Symmetry-Informed Geometric Representation for Molecules, Proteins, and Crystalline Materials

Shengchao Liu[1,2], **Weitao Du**[3], **Yanjing Li**[4], **Zhuoxinran Li**[5], **Zhiling Zheng**[6], **Chenru Duan**[7], **Zhiming Ma**[3], **Omar Yaghi**[6], **Anima Anandkumar**[8], **Christian Borgs**[6], **Jennifer Chayes**[6], **Hongyu Guo**[9,10], **Jian Tang**[1,11,12]

[1]Mila - Québec Artificial Intelligence Institute [2]Université de Montréal
[3]University of Chinese Academy of Sciences [4]Carnegie Mellon University
[5]University of Toronto [6]University of California, Berkeley [7]Massachusetts Institute of Technology
[8]California Institute of Technology [9]National Research Council Canada [10]University of Ottawa
[11]HEC Montréal [12]CIFAR AI Chair

Abstract

Artificial intelligence for scientific discovery has recently generated significant interest within the machine learning and scientific communities, particularly in the domains of chemistry, biology, and material discovery. For these scientific problems, molecules serve as the fundamental building blocks, and machine learning has emerged as a highly effective and powerful tool for modeling their geometric structures. Nevertheless, due to the rapidly evolving process of the field and the knowledge gap between science (*e.g.*, physics, chemistry, & biology) and machine learning communities, a benchmarking study on geometrical representation for such data has not been conducted. To address such an issue, in this paper, we first provide a unified view of the current symmetry-informed geometric methods, classifying them into three main categories: invariance, equivariance with spherical frame basis, and equivariance with vector frame basis. Then we propose a platform, coined Geom3D, which enables benchmarking the effectiveness of geometric strategies. Geom3D contains 16 advanced symmetry-informed geometric representation models and 14 geometric pretraining methods over 46 diverse datasets, including small molecules, proteins, and crystalline materials. We hope that Geom3D can, on the one hand, eliminate barriers for machine learning researchers interested in exploring scientific problems; and, on the other hand, provide valuable guidance for researchers in computational chemistry, structural biology, and materials science, aiding in the informed selection of representation techniques for specific applications. The source code is available on the GitHub repository.

1 Introduction

Artificial intelligence (AI) for molecule discovery has recently seen many developments, including small molecular property prediction [12, 16, 22, 36, 62, 73, 95, 97, 98, 113, 124, 126, 126, 129], small molecule design and optimization [6, 51, 53, 80, 131], small molecule reaction and retrosynthesis [37, 105, 110], protein property prediction [25, 134], protein folding and inverse folding [45, 61, 87], protein design [14, 38, 43, 83, 86], and crystalline material design [31, 119, 122]. One of the most fundamental building blocks for these tasks is the geometric structure of molecules. Exploring effective methods for robust representation learning to leverage such geometric information fully remains an open challenge that interests both machine learning (ML) and science researchers.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

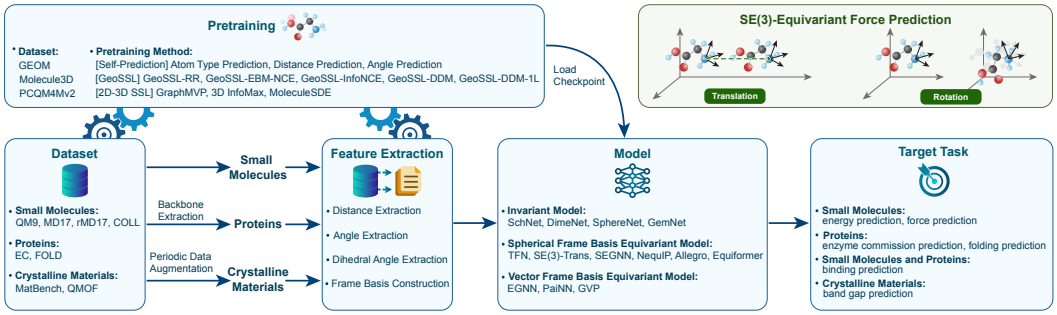


Figure 1: Pipeline for Geom3D, including dataset preprocessing, feature extraction, geometric pretraining and representation, and target tasks. We additionally demonstrate the **SE(3)-equivariant force prediction task**.

To this end, symmetry-informed geometric representation [1] has emerged as a promising approach. By leveraging physical principles (*i.e.*, group theory for depicting symmetric particles) into spatial representation, they facilitate a more robust representation of small molecules, proteins, and crystalline materials. Nevertheless, pursuing geometric learning research is still challenging due to its evolving nature and the knowledge gap between science (*e.g.*, physics) and machine learning communities. These factors contribute to a substantial barrier for machine learning researchers to investigate scientific problems and hinder efforts to reproduce results consistently. To overcome this, we introduce Geom3D, a benchmarking of the geometric representation with four advantages, as follows. [1]

(1) A unified and novel aspect in understanding symmetry-informed geometric models. The molecule geometry needs to satisfy certain physical constraints regarding the 3D Euclidean space. For instance, the molecules' force needs to be equivariant to translation and rotation (see SE(3)-equivariance in Fig. 1). In this work, we classify the geometric methods into three categories: *invariant* model, SE(3)-equivariant model with *spherical frame basis* and *vector frame basis*. The invariant models only consider features that are constant w.r.t. the SE(3) group, while the two families of equivariant models can be further unified using the *frame basis* to capture equivariant symmetry. An illustration of three categories is in Fig. 2. Building equivariant models on the *frame basis* provides a novel and unified view of understanding geometric models and paves the way for intriguing more ML researchers to explore scientific problems.

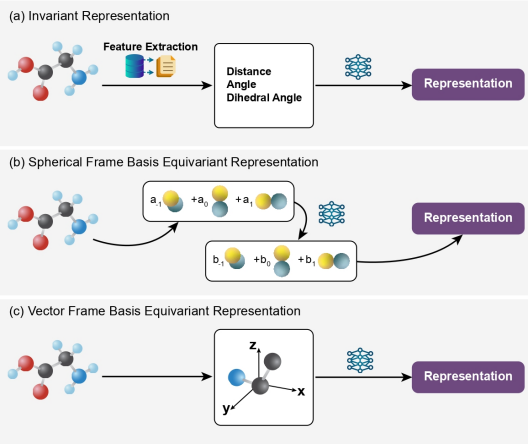
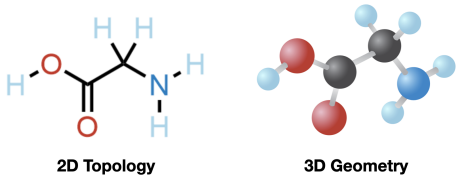


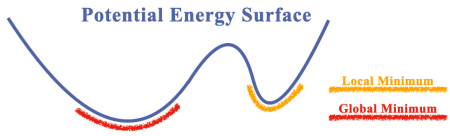
Figure 2: Three categories of geometric modules. (a) Invariant models only consider type-0 features. Equivariant models use either (b) spherical harmonics frames or (c) vector frames by projecting the coordinate vectors.

