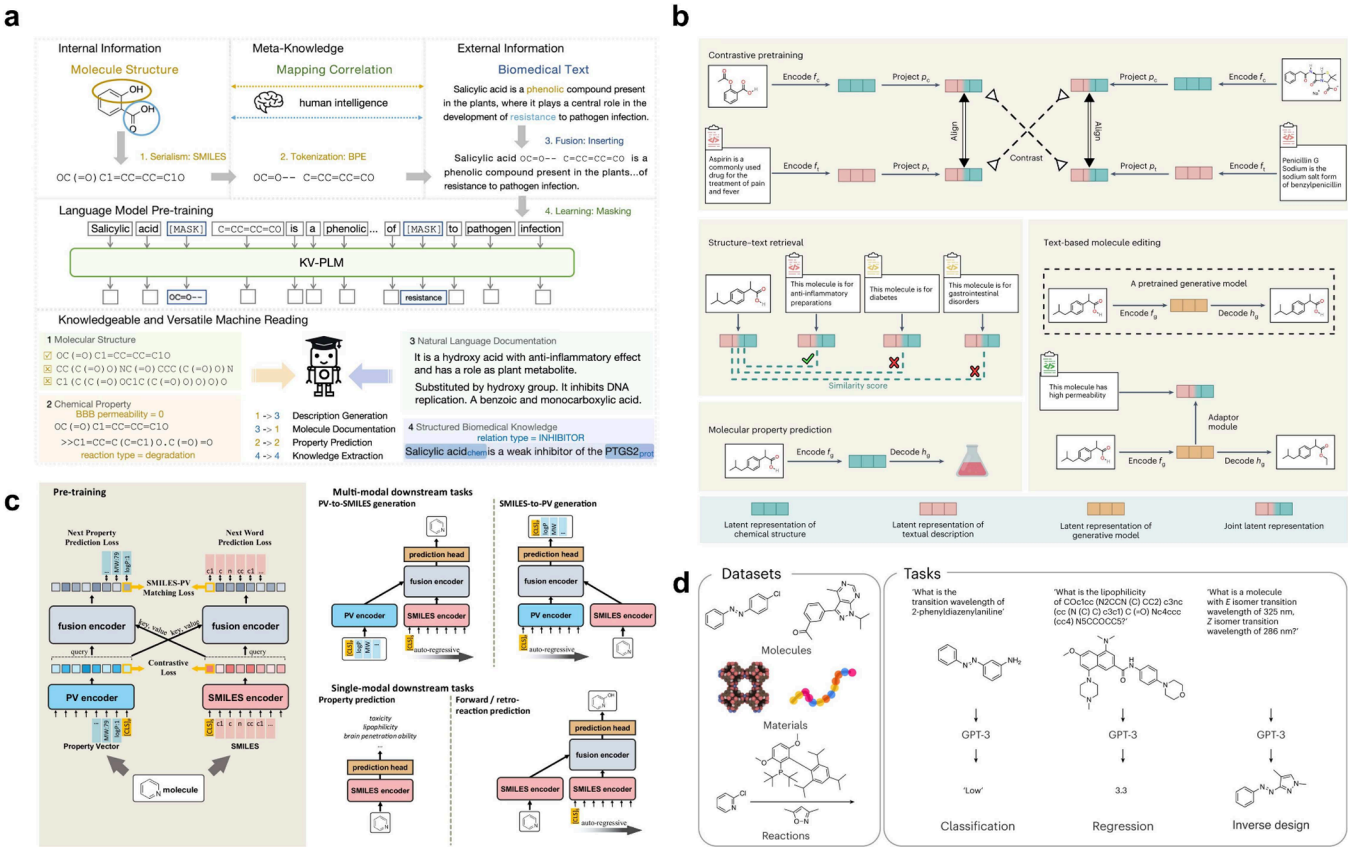


Figure 7. Multimodal foundation models. **a.** Schematic of multimodal KV-PLM, using SMILES and text by fusing them into unified data. Reprinted from ref 162 under the terms of the CC BY license. **b.** Multimodal model proposed by Liu et al., jointly training the structure and text encoders for each modality. Reprinted with permission from ref 40. Copyright 2023 Springer Nature. **c.** Workflow of SPMM proposed by Chang and Ye, which aims to learn joint representation of structure and properties. Reprinted from ref 41 under the terms of the CC BY license. **d.** Finetuning the public ChatGPT proposed by Jablonka et al. Reprinted from ref 175 under the terms of the CC BY license.

centric pretraining strategy combining denoising and supervised learning for equilibrium and nonequilibrium conformations, respectively.[37] This approach not only improved test set performance for force prediction but also enhanced stability in MD simulations and accurately reproducing interatomic distance distributions, demonstrating its practicality for MLIP. Additionally, Jia et al. demonstrated that explicitly learning the derivatives of total energy, i.e. forces and stress, generally results in superior performance across a variety of downstream tasks compared to training solely on energies or using denoising methods.[207]

Recently, Meta's FAIR team introduced a novel joint multi-domain pretraining (JMP) strategy, in which the model is trained simultaneously on diverse energy and force databases (Figure 6b).[200] To handle the inconsistencies in DFT calculation setups and differences in database sizes and system scales, they normalized the data and modified commonly used loss functions. This approach achieved SOTA performance in 34 out of 40 downstream tasks, spanning various databases and including properties and forces of molecules, crystals, and MOFs, underscoring the effectiveness of their pretraining method. The universal MLIP discussed in Section 3.2 can also be adapted for property prediction tasks. Indeed, Chen et al. and Yang et al. demonstrated that each universal MLIP's performance as a property predictor was comparable to or exceeded SOTA results.[36,137] These studies highlight the efficacy of learning force fields as a pretext task to learn the generic representations adaptable to a wide range of property predictions.

3.5. Multi-domain: Property Prediction and Inverse Design

While the property-directed inverse design may suggest novel materials beyond the known chemical space, it remains essential to evaluate the generated data using property prediction tools. In this sense, a *big* foundation model that is adaptable to both property prediction and inverse design would play a key role in materials discovery. This can be achieved by learning representations that are transferable across both domains, as discussed in Section 3.3. In this section, we focus on approaches to *big* foundation models, categorized into two areas: multimodal learning and leveraging LLMs.

