Deep-DFT: A Physics-ML Hybrid Approach to Predict Molecular Energy using Transformer

Youngwoo Cho[*]
KAIST
cyw314@kaist.ac.kr

Hongkee Yoon[*]
KAIST
hongkeeyoon7@gmail.com

Seunghoon Yi
Seoul National Univ.
jaguar6182@snu.ac.kr

Jaegul Choo
KAIST
jchoo@kaist.ac.kr

Myung Joon Han
KAIST
mj.han@kaist.ac.kr

Joonseok Lee
Seoul National Univ.
joonseok@snu.ac.kr

Sookyung Kim
Xerox PARC
sookim@parc.com

Abstract

Computing the energy of molecules plays a critical role for molecule design. Classical *ab-initio* methods using Density Functional Theory (DFT) often suffers from scalability issues due to its extreme computing cost. A growing number of data-driven neural-net-based DFT surrogate models have been proposed to address this challenge. After trained on the *ab-initio* reference data, these models significantly accelerate the energy prediction of molecular systems, circumventing numerically solving the Schrödinger equation. However, the performance of these models is often limited to the scope within the training data distribution. It is also challenging to discover physical insights from their prediction due to the lack of interpretability of neural networks. In this paper, we aim to design a physics-ML hybrid DFT surrogate model, which is physically interpretable as well as generalizable to beyond the training data distribution. To achieve these goals, we propose a physics-driven approach to fit the energy to an equation combining Coulomb and Lennard-Jones potentials by first predicting their sub-parameters, then computing the energy product by the equation. Our experimental results show the effectiveness of the proposed approach in its performance, generalizability, and interpretability.

1 Introduction

In computational chemistry, a common method to discover and design molecular structures is minimizing the energy between atoms in a given molecular system. Starting from the experimental geometry of the molecule, we calculate the total energy of the molecule by slightly perturbing the coordinates of each atom. The total energy is computed by *ab-initio* electronic structure calculation using Density Functional Theory (DFT) [9], numerically solving an approximated Schrödinger equation. As this process is computationally expensive, DFT is applied to a limited scale, typically up to 10^4 atoms. One of the grand challenges in computational physics is to design a model that can accelerate the energy calculation of DFT without loss of accuracy.

[*]Indicates equal contribution.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

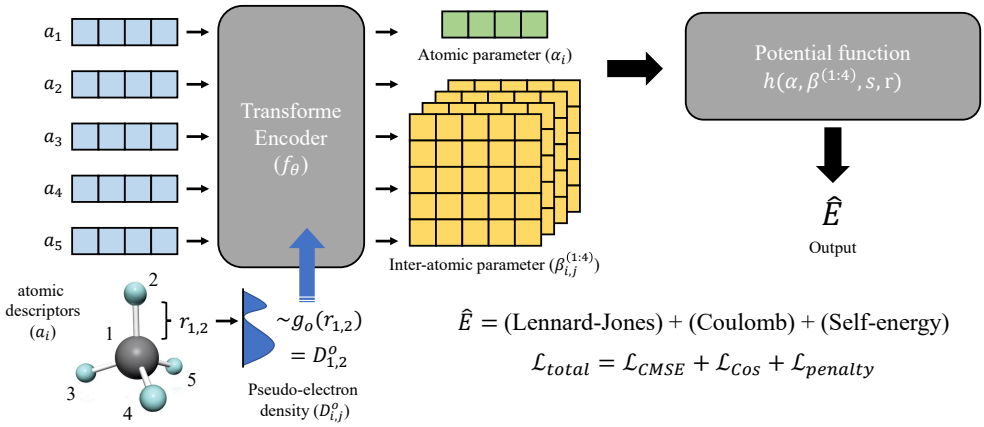

Figure 1: Overview of architecture: Deep-DFT

Recently, neural network models [4, 5] are proposed to learn the molecular representation and to predict the energetic properties of DFT calculation. ANI [13] and TensorMol [17], which are referred to as Behler-Parrinello networks [3], use deep neural networks (DNNs) to learn chemical representation and to predict energies based on hand-crafted atomic descriptors. DTNN [10], SchNet [11], HIP-NN [7], and PhysNet [14] learn chemical representations from nuclear charges and Cartesian coordinates of atoms by message-passing, where a DNN exchanges information between individual atoms by messages learned from the data. After trained on *ab-initio* reference data, these models significantly accelerate the energy prediction of molecular systems, circumventing numerically solving Schrödinger equation. Thanks to their computational efficiency and scalability, these DNN-based energy prediction models increasingly play an important role in molecular design and discovery.

