PHYSICS-EMPOWERED MOLECULAR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the energetic properties of molecular systems is a critical task in material design. With the trade-off between accuracy and computational cost, various methods have been used to predict the energy of materials, including recent neural-net-based models. However, most existing neural-net models are context-free (physics-ignoring) black-box models, limiting their applications to predict energy only within the distribution of the training set and thus preventing from being applied to the real practice of molecular design. Inspired by the physical mechanism of the interatomic potential, we propose a physics-driven energy prediction model using a Transformer. Our model is trained not only on the energy regression in the training set, but also with conditions inspired by physical insights and self-supervision based on Masked Atomic Modeling, making it adaptable to the optimization of molecular structure beyond the range observed during training, taking a step towards realizable molecular structure optimization.

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 have led to higher predictability of energy with greater accuracy than ever before. Despite the tremendous advances, however, many-body interactions between atoms are exponentially complex, so increasing raw computing power alone has fundamental limitations. Thus, it is unavoidable to put an effort into reducing computational cost, as well as improving the prediction accuracy (*e.g.*, quantum mechanics), which has been a grand challenge in computational material simulations.

Quantum mechanical electron structure simulation is aligned with the direction to enhance accuracy. Specifically, Density Functional Theory (DFT; Kohn & Sham (1965); Parr (1980)) is one of the most successful methods in terms of accuracy, describing and predicting the structure of a material, dynamics based on temperature, phase change, and chemical reaction. However, the number of atoms that DFT can practically handle is limited to a few thousand atoms. Thus, it is complicated to compute the dynamics at a larger scale with DFT, which demands an alternative method to scale up.

Another direction is to reduce the computational cost by expediting computation through approximation while considering many-body interactions. For atoms of much larger scale, so-called classical force field approximations Harrison et al. (2018) have been developed. Various classical force field potentials have been proposed, sharing a standard scheme: formulating the term-by-term energy equation using chemical intuition and performing parameterization for the targeting system. This approach also has drawbacks. First, building a classical force field requires extensive human effort to define energy terms by understanding the target system and parameterization. Also, the potential is described in the bonding terms of the material, so its applicability is restricted to non-reactive materials within similar phases, since the parameter should be changed depending on the chemical environment. Computationally heavier ReaxFF (Senftle et al., 2016; Gomzi et al., 2021) exists, but it focuses on chemical reactions that require parameters tailored to the specific reaction. The force field is challenged to secure accuracy despite a great deal of effort, and generalization to various chemical situations remains a challenging question. Nevertheless, it can mimic potentials with few parameters, since it is based on equations and thus scales better than quantum simulation, which requires energy optimization by self-consistent calculations.

Recently, machine learning has made an impact on the material simulation. Specifically, machine learning (ML) surrogate potentials, which predict the energies of given molecules, have drawn much attention from the scientific community as they provide an alternative solution to the aforementioned inevitable trade-off between cost and accuracy of physics-based models. Several state-of-the-art ML surrogate potentials achieved high accuracy close to that of quantum mechanics while requiring significantly less computational resources than traditional DFT-based methods.

One might argue that we can produce an infinite training set using DFT, and a fully data-driven approach would eventually work well on this problem, just like on other regression problems. However, considering the exponentially increasing complexity of this problem and limited resources, it is desirable to equip the surrogate potential model with domain knowledge from physics.

In this paper, we propose a hybrid approach that combines the powerful expressive power of Transformers (Vaswani et al., 2017) with classical force-field-style equations. In particular, our paper's main contributions are summarized as follows:

- We propose a *physics-empowered molecular representation learning* method, actually preserving the optimal structure instead of simply fitting the single energy value.
- Taking advantage of Transformers, our molecular representations are trained in a self-supervised manner by *Masked Atomic Modeling*, inspired by the approach of Masked Language Modeling.
- We conduct extensive experiments to evaluate the proposed model quantitatively and qualitatively, introducing several novel approaches to actually evaluate the model's ability to capture optimal structure of a molecule.

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 using deep neural networks.