3.5.1. Multimodal Learning. One of the earliest multimodal models is KV-PLM, which leveraged two modalities: molecular structure and text description (Figure 7a).[162] Specifically, molecular structure was represented using SMILES, tokenized with the byte pair encoding (BPE)[208] algorithm to capture frequent substring patterns in a molecule, while text descriptions were obtained from S2orc,[163] an academic paper corpus. The segmented SMILES tokens were inserted into the text, which was used to pretrain BERT as a backbone model via masked language modeling. The model successfully generated drug molecules from the input text and performed comparably to baselines such as D-MPNN,[209] showcasing the potential of language-based foundation models.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

MolT5, proposed by Edwards et al., took a similar approach, leveraging both text and SMILES to train the T5[136] model, with text-based inverse design and molecule captioning (rather than property prediction) investigated as downstream tasks.[210]

Su et al. employed molecular graphs instead of SMILES as a structural modality, with text description as another modality.[165] A pretrained BERT model[162,166] and a GNN model from GraphCL[105] were used as backbone text and graph encoders for their multimodal model, respectively, which was then trained using contrastive learning between the text and graph modalities. For inverse molecular design, they utilized an external model, MoFlow,[211] which transforms q sampled from the Gaussian distribution into molecular graphs. They optimized q to maximize the cosine similarity between the input text and generated molecular graph representations, successfully generating valid molecules based on descriptions at various levels. In property prediction, the graph encoder was finetuned, outperforming all compared SSL methods on average in the MoleculeNet benchmark.

Liu et al. proposed a similar multimodal learning framework, MoleculeSTM, which can be adapted to structure-text retrieval, text-based molecule editing, and property prediction (Figure 7b).[40] For the zero-shot molecule editing, another generative model is employed, by optimizing representations to maximize similarities to both representations encoded by the pretrained text encoder and the generative model. In the 4 editing tasks including single- and multi-objective editing, binding-affinity-based editing, and drug-relevance editing, MoleculeSTM achieved superior hit ratios compared to baselines including genetic search. Luo et al. proposed MolFM, where a knowledge graph was incorporated as regularization in addition to molecular graph and text encoders.[167] KV-PLM[162] and GraphMVP[111] were employed to initialize their encoders, with four pretraining objectives consisting of structure-text contrastive loss, cross-modal matching loss, masked language modeling loss, and knowledge graph embedding loss. In the downstream inverse design, the MolT5[210] decoder was employed to generate SMILES from the input text.

More recently, Chang and Ye proposed SPMM to learn the joint representation of structure and property by embedding SMILES and a property vector (PV) in a common embedding space, enabling bidirectional PV-to-SMILES generation and property prediction (Figure 7c).[41] PV consisted of 53 molecular properties calculated by RDKit and was treated as a 53-length sentence. Notably, PV was randomly masked to allow the use of partial properties in practice. SMILES generation takes up to 53 properties and requires no additional training, successfully generating valid and novel molecules with desired PVs in all tested cases, including scenarios with all 53 properties, a single property, four properties, and no properties provided. In addition, SPMM can predict the PVs of input SMILES without finetuning but can be finetuned for enhanced property prediction, achieving performance comparable to the SOTA on the MoleculeNet benchmark. These studies demonstrate that multimodal learning is a promising approach for developing *big* foundation models in chemistry, enabling a range of downstream tasks across domains and the use of more specific conditions to design the desired molecules.

3.5.2. Large Language Model. A large language model such as ChatGPT is renowned for its human-like performance and versatility in understanding context and generating text, making it one of the most popular foundation models. This has motivated researchers to leverage LLMs to solve scientific problems, leading to the development of science-specific LLMs trained on various scientific corpora.[166,212−214] Furthermore, more specialized LLMs have been developed that are adaptable to specific chemical problems, including text-based generation, property prediction, molecule captioning, molecule retrieval, reaction prediction, and more.[33,34,215,216]

Zhao et al. proposed a Dialogue Foundation Model for Chemistry (ChemDFM),[172] built upon the pretrained LLaMa-13B.[173] They trained LLaMa-13B on 3.8M research articles and 1.4K chemistry textbooks, followed by instruction tuning to better adapt the model to chemical languages, including molecular notations, using 2.7M instructions from chemical databases. In property prediction and inverse design tasks, the model performed better than 10-shot generalist LLMs such as GPT-4 and was comparable to conventional specialist models. Livne et al. introduced nach0, pretrained on both NLP and chemical domain datasets.[174] nach0 was finetuned with instruction tuning in a multi-task manner for cross-domain tasks such as text-guided molecule design and property prediction, outperforming SOTA baselines. Notably, these models are capable of interacting with humans as they were pretrained on large corpora, distinguishing them from the aforementioned multimodal language models pretrained purely on SMILES and text descriptions.

Instead of pretraining LLMs manually, researchers can leverage publicly available pretrained LLMs and adapt them for chemistry. Jablonka et al. proposed finetuning GPT-3 via the OpenAI API and benchmarked the model for property prediction and inverse design (Figure 7d).[175] The fine-tuned GPT-3 performed comparably to or even outperformed conventional specialized models in classification, particularly in low-data regimes. In a case study on the inverse design of molecules for photoswitches, the finetuned GPT-3 was able to generate valid molecules with higher novelty using a higher softmax temperature, albeit with an increased risk of invalidity. Choudhary employed GPT-2 and Mistral 7B[178] for property prediction and inverse design of crystals, respectively,[176] using text descriptions generated by ChemNLP.[76] These studies suggest a convenient approach to leveraging foundation models for materials discovery by simply finetuning public LLMs.

4. TRENDS AND FUTURE DIRECTIONS

4.1. Scope

The scope of foundation models can be categorized into three perspectives. The first is the type of materials the foundation models can handle, such as molecules, crystals, surfaces, MOFs, and so on. A large portion of reported foundation models have targeted molecules, and most others have focused on a single type of material. Meanwhile, a foundation model can leverage shared chemistry across different materials, hopefully allowing for complementary learning, especially for those with relatively limited data. This would enhance flexibility across diverse material types. Supervised learning of this concept was explored by Shoghi et al.,[200] but a similar approach could be promising in SSL with careful integration and utilization of different databases across a wide range of materials.

Second is the modality of data. Popular modalities for molecules include SMILES, graphs, text descriptions, and properties. These modalities complement each other and offer opportunities for more diverse tasks. For example, as 3D

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

geometric graphs provide richer structural information than 2D graphs and SMILES, leveraging 3D information can aid downstream tasks where only 2D graphs or SMILES are available.[111] In this sense, exploring additional modalities, such as images and spectral data, and investigating effective approaches to maximize their utility is a promising direction.[92] For example, Suzuki et al. utilized X-ray diffraction (XRD) patterns as a modality alongside crystal graphs to learn the structure−functionality relationship.[217] However, multimodal learning for crystals remains underexplored, warranting further investigation.[218]

Finally, the downstream tasks. The ultimate goal of a foundation model in chemistry is to maximize versatility across downstream tasks, regardless of domain. Progress in this direction is thought to depend on the previous two challenges, as each material type and modality constitutes a task in itself (e.g., predicting a molecule's properties from its SMILES). It is more effective to adapt a foundation model to a learned material type and modality, or at least related ones. Indeed, multimodal learning has proven to be one of the most promising approaches toward a *big* foundation model across domains, as discussed in Section 3.5.1. Furthermore, it is reasonable to expect that a foundation model trained on vast, diverse datasets could adapt well to new chemistry and tasks, an emergent capability.[29] Consequently, we anticipate that discovering effective methods to extend the range of materials and modalities will be key to developing *big* foundational models capable of encompassing any domain.

