JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

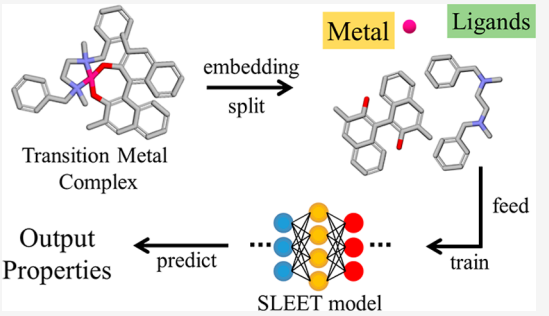
Feature-Driven Prediction of HOMO−LUMO Gaps in Transition-Metal Complexes Using the SLEET Model: A SMILES-Based Transformer Framework

Sheng-Hsuan Hung, Zong-Rong Ye, Chi-Feng Cheng, An-Cheng Yang,* Berlin Chen,* and Ming-Kang Tsai*

Cite This: J. Chem. Theory Comput. 2025, 21, 6410−6420 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A feature-driven model, SLEET, built upon the early reported SchNet-bs-RAN framework, that combines the approaches of SchNet and the bond-step representation weighted by the reduced atom number, is reported for evaluating the molecular electronic structure properties of transition-metal complexes (TMCs). Ligands were derived by segmenting purely two-dimensional SMILES representations, and metal−ligand interactions were modeled by using a Transformer-like architecture to construct a property prediction framework that aligns closely with chemical knowledge. This approach effectively captures the characteristics of the ligand field within TMCs. Consequently, the SLEET model delivers precise HOMO−LUMO gap predictions comparable to those achieved by three-dimensional information-based models while also demonstrating strong performance in predicting the molecular-weight-independent electronic properties.



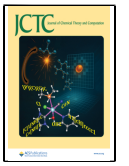
1. INTRODUCTION

Organometallic complexes play a critical role in modern chemical technology development, particularly in the fields of organic compound synthesis, porous metal−organic framework materials, pharmaceutical bioinorganic reagents, optoelectronic materials, and so forth. The fascinating chemical properties of organometallic complexes generally result from the interplay between the cationic metal center and its organic ligands. The metal−ligand interactions generally consisted of (A) electrostatic interactions between metal cations and ligand local negative charge density, (B) intramolecular bonding interactions of ligands, (C) the interligand nonbonding interactions, and (D) the electronic interactions between the d-orbitals of metals and high-lying occupied orbitals and low-lying virtual orbitals of ligands.[1−3] Due to the complex nature of the electronic structures of organometallic complexes, higher-level density functional theory (DFT) calculations often become computationally impractical, necessitating the use of molecular mechanics (MM)[4−6] or semiempirical quantum mechanical (SQM)[7] calculations for these systems at the nanoscale. Although some promising models can capture the relevant properties of organometallic complex ligands, such as terms A−C, these models have not yet achieved widespread application and exhibit various limitations in accuracy. These limitations arise due to the self-interaction error (SIE), leading to slight inaccuracies in the description of properties such as dissociation energies, band gaps, electron affinities, and barrier heights, making the description of term D challenging.[1−3,8] Nonetheless, efforts to customize functionals have improved the prediction of optical properties such as the HOMO−LUMO gap (ΔE_{HL}).[9,10]

Virtual high-throughput screening (VHTS) is a direct application of computational chemistry methods.[11−15] In typical applications such as small-molecule discovery, DFT can serve as a cost-effective method for initial screening. By employing DFT methods, reasonable accuracy predictions can be provided for organic molecules, with a lower consumption of resource and time compared to actual drug screening experimental processes.[11−15] This can yield desired molecular structures and property prediction data, significantly reducing the number of compounds that experimentalists need to test and thus lowering actual expenses. However, time and computational costs can be further saved through the use of machine learning (ML) methods, and ML-accelerated VHTS methods have gained considerable attention in recent years.[8,14,16−45]

For the characterizations of organic molecules, DFT calculations in describing structure−property relationships

Received: January 16, 2025
Revised: May 28, 2025
Accepted: June 16, 2025
Published: June 23, 2025



https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

have inspired the development of models using ML linear or nonlinear methods, which have progressed rapidly in recent years.[8] Gilmer et al. first proposed the interpretable model message passing neural network (MPNN) for constructing molecular structure−property relationships. This model fits the structure−property relationships obtained from DFT calculations using neural networks (NNs), establishing prediction models with DFT-level accuracy but with significantly lower computational costs.[46] Schütt et al. developed the MPNN model SchNet, which predicts molecular properties based on three-dimensional structures. Its predictive capability for small organic molecules has been validated, achieving an accuracy within 1 kcal/mol compared to DFT calculations.[47−51] Bjørn Jørgensen et al. developed an EdgeUpdate model for SchNet, introducing receiving atom information to improve predictive precision across all tested data sets compared to SchNet.[52] Gasteiger et al. researched the DimeNet and DimeNet++ MPNN models, enhancing the utilization of three-dimensional molecular structure information by incorporating bond angle descriptions to extract more molecular information, thus improving the NN model's predictive capabilities.[53,54] Choudhary et al. developed ALIGNN, which also uses bond angle descriptions and a more rational model architecture to better describe the structure−property relationships of molecules, achieving predictive capabilities for molecular properties that closely match DFT calculations.[55]

The accuracy of ML models in predicting the properties of organic molecules has improved annually.[46−55] However, these models have not yet reached an acceptable level for predicting the structure−property relationships of organometallic complexes.[56] ML models constructed using DFT data also struggle to provide effective prediction results due to the inherent source errors. Nevertheless, in the field of transition-metal complexes (TMCs), the synthesized compounds only occupy a small segment of the total chemical space, indicating a significant potential for discovery.[8] Duan et al., investigating the impact of density functional approximation (DFA) bias, used 23 DFAs to construct a data set of over 2k TMCs, which was employed to train various artificial neural networks (ANNs) to explore a large chemical space of over 187k TMCs.[57] They subsequently designed approximately 32.5 million TMC combinations.[58] Balcells and Skjelstad utilized DFT methods to perform extensive calculations on various properties of TMCs with 3d to 5d center metals from the Cambridge Structural Database (CSD), constructing the tmQM data set containing approximately 87k data points.[59] This data set appears to provide the opportunity to construct DFT-level accuracy models using ML methods for TMCs. Kneiding et al. further refined tmQM by filtering the center metals included and calculating additional quantum properties to create NatQG, resulting in the powerful data set tmQMg with around 60k data points. Utilizing the tmQMg data set with NatQG graph, several enn-s2s-based models were developed with good descriptive power for organic molecule data sets and the electronic properties of TMCs.[56] In 2023, Garrison et al. identified computational errors in the energy properties of tmQMg and reconstructed another organometallic complex data set, tmQM_wB97MV, containing only energy properties that have been corrected.[60]