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 of ML. 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 & Parrinello (2007) introduced an ML potential model that uses the atom-centered symmetry function (descriptor) to describe the local environment of each atom and passes each descriptor value to the simple feed-forward neural network to map the total energy. These descriptors (symmetry functions) process distance and angle information between paired atoms within a specific cutoff and produce a single value for each descriptor. Behler-Parinello Neural Network (BPNN; Behler & 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 realistic and the first 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. 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-based 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. As a specific example, ANI (Smith et al., 2017) extended BPNN by modifying its angular function. Gilmer et al. (2017) proposed a Message Passing Neural Networks (MPNN), specialized in learning from graph-structured data by updating hidden node states by combining messages from adjacent nodes. Since then, various graph-based approaches (Schütt et al., 2018; Gasteiger et al., 2020; Unke & Meuwly, 2019) have been proposed. MPNNs significantly improved accuracy in molecule-related tasks on QM9 (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). Recently, the Transformer (Vaswani et al., 2017) is applied to this problem as well (Cho et al., 2021; Thölke & 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 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 consisting 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, 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 L2 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

Each atom first needs to be represented as a vector before we feed it into our Transformer-based model. Here, we describe how we represent each atom in detail.

Atom-wise Representation. Atom i is embedded as a vector $\mathbf{x}_i^{(\text{self})} \in \mathbb{R}^d$ based on its type z_i and 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, and $[;]$ indicates concatenation.

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 & 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 & De Fabritiis, 2022) $\varphi(d_{i,j})$ is applied to guarantee continuity at the cutoff edges, *i.e.*, $\frac{\partial \psi(d)}{\partial d}|_{d=\tau} = 0$:

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

Neighbor Embedding. We adopt the idea of neighbor embedding (Thölke & 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, $\mathbf{x}_i = \mathbf{W}[\mathbf{x}_i^{(\text{self})}; \mathbf{x}_i^{(\text{neighbor})}]$.

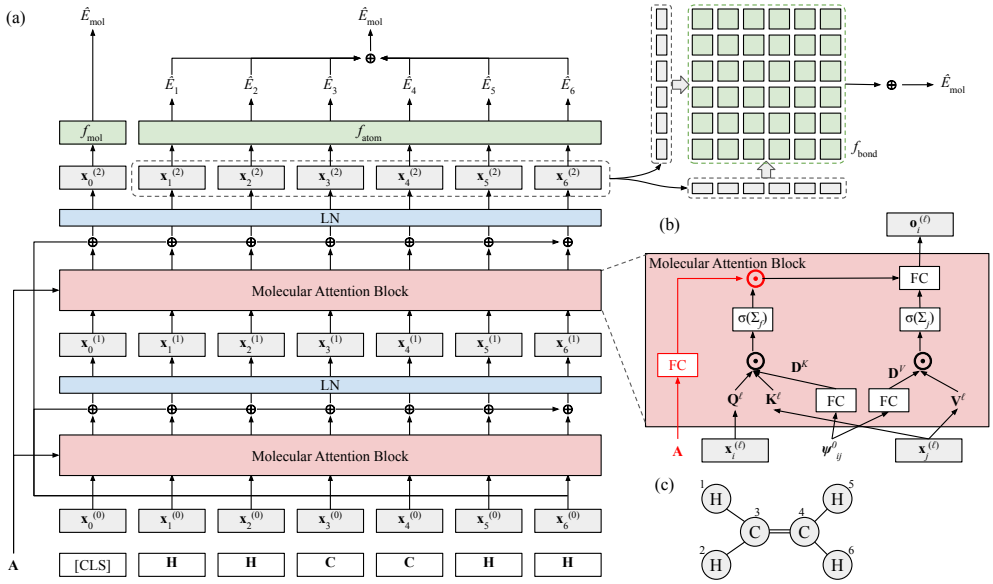

Figure 1: (a) Our model architecture. (b) Detailed Molecular Attention Block. (c) C_2H_4 example.

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), which will be detailed subsequently. We denote by $\mathbf{x}_i^{(\ell)}$ the atom embedding after $\ell = 0, ..., L$ stages of the MABs. After L blocks, the final sequence of atomic embeddings $\mathbf{x}_i^{(L)}$ are produced, and we estimate the overall molecule-level energy from them in two popular ways with Transformers. First, we may predict the atom-level energy E_i for atom i by passing $\mathbf{x}_i^{(L)}$ to a simple linear layer. 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. For f_{atom} and f_{mol}, we use a single fully-connected layer. 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 & 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 represented as $\mathbb{R}^m$ instead of a scalar to reflect dimension-wise relationship.

