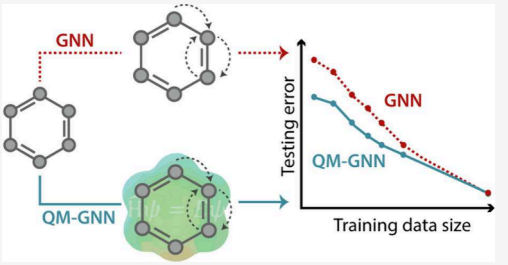
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

When Do Quantum Mechanical Descriptors Help Graph Neural Networks to Predict Chemical Properties?

Shih-Cheng Li,[¶] Haoyang Wu,[¶] Angiras Menon, Kevin A. Spiekermann, Yi-Pei Li,* and William H. Green*

Cite This: *J. Am. Chem. Soc.* 2024, 146, 23103−23120

ABSTRACT: Deep graph neural networks are extensively utilized to predict chemical reactivity and molecular properties. However, because of the complexity of chemical space, such models often have difficulty extrapolating beyond the chemistry contained in the training set. Augmenting the model with quantum mechanical (QM) descriptors is anticipated to improve its generalizability. However, obtaining QM descriptors often requires CPU-intensive computational chemistry calculations. To identify when QM descriptors help graph neural networks predict chemical properties, we conduct a systematic investigation of the impact of atom, bond, and molecular QM descriptors on the performance of directed message passing neural networks (D-MPNNs) for predicting 16 molecular properties. The analysis surveys computational and experimental targets, as well as classification and regression tasks, and varied data set sizes from several hundred to hundreds of thousands of data points. Our results indicate that QM descriptors are mostly beneficial for D-MPNN performance on small data sets, provided that the descriptors correlate well with the targets and can be readily computed with high accuracy. Otherwise, using QM descriptors can add cost without benefit or even introduce unwanted noise that can degrade model performance. Strategic integration of QM descriptors with D-MPNN unlocks potential for physics-informed, data-efficient modeling with some interpretability that can streamline *de novo* drug and material designs. To facilitate the use of QM descriptors in machine learning workflows for chemistry, we provide a set of guidelines regarding when and how to best leverage QM descriptors, a high-throughput workflow to compute them, and an enhancement to Chemprop, a widely adopted open-source D-MPNN implementation for chemical property prediction.



INTRODUCTION

Recent advances in the prediction of chemical reactivity and molecular properties using machine learning (ML) have sparked substantial interest in both academia and industry, particularly in pharmaceutical and material research and development, where rapid and accurate predictions of molecular properties can substantially accelerate the discovery and development of promising new drugs and materials. Furthermore, state-of-the-art deep learning architectures for chemical property prediction, especially directed message passing neural networks (D-MPNNs) as exemplified by the open-source Chemprop[1] software package, have shown success in predicting a variety of chemical properties, including solvation thermodynamics,[2−5] regioselectivity,[6−9] and reaction barriers.[10−14] However, ML models tailored for molecular property and chemical reactivity prediction often rely heavily on extensive data sets to achieve some degree of generalization. This dependence poses a critical challenge in closed-loop generative molecular discovery campaigns,[15,16] where achieving a robust generalization of property prediction models is crucial, but acquiring large-scale experimental data is often challenging and expensive. Therefore, developing strategies to improve the ML model performance, especially model robustness for novel molecules proposed by generative models, under significant data constraints is crucial in these domains.

The performance of ML models for chemical property prediction depends on having expressive representations of chemical species and reactions. Traditional approaches often utilize expert-guided fixed molecular fingerprints,[17−19] or Morgan fingerprints,[20] for molecular featurization (a.k.a. encoding). However, recent trends have shifted toward automatically learned molecular representations, such as those obtained from graph neural networks (GNNs), mainly due to the enhanced scalability, flexibility, expressiveness, and generalizability offered by learned representations.[21−24] Studies have shown that ML models that leverage learned molecular representations tend to outperform those that used fixed molecular fingerprints.[25−27] Yet, this transition to representation learning requires larger data sets, as the relationship

Received: April 4, 2024
Revised: July 18, 2024
Accepted: July 19, 2024
Published: August 6, 2024

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

between the features (the inputs to the model) and targets (the desired outputs) in deep neural networks becomes less transparent and more complex. Moreover, molecular representations can also be classified by the type of features used to generate them. For example, structure-based representations often require only very basic chemical and structural information, such as atom types, bond orders, and connectivities, which makes them widely used because of their simplicity and extensibility. However, because this featurization strategy tends to predominantly depend on basic molecular information, a more extensive training set is often necessary to learn representations capable of handling complex chemical property prediction tasks effectively.

Alternatively, to improve model performance with limited data, a common practice is to utilize specialized features—oftentimes computed descriptors—that are more relevant to the targets of interest. However, the best choice of descriptors is often task-specific, and obtaining these descriptors can require substantial effort. For instance, Sigman et al. utilized multivariate linear regression to establish a correlation between electronic and steric descriptors and reaction properties, including electrochemical potential, enantio-, site-, and chemo-selectivity.[9] Doyle and co-workers employed random forest to predict reaction yields of C−N cross-coupling reactions by selecting mechanistically relevant descriptors.[28] Boobier et al. employed a limited number of descriptors to predict solubility in both organic solvents and water.[2] This approach of selecting targeted descriptors has gained more traction and application in recent years,[4,6,8,13,29−32] with a growing inclination toward using quantum mechanical (QM) descriptors.[7,33−36] This shift is driven by the belief that QM descriptors can provide deeper physical insights and enhance model interpretability. Additionally, they are considered more extensible and generalizable across a wide range of chemical species and property prediction tasks.[37]

The significance of representation in chemical property prediction models has led to a growing interest in recent studies to develop methods for obtaining better representations. A notable example comes from Guan et al.,[7] where the authors combined machine-learned representation from molecular structures with QM descriptors to create a fused reactivity representation. Using this approach, they implemented a QM-augmented graph neural network (QM-GNN) using the Weisfeiler-Lehman network (WLN) as the GNN architecture to predict reaction regioselectivity for 3,003 electrophilic aromatic substitution reactions. They computed and utilized Hirshfeld partial charges, nucleophilic and electrophilic Fukui indices, and NMR shielding constants as atom descriptors, along with bond orders and bond lengths as bond descriptors. Since calculating QM descriptors for each new molecule would not be amenable to high-throughput inference, they also developed a multitask constrained D-MPNN model to quickly predict QM descriptors on the fly. Their findings demonstrated that incorporating QM descriptors can improve the ML model performance in terms of accuracy and reduce the need for extensive training data, although the benefits of QM descriptors diminish as more training data are used. In particular, replacing QM-calculated descriptors with ML-predicted ones (termed ml-QM-GNN in their work) maintained high performance and outperformed the baseline GNN model, especially in models trained on smaller data sets, showcasing the advantages of using hybrid representations to potentially achieve improved model performance with limited data.

However, subsequent studies suggest that the effectiveness of including QM descriptors varies with specific tasks. Stuyver and Coley further explored the performance, generalizability, and explainability of ml-QM-GNN in predicting regioselectivity for substitution reactions and predicting activation energy for competing E2/S_N2 reactions in the gas-phase.[33] Their results indicated that this approach could outperform purely structure-based GNNs in data-limited regimes, making it appealing for tasks with limited experimental data availability. Recently, they also successfully applied this approach to predict activation and reaction energies for [3 + 2] cycloadditions, identifying promising candidates for bioorthogonal click reactions through active learning.[38] These and other works indicate that QM descriptors can be beneficial for property prediction to varying degrees. However, contrasting findings from Spiekermann et al. showed that QM descriptors did not enhance activation energy prediction in a D-MPNN model trained on approximately 24,000 diverse gas-phase reactions.[14] The authors suggested that the underperformance of QM-GNN may be attributed to the large training set, although the learning curve had not reached asymptotic behavior in the baseline GNN model without QM descriptors. Similar observations were made by Guan et al., who recently utilized Fukui indices for predicting the regioselectivity of nucleophilic aromatic substitution reactions from a newly compiled in-house data set.[39] Likewise, Biswas et al. incorporated Hirshfeld partial charges and Mulliken total dipole moment into their D-MPNN and graph attention network models for predicting the acentric factors and critical properties of fluids.[34] Both studies concluded that the inclusion of QM descriptors had minimal impact on the overall performance of their respective models. Collectively, these studies indicate that while QM descriptors may be beneficial in certain scenarios, their effectiveness is highly context-dependent and seems to be particularly influenced by the size of the data set and the nature of the chemical prediction task.

Therefore, despite significant efforts to utilize QM descriptors in GNNs to enhance chemical property predictions, it remains unclear which descriptors are best suited for specific tasks. This uncertainty often leads to a trial-and-error approach in descriptor selection and to varied methods for incorporating these descriptors. Moreover, the effectiveness of using QM descriptors, particularly in relation to training data size and model chemistry, remains underexplored. To address these gaps, this work aims to systematically analyze the conditions under which the use of QM descriptors is beneficial, focusing on identifying the best strategies for incorporating QM descriptors for specific chemical property prediction tasks. More specifically, we conduct a comprehensive exploration of the impact of QM descriptors on the performance of D-MPNN models to predict various important chemical properties across 16 distinct molecular property data sets, ranging from electronic energy and solubility to toxicity and protein binding affinity. The data set sizes span from several hundred to hundreds of thousands of data points, and both classification and regression tasks are investigated. We adapt the work of Guan et al.[7] to compute QM descriptors with high throughput and also implement a flexible way of providing or predicting atom and bond QM descriptors in Chemprop,[1] the leading open-source D-MPNN software for chemical property prediction. We compute and curate a computational data set

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

that contains 37 distinct QM descriptors for approximately 65,000 molecules using 6 different model chemistries, and subsequently train models to predict these descriptors efficiently. We conduct an assessment of the QM-GNN and ml-QM-GNN models across various prediction tasks. Based on our results, we make recommendations about when and how QM descriptors should be used. It should be noted that, while in the rest of the manuscript we adopt the same acronyms for ML models (e.g., QM-GNN, ml-QM-GNN) as Guan et al.,[7] this work uses D-MPNN as the chosen GNN, not the WLN model they used.

BEST PRACTICES ON USING QM DESCRIPTORS IN GNN

Based on the results and insights gained in this study, we present a simple decision flowchart (Figure 1) to outline our recommended procedures when considering whether to augment GNN models with QM descriptors. The rationale and supporting evidence for the decision logic at each step are provided in detail in the Results and Discussion section. In general, decisions can be made within the following six steps.

(a) For data sets exceeding about 2,000 data points, we recommend GNN without QM descriptors. This is based on the observation that the advantages of additional QM descriptors tend to diminish rapidly as the data set size increases.

(b) With smaller data sets, the initial step should be to evaluate the feasibility of generating new data. If acquiring additional data is feasible, prioritize collecting more data and training a standard GNN. However, if new data generation is impractical, consider the inclusion of QM descriptors, but proceed with a careful selection and thorough assessment of these descriptors.

(c) It is crucial to determine if there is substantial evidence linking the chosen descriptors to the targets of interest. This evidence might come from quantitative analysis, physicochemical principles, and/or intuition. Conducting a pilot study to investigate the relationship between the selected descriptors and targets is highly recommended before starting a campaign to perform expensive quantum chemistry calculations on thousands of molecules or reactions.

(d) The accuracy of QM descriptors must be assessed. For instance, descriptors with a strong dependence on molecular conformers may be challenging to compute accurately, especially for large, conformationally diverse molecules. When in doubt, pilot studies can be useful in evaluating the feasibility of accurately obtaining the desired descriptors.

(e) Analyze the sensitivity of the model to the values of the descriptors through preliminary studies. If the model performance is highly sensitive to exact descriptor values, the QM-GNN approach is recommended to minimize errors that could arise from predicting descriptors using ML models.

(f) The inclusion of additional descriptors implies that new predictions are contingent on the availability of these descriptors for each inquiry molecule or reaction of interest. Therefore, consider the practicality and scalability of the models, alongside their accuracy. In high-throughput scenarios, the ml-QM-GNN approach is usually more cost-effective.

