JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

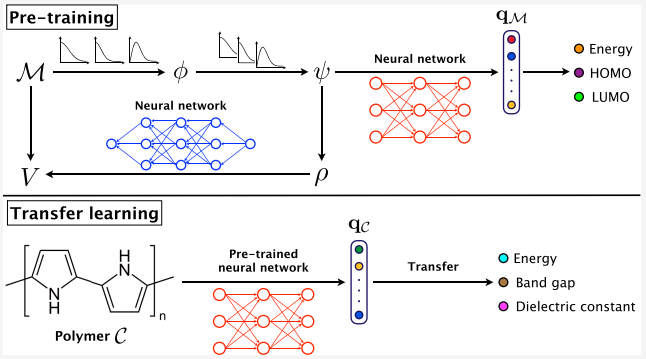
# Quantum Deep Descriptor: Physically Informed Transfer Learning from Small Molecules to Polymers

Masashi Tsubaki* and Teruyasu Mizoguchi

Cite This: J. Chem. Theory Comput. 2021, 17, 7814−7821 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** In this study, we propose a physically informed transfer learning approach for materials informatics (MI) using a quantum deep descriptor (QDD) obtained from the quantum deep field (QDF). The QDF is a machine learning model based on density functional theory (DFT) and can be trained with a large database of molecular properties. The pre-trained QDF model can provide an effective molecular descriptor that encodes the fundamental quantum-chemical characteristics (i.e., the wave function or orbital, electron density, and energies of a molecule) learned from the large database; we refer to this descriptor as a QDD. We show that a QDD pre-trained with certain properties of small molecules can predict different properties (e.g., the band gap and dielectric constant) of polymers compared with some existing descriptors. We believe that our DFT-based, physically informed transfer learning approach will not only be useful for practical applications in MI but will also provide quantum-chemical insights into materials in the future. All codes used in this study are available at https://github.com/masashitsubaki.






## INTRODUCTION

Quantum-chemical calculations, such as those in density functional theory (DFT),[1] can be approximated using machine learning (ML) approaches, including kernel methods, Gaussian process regression, and neural networks (NNs).[2−6] These approaches learn and predict the wave function $\psi$, electron density $\rho$, and properties of simple molecules (e.g., $H_2$ and $H_2O$)[2,6] or configurations of molecules (e.g., $C_2H_4$ and $C_4H_{10}$).[3] It is hoped that such models are transferable to more complicated molecular situations; Grisafi et al., 2018,[3] trained a GPR model for electron density prediction on small molecules and applied it to larger ones (e.g., $C_8H_{10}$ and $C_8H_{18}$). Capturing the fundamental quantum-chemical characteristics of molecules allows ML models to enhance not only transferability[7−9] but also generality and universality. That is, knowledge of $\psi$ and $\rho$, besides enabling extrapolation of the learned properties to large molecules,[10] allows prediction of different properties, such as the vibrational energy and dipole moment (e.g., SchNOrb[5]).

Very recently, Tsubaki and Mizoguchi[11] proposed the quantum deep field (QDF), an ML model based on DFT. The QDF model involves the linear combination of atomic orbitals (LCAOs) and two feedforward NNs: one learns the atomization energy $E$ in a supervised fashion, and the other learns the Hohenberg−Kohn map[2,12] in an unsupervised fashion, ensuring one-to-one correspondence between $\rho$ and the external potential $V$ (see Figure 1a). All parameters in the LCAO and both NNs are trained on the $E$ values of more than 130,000 small organic molecules provided by the QM9 database.[13] The Hohenberg−Kohn map serves as a physical constraint on $\psi$ and $\rho$; therefore, the QDF can be viewed as a self-consistent learning machine with a physically informed inductive bias[6,14−21] that allows the ML model to extrapolate its predictions of $E$ to molecules not appearing in the training database, even ones that are totally unknown, have a larger size, or have a different structure.

In this study, instead of treating the pre-trained QDF as an ML model for learning/predicting the properties of molecules, we assume that its final layer encodes the fundamental quantum-chemical characteristics (including $\psi$ and $\rho$) of molecules, learned from a large number of molecules in the QM9 database. The final layer, denoted as $\mathbf{q}_{\mathcal{M}} \in \mathbb{R}^N$ in Figure 1a, provides a fixed $N$-dimensional vector for any atomistic system, regardless of the size and structure of the molecule, and this vector can be reused as an input molecular descriptor for other tasks in materials informatics (MI). We refer to $\mathbf{q}_{\mathcal{M}}$ as the QDD and reuse it for transfer learning[22−25] of much larger molecules, such as the polymer in Figure 1b. We show that the