4.2. Performance

Central to the success of foundation models is scale, both in data and model size.[29] Hence, understanding the scalability of data and models in chemistry is crucial. It should be noted that although the high cost of labeling data makes SSL popular for pretraining, it is not straightforward to obtain large meaningful unlabeled data efficiently. Naively generating molecules and crystals risks producing a large amount of invalid and unstable material data. While generative models can successfully produce realistic data, they generally fill gaps within known chemical space rather than expand it, as they are trained on existing databases. Therefore, advanced techniques are needed to explore the configuration space and discover meaningful material structures beyond what is currently known, often requiring extensive domain expertise.[219] Nevertheless, unstable material data may still be useful to some extent, for example, in learning symmetries of crystals, which are important for determining their properties but challenging to capture.[35,220]

The integration of existing databases is one of the most straightforward ways to enlarge the dataset for training large-scale foundation models. However, data acquisition method must be compatible between existing databases to ensure seamless integration, which is often not the case, necessitating a proper approach to address inconsistencies.[200] For example, Shiota et al. recently proposed a total energy alignment method requiring minimal recalculations to integrate databases calculated using different *ab initio* packages (e.g., MP[132] using the Vienna Ab initio Simulation Package (VASP)[222] and OFF23[143] using Psi4[223]).[221] The importance of consistent data is further highlighted by Pengmei et al., who found that noisy labels, such as DFT-calculated HOMO−LUMO gaps with high uncertainty depending on the choice of DFT functionals and basis sets, could limit scalability and OOD performance.[224] In conclusion, both the quantity and quality of data are crucial for the effective scaling and performance of foundation models.

As with large amounts of data, it is important to design scalable models that can harness this information to learn meaningful knowledge. For example, transformers can scale up to tens or hundreds of billions of parameters while improving performance.[32] In this regard, language models for chemistry represent a promising architecture that has already been extensively studied. However, the scalability of GNNs has rarely been explored, with complex equivariant models typically having fewer than a million parameters.[38] Simply increasing the dimensionality and message-passing layers is not always beneficial due to the curse of dimensionality and oversmoothing, the latter of which Godwin et al. proposed mitigating by introducing noisy nodes.[225] Very recently, Sypetkowski et al. demonstrated that GNNs could scale up to 3 billion parameters with consistent improvements in molecular property prediction and highlighted the importance of model width in driving finetuning performance.[226] In another recent study, Pengmei et al. found distinct scaling characteristics during pretraining compared to other fields like language modeling.[224] Such studies should be conducted more actively to advance large-scale GNNs, paving the way for developing foundation models in chemistry.[224]

A common observation in many studies suggests that exploiting both local and global information about materials is effective in learning good representations. This is because properties may depend on local, global, or both types of information, providing such representations with maximum versatility in various downstream tasks involving properties. This can be achieved through contrastive learning between local and global views, predictive learning on local motifs and global properties, or generative learning to reconstruct masked nodes and entire graphs. Contrastive learning with data augmentation has also proven effective with carefully designed augmentation methods at both the node and graph levels. Additionally, supervised learning on force fields allows the model to effectively learn both local and global aspects, each represented by atomic forces and total energy. Overall, pretraining tasks that focus on both aspects represent a promising approach deserving further exploration.

While many reported foundation models demonstrate excellent performance compared to conventional approaches on benchmarks such as MoleculeNet,[182] Matbench,[94] and JARVIS-DFT,[140] more attention should be paid to their performance in the limited data regime, a typical challenge in chemistry and materials science.[35] For example, experimental datasets such as OQMD-EXP[131,184] and high-fidelity quantum mechanical calculation data (e.g., coupled cluster)[160,227,228] of relatively smaller amounts can serve as useful benchmarks. Performance tests on varying training data sizes would also provide insights into the scaling characteristics of the models.[153,184,202] Additionally, robust extrapolation capability is another key factor in chemistry that requires more focus.[37] In this regard, constructing benchmark datasets across various domains can help objectively assess a model's OOD performance. For example, Fu et al. reported a benchmark suite and metrics to evaluate the robustness of MLIPs during MD simulations, which is not necessarily guaranteed by low test set prediction error.[229] The Matbench Discovery also serves as a benchmark to assess the relaxation performance of universal MLIPs on unseen structures.[38] These benchmark datasets

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

would aid in validating foundation MLIPs in more practical settings.

4.3. Efficiency

Since the computational burden increases as the size of models grows, efforts have been made to improve model efficiency, particularly for LLMs.[230] For example, quantization reduces the number of bits required to represent model weights, pruning aims to identify and remove redundant weights and components, and matrix decomposition alleviates computational overhead in matrix multiplication. While these techniques are model-agnostic in principle, thorough experiments on GNNs for chemistry are necessary. Meanwhile, Sourek et al. proposed a technique to compress GNNs without loss of information by capturing symmetries, i.e., shared correlations in the data, taking advantage of the structured convolutional nature of GNNs.[231] Additionally, the necessity of equivariance in 3D geometric GNNs which incur large computational costs may not always manifest, requiring further investigation.[232] Studies on model compression tailored to GNNs would represent a promising direction for developing efficient foundation models for chemistry.

Efficiency is particularly critical for MLIPs, as large-scale simulations involving millions of timesteps are expected in MD simulations driven by MLIPs. However, foundation models can be very expensive even in the inference stage, limiting their utility as MLIPs. Knowledge distillation[233] may offer a solution, where knowledge from an expensive but accurate and generalized model (the teacher model) is distilled into a lighter and more efficient model (the student model).[233] Ekström Kelvinius et al. first proposed using knowledge distillation in GNN-based MLIPs by aligning embeddings obtained from the teacher and student models.[234] Gong et al. introduced ensemble distillation, achieved by training a single MLIP on snapshots sampled from MD simulations labeled by an ensemble of MLIPs.[235] While knowledge distillation has primarily been studied in classification models and rarely addressed in regression tasks, exploring what constitutes "knowledge" could unveil significant opportunities. For example, while the decomposition of total energy into atomic energies in the MLIP formalism is not unique and can be somewhat ad hoc if ill-trained, they can be correctly predicted in a well-generalized model, reflecting physical and chemical atomic interactions,[236−238] potentially serving as knowledge for MLIPs.