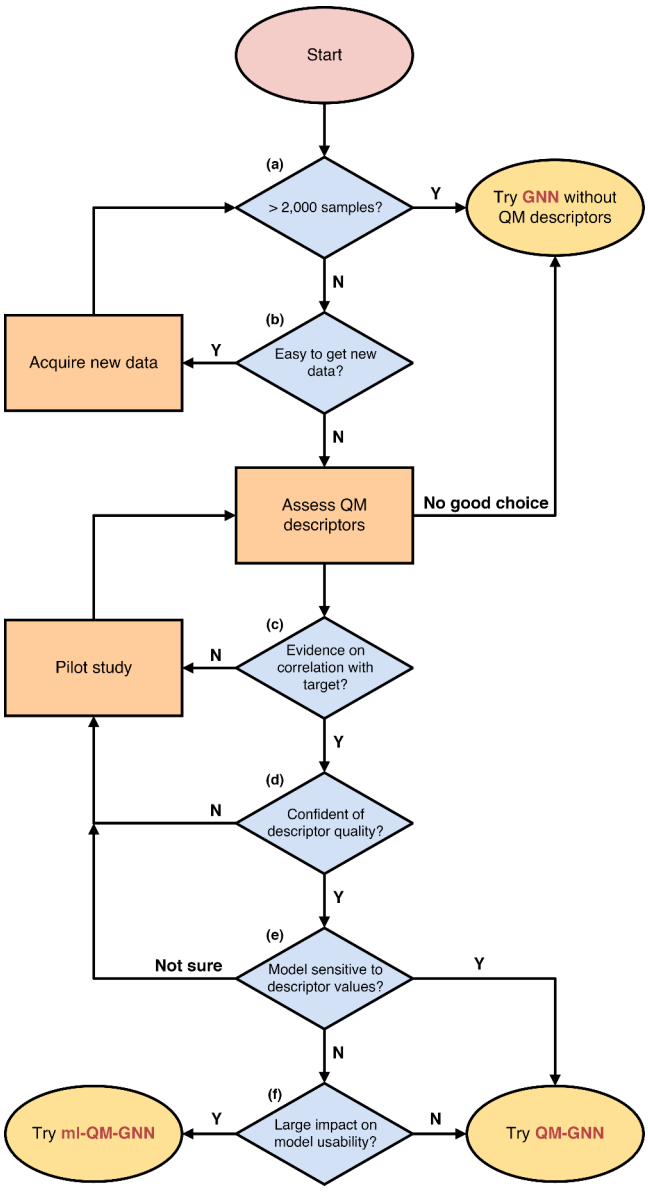


Figure 1. A decision flowchart illustrating recommended procedures for incorporating quantum mechanical (QM) descriptors in graph neural network (GNN) models to improve chemical property prediction on the basis of results and insights gained throughout this work. The selection of appropriate methods depends on the availability of resources and the desired balance between accuracy and efficiency. In the QM-GNN approach, QM descriptors are directly calculated from quantum chemical simulations, while in the ml-QM-GNN approach, machine learning (ML) models are employed to estimate QM descriptors to offer a more computationally efficient alternative.

Moreover, strategically selecting a refined set of the most relevant descriptors is crucial, for it can improve model performance, enhance interpretability by potentially revealing underlying physicochemical principles governing the descriptor-target relationships, and reduce computational costs by avoiding unnecessary descriptor calculations. Finally, if opting for QM descriptors, it is advisible to conduct benchmarking against alternative methods (e.g., comparing QM-GNN with a standard GNN as a baseline). This ensures that the benefits gained from QM descriptors justify their added cost.

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

METHODS

Integrating Descriptors into D-MPNN Molecular Encoding. D-MPNN[25] is one type of GNN.[40] This work uses Chemprop,[1] a widely used open-source D-MPNN package for the prediction of chemical properties. Chemprop utilizes a feature encoding module along with a D-MPNN to construct atomic embeddings. These embeddings are then aggregated into molecular representations, which are further processed by a feed-forward neural network (FFNN) for prediction of target properties. In the following, we briefly review how molecular encoding is achieved in Chemprop along with our modifications to incorporate additional descriptors as features. Notations are summarized in Table 1. In Chemprop, an input molecule M (typically input in SMILES format) is converted to a molecular graph $G = M(V, E, \mathbf{F}^{(n)})$ where atoms are treated as vertices (V, a.k.a. nodes) and bonds as edges (E, a.k.a. links), and $\mathbf{F}^{(n)}$ is a tensor that carries various node, edge, and graph attributes. The encoding of a molecular graph in Chemprop begins with the initialization of nodes and edges with features based on the structural topology of the molecule being represented and the chemical identities of the corresponding atoms and bonds. More specifically, the default atom features, $\boldsymbol{\alpha}_{\text{chem},i}$, consist of atomic number, formal charge, chirality, hybridization, aromaticity, total number of bonds and hydrogens to which an atom is connected, and scaled atomic mass. The default bond features, $\boldsymbol{\beta}_{\text{chem},ij}$, consist of bond order, the presence of conjugation or ring structures, and stereochemical information such as cis/trans isomerism. We added functionality in Chemprop to include additional QM descriptors in the initial feature vectors, and thus:

$$\boldsymbol{\alpha}_i, \boldsymbol{\beta}_{ij} = \begin{cases} \boldsymbol{\alpha}_{\text{chem},i}, \boldsymbol{\beta}_{\text{chem},ij} & \text{default Chemprop} \\ \text{concat}(\boldsymbol{\alpha}_{\text{qm},i}, \boldsymbol{\alpha}_{\text{chem},i}) & \\ \text{concat}(\boldsymbol{\beta}_{\text{qm},ij}, \boldsymbol{\beta}_{\text{chem},ij}) & \text{with QM descriptors} \end{cases} \tag{1}$$

where $\boldsymbol{\alpha}_i$ represents the initial atom features of the i^{th} atom and $\boldsymbol{\beta}_{ij}$ represents the initial undirected bond features between the i^{th} atom and the j^{th} atom before messaging passing. Notice that we used the subscript "chem" to distinguish the default atom and bond features in Chemprop from the QM descriptors (denoted by subscript "qm") added in this work.

These initial features are then passed on to the message passing phase, which utilizes directed bonds to share and update the hidden messages throughout the molecular graph. More specifically, the initial hidden state of an edge, $\boldsymbol{h}^0_{\beta,ij}$, is computed as

$$\boldsymbol{h}^0_{\beta,ij} = \Omega[\mathbf{W}_1 \times \text{concat}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_{ij})] \tag{2}$$

where $\text{concat}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_{ij})$ denotes the concatenation of atom and bond features to create the initial directed bond feature vector, $\mathbf{W}_1$ is a learned weight matrix, and Ω is a nonlinear activation function. The hidden state of each edge will be updated iteratively T times based on information carried by all incoming bonds according to the following message passing algorithm:

$$\boldsymbol{h}^{t+1}_{\beta,ij} = \Omega\left[\boldsymbol{h}^0_{\beta,ij} + \mathbf{W}_2 \times \sum_{k\in\{V|\mathcal{N}(i)\setminus j\}} \boldsymbol{h}^t_{\beta,ki}\right] \tag{3}$$

where t is the iteration step counter in message passing, $\mathbf{W}_2$ is a learned weight matrix, and $\{V|\mathcal{N}(i)\setminus j\}$ refers to the set of neighboring vertices or atoms connected to the i^{th} atom, excluding the j^{th} atom. After all message passing steps, the learned edge hidden information ($\boldsymbol{h}_{\beta,ij}$=$\boldsymbol{h}^T_{\beta,ij}$, a.k.a. learned bond representation) is further converted to the learned atomic embedding, $\boldsymbol{h}_{\alpha,i}$, by concatenating the sum of all the incoming bond features with the original vertex features:

$$\boldsymbol{h}_{\alpha,i} = \Omega\left[\mathbf{W}_3 \times \text{concat}\left(\boldsymbol{\alpha}_i, \sum_{k\in\{V|\mathcal{N}(i)\}} \boldsymbol{h}^T_{\beta,ki}\right)\right] \tag{4}$$

where $\mathbf{W}_3$ is a learned weight matrix. The learned atomic ($\boldsymbol{h}_{\alpha,i}$) and bond ($\boldsymbol{h}_{\beta,ij}$) embedding from messaging passing can be used alone or optionally combined with additional QM or non-QM features in a way similar to the initial feature vectors shown in eq 2 with an additional linear layer, and therefore

$$\boldsymbol{r}_{\alpha,i}, \boldsymbol{r}_{\beta,ij} = \begin{cases} \boldsymbol{h}_{\alpha,i}, \boldsymbol{h}_{\beta,ij} & \text{default Chemprop} \\ \text{Linear}[\text{concat}(\boldsymbol{h}_{\alpha,i}, \boldsymbol{\alpha}_{\text{qm},i})] & \\ \text{Linear}[\text{concat}(\boldsymbol{h}_{\beta,ij}, \boldsymbol{\beta}_{\text{qm},ij})] & \text{with QM descriptors} \end{cases} \tag{5}$$

where the resulting atom ($\boldsymbol{r}_{\alpha,i}$) and bond ($\boldsymbol{r}_{\beta,ij}$) representations can be useful in predicting atom and bond targets (more discussion in the next section). Moreover, atom representations can be further transformed into a learned molecular representation, $\boldsymbol{h}_\mu$, via an aggregation function:

Table 1. Notations and Symbols for Describing Chemprop's D-MPNN Architecture

notation	description
a, $\mathbf{a}$, $\mathbf{A}$, $\mathbf{A}^{(n)}$	A scalar (n = 0), vector (n = 1), matrix (n = 2), n^{th}-order tensor.
$\text{concat}(\mathbf{a}, \mathbf{b}) = \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}$	Concatenation of vectors.
$\text{aggreg}(\mathbf{a}, \mathbf{b})$	Aggregation of vectors, e.g., summation, average, etc.
D-MPNN	A directed message passing neural network unit.
FFNN	A feed-forward neural network unit.
M	A molecule or molecular complex.
$G = M(V, E, \mathbf{F}^{(n)})$	A molecular graph with features or attributes.
V	Set of vertices or nodes or atoms in a molecular graph.
E	Set of edges or links or bonds in a molecular graph.
$\mathbf{F}^{(n)}$	The feature or attribute tensor of a molecular graph.
n_α	Number of nodes or atoms.
n_β	Number of edges or bonds.
$\boldsymbol{\alpha}_i$	Initial atom feature vector for i^{th} atom.
$\boldsymbol{\alpha}_{\text{chem},i}$	Atom chemical identity feature vector for i^{th} atom (Chemprop default).
$\boldsymbol{\alpha}_{\text{qm},i}$	Atom QM descriptor vector for i^{th} atom.
$\boldsymbol{\beta}_{ij}$	Initial bond feature vector between i^{th} and j^{th} atom.
$\boldsymbol{\beta}_{\text{chem},ij}$	Bond chemical identity feature vector between i^{th} and j^{th} atom (Chemprop default).
$\boldsymbol{\beta}_{\text{qm},ij}$	Bond QM descriptor vector between i^{th} and j^{th} atom.
$\boldsymbol{\mu}$	Additional molecular feature vector.
$\boldsymbol{h}_{\alpha,i}$	Learned D-MPNN atom embedding for the i^{th} atom.
$\boldsymbol{h}_\mu$	Learned D-MPNN embedding for molecule M.
$\boldsymbol{h}^0_{\beta,ij}$, $\boldsymbol{h}^t_{\beta,ij}$	Hidden states of directed bond ij initially, and at message passing step t.
$\boldsymbol{h}_{\beta,ij}$	Learned D-MPNN embedding for directed bond ij.
$\boldsymbol{r}_{\alpha,i}$	Representation of atom i as input to the FFNN.
$\boldsymbol{r}_{\beta,ij}$	Representation of directed ij as input to the FFNN.
$\boldsymbol{r}_\lambda = \boldsymbol{r}_{\alpha,i}$ or $\boldsymbol{r}_{\beta,ij}$	Representation of atom or bond as input to the FFNN.
$\boldsymbol{r}_\mu$	Representation of molecule M as input to the FFNN.
$\mathbf{W}_*$	A learnable weight matrix.
Ω	A nonlinear activation function.
$q_\lambda = q_i$ or q_{ij}, Q	Predicted atom or bond target value (unconstrained), and expected net value for molecule M.
$\varphi(x)$, ν, δ, n, $\mathbf{y}$	Radial basis function of scalar x, three associated parameters, and output vector $\mathbf{y}$.
$\mathbf{f}_\lambda$	Learned FFNN embedding of atom or bond.
k_λ, w_λ	Attention mechanism parameters.
q^{final}_λ	Final overall molecular constraint-corrected predicted target value.