In addition to the changes introduced by Thölke & 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 by a low value, reducing its relationship even if semantic relevance is estimated high. This part can be optionally turned on or off, 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 simple form that can simultaneously reflect 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):

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

where β_0 corresponds to $q_i q_j$, influence of charges 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 following the 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 given 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

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

We summarize what to expect from modeling to predict the parameters of a physical formulation as follows. First, we aim to make sure that we satisfy physical conditions when we learn from given information and extrapolate to unseen cases. Second, by observing the predicted parameters, we can monitor whether the model actually captures the 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.

Initially, the model is subjected to minimize the MSE Loss between the predicted molecule energy $\hat{E}_{\text{mol}}$ and its ground truth E_{mol}. In other words, the energy loss $\mathcal{L}_{\text{energy}}$ is defined as $\|\hat{E}_{\text{mol}} - E_{\text{mol}}\|^2$.

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 multimodal models and tasks (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 input 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 would like to 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 0.15, same as Devlin et al. (2018)), replacing the masked tokens to `[mask]`. In the resulting matrix, the rows of $\mathbf{X}$ corresponding to the masked tokens, are replaced with the mask embedding. 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,

$$\mathcal{L}_{\text{force}} = \sum_{i=1}^{N} \|\hat{\mathbf{F}}_i\|^2 = \sum_{i=1}^{N} \left(\frac{\partial E_i}{\partial x}\right)^2 + \left(\frac{\partial E_i}{\partial y}\right)^2 + \left(\frac{\partial E_i}{\partial z}\right)^2, \tag{8}$$

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

Inequality Bound Condition. The definition of a stable equilibrium structure is a structure corresponding 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 changing the position based on force information. Naturally, if there is any local deviation from the optimal structure, the energy is always greater than its ground-state energy. This sounds obvious physically, but a machine learning model is unaware of this and its estimation may be invalid. Therefore, we apply an additional condition to narrow down the solution space. The energy should be greater than the ground state when locally deviating from the stable structure. During training, small Gaussian random noise with an amplitude of 0.3 Å is applied to the optimal structure. This is implemented by an additional loss term $\mathcal{L}_{\text{bound}}$ based on the energy inequality condition:

$$\mathcal{L}_{\text{bound}} = \begin{cases} \hat{E}_{\text{mol}} - \hat{E}^*_{\text{mol}} & \text{if } \hat{E}^*_{\text{mol}} \leq \hat{E}_{\text{mol}} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

3.7 OVERALL OBJECTIVE

Combining all together, our model minimizes three loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{energy}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{force}}\mathcal{L}_{\text{force}} + \lambda_{\text{bound}}\mathcal{L}_{\text{bound}}, \tag{10}$$

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**. How does our model perform on a basic structural optimization experiment compared to other models? (Sec. 4.3) **Q3**. How much physics-driven constraints affect the prediction? (Sec. 4.5)

4.1 EXPERIMENTAL SETTINGS

Datasets. QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) dataset is a collection of optimal structures of 130,000 molecules with up to nine atoms of {C, H, O, N, F}, selected from GDB-17 (Ruddigkeit et al., 2012). We use 80% for training, 5% for validation, and 15% for testing.

Baselines. We compare state-of-the-art energy prediction models on the QM9 dataset: SchNet (Schütt et al., 2018), DimeNet (Gasteiger et al., 2020), TorchMDNet(ET) (Thölke & De Fabritiis, 2022), and MXMNet (Zhang et al., 2020b). Except for MXMNet, publicly available pretrained parameters are utilized in the experiment.

Evaluation Metric. We report the mean average error (MAE) between the prediction and the ground truth for the estimation of energy (MAE_{E}, in meV/mol) and force (MAE_{F}, in eV/Å) estimation, following existing studies.

