Towards Physically Reliable Molecular Representation Learning

Abstract

Estimating the energetic properties of molecular systems is a critical task in material design. Machine learning has shown remarkable promise on this task over classical force-fields, but a fully data-driven approach suffers from limited labeled data; not just the amount of available data lacks, but the distribution of labeled examples is highly skewed to stable states. In this work, we propose a molecular representation learning method that extrapolates well beyond the training distribution, powered by physics-driven parameter estimation from classical energy equations and self-supervised learning inspired from masked language modeling. To ensure reliability of the proposed model, we introduce a series of novel evaluation schemes in multifaceted ways, beyond the energy or force accuracy that has been dominantly used. From extensive experiments, we demonstrate that the proposed method is effective in discovering molecular structures, outperforming other baselines. Furthermore, we extrapolate it to the chemical reaction pathways beyond stable states, taking a step towards physically reliable molecular representation learning.

1 INTRODUCTION

Material simulation is a vast research field that spans understanding material's optimal structure, simulating microscopic dynamics depending on time, temperature, and pressure beyond the experimental resolution, and reducing trial-error loops in designing new materials. The foundation of this simulation is defining the energy at the atomic level considering interactions between numerous atoms, so-called many-body problem. Advances in theory and computational capability, *e.g.*, Density Functional Theory (DFT; Kohn and Sham [1965], Parr [1980]), have led to higher predictability of energy with greater accuracy. Despite the tremendous advances, however, many-body interactions between atoms have exponential complexity over the number of atoms, and it has been a grand challenge in computational material simulations to reduce computational cost while improving the prediction accuracy.

Recently, machine learning approaches have drawn attention as an alternative to classical force fields that rely on physical principles and human intuition. However, pure data-driven approaches often suffer from limited amount and quality of available data. Sometimes one may benefit from simulations, which provide data at larger scale than actual experiments. It still requires, however, expensive and time-consuming DFT or molecular dynamics (MD) simulations, accompanying with significant human analysis due to our limited knowledge.

Furthermore, for some specific system of interest (*e.g.*, a drug candidate), it is often essential to accurately estimate the molecular dynamics across the reaction pathway, not just the stable states before and after the reaction. Molecular structure data, however, are vastly available only at their stable states, while it is extremely costly to collect data on their transition states during a chemical reaction. Therefore, it is vital to have strategies for building a stable model that extrapolates well from stable structures to unstable intermediate ones. If we can train a physically-reasonable model that performs reasonably even at unstable states from a stable-state-only dataset, we may be able to transfer it to the reaction pathway reconstruction problem, which severely suffers from data scarcity.

Another challenge in ML-based molecular modeling is validation. It is often challenging to verify if the model truly learns physically reasonable potential energy surface, which is essential for comprehending molecular structural dynamics and constructing chemical reaction pathways. In previous works, energy estimation accuracy in a stable state has been commonly used, expecting that discovering the actual potential energy surface is needed for the model to

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

precisely estimate its energy. However, since the test cases are confined to stable states, it is questionable whether the model captures the true geometry of the potential energy surface, or has merely fitted to the energy values. In other words, the meta-stability of potential energy surface cannot be verified solely through the stable-state energy estimation. Therefore, additional metrics and evaluation schemes that compensate the current scheme would benefit the community by providing crosscheck validity of existing and future methods.

In this paper, we tackle the aforementioned challenges in molecular structure modeling as follows:

- A natural direction to tackle the data scarcity issue in data-driven models is to incorporate as much physical intuitions and knowledge as possible. In this paper, we propose a *physics-empowered hybrid model for molecular representation learning*, which combines the expressive power of a Transformer [Vaswani et al., 2017] with classical force-field-style equations.
- To build a physically reliable model that generalizes well beyond the steady-state-only training data, we design a *self-supervised learning approach* that the model can learn underlying chemical rules without overly relying on scarcely available labels provided only at stable states. To be specific, we propose an effective *masked atomic modeling* idea, inspired by masked language modeling.
- We examine the possibility of *transfer learning* from our model trained only on stable structures to *chemical reaction pathways*, which requires energy estimation of molecules at transition states, unseen during the training at all. A general understanding of the physical rules would be essential for this challenging generalization problem.
- We design a series of *novel evaluation schemes* to measure reliability of the molecular potential energy surface learned by the model. To be specific, we propose to recover molecular structure from perturbation, to reassemble molecules from broken bonds, and to predict the entire chemical reaction pathway mentioned above. Together with the existing energy estimation accuracy, our evaluation methods verify the models in multifaceted ways, preventing from overfitting to a single objective.

2 RELATED WORK

ML potentials can be categorized into three types based on model complexity and history: kernel-based descriptors, fixed atomic descriptors, and learnable descriptors.

Kernel-based Methods. Kernel-regression-based potentials are mainly applied to a single atom or a few elemental species, where the kernel method is one of the lightest forms. Gaussian approximation potential (GAP; Bartók et al. [2010]), smooth overlap of atomic potential (SOAP; Bartók et al. [2013]), and spectral neighbor analysis potential (SNAP; Chen et al. [2017]) are representative examples. These models can be trained on a small amount of data, but it is difficult to be extended to chemically complex cases.

Fixed descriptors. Behler and Parrinello [2007] uses an atom-centered symmetry function to describe the local environment of each atom and passes each descriptor value to the simple feed-forward network to map the total energy. They estimate the energy for each descriptor from the distance and angle information between paired atoms within a specific cutoff. Behler-Parinello neural-net (BPNN; Behler and Parrinello [2007]) series are the representative practical examples that increase model complexity for high-dimensional Potential Energy Surface (PES) compared to previous kernel-based methods. BPNN was the first realistic attempt to decompose the total energy as a sum of each individual atom's energy. A fundamental limitation of this approach is that fixed descriptors are insufficient to cover complex spatial patterns (*e.g.*, ring structures, bond types, or chemical functional groups), limiting the knowledge transferability between different molecules and atoms. Also, the original symmetry function does not reflect the chemical environment outside the cutoff at all [Kulichenko et al., 2021]. Despite these limitations, it achieved accuracy that no previous classical force field reached. It has been shown to work for systems with many atoms in a dense system with a few species [Behler, 2015, Kulichenko et al., 2021].