4.4. Interpretability

As many researchers rely on the use of foundation models, the limitations of foundation models could be easily ignored. As an example, some foundation models are still vulnerable to generating hallucination, generating plausible yet nonfactual content,[239] and the model collapsed[240] outputs, which lack sample diversity. Thus, when we use a foundation model, it is important to verify that the generated outputs have valid and safe meaning in chemistry domain. Especially, when the generated contents are biased to a narrow distribution of training data, the novel discovery of chemical compounds or properties is limited. To solve these problems, it is recommended to utilize the model interpretability techniques which explain sample generation mechanisms in the foundation models.[241] For example, interpretable features from a foundation model, Claude 3, were extracted and visualized.[242] Moreover, it was found that specific concepts learned in foundation models can be localized and edited at the word level.[243,244] Recently, a method to accurately identify source documents used to train foundation models was proposed.[245] Such advances in model interpretability make it possible to better understand foundation models,[246] paving the way for their more reliable use in chemistry.

5. CONCLUSION

A foundation model in chemistry is expected to solve various problems in the field, guiding researchers to achieve the long-standing goal of materials discovery and understanding the behavior of matter. The current status of foundation models in chemistry is still premature, and many challenges must be overcome for these models to be implemented at a practical level. However, there are clear trends in effective pretraining approaches and learning schemes, and importantly, databases are expanding. These offer promising opportunities for foundation models in chemistry in the future, toward new possibilities beyond the domains focused in this perspective.

AUTHOR INFORMATION

Corresponding Author

Yousung Jung − *Department of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea; Institute of Engineering Research, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea;* orcid.org/0000-0003-2615-8394; Email: yousung.jung@snu.ac.kr

Authors

Junyoung Choi − *Department of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea*

Gunwook Nam − *Department of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea*

Jaesik Choi − *Graduate School of Artificial Intelligence, KAIST Daejeon, Daejeon 34141, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c01160

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Y.J. acknowledges support from NRF (RS-2023-00283902, RS-2024-00464386) and IITP (RS2021-II211343) of Korea.

REFERENCES

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(2) Mitchell, J. B. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 468−481.

(3) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.

(4) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(5) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. *Nature Catalysis* **2018**, *1*, 696−703.
(6) Li, J.; Zhou, M.; Wu, H.-H.; Wang, L.; Zhang, J.; Wu, N.; Pan, K.; Liu, G.; Zhang, Y.; Han, J.; Liu, X.; Chen, X.; Wan, J.; Zhang, Q. Machine Learning-Assisted Property Prediction of Solid-State Electrolyte. *Adv. Energy Mater.* **2024**, *14*, 2304480.
(7) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **2021**, *20*, 750−761.
(8) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine learning force fields. *Chem. Rev.* **2021**, *121*, 10142−10186.
(9) Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(10) Qi, J.; Banerjee, S.; Zuo, Y.; Chen, C.; Zhu, Z.; Holekevi Chandrappa, M. L.; Li, X.; Ong, S. P. Bridging the gap between simulated and experimental ionic conductivities in lithium superionic conductors. *Mater. Today Phy.* **2021**, *21*, 100463.
(11) Cheng, D.; Zhao, Z.-J.; Zhang, G.; Yang, P.; Li, L.; Gao, H.; Liu, S.; Chang, X.; Chen, S.; Wang, T.; Ozin, G. A.; Liu, Z.; Gong, J. The nature of active sites for carbon dioxide electroreduction over oxide-derived copper catalysts. *Nat. Commun.* **2021**, *12*, 395.
(12) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse design of solid-state materials via a continuous representation. *Matter* **2019**, *1*, 1370−1384.
(13) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS central science* **2020**, *6*, 1412−1420.
(14) Ren, Z.; et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **2022**, *5*, 314−335.
(15) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. S. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. *Int. Conf. Learn. Representations* **2022**, 1−20.
(16) Zeni, C. et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687* **2023**.
(17) Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **2024**, *15*, 1−16.
(18) Park, H.; Li, Z.; Walsh, A. Has generative artificial intelligence solved inverse materials design? *Matter* **2024**, *7*, 2355−2367.
(19) Vo, T. H.; Nguyen, N. T. K.; Le, N. Q. K. Improved prediction of drug-drug interactions using ensemble deep neural networks. *Medicine in Drug Discovery* **2023**, *17*, 100149.
(20) Le, N. Q. K. Predicting emerging drug interactions using GNNs. *Nature Computational Science* **2023**, *3*, 1007−1008.
(21) Chen, S.; Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **2021**, *1*, 1612−1620.
(22) He, T.; Huo, H.; Bartel, C. J.; Wang, Z.; Cruse, K.; Ceder, G. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Sci. Adv.* **2023**, *9*, No. eadg8180.
(23) Kim, S.; Noh, J.; Gu, G. H.; Chen, S.; Jung, Y. Predicting synthesis recipes of inorganic crystal materials using elementwise template formulation. *Chemical Science* **2024**, *15*, 1039−1045.
(24) Chen, S.; Jung, Y. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence* **2022**, *4*, 772−780.
(25) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **2019**, *5*, 1572−1583.
(26) Szymanski, N. J.; Fu, S.; Persson, E.; Ceder, G. Integrated analysis of X-ray diffraction patterns and pair distribution functions for machine-learned phase identification. *npj Comput. Mater.* **2024**, *10*, 45.
(27) Riesel, E. A.; Mackey, T.; Nilforoshan, H.; Xu, M.; Badding, C. K.; Altman, A. B.; Leskovec, J.; Freedman, D. E. Crystal structure determination from powder diffraction patterns with generative machine learning. *J. Am. Chem. Soc.* **2024**, *146*, 30340−30348.
(28) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **2023**, *123*, 8736−8780.
(29) Bommasani, R. et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
(30) Wang, H.; Kaddour, J.; Liu, S.; Tang, J.; Lasenby, J.; Liu, Q. Evaluating self-supervised learning for molecular graph embeddings. *Adv. Neur. Inf. Process. Syst.* **2024**, *36*, 68028−68060.
(31) Batatia, I. et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096* **2023**.
(32) Brown, T.; et al. Language models are few-shot learners. *Adv. Neur. Inf. Process. Syst.* **2020**, *33*, 1877−1901.
(33) Lei, G.; Docherty, R.; Cooper, S. J. Materials science in the era of large language models: a perspective. *Digital Discovery* **2024**, *3*, 1257−1272.
(34) Ramos, M. C.; Collison, C.; White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science* **2025**, *16*, 2514−2572.
(35) Das, K.; Samanta, B.; Goyal, P.; Lee, S.-C.; Bhattacharjee, S.; Ganguly, N. Crysgnn: Distilling pre-trained knowledge to enhance property prediction for crystalline materials. *Proceedings of the AAAI Conference on Artificial Intelligence* **2023**, *37*, 7323−7331.
(36) Yang, H. et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* **2024**.
(37) Feng, R.; Zhu, Q.; Tran, H.; Chen, B.; Toland, A.; Ramprasad, R.; Zhang, C. May the force be with you: Unified force-centric pre-training for 3d molecular conformations. *Adv. Neur. Inf. Process. Syst.* **2024**, *36*, 72750−72760.
(38) Riebesell, J.; Goodall, R. E. A.; Benner, P.; Chiang, Y.; Deng, B.; Lee, A. A.; Jain, A.; Persson, K. A. Matbench Discovery − A framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920* **2024**.
(39) Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2024**, *46*, 5625−5644.
(40) Liu, S.; Nie, W.; Wang, C.; Lu, J.; Qiao, Z.; Liu, L.; Tang, J.; Xiao, C.; Anandkumar, A. Multi-modal molecule structure−text model for text-based retrieval and editing. *Nature Machine Intelligence* **2023**, *5*, 1447−1457.
(41) Chang, J.; Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nat. Commun.* **2024**, *15*, 2323.
(42) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.
(43) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.-R.; Gross, E. K. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 205118.
(44) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(45) Sodeyama, K.; Igarashi, Y.; Nakayama, T.; Tateyama, Y.; Okada, M. Liquid electrolyte informatics using an exhaustive search with linear regression. *Phys. Chem. Chem. Phys.* **2018**, *20*, 22585−22591.
(46) Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K.-A. N.; Cui, Y.; Reed, E. J. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **2017**, *10*, 306−320.
(47) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* **2010**, *104*, 136403.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