Implementation Details. We try $L \in \{4, 6, 8\}$ to stack Molecule Attention Blocks after the embedding layer. We set the embedding size $d = 256$, which is same as (number of heads) $\times$ n_b. Here, n_b is the same as the dimension of the query, key, and value in the attention block. For activation, we use LeakyRELU (Nair & Hinton, 2010; Sun et al., 2014) function after f_{mol} and ELU (Clevert et al., 2015) after f_{bond}. To enforce the positive base and exponents in the parameterized LJP and to avoid numerical errors, we add $1 + \epsilon$ to β_3, β_4. We set the cutoff threshold $\tau = 5$Å, and the

Model \Metric	MAE_E (↓)	MAE_F (↓)	ΔP (Å) (↓)
SchNet	14.00	2.64	0.47
DimeNet	8.02	1.79	0.58
ET	6.15	1.15	0.32
MXMNet	**5.90**	1.83	1.57
Ours ($\mathcal{L}_{\text{energy}}$ only)	8.35	1.28	1.23
Ours (full model)	14.80	**0.01**	**0.12**

Table 1: Comparison with baseline models for energy and force accuracy (in MAE) and average distortion ΔP after structure optimization experiment.

number of RBFs $n_b = 16$. We use a single linear layer for f_{atom} and f_{bond}, while a two-layer MLP for the MAM task. Specifically, the MLP outputs the estimated likelihood score for 64 atoms for each masked input token. For the overall objective function, we choose weights as $\lambda_{\text{force}} = 0.2$, $\lambda_{\text{mask}} = 0.7$, and $\lambda_{\text{bound}} = 1$. The $\beta_{z_i,k}$ and $\mu_{z_i,k}$ are initialized to $(2n_b^{-1}(1 - \exp(-\tau))^{-2}$ and uniformly within $[0, 1]$, respectively. In the training phase, we used a learning rate of 5×10^{-4} with ADAM optimizer (Kingma & Ba, 2014). We use a warmup of 20 epochs, total training epochs of 900, a decay patience of 24 with a ratio of 0.6. The minimum learning rate was set to 10^{-7}.

4.2 Comparison with Baselines

Tab. 1 compares the performance of our model with baselines on the tasks of energy prediction, force prediction, and structure optimization. We report the performance of our model with two configurations; one with the full model and the other with $\mathcal{L}_{\text{energy}}$ only, by setting λ_{mask}, λ_{force}, and λ_{bound} to zero.

In this line of research, the Energy MAE has been the most widely used metric. At a glance to the first column, Energy MAE, we observe that our proposed model estimates the molecule energy comparably with most baselines, slightly lagging behind the current state-of-the-art, MXMNet.

We ask a question here: does this mean that the models actually understand the molecular structure and estimate the energy precisely from it? We are not able to answer this question solely from the energy MAE, since one energy value in the optimal structure does not contain all information on how the PES is reflected according to the degree of freedom with the position of the atoms. Therefore, we further investigate with additional metrics which reflect the structure or physical conditions: the force and simple structural optimization task.

The second column of Tab. 1 reports the MAE in force estimation of each model by differentiating the potential with respect to the position. Since the QM9 contains only optimal structures, even if force information is not explicitly provided, the force applied to each atom should be all 0. Thus, lower Force MAE means higher accuracy, as long as it is positive. Ideally, lower energy MAE should indicate better molecular structure, and thus it should lead to lower force MAE as well. This is the underlying expectation that the entire research community has solely focused on energy MAE optimization.

Interestingly, we observe that energy MAE and force MAE do not have strong one-to-one correspondence. All existing models and our $\mathcal{L}_{\text{energy}}$ model is optimized only for the energy prediction while showing significantly worse performance on the force estimation. In other words, those previous models are significantly overfitting on the energy estimation, without precisely understanding the molecular structure. However, our full model dramatically reduces the force error through the additional terms based on scientific insights. Although the energy MAE is slightly higher than other models like MXMNet, a difference of about 10 meV/mol is acceptable in terms of the overall quality.

4.3 Qualitative Analysis with Structure Optimization

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).