However, there are two main challenges in the existing DFT surrogate models. Firstly, they tend to perform well only within the distribution of the training data. Slight perturbation of the molecular structure often makes the geometry far from the training data distribution, leading to poor energy prediction. Secondly, it is challenging to discover physical insights from the prediction due to the lack of interpretability of DNNs.

To tackle these challenges, we propose **Deep-DFT** to predict DFT energy of a molecule enabling physical interpretation as well as generalization beyond the training data distribution. To achieve these goals, instead of directly estimating molecule energy, we adopt a physics-ML hybrid scheme: fitting to a function that describes the energy of the chemical system combining Coulomb potential and Lennard-Jones potential. In this way, the energy of a molecule is assumed to be a function of inter-atomic distances and sub-parameters of the energy function (Sec. 2.2). Most of the existing neural-net potentials have focused on explaining short-range interactions while ignoring charge effects rather than considering long-range interactions taking into account atomic charge effects. Thus, there are obvious limitations to account for charge effects or long-range interactions from reactions. Recently, a method of indirectly predicting the electronegativity of an atom and calculating the atomic charge based on a neural network has been introduced. To estimate atomic and inter-atomic parameters, we adopt Transformers [15], learning a wide range of contextualized representations for a set of tokens (in our case, the tokens are a set of atom features such as coordinate values or types of atoms). To the best of our knowledge, our model is the first attempt to adopt Transformers to directly acquire a charge [6].

We summarize our contributions as three folds:

- **Generalizablity**: Since we fit the energy with a fixed formation of a function, the predicted energy is unlikely to be far off from the overall physical trend defined in the equation. Therefore, even for an unseen structure beyond the training distribution, the proposed model estimates the energy within a physically feasible range.
- **Interpretablity**: As we obtain a complete mathematical function describing the potential energy of the chemical system unlike existing black-box models, our model conveys fully interpretable results, potentially expected to discover physical insights for scientific research.
- **Expandability**: Our model can utilize model parameters learned from a system with fewer atoms, even when new atoms are added. Therefore, it naturally enables curriculum learning by starting with a small number of atom types and gradually increasing the types of atoms.

2 Proposed Model

As illustrated in Figure 1, the proposed model involves two main components: 1) the **Transformer Molecular Encoder (f_θ)**, which predicts the atomic and inter-atomic parameters from a sequence of atomic descriptors of the target molecule (Sec. 2.1), and 2) the **Parameterized Energy Function (h)**, which computes the energy product from the predicted atomic and inter-atomic parameters (Sec. 2.2). As the energy function is continuous and differentiable, we train f_θ in an end-to-end fashion based on the loss derived from the predicted energy, sub-parameters of the energy function, and the overall distribution of the dataset.

2.1 Transformer Molecular Encoder

Transformers [15] effectively capture interactions among the elements of a sequence with self-attention. Following recent works [12, 8] that applied Transformers to learn the molecular representations from handcrafted descriptors, we adopt Transformers as the basis of our model.

Atomic Descriptors. Based on chemical properties and coordinates within the molecule, each atom i is encoded by atomic descriptors $a_i = \left\{C_i, \{D_{ij}^o\}_{0\leq j<n, o\in\{s,p,d,\ldots\}}\right\}$, where $C_i = \{z_i, s_i\}$ is the set of chemical descriptors with atomic number z_i and atomic self-energy s_i, D_{ij}^o is the set of inter-atomic pseudo-electron density associated with atom i.

Transformer. Transformer encoder [15] f_θ in Figure 1 is composed of l blocks of alternating layers of multi-head self-attention and MLP, with layer-norm [1] in-between and residual connections after every block [16, 2]. As the order in the input sequence does not matter for computing the energy, the positional embedding is removed. Instead, relative positions between atoms are considered in this problem by modifying the self-attention module in the Transformer as follows. The relevance between a query $Q_i \in \mathbb{R}^d$ and a key $K_j \in \mathbb{R}^d$ is weighted by pseudo-electron density D_{ij} between them to reflect orbital information of each atom. For the query, key, value $Q, K, V \in \mathbb{R}^{n\times d}$ with n atoms, the self-attention layer learns the atom representations by the attention operation

$$\text{Attn}(Q, K, V, D) = \text{softmax}\left(\frac{QK^\top \odot D}{\sqrt{d}}\right)V, \tag{1}$$

where d is dimension of the key and $\odot$ indicates element-wise matrix multiplication.