Received: June 7, 2021
Published: November 30, 2021

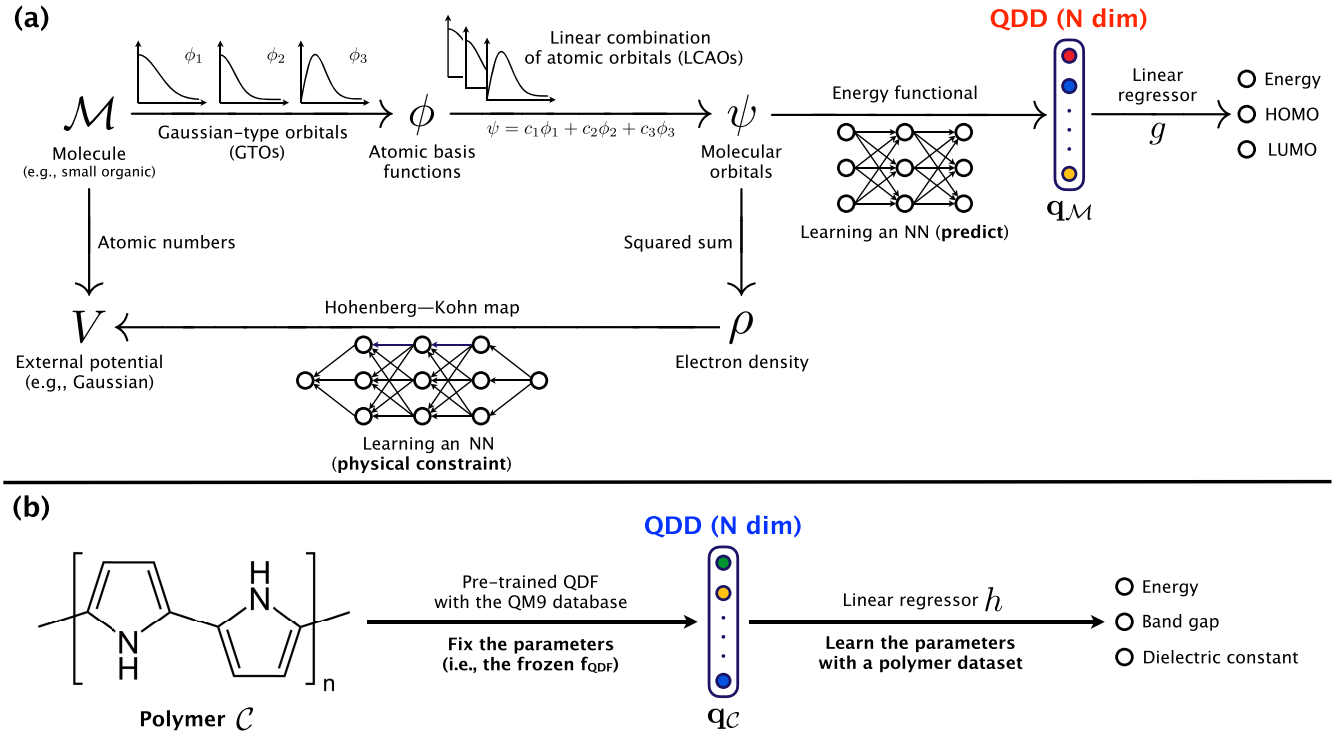



**Figure 1.** (a) Computational components in the QDF model, which inputs a molecule $\mathcal{M}$ and learns its properties. The final layer, which we refer to as the quantum deep descriptor (QDD), is an $N$-dimensional vector $\mathbf{q}_{\mathcal{M}}$ that includes the fundamental quantum-chemical characteristics (i.e., the molecular orbital $\psi$, electron density $\rho$, and target properties) of $\mathcal{M}$. (b) Given a new molecule $C$, such as a polymer, the QDD $\mathbf{q}_C$ obtained by the pre-trained $f_{\mathrm{QDF}}$ can be reused as the molecular descriptor. We replace the regressor $g$ by another regressor $h$ on $\mathbf{q}_C$ and then learn the new $h$ for predicting new properties.

QDD, which was pre-trained using the atomization energy, highest occupied molecular orbital (HOMO), and lowest unoccupied molecular orbital (LUMO) of small molecules simultaneously, can predict not only the atomization energy and band gap but also different properties (e.g., dielectric constants) of polymers.

In the following, we first view the QDF as a model to learn/predict molecular properties and then as a model to generate QDDs. Finally, we describe transfer learning with the QDD for polymer property prediction and discuss its results.

## TRAINING THE QDF

**Molecular Data Definition.** First, we consider a training database (e.g., QM9) denoted by $\mathcal{D}_{\mathrm{train}} = \{(\mathcal{M}_1, E_1), (\mathcal{M}_2, E_2), ..., (\mathcal{M}_D, E_D)\}$, where $\mathcal{M}_d$ is the $d$th molecule, $E_d$ is a property (e.g., the atomization energy) of, and $D$ is the number of data samples. Each molecule is defined as $\mathcal{M} = \{(a_1, \mathbf{R}_1), (a_2, \mathbf{R}_2), ..., (a_M, \mathbf{R}_M)\}$, where $a_m$ is the $m$th atom type, $\mathbf{R}_m$ is the 3D coordinate vector of $a_m$, and $M$ is the number of atoms in $\mathcal{M}$.

Given a molecule $\mathcal{M} \in \mathcal{D}_{\mathrm{train}}$, we first place a sphere on each atom, where the sphere has radius $s$ Å. Then, we divide each sphere into grids in intervals of $g$ Å. Therefore, we have a set of grid points of $\mathcal{M}$ denoted by $\mathcal{G}_{\mathcal{M}} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_G\}$, where $\mathbf{r}_j$ is the 3D coordinate vector of the $j$th grid point and $G$ is the number of points. We provide some molecular grid examples in the Supporting Information.

**Linear Combination of Atomic Orbitals.** Given a molecule $\mathcal{M} = \{(a_m, \mathbf{R}_m)\}_{m=1}^{M}$ and its field $\mathcal{G}_{\mathcal{M}} = \{\mathbf{r}_j\}_{j=1}^{G}$, we consider the set of $N$ molecular orbitals at the $j$th grid point, that is, $\{\psi_1(\mathbf{r}_j), \psi_2(\mathbf{r}_j), ..., \psi_N(\mathbf{r}_j)\}$. The LCAO method provides $\psi_n(\mathbf{r}_j)$, the $n$th molecular orbital at $\mathbf{r}_j$, as follows

$$\psi_n(\mathbf{r}_j) = \sum_{i=1}^{N} c_{ni}\phi_i(\mathbf{r}_j - \mathbf{R}_i) \; s.\,t. \sum_{i=1}^{N} {c_{ni}}^2 = 1 \tag{1}$$