(2) A unified platform for various scientific domains. There exist multiple platforms and tools for molecule discovery, but they are (1) mainly focusing on molecule's 2D graph representation [72, 96, 137]; (2) using 3D geometry with customized data structures or APIs [3, 99]; or (3) covering only a few geometric models [71]. Thus, it is necessary to have a platform benchmarking the geometric models, especially for researchers interested in solving scientific problems. In this work, we propose Geom3D, a geometric modeling framework based on PyTorch Geometric (PyG) [29], one of the most widely-used platforms for graph representation learning. Geom3D benchmarks 16 geometric models on solving 46 scientific tasks, and these tasks include the three most fundamental molecule types: small molecules, proteins, and crystalline materials. Each of them requires distinct domain-specific preprocessing steps, *e.g.*, crystalline materials molecules possess periodic structures and thus need a particular periodic data augmentation. By leveraging such a unified framework, Geom3D serves as a comprehensive benchmarking tool, facilitating effective and consistent analysis components to interpret the existing geometric representation functions in a fair and convenient comparison setting.

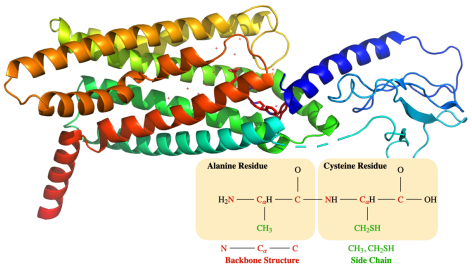
[1] In what follows, we may use "molecule" to refer to "small molecule" for brevity.



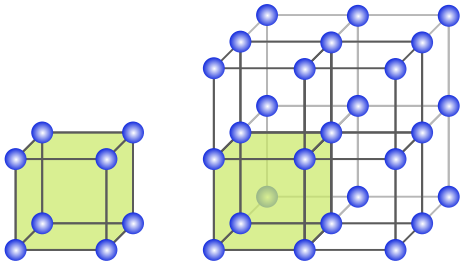
(a) An example of 2D topology and 3D geometry.



(b) An illustration of the potential energy surface.



(c) An example of protein structure.



(d) An example of crystalline material.

Figure 3: Fig. 3(a) illustrates 2D topology and 3D conformation for molecule Glycine. Fig. 3(c) displays the 3D structure of protein. Fig. 3(d) shows a simple cubic crystal of the element Po. Fig. 3(b) is a demo of PES.

(3) A framework for a wider range of ML tasks. The geometric models in Geom3D can serve as a building block for exploring extensive ML tasks, including but not limited to studying the molecule dynamic simulation and scrutinizing the transfer learning effect on molecule geometry. For example, pretraining is an important strategy to quickly transfer knowledge to target tasks, and recent works explore geometric pretraining on 3D conformations (including supervised and self-supervised) [55, 75, 130] and multi-modality pretraining on 2D topology and 3D geometry [28, 74, 81]. Other transfer learning venues include multi-task learning [77, 79] and out-of-distribution or domain adaptation [54, 127, 128], yet no geometry information has been utilized. All of these directions are promising for future exploration, and Geom3D serves as an auxiliary tool to accomplish them. For example, as will be shown in Sec. 4, we leverage Geom3D to effectively evaluate 14 pretraining methods with benchmarks.

(4) A framework for exploring data preprocessing and optimization tricks. When comparing different symmetry-informed geometric models, we find that in addition to the model architecture, there are two important factors affecting the performance: the data preprocessing (*e.g.*, energy and force rescaling and shift) and optimization methods (*e.g.*, learning rate, learning rate schedule, number of epochs, random seeds). In this work, we explore the effect of four preprocessing tricks and around 2-10 optimization hyperparameters for each model and task. In general, we observe that each model may benefit differently in different tasks regarding the preprocessing and optimization tricks. However, data normalization is found to help improve performance hugely in most cases. We believe that Geom3D is an effective tool for exploring and understanding various engineering tricks.

2 Data Structures for Geometric Data

Small molecule 3D conformation. Molecules are sets of points in the 3D Euclidean space, and they move in a dynamic motion, as known as the potential energy surface (PES). The region with the lowest energy corresponds to the most stable state for molecules, and molecules at these positions are called **conformations**, as illustrated in Fig. 3(b). For notation, we mark each 3D molecular graph as $\boldsymbol{g} = (\boldsymbol{X}, \boldsymbol{R})$, where $\boldsymbol{X}$ and $\boldsymbol{R}$ are for the atom types and positions, respectively.

Crystalline material with periodic structure. The crystalline materials or extended chemical structures possess a characteristic known as periodicity: their atomic or molecular arrangement repeats in a predictable and consistent pattern across all three spatial dimensions. This is the key aspect that differentiates them from small molecules. In Fig. 3(d), we show an original unit cell (marked in green) that can repeatedly compose the crystal structure along the lattice. To model such a periodic structure, we adopt the data augmentation from CGCNN [123]: for each original unit cell, we shift it along the lattice in three dimensions and connect edges within a cutoff value (hyperparameter). For more details on the two augmentation variants, please check Appendix A.

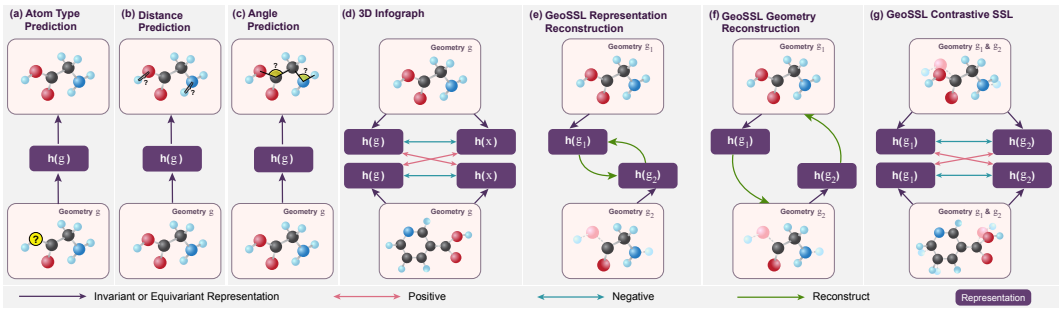


Figure 4: Pipelines for seven single-modal geometric pretraining methods. (a-c) conduct self-prediction. (d) maximizes the MI between nodes and graphs. (e-g) are GeoSSL, maximizing the MI between views $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$.

Protein with backbone structure. Protein structures can be classified into four primary levels, *e.g.*, the primary structure represents the linear arrangement of *amino acids* within a polypeptide chain, where each amino acid is a small molecule. The geometric models can be naturally adopted to the higher-order structures, and in Geom3D, we consider the tertiary structure, which encompasses the complete three-dimensional organization of a single protein. Regarding the data structure, we consider the tertiary structure: each amino acid has an important *backbone structure* $N - C_\alpha - C$ structure, and the C_α is bonded to the side chain. There are 20 common types of side chains corresponding to 20 amino acids, as illustrated in Fig. 3(c). Considering the long-sequence issue in proteins, existing works [59, 116] mainly model the backbone structures for computational efficiency.