Four early models, including SchNet, EdgeUpdate, DimeNet ++, and ALIGNN, tested by Balcells and co-workers[56] appeared to show inconsistent prediction performance in between QM9 and tmQMg data sets, and this indicates the missing link of the coordination bonding description. The atoms of the first coordination sphere ligands usually provide the most substantial influence to the d-orbital electronic structure of TMCs, as being essentially stated by the ligand field theory. Such a directional (or could be symmetric) effect toward the central metal ions appears to have a significant difference from that of organic molecules. Consequently, the effective feature description methods for organic molecules do not apply to inorganic molecules.[61] Among the three additional models showcased previously, those models utilizing the extra calculated NatQG data exhibited a significantly better performance in predicting the structure−property relationship of TMCs.[56] However, this improvement comes with the increased data preparation effort for collecting natural bond analysis. This consequently hinders the efficiency of conducting VHTS application to explore the new TMCs—those not being included in the data set.

Although we are encouraged to see that ML models are progressively improving their ability to describe TMCs, their cost-saving advantage to possibly substituting DFT calculations might be offset by the increasing data preparation workload. Using low-dimensional information directly derived from chemical language like SMILES to train and construct an ML model might be a way to reclaim the advantage of ML models in VHTS. Training with low-dimensional information is often related to the use of descriptors, which are generated using chemistry-related libraries, software, or models[11,62−67] to produce a feature set sufficient to describe molecular structures or properties and then using this descriptor set to construct ML models.[41,43,61,68,69] For simple systems (such as organic molecules), SMILES can be used to generate descriptors that adequately describe the features of the organic molecules. SMILES can be constructed through simple organic bonding rules, making it very suitable for the development of a VHTS protocol to sample new TMCs (equivalently as the new texts) and accessible integration with the large language model. In our previous work, we proposed the SchNet-bs-RAN model, which replaced the Cartesian coordinates required by SchNet with the bond steps derived from SMILES where the interatomic bond steps are quasi-physically weighted by the neighboring-group atomic numbers adapting the format of reduced mass.[70] Such a model design does not require the handcrafted feature selection with the advantage of easy data set preparation, allowing the model to be trained and make reasonably accurate predictions using the low-dimensional information.

In this study, we focus on describing the relationship between the structures of TMCs and the corresponding HOMO−LUMO gaps. Vertical photophysical adsorption is a common spectroscopic measurement applied to characterize the electronic structure of TMCs for understanding their optical properties. Herein, we reported the development of SchNet-based ligand-embedding extending transformer (SLEET) model, which combines the characteristics of the SchNet-bs-RAN model and Transformer architecture, being inspired by the success of using Transformer to describe protein−ligand interactions.[71] A very recent study also employed the attention mechanisms to incorporate the interactions between metal centers and ligands into GNN architectures, where the description of ligands was built upon the representation learner scheme versatilely utilizing several types of atom-level inputs.[72] A quasi-analogous scenario is attempted in this study to describe the interplay between the