Specifically, we encode inter-atomic distance with the estimated electron density from the radial distribution function (RDF) g_o of orbital o. As RDFs describe how the electron density varies with the distance from each atom, we pass the radial distance of each atom instead of absolute coordinates. This better represents the local atomic environment, invariant under SE(3) transformation, including rotation and translation. We set the number of heads $\eta = |o|$ with the number of orbitals (RDFs).

Atomic and Inter-atomic parameter. In the proposed model, the Transformer encoder takes a sequence of atomic descriptors (*i.e.*, $[a_0, a_1, ..., a_{n-1}]$) as its input, then encodes a molecule as a joint representation of atomic and inter-atomic features. It outputs a sequence $\{X_i\}$ with transformed embeddings from $\{a_i\}$. The atomic feature α_i of the i-th atom and the inter-atomic features $\beta_{i,j}^{1:4}$ between i-th atom and its neighbor atom j are obtained by

$$\alpha_i = \text{ReLU}(\text{FC}(X_i)), \tag{2}$$

$$\beta_{i,j}^m = \sigma\left(X_i^{(m)} \cdot X_j^{(m)}\right) + \text{ELU}\left(X_i^{(m)} \cdot X_j^{(m)}\right) + 1, \tag{3}$$

where σ and ELU are *sigmoid* and *exponential linear unit*, respectively. Note that $X_i^{(m)}$ indicates the m-th head of the i-th embedding (our model uses four heads as four inter-atomic features are used).

2.2 Parameterized Energy Function

Using the predicted atomic parameters α_i and inter-atomic parameters $\beta_{ij}^{(1:4)}$, we compute the total energy using the following potential function $\widehat{E} = h(\alpha, \beta^{(1:4)}, r, s)$:

$$\widehat{E} = \sum_{1\leq i<j<n}\left\{\underbrace{-\beta_{ij}^{(1)}\frac{\alpha_i\alpha_j}{r_{ij}}}_{\text{Coulomb potential}} + \underbrace{\beta_{ij}^{(2)}\tanh\left(\left(\frac{\beta_{ij}^{(4)}}{r_{ij}}\right)^{2\beta_{ij}^{(3)}} - 2\left(\frac{\beta_{ij}^{(4)}}{r_{ij}}\right)^{\beta_{ij}^{(3)}}\right)}_{\text{Lennard-Jones like potential}}\right\} + \sum_{i=1}^{n}\underbrace{s_i}_{\text{Self energy}}. \tag{4}$$

Here, data from different molecules are used all together for training. As the energy variation among different 3D structures is different molecule by molecule, the energy tends to be predicted as the mean energy of all variants of the target molecule if its energy variation is relatively small. To tackle this, we apply two loss terms:

- 1) $\mathcal{L}_{\text{CMSE}}$ **(Calibrated MSE)**: Variants of each molecule are grouped as a set, and we individually predict energy for the structures in the set. For the prediction $\mathbb{E}' = \{E'_0, .., E'_k\}$ and its corresponding ground truth $\mathbb{E} = \{E_1, .., E_k\}$ for k variants, we take the minimum of prediction and ground-truth, and subtract as offset in MSE loss. By doing this, the model better learns the relative variation of energy between different structures in the same molecule.
- 2) $\mathcal{L}_{\text{COS}}$ **(Cosine similarity)**: the cosine similarity between $\mathbb{E}'$ and $\mathbb{E}$ is regularized to be close to 1, making the distribution of ground-truth and that of prediction be similar.

In addition, we add **Rule-based penalty loss** $\mathcal{L}_p$ to induce physics-driven constraints, such as preventing inter-atomic distance from being too small, constraining the charge value α lower than 5, regularizing the distribution of $\beta^{(3)}$ centered around 6, and constraining the sum of all charges to be 0. The overall loss $\mathcal{L}$ of Deep-DFT is given by

$$\mathcal{L} = \mathcal{L}_{\text{CMSE}} + \mathcal{L}_{\text{COS}} + \mathcal{L}_p = \|(E - E_{min}) - (\widehat{E} - \widehat{E}_{min})\|_2 + (1 - \mathbb{E} \cdot \mathbb{E}') + \mathcal{L}_p. \tag{5}$$

After trained on *ab-initio* DFT data for the reference molecule, the learned sub-parameters in the energy function represent quantum-mechanical inter-atomic interactions in the system, conveying meaningful physical insights.