Deep Learning Models. Recently, deep neural networks have been actively applied to construct surrogate potentials. Most models in this category allow the chemical environmental information can be transferred between atoms over a greater distance than traditional models, providing a higher degree of freedom. ANI [Smith et al., 2017] extends BPNN by modifying its angular function. Message Passing Neural Networks (MPNN) [Gilmer et al., 2017] is specialized in learning from a graph-structured representation by updating hidden node states using messages from adjacent nodes. MPNN significantly improves accuracy in molecule-related tasks on QM9 dataset [Ruddigkeit et al., 2012, Reymond, 2015, Ramakrishnan et al., 2014], while the increased model capability nest a risk of overfitting [Hawkins, 2004, Zuo et al., 2020]. Since then, various graph-based approaches [Schütt et al., 2018, Gasteiger et al., 2020, Unke and Meuwly, 2019] have been proposed. Recently, the Transformer [Vaswani et al., 2017] is applied to this problem [Cho et al., 2021, Thölke and De Fabritiis, 2022], following its success on natural language processing Devlin et al. [2018] and computer vision [Dosovitskiy et al., 2021, Lu et al., 2019, Sun et al., 2019].

3 THE PROPOSED METHOD

3.1 PROBLEM DEFINITION AND NOTATIONS

Given a molecular structure graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of N atoms constructing the molecule and $\mathcal{E}$ is a set of

bonds between a pair of atoms with direct interaction, we aim at a regression problem to estimate the energy $E_{\text{mol}} \in \mathbb{R}$ of the molecule. The total energy at the molecule level E_{mol} is decomposed into the atomic-level energies, denoted by E_i for each atom $i = 1, ..., N$, where $E_{\text{mol}} = \sum_i E_i$. Each atom i in the molecule is represented by its atomic number $z_i \in \mathbb{R}$, its position $\mathbf{p}_i \in \mathbb{R}^3$ in Cartesian coordinates, and electro-negativity $n_{z_i} \in \mathbb{R}$ of the atom type. We denote the pairwise L_2 distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ between atoms, computed from $\{\mathbf{p}_i\}$. Here, the element $d_{i,j}$ is the radial distance between two atoms i and j. Adjacency matrix that represents bond information of the molecule denoted by $\mathbf{A} \in \{0, 1\}^{N \times N}$.

3.2 ATOM REPRESENTATIONS

We represent each atom based on its atom-wise characteristics and relation with neighboring atoms in the molecule.

Atom-wise Representation. Atom i is represented as an embedding $\mathbf{x}_i^{(\text{self})} \in \mathbb{R}^d$ based on its type z_i, concatenated with its electro-negativity n_{z_i}:

$$\mathbf{x}_i^{(\text{self})} = [E(z_i); n_{z_i}], \tag{1}$$

where E is an embedding layer.

Radial Basis Functions. Inspired by the localized orbitals in DFT, we start with a simple Gaussian basis to represent the relationship between two atoms. For a pair of two atoms i and j in the molecule, we assign n_b basis functions following Unke and Meuwly [2019]:

$$\psi_{i,j,k}(d_{i,j}) \equiv \varphi(d_{i,j}) \exp\left\{-\beta_{z_i,k}\left(\exp(-d_{i,j}) - \mu_{z_i,k}\right)^2\right\} \tag{2}$$

where $i = 1, ..., N$ is the center atom index, $j = 1, ..., N$ is a neighboring atom index, z_i is the atomic number of atom i, and $k = 1, 2, ..., n_b$ denotes the index of the basis for each center atom type z_i. For a predefined distance threshold τ, $\varphi(d) = 1$ if $d < \tau$ and 0 otherwise. With a reasonable n_b, we can enhance expressibility of the model, generating more accurate potentials. $\beta_{z_i,k}$ and $\mu_{z_i,k}$ are the learnable parameters for each atom type z_i, which control the center and width of each individual basis. Finally, a cosine envelope function [Thölke and De Fabritiis, 2022] $\varphi(d_{i,j})$ is applied to guarantee continuity at the cutoff edges, *i.e.*, $\frac{\partial \psi(d)}{\partial d}|_{d=\tau} = 0$:

$$\varphi(d_{i,j}) = \begin{cases} \frac{1}{2}(\cos(\frac{\pi d_{i,j}}{\tau}) + 1) & \text{if } 0 \le d_{i,j} \le \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Neighbor Embedding. We adopt the idea of neighbor embedding [Thölke and De Fabritiis, 2022], which represents relative information from nearby atoms under the distance of some threshold τ, denoted by $\mathbf{x}^{(\text{neighbor})} \in \mathbb{R}^d$:

$$\mathbf{x}_i^{(\text{neighbor})} = \sum_{j=1}^{n_b} \mathbf{U}\left[\mathbf{x}_j^{(\text{self})} \odot \mathbf{V}\boldsymbol{\psi}_{i,j}^0\right], \tag{4}$$

where $\boldsymbol{\psi}_{i,j} = [\psi_{i,j,1}, ..., \psi_{i,j,n_b}] \in \mathbb{R}^{n_b}$, $\mathbf{V} \in \mathbb{R}^{d \times n_b}$ is a projection matrix from radial basis functions to the atomic embedding space, and $\odot$ indicates element-wise multiplication. $\mathbf{U} \in \mathbb{R}^{d \times d}$ is another linear projection matrix. As a result, $\mathbf{x}_i^{(\text{neighbor})} \in \mathbb{R}^d$, the neighbor embedding of atom i, is in the same atomic embedding space. For each atom i, we combine the atomic and neighbor embeddings, then they are projected back to the same dimensionality by $\mathbf{W} \in \mathbb{R}^{d \times 2d}$. That is, $\mathbf{x}_i = \mathbf{W}[\mathbf{x}_i^{(\text{self})}; \mathbf{x}_i^{(\text{neighbor})}]$.