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

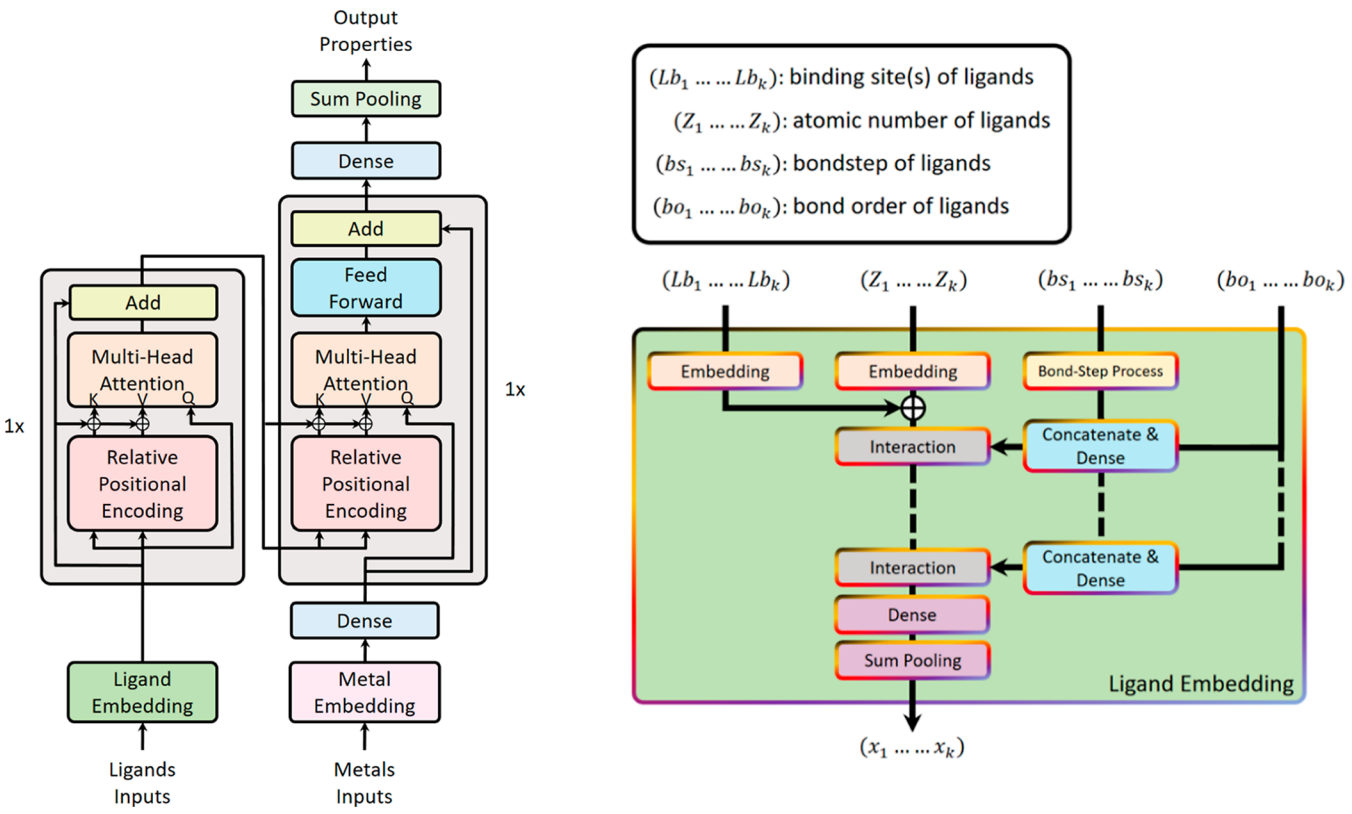


Figure 1. Architectures of SLEET (left) and ligand-embedding layer (right).

metal−ligand interactions and interligand interaction through the use of attention mechanism. SLEET is specifically designed for predicting the properties of transition-metal complexes, handling metal and ligand information separately and retaining the feature of using SMILES as the sole input for low-dimensional training and prediction from the SchNet-bs-RAN model. For modeling interactions between the central metal and ligands, we utilize a Transformer-like architecture and employ relative position encoding (RPE)[73] to differentiate the spatial positions occupied by ligands. We selected the tmQMg data set for model testing because it offers sufficient advanced model evaluation data, enabling direct comparison of different model performances.[56]

2. METHODS

SchNet model utilizes the interatomic distances and atom types as the fundamental molecular representations.[47] However, such a model framework is not sensitive to the complex structure symmetry and does not clearly label the atoms of individual fragments (or the said ligand). For the interatomic distances are initially equivalent and could be distinguished into ligand-based contribution through the learning process, it is anticipated to be very challenging. The chemical properties of TMCs are consensually and significantly associated with the field strength and spatial position of the ligands, as being well expounded by the ligand field theory of the inorganic chemistry community. Consequently, we introduce the SchNet-based ligand-embedding extending transformer (SLEET) model, employing an attention mechanism to model the interactions of metal−ligands and interligands. SLEET separates all ligand-related information and incorporates it through a ligand-embedding layer, enabling the model to learn the relative attention weights between each ligand and the central metal atom from the data. Ultimately, the central metal determines the overall energy performance of the TMCs, analogous to the concept of ligand field theory.

Although SLEET was designed with SchNet-bs-RAN as its ligand-embedding method,[70] the model was conceptually proposed as a framework for ligand−metal interactions. It can utilize any type of embedding method, including neural message passing models or models that introduce ligand-related information, as the ligand-embedding mechanism for SLEET.

2.1. Using Open Babel[63] and RDKit[62] to Handle SMILES Data. In order to prepare the data set using the low-dimensional information, we utilized the automatic conversion feature of Open Babel to convert the Cartesian coordinate of transition-metal complexes into SMILES, which contains the metal−ligand coordination and ligand-based valence bond information. If the SMILES generated by Open Babel cannot satisfy the valence bond rule of RDKit, manual modification on SMILES would be conducted through the procedures of adding/subtracting charges and hydrogen. The obtained SMILES were subsequently processed by RDKit to obtain the ligand-based bond step and bond order information based on the underwritten valence bond rules These manual modifications for facilitating bond step and bond order generation may introduce potential errors to the predictive properties of the model; however, the percentage of these modified data points is very low. Table S1 provides five examples of TMCs for the information generated by Open Babel and RDKit.

2.2. SLEET Architecture. We attempt to organize and simplify the functions of the SLEET architecture through a straightforward description. SLEET uses node features x_m for a central metal atom and x_{li}, where i denotes the ith atom within

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

an individual ligand, and the edge features e_{lij} corresponding to i and j ($i \neq j$) atom pairs within an individual ligand. Each TMC consists of a central metal assembly M and a ligand assembly L, which are composed of the central metal atom m and ligands l formed by SMILES fragments, respectively. Each ligand l is embedded into a hidden state h_l, representing ligand nodes, through T time steps of the ligand-embedding block E_L. Similarly, the central metal atom m is embedded into a hidden state h_m via the metal-embedding block E_M.

In the node features x_{li}, in addition to the atomic type, we introduced binding site information recognized by RDKit, which identifies the atoms bonded to the metal. This information is presented in a one-hot encoding format, incorporated using the same embedding method as that for atomic types, and then added to the embedding vector of the atomic types.

$$h_m = E_M(x_m) \tag{1}$$

$$h_{l_i}^{t+1} = E_L(h_{l_i}^t, h_{l_j}^t, e_{l_{ij}}) \tag{2}$$

As shown in eq 2, the ligand hidden state $h_{l_i}^t$ at step t is computed through E_L, resulting in $h_{l_i}^{t+1}$ over a total of T time steps, followed by

$$h_m = \mathrm{MLI}(\{h_m | m \in M\}, \{h_{l_i}^T | l \in L\}) \tag{3}$$

The hidden state h_m of m interacts with the ligand-hidden state $h_{l_i}^T$ through the MLI (metal−ligand interaction), resulting in the updated h_m. The MLI function is responsible for managing how the ligand field is simulated within the TMCs. We posit that the hidden state of the central metal can ultimately represent how the ligand field influences chemical properties. Therefore, the hidden state of the central metal is the primary focus.

$$\hat{y} = P(\{h_m | m \in M\}) \tag{4}$$

Finally, the property prediction value is obtained from h_m through a pooling function P, completing the forward process of the SLEET model. Although the above formulation defines M as the ensemble of central metals, implying that the SLEET architecture is capable of predicting the properties of multicentral metal systems, this study implements SLEET only for property prediction in single-metal-center systems. Accordingly, subsequent discussions will focus primarily on single-metal-center systems. Details for the ligand-embedding block and the corresponding hidden state are provided in the Supporting Information.