(48) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation* **2016**, *14*, 1153−1173.
(49) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436−444.
(50) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; http://www.deeplearningbook.org.
(51) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3*, 93.
(52) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learning Syst.* **2021**, *32*, 4−24.
(53) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *Int. Conf. Mach. Learn.* **2017**, 1263−1272.
(54) Satorras, V. G.; Hoogeboom, E.; Welling, M. E (n) equivariant graph neural networks. *Int. Conf. Mach. Learn.* **2021**, 9323−9332.
(55) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219* **2018**.
(56) Han, J.; Rong, Y.; Xu, T.; Huang, W. Geometrically equivariant graph neural networks: A survey. *arXiv preprint arXiv:2202.07230* **2022**.
(57) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.
(58) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neur. Inf. Process. Syst.* **2022**, *35*, 11423−11436.
(59) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Int. Conf. Mach. Learn.* **2021**, 9377−9388.
(60) Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Adv. Neur. Inf. Process. Syst.* **2021**, *34*, 6790−6802.
(61) Chang, Y.; et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1−45.
(62) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is all you need. *Adv. Neur. Inf. Process. Syst.* **2017**, *30*, 1−11.
(63) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* **2019**.
(64) Zhao, W. X. et al. A survey of large language models. *arXiv preprint arXiv:2303.18223* **2023**.
(65) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform badly for graph representation? *Adv. Neur. Inf. Process. Syst.* **2021**, *34*, 28877−28888.
(66) Thölke, P.; Fabritiis, G. D. Equivariant Transformers for Neural Network based Molecular Potentials. *Int. Conf. Learn. Representations* **2022**, 1−20.
(67) Liao, Y.-L.; Smidt, T. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. *11th Int. Conf. Learn. Representations* **2023**, 1−32.
(68) Wang, R.; Ji, Y.; Li, Y.; Lee, S.-T. Applications of Transformers in Computational Chemistry: Recent Progress and Prospects. *J. Phys. Chem. Lett.* **2025**, *16*, 421−434.
(69) Han, Y.; Xu, X.; Hsieh, C.-Y.; Ding, K.; Xu, H.; Xu, R.; Hou, T.; Zhang, Q.; Chen, H. Retrosynthesis prediction with an iterative string editing model. *Nat. Commun.* **2024**, *15*, 6404.
(70) Wei, L.; Li, Q.; Song, Y.; Stefanov, S.; Dong, R.; Fu, N.; Siriwardane, E. M.; Chen, F.; Hu, J. Crystal Composition Transformer: Self-Learning Neural Language Model for Generative and Tinkering Design of Materials. *Adv. Sci.* **2024**, *11*, 2304305.
(71) Wan, Y.; Xie, T.; Wu, N.; Zhang, W.; Kit, C.; Hoex, B. From Tokens to Materials: Leveraging Language Models for Scientific Discovery. *arXiv preprint arXiv:2410.16165* **2024**.
(72) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31−36.
(73) Krenn, M.; et al. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3*, 100588.
(74) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1603.
(75) Ganose, A. M.; Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Commun.* **2019**, *9*, 874−881.
(76) Choudhary, K.; Kelley, M. L. ChemNLP: a natural language-processing-based library for materials chemistry text data. *J. Phys. Chem. C* **2023**, *127*, 17545−17555.
(77) Xiao, H.; Li, R.; Shi, X.; Chen, Y.; Zhu, L.; Chen, X.; Wang, L. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nat. Commun.* **2023**, *14*, 7027.
(78) Chen, Y.; Wang, X.; Deng, X.; Liu, Y.; Chen, X.; Zhang, Y.; Wang, L.; Xiao, H. MatterGPT: A Generative Transformer for Multi-Property Inverse Design of Solid-State Materials. *arXiv preprint arXiv:2408.07608* **2024**.
(79) Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857−876.
(80) Wu, L.; Lin, H.; Tan, C.; Gao, Z.; Li, S. Z. Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 4216−4235.
(81) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. *Int. Conf. Mach. Learn.* **2020**, 1597−1607.
(82) Oord, A. v. d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* **2018**.
(83) Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *Int. Conf. Learn. Representations* **2019**, 1−24.
(84) He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.* **2020**, 9729−9738.
(85) Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y. Multimodal deep learning. *Proc. 28th Int. Conf. Mach. Learn. (ICML-11)* **2011**, 689−696.
(86) Lahat, D.; Adali, T.; Jutten, C. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE* **2015**, *103*, 1449−1477.
(87) Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning transferable visual models from natural language supervision. *Int. Conf. Mach. Learn.* **2021**, 8748−8763.
(88) Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *Int. Conf. Mach. Learn.* **2021**, 4904−4916.
(89) Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Computer Vision − ECCV* **2020**, *2020*, 121−137.
(90) Fei, N.; Lu, Z.; Gao, Y.; Yang, G.; Huo, Y.; Wen, J.; Lu, H.; Song, R.; Gao, X.; Xiang, T.; Sun, H.; Wen, J.-R. Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.* **2022**, *13*, 3094.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