3 Symmetry-Informed Geometric Representation

3.1 Group Symmetry and Equivariance

Symmetry means the object remains invariant after certain transformations [121], and it is everywhere on Earth, such as in animals, plants, and molecules. Formally, the set of all symmetric transformations satisfies the axioms of a group. Therefore, the group theory and its representation theory are common tools to depict such physical symmetry. **Group** is a set G equipped with a group product $\times$ satisfying:

$$(1)\ \exists \boldsymbol{e} \in G,\ \boldsymbol{a}\times\boldsymbol{e}=\boldsymbol{e}\times\boldsymbol{a}, \forall \boldsymbol{a}\in G; \quad (2)\ \boldsymbol{a}\times\boldsymbol{a}^{-1}=\boldsymbol{a}^{-1}\times\boldsymbol{a}=\boldsymbol{e}; \quad (3)\ \boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c})=\boldsymbol{a}\times\boldsymbol{b}\times\boldsymbol{c}. \tag{1}$$

Group representation is a mapping from the group G to the group of linear transformations of a vector space X with dimension d (see [132] for more rigorous definition):

$$\rho_X(\cdot): G \to \mathbb{R}^{d\times d} \qquad \text{s.t.} \quad \rho(\boldsymbol{e})=1 \ \wedge\ \rho_X(\boldsymbol{a})\rho_X(\boldsymbol{b})=\rho_X(\boldsymbol{a}\times\boldsymbol{b}),\ \forall \boldsymbol{a},\boldsymbol{b}\in G. \tag{2}$$

During modeling, the X space can be the input 3D Euclidean space, the equivariant vector space in the intermediate layers, or the output force space. This enables the definition of equivariance as below.

Equivariance is the property for the geometric modeling function $f: X \to Y$ as:

$$f(\rho_X(\boldsymbol{a})\boldsymbol{x})=\rho_Y(\boldsymbol{a})f(\boldsymbol{x}),\ \ \forall \boldsymbol{a}\in G, \boldsymbol{x}\in X. \tag{3}$$

As displayed in Fig. 1, for molecule geometric modeling, the property should be rotation-equivariant and translation-equivariant (*i.e.*, SE(3)-equivariant). More concretely, $\rho_X(\boldsymbol{a})$ and $\rho_Y(\boldsymbol{a})$ are the SE(3) group representations on the input (*e.g.*, atom coordinates) and output space (*e.g.*, force space), respectively. SE(3)-equivariant modeling in Eq. (3) is essentially saying that the designed deep learning model f is modeling the whole transformation trajectory on the molecule conformations, and the output is the transformed $\hat{y}$ accordingly. Further, we want to highlight that, in addition to the network architecture or representation function, the input features can also be represented as an equivariant feature mapping from the 3D mesh to $\mathbb{R}^{\tilde{d}}$ [11], where $\tilde{d}$ depends on input data, *e.g.*, $\tilde{d}$ = 1 (for atom type dimension) + 3 (for atom coordinate dimension) on small molecules. Such features are called steerable features in [5, 11] when only considering the subgroup SO(3)-equivariance.

Invariance is a special type of equivariance, defined as:

$$f(\rho_X(\boldsymbol{a})\boldsymbol{x})=f(\boldsymbol{x}),\ \ \forall \boldsymbol{a}\in G, \boldsymbol{x}\in X, \tag{4}$$

with $\rho_Y(\boldsymbol{a})$ as the identity $\forall \boldsymbol{a}\in G$. The group representation helps define the equivariance condition for f to follow. Then, the question boils down to how to design such an equivariant f. In the following, we will discuss geometric modelings from a novel and unified perspective using the frame.

3.2 Invariant Geometric Representation Learning

The simple way of achieving SE(3) group symmetric in molecule geometry is invariant modeling. It means the model considers only the invariant features or type-0 features [106] when modeling, and such type-0 features are invariant with respect to rotation and translation. Specifically, several works have been adopting the invariant features for modeling, including but not limited to pairwise distance (SchNet [103]), bond angles (DimeNet [64]), and torsion angles (SphereNet [84] and GemNet [63]). Note that the torsion angles are angles between two planes defined by pairwise bonds. We also want to highlight that, from a mathematical perspective, equivariance and invariance can be transformed to each other by the scalarization technique. Please check [46] for details.

3.3 Equivariant Geometric Representation Learning

Invariant modeling only captures the type-0 features. However, equivariant modeling of higher-order particles may bring in extra expressiveness. For example, the elementary particles in high energy physics [92] inherit higher order symmetries in the sense of SO(3) representation theory, which makes the equivariant modeling necessary. Such higher-order particles include type-1 features like coordinates and forces in molecular conformation. There are many approaches to design such SE(3)-equivariant model satisfying Eq. (3). There are two main venues, as will be discussed below.

Spherical Frame Basis. This research line utilizes the irreducible representations [35] for building SO(3)-equivariant representations, and the first work is TFN [106]. Its main idea is to project the 3D Euclidean coordinates into the spherical harmonics space, which transforms equivariantly according to the irreducible representations of SO(3), and the translation-equivariant can be trivially guaranteed using the relative coordinates. Following this, there have been variants combining it with the attention module (Equiformer [68]) or with more expressive network architectures (SEGNN [4], Allegro [89]).

Vector Frame Basis. This is an alternative solution by using the vector (in physics) frame basis. It builds the frame in the vector space, and the SO(3)-equivariance can be satisfied with the Gram-Schmidt process. Works along this line for molecule discovery include EGNN [102] and PaiNN [104] for geometric representation, 3D-EMGP [55] and MoleculeSDE [74] for geometric pretraining, and ClofNet [19] for conformation generation. For macromolecules like protein, the equivariant vector frame has been used for protein design (StructTrans [50]) and protein folding (AlphaFold2 [61]).

The spherical frame basis can be easily extended to higher-order particles, yet it may suffer from the high computational cost. On the other hand, the vector frame basis is specifically designed for the 3D point clouds; thus, it is more efficient but cannot generalize to higher-order particles. Meanwhile, we would like to acknowledge other equivariant modeling paradigms, including using orbital features [93] and elevating 3D Euclidean space to SE(3) group [30, 49]. Please check Appendix F for details.

3.4 Geometric Pretraining

Recent studies have started to explore **single-modal geometric pretraining** on molecules. The GeoSSL paper [75] covers a wide range of geometric pretraining algorithms. The type prediction, distance prediction, and angle prediction predict the masked atom type, pairwise distance, and bond angle, respectively. The 3D InfoGraph predicts whether the node- and graph-level 3D representation are for the same molecule. GeoSSL is a novel geometric pretraining paradigm that maximizes the mutual information (MI) between the original conformation $\boldsymbol{g}_1$ and augmented conformation $\boldsymbol{g}_2$, where $\boldsymbol{g}_2$ is obtained by adding small perturbations to $\boldsymbol{g}_1$. RR, InfoNCE, and EBM-NCE optimize the objective in the latent representation space, either generative or contrastive. GeoSSL-DDM [75, 130] optimizes the same objective function using denoising score matching. 3D-EMGP [56] has the same strategy and utilizes an equivariant module to denoise the 3D noise directly. We illustrate these algorithms in Fig. 4. Another research line is the **multi-modal pretraining on topology and geometry**. GraphMVP [81] first proposes one contrastive objective (EBM-NCE) and one generative objective (VRR) to optimize the MI between the 2D topologies and 3D geometries in the representation space. 3D InfoMax [108] is a special case of GraphMVP, with the contrastive part only. MoleculeSDE [74] extends GraphMVP by introducing two SDE models for solving the 2D and 3D reconstruction.

Table 1: Results of 26 models on 12 quantum mechanics prediction tasks in QM9, with 110K for training, 10K for validation, and 11K for testing. The task unit is specified, and the evaluation is the mean absolute error (MAE).