3.3 OUR TRANSFORMER MODEL

As illustrated in Fig. 1, our model is based on a Transformer. Given a molecule as a set of its N atoms, encoded as $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, ..., N$, our model adds an additional `[CLS]` token, denoted by $\mathbf{x}_0 \in \mathbb{R}^d$, to explicitly learn to represent the overall molecule embedding. On this input sequence, the model stacks L Molecular Attention Blocks (MAB) to contextualize each atom representation across the molecule (within the cutoff distance τ). The atom embedding after $\ell = 0, ..., L$ stages of the MABs is denoted by $\mathbf{x}_i^{(\ell)}$. After L blocks, the final sequence of atomic embeddings $\{\mathbf{x}_i^{(L)}\}$ are produced.

From this, we estimate the overall molecule-level energy from them in two popular ways in Transformers. First, we predict the atom-level energy E_i for atom i by passing $\mathbf{x}_i^{(L)}$ through an MLP. That is, $\hat{E}_i = f_{\text{atom}}(\mathbf{x}_i^{(L)})$, where $f_{\text{atom}} : \mathbb{R}^d \to \mathbb{R}$ is an atom-level energy regressor, and then, summation over all atoms $i = 1, ..., L$ gives the molecule-level energy; that is, $\hat{E}_{\text{mol}} = \sum_{i=1}^{N} \hat{E}_i$. Another approach is directly computing the molecule-level energy from the `[CLS]` by $\hat{E}_{\text{mol}} = f_{\text{mol}}(\mathbf{x}_0^{(L)})$, where $f_{\text{mol}} : \mathbb{R}^d \to \mathbb{R}$ is a molecule-level energy regressor. Both approaches are evaluated in Sec. 4. In Sec. 3.4, we will introduce our main approach for this regression to take advantage of domain knowledge from physics.

Details on Molecular Attention Block. Each Molecular Attention Block (MAB) at level ℓ takes a sequence of atomic embeddings $\{\mathbf{x}_i^{(\ell-1)} : i = 0, ..., N\}$ from the previous level. For each atom $\mathbf{x}_i^{(\ell-1)}$ as query and all atoms including i as the context (keys and values), it performs self-attention as in Fig. 1(b). Following TorchMDNet [Thölke and De Fabritiis, 2022], we modify from the vanilla Transformer [Vaswani et al., 2017] to explicitly reflect the relation arisen from the physical distance between two atoms i and j, in addition to the semantic relevance between them modeled by regular Transformers. Specifically, from the radial basis $\boldsymbol{\psi}_{i,j}^0$ [Orr et al., 1996], we compute $\mathbf{D}^K, \mathbf{D}^V \in \mathbb{R}^{N \times N \times m}$, where m is the embedding dimensionality used for query, key, and value. An element $d_{i,j}^K, d_{i,j}^V \in \mathbb{R}^m$ represents physical tendency to attract each other between atom i and j for key-purpose and value-purpose, respectively. These are mapped from the radial basis function $\boldsymbol{\psi}_{i,j}^0$ by a linear layer, followed by SiLU [Elfwing et al., 2018] activation. This relation is

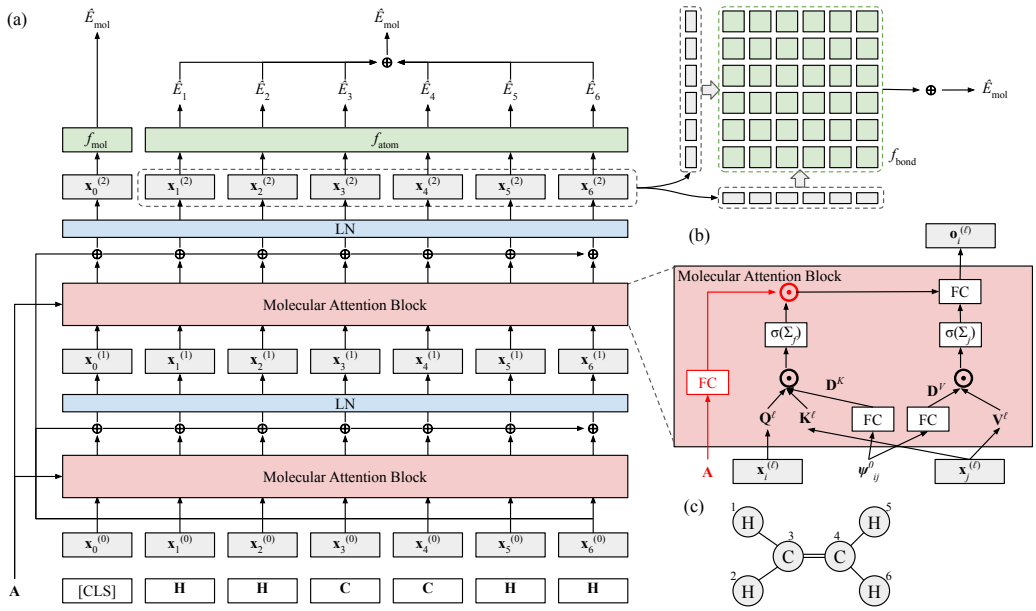

Figure 1: (a) Our model architecture. (b) Detailed Molecular Attention Block. (c) C_2H_4 example.

represented as $\mathbb{R}^m$ instead of a scalar to reflect dimension-wise relationship.