where $c_{ni}$ is the $i$th coefficient, $\phi_i(\mathbf{r}_j - \mathbf{R}_i)$ is the $i$th atomic basis function (with origin $\mathbf{R}_i$), and $N$ is the number of atomic basis functions. In this study, Gaussian-type orbitals (GTOs) were used as atomic basis functions

$$\phi_i(\mathbf{r}_j - \mathbf{R}_i) = D_{ij}^{(q_i - 1)} e^{-\zeta_i {D_{ij}}^2} \tag{2}$$

where $D_{ij} = \|\mathbf{r}_j - \mathbf{R}_i\|$, $q_i$ is the principal quantum number, and $\zeta_i$ is the orbital exponent. This study used the 6-31G basis set, but the GTO in eq 2 is simplified in terms of the spherical harmonics (i.e., only the radial parts of the orbitals were considered). For a calculation that involves 2px, 2py, 2pz, and 3d orbitals, we must learn/optimize the three-dimensional orbital orientations of every atom in the molecule; this is difficult in the current learning framework of the QDF, and we reserve attempting it for future work.

Finally, we represent the $N$ molecular orbitals $\{\boldsymbol{\psi}_n(\mathbf{r}_j)\}_{n=1}^{N}$, as a single $N$-dimensional vector $\boldsymbol{\psi}(\mathbf{r}_j)$. Therefore, the LCAO at the grid point $\mathbf{r}_j$ can be written as

$$\boldsymbol{\psi}(\mathbf{r}_j) = \sum_{i=1}^{N} \mathbf{c}_i\phi_i(\mathbf{r}_j - \mathbf{R}_i) \tag{3}$$



https://doi.org/10.1021/acs.jctc.1c00568
J. Chem. Theory Comput. 2021, 17, 7814−7821

where $\boldsymbol{\psi}(\mathbf{r}_j) \in \mathbb{R}^N$ has the $n$th molecular orbital, $\boldsymbol{\psi}_n(\mathbf{r}_j)$, as its $n$th element, and $\mathbf{c}_i \in \mathbb{R}^N$ has the $n$th coefficient, $c_{ni}$, as its $n$th element. The orbital exponents, $\{\zeta_1, \zeta_2, ..., \zeta_N\}$, and coefficient vectors, $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N\}$, are randomly initialized and then learned/optimized for predicting a property by backpropagation and stochastic gradient descent (SGD). The LCAO calculation for all grid points in the molecular field can be implemented as illustrated in the figure in the Supporting Information.

**NN for the Energy Functional.** We form a large matrix from the vector-format molecular orbitals on all grid points: $\boldsymbol{\Psi} = \{\boldsymbol{\psi}(\mathbf{r}_j)\}_{j=1}^{G}$. Using this $\boldsymbol{\Psi}$ as the input, we consider an NN-based energy functional, $\mathcal{F}_{\mathrm{NN}}$, that provides the final layer of the NN and outputs the property of interest

$$\mathbf{q}_{\mathcal{M}} = \mathcal{F}_{\mathrm{NN}}[\boldsymbol{\Psi}] \tag{4}$$

$$E'_{\mathcal{M}} = \mathbf{w}_E^{\mathrm{T}}\mathbf{q}_{\mathcal{M}} + b_E \tag{5}$$

where $\mathbf{q}_{\mathcal{M}} \in \mathbb{R}^N$ is the final layer (i.e., the $N$-dimensional vector) of the NN, $E'_{\mathcal{M}}$ is the predicted property of $\mathcal{M}$, $\mathbf{w}_E \in \mathbb{R}^N$ is the trainable vector, and $b_E \in \mathbb{R}$ is the trainable scalar. We minimize the loss between $E'_{\mathcal{M}}$ and $E_{\mathcal{M}}$. Details about the NN architecture of $\mathcal{F}_{\mathrm{NN}}$ are described in the Supporting Information.

**NN for the Hohenberg–Kohn Map.** Unfortunately, simply minimizing $\mathcal{L}_E$ does not lead to a physically accurate ML model. Because the NN has strong nonlinearity, $\mathcal{F}_{\mathrm{NN}}$ will output the actual $E$ even if its input $\boldsymbol{\Psi}$ is not correct; in other words, the model overfits to the training database and does not learn a physically valid function. In particular, the Kohn–Sham orbitals provide the electron density by $\rho(\mathbf{r}) = \sum_{n=1}^{N} |\psi_n(\mathbf{r})|^2$; however, the model does not guarantee this.

To address this problem, when learning the model, we impose a physical constraint based on the first Hohenberg–Kohn theorem, which states that the external potential $V(\mathbf{r})$ is a unique function of $\rho(\mathbf{r})$; note that this function is nonlinear, so long as it is a one-to-one correspondence. Here, we use an NN for this nonlinear function; it is called the Hohenberg–Kohn map.[2,12] It acts as a physical constraint on $\rho(\mathbf{r}) = \sum_{n=1}^{N} |\psi_n(\mathbf{r})|^2$ and therefore ensures that the DFT-based model as a whole is based on physics (see again Figure 1a).

Specifically, we consider, as a simplification, the following Gaussian external potential[2,26] at $\mathbf{r}_j$

$$V_{\mathcal{M}}(\mathbf{r}_j) = -\sum_{m=1}^{M} Z_m e^{-\left|\left|\mathbf{r}_j - \mathbf{R}_m\right|\right|^2} \tag{6}$$

where $Z_m$ is the nuclear charge of $a_m$. We assume this $V_{\mathcal{M}}(\mathbf{r}_j)$ to be the ground-truth external potential of $\mathcal{M}$, to be used as the target for learning. The electron density is given by $\rho(\mathbf{r}_j) = \sum_{n=1}^{N} |\psi_n(\mathbf{r}_j)|^2$. Then, the NN-based Hohenberg–Kohn map, $\mathcal{HK}_{\mathrm{NN}}$, relates $\rho\mathbf{r}_j$ to $V'_{\mathcal{M}}(\mathbf{r}_j)$, the predicted external potential of $\mathcal{M}$

$$V'_{\mathcal{M}}(\mathbf{r}_j) = \mathcal{HK}_{\mathrm{NN}}(\rho(\mathbf{r}_j)) \tag{7}$$

Finally, we minimize the loss between $V'_{\mathcal{M}}$ and $V_{\mathcal{M}}$. Details about the NN architecture of the $\mathcal{HK}_{\mathrm{NN}}$ are described in the Supporting Information.