Featurization	Model	$\alpha \downarrow$ a_0^3	$\nabla\mathcal{E} \downarrow$ meV	$\mathcal{E}_{\text{HOMO}} \downarrow$ meV	$\mathcal{E}_{\text{LUMO}} \downarrow$ meV	$\mu \downarrow$ D	$C_v \downarrow$ $\frac{cal}{mol \cdot K}$	$G \downarrow$ meV	$H \downarrow$ meV	$R^2 \downarrow$ a_0^2	$U \downarrow$ meV	$U_0 \downarrow$ meV	ZPVE $\downarrow$ meV
1D FPs	MLP	2.231	196.72	131.27	164.94	0.526	0.919	2158.64	2358.23	68.621	2340.61	2314.77	155.921
	RF	3.801	207.02	165.72	183.04	0.534	1.485	3391.79	3729.94	94.512	3705.75	3678.25	253.132
	XGB	2.748	199.71	139.88	165.43	0.516	1.062	2563.93	2804.27	82.959	2786.28	2769.29	180.989
1D SMILES	CNN	0.364	165.22	124.65	114.81	0.566	0.173	156.66	170.59	20.403	166.18	169.89	10.070
	BERT	0.313	117.50	84.93	98.88	0.446	0.176	170.01	183.43	18.002	183.84	188.60	13.410
1D SELFIES	CNN	0.345	157.04	115.51	113.00	0.499	0.168	136.42	146.56	20.080	143.00	140.01	10.149
	BERT	0.348	123.11	91.15	90.80	0.461	0.203	168.20	187.50	19.125	204.93	195.98	17.328
2D Graph	GCN	1.338	145.82	96.21	106.66	0.434	0.526	1198.12	1291.57	37.585	1281.03	1303.39	85.103
	ENN-S2S	1.401	270.59	129.18	132.84	0.577	0.760	1487.21	955.24	34.609	1800.79	1521.32	51.226
	GraphSAGE	1.601	131.45	88.78	93.21	0.402	0.544	1473.42	1617.73	38.112	1553.01	1565.65	95.344
	GAT	1.132	135.90	94.70	98.52	0.406	0.291	911.82	991.31	26.583	1161.29	592.67	55.061
	GIN	1.165	175.82	90.66	110.74	0.539	0.691	848.24	1090.36	35.110	1498.23	1364.18	108.331
	D-MPNN	0.568	118.42	85.01	86.20	0.441	0.241	423.14	458.39	24.816	470.01	445.91	29.291
	PNA	0.681	148.88	88.72	97.31	0.361	0.409	664.98	692.74	23.855	616.70	694.92	57.217
	Graphormer	2.836	79.27	54.24	52.42	0.330	0.080	2066.28	2546.01	131.158	2229.88	2525.51	144.595
	AWARE	0.297	144.91	133.89	98.86	0.602	0.129	86.62	94.47	22.180	93.59	95.73	5.275
	GraphGPS	0.209	75.98	54.75	54.53	0.288	0.089	528.50	693.19	12.488	296.00	411.16	49.888
3D Graph	SchNet	0.060	44.13	27.64	22.55	0.028	0.031	14.19	14.05	0.133	13.93	13.27	1.749
	DimeNet++	0.044	36.22	20.01	16.66	0.028	0.022	7.45	6.14	0.323	6.33	7.18	1.118
	SE(3)-Trans	0.137	56.52	34.65	34.41	0.050	0.063	65.28	70.70	1.747	68.92	68.88	5.428
	EGNN	0.062	49.56	30.08	24.98	0.029	0.030	10.01	9.14	0.089	9.28	9.08	1.519
	PaiNN	0.049	42.73	24.46	20.16	0.016	0.025	8.43	7.88	0.169	8.18	7.63	1.419
	GemNet-T	0.041	35.46	17.85	15.86	0.021	0.023	7.61	7.08	0.271	6.42	5.88	1.232
	SphereNet	0.047	38.93	21.45	18.25	0.027	0.025	8.16	13.68	0.288	6.77	7.43	1.295
	SEGNN	0.048	33.61	17.66	17.01	0.021	0.026	11.60	12.45	0.404	11.29	12.20	1.590
	Equiformer	0.051	33.46	17.93	16.85	0.015	0.023	14.49	14.60	0.433	14.88	13.78	2.342

3.5 Discussion: Reflection-antisymmetric in Geometric Learning

Till now, we have discussed the SE(3)-equivariance, *i.e.*, the translation and rotation equivariance. As highlighted in the recent work [57, 74], the molecules needlessly satisfy the reflection-equivariant, but instead, they should be reflection-antisymmetric [74]. One standard example is that the energy of small molecules is reflection-antisymmetric in a binding system. Each of the two equivariant categories discussed in Sec. 3.3 can solve this problem easily. The spherical frame basis can achieve this by adding the reflection into the Wigner-D matrix [4]. The vector frame basis can accomplish this using the cross-product during frame construction [74].

4 Geometric Datasets and Benchmarks

In Sec. 3, we introduce a novel aspect for understanding symmetry-informed geometric models. In this section, we discuss utilizing Geom3D framework for benchmarking 16 geometric models over 46 tasks. For the detailed dataset acquisitions and task specifications (*e.g.*, *dataset size*, *splitting*, and *task unit*), please check Appendix B. Geom3D also covers 7 1D models and 10 2D graph neural networks (GNNs) and benchmarks the 14 pretraining algorithms to learn a robust geometric representation. Additionally, we want to highlight Geom3D enables exploration of important data preprocessing and optimization tricks for performance improvement, as will be introduced next.

4.1 Small Molecules: QM9

QM9 [94] is a dataset consisting of 134K molecules, each with up to 9 heavy atoms. It includes 12 tasks that are related to the quantum properties. For example, U0 and U298 are the internal energies at 0K and 298.15K, respectively, and U298 and G298 are the other two energies that can be calculated from H298. The other 8 tasks are quantum mechanics related to the density functional theory (DFT) process. On the QM9 dataset, we can easily get the 1D descriptors (Fingerprints/FPs [100], SMILES [120], SELFIES [66]), 2D topology, and 3D conformation. This enables us to build models on each of them respectively: (1) We benchmark 7 models on 1D descriptors, including multi-layer perception (MLP), random forest (RF), XGBoost (SGB), convolution neural networks (CNN), and BERT [17]. (2) We benchmark 10 2D GNN models on the molecular topology, including GCN [22, 62], ENN-S2S [36], GraphSAGE [40], GAT [113], GIN [124], D-MPNN [126], PNA [12], Graphormer [129], AWARE [16], GraphGPS [95]. (3) We benchmark 9 3D geometric models on the molecular conformation, including SchNet [103], DimeNet++ [64], SE(3)-Trans [33], EGNN [102], PaiNN [104], GemNet-T [63], SphereNet [84], SEGNN [4], Equiformer [68]. The evaluation metric is the mean absolute error (MAE). The detailed training tricks are in Appendix B.

Table 2: Results on 6 energy ($\frac{kcal}{mol}$) and force ($\frac{kcal}{mol \cdot Å}$) prediction tasks in MD17 and rMD17 (w/o normalization), and the metric is the mean absolute error (MAE). The data split and complete results are in Appendices B and I.