In addition to the changes introduced by Thölke and De Fabritiis [2022], we additionally feed the adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$, followed by a linear layer and SiLU activation. This $\mathbf{A}$-mask is multiplied element-wise with the inferred attention weights, in order to additionally control this semantic relevance based on physical adjacency. For instance, two atoms that are far away will be likely multiplied with a low value, reducing its relationship even if semantic relevance is estimated high. This part is optional, and we provide ablation study in Sec. 4.

3.4 PHYSICS-DRIVEN PARAMETRIC ENERGY PREDICTION

Instead of directly regressing to the atom or molecule energy as described in Sec. 3.3, we propose to design a parametric model that reflects physical insights. For this formulation, we use a form that simultaneously reflects the repulsive and attractive forces between two atoms i, j within the bond energy $E_{i,j}$; namely, Coulomb's law and Lennard-Jones Potential (LJP):

$$E_{i,j} = -\beta_1 \frac{\beta_0}{d_{i,j}} + \beta_2 \left[\left(\frac{\beta_4}{d_{i,j}} \right)^{2\beta_3} - 2 \left(\frac{\beta_4}{d_{i,j}} \right)^{\beta_3} \right]. \quad (5)$$

β_0 corresponds to the influence of charges ($q_i q_j$) between two atoms in Coulomb potential. β_4 is the equilibrium distance between atom i and j, where the repulsive and attractive forces become equivalent, and thus the atom-atom potential energy becomes zero. The energy becomes minimal at this point. β_1 and β_2 are linear coefficients for the Coulomb and LJP parts. It is known that $\beta_3 \approx 6$ under the condition of London dispersion force [London, 1930, Cornell et al., 1995], but the repulsive equivalence of $2\beta_3 \approx 12$ is much more an approximate term (square of the attractive term), so we leave β_3 as an open parameter to be learned from the data. These 5 parameters, denoted by $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]$, are estimated by a regressor $f_{\text{bond}} : \mathbb{R}^{2d} \to \mathbb{R}^5$; that is, $\hat{\boldsymbol{\beta}} = f_{\text{bond}}([\mathbf{x}_i; \mathbf{x}_j])$.

The overall molecule-level energy is calculated by the sum of all pair-wise bond energies and the atomic self-energies; that is, $\hat{E}_{\text{mol}} = \hat{E}_{\text{bond}} + \hat{E}_{\text{atom}}$, where $\hat{E}_{\text{bond}}$ and $\hat{E}_{\text{atom}}$ are defined as

$$\hat{E}_{\text{bond}} = \sum_{i=1}^{N} \sum_{j>i}^{N} \hat{E}_{i,j}, \quad \text{and} \quad \hat{E}_{\text{atom}} = \sum_{i=1}^{N} f_{\text{atom}} \left(\mathbf{x}_i^{(L)} \right). \quad (6)$$

We summarize what to expect from this physics-based modeling as follows. First, we aim to satisfy physical conditions so that the model better extrapolates to unseen cases. Second, by observing the predicted parameters, we can monitor whether the model actually captures the known physical properties of the molecule. Lastly, we expect the model to predict the energy directly from $\hat{E}_{\text{atom}}$ if the given formula is difficult to follow. In Eq. (5), for instance, if the inter-atomic potential does not fit well with LJP, the model assigns $\beta_3 \approx 0$, relying solely on the Coulombic potential.

The model minimizes the MSE loss between the predicted molecule energy $\hat{E}_{\text{mol}}$ and its ground truth E_{mol}; that is, $\mathcal{L}_{\text{energy}} = \|\hat{E}_{\text{mol}} - E_{\text{mol}}\|^2$.

3.5 MASKED ATOMIC MODELING

Masked Language Modeling (MLM), originally introduced by BERT [Devlin et al., 2018] for language modeling, has been successfully utilized as a pre-training task for various

models [Lu et al., 2019, Sun et al., 2019, Zhang et al., 2020a]. The main idea is to randomly mask a subset of tokens and let the model recover them from its contexts, *i.e.*, the other textual or visual tokens in the input sequence. This concept naturally supports self-supervised learning as long as the elements in the sequence are contextually relevant, requiring no human labeling.

In this paper, we propose Masked Atomic Modeling (MAM) in a similar spirit. All chemical materials are composed of multiple atoms, often with more than one type. When a majority of atoms in a valid molecule is known, a set of possible atoms in the rest is significantly reduced when considering the properties of each atom according to the law of chemistry, *e.g.*, the octet rule, Lewis symbol analysis. With MAM, we train our Transformer to discover such chemical restrictions purely by observing a set of valid molecules in the training examples without direct supervision.

Formally, on a sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$ with N atoms, we randomly mask each token by a probability of ρ (we use 0.3, twice as Devlin et al. [2018]), replacing the masked tokens to `[MASK]`. The model is trained to minimize the log loss over the masked tokens:

$$\mathcal{L}_{\text{mask}} = -\log p(\mathbf{X} \otimes \mathbf{m} | \mathbf{X} \otimes (\mathbf{1} - \mathbf{m})), \tag{7}$$

where $\mathbf{m} \in \{0, 1\}^N$ is a binary mask vector for atoms, $\mathbf{1}$ is a one-valued vector, and $\otimes$ indicates row-wise multiplication. p is estimated by a binary classifier, where we use a two-layer MLP.

3.6 COMBINING PHYSICAL CONSTRAINTS

Zero-Force Regularization. When a molecule is in its equilibrium state, the net force on each atom should be at zero. This condition may provide a strong hint for the model to find the valid and optimal molecular structure, but this has not been utilized well in existing studies. Thus, we additionally regularize to minimize the force, computed by the partial gradients of the predicted energy with respect to the 3-dimensional axis (x, y, z). Formally,

$$\mathcal{L}_{\text{force}} = \sum_{i=1}^{N} \|\hat{\mathbf{F}}_i\|^2 = \sum_{i=1}^{N} \left(\frac{\partial E_i}{\partial x}\right)^2 + \left(\frac{\partial E_i}{\partial y}\right)^2 + \left(\frac{\partial E_i}{\partial z}\right)^2,$$

where $\hat{\mathbf{F}} \in \mathbb{R}^3$ is the predicted force of atom i.