3 Experiments

We perform experiments to answer the following two questions: 1) Can our model be generalized beyond the training distribution of the molecular conformation? 2) Can we explain physics of the molecular system from the obtained energy function?

Experimental Settings. We use ANI-1 dataset [13][2] which consists of more than 20M off equilibrium conformations for 57,462 small organic molecules with C, H, O, N atoms. For effective training, we adopt *curriculum learning* strategy; we first pre-train our model with four smallest molecules, where each containing only the following elements ($\{\text{C}, \text{H}\}$, $\{\text{C}, \text{H}, \text{O}\}$, and $\{\text{C}, \text{H}, \text{O}, \text{N}\}$), followed by training on the full dataset. In the second phase, we iterate training between 'optimal-batch' (taking equilibrium structures only) and 'full-batch' (with both equilibrium and off-equilibrium structures). Figure 2 shows the training curve of the pre-training phase (a) and the second phase (b). After full-training, our model achieves 150 meV of training MSE and 200 meV of test MSE.

Generalizability. To see if our model is generalizable beyond the training distribution, we take a simple molecule (CH_4). We pick one Carbon atom and change its position towards x-axis in 3D Euclidean space, starting from the equilibrium position ($x = 0$, $y = 0$). As shown in Figure 3(a), we pick a Carbon atom and tweak it from equilibrium geometry, moving its position through the x direction.

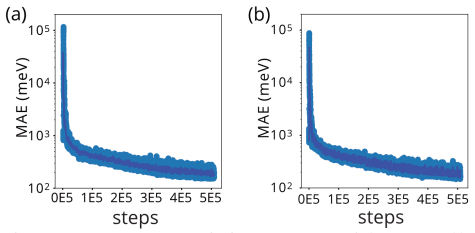

Figure 2: (a) pre-training curve with 4 smallest molecules (b) training curve with full dataset.

As the training data covers within 4 Å off from the equilibrium position, we compare performance of our model against ANI-1 model [13] beyond this scope, illustrated in Figure 3, (b,d) and (c,e), respectively. ANI-1 model fails to predict energy outside of the training range. With 5 Å or farther from the optimal structure, ANI-1 outputs unreasonable total energies that are even lower than that with the optimal structure. On the other hand, our model always predicts physically feasible energy, showing increasing pattern as it gets farther from the equilibrium. This result illustrates that our model is better generalized beyond the training distribution.

Interpretability. The final equation obtained from our model contains complete information describing soft repulsive and attractive interactions between each atom in the target molecular system. Specifically, the complicated inter-atomic dynamics and interplay between Coulomb force and

[2] `https://github.com/isayev/ANI1_dataset`

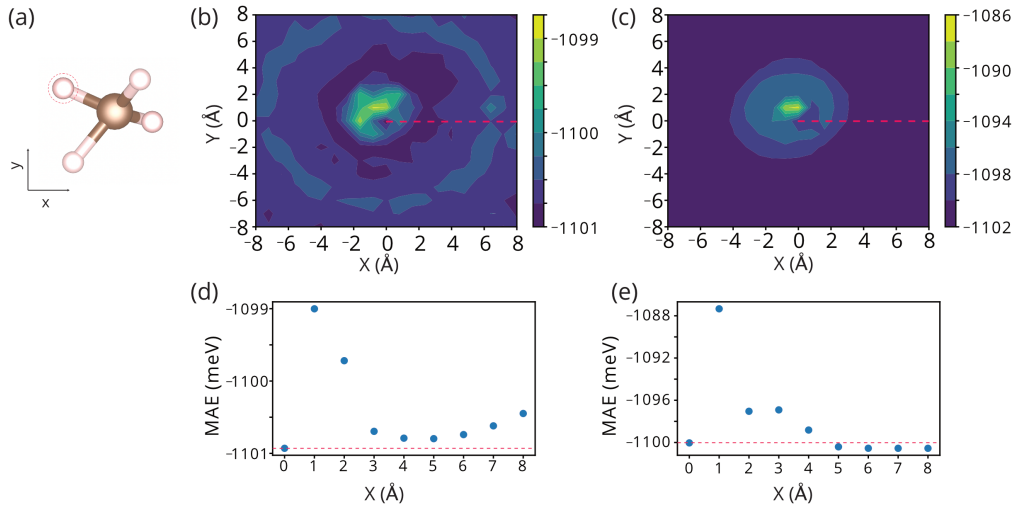

Figure 3: (a) 3D geometry of equilibrium CH_4 molecule. A picked Carbon atom to move is denoted with the dotted circle. {(b),(d)} Energy prediction (in meV) from our model with movement of the picked Carbon atom, along the red dotted lines in (b) from an equilibrium on $x = 0$ and $y = 0$. {(c),(e)} Energy prediction by ANI-1 [13] with the same movement, along the red dotted lines in (c).