**Learning.** In learning the model, it is important to note that we must consider the physical condition in terms of the total number of electrons $N_{\mathrm{elec}}$

$$N_{\mathrm{elec}} = \int \rho(\mathbf{r})\mathrm{d}\mathbf{r} = \int \sum_{n=1}^{N} |\psi_n(\mathbf{r})|^2 \mathrm{d}\mathbf{r} \approx \sum_{j=1}^{G}\sum_{n=1}^{N} |\psi_n(\mathbf{r}_j)|^2 \tag{8}$$

In other words, this $N_{\mathrm{elec}}$ must be kept unchanged when learning/updating the molecular orbitals with the iterative SGD algorithm. We implement this by simply normalizing and transforming $\psi_n$ in SGD as follows

$$\boldsymbol{\psi}_n \leftarrow \sqrt{\frac{N_{\mathrm{elec}}}{N}}\frac{\boldsymbol{\psi}_n}{|\boldsymbol{\psi}_n|} \tag{9}$$

**Generating the QDD.** The QDF model, which inputs a molecule $\mathcal{M}$ and outputs its property $E_{\mathcal{M}}$, can be divided into two parts as follows

$$\mathbf{q}_{\mathcal{M}} = f_{\mathrm{QDF}}(\mathcal{M}) \tag{10}$$

$$E_{\mathcal{M}} = g(\mathbf{q}_{\mathcal{M}}) \tag{11}$$

where the function $f_{\mathrm{QDF}}$ takes $\mathcal{M}$ as input and produces the $N$-dimensional final layer $\mathbf{q}_{\mathcal{M}}$ in eq 4 as output, and $g$ is the linear regressor in eq 5. We refer to $\mathbf{q}_{\mathcal{M}}$ as the QDD of $\mathcal{M}$.

We first emphasize that the QDD is learned to encode $\psi$, $\rho$, and property $E$ (e.g., the atomization energy, HOMO, or LUMO), of a molecule into an $N$-dimensional vector. Additionally, once $f_{\mathrm{QDF}}$ is trained on a given $\mathcal{D}_{\mathrm{train}}$, we can easily obtain $\mathbf{q}_{\mathcal{M}}$ for any size and structure of the molecule $\mathcal{M}$ from eq 10 (there is, however, a limitation derived from $\mathcal{D}_{\mathrm{train}}$; we will discuss this later). Furthermore, the $\mathbf{q}_{\mathcal{M}}$ obtained by pre-trained $f_{\mathrm{QDF}}$ can be reused as a molecular descriptor; that is, we can replace the regressor $g$ in eq 11 with another regressor $h$ on $\mathbf{q}_{\mathcal{M}}$ and then learn the new $h$ to make predictions about other molecules and properties. In the following, we consider how transfer learning with the QDD makes it possible to treat much larger and differently structured molecules (e.g., polymers) and their various properties (e.g., band gap and dielectric constants).

**Transfer Learning with the Pre-Trained QDD.** We consider a dataset for transfer learning $\mathcal{D}_{\mathrm{transfer}} = \{(C_1, Z_1), (C_2, Z_2), ..., (C_T, Z_T)\}$, where $C_t$ is the $t$th molecule, $Z_t$ is a property of $C_t$, and $T$ is the number of data samples. $C$ may be, for example, a polymer, but its data structure is the same as that of $\mathcal{M} \in \mathcal{D}_{\mathrm{train}}$ (i.e., atom types and their 3D coordinates); however, $Z$ is not necessarily the same property as that of $E \in \mathcal{D}_{\mathrm{train}}$.

On the other hand, the atom types in $C \in \mathcal{D}_{\mathrm{transfer}}$ must be the same as those in $\mathcal{M} \in \mathcal{D}_{\mathrm{train}}$; for example, when $\mathcal{D}_{\mathrm{train}}$ is the QM9 database,[13] which includes only molecules composed of H, C, N, O, and F, $C$ must be composed of the same elements. This is because the learning parameters in the GTO and LCAO, that is, the orbital exponent $\zeta$ and coefficient $c$ in eqs 2 and 3, respectively, are characterized by the atomic-orbital type, randomly initialized and then trained on $\mathcal{D}_{\mathrm{train}}$. If $C$ has unknown atom types (e.g., Cl) that did not appear in


https://doi.org/10.1021/acs.jctc.1c00568
J. Chem. Theory Comput. 2021, 17, 7814−7821

$\mathcal{D}_{\text{train}}$, $\zeta$ and $c$ for the Cl orbitals will be randomly initialized, and therefore, the prediction cannot be reliable.