Inequality Bound Condition. A stable equilibrium structure of a molecule corresponds to the lowest energy under the given composition. Such an optimal structure can be found by estimating energy from the given structure, differentiating it with respect to the position, and deviating the position based on the force. Naturally, if there is any local deviation from the optimal structure, the energy is always higher than its ground state. This sounds obvious physically, but a machine learning model is unaware of this and thus its estimation may be invalid. Thereby, we apply an additional condition that the energy should be greater than the ground state when locally deviating from the stable structure, to narrow down the solution space. During training, small Gaussian noise with an amplitude of 0.5Å is applied to the optimal structure. This is implemented by an additional loss $\mathcal{L}_{\text{bound}}$ based on the energy inequality condition:

$$\mathcal{L}_{\text{bound}} = \begin{cases} \hat{E}_{\text{mol}} - \hat{E}^*_{\text{mol}} & \text{if } \hat{E}^*_{\text{mol}} \leq \hat{E}_{\text{mol}}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

3.7 OVERALL OBJECTIVE

Combining all together, our model minimizes

$$\mathcal{L} = \mathcal{L}_{\text{energy}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{force}}\mathcal{L}_{\text{force}} + \lambda_{\text{bound}}\mathcal{L}_{\text{bound}}, \tag{9}$$

where λ_{force}, λ_{mask}, and λ_{bound} are coefficients controlling relative importance of each loss term.

4 EXPERIMENTS AND RESULTS

We conduct experiments to answer the following questions: **Q1**. How does our model perform on energy estimation compared to other models? (Sec. 4.2) **Q2**. Do our and baseline models truly comprehend the molecular potential energy surface structure? (Sec. 4.3–4.4) **Q3**. How much physics-driven constraints affect the prediction? (Appendix C)

4.1 EXPERIMENTAL SETTINGS

Datasets. We use three public datasets to evaluate the proposed model. QM9 dataset [Ruddigkeit et al., 2012, Ramakrishnan et al., 2014] is a collection of optimal structures of 130,000 molecules with up to 9 atoms of {C, H, O, N, F}, selected from GDB-17 [Ruddigkeit et al., 2012]. This dataset contains only the stable structure of molecules. We use 80% for training, 5% for validation, and 15% for testing. OC20 dataset [Chanussot et al., 2021] contains stable structures and relaxation trajectories for systems of 15K bulk catalysts and 82 adsorbates. We evaluate our model on two tasks, energy and force regression based on the given structure (S2EF) and relaxed energy prediction with given initial structure (IS2RE). To evaluate performance on non-equilibrium molecular conformations and reactions, we use Transition1x dataset [Schreiner et al., 2022], which contains reaction paths from 10k organic reactions, with 10M molecular conformations.

Baselines. We compare our model to several state-of-the-art energy prediction models: SchNet [Schütt et al., 2018], DimeNet [Gasteiger et al., 2020], TorchMDNet(ET) [Thölke and De Fabritiis, 2022], ForceNet [Hu et al., 2021], and MXMNet [Zhang et al., 2020b].

Evaluation Metric. We report the mean average error (MAE) between the ground truth and predicted energy (MAE_E, in meV/mol) and force (MAE_F, in eV/Å), following existing studies.

More implementation details are provided in Appendix A.

Dataset (Task)	QM9			OC20 (S2EF)	OC20 (IS2RE)	
Model	$\mathrm{MAE_E}$	$\mathrm{MAE_F}$	ΔP	$\mathrm{MAE_F}$	$\mathrm{MAE_F}$	ΔP
SchNet [Schütt et al., 2018]	14.00	2.64	0.47	0.0743	1.059	0.60
CGCNN [Xie and Grossman, 2018]	–	–	–	0.0673	0.988	0.58
MXMNet [Zhang et al., 2020b]	**5.90**	1.83	1.57	–	–	–
DimeNet [Gasteiger et al., 2020]	8.02	1.79	0.58	0.0693	1.012	0.55
ForceNet [Hu et al., 2021]	18.62	0.41	0.21	–	–	–
TorchMDNet (ET) [Thölke and De Fabritiis, 2022]	6.15	1.15	0.32	–	–	–
Ours ($\mathcal{L}_{\text{energy}}$ only)	8.35	1.28	1.23	–	–	–
Ours (full model)	15.16	**0.005**	**0.025**	**0.0549**	**0.887**	**0.10**

Table 1: Comparison with baseline models for energy and force accuracy (in MAE) and average distortion ΔP after structure optimization experiment. Lower values mean better performance for all metrics. We report $\mathrm{MAE_E}$ in meV/mol, $\mathrm{MAE_F}$ in eV/Å, and ΔP in Å.

4.2 COMPARISON WITH BASELINES

In this line of research, the MAE in energy estimation has been most widely used. A primary application for calculating molecular energy is to search for a stable structure and to perform molecular dynamics (MD) simulations of structural changes over temperature and time. All of these works are the foundation for the design and discovery of new materials [Friederich et al., 2021, Louie et al., 2021].

At a glance to the $\mathrm{MAE_E}$ column on QM9 dataset in Tab. 1, we observe that our proposed model estimates the molecule energy comparably with baselines, slightly lagging behind the current state-of-the-art. An underlying assumption for relying on the energy estimation accuracy to evaluate molecule representation learning is that the model would need to understand the actual molecular structure in order to precisely estimate its energy. We raise a question about this assumption: Although energy estimation and structure understanding are positively correlated, the model might overfit to energy estimation if we solely rely on this, optimizing beyond the physical rules permit. This is because, with data-driven approaches, the model is not fully informed with physical constraints and just optimizes over the objective from limited amount of data.