Lenard-Jones force can be easily understood by plotting the energy vs. inter-atomic distance. Physical parameters critical to understand the target system can also be easily obtained from the equation itself. For example, the inter-atomic bond distances in the equilibrium state can be analytically computed with r_{ij} value at the minimum energy. Coulomb's constant $\beta_{ij}^{(1)}$ (C-H: $\sim -4.1 \times 10^3$ meV $\cdot$ Å$\cdot\ e^{-2}$, H-H: 0.0 meV $\cdot$ Å$\cdot\ e^{-2}$), charge values α_i, α_j (C: 0.14, H: -0.03), and dispersion energy $\frac{\beta_{ij}^{(2)}}{4}$ (C-H: $\sim$ 0.15 meV, H-H $\sim$ 0.47 meV) can be extracted from the equation itself.

The stable methane is a representative non-polar system, so it shows that the distance of the H-H bond does not need to be counted the Coulomb interaction. It also shows that only a minimal charge is assigned to capture some Coulomb interaction when the C-H bond is away from the ground state. The extracted numbers are used to verify the validity of our model from physically interpretable results.

Expandability. Our model is scalable to the number of atom types regardless of whether it is a binary element or a ternary element or even more, at the design time. Specifically, it can be achieved by replacing the number of electrons in each atom with a one-hot vector in the input sequence to the Transformer. This one-hot vector encoding potentially enables our model to learn new types of atoms flexibly. For example, after learning a molecule composed only of C and H, it is possible to start training with the same model parameter for molecules with C, H, N or with C, H, O, N. This not only enables curriculum learning but is also advantageous when applied to complex atom-type composition systems. Our model has advantages over many other descriptor-based neural network potential models that reveal difficulty in utilizing pre-trained information when the composition of elements changes.

4 Summary and Impact

We propose a novel DFT emulator beyond physically naive machine learning models by applying physical regularization. It is also the first attempt which the ML model directly predicts the atomic charge. In the context of this work, physical regularization is the process of applying knowledge in Physics into an otherwise physically naive neural-network models. By regularizing the model to fit into a physical equation, the model performs better on unseen test sets beyond the scope of training distribution. The proposed model is fully **generalizable** which can predict energy outside of the training regime, and **interpretable** which can provide human understandable mathematical equations. Therefore, our model can be used as a novel tool providing physical insights for scientific discovery.

Acknowledgement

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021H1D3A2A03038607, Brain Pool Plus Program) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)). The corresponding author is Sookyung Kim (sookim@parc.com).

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

[2] A. Baevski and M. Auli. Adaptive input representations for neural language modeling. *arXiv:1809.10853*, 2018.

[3] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[4] Y. Cho, S. Kim, P. P. Li, M. P. Surh, T. Y.-J. Han, and J. Choo. Physics-guided reinforcement learning for 3d molecular structures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] S. Kim, P. Li, J. T. Kim, P. Karande, and Y. Han. Optimizing 3D structure of h_2o molecule using DDPG. In *International Conference on Machine Learning (ICML)*, 2019.

[6] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nature communications*, 12(1):1–11, 2021.

[7] N. Lubbers, J. S. Smith, and K. Barros. Hierarchical modeling of molecular energies using a deep neural network. *The Journal of chemical physics*, 148(24):241715, 2018.

[8] Ł. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, and S. Jastrzębski. Molecule attention transformer. *arXiv:2002.08264*, 2020.

[9] R. G. Parr. Density functional theory of atoms and molecules. In *Horizons of Quantum Chemistry*. Springer, 1980.

[10] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1):1–8, 2017.

[11] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[12] B. Shin, S. Park, K. Kang, and J. C. Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, pages 230–248. PMLR, 2019.

[13] 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.

[14] O. T. Unke and M. Meuwly. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NIPS)*, 2017.

[16] X. Wang, Z. Tu, L. Wang, and S. Shi. Self-attention with structural position representations. *arXiv:1909.00383*, 2019.

[17] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, and J. Parkhill. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical science*, 9(8):2261–2269, 2018.