Model	Energy /Force	MD17						rMD17					
		Aspirin ↓	Ethanol ↓	Malonaldehyde ↓	Naphthalene ↓	Salicylic ↓	Toluene ↓	Aspirin ↓	Ethanol ↓	Malonaldehyde ↓	Naphthalene ↓	Salicylic ↓	Toluene ↓
SchNet	Energy	0.475	0.109	0.3	0.167	0.212	0.149	0.534	1.757	0.26	0.124	2.618	0.119
	Force	1.203	0.386	0.794	0.587	0.826	0.568	1.243	0.449	0.862	0.587	0.878	0.574
EGNN	Energy	17.892	0.436	0.896	12.177	6.964	4.051	17.35	0.402	0.534	12.164	7.794	15.021
	Force	3.042	0.924	1.566	1.136	1.177	1.202	3.825	0.989	1.334	1.183	1.571	1.165
PaiNN	Energy	27.626	0.063	0.102	0.622	0.371	0.165	30.156	1.17	0.07	5.297	5.219	0.045
	Force	0.572	0.23	0.338	0.132	0.288	0.141	0.573	0.316	0.377	0.161	0.321	0.231
GemNet-T	Energy	0.684	4.598	4.966	0.482	0.128	0.098	5.389	1.615	9.496	0.031	21.411	959.745
	Force	0.558	0.219	0.433	0.212	0.326	0.174	0.555	0.233	0.337	0.154	0.371	0.4
SphereNet	Energy	0.244	1.603	1.559	0.167	0.188	0.113	0.304	0.072	0.138	0.093	0.771	20.479
	Force	0.546	0.168	0.667	0.315	0.479	0.194	0.622	0.217	0.5	0.279	2.088	0.254
SEGNN	Energy	17.774	0.151	0.247	0.655	2.173	0.624	15.721	0.13	0.182	1.11	1.494	0.814
	Force	9.003	0.893	1.249	0.895	2.22	1.138	8.549	0.846	1.185	0.926	2.056	1.241
NequIP	Energy	8.333	0.971	2.293	1.032	2.952	1.303	9.618	0.936	2.313	2.089	3.302	1.306
	Force	23.769	5.832	12.099	5.247	14.048	6.8	22.904	6.027	12.372	5.529	15.693	7.094
Allegro	Energy	1.138	0.258	1.33	0.824	1.114	0.441	1.366	1.002	0.417	1.756	1.035	0.437
	Force	3.405	1.412	4.191	3.743	4.934	1.968	3.186	2.799	2.125	3.815	4.781	2.048
Equiformer	Energy	0.308	0.096	0.183	0.097	0.189	0.209	0.375	0.064	0.085	0.069	0.143	0.104
	Force	0.286	0.142	0.23	0.068	0.2	0.08	0.305	0.162	0.24	0.07	0.218	0.077

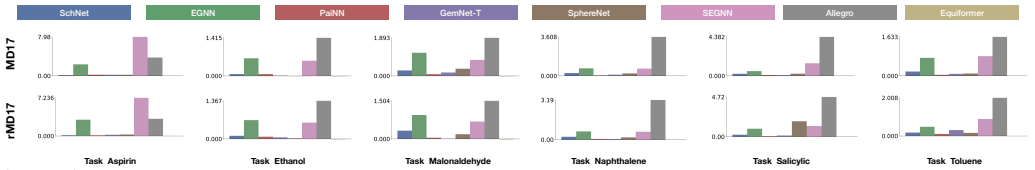


Figure 5: Ablation study on the effect of data normalization. Here are visualizations on performance differences on 6 tasks and 2 datasets, with MAE(force pred w/o normalization) - MAE(force pred w/ normalization).

The results of these 26 models are in Table 1, and two important insights are below: (1) There is no one universally best geometric model, yet PaiNN, GemNet, and SphereNet perform well in most tasks. However, GemNet-T and SphereNet take up to 5 GPU days per task, and PaiNN takes less than 20 GPU hours. (2) The geometric conformation is important for quantum property prediction. The performance of using 3D conformation is better than all the 1D and 2D models *by orders of magnitudes*.

4.2 Small Molecules: MD17 and rMD17

MD17 [8] is a dataset of molecular dynamics simulation. It has 8 tasks corresponding to eight organic molecules, and each task includes the molecule positions along the PES (see Fig. 3(b)). The goal is to predict each atom's energy and interatomic forces for each molecule's position. We follow the literature [64, 84, 103, 104] of using 8 subtasks, 1K for training and 1K for validation, while the test set (from 48K to 991K) is much larger. However, the MD17 dataset contains non-negligible numerical noises [9], and it is corrected by the revised MD17 (rMD17) dataset [10]. 100K structures were randomly chosen for each task/molecule in MD17, and the single-point force and energy calculations were performed for each structure using the PBE/def2-SVP level of theory. The calculations were conducted with tight SCF convergence and a dense DFT integration grid, significantly minimizing the computational noises. The results on MD17 and rMD17 are in Table 2. We select 12 subtasks for illustration, and more comprehensive results can be found in Appendix I. We can observe that, in general, PaiNN and Equiformer perform well on MD17 and rMD17 tasks. We also report **ablation study on data normalization**. NequIP [3] and Allegro [89] introduce a normalization trick: multiplying the predicted energy with the mean of ground-truth force (reproduced results in Appendix J). We plot the performance gap, MAE(w/o normalization) - MAE(w/ normalization), in Fig. 5, and observe most of the gaps are positive, meaning that adding data normalization can lead to generally better performance.

4.3 Small Molecules: COLL

The COLL dataset [34] comprises energy and force data for 140K random snapshots obtained from molecular dynamics simulations of molecular collisions. These simulations were conducted using the semiempirical GFN2-xTB method. To obtain the data, DFT calculations were performed utilizing the revPBE functional and def2-TZVP basis set, which also incorporated D3 dispersion corrections. The task is to predict the energy and force for each atom in each molecule, and we consider 8 most advanced geometric models for benchmarking. The results are in Table 3, and two invariant models (GemNet and SphereNet) reach more optimal performance.

Table 3: Results on energy and force prediction in COLL. 120k for training, 10k for val, 9.48k for test. The metric is the mean absolute error (MAE).

Model	Energy (eV) ↓	Force (eV/Å) ↓
SchNet	0.178	0.130
DimeNet++	0.036	0.049
EGNN	1.808	0.234
PaiNN	0.030	0.052
GemNet-T	0.017	0.028
SphereNet	0.032	0.047
SEGNN	7.085	0.642
Equiformer	0.036	0.030

Table 4: Results on 2 binding affinity prediction tasks. We select three evaluation metrics for LBA: the root mean squared error (RMSD), the Pearson correlation (R_p) and the Spearman correlation (R_S). LEP is a binary classification task, and we use the area under the curve for receiver operating characteristics (ROC) and precision-recall (PR) for evaluation. We run cross-validation with 5 seeds, and the mean and std are reported.