2.3. Metal-Embedding Block. For describing the central metal atoms, we use embedding in a manner similar to that for atom labels, as shown on the right in Figure 1. The imbalanced metal element distribution in the tmQMg data set across the periodic table, as shown in Figure 2, may result in poor generalization of the model. To mitigate this, the metal-embedding layer not only uses standard atom label embeddings but also incorporates group and period information for all metal atoms, embedding these features in a similar label-based manner.

The parameter that governs the dimensionality of the metal-embedding features in the model is referred to as n_metal_basis. The embeddings for metal atoms, groups, and periods are all assigned an n_metal_basis of 256, consistent with the n_atom_basis. Subsequently, the embedding vectors for the metal atoms, groups, and periods are concatenated to form a vector with a dimensionality of $3 \times n$_metal_basis. This vector is then passed through a dense layer to reduce its dimensionality to n_embed_out, aligning it with the node representation of ligand-embedding.

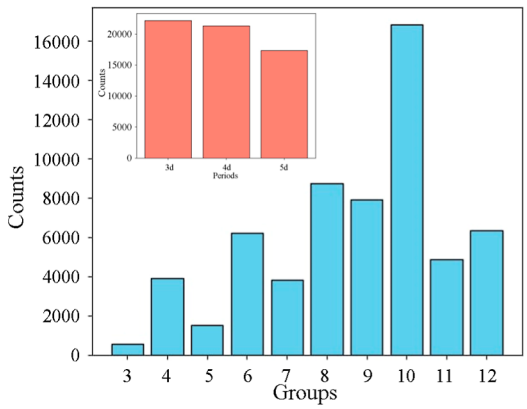


Figure 2. Distribution of the central metal elements across the tmQMg data set, categorized by groups. The inset shows the categorization by period.

2.4. Relative Positional Encoding. In this model, we adopt the relative positional encoding (RPE) technique introduced by Shaw et al.,[73] employing it before the attention computation for encoding the intraligand position and the intracomplex (as the interligand) position, as shown in Figure 1. The attention mechanism serves to capture the similarity between two different entities by utilizing the query (Q), key (K), and value (V) vectors. The rightmost arrow in the multihead attention module corresponds to Q, while the other two arrows represent K and V. Once the resulting attention matrix is multiplied back to the original V vector, the updated vector essentially represents the influence exerted by the Q vector on it. The detailed mathematical formulation for these vectors has been well demonstrated by Vaswani et al.[74]

RPE assigns distinct offsets to ligands at each relative sequence position based on their respective locations. This enables the differentiation of complexes that share the same composition but have different stereogeometric arrangements if the isomeric difference has been priorly encrypted in the input strings. The same beneficial effect is also anticipated for employing RPE in the ligand interaction block for constructing the representative coordination field. Table S2 provides the comparison of using positional encoding (as the absolute sequential order of ligands being predetermined by SMILES) and relative position encoding, and the RPE approach gives a minorly better performance than the PE case. Physically speaking, the ligands are spatially distributed entities surrounding the central metal in a three-dimensional space, and this distribution may be able to project to lower-dimensional expression if some symmetry constraints are applied. Despite our intentions to mimic the spatial distribution of ligands through bookkeeping the relative sequence order between each ligand pair, as implemented in RPE, the current chemical expression—the low-dimensional SMILES sequence—is still the bottleneck factor unable to distinguish stereoisomers like cis vs trans or mer vs fac complexes. Therefore, the prediction performance using PE or RPE was not substantially differentiated. More development in the geometric expression would be critical if one would like to