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

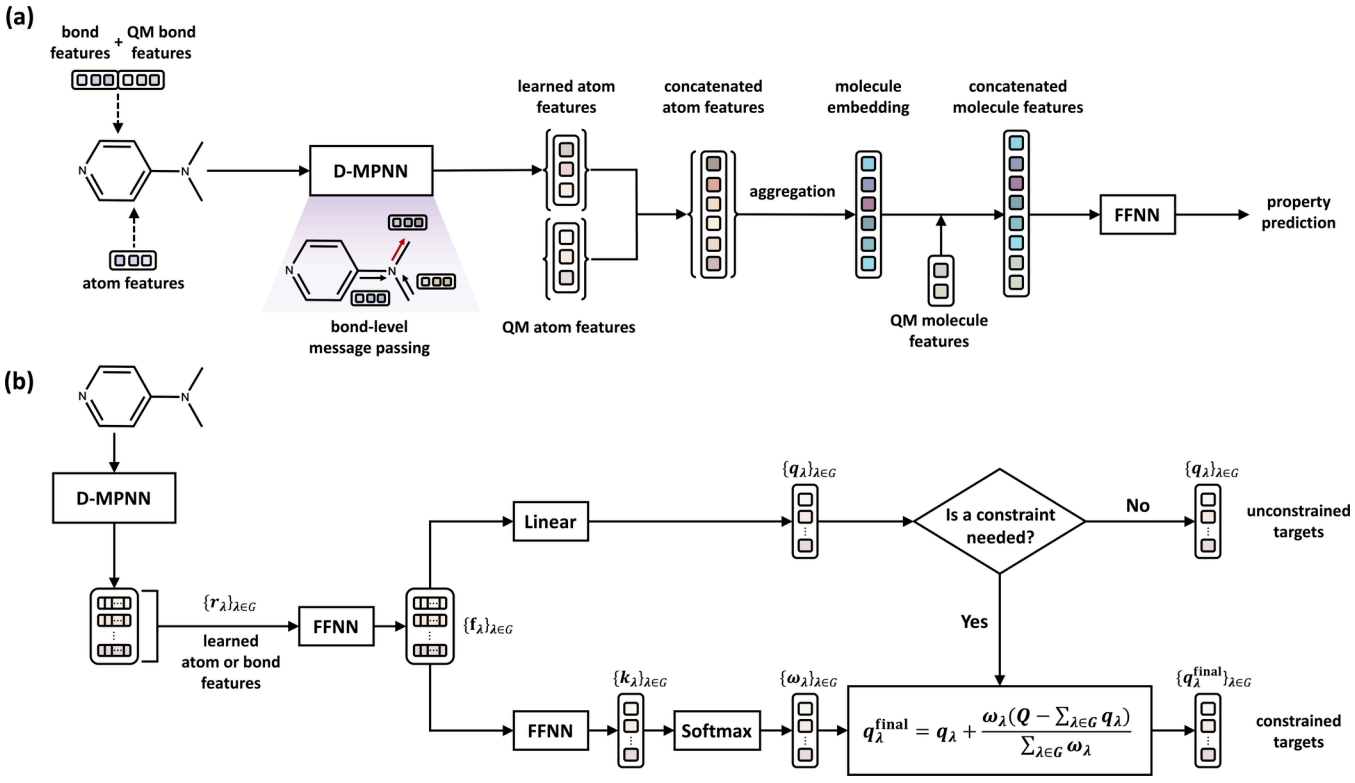


Figure 2. Schematic of the two machine learning model architectures employed in this study. (a) A directed message passing neural network (D-MPNN) augmented with quantum mechanical (QM) descriptors to enhance the prediction of chemical properties. (b) A multitask D-MPNN with an attention-based constraining method designed for simultaneously predicting multiple atom and bond properties while maintaining consistency with physical constraints. The notations and symbols used in the diagrams are defined in Table 1.

$$\boldsymbol{h}_\mu = \underset{i \in V}{\text{aggreg}}(\boldsymbol{r}_{\alpha,i}) \tag{6}$$

Similarly, additional molecular descriptors ($\boldsymbol{\mu}$) can be concatenated with the MPNN-learned molecular embedding ($\boldsymbol{h}_\mu$) to augment the representation:

$$\boldsymbol{r}_\mu = \begin{cases} \boldsymbol{h}_\mu & \text{default Chemprop} \\ \text{concat}(\boldsymbol{h}_\mu, \boldsymbol{\mu}) & \text{with QM descriptors} \end{cases} \tag{7}$$

The resulting molecular representation $\boldsymbol{r}_\mu$ can then be passed to the FFNN layer in Chemprop to predict the desired molecular properties.

This update to Chemprop software provides users with great flexibility and ease in providing additional atom, bond, and molecular descriptors as additional features at various stages of embedding learning. In general, the particular implementation choice should depend on specific tasks, and we leave this to the user to decide. The QM-augmented D-MPNN architecture used in this work is illustrated in Figure 2(a). This work focuses on molecular property prediction, and therefore additional atom features are concatenated with learned atom representations instead of initial atom features to minimize error propagation during message passing. In contrast, additional bond features are concatenated with initial bond features rather than the learned bond representation to ensure that information is passed to the learned molecular representation.

Although the inclusion of additional descriptors has the potential to improve model performance, a particular challenge emerges when integrating atom features that need to satisfy constraints on their summative value across an entire molecule, for the learned molecular representation is aggregated from atom representations (eq 6). In the current Chemprop framework, the aggregation functions, whether they involve averaging, summing, or normalizing by a constant, can lead to information loss in this scenario. To overcome this limitation, we transformed relevant QM atom and bond features from continuous values to expanded vectors by radial basis function (RBF) expansion[7] before incorporating these descriptors into D-MPNN as follows (this step is not automatically done in Chemprop):

$$\mathbf{y} = \varphi(x) = \left\{ \exp\left[-\frac{x - (\nu + \delta k)^2}{\delta} \right] \right\}_{k \in [0,1,\ldots,n-1]} \tag{8}$$

where x represents any atom or bond property that takes continuous values and $\mathbf{y}$ denotes its corresponding expanded vector. The RBF expansion involves three parameters: ν, the starting point of the vector, δ, the interval between the basis functions, and n, the total number of basis functions. By adjusting these parameters, we can achieve different cover ranges and resolutions in the expanded vector. The specific parameter values used in this study are detailed in Table S2.

Multitask Constrained D-MPNN for Atom and Bond Descriptors Prediction. A multitask constrained D-MPNN for atom and bond descriptor prediction has previously been developed by Guan et al.[7] Their implementation allows a D-MPNN to be trained on multiple atom and bond properties simultaneously. An attention-based constraining method is used to correct for the discrepancy between the sum of a predicted atom property for all atoms (or predicted bond property for all bonds, where applicable) in a molecule and its expected net value at the molecular level, such as the relationship between atomic partial charges and the net charge of a molecule. However, Guan et al.'s code implementation supports only neutral species, and constraints can only be applied to atom properties but not to bond properties. Moreover, the total loss across different targets is not scaled automatically; instead, the weights of each target need to be specified. Due to these limitations, we have modified the algorithm and implemented a more flexible version in Chemprop[1] release 1.6, which is publicly available on GitHub.[41] As shown in Figure 2(b), the atom and bond embeddings are obtained from the D-MPNN procedure as mentioned in the previous section. To predict atom and bond properties, atom ($\boldsymbol{r}_{\alpha,i}$) and bond ($\boldsymbol{r}_{\beta,ij}$) embeddings are

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

Table 2. Overview of the Benchmark Data Sets Employed in This Study, Highlighting Their Key Characteristics and the Nature of the Data They Contain[a]

data set	category	description of target properties	task type	no. of tasks[b]	no. of data points	ref
QM7	quantum mechanics	atomization energy	regression	1	6,830	65
QM8	quantum mechanics	electronic spectra and excited state energy	regression	12	21,786	66
QM9	quantum mechanics	energetic, electronic and thermodynamic properties	regression	12	133,885	67
ESOL	physical chemistry	water solubility	regression	1	1,040	68
FreeSolv	physical chemistry	hydration free energy in water	regression	1	630	69
Lipophilicity	physical chemistry	octanol/water distribution coefficients	regression	1	4,200	70
IP	physical chemistry	ionization potential	regression	1	2,147	64
CritProp	physical chemistry	critical properties and phase change properties	regression	8	5,542	34
PDBbind-F	biophysics	protein binding affinity (full subset of PDBbind)	regression	1	9,880	71
PDBbind-C	biophysics	protein binding affinity (core subset of PDBbind)	regression	1	168	71
PDBbind-R	biophysics	protein binding affinity (refined subset of PDBbind)	regression	1	3,040	71
HIV	biophysics	inhibition of HIV replication	classification	1	41,127	72
BACE	biophysics	inhibition of human β-secretase 1	classification	1	1,513	73
BBBP	physiology	ability to penetrate the blood-brain barrier	classification	1	2,039	74
Tox21	physiology	toxicity	classification	12	7,831	44
ClinTox	physiology	toxicity	classification	2	1,478	21

[a]The data in the QM7, QM8, and QM9 data sets are calculated using quantum mechanics, while the rest are measured experimentally. [b]For data sets with multiple targets, the model is cotrained on all the targets.

linked with the corresponding targets by feedforward neural networks. Hence, the predicted atom (q_i) or bond (q_{ij}) property can be described as

$$q_i = \mathrm{FFNN}(\boldsymbol{r}_{\alpha,i}) \quad \text{atom target} \tag{9}$$

$$q_{ij} = \mathrm{FFNN}[\mathrm{concat}(\boldsymbol{r}_{\beta,ij}, \boldsymbol{r}_{\beta,ji}), \mathrm{concat}(\boldsymbol{r}_{\beta,ji}, \boldsymbol{r}_{\beta,ij})] \quad \text{bond target} \tag{10}$$

For an atom target, the input size of FFNN is equal to the hidden size (i.e., the dimension of the embedding). However, for bond property prediction, the input size is twice the hidden size because $\boldsymbol{r}_{\beta,ij}$ and $\boldsymbol{r}_{\beta,ji}$ are two distinct vectors as a result of using directed bonds during message passing, but both latent representations are connected to the same bond target that does not depend on direction. Therefore, we concatenate the embeddings of the same bond into concat($\boldsymbol{r}_{\beta,ij}$, $\boldsymbol{r}_{\beta,ji}$) and concat($\boldsymbol{r}_{\beta,ji}$, $\boldsymbol{r}_{\beta,ij}$), and subsequently pass both into the same FFNN and average the two outputs to ensure that the concatenation of forward and backward bonds is order-invariant.

It is important to note that in the framework described above, for certain properties, such as atomic partial charge, the sum of the predicted values $\sum_{\lambda\in G} q_\lambda$ for all atoms (or bonds, where applicable) may deviate from the expected net value Q of a molecule because each q_λ is predicted without knowledge of the overall constraint:

$$\mathbf{f}_\lambda = \mathrm{FFNN}(\boldsymbol{r}_\lambda) \tag{11}$$

$$q_\lambda = \mathrm{Linear}(\mathbf{f}_\lambda) \tag{12}$$

where $\boldsymbol{r}_\lambda$ denotes the D-MPNN learned representation (can be augmented by QM descriptors as shown in eq 5) of an atom or a bond, and $\mathbf{f}_\lambda$ is an atom or bond embedding subsequently learned from a FFNN. The unconstrained prediction of q_λ for an individual atom or bond target is then computed via a linear transformation on $\mathbf{f}_\lambda$. Because q_λ is predicted independently for each individual atom or bond without imposing any overall constraints at the molecular level, additional engineering of the model architecture is required to ensure that the predictions for all atoms or bonds as a whole satisfy corresponding physical constraints for an entire molecule whenever applicable. To resolve this issue, a seemingly intuitive approach at first glance would be to spread the error evenly across each atom, but the contribution of each atom to the aggregated value is not the same. Therefore, to better address the discrepancy between $\sum_{\lambda\in G} q_\lambda$ and Q, we adopt an attention-based constraining technique from Guan et al.[7] to determine the weight of correction for each atom, which is inspired by the attention mechanism in natural language processing.[42] Finally, the constraint-corrected value of $q_\lambda^{\mathrm{final}}$ is obtained by

$$k_\lambda = \mathrm{FFNN}(\mathbf{f}_\lambda) \tag{13}$$

$$w_\lambda = \frac{\exp(k_\lambda)}{\sum_{\lambda\in G}\exp(k_\lambda)} \tag{14}$$

$$q_\lambda^{\mathrm{final}} = q_\lambda + \frac{w_\lambda\left(Q - \sum_{\lambda\in G} q_\lambda\right)}{\sum_{\lambda\in G} w_\lambda} \tag{15}$$

where k_λ is a vector learned from feeding the same embedding $\mathbf{f}_\lambda$ used to predict q_λ into a separate FFNN, and k_λ can be seen as a high level representation of an inquiry about how to adjust the predicted values of each individual atom or bond to satisfy the overall molecular constraints. The final, constraint-adjusted prediction for each atom or bond target, $q_\lambda^{\mathrm{final}}$, is determined by correcting the initial unconstrained prediction, q_λ, through a weighted adjustment. For each target property, this adjustment is based on the discrepancy between the sum of the predictions for all atoms or bonds, $\sum_{\lambda\in G} q_\lambda$, and the overall expected value, Q, of the molecule. The adjustment for each individual atom or bond prediction is proportionally allocated using a normalized weight, w_λ, calculated via a softmax function to represent the fraction of the total discrepancy that each individual adjustment accounts for.

Data Preparation. To provide a diverse exploration of the chemical space, we select 64,921 representative molecules containing C, H, O, N, P, S, Si, F, Cl, and Br atoms from a functional group-based sampling of molecules in publicly available databases including GDB-17,[43] Tox21,[44] CPDat,[45] ChEMBL,[46] ZINC20,[47] DrugBank,[48] as well as novel dye-like substructures[15] and amines for carbon capture in other internal data sets. The selected representative molecules embody the functional groups most frequently observed across all the databases we sampled from. Atom, bond, and molecular QM descriptors of these representative molecules are computed utilizing a refined version of the automated workflow originally developed by our group.[49] This updated version offers enhanced atom/bond descriptors calculation capabilities, along with improved support for charged molecules. More specifically, a total of 37 QM descriptors are generated for each of the ~65,000 species in the curated data set of this work. These descriptors consist of 13 atom descriptors, 4 bond descriptors, and 20 molecular descriptors. The atom descriptors include NPA charges, Parr functions,[50] NMR shielding constants, and valence orbital occupancies. The bond

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

descriptors consist of bond order, bond length, bonding electrons, and bond natural ionicity. The molecular descriptors encompass energy gaps, ionization potential (IP), electron affinity (EA), and dipole and quadrupole moments. Each descriptor is detailed in Table S1, selected for its universality to molecules, feasibility and ease of calculation via Density Functional Theory (DFT), and relevance to important applications (e.g., partial charge's role in molecular dynamics).