Model	LBA			LEP	
	RMSD ↓	R_P ↑	R_C ↑	ROC ↑	PR ↑
SchNet	1.521 ± 0.02	0.474 ± 0.01	0.452 ± 0.01	0.450 ± 0.03	0.379 ± 0.03
DimeNet++	1.672 ± 0.09	0.550 ± 0.01	0.556 ± 0.01	0.590 ± 0.06	0.496 ± 0.05
EGNN	1.494 ± 0.04	0.503 ± 0.04	0.483 ± 0.05	0.657 ± 0.05	0.559 ± 0.05
PaiNN	1.434 ± 0.02	0.583 ± 0.02	0.580 ± 0.02	0.585 ± 0.02	0.432 ± 0.03
GemNet-T	–	–	–	0.674 ± 0.04	0.565 ± 0.05
SphereNet	1.581 ± 0.02	0.538 ± 0.01	0.529 ± 0.01	0.523 ± 0.04	0.432 ± 0.05
SEGNN	1.416 ± 0.03	0.566 ± 0.02	0.550 ± 0.02	0.574 ± 0.03	0.485 ± 0.03

4.4 Small Molecules & Proteins: LBA & LEP

The binding affinity measures the strength of the binding interaction between a small molecule (ligand) to the target protein. In Geom3D, we consider modeling both the ligands and proteins with their 3D structures. During binding, a cavity in a protein can potentially possess suitable properties for binding a small molecule, and it is called a pocket [107]. Due to the large volume of protein, Geom3D follows existing works [112] by only taking the binding pocket instead of the whole protein structure. Specifically, Geom3D models up to 600 atoms for each ligand and protein pair. For the benchmarking, we consider two binding affinity tasks. (1) The first task is ligand binding affinity (LBA) [117]. It is gathered from [118], and the task is to predict the binding affinity strength between a ligand and a protein pocket. (2) The second task is ligand efficacy prediction (LEP) [32]. The input is a ligand and both the active and inactive conformers of a protein, and the goal is to classify whether or not the ligand can activate the protein's function. The results on two binding tasks are in Table 4, and we can observe that PaiNN, GemNet, and SEGNN are generally outstanding on the two tasks.

4.5 Proteins: EC and Fold

An essential aspect of proteins is their ability to serve as bio-catalysts, known as enzymes. Enzyme Commission (EC) number [60] is a numerical classification scheme that describes the enzyme functionalities. Here we follow a recent work [42] in predicting 37K proteins with 384 EC types. Another protein geometric task we consider is protein folding. It is an important biological task in predicting the 3D structures from 1D amino acid sequences. Here we apply the folding pattern classification task [44, 69], comprising 16K proteins and 1,195 fold patterns. We further consider three testsets (Fold, Superfamily, and Family) based on the sequence and structure similarity [88]. The detailed specifications are in Appendix B. The results of 5 models are in Table 5, and CDConv [27] outperforms other methods by a large margin.

Table 5: Results on EC and Fold classification, and metric is the accuracy. The data splits are in Appendix B.

Model	EC (%)	Fold			
		Fold (%)	Sup (%)	Fam (%)	Avg (%)
GVP-GNN [58]	63.936	34.819	52.711	95.047	60.859
GearNet-IEConv [42]	–	39.694	59.330	98.506	65.843
GearNet [135]	78.836	29.109	43.062	95.991	56.054
ProNet [115]	84.251	52.089	69.378	98.270	73.246
CDConv [27]	86.887	60.028	79.904	99.528	79.820

4.6 Crystalline Materials: MatBench and QMOF

MatBench [20] is created specifically to evaluate the performance of machine learning models in predicting properties of inorganic bulk materials covering mechanical, electronic, and thermodynamic material properties [20]. Here we consider 8 regression tasks with crystal structures, including predicting the formation energy (Perovskites, E_{form}), exfoliation energies (E_{exfo}), band gap, shear and bulk modulus ($log_{10}G$ and $log_{10}K$), etc. Please check Appendix B for more details. Quantum MOF (QMOF) [101] is a dataset of over 20K metal-organic frameworks (MOFs) and coordination polymers derived from DFT. The task is to predict the band gap, the energy gap between the valence band and the conduction band. The results of 8 geometric models on 8 MatBench tasks and 1 QMOF task are in Table 6, and we can observe that the performance of all the models are very close; only PaiNN, GemNet, and Equiformer are slightly better. We also conduct **ablation study on periodic data augmentation.** We note that there are two data augmentation (DA) methods: gathered and expanded. Gathered DA means that we shift the original unit cell along three dimensions, and the translated unit cells will have the *same* node indices as the original unit cell, *i.e.*, a multi-edge graph.

Table 6: Results on the 8 tasks from MatBench and 1 task from QMOF (with optimal DA). The data split and task unit are in Appendix B, and the metric is the mean absolute error (MAE).

Model	MatBench								QMOF
	Per. $E_{\text{form}}\downarrow$ 18,928	Dielectric $\downarrow$ 4,764	$log_{10}G\downarrow$ 10,987	$log_{10}K\downarrow$ 10,987	$E_{\text{exfo}}\downarrow$ 636	Phonons $\downarrow$ 1,265	$E_{\text{form}}\downarrow$ 132,752	Band Gap $\downarrow$ 106,113	Band Gap $\downarrow$ 20,425
SchNet	0.040	0.334	0.081	0.060	65.201	42.586	0.026	0.327	0.236
DimeNet++	0.037	0.357	0.081	0.058	68.685	38.339	0.025	0.208	0.240
EGNN	0.039	0.306	0.089	0.064	78.205	76.143	0.026	0.211	0.256
PaiNN	0.038	0.317	0.080	0.053	67.752	44.602	0.022	0.190	0.235
GemNet-T	0.042	0.325	0.088	0.061	68.425	48.986	0.026	0.186	0.207
SphereNet	0.043	0.388	0.087	0.061	72.987	36.300	0.029	0.217	0.251
SEGNN	0.046	0.360	0.087	0.059	65.052	43.638	0.047	0.330	0.330
Equiformer	0.046	0.280	0.087	0.057	62.977	37.381	0.031	0.216	0.238

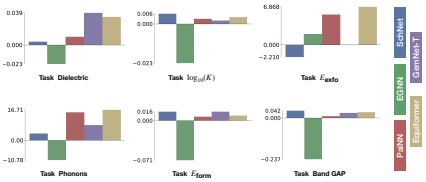


Figure 6: Ablation study on the performance gap with DA: MAE(expanded DA) - MAE(gathered DA).

Table 7: Pretraining results on 12 quantum mechanics prediction tasks from QM9, and the backbone model is SchNet. We take 110K for training, 10K for validation, and 11K for testing. The evaluation metric is the mean absolute error (MAE), and the best and the second best results are marked in **bold** and **bold**, respectively.

Pretraining	$\alpha\downarrow$	$\nabla\mathcal{E}\downarrow$	$\mathcal{E}_{\text{HOMO}}\downarrow$	$\mathcal{E}_{\text{LUMO}}\downarrow$	$\mu\downarrow$	$C_v\downarrow$	$G\downarrow$	$H\downarrow$	$R^2\downarrow$	$U\downarrow$	$U_0\downarrow$	ZPVE$\downarrow$
– (random init)	0.060	44.13	27.64	22.55	0.028	0.031	14.19	14.05	0.133	13.93	13.27	1.749
Supervised	0.062	40.31	25.57	21.69	0.030	0.030	14.36	14.68	0.308	15.21	16.13	1.638
Type Prediction	0.073	45.38	28.76	24.83	0.036	0.032	16.66	16.28	0.275	15.56	14.66	2.094
Distance Prediction	0.065	45.87	27.61	23.34	0.031	0.033	14.83	15.81	0.248	15.07	15.01	1.837
Angle Prediction	0.066	48.45	29.02	24.40	0.034	0.031	14.13	13.77	0.214	13.50	13.47	1.861
3D InfoGraph	0.062	45.96	29.29	24.60	0.028	0.030	13.93	13.97	0.133	13.55	13.47	1.644
GeoSSL-RR	0.060	43.71	27.71	22.84	0.028	0.031	14.54	13.70	**0.122**	13.81	13.75	1.694
GeoSSL-InfoNCE	0.061	44.38	27.67	22.85	**0.027**	0.030	13.38	13.36	**0.116**	13.05	13.00	1.643
GeoSSL-EBM-NCE	0.057	43.75	27.05	22.75	0.028	0.030	12.87	12.65	0.123	13.44	12.64	1.652
3D InfoMax	0.057	42.09	25.90	21.60	0.028	0.030	13.73	13.62	0.141	13.81	13.30	1.670
GraphMVP	0.056	41.99	25.75	**21.58**	**0.027**	**0.029**	13.43	13.31	0.136	13.03	13.07	1.609
GeoSSL-DDM-1L	0.058	42.64	26.32	21.87	0.028	0.030	12.61	12.81	0.173	12.45	12.12	1.696
GeoSSL-DDM	0.056	42.29	**25.61**	21.88	**0.027**	**0.029**	**11.54**	**11.14**	0.168	**11.06**	**10.96**	1.660
MoleculeSDE (VE)	**0.056**	**41.84**	25.79	21.63	**0.027**	**0.029**	**11.47**	**10.71**	0.233	**11.04**	**10.95**	**1.474**
MoleculeSDE (VP)	**0.054**	**41.77**	**25.74**	**21.41**	**0.026**	**0.028**	13.07	12.05	0.151	12.54	12.04	**1.587**