(91) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.* **2016**, 770−778.
(92) Takeda, S.; Kishimoto, A.; Hamada, L.; Nakano, D.; Smith, J. R. Foundation model for material science. *Proceedings of the AAAI Conference on Artificial Intelligence* **2023**, *37*, 15376−15383.
(93) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **2021**, *7*, 84.
(94) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **2020**, *6*, 138.
(95) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **2022**, *8*, 59.
(96) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.
(97) Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185.
(98) Gasteiger, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *Int. Conf. Learn. Representations* **2020**, 1−13.
(99) Pan, S. J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **2010**, *22*, 1345−1359.
(100) Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and avoiding negative transfer. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.* **2019**, 11293−11302.
(101) Chen, X.; Lu, S.; Chen, Q.; Zhou, Q.; Wang, J. From bulk effective mass to 2D carrier mobility accurate prediction via adversarial transfer learning. *Nat. Commun.* **2024**, *15*, 5391.
(102) Velickovic, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; Hjelm, R. D. Deep graph infomax. *ICLR (Poster)* **2019**, *2*, 4.
(103) Sun, F.-Y.; Hoffman, J.; Verma, V.; Tang, J. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. *Int. Conf. Learn. Representations* **2020**, 1−16.
(104) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 1−7.
(105) You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; Shen, Y. Graph contrastive learning with augmentations. *Adv. Neur. Inf. Process. Syst.* **2020**, *33*, 5812−5823.
(106) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? *Int. Conf. Learn. Representations* **2019**, 1−17.
(107) Sterling, T.; Irwin, J. J. ZINC 15−ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(108) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **2022**, *4*, 279−287.
(109) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *Int. Conf. Learn. Representations* **2017**, 1−14.
(110) 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.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **2019**, *47*, D1102−D1109.
(111) Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; Tang, J. Pre-training Molecular Graph Representation with 3D Geometry. *Int. Conf. Learn. Representations* **2022**, 1−32.
(112) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neur. Inf. Process. Syst.* **2017**, *30*, 1−11.
(113) Axelrod, S.; Gomez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **2022**, *9*, 185.
(114) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. *Int. Conf. Learn. Representations* **2020**, 1−22.
(115) Hamilton, W. L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* **2017**.
(116) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, *40*, D1100−D1107.
(117) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science* **2018**, *9*, 5441−5451.
(118) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Adv. Neur. Inf. Process. Syst.* **2020**, *33*, 12559−12571.
(119) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Informatics* **2019**, 429−436.
(120) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(121) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* **2022**.
(122) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
(123) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **2022**, *4*, 1256−1264.
(124) Irwin, J. J.; Shoichet, B. K. ZINC- a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(125) Kang, Y.; Park, H.; Smit, B.; Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal−organic frameworks. *Nature Machine Intelligence* **2023**, *5*, 309−318.
(126) Magar, R.; Wang, Y.; Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Comput. Mater.* **2022**, *8*, 231.
(127) Ward, L.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(128) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal−organic frameworks. *Nature Chem.* **2012**, *4*, 83−89.
(129) Das, K.; Samanta, B.; Goyal, P.; Lee, S.-C.; Bhattacharjee, S.; Ganguly, N. CrysXPP: An explainable property predictor for crystalline materials. *npj Comput. Mater.* **2022**, *8*, 43.
(130) Louis, S.-Y.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; Hu, J. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18141−18148.
(131) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1*, 1−15.
(132) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

(133) Fu, N.; Wei, L.; Hu, J. Physics-Guided Dual Self-Supervised Learning for Structure-Based Material Property Prediction. *J. Phys. Chem. Lett.* **2024**, *15*, 2841−2850.
(134) Omee, S. S.; Louis, S.-Y.; Fu, N.; Wei, L.; Dey, S.; Dong, R.; Li, Q.; Hu, J. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns* **2022**, *3*, 100491.
(135) Rubungo, A. N.; Arnold, C.; Rand, B. P.; Dieng, A. B. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029* **2023**.
(136) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1−67.
(137) Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2022**, *2*, 718−728.
(138) Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C. J.; Ceder, G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence* **2023**, *5*, 1031−1041.
(139) Choudhary, K.; DeCost, B.; Major, L.; Butler, K.; Thiyagalingam, J.; Tavazza, F. Unified graph neural network force-field for the periodic table: solid state applications. *Digital Discovery* **2023**, *2*, 346−355.
(140) Choudhary, K.; et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **2020**, *6*, 173.
(141) Takamoto, S.; et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat. Commun.* **2022**, *13*, 2991.
(142) Takamoto, S.; Izumi, S.; Li, J. TeaNet: Universal neural network interatomic potential inspired by iterative electronic relaxations. *Comput. Mater. Sci.* **2022**, *207*, 111280.
(143) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable machine learning force fields for organic molecules. *arXiv preprint arXiv:2312.15211* **2023**.
(144) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80−85.
(145) Park, Y.; Kim, J.; Hwang, S.; Han, S. Scalable Parallel Algorithm for Graph Neural Network Interatomic Potentials in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2024**, *20*, 4857−4868.
(146) Barroso-Luque, L.; Shuaibi, M.; Fu, X.; Wood, B. M.; Dzamba, M.; Gao, M.; Rizvi, A.; Zitnick, C. L.; Ulissi, Z. W. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. *arXiv preprint arXiv:2410.12771* **2024**.
(147) Liao, Y.-L.; Wood, B.; Das, A.; Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059* **2023**.
(148) Schmidt, J.; Wang, H.-C.; Cerqueira, T. F.; Botti, S.; Marques, M. A. A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals. *Sci. Data* **2022**, *9*, 64.
(149) Ross, J.; Belgodere, B.; Hoffman, S. C.; Chenthamarakshan, V.; Mroueh, Y.; Das, P. GP-MoLFormer: A Foundation Model For Molecular Generation. *arXiv preprint arXiv:2405.04912* **2024**.
(150) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic acids research* **2016**, *44*, D1202−D1213.
(151) Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. W. Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. *12th Int. Conf. Learn. Representations* **2024**, 1−20.
(152) Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* **2023**.
(153) Zaidi, S.; Schaarschmidt, M.; Martens, J.; Kim, H.; Teh, Y. W.; Sanchez-Gonzalez, A.; Battaglia, P.; Pascanu, R.; Godwin, J. Pre-training via Denoising for Molecular Property Prediction. *11th Int. Conf. Learn. Representations* **2023**, 1−26.
(154) Sanchez-Gonzalez, A.; Godwin, J.; Pfaff, T.; Ying, R.; Leskovec, J.; Battaglia, P. Learning to simulate complex physics with graph networks. *Int. Conf. Mach. Learn.* **2020**, 8459−8468.
(155) Nakata, M.; Shimazaki, T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300−1308.
(156) Liu, S.; Guo, H.; Tang, J. Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching. *11th Int. Conf. Learn. Representations* **2023**, 1−27.
(157) Xu, Z.; Luo, Y.; Zhang, X.; Xu, X.; Xie, Y.; Liu, M.; Dickerson, K.; Deng, C.; Nakata, M.; Ji, S. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs. *arXiv preprint arXiv:2110.01717* **2021**.
(158) Jiao, R.; Han, J.; Huang, W.; Rong, Y.; Liu, Y. Energy-motivated equivariant pretraining for 3d molecular graphs. *Proceedings of the AAAI Conference on Artificial Intelligence* **2023**, *37*, 8096−8104.
(159) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(160) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.
(161) Ni, Y.; Feng, S.; Hong, X.; Sun, Y.; Ma, W.-Y.; Ma, Z.-M.; Ye, Q.; Lan, Y. Pre-training with fractional denoising to enhance molecular property prediction. *Nat. Mach. Intell.* **2024**, *6*, 1−10.
(162) Zeng, Z.; Yao, Y.; Liu, Z.; Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat. Commun.* **2022**, *13*, 862.
(163) Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; Weld, D. S. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* **2019**.
(164) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research* **2009**, *37*, W623−W633.
(165) Su, B.; Du, D.; Yang, Z.; Zhou, Y.; Li, J.; Rao, A.; Sun, H.; Lu, Z.; Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481* **2022**.
(166) Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* **2019**.
(167) Luo, Y.; Yang, K.; Hong, M.; Liu, X. Y.; Nie, Z. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484* **2023**.
(168) Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Adv. Neur. Inf. Process. Syst.* **2013**, *26*, 1−9.
(169) Wishart, D. S.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* **2018**, *46*, D1074−D1082.
(170) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.
(171) 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.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research* **2021**, *49*, D1388−D1395.
(172) Zhao, Z.; Ma, D.; Chen, L.; Sun, L.; Li, Z.; Xia, Y.; Xu, H.; Zhu, Z.; Zhu, S.; Fan, S.; Shen, G.; Yu, K.; Chen, X. ChemDFM: A Large Language Foundation Model for Chemistry. *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*; 2024, *1−24*.
(173) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