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

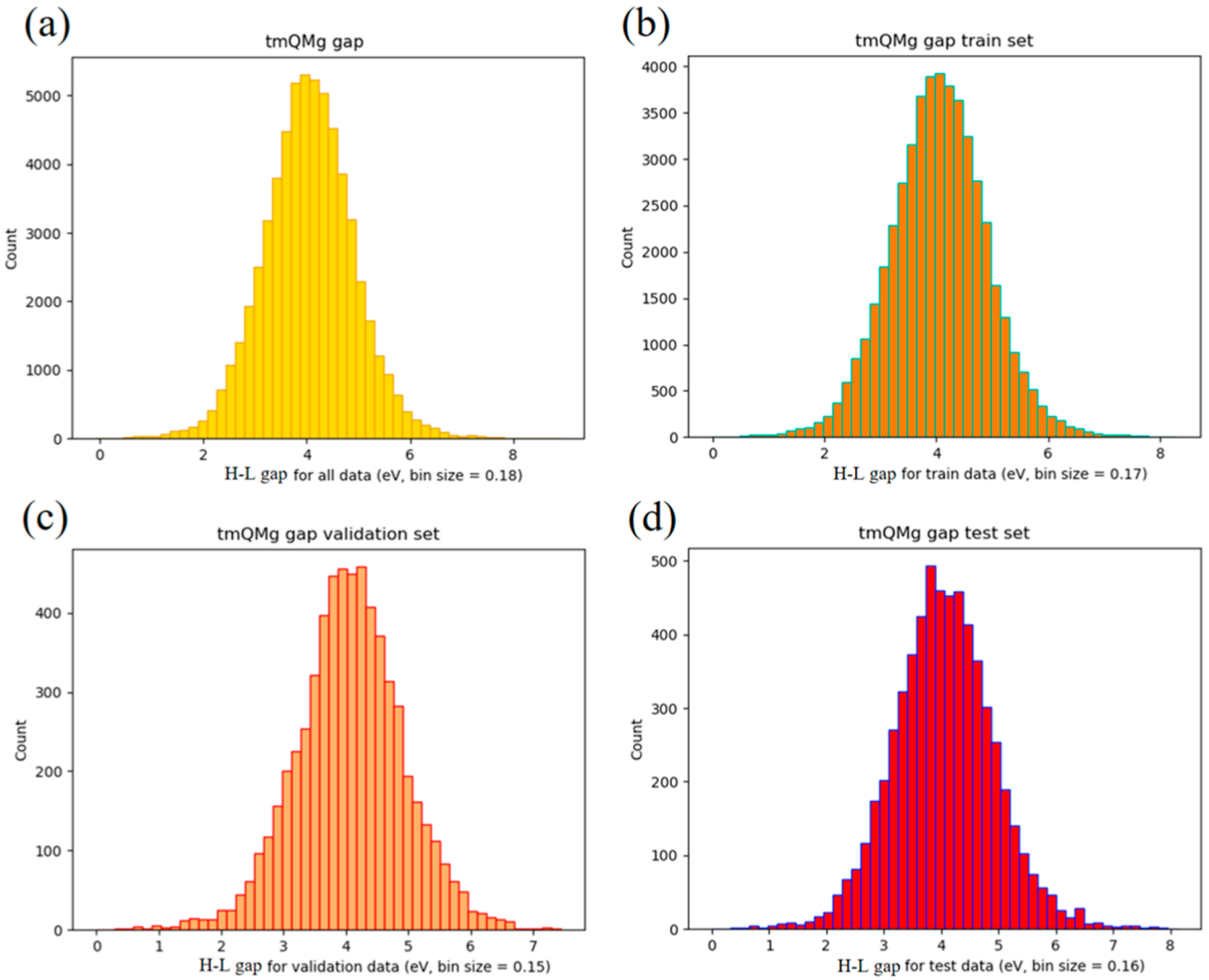


Figure 3. Data distribution plot of the tmQMg data set. (a) Display of the overall H−L gap distribution. (b−d) Distributions of the training, validation, and test sets, respectively. It can be observed that the data sampled randomly exhibit nearly identical distributions among the three different subsets.

utilize the language expression model for the TMC discovery.[75]

2.5. Metal−Ligand Interaction (MLI) Approach. We calculate the attention between the ligand-embedding vector and the metal-embedding vector using an architecture similar to Transformer. The resulting attention-modulated vector is then added back to the metal-embedding vector. This process closely resembles the Transformer mechanism,[74] except that self-attention for the metal-embedding vector is omitted. MLI is precisely the key component in SLEET that enables implementation of the ligand field concept.

Furthermore, due to the low complexity of the information and the limited availability of data, the computation block for MLI is executed only once, unlike the Transformer architecture, which typically repeats the operation multiple times. It has been demonstrated by Shoghi et al. that attempting to train an overly parametrized model under conditions of insufficient data can result in diminished predictive performance compared to a model of the same architecture with fewer parameters.[76] The self-attention mechanism assigns weights to the relative positional relationships among ligands, where Q, K, and V are obtained by mapping the ligand embedding through the independent linear layers.

First, the encoding matrix of relative positional relationships between ligands is computed using their embedding vectors. This encoding matrix is an able-to-learn matrix indexed by relative positions. K and V each correspond to distinct encoding matrices. After the encoding matrix is added to K, it is then multiplied by Q, yielding the attention weight matrix that characterizes ligand−ligand interaction forces. At this stage, K has incorporated the positional offsets from the relative positional encoding, meaning that when K and Q are used to compute attention weights, the model is already informed of the relative positional relationships between ligands. This weight matrix is a 2D matrix where the sum of the values in each column equals 1, indicating the relative influence of each ligand on the others. For example, if a TMC contains six ligands, the attention weight matrix would be a 6×6 matrix, where the total sum of the values is 6.

Next, the attention weights are multiplied by V, resulting in a final feature matrix that represents the environmental influence

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

Table 1. Definition of the Graph for the NatQG Series, SchNet, and SLEET Models

graph	description
base (enn-s2s)	being composed of nodes and edges; the nodes are characterized by atomic number, valence state, covalent radius, and Pauling electronegativity; the edges are defined by bond order and bond distance
NatQG	using natural bond orbital (NBO) data to predict quantum properties, incorporating key information such as natural charges, bond orders, donor−acceptor interactions, potential energies, and more
u-NatQG	building on top of NatQG, using undirected edges represent symmetric relationships between nodes, like conventional molecular graphs
CRG (SchNet)	representing molecular interactions using atom-centered graphs within a specific radius
CBSG (SLEET)	similar to CRG, however, using bond steps instead of radius for graph construction; for ligand nodes, atomic number and metal-bonding annotation are used; for metal nodes, the group and period information for the metal elements are included; the edges are characterized by bond order and bond step; all information can be extracted from parsing SMILES

Table 2. Model Comparison of NatQG Series, SchNet, and SLEET in the Various Electronic Structure Properties[c]

properties	ΔE_{HL}	E_H	E_L	polarizability	dipole moment
[b]models			[a]MAE (testing set)		
NatQG(base)	0.227	0.356	0.354	5.870	1.710
u-NatQG	**0.164**	**0.087**	**0.096**	**4.940**	0.895
d-NatQG	0.196	0.103	0.110	4.960	0.981
[d]SchNet	0.248 ± 0.006	0.328 ± 0.005	0.358 ± 0.004	6.137 ± 0.281	**0.826 ± 0.015**
[d]SLEET	0.208 ± 0.002	0.272 ± 0.008	0.276 ± 0.002	8.082 ± 0.117	1.342 ± 0.008
models			R^2 (testing set)		
NatQG(base)	0.835	0.734	0.722	0.993	0.537
u-NatQG	**0.910**	**0.991**	**0.988**	**0.995**	0.858
d-NatQG	0.877	0.987	0.984	0.994	0.845
SchNet	0.838	0.817	0.801	0.990	**0.904**
SLEET	*0.860*	*0.793*	*0.803*	*0.982*	*0.702*

[a]The units are eV for energy, D for dipole moment, bohr^3 for polarizability. [b]The results of baseline, u-NatQG, and d-NatQG models are extracted from ref 56. The results of SchNet and SLEET are carried out in this study. [c]The best results are highlighted in bold among the compared models; the italic highlight denotes better performance of SLEET in respect to the parent framework—SchNet. [d]Statistics of three repetitions.

exerted by different ligands on each other. This matrix is then added back to ligand embedding, producing a ligand embedding that implicitly encodes ligand−ligand interaction information.

Subsequently, cross-attention is performed, where the metal embedding serves as the query source and the weighted ligand embedding is incorporated as the environmental influence on the central metal. Cross-attention assigns weighted influences to different ligand arrangements around the central metal. Here, K and V are derived from the ligand embedding, while Q originates from the metal embedding, all mapped through linear layers.