Considering the abovementioned condition on $\mathcal{D}_{\text{transfer}}$, we learn a function that inputs a molecule $C$ and outputs its property $Z$. Here, instead of learning the function from scratch, we insert the $f_{\text{QDF}}$ pre-trained on $\mathcal{D}_{\text{train}}$ into eq 10

$$\mathbf{q}_C = f_{\text{QDF}}(C) \tag{12}$$

$$Z'_C = h(\mathbf{q}_C) \tag{13}$$

where $\mathbf{q}_C$ is the $N$-dimensional QDD of $C$, $Z'_C$ is the predicted property, and $h$ is a regressor specific to $Z$. This study used the ridge regressor for $h$

$$Z'_C = w_0 + \sum_{i=1}^{N} w_i q_i \tag{14}$$

where $\{w_i\}_{i=0}^{N}$ are the weights ($w_0$ is the bias term or intercept) to be learned and $q_i$ is the $i$th element of $\mathbf{q}_C$. Finally, we minimize the following loss function

$$\mathcal{L}_Z = \sum_{t=1}^{T} \left\| Z_{C_t} - Z'_{C_t} \right\|^2 + \lambda \sum_{i=1}^{N} w_i^2 \tag{15}$$

where $\lambda$ is the L2 regularization hyperparameter. Note that we only learn the parameters of $h$ and fix those of $f_{\text{QDF}}$, as shown in Figure 1b; this is the simplest type of linear transfer learning that can be built on the frozen pre-trained model.

## RESULTS AND DISCUSSION

**Dataset and the Comparative Descriptor/Model.** For training the QDF model on $\mathcal{D}_{\text{train}}$, we used the atomization energy, HOMO, and LUMO in the QM9 database[13] as the target properties to be learned simultaneously by the regressor; thus, the output dimension of $g$ was three. The properties were calculated from DFT at the B3LYP/6-31G(2df,p) level of theory in Gaussian 09. QM9 contains approximately 130,000 small organic molecules; the number of atoms in each molecule ranges from 3 to 29.

For transfer learning with the QDD on $\mathcal{D}_{\text{transfer}}$, we used the atomization energy, band gap, and two types of dielectric constants (total and ionic) in the polymer dataset created by Huan et al.[27] Each polymer datum is given as its repeat unit, and the properties were calculated via DFT using the projector-augmented wave in the Vienna Ab initio Simulation Package (further details are provided in Huan et al.[27]). The polymer dataset contains 348 polymers; the number of atoms in each repeat unit ranges from 4 to 156. Note that the molecules in both $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{transfer}}$ are composed of H, C, N, O, and F atoms. QM9 includes only molecules containing less than 9 heavy elements (C, N, O, and F), whereas the polymer dataset has molecules with up to 72 heavy elements.

To evaluate transfer learning performance, we compared the QDD with the many-body tensor representation (MBTR),[28] smooth overlap of atomic positions (SOAP),[29] and graph neural network (GNN) (Table 1). The MBTR and SOAP are standard, widely used, and successful descriptors in MI,[30−32] and for creating these descriptors, we used DScribe software.[33] Specifically, we trained the MBTR and SOAP with a simple NN on the QM9 database and then used its pre-trained final layer as the descriptor for transfer learning on the polymer dataset. In the same way, we trained the GNN model [also called a graph convolutional network (GCN) and message-passing neural network (MPNN) in the ML community] and then used its pre-trained final layer as the descriptor. We provide further details about the MBTR, SOAP, and GNN in the Supporting Information.

**Table 1. Sizes of Four Models: MBTR, SOAP, GNN, and QDF[a]**

| descriptor | dimension | NN parameters |
|---|---|---|
| MBTR | 475 | 740,503 |
| SOAP | 480 | 742,001 |
| GNN | 350 | 739,903 |
| QDF | 350 | 754,304 |

[a]All NNs used in these models are the same shallow architectures and their depth is three. For a fair comparison, we kept the number of these NN parameters (about 750,000). In the MBTR, we set $n$ = 5 and in the SOAP, we set the GTO, $r_{\text{cut}}$ = 6.0, $n_{\text{max}}$ = 3, $l_{\text{max}}$ = 3, and $\sigma$ = 1.0. (see our source code). Other results with different SOAP settings are provided in the Supporting Information.

**Learning of the Atomization Energy, HOMO, and LUMO in QM9.** Figure 2 displays the learning curves of three properties in QM9. The NN model using the SOAP descriptor showed the highest performance (in particular in HOMO and LUMO predictions); however, the MBTR and our QDF also achieved high performance and these results were competitive. Although these accuracies will vary by hyperparameter tuning, our aim is not to attain state-of-the art (SoTA) interpolation accuracy on QM9 but to evaluate the transferability and physical validity of the learned models. Note that some MAEs indicated chemical accuracies comparable to those reported in the literature,[34] and overall, we believe that these accuracies were reasonable on QM9.

**Transfer to Atomization Energy and Band Gap of Polymers.** We first evaluated the transfer learning/prediction performance for the atomization energy and band gap of polymers. As shown in Figure 3, the QDD achieved high performance in both properties, whereas the MBTR and GNN could not, and the SOAP was competitive with the QDF in the atomization energy but could not outperform in the band gap. We emphasize that, as shown in Figure 2, some MAEs of the QDF were almost the same as those of the MBTR and SOAP in QM9; nonetheless, the QDF did not overfit to the QM9 molecules, and the QDD further achieved better transfer performance to the polymers.

These results tell us that even if an ML model achieves high interpolation accuracy in a large database, the physical validity of the learned model is not guaranteed. In particular, if we divide the large database into a training set (80%) and a test set (20%), eventually the test set will contain data samples that are very similar to (or almost the same as) the training set. The MBTR apparently achieved high accuracy (Figure 2), but this was merely the result of overfitting to the QM9 molecular properties; the failure to learn the physically valid function leads to failure in transferring to the polymer properties.