In this study, 3D molecular conformers are generated using the MMFF94s[51] force field in RDKit[52] from input SMILES strings. For each molecule, we generate 20 conformers, and the conformer with the lowest MMFF94s energy undergoes further optimization at the GFN2-xTB[53] level of theory, followed by a frequency calculation at the same level to ensure it had no imaginary frequencies. Furthermore, we perform three DFT single point calculations (neutral, cation +1, anion −1) using six different functionals (ωB97XD,[54] B3LYP,[55] M06-2X,[56] PBE0,[57] TPSS,[58] and BP86[59,60]) with the def2-SVP[61] basis set in Gaussian 16.[62] This selection includes popular pure and hybrid functionals used in a variety of chemical domains. Structural and vibrational frequency checks are implemented to ensure the convergence of the final optimized geometries. Natural bond orbitals and related descriptors are computed using NBO 7.0.[63] The QM calculation outcomes are processed to obtain the desired descriptors through the use of scripts integrated into the automated workflow.

To investigate the impact of QM descriptors on D-MPNN performance, we evaluate the performance of the QM-augmented D-MPNN on 16 public data sets.[21,34,64] This includes classification and regression tasks on data sets spanning various fields such as quantum mechanics, physical chemistry, biophysics, and physiology. The data set sizes vary from several hundred to one hundred thousand data points. Detailed descriptions of each data set are listed in Table 2. It is important to note that the number of data points in some data sets does not match the original numbers from Wu et al.,[21] due to the existence of duplicated molecules, failed processing by RDKit,[52] and failed SMILES conversion, as detailed in Yang et al.[25] Additionally, we have fewer molecules in the ESOL and FreeSolv data sets because we additionally calculate QM descriptors for these species as part of our analysis, and some of these fail in the calculations.

Model Training. To predict QM descriptors on the fly, three separate D-MPNN models are trained on three subsets of QM descriptors computed using ωB97XD/def2-SVP: one for atom and bond properties, another for molecular properties of dipole and quadrupole moments, and the last for energy gaps, IP, and EA. For the model trained on atom/bond properties, the summation constraint for partial charges is set as the net charge for each species. Additionally, the nucleophilic/electrophilic Parr functions have a summation constraint of 1, while there are no constraints for other atom/bond descriptors. We split the data set of 64,921 data points into 80% training, 10% validation, and 10% test splits for hyperparameter search with 30 Bayesian (Tree Parzen Estimator as implemented in the hyperopt package[75]) search iterations. The fine-tuned hyperparameters include the depth and hidden size of message passing layers, the number of layers and hidden size in the FFNNs, and the dropout rate. We evaluate the models' ability to predict QM descriptors using random and Bemis-Murcko scaffold splits.[76] The scaffold split is considered to be a more rigorous approach to measuring a model's generalizability, as it is based on molecular backbones.[77,78] Each model is trained for 50 epochs during hyperparameter optimization and final model training.

For the benchmark data sets, we conduct hyperparameter optimizations using 20 iterations on 10 randomly seeded 80:10:10 data splits for training, validation, and testing. However, due to computational cost, models trained on QM9 and HIV use only three splits. Each model is also trained for 50 epochs during both hyperparameter optimization and final model training. In addition, the final models are trained with an ensemble size of 5. Scaled sums are used to aggregate atom features into molecular feature vectors for both hyperparameter tuning and model production.

We use the models trained on the QM descriptors data set to predict QM descriptors for each molecule in the 16 benchmark data sets. For data sets with more than one target, the model is cotrained on all the targets. The predicted atom/bond QM features are further expanded via RBF expansion, as described earlier. As depicted in Figure 2(a), the expanded atom descriptors are concatenated with the atom hidden representation, the expanded bond properties are concatenated with the bond features, and the molecular properties are concatenated with the learned molecular fingerprint in D-MPNN. We train D-MPNN models with QM features predicted from ML models, which we refer to as ml-QM-GNN models in the following sections. On the other hand, the D-MPNN models trained with the QM descriptors directly calculated by ωB97XD/def2-SVP are referred to as QM-GNN models. We use the same optimized hyperparameters as used in GNN models for ml-QM-GNN and QM-GNN models to distinguish the effect of QM descriptors within the GNN model from hyperparameter choices.

Shapley Value Analysis. Shapley value analysis is a widely used method for explaining machine learning models by evaluating the expected (average) marginal contribution of including individual features. This approach allows for the determination of feature importance and enables model interpretation through a linear combination of features. For a detailed theoretical background on Shapley value analysis, we refer readers to the work of Lundberg and Lee.[79] In this work, to elucidate why certain QM descriptors appear to enhance or do not affect the performance of the GNN models, we implemented a customized version of Chemprop v1 software capable of performing Shapley value analysis. For the implementation, we utilized the popular SHAP (SHapley Additive exPlanations) package,[79] specifically the *PermutationExplainer* within SHAP.

However, direct application of SHAP to Chemprop is not straightforward, as the *PermutationExplainer* typically expects the input to be a single feature vector for neural networks like FFNNs. In contrast, Chemprop's molecular fingerprints are featurized in several steps, as explained earlier, and thus the input to default Chemprop cannot be a single vector containing all default and additional QM features. To perform Shapley value analysis in a molecular context, we made modifications to the Chemprop codebase and developed additional wrapper classes to enable Shapley value computation using Chemprop with the SHAP package. This implementation is open-source and available on GitHub with examples (see Code and Data Availability), allowing interested users to perform similar SHAP analyses. Currently, SHAP is implemented with Chemprop v1, as the models in our paper are based on this version. Future work will include Shapley value analysis implementation for Chemprop v2.

For our Shapley value analysis, we focused on the ESOL and FreeSolv data sets, as we have calculated true QM descriptors for these data sets. The analysis was conducted independently on ML models trained on 80% of each data set. For each ML model, we computed Shapley values for all features across all molecules in the corresponding data set. Specifically, for each molecule, we calculated the expected Shapley value of each feature using the *PermutationExplainer*, sampling 1,000 different combinations of included and excluded features during prediction time. This process was repeated for all molecules in both data sets. We then averaged the absolute Shapley values of each feature across all molecules to determine the expected average effect of each feature on the model outcome. As an implementation note, we computed Shapley values using scaled feature and target values to ensure fair comparison across different property scales. The scaling procedure follows Chemprop's default implementation, resulting in scaled data with a mean of zero and standard deviation of one. Consequently, the computed Shapley values should be interpreted relative to the scaled data statistics to determine their effect on the model. However, regardless of scaling, the relative importance of features can be compared directly via the magnitude of their Shapley values.

RESULTS AND DISCUSSION

Impact of DFT Level of Theory on QM Descriptors. As mentioned in the Methods section, we compute QM descriptors at six DFT levels of theory (LOT): two pure

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

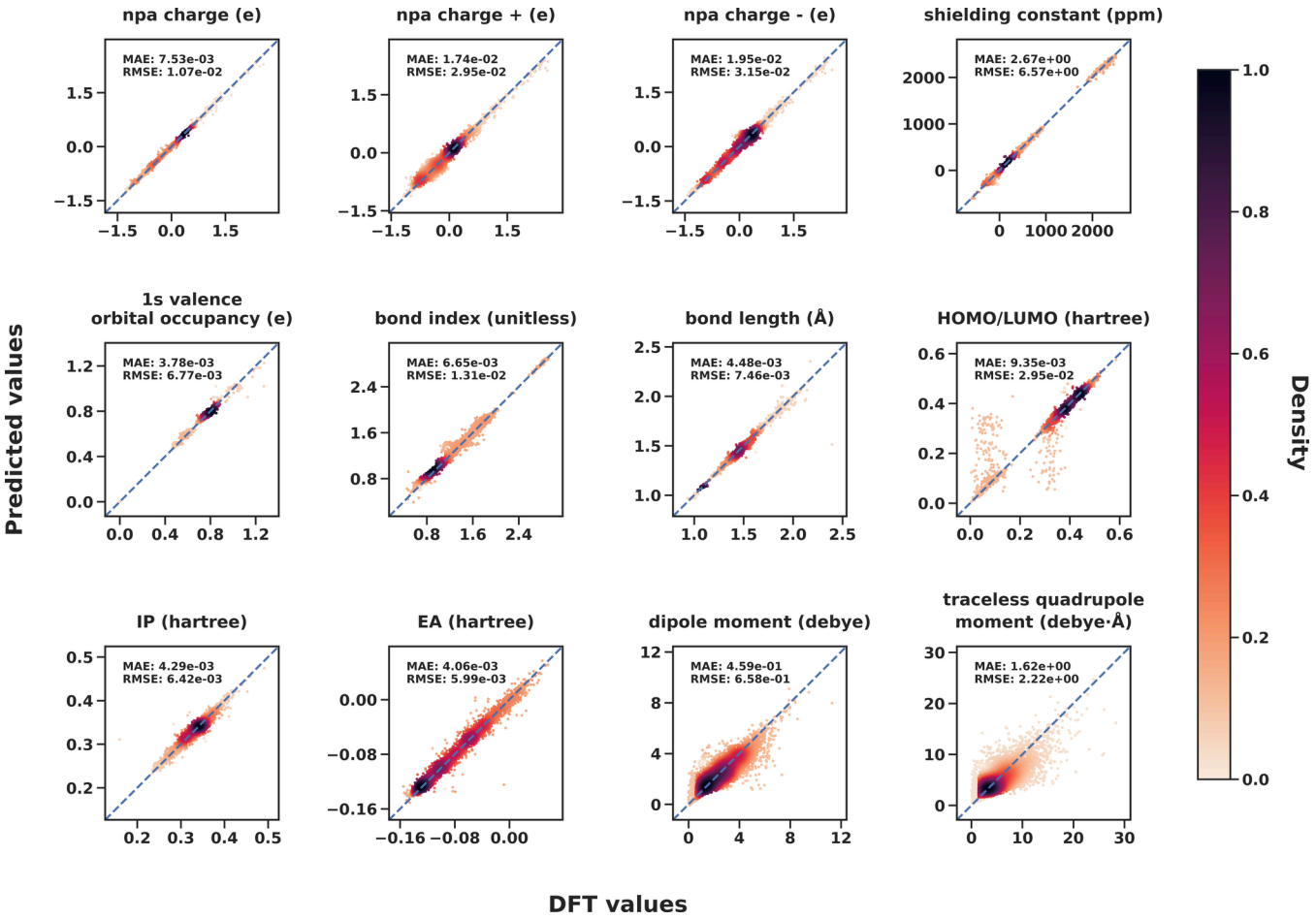


Figure 3. Parity plots comparing predicted QM descriptors against their corresponding DFT calculated values for 12 selected features (full version in Supporting Information). Each data point in the plots represents a molecule from the held-out test sets obtained through random splitting of the full data set. The color scale indicates the relative density of the data points in each region of the plot with a density value of 1 representing the highest concentration of points. Note that this measure of local data density does not reflect the absolute percentage of data points in the entire data set. The proximity of the data points to the diagonal line indicates a strong agreement between the predicted and DFT-calculated values.

functionals, TPSS and BP86, two global hybrid functionals, B3LYP and PBE0, a range-separated hybrid functional ωB97XD, and a hybrid meta-GGA functional, M06-2X. The goal is to examine the effect of different DFT levels on QM descriptor values, as these descriptors are often calculated at various levels in different studies, making it crucial to understand how choice of DFT method affects the descriptors. The correlation matrix for each QM descriptor across six DFT levels is shown in Figure S2. The correlations for atom or bond descriptors at different LOTs are high, with correlation coefficients often larger than 0.9. The correlation coefficients for bond length are 1 since identical geometries optimized at GFN2-xTB are used to run single-point calculations with different functionals to derive the QM descriptors. In addition to atom and bond descriptors, molecular descriptors including dipole moment, quadrupole moment, and EA also show very strong correlations across different LOTs.

We observe slightly lower correlations in energy gap and IP. The results suggest that ωB97XD and M06-2X correlate well with each other, as well as B3LYP and PBE0. This is potentially due to ωB97XD and M06-2X having similar percentages of exact exchange to one another, with the same being true for B3LYP and PBE0. For IP calculations, all hybrid functionals correlate with each other quite well, with the correlation coefficient larger than 0.99. Energy gaps, such as the HOMO/LUMO gap, are determined by the molecular orbitals, which are affected by the construction of the self-consistent field (SCF) density. Furthermore, IP calculation is based on the difference in SCF energy between the neutral and charged states. During the SCF procedure, the SCF density is updated iteratively until the SCF energy converges. Given the dependence of these descriptors on the SCF density, it is reasonable to expect these descriptors to be sensitive to the choice of method, since different DFT methods utilize various approximations to solve the Kohn−Sham equations. However, it is surprising that the correlation coefficients for EA calculated by different levels of theories are quite high. It is important to note that, while the values are correlated, there is an offset (Figure S3), as indicated by the R^2 values. In summary, the high correlation observed among the values of the chosen QM descriptors across different DFT levels indicates that using a single DFT level for their calculation is both practical and sufficient in most studies.