(174) Livne, M.; Miftahutdinov, Z.; Tutubalina, E.; Kuznetsov, M.; Polykovskiy, D.; Brundyn, A.; Jhunjhunwala, A.; Costa, A.; Aliper, A.; Aspuru-Guzik, A.; Zhavoronkov, A. nach0: Multimodal natural and chemical languages foundation model. *Chemical Science* **2024**, *15*, 8380−8389.
(175) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **2024**, *6*, 161−169.
(176) Choudhary, K. AtomGPT: Atomistic Generative Pretrained Transformer for Forward and Inverse Materials Design. *J. Phys. Chem. Lett.* **2024**, *15*, 6909−6917.
(177) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
(178) Jiang, A. Q. et al. Mistral 7B. *arXiv preprint arXiv:2310.06825* **2023**.
(179) You, Y.; Chen, T.; Shen, Y.; Wang, Z. Graph contrastive learning automated. *Int. Conf. Mach. Learn.* 2021, pp 12121−12132.
(180) Thakoor, S.; Tallec, C.; Azar, M. G.; Azabou, M.; Dyer, E. L.; Munos, R.; Veličković, P.; Valko, M. *Large-Scale Representation Learning on Graphs via Bootstrapping*; Int. Conf. Learn. Representations, 2022, pp 1−21.
(181) Kim, D.; Baek, J.; Hwang, S. J. Graph self-supervised learning with accurate discrepancy learning. *Adv. Neur. Inf. Process. Syst.* **2022**, *35*, 14085−14098.
(182) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513−530.
(183) Vapnik, V.; Izmailov, R. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.* **2015**, *16*, 2023−2049.
(184) Song, Z.; Meng, Z.; King, I. A Diffusion-Based Pre-training Framework for Crystal Property Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* **2024**, *38*, 8993−9001.
(185) Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research* **2022**, *2022*, 0004.
(186) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X.; Hou, T.; Cao, D. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic acids research* **2021**, *49*, W5−W14.
(187) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* **2020**.
(188) Landrum, G. *RDKit: Open-source cheminformatics*; 2006.
(189) Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics* **2020**, *8*, 64−77.
(190) Focassio, B. M.; Freitas, L. P.; Schleder, G. R. Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Appl. Mater. Interfaces* **2024**, 1−11.
(191) Gonzales, C.; Fuemmeler, E.; Tadmor, E. B.; Martiniani, S.; Miret, S. Benchmarking of Universal Machine Learning Interatomic Potentials for Structural Relaxation. *AI for Accelerated Materials Design - NeurIPS* **2024**, *2024*, 1−7.
(192) Jun, K.; Xiao, Y.; Sun, W.; Byeon, Y.-W.; Kim, H.; Ceder, G. Nitride Lithium-ion Conductors with Enhanced Oxidative Stability. *J. Electrochem. Soc.* **2024**, *171*, 090518.
(193) Lu, P.; Gong, S.; Wang, C.; Yu, Z.; Huang, Y.; Ma, T.; Lian, J.; Jiang, Z.; Chen, L.; Li, H.; Wu, F. Superior Low-Temperature All-Solid-State Battery Enabled by High-Ionic-Conductivity and Low-Energy-Barrier Interface. *ACS Nano* **2024**, *18*, 7334−7345.
(194) Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. *Zero-shot text-to-image generation*. Int. Conf. Mach. Learn. 2021, 8821−8831.
(195) Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. *High-resolution image synthesis with latent diffusion models*. Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recogn. 2022, 10684−10695.
(196) Anstine, D. M.; Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736−8750.
(197) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
(198) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neur. Inf. Process. Syst.* **2020**, *33*, 6840−6851.
(199) 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.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268−276.
(200) Shoghi, N.; Kolluru, A.; Kitchin, J. R.; Ulissi, Z. W.; Zitnick, C. L.; Wood, B. M. *From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction*. The Twelfth Int. Conf. Learn. Representations 2024, 1−30.
(201) Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation* **2011**, *23*, 1661−1674.
(202) Wang, Y.; Xu, C.; Li, Z.; Barati Farimani, A. Denoise pretraining on nonequilibrium molecules for accurate and transferable neural potentials. *J. Chem. Theory Comput.* **2023**, *19*, 5077−5087.
(203) Shen, S.; Liu, K.; Zhu, M.; Chen, H. Boost Your Crystal Model with Denoising Pre-training. *ICML 2024 AI for Science Workshop* **2024**, 1−12.
(204) New, A.; Le, N. Q.; Pekala, M.; Stiles, C. D. Self-supervised learning for crystal property prediction via denoising. *ICML 2024 AI for Science Workshop* **2024**, 1−9.
(205) Gao, X.; Gao, W.; Xiao, W.; Wang, Z.; Wang, C.; Xiang, L. Supervised Pretraining for Molecular Force Fields and Properties Prediction. *NeurIPS 2022 AI for Science: Progress and Promises* **2022**, 1−12.
(206) Nakata, M.; Shimazaki, T.; Hashimoto, M.; Maeda, T. PubChemQC PM6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *J. Chem. Inf. Model.* **2020**, *60*, 5891−5899.
(207) Jia, S.; Parthasarathy, A. R.; Feng, R.; Cong, G.; Zhang, C.; Fung, V. Derivative-based pre-training of graph neural networks for materials property predictions. *Digital Discovery* **2024**, *3*, 586−593.
(208) Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* **2015**.
(209) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(210) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817* **2022**.
(211) Zang, C.; Wang, F. *Moflow: an invertible flow model for generating molecular graphs*. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining 2020, 617−626.
(212) Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* **2022**.
(213) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234−1240.
(214) Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. Mausam, MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8*, 102.
(215) Liu, P.; Tao, J.; Ren, Z. Scientific language modeling: A quantitative review of large language models in molecular science. *arXiv preprint arXiv:2402.04119* **2024**.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518