In cross-attention, the attention weight matrix obtained from the multiplication of K and Q represents the model's assessment of the degree to which each ligand's embedding should influence the central metal embedding. The V matrix, which is then multiplied by this attention weight matrix, represents the actual ligand contributions to the metal embedding. Finally, the weighted V is added to the metal embedding, completing the model's simulation of the ligand field.

2.6. Training. We utilized Schnetpack2.0 for the use of the SchNet model by removing the process of property normalization originally implemented in this model. With this modification, the predictive performance for the HOMO−LUMO gap of the QM9 data set was able to replicate in comparison with the reported results.[47] Most hyperparameter settings followed the default configurations except for the following modifications: learning rate decay = 0.5, patience = 25 (in scheduler_args), and early stop patience = 50.

For SLEET, we similarly built and executed the model using Schnetpack2.0 as the foundation.[49] Its hyperparameter settings were set as follows: n_embed_out = 512; n_atom_basis = 256, and batch size = 32, all being inherited from the work by Kneiding et al.,[56] while the settings for learning rate, learning rate decay, filters, patience, and early stop patience were identical to those used in the SchNet model reproduction experiments.

2.7. Data Set. The tmQMg data set was used for the predictions of he tH−L gap. The partition sizes of training set, validation set, and testing set were 48,639, 6079, and 6081 (80:10:10 split) of the tmQMg data set, respectively, and remained identical with the models published in the same paper as the tmQMg data set,[56] and the data distributions are shown in Figure 3. We represented all H−L gaps using the unit of eV.

3. RESULTS AND DISCUSSION

3.1. Comparison of Two-Dimensional and Three-Dimensional Information Models. This study compares the predictive performance of various models, developed based on the graph neural network (GNN) theory, in estimating the molecular electronic structure properties for the tmQMg data set. Models using two-dimensional information are represented by the SLEET model, whereas three-dimensional information (and beyond) models include the NatQG series and SchNet. The classification is determined by whether the graph information used originates from 2D or three-dimensional sources. The pioneering work reported by Kneiding et al.[56] compared the various three-dimensional information models, including enn-s2s,[46] MXMNet,[38] SchNet,[47] EdgeUpdate,[52] DimeNet++,[54] and ALIGNN.[55] The u-NatQG model adapting the undirected-edge natural quantum graph was reported to outperform the tested counterparts in predicting ΔE_{HL}, while

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

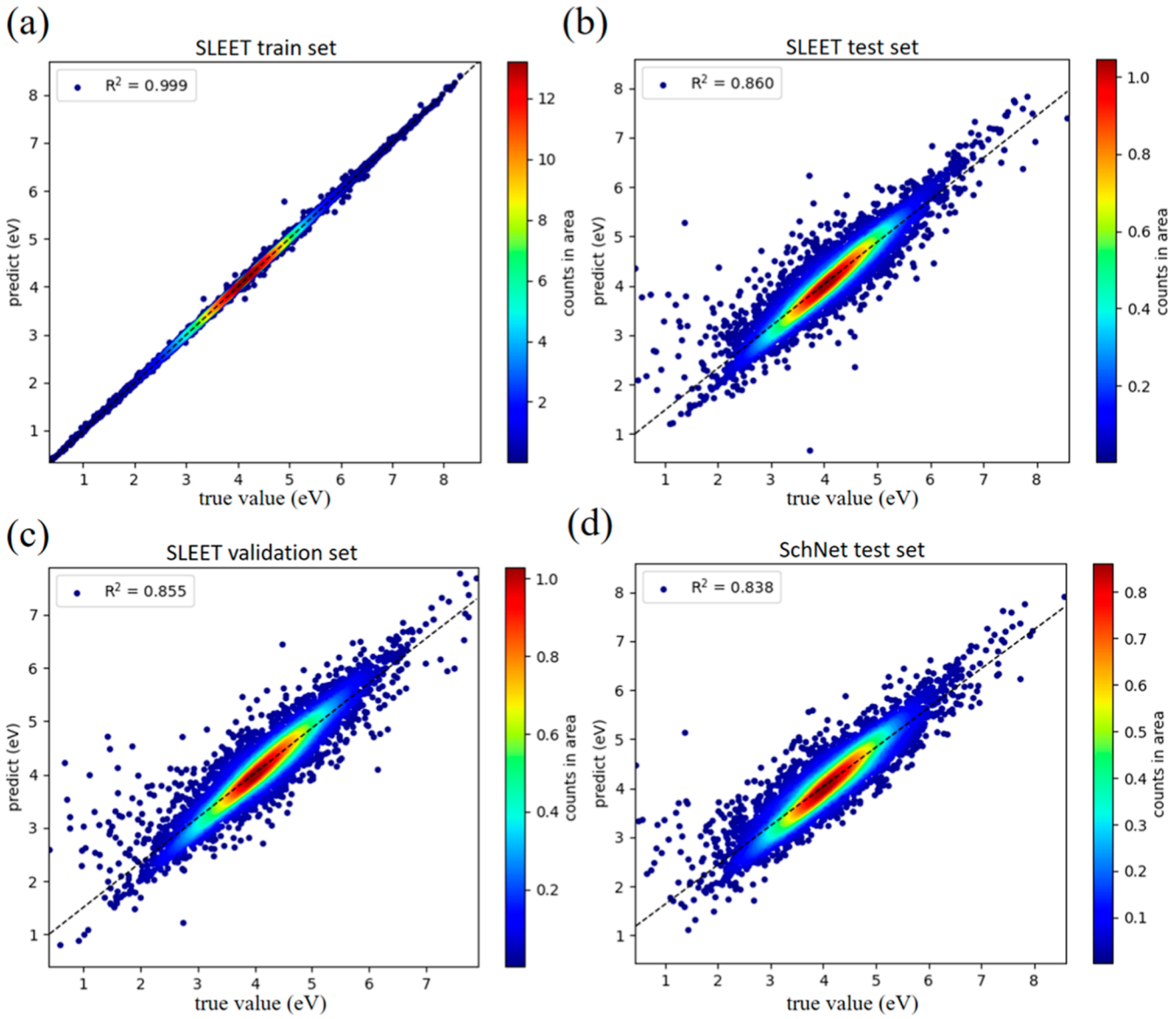


Figure 4. Scatter plot of data points predicted by SLEET and SchNet for the tmQMg data set. (a−c) Scatter plots of the SLEET predictions for the training, validation, and test sets of tmQMg, respectively, (d) scatter plot of SchNet predictions for the test set. The *x*-axis denotes the property values stored in the tmQMg data set, while the *y*-axis represents the model's predicted values. The color indicates data density, and the top-left corner displays the R^2 value. The unit is in electron volts (eV).