Indeed, in our implemented GNN model, we normalized all atom vectors in the NN-based message-passing algorithm. The GNN without such normalization also achieved high accuracy on QM9, but it exhibited very poor transfer performance on the polymers (see the Supporting Information for details). This shows that it is impossible to ensure the physical validity of an ML model by evaluating it on a single benchmark database, even a large one.



https://doi.org/10.1021/acs.jctc.1c00568
J. Chem. Theory Comput. 2021, 17, 7814−7821

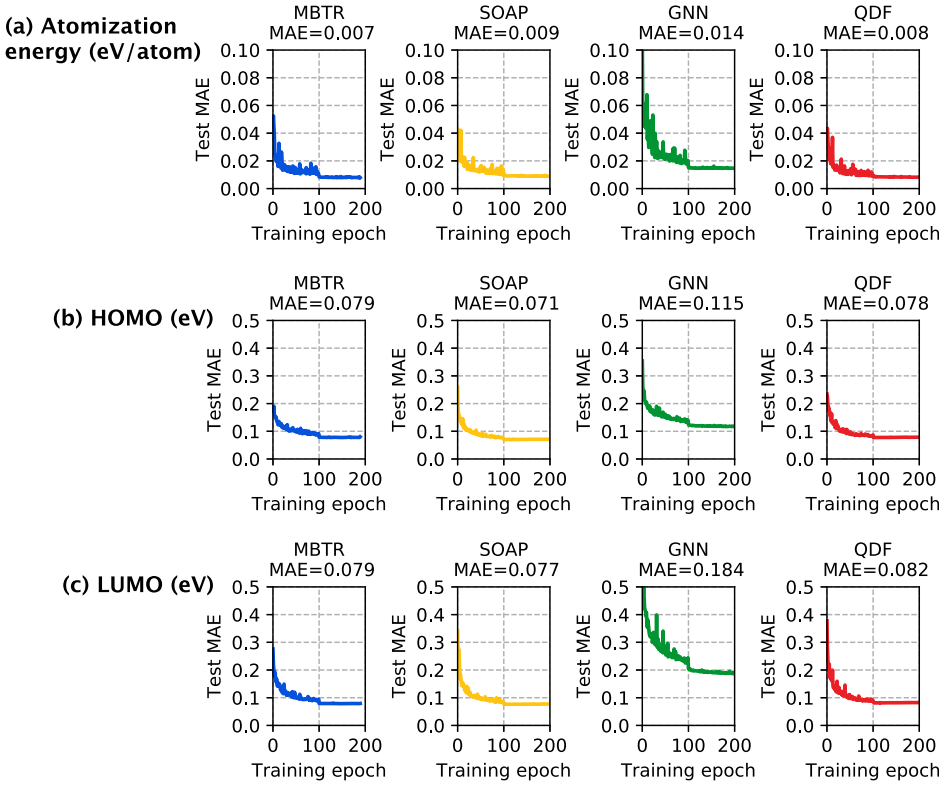



**Figure 2.** Learning curves (mean absolute error (MAE) vs number of training iterations) of four models (MBTR, SOAP, GNN, and QDF) for three kinds of molecular properties on the QM9 database.

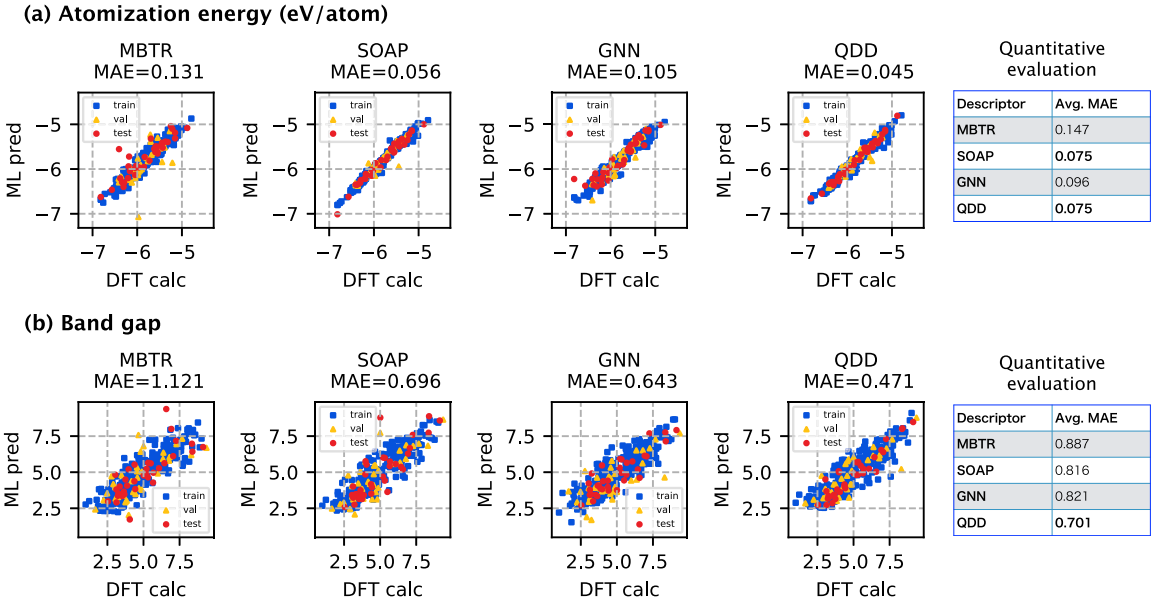



**Figure 3.** Transfer learning and prediction performances for the atomization energy and band gap of polymers. In the plot graph, the *x*-axis is the DFT result and the *x*-axis is the predicted result by the pre-trained descriptor (MBTR, SOAP, GNN, and QDD). This is an example of the split pattern of the training, validation, and test sets. As a quantitative evaluation, we prepared 1000 random split patterns of the training, validation, and test sets (i.e., bootstrap) and averaged 1000 MAEs. We provide this averaged MAE score in the right table.