In order to see if the models actually capture the optimal structure of molecules, we conduct an additional structure optimization experiment. Starting from the stable structures in the QM9, we slightly perturb each atom's position from its original optimum and optimize the structure again, ex-

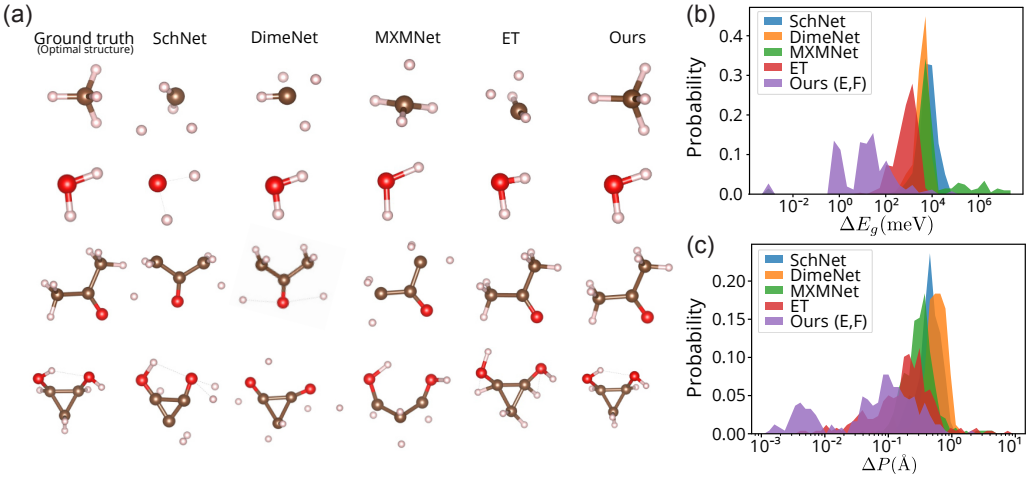

Figure 2: (a) Structural optimization results. The left-most column is the initial stable structure in QM9, followed by recovery results by competing models sequentially. For more structural optimization results, see Appendix Fig. I. (b-c) Distribution of energy difference (ΔE_{g}) and structural change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

Model setup	$\mathrm{MAE_E}$	$\mathrm{MAE_F}$	ΔP
Base	11.83	0.77	1.76
Base + `[CLS]`	9.55	0.90	1.11
Base + `[CLS]` + MAM	9.39	1.27	1.29
Base + `[CLS]` + LJP	9.09	1.09	0.39
Our full model	14.76	0.01	0.12

(a) Ablation study on major components

Layers	4 (Base)		6 (Large)		8 (Huge)	
Method	$\mathrm{MAE_E}$	$\mathrm{MAE_F}$	$\mathrm{MAE_E}$	$\mathrm{MAE_F}$	$\mathrm{MAE_E}$	$\mathrm{MAE_F}$
Base	11.86	0.91	11.83	0.77	11.33	0.72
+ `[CLS]`	11.70	0.78	9.03	0.90	9.70	0.78
+ **A**-mask	9.89	0.98	9.55	1.33	9.33	0.88
+ MAM	10.77	1.43	9.38	1.27	8.35	1.28

(b) Ablation study on SSL methods

Table 2: Ablation Study Results

pecting it to converge back to the original optimal one. Upon convergence, we measure the structure distortions ΔP, in average Euclidean distance, as each atom moves from its optimal structure.

The right-most column of Tab. 1 reports quantitative comparison results of this experiment. Our physics-driven model achieves a more accurate score compared to other models, indicating that our proposed model indeed understands and is capable of reproducing a stable structure instead of simply over-optimizing on the energy estimation.

Fig. 2 (a) illustrates a few examples of optimized structures by baseline models and ours. The initial stable structures in the left-most column are not well-maintained by the baselines. For the simplest CH_4, the Hydrogen atoms around the Carbon atom are expected to be symmetrically arranged, but the optimized structures by the baselines are not symmetric. In more complex molecules, most potentials do not preserve the stable structure.