(216) Van Herck, J.; et al. Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chemical Science* **2025**, *16*, 670−684.
(217) Suzuki, Y.; Taniai, T.; Saito, K.; Ushiku, Y.; Ono, K. Self-supervised learning of materials concepts from crystal structures via deep neural networks. *Machine Learning: Science and Technology* **2022**, *3*, 045034.
(218) Das, K.; Goyal, P.; Lee, S.-C.; Bhattacharjee, S.; Ganguly, N. CrysMMNet: Multimodal Representation for Crystal Property Prediction. *Proc. 39th Conf. Uncertainty in Artificial Intelligence* **2023**, *216*, 507−517.
(219) Oganov, A. R. *Modern methods of crystal structure prediction*; John Wiley & Sons, 2011.
(220) Gong, S.; Yan, K.; Xie, T.; Shao-Horn, Y.; Gomez-Bombarelli, R.; Ji, S.; Grossman, J. C. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Sci. Adv.* **2023**, *9*, No. eadi3245.
(221) Shiota, T.; Ishihara, K.; Do, T. M.; Mori, T.; Mizukami, W. Taming Multi-Domain,-Fidelity Data: Towards Foundation Models for Atomistic Scale Simulations. *arXiv preprint arXiv:2412.13088* **2024**.
(222) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169.
(223) Smith, D. G. A.; et al. PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **2020**, *152*, 184108.
(224) Pengmei, Z.; Shen, Z.; Wang, Z.; Collins, M.; Rangwala, H. Pushing the Limits of All-Atom Geometric Graph Neural Networks: Pre-Training, Scaling and Zero-Shot Transfer. *arXiv preprint arXiv:2410.21683* **2024**.
(225) Godwin, J.; Schaarschmidt, M.; Gaunt, A.; Sanchez-Gonzalez, A.; Rubanova, Y.; Veličković, P.; Kirkpatrick, J.; Battaglia, P. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971* **2021**.
(226) Sypetkowski, M.; Wenkel, F.; Poursafaei, F.; Dickson, N.; Suri, K.; Fradkin, P.; Beaini, D. On the Scalability of GNNs for Molecular Graphs. *arXiv preprint arXiv:2404.11568* **2024**.
(227) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science* **2021**, *1*, 46−53.
(228) Ruth, M.; Gerbig, D.; Schreiner, P. R. Machine learning of coupled cluster (T)-energy corrections via delta (Δ)-learning. *J. Chem. Theory Comput.* **2022**, *18*, 4846−4855.
(229) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237* **2022**.
(230) Ganesh, P.; Chen, Y.; Lou, X.; Khan, M. A.; Yang, Y.; Sajjad, H.; Nakov, P.; Chen, D.; Winslett, M. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 1061−1080.
(231) Sourek, G.; Zelezny, F.; Kuzelka, O. Lossless compression of structured convolutional models via lifting. *arXiv preprint arXiv:2007.06567* **2020**.
(232) Ko, T. W.; Ong, S. P. Recent advances and outstanding challenges for machine learning interatomic potentials. *Nature Computational Science* **2023**, *3*, 998−1000.
(233) Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* **2015**.
(234) Ekström Kelvinius, F.; Georgiev, D.; Toshev, A.; Gasteiger, J. Accelerating molecular graph neural networks via knowledge distillation. *Adv. Neur. Inf. Process. Syst.* **2024**, *36*, 25761−25792.
(235) Gong, S.; Zhang, Y.; Mu, Z.; Pu, Z.; Wang, H.; Yu, Z.; Chen, M.; Zheng, T.; Wang, Z.; Chen, L.; Wu, X.; Shi, S.; Gao, W.; Yan, W.; Xiang, L. BAMBOO: a predictive and transferable machine learning force field framework for liquid electrolyte development. *arXiv preprint arXiv:2404.07181* **2024**.
(236) Yoo, D.; Lee, K.; Jeong, W.; Lee, D.; Watanabe, S.; Han, S. Atomic energy mapping of neural network potential. *Phys. Rev. Mater.* **2019**, *3*, 093802.
(237) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-driven learning of total and local energies in elemental boron. *Physical review letters* **2018**, *120*, 156001.
(238) Wang, S.; Liu, Y.; Mo, Y. Frustration in Super-Ionic Conductors Unraveled by the Density of Atomistic States. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202215544.
(239) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 1−11.
(240) Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755−759.
(241) Kim, J.; Gu, G. H.; Noh, J.; Kim, S.; Gim, S.; Choi, J.; Jung, Y. Predicting potentially hazardous chemical reactions using an explainable neural network. *Chemical Science* **2021**, *12*, 11028−11037.
(242) Templeton, A. et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. 2024; https://transformer-circuits.pub/2024/scaling-monosemanticity/ (accessed: 2025−01−19).
(243) Meng, K.; Bau, D.; Andonian, A.; Belinkov, Y. Locating and editing factual associations in GPT. *Adv. Neur. Inf. Process. Syst.* **2022**, *35*, 17359−17372.
(244) Meng, K.; Sharma, A. S.; Andonian, A. J.; Belinkov, Y.; Bau, D. *Mass-Editing Memory in a Transformer*. The Eleventh Int. Conf. Learn. Representations 2023, 1−21.
(245) Park, B.; Choi, J. Memorizing Documents with Guidance in Large Language Models. *Proc. Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24* **2024**, 6460−6468.
(246) Hutson, M. How does ChatGPT 'think'? Psychology and neuroscience crack open AI large language models. *Nature* **2024**, *629*, 986−988.

https://doi.org/10.1021/jacsau.4c01160
JACS Au 2025, 5, 1499−1518