For this reason, we additionally check the MAE in force prediction. Ideally, the net force should be 0 for a molecule in a stable state. If a model has learned the correct molecular structure, the estimated net force should be close to 0 as well. The $\mathrm{MAE_F}$ columns of Tab. 1 report the force estimation accuracy of each model by differentiating the energy with respect to the position. On QM9, our full model precisely estimates zero net force ($\mathrm{MAE_F} \approx 0$), indicating that our $\mathcal{L}_{\text{force}}$ introduced in Sec. 3.6 plays its expected role and the learned force condition generalizes well to the unseen test set.

Interestingly, however, other baselines achieving better energy accuracy, including our model only with $\mathcal{L}_{\text{energy}}$, catastrophically fail to estimate zero net force. This contradicts to the common assumption that precise energy estimation relies on general understanding of the molecular structure and underlying physical rules. This result indicates that overly optimizing only on the single energy criterion leads to break the basic constraints that the models must satisfy for a valid structure, making the achieved energy accuracy meaningless as well.

The rest of Tab. 1 reports performance on S2EF and IS2RE tasks on OC20, comparing against a few baselines using scores reported in Open-Catalyst-Project[1]. Our method is competent on both tasks, outperforming all baselines. Note that the difference in $\mathrm{MAE_F}$ is not as dramatic as in QM9, since both energy and force information are included in OC20 and utilized by all models.

4.3 QUALITATIVE ANALYSIS WITH STRUCTURE OPTIMIZATION

In order to see if the models actually capture the optimal structure of molecules, we design an additional structure optimization experiment. Starting from the stable structures in the dataset, we slightly perturb each atom's position from its original optimum and optimize the structure again, expecting it to converge back to the original optimum. Upon convergence, we measure the average Euclidean distance ΔP of each atom's distortion from its optimal position in the ground truth.

The ΔP columns of Tab. 1 compare the performance of each model on this experiment. Our physics-driven model attains a higher level of accuracy when compared to other models, thus demonstrating its proficiency in learning the potential energy surface of the target molecule. Moreover, it is capable of reproducing a stable structure rather than over-optimizing solely on the energy estimation.

Fig. 2(a) shows optimized structures by baseline models and ours. The left-most column displays the initial stable structures, which the baselines fail to maintain. For instance, in the case of CH_4 (top row), the Hydrogen atoms surrounding the Carbon atom should be arranged symmetrically, but the optimized structures by the baselines lack symmetry. In contrast, our model successfully recovers the optimal structure even in complex scenarios.

Fig. 2(b-c) show the average difference in energy distribution ΔE_g and distance deviation ΔP before and after reoptimization, calculated over 256 molecules (comprising 128 smallest and 128 randomly sampled larger molecules) from QM9. Our model achieves a center value of ΔE_{g} that is two orders of magnitude smaller than other potentials, indicating its superior ability to recover the optimal struc-

[1] https://github.com/Open-Catalyst-Project

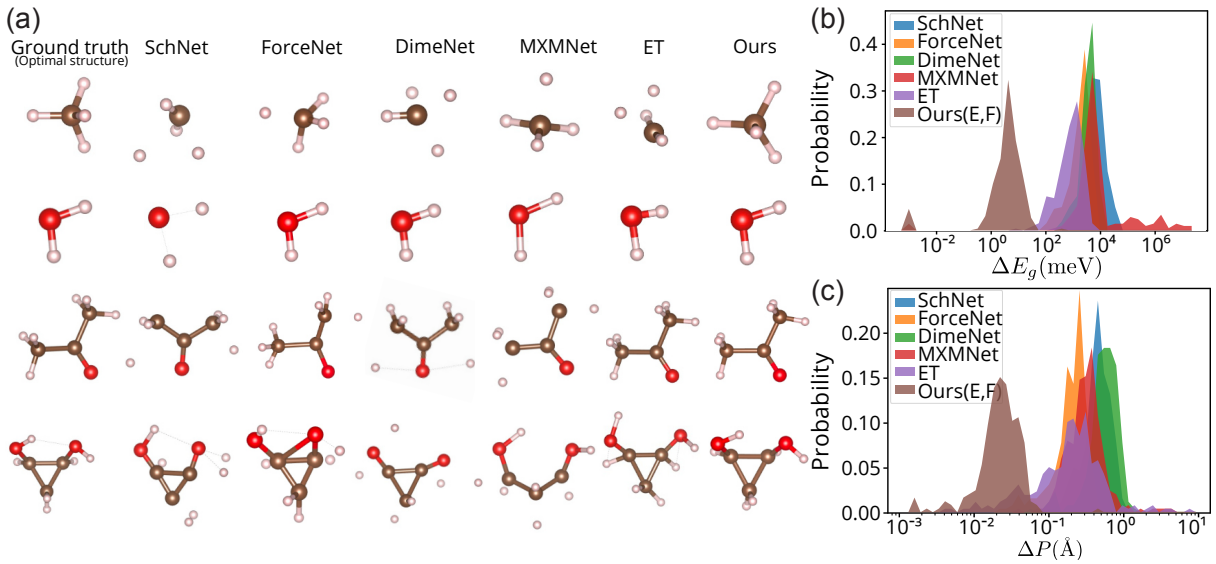

Figure 2: (a) Structural optimization results. The left-most column is the initial stable structure in QM9, followed by recovery results by competing models sequentially. For more structural optimization results, see Appendix Fig. I. (b-c) Distribution of energy difference (ΔE_{g}) and structural change (ΔP) before and after structural optimization, in log scale.