However, expanded DA will assume the translated unit cells have different node indices from the original unit cell. (A visual demonstration is in Appendix A). We conduct an ablation study on the effect of these two DAs, and we plot MAE(expanded DA) - MAE(gathered DA) on six tasks in Fig. 6. It reveals that for most of the models (except EGNN), using gathered DA can lead to consistently better performance, and thus it is preferred. For more qualitative analysis, please check Appendix J.

4.7 Geometric Pretraining on Small Molecules

We run 14 pretraining algorithms, including one supervised pretraining: the pretraining dataset (*e.g.*, PCQM4Mv2 [48]) possess the energy or energy gap label for each conformation, which can be naturally adopted for pretraining. The benchmark results of using SchNet as the backbone model pretrained on PCQM4Mv2 and fine-tuning on QM9 tasks are in Table 7. We observe that MoleculeSDE and GeoSSL-DDM utilizing the geometric denoising diffusion models outperform other pretraining methods in most cases. On the other hand, supervised pretraining (pretrained on energy gap $\nabla\mathcal{E}$) reaches outstanding performance on $\nabla\mathcal{E}$ downstream task, yet the generalization to other tasks is modest. Please check Appendix I for more pretraining results with different backbone models.

5 Conclusion and Future Directions

Geom3D provides a unified view on the SE(3)-equivariant models, together with the implementations. Indeed these can serve as the building blocks to various tasks, such as geometric pretraining (as displayed in Sec. 4.7) and the conformation generation (ClofNet [19], MoleculeSDE [74]), paving the way for building more foundational models and solving more challenging tasks.

Limitations on models and tasks. Geom3D includes 10 2D graph models, 16 geometric models, 14 pretraining methods, and 46 diverse tasks. We would also like to acknowledge there exist many more tasks (*e.g.*, Atom3D [112], Molecule3D [125], OC20 [7]) and more geometric models (*e.g.*, OrbNet [93], MACE [2] and LieTransformer [49]). We will continue adding them in the future.

Multi-modality as future exploration. Recently, there have been quite some explorations on building multi-modal applications on molecules, especially by incorporating textual data [23, 24, 78, 82, 83, 109, 133, 136]. However, these works mainly focus on the 1D sequence or 2D topology, and 3D geometry is rarely considered. We believe that Geom3D can support this for future exploration.

Acknowledgement

The authors would like to thank Zichao Rong, Chengpeng Wang, Jiarui Lu, Farzaneh Heidari, Zuobai Zhang, Limei Wang, and Hanchen Wang for their helpful discussions. This project is supported by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, the Canada CIFAR AI Chair Program, collaboration grants between Microsoft Research and Mila, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund, and a National Research Council of Canada (NRC) Collaborative R&D Project. This project was also partially funded by IVADO Fundamental Research Project grant PRF-2019-3583139727.

References

[1] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021. 2

[2] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e(3)-equivariant atom-centered interatomic potentials, 2022. 9

[3] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):1–11, 2022. 2, 7, 32, 34, 42, 45, 48

[4] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021. 5, 6, 32, 48

[5] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 4, 30

[6] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019. 1

[7] Lowik Chanussot*, Abhishek Das*, Siddharth Goyal*, Thibaut Lavril*, Muhammed Shuaibi*, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 2021. 9

[8] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 7, 22

[9] Anders Christensen and O. Anatole von Lilienfeld. Revised md17 dataset. *Materials Cloud Archive*, 2020. 7

[10] Anders S Christensen and O Anatole von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *arXiv.org*, 2020. 7, 22

[11] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016. 4

[12] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33:13260–13271, 2020. 1, 6, 20

[13] Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, 47(D1):D482–D489, 2019. 23

[14] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. 1

[15] Pierre-Paul De Breuck, Matthew L Evans, and Gian-Marco Rignanese. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on modnet. *Journal of Physics: Condensed Matter*, 33(40):404002, 2021. 24

[16] Mehmet F Demirel, Shengchao Liu, Siddhant Garg, Zhenmei Shi, and Yingyu Liang. Attentive walk-aggregating graph neural networks. *TMLR*, 2022. 1, 6, 20

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[18] Weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3d equivariant graph neural networks, 2023. 27, 30

[19] Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pages 5583–5608. PMLR, 2022. 5, 9, 27, 30

[20] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *arXiv.org*, 6, 2020. 8, 24, 48

[21] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020. 24, 45

[22] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*, 2015. 1, 6, 20

[23] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022. 9

[24] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021. 9

[25] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[26] Thomas Engel and Johann Gasteiger. *Applied chemoinformatics: achievements and future opportunities*. John Wiley & Sons, 2018. 47

[27] Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2023. 8, 32, 47, 48

[28] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Chemrl-gem: Geometry enhanced molecular representation learning for property prediction. *arXiv preprint arXiv:2106.06130*, 2021. 3

[29] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 2

[30] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020. 5, 30

[31] Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*, 2023. 1, 30

[32] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004. 8, 23

[33] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020. 6, 32, 34, 48

[34] Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. 2020. 7, 23

[35] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022. 5, 28, 32, 48

[36] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. 1, 6, 20, 28

[37] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning*, pages 3668–3679. PMLR, 2020. 1

[38] Nate Gruver, Samuel Stanton, Nathan C Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion. *arXiv preprint arXiv:2305.20009*, 2023. 1

[39] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996. 47

[40] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017. 6, 20

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 34

[42] Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. 2020. 8, 23, 32, 48

[43] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023. 1

[44] Jie Hou, Badri Adhikari, and Jianlin Cheng. Deepsf: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34, 2018. 8, 24

[45] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. 1

[46] Elton P Hsu. *Stochastic analysis on manifolds*. Number 38. American Mathematical Soc., 2002. 5, 27

[47] Qian-Nan Hu, Hui Zhu, Xiaobing Li, Manman Zhang, Zhe Deng, Xiaoyan Yang, and Zixin Deng. Assignment of ec numbers to enzymatic reactions with reaction difference fingerprints. *PloS one*, 7(12):e52901–e52901, 2012. 23

[48] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020. 9

[49] Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pages 4533–4543. PMLR, 2021. 5, 9, 30