Fig. 2 (b,c) show the distribution of the energy difference ΔE_g before and after the reoptimization, averaged over 256 molecules (128 smallest and 128 randomly sampled molecules) from the QM9 dataset. Fig. 2 (b) demonstrates the ΔE_{g} in log scale, which clearly shows the difference in the performance of each model. With our model, the center of ΔE_{g} is at least two orders of magnitude smaller compared to other potentials, indicating that our model recovers the optimal structure far better than all other models. Also, in Fig. 2 (c), the distance deviation ΔP is mostly less than 0.1Å, and even compared to other potentials, it can be seen that ΔP is at least 3 times smaller. This structure optimization benchmark is a tough task because the QM9 dataset consists only of stable structures. Furthermore, it is unreasonable to expect molecular dynamics (MD) to work appropriately, reflecting the temperature and dynamics on this stable structure-only dataset. On the contrary, great performance on this challenging optimization task demonstrates that our physics-driven model captures basic physical information, such as distance symmetry, even from limited information.

4.4 Ablation Study

Tab. 2a reports the result of the ablation study on major components we introduce; that is, physics-driven parametric LJP + Coulomb potential, physics-driven regularization (force-preserving loss, inequality condition), and Masked Atomic Modeling (MAM). The 'Base' model indicates our Trans-

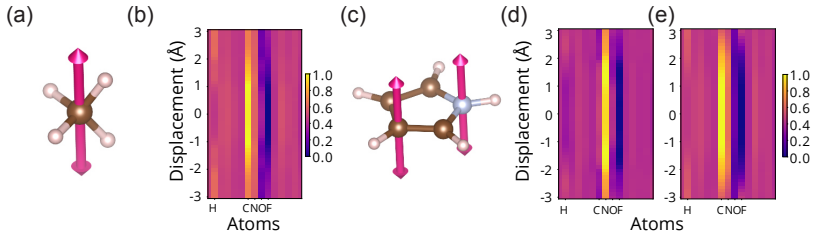

Figure 3: Visualization of MAM. (a), (c) The masked atom is moved along the pink arrow (z-axis), and (b), (d-e) illustrate likelihood score along corresponding positions.

former model described in Sec. 3.3 without using `[CLS]` token. The `[CLS]` token turns out to be effective, reducing the energy error. It is also shown that the energy loss is slightly reduced by adding MAM. The adoption of parametric LJP also reduces MAE, and in particular, we observe that ΔP is significantly reduced in structure optimization tasks. From this, we conclude that physics-driven guidance actually helps the model to learn the structure of molecules. Finally, force information and the inequality-bound condition are utilized in the last row, marked as our full model. In spite of the slight increase in energy MAE, it is clearly demonstrated that the error of ΔP in structural prediction is reduced when learning is performed by utilizing force information and inequality-bound condition. Tab. 2b shows that the $\mathbf{A}$-mask we introduce in Fig. 1 indeed helps in most cases. Also, we observe that using more MABs up to 8 tend to improve the overall performance.

4.5 Qualitative Analysis

Self-supervised Learning with MAM. Fig. 3 illustrates the effect of self-supervised learning with MAM, depending on the position of atoms. For example, Fig. 3 (a) shows the example of CH_4, where we perform MAM inference to figure out an appropriate atom type through the vertical direction. Fig. 3 (b) shows the inferred atom type at each position, from atomic number 1 to 14. The atoms that the QM9 covers, H, C, N, O, and F, are marked in the figure.

Fig. 3 (b) shows that around ± 2Å from the center, the Carbon is strongly favored. On the other hand, Fluorine (F), which is not completely chemically favored, MAM shows a very low affinity. The Nitrogen and Carbon of C_4NH_5 also show a similar trend as shown in Fig. 3 (c-e). In Fig. 3 (e), Carbon is favored by MAM as expected, and interestingly, Nitrogen is also weakly favored, unlike CH_4. Presumably, it is due to the shape of the C_4NH_5 molecule. Note that the amplitude of the atom recommendation through MAM is maximized at the most stable energy position. This reveals that the model self-learns the relationship between surrounding atoms from energy and the positions through MAM. In molecule generation tasks, MAM would be more efficient than randomly connecting atoms and repeating structural optimization iteratively.

Physics-driven Modeling and Regularization. We design our model to predict the parameters of a physics-inspired equation (Sec. 3.4). In Eq. (5), if both β_2 and β_3 are a finite number greater than 0, this implies that the equation is fitted to the distance. In particular, having $\beta_3 \approx 6$ indicates that it has similar behavior to the LJ potential. Since we have two freedoms of Coulomb's terms and LJP-like terms, there is no reason to converge to a single β_3; based on training, β_3 seen to be distributed between 4 and 16, which is close to the 6 of LJ potential.