**Transfer to Dielectric Constants of Polymers.** Finally, we evaluated the transfer learning/prediction performance for the two types of polymer dielectric constants (total and ionic). Unlike the atomization energy, these total and ionic dielectric constants were not target properties in pre-training. Figure 4 shows that the transfer accuracy of the QDD (avg. MAE is 0.310 eV) outperformed that of the SOAP (0.374 eV) in the total, and the transfer accuracy of the SOAP (0.258 eV) outperformed that of the QDD (0.272 eV) in the ionic; the performances in predicting the dielectric constants are comparable. The QDF could perhaps be improved to predict these polymer properties in the future. In the following, we discuss how to improve the MBTR, SOAP, GNN, and QDF approaches.



https://doi.org/10.1021/acs.jctc.1c00568
J. Chem. Theory Comput. 2021, 17, 7814−7821

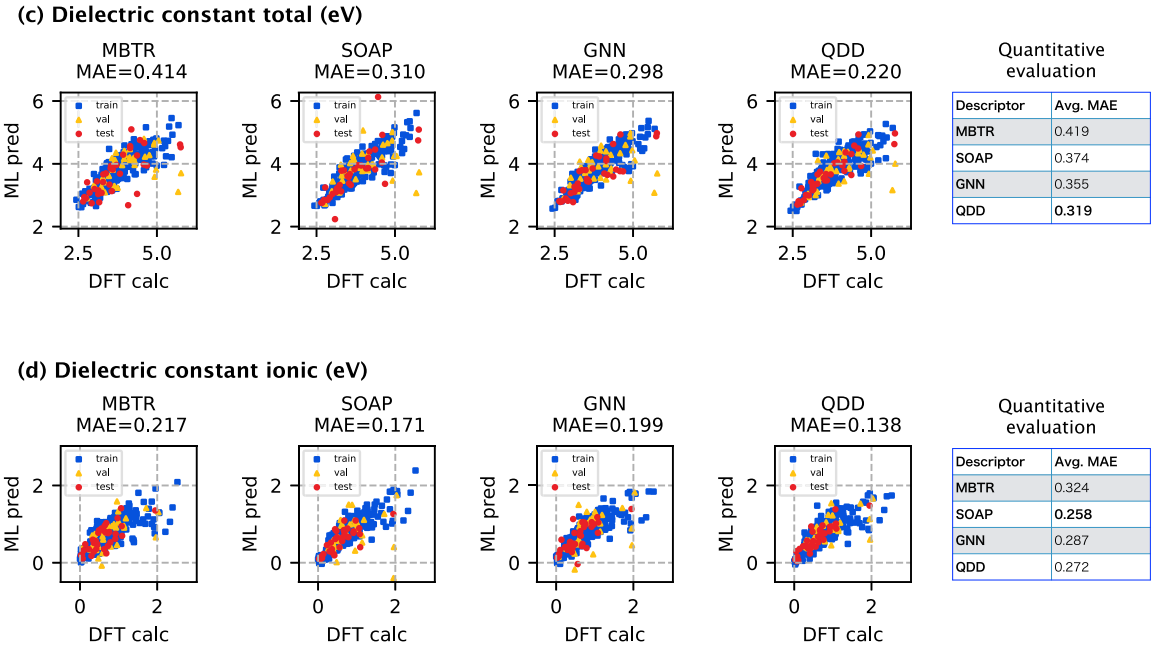



**Figure 4.** Transfer learning and prediction performances for the dielectric constants (total and ionic) of polymers.

**Discussion for Improving the Current QDF.** Although the QDD achieved high performance in transfer learning of dielectric constants, it may not have the same accuracy in transferring the atomization energy and band gap; the current QDF model should be improved. Here, we emphasize that the QDF can be improved in ways that incorporate physical understanding, whereas improvements to the MBTR, SOAP, and GNN are largely formal. For example, we may be able to improve the MBTR, SOAP, and GNN models by (1) enlarging the dimensions and layers, (2) finding other effective molecular descriptors and NN architectures (e.g., deeper networks and transformers), and (3) pre-training on other large databases. Such trial-and-error, however, just results in tuning and selecting the hyperparameters, descriptors/architectures, and databases. By contrast, we can improve the QDF model by (1) using Slater-type orbitals that satisfy the cusp condition in place of the current simplified GTOs in the LCAO, (2) designing a model that involves the spherical harmonics (i.e., 2px, 2py, 2pz, 3d, and 4f orbitals) for heavy atoms, and (3) considering other forms of the external potential in the Hohenberg–Kohn map constraint, such as the Coulomb and pseudo-potentials. We believe that these are physically meaningful improvements rather than hyperparameter tuning in ML and thus that the QDF model and the QDD descriptor have the potential for further improvement.

## CONCLUSIONS AND FUTURE DIRECTIONS

In this study, we proposed a physically informed transfer learning approach using the QDD obtained from the pre-trained QDF model. We have evaluated this approach with polymer property prediction. Our QDD has achieved higher performance than existing approaches.

The principal future goal should not be the achievement of SoTA performance on a single benchmark database. It is more important (and will yield more practical applications) to consider other materials [including additional polymers, biologically active materials (e.g., drugs), and catalysts] .[24,25,35] The properties of these materials should be predicted by transferring the physically informed ML models, such as our QDF, electron density–based ML,[3,4] molecular orbital–based ML (MOB-ML),[7−10] and differential Kohn–Sham ML models and[6,21] training them on various quantum-chemical property databases, such as PubChemQC,[36] tmQM,[37] Alchemy,[38] and others .[31,39,40] The key questions are as follows: (1) On what data (e.g., excited states) do we train a physically informed ML model? (2) What data (e.g., catalysis) do we transfer and predict with the trained ML model? (3) What quantum-chemical insights do we obtain from (1) and (2)? We believe that such data-driven ML approaches will not only be useful for applications in MI but also provide new physical insights into materials science.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00568.