[50] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019. 5

[51] Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023. 1

[52] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013. 24, 48

[53] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019. 1

[54] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022. 3

[55] Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. 3d equivariant molecular graph pretraining. *arXiv preprint arXiv:2207.08824*, 2022. 3, 5, 32, 48

[56] Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. Energy-motivated equivariant pretraining for 3d molecular graphs. *arXiv preprint arXiv:2207.08824*, 2022. 5, 37

[57] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022. 6

[58] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. 2020. 8, 32, 48

[59] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020. 4

[60] J.M. Enzyme nomenclature: prepared by edwin c. webb, academic press, 1992. £34.00 (xiii + 862 pages) isbn 0 12 227165 3. *Trends in biochemical sciences (Amsterdam. Regular ed.)*, 18, 1993. 8

[61] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1, 5

[62] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1, 6, 20

[63] Johannes Klicpera, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 5, 6, 32, 33, 48

[64] Johannes Klicpera, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020. 5, 6, 7, 22, 32, 48

[65] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020. 33

[66] Mario Krenn, Florian H"ase, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. 6, 20

[67] Greg Landrum et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013. 47

[68] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022. 5, 6, 32, 34, 48

[69] Chen Lin, Ying Zou, Ji Qin, Xiangrong Liu, Yi Jiang, Caihuan Ke, and Quan Zou. Hierarchical classification of protein folds using a novel ensemble classifier. *PloS one*, 8(2):e56499, 2013. 8

[70] Chen Lin, Ying Zou, Ji Qin, Xiangrong Liu, Yi Jiang, Caihuan Ke, and Quan Zou. Hierarchical classification of protein folds using a novel ensemble classifier. *PloS one*, 8(2):e56499–e56499, 2013. 24

[71] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, et al. Dig: A turnkey library for diving into graph deep learning research. *The Journal of Machine Learning Research*, 22(1):10873–10881, 2021. 2

[72] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, Keqiang Yan, Haoran Liu, Cong Fu, Bora M Oztekin, Xuan Zhang, and Shuiwang Ji. DIG: A turnkey library for diving into graph deep learning research. *Journal of Machine Learning Research*, 22(240):1–9, 2021. 2

[73] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *NeurIPS*, 32, 2019. 1, 20

[74] Shengchao Liu, Weitao Du, Zhiming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, 2023. 3, 5, 6, 9, 32, 37, 48

[75] Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022. 3, 5, 37

[76] Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *ICLR*, 2023. 32, 48

[77] Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9977–9978, 2019. 3

[78] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*, 2022. 9

[79] Shengchao Liu, Meng Qu, Zuobai Zhang, Huiyu Cai, and Jian Tang. Structured multi-task learning for molecular property prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 8906–8920. PMLR, 2022. 3

[80] Shengchao Liu, Chengpeng Wang, Weili Nie, Hanchen Wang, Jiarui Lu, Bolei Zhou, and Jian Tang. Graphcg: Unsupervised discovery of steerable factors in graphs, 2023. 1

[81] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021. 3, 5, 32, 37, 48

[82] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023. 9

[83] Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023. 1, 9

[84] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021. 5, 6, 7, 22, 32, 33, 48

[85] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 22

[86] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020. 1

[87] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. 1

[88] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995. 8, 24

[89] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *arXiv preprint arXiv:2204.05249*, 2022. 5, 7, 32, 34, 42, 45, 48

[90] Emmy Noether and M. A. Tavel. Invariant variation problems. 2005. 26

[91] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. 21

[92] Donald H Perkins. *Introduction to high energy physics*. CAMBRIDGE university press, 2000. 5

[93] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R Manby, and Thomas F Miller. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of chemical physics*, 153(12), 2020. 5, 9, 30

[94] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014. 6, 22

[95] Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a General, Powerful, Scalable Graph Transformer. *Advances in Neural Information Processing Systems*, 35, 2022. 1, 6, 20

[96] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837. 2

[97] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. *bioRxiv*, 2021. 1

[98] Roshan M Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. 1

[99] Patrick Reiser, Andre Eberhard, and Pascal Friederich. Graph neural networks in tensorflow-keras with raggedtensor representation (kgcnn). *Software Impacts*, page 100095, 2021. 2, 45, 48

[100] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. 6, 20

[101] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021. 8, 24

[102] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. *arXiv preprint arXiv:2102.09844*, 2021. 5, 6, 32, 34, 48

[103] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. 5, 6, 7, 22, 32, 33, 48

[104] Kristof T Schütt, Oliver T Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *arXiv preprint arXiv:2102.03150*, 2021. 5, 6, 7, 22, 32, 35, 48

[105] Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning*, pages 8818–8827. PMLR, 2020. 1

[106] Tess Smidt, Nathaniel Thomas, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 5, 28, 33, 48

[107] Antonia Stank, Daria B Kokh, Jonathan C Fuller, and Rebecca C Wade. Protein binding pocket dynamics. *Accounts of chemical research*, 49(5):809–815, 2016. 8, 23

[108] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pages 20479–20502. PMLR, 2022. 5, 32, 37, 48

[109] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022. 9

[110] Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Energy-based view of retrosynthesis. *arXiv preprint arXiv:2007.13437*, 2020. 1

[111] Ruoxi Sun, Hanjun Dai, and Adams Wei Yu. Rethinking of graph pretraining on molecular representation. 2022. 39

[112] Raphael JL Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Brandon Anderson, Stephan Eismann, Risi Kondor, et al. Atom3d: Tasks on molecules in three dimensions. *arXiv preprint arXiv:2012.04035*, 2020. 8, 9, 23

[113] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 1, 6, 20

[114] Hanchen Wang, Jean Kaddour, Shengchao Liu, Jian Tang, Matt Kusner, Joan Lasenby, and Qi Liu. Evaluating self-supervised learning for molecular graph embeddings. *arXiv preprint arXiv:2206.08005*, 2022. 39

[115] Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks. 2022. 8, 32, 48

[116] Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[117] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004. 8, 23

[118] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005. 8, 23

[119] Shiyu Wang, Xiaojie Guo, and Liang Zhao. Deep generative model for periodic graphs. *arXiv preprint arXiv:2201.11932*, 2022. 1

[120] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. 6, 20

[121] Hermann Weyl. *Symmetry*, volume 47. Princeton University Press, 2015. 4

[122] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021. 1

[123] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018. 3, 20

[124] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 1, 6, 20

[125] Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs. *arXiv preprint arXiv:2110.01717*, 2021. 9

[126] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019. 1, 6, 20, 33

[127] Huaxiu Yao, Ying Wei, Long-Kai Huang, Ding Xue, Junzhou Huang, and Zhenhui Jessie Li. Functionally regionalized knowledge transfer for low-resource drug discovery. *Advances in Neural Information Processing Systems*, 34:8256–8268, 2021. 3

[128] Huaxiu Yao, Xinyu Yang, Xinyi Pan, Shengchao Liu, Pang Wei Koh, and Chelsea Finn. Leveraging domain relations for domain generalization. *arXiv preprint arXiv:2302.02609*, 2023. 3

[129] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021. 1, 6, 20

[130] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022. 3, 5, 32, 37, 48

[131] Chengxi Zang and Fei Wang. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 617–626, 2020. 1

[132] Anthony Zee. *Group theory in a nutshell for physicists*, volume 17. Princeton University Press, 2016. 4, 26

[133] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022. 9

[134] Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022. 1

[135] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. 2022. 8, 32, 48

[136] Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *bioRxiv*, pages 2023–05, 2023. 9

[137] Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022. 2, 48