QM Descriptors Prediction. Given the strong correlations observed among the values of most QM descriptors calculated using different model chemistry, we hypothesize that the choice of DFT level used to obtain the QM descriptors considered in this study would have minimal impact on the performance of the QM-augmented model. Because performing QM calculations to derive descriptors for each new

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

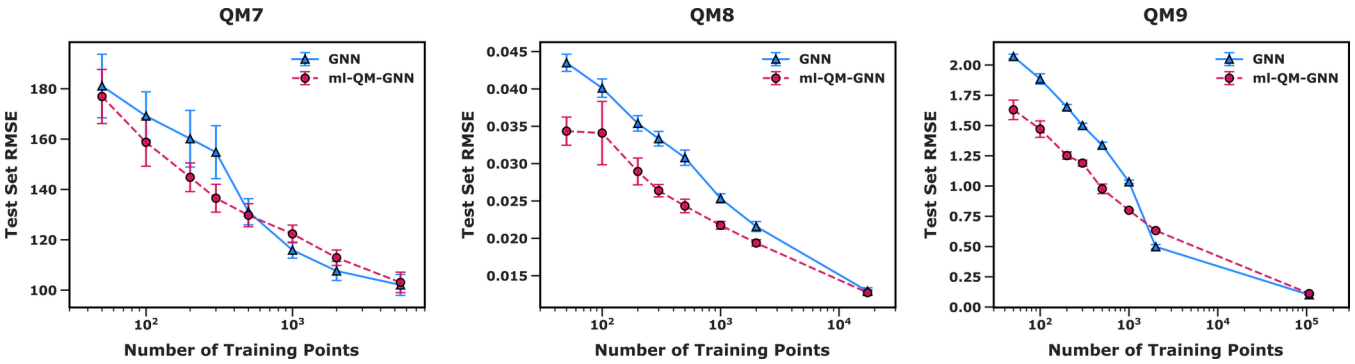


Figure 4. Test RMSE of models trained on random split training sets of varying sizes sampled from three computational chemistry data sets detailed in Table 2: QM7, QM8, and QM9. The error bars represent the standard deviation of the RMSE values across 10 folds for the QM7 and QM8 data sets and 3 folds for the QM9 data set. QM descriptors are more likely to provide a greater reduction in prediction error when the size of the training set is limited.

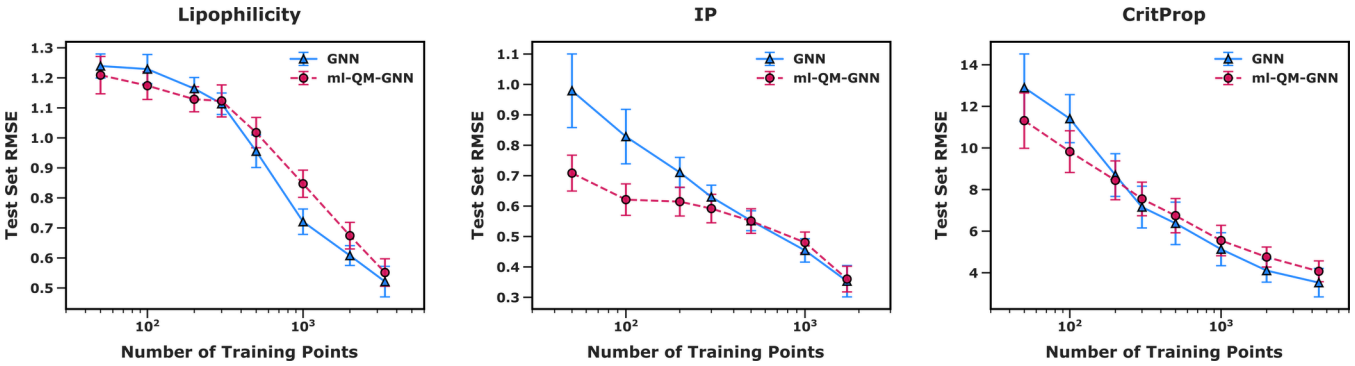


Figure 5. Test RMSE of models trained on random split training sets of varying sizes sampled from three experimental physical chemistry data sets detailed in Table 2: lipophilicity, IP, and CritProp. The error bars represent the standard deviation of the RMSE values across 10 folds for each data set. QM descriptors can help predict experimentally measured properties, particularly when training data is limited.

molecule can be expensive and time-consuming, we train ML models to predict these descriptors easily and efficiently. Herein, we train three models on different subsets of QM descriptors using ωB97XD/def2-SVP//GFN2-xTB as the model chemistry. We also tested using a single model cotrained on all descriptors, but this single model had worse performance than the three individual models ultimately used in this work. (Table S4). More specifically, one model is trained on atom/bond descriptors, another on molecular descriptors of dipole and quadrupole moments, and the third one on energy gaps, IP, and EA. The performance of each model is tested on held-out test sets for both random and scaffold splits. As shown in Figures 3 and S4, the models generally show a good correlation between predicted values and DFT values for most descriptors, suggesting that our D-MPNN models can predict QM descriptors well. For some properties, such as Parr functions and bonding electrons, significant errors can be observed within certain data regimes. These errors can be attributed to the underlying data distribution of these descriptors. It is probable that the model predictions tend to gravitate toward the values around which the majority of the data points are concentrated, as it seeks to optimize its performance based on the prevalent patterns in the data. Table S3 summarizes the model performance on each QM descriptor prediction for random and scaffold splits. The testing mean absolute error (MAE) and root-mean-square error (RMSE) for scaffold splits are also close to random splits, suggesting that the presented models can generalize beyond their training set and be used to predict QM descriptors for molecules with backbones that have not been seen during model training. The D-MPNN models for various QM descriptor predictions developed here are highly valuable in high-throughput screening workflow, as performing QM calculations is usually very costly. For example, DFT single point calculations of the ~65,000 chemical species using ωB97XD/def2-SVP in this work consumed approximately 20,000 CPU hours, while predicting all these values using ML models takes less than 0.5 CPU hour on the same computer.

Impact of QM Descriptors on D-MPNN Performance for Molecular Property Prediction. To examine the effect of integrating QM descriptors into D-MPNNs for diverse molecular property prediction tasks, we first predict the QM descriptors (listed in Table S1) for each compound in the 16 benchmark data sets (described in Table 2). Subsequently, we employ the predicted descriptors as supplementary features in the D-MPNN models (as described in Methods) to predict the corresponding target properties. The following analysis focuses on three key aspects: (1) the influence of training data set size and the correlation between descriptors and targets on model performance, (2) the importance of refined feature selection in improving model accuracy and interpretability, and (3) the effects of incorporating QM descriptors on model sensitivity and generalizability.

Impact of Training Data Set Sizes and Descriptor Correlation with Targets. The results indicate that the ml-QM-GNN models generally match the accuracy of the standard GNN models in most data sets (Table S8). However, one of our main focuses is to determine whether the benefits of

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

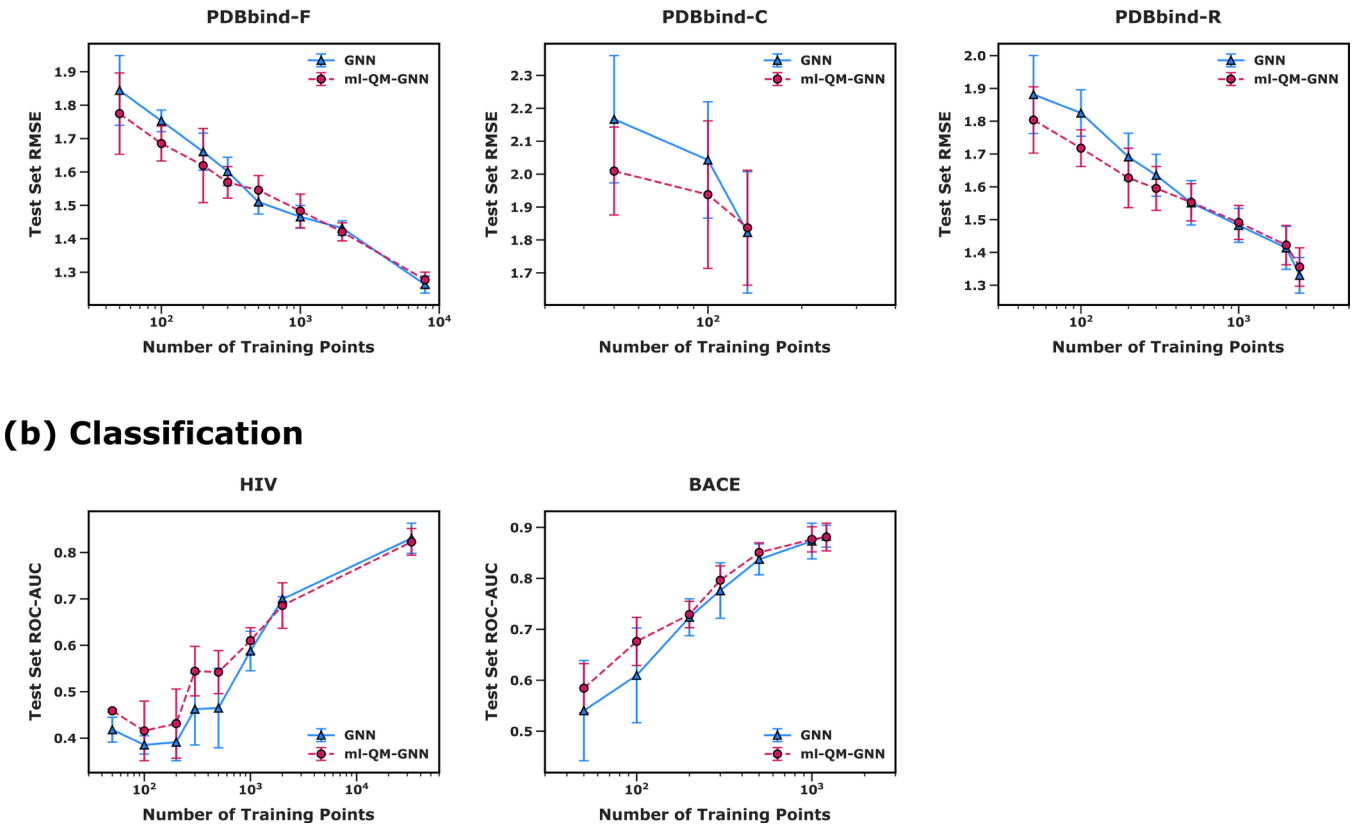


Figure 6. Test performance of models trained on random split downsampled subsets sampled from different biophysics data sets detailed in Table 2, separated into two categories: (a) regression data sets (PDBbind-F, PDBbind-C, and PDBbind-R) using RMSE as metric (lower the better) and (b) classification data sets (HIV and BACE) using receiver operating characteristic—area under the curve (ROC-AUC) as metric (higher the better). The error bar is calculated as the standard deviation of test errors across 3 folds for HIV data set and 10 folds for all other data sets. QM descriptors can provide some benefit even for targets lacking strong physical relationships, particularly with limited training data.

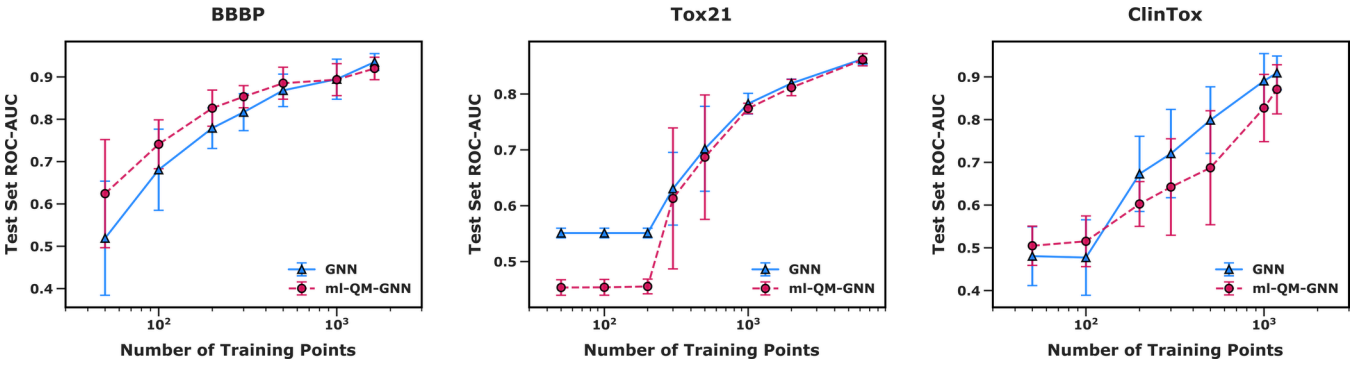


Figure 7. Test ROC-AUC of models trained on random split training sets of varying sizes sampled from three experimental physiology data sets detailed in Table 2: BBBP, Tox21, and ClinTox. The error bars represent the standard deviation of the errors across 10 folds for each data set. QM descriptors used in this work do not appear to benefit toxicity prediction for the Tox21 data set.

introducing QM descriptors are more significant in smaller data sets, a scenario commonly encountered in experimental and *de novo* molecular discovery studies, where data sets are often limited to a few hundred molecules or reactions due to the intensive labor involved in the synthesizing and analyzing new compounds. By selectively reducing the size of training and validation data sets (a.k.a. downsampling), we demonstrate that QM descriptors can consistently improve model performance across various tasks, especially for small data sets up to around 2000 entries (Figures 4 to 7). For instance, the ml-QM-GNN model surpasses the performance of the standard GNN model when the training data set contains fewer than 500 samples for the QM7 data set and fewer than 2000 samples for both the QM8 and QM9 data sets (Figure 4).