SchNet and DimeNet++ provided the best prediction for polarizability and dipole moment, respectively. In Table 1, we provide a concise overview of the types of information represented by different graphs. The graphs in the NatQG series are meticulously crafted in terms of the input graph information, incorporating numerous quantum properties that directly or indirectly indicate atomic behaviors, interactions, and even properties related to stability energy. For the SchNet and SLEET models, these adapt the graphs of atom-centered cutoff radius (CRG) and atom-centered cutoff bond step (CBSG), respectively. The CRG graph requires the predetermined three-dimensional molecular coordinates, and the CBSG can be derived simply from SMILES.

In Table 2, we compare the predicted results for the H−L gap, HOMO energy (E_H), LUMO energy (E_L), polarizability, and dipole moment between u-NatQG, SchNet, and SLEET models. It is apparent that u-NatQG can provide the most accurate prediction for the H−L gap, E_H, E_L, and polarizability. The use of quantum-graph features significantly enhance the predictive performance of the enn-s2s model in comparison with its baseline model (containing the generic atomic and bond properties as well as bond distances). Interestingly, it can be found that SLEET demonstrates comparable, and sometimes better, performance with respect to other three-dimensional information models (see the summary of Table S3) for the task of predicting ΔE_{HL} of TMCs. Specifically, it ranks fifth out of a total of 12 three-dimensional information models in the statistical analysis. If one excludes models that use NatQG as the input graph information, SLEET outperforms all other models for ΔE_{HL} gap prediction. This may suggest that, for TMCs, three-dimensional information may not be the only solution for predicting ΔE_{HL}, as two-dimensional information can achieve comparable levels of predictive accuracy if the ligand−metal interactions are appropriately employed. Additionally, the comparison of E_H, E_L, polarizability, and dipole moment between enn-s2s⊕G-

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

(base), SchNet, and SLEET was also done and tabulated in Table 2, where SLEET consistently provided a comparable performance to other models. Further error analysis, including MAEs of ΔE_{HL}, E_H, and E_L, is summarized in Figure S3a, where the size of training data per metal element appears to be inversely proportional to the size of prediction errors. Once the size of the training set is reduced, the prediction error increases notably, as seen in Figure S3b. The correlation of the prediction error of E_H versus E_L is assessed using mean-signed error (MSE) in Figure S4 where some TMCs represented by the same-signed MSE may imply the presence of cancellation effect in predicting ΔE_{HL}, despite these properties being trained using three independent neural networks. Figure S5 provides the categorized error analysis for the prediction of ΔE_{HL}, E_H, and E_L, and there may be a trivial cancellation effect for the Y−Zr−Nb testing set, whereas both E_H and E_L predictions were positively biased favoring good ΔE_{HL} prediction. Nonetheless, the Y−Zr−Nb data points only represent a small fraction out of the whole data set (see Figure 2).

3.2. Comparison of Performance between SLEET and SchNet Models. We further present the scatter plots of SLEET in three subsets and the results of reproducing SchNet model's prediction on the tmQMg data set in Figure 4. In our experiments, we found that the reproduced SchNet model performed significantly better than the experimental results reported by Kneiding et al.[56] We adopted the official Schnetpack framework developed by the original SchNet authors. Therefore, when comparing the model performance, we relied on our own reproduced SchNet model for benchmarking. The corresponding scatter plots of E_H and E_L are provided in Figures S1 and S2.

From the comparative results, we found that despite SLEET being built upon a modified SchNet-bs-RAN model (which means we utilized solely two-dimensional information and incorporated only bond step and bond order in the edge features), it demonstrated significantly superior performance compared to the SchNet model. Based on our previous findings, simplifying the SchNet model into the SchNet-bs variant generally results in reduced performance. Although the simple application of the RAN approach can improve the performance of the SchNet-bs model, it still lags behind that of the full SchNet model. However, SLEET, which is based on the SchNet-bs-RAN framework, surpasses the original SchNet model in terms of performance. This suggests that the MLI method and the inclusion of bond order information are highly effective for predicting the H−L gap of the TMCs. These features enable the model to better understand the relationship between ligand configurations and the H−L gap, provide explicit insights into metal−ligand interactions, and enhance the differentiation between individual ligands. Consequently, the model achieves performance levels that exceed those of the original base model.

3.3. Selection of Foundation Model for Ligands. Organic ligands play the critical role for providing the ligand field effect to the central metal valence orbitals in the formation of transition-metal complexes. In order to confirm if the current two-dimensional ligand embedding scheme can provide the reasonable description to the electronic properties of organic molecules, we compared the prediction of E_H, E_L, ΔE_{HL}, zero-point vibration energy (ZPVE), dipole moment (μ), polarizability (α), electron spatial extent $\langle r^2 \rangle$, and heat capacity using the QM9 data set as summarized in Table S4. The SchNet bond-step approach using the reduced atom-number-based formulation can actually provide the best prediction among the compared models for all of the categories. This consequently supports our selection of the foundation model of the ligand embedding block.

Two types of reduced atom number formulations were introduced by Hung et al.,[70] i.e., reduced atom number (RAN) approach and reduced atom neighbor (RANe) approach. The RAN approach mainly combines the reduced-mass-like description into the bond step of each atom pair using the atomic number, while the RANe approach primarily sums up the RAN values of the neighboring atom and appends to the bond-step description.

However, for the description of the metal−ligand interaction using SLEET, the use of RAN was found to be more effective than RANe (refer to Table S5). This inconsistency could be attributed to RANe introducing excessive redundant information, resulting in confusion during the generation of ligand node representations. Consequently, the predictive capability of models using RANe is inferior to that employing RAN.