5 Conclusion

In this work, we propose a physics-driven model and regularization scheme for molecular energy prediction. The physics-driven design enables our model to generate a physically meaningful structure. We also propose a self-supervised learning method inspired by masked language modeling. Our observation indicates that the state-of-the-art models may not be robust under the structure optimization task with the QM9 dataset, which only contains optimized stable structures. We utilize inequality constraints and force information of the optimal conformations, which results in physically more reasonable outcomes. With the combination of physics-driven modeling and regularization, our model outperforms the state-of-the-art models in structure optimization tasks with the QM9 dataset. We also demonstrate that our approach can be used for structure optimization with a few meV scale differences in the initial structure of small molecules. Furthermore, we find that much improvements are needed to achieve a reasonable PES that not only matches a single energy value for large molecules. When constructing an ML potential, it is always challenging to have a sufficient dataset; thus, our approach that maximally utilizes the information in both model design and training would shed light on future research.

REFERENCES

Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13): 136403, 2010.

Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

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

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

Chi Chen, Zhi Deng, Richard Tran, Hanmei Tang, Iek-Heng Chu, and Shyue Ping Ong. Accurate force field for molybdenum by machine learning large materials data. *Physical Review Materials*, 1(4):043603, 2017.

Youngwoo Cho, Hongkee Yoon, Seunghoon Yi, Jaegul Choo, Myung Joon Han, Joonseok Lee, and Sookyung Kim. Deep-DFT: A physics-ml hybrid approach to predict molecular energy using transformer. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2021.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv:1511.07289*, 2015.

Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

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

Pascal Friederich, Florian Häse, Jonny Proppe, and Alán Aspuru-Guzik. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.*, 20:750–761, 2021.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proc. of the International conference on machine learning (ICML)*, 2017.

Vjeran Gomzi, Iva Movre Šapić, and Andrej Vidak. ReaxFF Force Field Development and Application for Toluene Adsorption on MnMOx (M = Cu, Fe, Ni) Catalysts. *J. Phys. Chem. A*, 125: 10649–10656, December 2021. ISSN 1089-5639.

Judith A. Harrison, J. David Schall, Sabina Maskey, Paul T. Mikulski, M. Todd Knippenberg, and Brian H. Morrow. Review of force fields and intermolecular potentials used in atomistic computational materials research. *Applied Physics Reviews*, 5:031104, August 2018.

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

Diederik P Kingma and Jimmy 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, November 1965.

Maksim Kulichenko, Justin S. Smith, Benjamin Nebgen, Ying Wai Li, Nikita Fedik, Alexander I. Boldyrev, Nicholas Lubbers, Kipton Barros, and Sergei 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, March 1930. ISSN 0044-3328.

Steven G. Louie, Yang-Hao Chan, Felipe H. da Jornada, Zhenglu Li, and Diana Y. Qiu. Discovering and understanding materials through computation. *Nat. Mater.*, 20:728–735, June 2021. ISSN 1476-4660.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems (NeurIPS)*, 32, 2019.

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

Mark JL Orr et al. Introduction to radial basis function networks, 1996.

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

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

Jean-Louis Reymond. The Chemical Space Project. *Acc. Chem. Res.*, 48:722–730, March 2015. ISSN 0001-4842.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis 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.

Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet - a deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148:241722, June 2018. ISSN 0021-9606, 1089-7690.

Thomas P. Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B. Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J. Janik, Hasan Metin Aktulga, Toon Verstraelen, Ananth Grama, and Adri C. T. van Duin. The ReaxFF reactive force-field: Development, applications and future directions. *Npj Comput. Mater.*, 2:1–14, March 2016. ISSN 2057-3960.

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

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust, December 2014.

Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. *ArXiv220202541 Phys.*, February 2022.

Oliver T. Unke and Markus 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv:2011.09046*, 2020a.

Shuo Zhang, Yang Liu, and Lei Xie. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2020b.

Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A*, 124:731–745, January 2020. ISSN 1089-5639.