Examples of molecular gird data, implementation of LCAOs, NN for the energy functional, NN for the Hohenberg–Kohn map, NN on the MBTR, NN on the SOAP, GNN, and hyperparameters of the QDF (PDF)

## AUTHOR INFORMATION

### Corresponding Author

**Masashi Tsubaki** − *National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan;* orcid.org/0000-0002-5466-2195; Email: tsubaki.masashi@aist.go.jp

### Author

**Teruyasu Mizoguchi** − *Institute of Industrial Science, The University of Tokyo, Tokyo 113-0033, Japan;* orcid.org/0000-0003-3712-7307

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00568

## Notes

The authors declare no competing financial interest.

## REFERENCES


(1) Giustino, F. *Materials Modelling Using Density Functional Theory: Properties and Predictions*; Oxford University Press, 2014.

(2) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn−Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872−10.

(3) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **2018**, *5*, 57−64.

(4) Zhang, Y.; Hu, C.; Jiang, B. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962−4967.

(5) Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(6) Li, L.; Hoyer, S.; Pederson, R.; Sun, R.; Cubuk, E. D.; Riley, P.; Burke, K.; et al. Kohn−Sham equations as regularizer: Building prior knowledge into machine-learned physics. *Phys. Rev. Lett.* **2021**, *126*, 036401.

(7) Welborn, M.; Cheng, L.; Miller, T. F., III. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772−4779.

(8) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F., III. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **2019**, *150*, 131103.

(9) Cheng, L.; Kovachki, N. B.; Welborn, M.; Miller, T. F., III. Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning. *J. Chem. Theory Comput.* **2019**, *15*, 6668−6677.

(10) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J. R.; Miller, T. F., III. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021**, *154*, 064108.

(11) Tsubaki, M.; Mizoguchi, T. Quantum deep field: data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning. *Phys. Rev. Lett.* **2020**, *125*, 206401.

(12) Moreno, J. R.; Carleo, G.; Georges, A. Deep learning the Hohenberg−Kohn maps of density functional theory. *Phys. Rev. Lett.* **2020**, *125*, 076402.

(13) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(14) Raissi, M.; Karniadakis, G. E. Hidden physics models: Machine learning of nonlinear partial differential equations. *J. Comput. Phys.* **2018**, *357*, 125−141.

(15) Raissi, M. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *J. Mach. Learn. Res.* **2018**, *19*, 932−955.

(16) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686−707.

(17) Nagai, R.; Akashi, R.; Sasaki, S.; Tsuneyuki, S. Neural-network Kohn−Sham exchange-correlation potential and its out-of-training transferability. *J. Chem. Phys.* **2018**, *148*, 241737.

(18) Nagai, R.; Akashi, R.; Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Comput. Mater.* **2020**, *6*, 43.

(19) Pun, G. P.; Batra, R.; Ramprasad, R.; Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **2019**, *10*, 2339.

(20) Pfau, D.; Spencer, J. S.; Matthews, A. G.; Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2020**, *2*, 033429.

(21) Kasim, M. F.; Vinko, S. M. Learning the exchange-correlation functional from nature with fully differentiable density functional theory. *arXiv (Physics.Chemical Physics)*, March 18, 2021, 2102.04229, ver. 3. https://arxiv.org/abs/2102.04229 (accessed 2021-06-07).

(22) Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming data scarcity with transfer learning. *arXiv (Computer Science.Machine Learning)*, November 2, 2017, 1711.05099, ver. 1. https://arxiv.org/abs/1711.05099 (accessed 2021-06-07).

(23) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.

(24) Sarullo, K.; Matlock, M. K.; Swamidass, S. J. Site-level bioactivity of small-molecules from deep-learned representations of quantum chemistry. *J. Phys. Chem. A* **2020**, *124*, 9194−9202.

(25) Lee, C.-K.; Lu, C.; Yu, Y.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Liu, Q.; Shi, L. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *J. Chem. Phys.* **2021**, *154*, 024906.

(26) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(27) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, 160012.

(28) Huo, H.; Rupp, M. Unified representation for machine learning of molecules and crystals. *arXiv (Physics.Chemical Physics)*, January 2, 2018, 1704.06439, ver. 3. https://arxiv.org/abs/1704.06439 (accessed 2021-06-07).

(29) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(30) Nyshadham, C.; Rupp, M.; Bekker, B.; Shapeev, A. V.; Mueller, T.; Rosenbrock, C. W.; Csányi, G.; Wingate, D. W.; Hart, G. L. Machine-learned multi-system surrogate models for materials prediction. *npj Comput. Mater.* **2019**, *5*, 51.

(31) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J. Chem. Phys.* **2019**, *150*, 204121.

(32) Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **2020**, *11*, 4428.

(33) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(34) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(35) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W., et al. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *arXiv (Condensed Matter.Materials Science)*, March 16, 2021, 2010.09990, ver. 4. https://arxiv.org/abs/2010.09990 (accessed 2021-06-07).

(36) Nakata, M.; Shimazaki, T. PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300−1308.

(37) Balcells, D.; Skjelstad, B. B. tmQM Dataset-Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135−6146.




https://doi.org/10.1021/acs.jctc.1c00568
*J. Chem. Theory Comput.* 2021, 17, 7814−7821

(38) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J., et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv (Computer Science.Machine Learning)*, June 22, 2019, 1906.09427, ver. 1. https://arxiv.org/abs/1906.09427 (accessed 2021-06-07).
(39) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **2016**, *3*, 160009.
(40) Lu, J.; Xia, S.; Lu, J.; Zhang, Y. Dataset Construction to Explore Chemical Space with 3D Geometry and Deep Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1095.



https://doi.org/10.1021/acs.jctc.1c00568
*J. Chem. Theory Comput.* 2021, 17, 7814−7821