Leveraging QM descriptors in D-MPNNs proves advantageous not only for QM data sets, where a significant benefit is not surprising because both the target properties and input descriptors are derived from similar computational chemistry calculations, but also for predicting experimental physical chemistry properties (Figure 5). For instance, augmenting the

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

D-MPNN model with QM descriptors substantially enhances its performance in predicting ionization potential (IP) when the training set contains fewer than 500 samples. However, this marked improvement in predicting experimental IP is partially anticipated, as calculated IP values are included as additional molecular features in the model. Similarly, the QM-augmented model demonstrates improved performance in predicting lipophilicity when the training set comprises fewer than 300 samples. We also observe that the ml-QM-GNN model outperforms the standard GNN in the small-data regime for predicting critical properties of fluids. These findings indicate that QM descriptors can potentially assist D-MPNN in capturing relevant information that improves its predictive performance for experimentally measured properties, beyond those derived from QM calculations, particularly when data is scarce.

In the context of biophysics (Figure 6) and physiology (Figure 7), QM-augumented models exhibit a modest reduction in prediction errors for various tasks when data is limited, except for the Tox21 data set. The positive influence of QM descriptors on certain property prediction tasks in biophysics and physiology is particularly intriguing, considering the seemingly weak physical correlation or intuitive relationship between generic DFT descriptors and specific targets, such as inhibition of HIV replication. This observation suggests that while the improvements due to QM descriptors are not always substantial, there are indications that QM descriptors might assist D-MPNN models in capturing certain fundamental molecular characteristics that may potentially contribute to the prediction of more complex biological properties, even in cases where a strong direct physical relationship is not immediately apparent. However, further research is needed to confirm this hypothesis and understand the precise mechanisms involved.

In addition to models trained on randomly downsampled data sets (Figures 4−7), we also evaluated model performance when downsampling and using scaffold splits (Figure S6). Notably, for the scaffold-split IP data set, QM descriptors can offer advantage at a larger data set size compared to random splits. Moreover, the ml-QM-GNN models tend to exhibit narrower error bars compared to their conventional GNN counterparts in scaffold-split QM8, QM9, and IP data sets, suggesting that the inclusion of additional descriptors can potentially improve the reliability of the predictions. However, for most other data sets, the performance trends closely mirrored those observed in random splits. This similarity may be attributed to the inherent uncertainty in predicted descriptors, which could limit the model's ability to improve on more challenging scaffold splits compared to random splits. We explore the impact of QM descriptor quality on model performance in greater detail in the following sections.

To further corroborate our hypothesis that the choice of DFT method for computing QM descriptors has minimal impact on our analysis, we conducted a comparative study using B3LYP/def2-SVP. Figure S5 demonstrates that the results obtained with B3LYP/def2-SVP are remarkably consistent with those generated using ωB97XD/def2-SVP, further confirming our hypothesis.

We further quantified the enhancements provided by QM descriptors across different regression tasks using models trained on 100 downsampled data points, and the findings indicate that QM descriptors generally offer greater benefits for targets that exhibit a stronger correlation or intuitive relationship with the descriptors (Figure 8). However, the advantage of employing QM descriptors tends to gradually

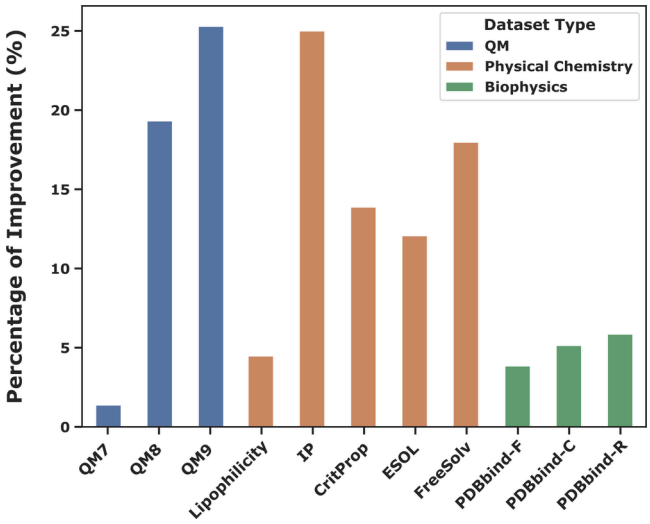


Figure 8. Percentage improvement in test RMSE across various regression data sets resulting from the transition from GNN to QM-augmented GNN models trained on 100 data points. The ESOL and FreeSolv models utilize calculated QM descriptors, while the other models employ ML-predicted QM descriptors. Greater improvements are observed for targets that exhibit stronger correlations or intuitive relationship with the QM descriptors.

diminish as more training data become available. This observation suggests that with an expanded data set, model performance may increasingly depend on the relationship between fundamental molecular graph information and prediction targets, rather than additional descriptors. Consequently, computing and including QM descriptors may not be cost-effective in scenarios where the training set is already sufficiently large. However, the ability of QM descriptors to provide meaningful improvements to the prediction of many molecular properties, despite the difference in the nature and origin of data shown, highlights their versatility and potential to improve the performance of predictive models across various chemical domains, especially in scenarios where limited training data is available.

Impact of Refined Feature Selection on Model Accuracy and Interpretability. The analysis above suggests that QM-GNN models tend to perform better when the descriptors correlate with the targets. This observation leads us to hypothesize that carefully selecting subsets of QM descriptors with high correlations to the targets could further improve model performance. Intuitively, using too many irrelevant features may introduce noise and distract the models, particularly in the small-data regime. Furthermore, revealing which descriptors are helpful for specific tasks can potentially provide some insights into the physicochemical principles underlying the descriptor-target relationships and subsequently facilitate a better interpretability and comprehension of model decisions. However, conducting a full ablation study on all descriptors and data sets would be computationally expensive and beyond the scope of this work. Therefore, we perform a preliminary assessment to investigate whether particular types of descriptor can explain most of the improvement.

Specifically, we train models using subsets of descriptors based on whether they are atom, bond, or molecular features,

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

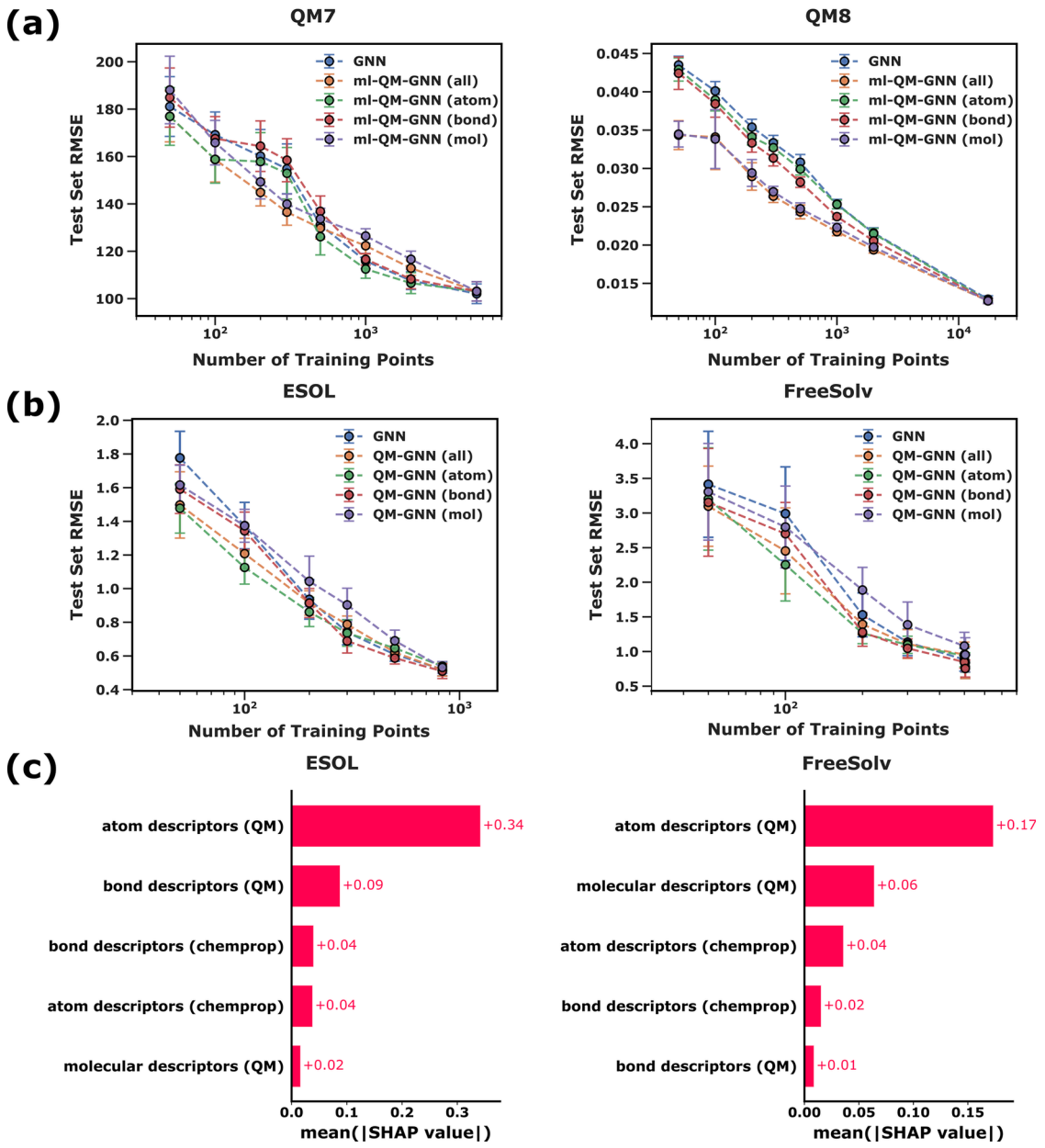


Figure 9. Test RMSE for models trained on downsampled random split training sets sampled from selected data sets: (a) QM7 and QM8 and (b) ESOL and FreeSolv. These models are augmented with different subsets of QM descriptors as additional features. The subsets include all QM descriptors (all), only atom descriptors (atom), only bond descriptors (bond), and only molecular descriptors (mol). Error bars represent the standard deviation of errors across 10 folds. Descriptor selection can greatly impact model performance, with the optimal subset varying across tasks, thereby underscoring the importance of a strategic approach to descriptor selection. (c) The mean absolute Shapley (SHAP) values for each descriptor subset illustrating their relative importance in predicting outcomes for the ESOL and FreeSolv data sets. Descriptor subsets are arranged in descending order of importance from top (most) to bottom (least). The values represent the average magnitude of each subset's impact on model predictions and are calculated using scaled data to ensure fair comparison across different property scales; see the Methods section for detailed computational procedures.

and compare their performance (Figures 9 and S8). The results suggest that models trained with the full set of descriptors generally perform comparably to models trained on the optimal subset of descriptors. This indicates that D-MPNNs are capable of identifying relevant features from all input descriptors, provided that the optimal subset is included within the full set. In some cases, a carefully chosen subset of descriptors can slightly outperform the full set, likely due to the reduction of extraneous noise from irrelevant descriptors.

Notably, molecular descriptors are identified as particularly beneficial for most of the molecular property prediction tasks we examine. However, variations in the utility of descriptor types are also observed across different data sets. For example, atom and bond descriptors seem to benefit ESOL and FreeSolv model performance to varying degrees (Figure 9b). To further quantify and elucidate the impact of different descriptors on model predictions, we calculated Shapley values for each input feature in the ESOL and FreeSolv data sets (Figure 9c and S7), following the procedures outlined in the Methods section. The Shapley value analysis yields several important insights. First, atom QM descriptors emerge as the most influential predictors for both data sets, agreeing with the results from models

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

trained on subsets of descriptors (Figure 9b). This finding also strongly suggests a critical role for atom-level QM properties in capturing molecular characteristics related to solvation. Moreover, QM descriptors demonstrate a more substantial impact on model predictions compared to Chemprop's default features, highlighting the significant value of incorporating QM information into the models. Interestingly, despite the analysis being conducted on models trained with 80% of the data set, a regime where the performance gap between QM-GNN and standard GNN models narrows (Figure 11), the results reveal that QM descriptors continue to exert a meaningful influence on model predictions.

Although the results from models trained on subsets of descriptors seem to suggest that using a comprehensive set of descriptors generally performs comparably to models trained on the optimal subset, this parity is contingent upon the inclusion of the optimal subset within the full set. Crucially, selecting an inappropriate subset of QM descriptors can significantly impair model performance, emphasizing the importance of a detailed analysis to identify the most effective descriptors. Given these insights, we recommend a strategic approach to descriptor selection, starting with a broad set for initial rapid testing and then refining the set through systematic evaluation. Adopting this strategy can potentially not only enhance model performance by minimizing errors due to irrelevant features but also help elucidate the underlying reasons contributing to the effectiveness of specific descriptors on particular targets. Moreover, a refined descriptor set can streamline model application and future investigations by lowering the computational demands of descriptor calculation for new molecular inquiries.

Various methods can be employed to explain model performance due to particular features and potentially be implemented with Chemprop. In this study, we demonstrated the use of Shapley value analysis as a valuable tool for identifying meaningful descriptors and building more explainable models. Results from the Shapley value analysis not only corroborates the importance of including QM descriptors but also provides a quantitative measure of their impact, even in scenarios where their benefits might not be immediately apparent from overall model performance metrics. Furthermore, the analysis offers valuable insights into the relative importance of different descriptor types, potentially guiding future efforts in feature selection and model interpretability. It is worth noting that there are many other methods available for such tasks, including ablation studies,[80] Local Interpretable Model-agnostic Explanations (LIME),[81] and counterfactual explanations.[82] Each of these techniques offers unique perspectives on feature importance and model interpretability, and their application could provide further insights into the role of QM descriptors in molecular property prediction models.