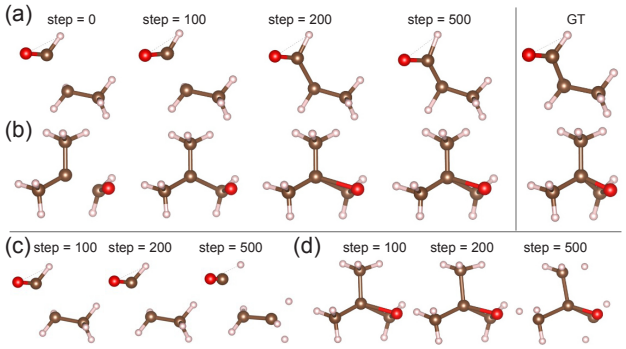

Figure 3: Molecule assembly results on (a) GDB-35 and (b) GDB-87. The original stable structure (GT) is recovered at 500 steps, connecting the broken bond. (c-d) Failure results by our model trained *without bound conditions*.

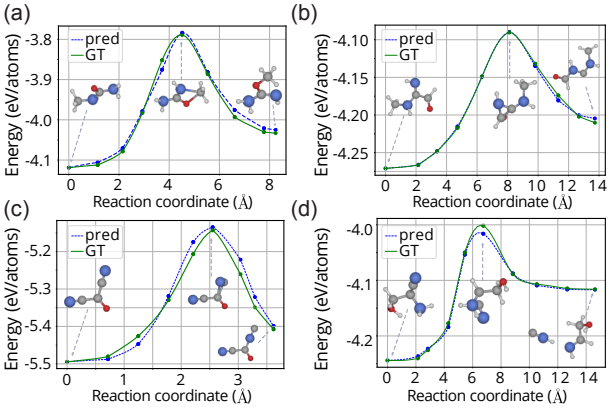

Figure 4: Examples of energy prediction following the reaction pathways on Transition1x. The three structures in each panel correspond to the representative structures along each reaction coordinate: the reactant, transition state, and product structure, respectively.

ture. Also, in Fig. 2(c), the distance deviation ΔP is mostly less than 0.1Å, and our model's ΔP values are at least 10 times smaller than other models. Despite being a challenging task even for molecular dynamics, our model's excellent performance on this stable structure-only dataset like QM9 signifies its capability of capturing fundamental physical principles such as distance symmetry from limited information. Additional examples are presented in Appendix D.

4.4 MOLECULAR ASSEMBLY AND CHEMICAL REACTION PATHWAY PREDICTION

We employ our approach for a couple of additional tasks, including the assessment of potential stability in non-equilibrium structures; namely, the molecular assembly and the chemical reaction pathway prediction. For molecular assembly, the energy profile continuously decreases from the initial structure to the optimal one, whereas chemical reactions require overcoming an activation barrier.

Molecular Assembly. The molecule assembly task presents an additional challenge beyond the structure optimization presented in Sec.4.3, where the objective is to recover the stable structure from an (almost) optimal structure. This task involves breaking one or more bonds in the molecule by moving functional groups far away, and recovering the original stable structure from this completely broken one. To accomplish this, we randomly select one or two functional groups in a molecule and disconnect the bonds between them by translating each towards different directions, with a displacement of 0.7Å. We begin with the distorted structure and optimize it using the energy profile of our model to determine if it can regain the original stable structure. Since

No.	Base	[CLS]	LJP	Mask	Force	Bound	MAE_E ↓	MAE_F ↓	ΔP ↓
1	✓						11.83	0.77	1.76
2	✓	✓					9.03	0.90	1.11
3	✓	✓	✓	✓			9.70	1.91	0.814
4	✓	✓	✓		✓		10.18	0.016	0.141
5	✓	✓	✓			✓	16.34	0.007	0.038
6	✓	✓		✓	✓	✓	20.67	0.004	0.022
7	✓	✓	✓		✓	✓	17.50	0.005	0.027
8	✓	✓	✓	✓		✓	17.34	0.013	0.044
9	✓	✓	✓	✓	✓		9.65	0.015	0.083
10	✓	✓	✓	✓	✓	✓	15.16	0.005	0.025

Table 2: Ablation study results, adding or subtracting components in the loss function. Red figures indicate unacceptably inferior results (MAE_F, $\Delta P \gg 0.1$).

the training dataset does not contain non-equilibrium information, it is challenging for the model to accurately discover the energy values along the pathway in which molecules are combined.

As shown in Fig. 3, only our method succeed in recovering the original structure, while others show catastrophic failure. We experiment with our model without the bound conditions on the same task. Fig. 3(c-d) illustrates that our model also fails in this case. This highlights the importance of the bound conditions to learn a physically reasonable potential, even with a limited dataset consisting only of optimal structures.

Chemical Reaction Pathway Energy Prediction. Lastly, we conduct an even more complex task of predicting energies across the complete chemical reaction pathway, encompassing the structures of reactants, transition states, and products. To accomplish this task, we adopt a transfer learning approach, by initializing the weights from a pre-trained model on QM9 and subsequently fine-tuning on Transition1x. This is because the two datasets provide different angles of information. QM9 contains 13× types of molecules than Transition1x, so the model is pretrained on QM9 to learn general molecular structures at equilibrium state. The model is then fine-tuned to learn the transition dynamics on Trainsition1x, covering fewer types of molecules than QM9.

Fig. 4 shows a few examples of energy profiles, following the reaction pathway on the validation set of Transition1x. Our model accurately predicts not just the energy of the most stable structure (product) but also that of reactant and transition state structures. Slightly higher error in energy estimation is observed near the transition state, but it is not significant enough to alter the activation barrier height that defines the chemical reaction rates. From this result, we conclude that our approach is effective to create a more general potential energy surface from limited information.