APPENDIX

A ADDITIONAL QUALITATIVE EXAMPLES

We report the structural optimization results of random molecules in the QM9 dataset. We observe that our model and ET mostly preserve the optimal structure, while other baselines significantly destroy structures.

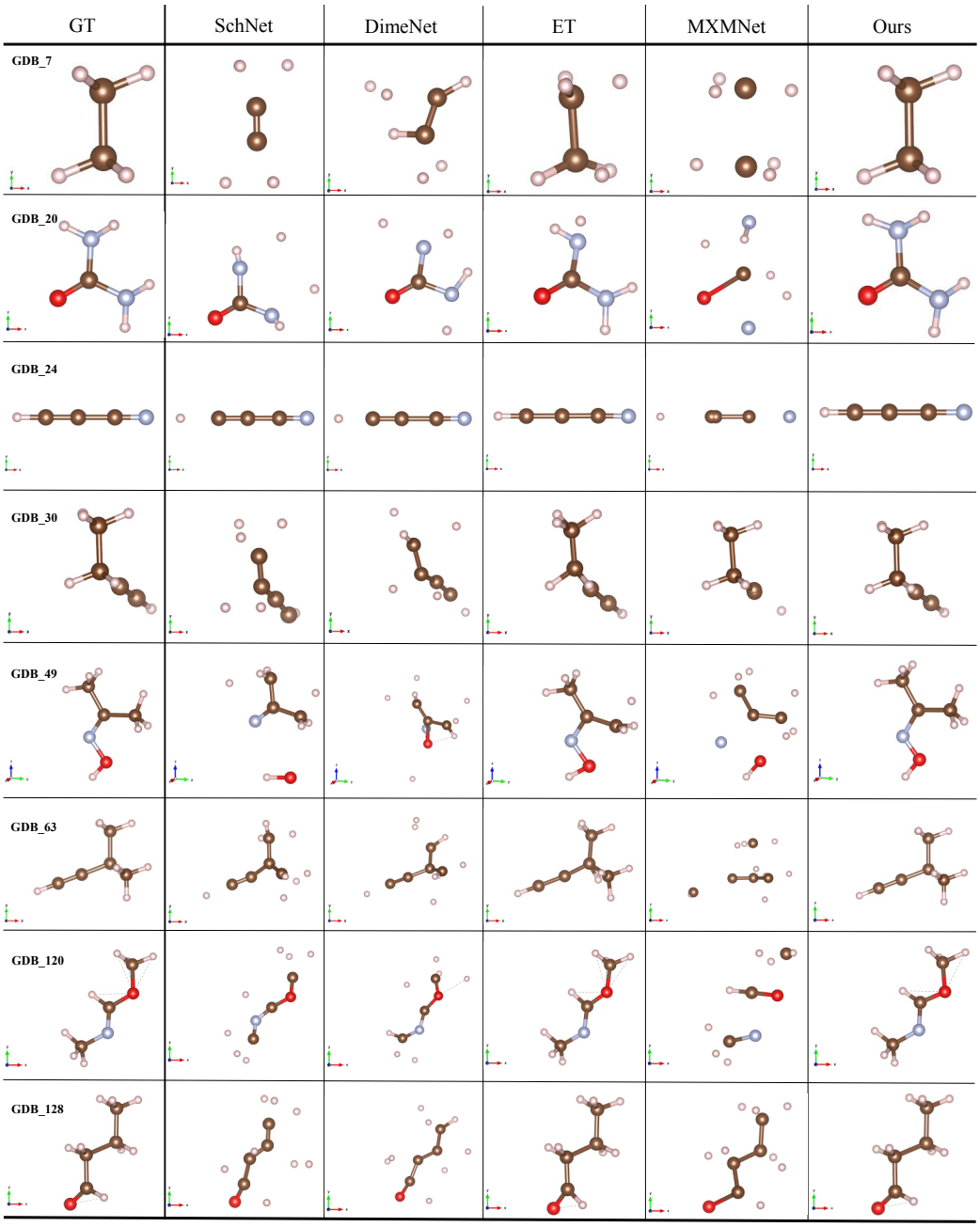

Figure I: Additional structural optimization results by ours and baselines.