Impact on Model Sensitivity and Generalizability. The effectiveness of QM descriptors appears to depend on their correlation with the targets, as evident from the above analyses. However, the advantages of QM descriptors may not always be readily apparent, particularly in real-world applications where QM descriptors are often used to predict targets lacking obvious physical correlations or intuitive relationships. Therefore, it is crucial to further analyze and discuss the impact of incorporating QM descriptors on the sensitivity and generalizability of D-MPNN models for molecular property prediction.

To establish a baseline for further comparisons, we conducted a control experiment to investigate whether including the actual target value as a descriptor improves the model's performance in predicting the target itself. Considering that the HOMO/LUMO gap serves both as a molecular descriptor and a target in the QM9 data set, we posed the following question: Does the inclusion of the true gap value in the model decrease the error in predicting the gap itself and its associated properties? In our first experiment, we introduced the true HOMO/LUMO gap value as an additional descriptor for the model trained on the full training set of the QM9 data set. The results showed a considerable reduction in the prediction error of the gap itself (Table S6), confirming that including correlated (i.e., in this case, perfectly correlated) QM descriptors can enhance model accuracy. In the second experiment, we used the true HOMO/LUMO gap value to train a model for predicting other targets in the QM9 data set, which include multiple energetic properties (e.g., enthalpy, Gibbs free energy, etc.) closely related to the HOMO/LUMO gap. The results confirmed that the QM-GNN approach improves accuracy in predicting related properties, further supporting the observation that augmenting D-MPNN with target-correlated QM descriptors can enhance model performance (Table S6).

However, computing QM descriptors for each new species in real-life applications can be costly; therefore, we replicated the analysis using ML-predicted HOMO/LUMO gaps. The resulting mL-QM-GNN model substantially underperformed compared to both the QM-GNN model and the baseline GNN model using standard D-MPNN (Table S6). This raises a critical question: How does uncertainty in QM descriptor values affect model performance? To address this, we conducted an error-sensitivity analysis for the QM-GNN models by introducing varying levels of noise to the true HOMO/LUMO gap descriptor values. The results suggest that, in this case, perturbing the descriptor values by only approximately 4% from their true values could severely impair the model performance (Table S7). Furthermore, results from models trained on downsampled sets indicate that, compared to using exact values of QM descriptors, increased uncertainties in QM descriptors cause the benefit of using additional descriptors to diminish more rapidly as the number of training data points increases (Figure 10).

Consequently, errors in on-the-fly QM descriptor prediction can hinder the performance of ml-QM-GNN models, especially for tasks with small training sets. This observation could explain why the ml-QM-GNN models underperformed for the ESOL and FreeSolv data sets (Figure 11). To further investigate, we computed QM descriptors for these two data sets using the ωB97XD/def2-SVP model chemistry and the same procedure as discussed in the Methods section. Subsequently, we assessed the performance of the corresponding D-MPNN models using these calculated QM descriptors. The results show that the QM-GNN model consistently outperforms ml-QM-GNN models in all random split downsampled scenarios and can substantially outperform the baseline GNN models when trained with fewer than 200 data points (Figure 11). Additional results and analyses of the errors associated with predicted QM descriptors and the distribution of molecular motifs in the ESOL and FreeSolv data sets compared to the QM descriptors data set can be found in Section S11.

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

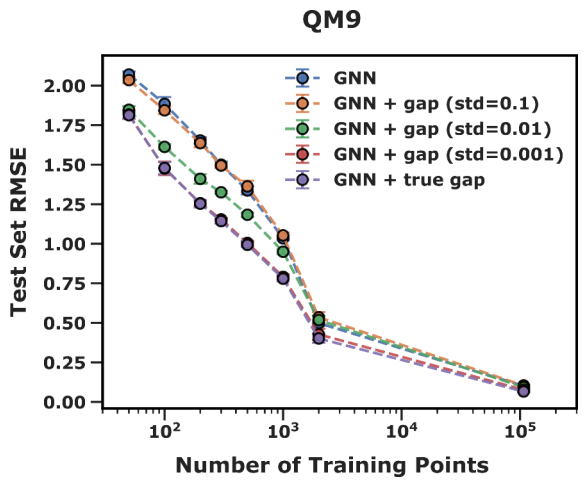


Figure 10. Test RMSE for models trained on downsampled subsets of the QM9 data set with the HOMO/LUMO gap as an additional feature. The GNN model, trained without the additional gap feature, serves as a baseline. The "GNN + true gap" model, trained with the true gap value, represents the best-case scenario for comparison. The remaining models are trained with a noisy gap feature where different levels of Gaussian noise are introduced by varying the standard deviation of the added noise. Increasing noise levels leads to a more rapid diminishing of the benefit provided by the additional descriptor as the training set size increases.

To further assess the generalizability of the QM-GNN model compared to the baseline GNN model, we evaluated test performance of models trained on both random split and scaffold split subsets containing 100 data points each sampled from ESOL and FreeSolv, respectively. The scaffold split provides a more rigorous assessment of a model's generalizability compared to the random split because it often results in a greater structural divergence between the training and test-set molecules.[77,78] More specifically, in the scaffold split experiment, we trained the models using only molecules without any Bemis-Murcko scaffolds and then tested them on molecules containing scaffolds. This is a challenging test because molecules without scaffolds have no rings in their backbones, essentially requiring the models to learn from nonring species to predict the properties of ring-containing species. Rings can have a significant effect on solvation energies, making this a demanding extrapolation task. The results demonstrate that incorporating QM descriptors leads to improved performance for both random and scaffold splits (Figure 12). Notably, the addition of QM descriptors results in a substantially greater performance gain in the more challenging scaffold split experiment, suggesting that QM descriptors can significantly enhance the extrapolative capability and generalizability of D-MPNN models for molecular property predictions.

In summary, our findings demonstrate that incorporating QM descriptors into D-MPNN models can substantially improve their performance for molecular property prediction, particularly when dealing with small data sets. However, it is crucial to consider the trade-off between the potential benefits and the computational cost and uncertainty associated with obtaining these descriptors. For small data sets with easily computable QM descriptors, we recommend directly calculating the descriptors using quantum mechanics for each inquiry molecule during inference. This approach minimizes the likelihood of model underperformance due to errors in descriptor values introduced by on-the-fly prediction. In contrast, when dealing with data sets for which QM descriptors are difficult to compute or estimate accurately, caution should be exercised when applying them within GNN models. In such cases, thorough pilot benchmark studies are advised to better understand the correlation and uncertainty of selected QM descriptors on target prediction model performance before embarking on costly QM calculations for a large number of molecules or reactions. Moreover, descriptor selection can have a decisive impact on model performance, with the optimal subset often varying depending on the specific task, underscoring the importance of a well-planned selection strategy. Judiciously choosing the most informative descriptors not only improves model accuracy but also enhances interpretability by potentially revealing underlying physicochemical principles governing descriptor-target relationships while reducing computational costs by avoiding unnecessary calculations. Adopting this targeted approach can facilitate the development of more accurate, robust, and interpretable models for

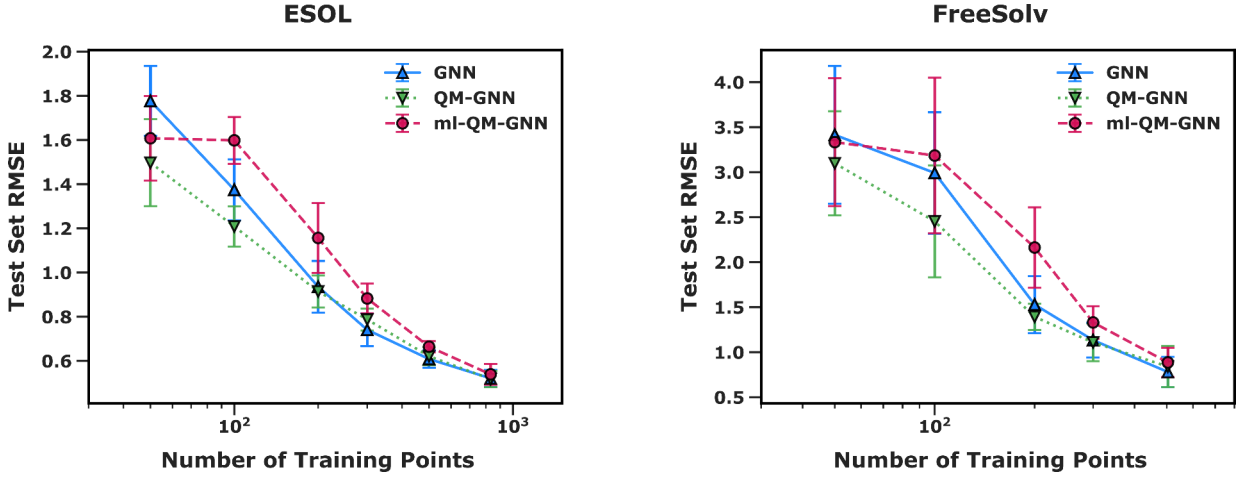


Figure 11. Test RMSE for models trained on downsampled subsets of random split training sets from the ESOL and FreeSolv physical chemistry data sets. The QM-GNN model uses directly calculated QM descriptors, while the ml-QM-GNN model employs ML-predicted QM descriptors. The error bars represent the standard deviation of the RMSE values across 10 folds. Substantial difference in performance of the models, particularly when trained on smaller data sets, highlights the importance of using accurate QM descriptors for improved model performance.

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

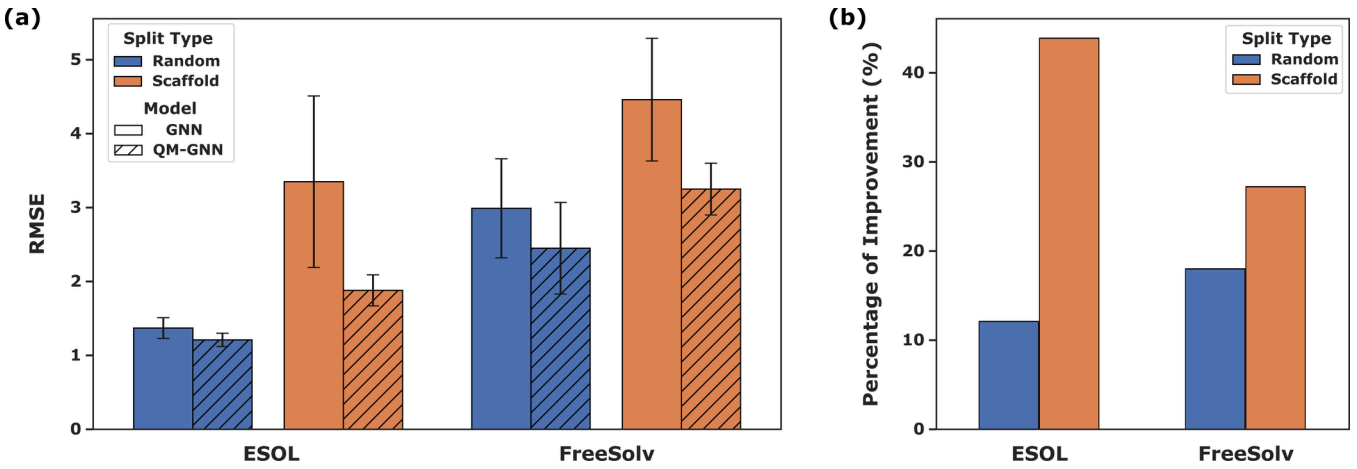


Figure 12. Extrapolation performance comparison between GNN and QM-GNN models trained on 100 data points. (a) Test RMSE on the ESOL and FreeSolv data sets for GNN and QM-GNN models with error bars representing the standard deviation from 10-fold cross-validation. (b) Percentage improvement in test RMSE for the ESOL and FreeSolv data sets resulting from the transition from GNN to QM-GNN models. The substantial performance gain achieved by QM-GNN models, particularly in the more challenging scaffold split scenario, highlights its superior extrapolative capability and generalizability compared to the baseline GNN models.

molecular property prediction. Our analysis of model generalizability using scaffold splitting highlights the potential of QM descriptors to enhance the extrapolative capability of D-MPNN models. By incorporating QM features, the models can potentially better capture the underlying physical principles governing molecular properties, enabling them to make more accurate predictions on molecules with distinct structural scaffolds that are not present in the training set. This improved extrapolative performance underscores the value of QM descriptors in enhancing the generalizability and transferability of D-MPNN models across diverse chemical spaces. Overall, the judicious use of QM descriptors in D-MPNN models can lead to significant improvements in molecular property prediction, especially for small data sets and when extrapolating to novel chemical structures. However, careful consideration of computational cost, uncertainty associated with obtaining these descriptors, and the strategic selection of relevant descriptors is necessary to ensure optimal model performance, efficiency, and interpretability.

CONCLUSIONS

This study highlights the value of integrating quantum mechanical (QM) descriptors into directed message passing neural networks (D-MPNNs) to enhance the accuracy of molecular property predictions, particularly when training data is limited. By incorporating physics-based information, this approach enables data-efficient, interpretable machine learning for chemistry. Our findings suggest that QM descriptors are most beneficial when: (1) the training set has fewer than 2000 data points, (2) the selected descriptors correlate well with the target properties, and (3) the descriptors can be accurately computed at a reasonable cost.