4.5 ABLATION STUDY

We conduct ablation study to see which component contributes to improve which metric. We start from a 'Base' model which indicates our Transformer model described in Sec. 3.3 without using any physics-empowered components. Tab. 2 compares multiple configurations of our model using a subset of components. Comparing #1 and #2, the [CLS] token turns out to be effective, reducing the energy error. The rest compares by adding each component separately starting from our base + LJP equation model (#3–5) and by eliminating each component from the full model (#6–10). We observe the following:

- **Mask** plays its role in improving the energy estimation. Comparing #7 and #10, having Mask helps the model to improve MAE_E without affecting MAE_F or ΔP. Solely with Mask (#3), it achieves a nice MAE_E, but its structure is suboptimal implied by inferior MAE_F and ΔP.
- **Bound** condition is the most important component for understanding the overall structure. Without it (#9), ΔP gets significantly worse than the full model (#10), while MAE_E gets (probably illegally) better by focusing more on the energy like baseline models. With Bound only (#5), it achieves reasonable MAE_F and ΔP, which is not possible only with Mask (#3) or Force (#4).
- **Force** affects all metrics slightly at the same time. Without Force (#8), all metrics get slightly worse compared to the full model (#10). With the Force only (#4), however, the ΔP is suboptimal. We conclude that the Bound condition is also needed to get an acceptable ΔP.

Appendix B presents additional ablation study on model size and MAM masking ratio. Also, Appendix C provides additional qualitative analysis on MAM and physics-driven modeling.

5 CONCLUSION

In this study, we present a molecular representation learning approach that harnesses physics-driven parameter estimation from classical energy equations and self-supervised learning via masked atomic modeling. This method addresses the challenges posed by data scarcity and facilitates extrapolation predictions beyond the training distribution.

Furthermore, we introduce a set of innovative evaluation schemes to assess the model's ability to generalize the structure of molecular potential energy surfaces beyond stable-state energies in the training set. Specifically, we evaluate the molecular structure optimization, molecular assembly, and chemical reaction pathway prediction capabilities of the model. Our extensive experiments on multiple benchmark datasets demonstrate that this multifaceted evaluation approach is advantageous, in addition to the widely-used evaluation scheme that relies on energy or force estimation accuracy in stable states, to ensure the reliability of the learned potential energy surface.

To conclude, we take a step towards physically reliable molecular representation learning under limited data availability. Maximally utilizing information in both model design and training would shed light on future research.

REFERENCES

A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.

A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Physical Review B*, 87(18): 184115, 2013.

J. Behler. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.*, 115: 1032–1050, 2015. ISSN 1097-461X.

J. Behler and M. Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98:146401, Apr. 2007.

L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, and Z. Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 2021.

C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong. Accurate force field for molybdenum by machine learning large materials data. *Physical Review Materials*, 1(4):043603, 2017.

Y. Cho, H. Yoon, S. Yi, J. Choo, M. J. Han, J. Lee, and S. Kim. Deep-DFT: A physics-ml hybrid approach to predict molecular energy using transformer. In *NeurIPS Workshop*, 2021.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv:1511.07289*, 2015.

W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117:5179–5197, May 1995. ISSN 0002-7863.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.*, 20:750–761, June 2021. ISSN 1476-4660.

J. Gasteiger, J. Groß, and S. Günnemann. Directional message passing for molecular graphs. In *ICLR*, 2020.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

D. M. Hawkins. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.*, 44:1–12, Jan. 2004. ISSN 0095-2338.

W. Hu, M. Shuaibi, A. Das, S. Goyal, A. Sriram, J. Leskovec, D. Parikh, and C. L. Zitnick. ForceNet: A graph neural network for large-scale quantum calculations. *arXiv:2103.01436*, 2021.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov. 1965.

M. Kulichenko, J. S. Smith, B. Nebgen, Y. W. Li, N. Fedik, A. I. Boldyrev, N. Lubbers, K. Barros, and S. Tretiak. The Rise of Neural Networks for Materials and Chemical Dynamics. *J. Phys. Chem. Lett.*, 12:6227–6243, July 2021.

F. London. Zur Theorie und Systematik der Molekularkräfte. *Z. Physik*, 63:245–279, Mar. 1930. ISSN 0044-3328.

S. G. Louie, Y.-H. Chan, F. H. da Jornada, Z. Li, and D. Y. Qiu. Discovering and understanding materials through computation. *Nat. Mater.*, 20:728–735, 2021. ISSN 1476-4660.

J. Lu, D. Batra, D. Parikh, and S. Lee. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, ICML'10, pages 807–814, Madison, WI, USA, June 2010. Omnipress.

M. J. Orr et al. Introduction to radial basis function networks, 1996.

R. G. Parr. Density functional theory of atoms and molecules. In *Horizons of quantum chemistry*, pages 5–15. Springer, 1980.

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

J.-L. Reymond. The Chemical Space Project. *Acc. Chem. Res.*, 48:722–730, Mar. 2015. ISSN 0001-4842.

L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling*, 52(11): 2864–2875, 2012.

M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci Data*, 9:779, Dec. 2022. ISSN 2052-4463.

K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet - a deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148:241722, June 2018. ISSN 0021-9606, 1089-7690.

J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4): 3192–3203, 2017.

C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019.

Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust, Dec. 2014.

P. Thölke and G. De Fabritiis. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. *ArXiv220202541 Phys.*, Feb. 2022.

O. T. Unke and M. Meuwly. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments and Partial Charges. *J. Chem. Theory Comput.*, 15:3678–3693, June 2019. ISSN 1549-9618, 1549-9626.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

T. Xie and J. C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14): 145301, 2018.

B. Zhang, H. Hu, J. Lee, M. Zhao, S. Chammas, V. Jain, E. Ie, and F. Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv:2011.09046*, 2020a.

S. Zhang, Y. Liu, and L. Xie. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. In *NeurIPS Workshop*, 2020b.

Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A*, 124:731–745, Jan. 2020. ISSN 1089-5639.