In light of these demands, we recommend a comprehensive decision process to achieve optimal use of QM descriptors. More specifically, given the usually high computational cost associated with generating descriptors through quantum chemistry calculations, and the significant advantage of a larger training set, we recommend prioritizing the acquisition of new training data whenever possible. However, if obtaining additional training data is not feasible or comparatively more expensive, employing QM descriptors can be a suitable alternative. To reduce computational costs, ML-estimated QM descriptors can sometimes be a practical substitute for direct QM calculations, but they may compromise the accuracy of model predictions due to inherent estimation errors. Importantly, the accuracy of QM descriptors is often crucial to model success, as even a modest level of uncertainty in the descriptors can potentially significantly reduce their effectiveness in improving model performance. Moreover, simply using a large number of QM descriptors without careful analysis is not a reliable strategy, as irrelevant descriptors can introduce noise and confuse the model, particularly when training data are scarce. While this study extensively evaluates a set of general density functional theory (DFT) QM descriptors that can be readily computed, there remains untapped potential in unexplored descriptors that could significantly benefit specific predictive tasks. However, the use of domain-specific descriptors often requires specialized expertise, which may affect model accessibility and usability. The interplay of numerous factors underscores the significance of a well-conceived descriptor selection strategy that carefully balances model reliability, robustness, explainability, and computational efficiency. To help the reader navigate these considerations more effectively, we propose a comprehensive decision process for the optimal use of QM descriptors, which is summarized in a user-friendly decision flowchart provided in the manuscript. Looking forward, future research directions include a deeper exploration of the diverse range of QM and non-QM descriptors not covered in this study, particularly those that may offer novel insights into molecular property and chemical reactivity prediction. The choice of computational methods and the impact of molecular conformers on the effectiveness of descriptors also warrant further examination. Additionally, while this work focuses on D-MPNNs, the insights and guidelines provided are likely applicable to other machine learning frameworks for chemistry.

In conclusion, the strategic application of QM descriptors can substantially enhance the predictive performance of D-MPNNs. However, this approach requires careful consideration of training data availability, the nature of each prediction task, and computational constraints. New data acquisition should be prioritized when it is practical and economical.

https://doi.org/10.1021/jacs.4c04670
J. Am. Chem. Soc. 2024, 146, 23103−23120

Nevertheless, QM descriptors can provide crucial predictive insights in scenarios where gathering experimental data is slow and expensive, such as in design-make-test-analyze molecular discovery cycles. They can also be invaluable in navigating the complex landscape of chemistry in research areas often overwhelmed by the vast number of potential reaction pathways or molecules for screening, such as in retrosynthesis pathway exploration and high-throughput virtual screening of drug candidates. Furthermore, in active learning autonomous material design campaigns, the iterative nature of these projects means that even minor enhancements in model accuracy per cycle can culminate in significant improvements over time. Consequently, any marginal improvements offered by QM descriptors, especially in the early phases of molecular discovery projects, can markedly boost the effectiveness of the active learning and molecular synthesis loop. Therefore, in these diverse contexts, leveraging QM descriptors with D-MPNNs has the potential to be a valuable tool for innovative chemical design.

■ ASSOCIATED CONTENT

Data Availability Statement

The curated data set encompassing 37 QM descriptors for 64 921 distinct molecules across six levels of theory (i.e., ωB97XD, B3LYP, M06-2X, PBE0, TPSS, and BP86), along with the D-MPNN models for QM descriptor prediction, the associated data splits used for model training, and scripts are available for free on Zenodo at doi.org/10.5281/zenodo.10668491. The data set and scripts are open access and distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/). Online predictions using the trained QM descriptor models are made available through an easily accessible web tool via the QM Descriptors Utility page within MIT's ASKCOS platform (https://askcos.mit.edu/qm). Furthermore, the refined workflow designed for high-throughput QM descriptors calculations is accessible through GitHub (https://github.com/oscarwumit/QM_descriptors_calculation). The customized Chemprop v1 capable of performing Shapley value analysis is available via https://github.com/oscarwumit/chemprop_developing/tree/shap_v1. Citations of data set, scripts, and software programs should include reference to this article. Additional results and discussions can be found in the Supporting Information document.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c04670.

> Comprehensive analysis of the curation of the QM descriptors data set; results from training machine learning models to predict QM descriptors and molecular properties; additional findings from the sensitivity analysis of QM-GNN, the conformational effects on QM descriptors; and the extrapolative performance of QM-augmented models (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Yi-Pei Li − *Department of Chemical Engineering, National Taiwan University, Taipei 10617, Taiwan;* orcid.org/0000-0002-1314-3276; Email: yipeili@ntu.edu.tw

William H. Green − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-2603-9694; Email: whgreen@mit.edu

Authors

Shih-Cheng Li − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Chemical Engineering, National Taiwan University, Taipei 10617, Taiwan*

Haoyang Wu − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-0644-7554

Angiras Menon − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Kevin A. Spiekermann − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-9484-9253

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c04670

Author Contributions

¶S.-C.L. and H.W. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Y.-P.L. is supported by Taiwan NSTC Young Scholar Fellowship Einstein Program (112-2636-E-002-005) and is grateful to the National Center for High-performance Computing (NCHC) for their support with computing facilities. S.-C.L. gratefully acknowledges a scholarship from Y.L. Lin Hung Tai Education Foundation. H.W. gratefully acknowledges fellowship support from Takeda Pharmaceuticals. Most of the financial support for this project was provided by the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS) at MIT and the DARPA Accelerated Molecular Discovery (AMD) program (DARPA HR00111920025). The authors acknowledge the MIT SuperCloud[83] for providing high-performance computing resources and consultation. We are particularly grateful for input and help from Lauren E. Milechin, Chansup Byun, and Jeremy Kepner of SuperCloud. We thank Prof. Thijs Stuyver and Prof. Kevin P. Greenman for helpful discussions on this work. The authors thank Kariana Moreno Sader for her invaluable expertise and meticulous attention to detail in enhancing the clarity and quality of the figures in this research paper.

■ REFERENCES

(1) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9−17.

(2) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11*, 5753.

(3) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group contribution and machine learning

approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433−446.
(4) Fowles, D. J.; Palmer, D. S.; Guo, R.; Price, S. L.; Mitchell, J. B. Toward Physics-Based Solubility Computation for Pharmaceuticals to Rival Informatics. *J. Chem. Theory Comput.* **2021**, *17*, 3700−3709.
(5) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144*, 10785−10797.
(6) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58*, 4515−4519.
(7) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regioselectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **2021**, *12*, 2198−2208.
(8) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting regioselectivity in radical C-H functionalization of heterocycles through machine learning. *Angew. Chem., Int. Ed.* **2020**, *59*, 13253−13259.
(9) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Acc. Chem. Res.* **2016**, *49*, 1292−1301.
(10) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep learning of activation energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992−2997.
(11) Heid, E.; Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *J. Chem. Inf. Model.* **2022**, *62*, 2101−2110.
(12) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space. *J. Chem. Phys.* **2021**, *155*, 064105.
(13) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **2021**, *12*, 1163−1175.
(14) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast predictions of reaction barrier heights: toward coupled-cluster accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976−3986.
(15) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **2023**, *382*, No. eadi1407.
(16) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, No. e1608.
(17) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(18) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495−6502.
(19) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(20) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(21) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(22) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask prediction of site selectivity in aromatic C−H functionalization reactions. *React. Chem. Eng.* **2020**, *5*, 896−902.
(23) Duan, Y.-J.; Fu, L.; Zhang, X.-C.; Long, T.-Z.; He, Y.-H.; Liu, Z.-Q.; Lu, A.-P.; Deng, Y.-F.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. Improved GNNs for Log D 7.4 Prediction by Transferring Knowledge from Low-Fidelity Data. *J. Chem. Inf. Model.* **2023**, *63*, 2345−2359.
(24) Guo, J.; Sun, M.; Zhao, X.; Shi, C.; Su, H.; Guo, Y.; Pu, X. General Graph Neural Network-Based Model To Accurately Predict Cocrystal Density and Insight from Data Quality and Feature Representation. *J. Chem. Inf. Model.* **2023**, *63*, 1143−1156.
(25) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(26) Magar, R.; Wang, Y.; Lorsung, C.; Liang, C.; Ramasubramanian, H.; Li, P.; Farimani, A. B. AugLiChem: data augmentation library of chemical structures for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045015.
(27) Aldeghi, M.; Coley, C. W. A graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **2022**, *13*, 10486−10498.
(28) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360*, 186−190.
(29) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1339−1345.
(30) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363*, No. eaau5631.
(31) Hao, Y.; Hung, Y. C.; Shimoyama, Y. Investigating Spatial Charge Descriptors for Prediction of Cocrystal Formation Using Machine Learning Algorithms. *Cryst. Growth Des.* **2022**, *22*, 6608−6615.
(32) Tayyebi, A.; Alshami, A. S.; Rabiei, Z.; Yu, X.; Ismail, N.; Talukder, M. J.; Power, J. Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *J. Cheminform.* **2023**, *15*, 99.
(33) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *J. Chem. Phys.* **2022**, *156*, 084104.
(34) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning. *J. Chem. Inf. Model.* **2023**, *63*, 4574−4588.
(35) Abarbanel, O. D.; Hutchison, G. R. QupKake: Integrating Machine Learning and Quantum Chemistry for micro-pKa Predictions. *J. Chem. Theory Comput.* **2024**, DOI: 10.1021/acs.jctc.4c00328.
(36) Shimakawa, H.; Kumada, A.; Sato, M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *npj Comput. Mater.* **2024**, *10*, 11.
(37) Von Lilienfeld, O. A. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676−1689.
(38) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chem. - Eur. J.* **2023**, *29*, No. e202300387.
(39) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. S_NAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *J. Chem. Inf. Model.* **2023**, *63*, 3751−3760.
(40) Bacciu, D.; Errica, F.; Micheli, A.; Podda, M. A Gentle Introduction to Deep Learning for Graphs. *Neural Netw* **2020**, *129*, 203−221.
(41) Chemprop: Message Passing Neural Networks for Molecule Property Prediction. https://www.github.com/chemprop/chemprop (accessed 2024-6-29).
(42) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is All you Need. *Advances in Neural Information Processing Systems* **2017**, 5998−6008.

(43) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(44) Tox21 Challenge. https://tripod.nih.gov/tox21/challenge/ (accessed 2024-6-29).
(45) Dionisio, K. L.; Phillips, K.; Price, P. S.; Grulke, C. M.; Williams, A.; Biryol, D.; Hong, T.; Isaacs, K. K. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Sci. Data* **2018**, *5*, 180125.
(46) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(47) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065−6073.
(48) Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.
(49) *QM descriptors calculation*. https://github.com/yanfeiguan/QM_descriptors_calculation (accessed 2024-6-29).
(50) Domingo, L. R.; Pérez, P.; Sáez, J. A. Understanding the local reactivity in polar organic reactions through electrophilic and nucleophilic Parr functions. *RSC Adv.* **2013**, *3*, 1486−1494.
(51) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(52) RDKit: Open-Source Cheminformatics. https://www.rdkit.org (accessed 2024-6-29).
(53) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(54) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom−atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.
(55) Becke, A. D. A new mixing of Hartree−Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.
(56) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.
(57) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(58) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta−Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
(59) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.
(60) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822−8824.
(61) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
(62) Frisch, M. J. et al. *Gaussian 16 Revision C.02.*; Gaussian Inc: Wallingford CT, 2016.
(63) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Karafiloglou, P.; Landis, C. R.; Weinhold, F. NBO 7.0.; *Theoretical Chemistry Institute*; University of Wisconsin: Madison, 2018.
(64) Liu, Y.; Li, Z. Predict Ionization Energy of Molecules Using Conventional and Graph-Based Machine Learning Models. *J. Chem. Inf. Model.* **2023**, *63*, 806−814.
(65) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(66) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.
(67) 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.
(68) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(69) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711−720.
(70) Wenlock, M.; Tomkinson, N. Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds. **2015**. DOI: 10.6019/CHEMBL3301361 (accessed 2024-6-29).
(71) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.
(72) AIDS Antiviral Screen Data. https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data (accessed 2024-6-29).
(73) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936−1949.
(74) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686−1697.
(75) Bergstra, J.; Yamins, D.; Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *Proceedings of the 30th International Conference on Machine Learning*. 2013; pp 115−123.
(76) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(77) Greenman, K. P.; Green, W. H.; Gomez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13*, 1152−1162.
(78) Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H. Comment on 'Physics-based representations for machine learning properties of chemical reactions. *Mach. Learn.: Sci. Technol.* **2023**, *4*, 048001.
(79) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv*, November 25, 2017, 1705.07874, ver. 2. DOI: 10.48550/arXiv.1705.07874.
(80) Meyes, R.; Lu, M.; de Puiseau, C. W.; Meisen, T. Ablation Studies in Artificial Neural Networks. *arXiv*, February 18, 2019, 1901.08644, ver. 2. DOI: 10.48550/arXiv.1901.08644.
(81) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv*, August 9 2016, 1602.04938, ver. 3. DOI: 10.48550/arXiv.1602.04938.
(82) Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K. E.; Dickerson, J. P.; Shah, C. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv*, November 25, 2022, 2010.10596, ver. 3. DOI: 10.48550/arXiv.2010.10596.
(83) Reuther, A. et al. Interactive supercomputing on 40,000 cores for machine learning and data analysis. *2018 IEEE High Performance extreme Computing Conference (HPEC)*. 2018; pp 1−6.