3.4. Discussion of RDKit Errors. As previously mentioned, using RDKit to generate SMILES for TMCs introduces certain inaccuracies primarily because it applies the organic chemistry rules. This likely requires the model to inevitably handle these errors within the data set, which may in turn impact the predictive performance of the model. However, since all ionic fragments derived from TMCs are processed under the same preprocessing procedure, the systemic errors resulting from these intrinsic shortcomings appear to have had a limited effect on the outcomes. Nevertheless, identifying a more effective method for generating SMILES and recognizing ligands could potentially lead to improvements in the model performance.

4. CONCLUSION

In this study, we report the SLEET model designed for predicting the electronic properties of TMCs, solely using fragmented SMILES as the input information for feature generation. The model demonstrates good predictive performance, surpassing some three-dimensional information-based models, like enn-s2s⊕G(base), MXMNet, SchNet, EdgeUpdate, and DimeNet++, being underperformed than enn-s2s⊕G(u-NatQG) and enn-s2s⊕G(d-NatQG) models incorporating the quantum information, particularly for the task of ΔE_{HL} prediction. We utilized the SchNet-bs-RAN model as the ligand-embedding framework to generate information about ligand nodes and incorporate bond order details.

Using the Transformer-like metal−ligand interaction block, SLEET can account for metal−ligand interactions and ligand−ligand interactions and represent the relative positional differences of ligands via RPE. SLEET tends to simulate the concept of ligand field theory, thereby establishing a strong correlation between two-dimensional information and ΔE_{HL}.

SLEET provides a structured approach for the property prediction of TMCs. Currently, SLEET can suggest potential new TMCs by generating random or specified combinations of metal centers and ligands and predicting their properties, making it applicable to VHTS. In the future, SLEET could be further enhanced by integrating the generative model logic or introducing chemical knowledge via natural language embeddings, expanding its range of applications.

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00085.

Example SMILES generated by Open Babel and RDKit; PE vs RPE in SLEET; model comparison using 2D graph information and QM9 data set; comparison of RAN vs RANe using SLEET; supplementary comparison of ΔE_{HL} prediction with the various models; hyperparameters for the SLEET model; training cost of SLEET vs SchNet; and the python code for SLEET model (https://github.com/MKTMLTeam/SLEET) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

An-Cheng Yang − *National Center for High-performance Computing, National Applied Research Laboratories, Hsinchu City 30076, Taiwan; Intelligent Computing for Sustainable Development Research Center, National Taiwan Normal University, Taipei 11677, Taiwan*; Email: acyang@narlabs.org.tw

Berlin Chen − *Department of Computer Science and Information Engineering and Intelligent Computing for Sustainable Development Research Center, National Taiwan Normal University, Taipei 11677, Taiwan*; Email: berlin@csie.ntnu.edu.tw

Ming-Kang Tsai − *Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan; Intelligent Computing for Sustainable Development Research Center, National Taiwan Normal University, Taipei 11677, Taiwan; Department of Chemistry, Fu-Jen Catholic University, New Taipei City 24205, Taiwan;* orcid.org/0000-0001-9189-5572; Email: mktsai@ntnu.edu.tw

Authors

Sheng-Hsuan Hung − *Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan*

Zong-Rong Ye − *Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan*

Chi-Feng Cheng − *Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan; National Center for High-performance Computing, National Applied Research Laboratories, Hsinchu City 30076, Taiwan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00085

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study is supported by the National Science and Technology Council of Taiwan through Grant 113-2113-M-003-018, the Higher Education Sprout Project of NTNU (113J1A47) by the Ministry of Education in Taiwan, and the Physics Division of National Center for Theoretical Sciences of Taiwan. The authors are also grateful for the computational resources provided by the National Center for High-Performance Computing of Taiwan.

■ REFERENCES

(1) Cramer, C. J.; Truhlar, D. G. Density functional theory for transition metals and transition metal chemistry. *Phys. Chem. Chem. Phys.* **2009**, *11* (46), 10757−10816.

(2) Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L. Beyond Density Functional Theory: The Multiconfigurational Approach To Model Heterogeneous Catalysis. *ACS Catal.* **2019**, *9* (9), 8481−8502.

(3) Kulik, H. J. Perspective: Treating electron over-delocalization with the DFT+U method. *J. Chem. Phys.* **2015**, *142* (24), 240901.

(4) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024−10035.

(5) Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. Molecular Mechanics for Coordination Complexes: The Impact of Adding d-Electron Stabilization Energies. *J. Am. Chem. Soc.* **1995**, *117* (32), 8407−8415.

(6) Li, P.; Merz, K. M., Jr Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **2017**, *117* (3), 1564−1686.

(7) Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Mueller-Westerhoff, U. T. An intermediate neglect of differential overlap technique for spectroscopy of transition-metal complexes. Ferrocene. *J. Am. Chem. Soc.* **1980**, *102* (2), 589−599.

(8) Janet, J. P.; Duan, C.; Nandy, A.; Liu, F.; Kulik, H. J. Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design. *Acc. Chem. Res.* **2021**, *54* (3), 532−545.

(9) Kronik, L.; Stein, T.; Refaely-Abramson, S.; Baer, R. Excitation Gaps of Finite-Sized Systems from Optimally Tuned Range-Separated Hybrid Functionals. *J. Chem. Theory Comput.* **2012**, *8* (5), 1515−1531.

(10) Cytter, Y.; Nandy, A.; Duan, C.; Kulik, H. J. Insights into the deviation from piecewise linearity in transition metal complexes from supervised machine learning models. *Phys. Chem. Chem. Phys.* **2023**, *25* (11), 8103−8116.

(11) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.

(12) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12* (3), 191−201.

(13) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10* (3), 2354−2377.

(14) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15* (10), 1120−1127.

(15) Shu, Y.; Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **2015**, *142* (10), 104104.

(16) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336−2347.

(17) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.

(18) Chen, A.; Zhang, X.; Zhou, Z. Machine learning: Accelerating materials development for energy storage and conversion. *InfoMat* **2020**, *2* (3), 553−576.

(19) Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2021**, *121* (16), 9873−9926.

(20) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89* (9), 094104.

(21) Zhao, L.; Zhang, J.; Zhang, Y.; Ye, S.; Zhang, G.; Chen, X.; Jiang, B.; Jiang, J. Accurate Machine Learning Prediction of Protein Circular Dichroism Spectra with Embedded Density Descriptors. *JACS Au* **2021**, *1* (12), 2377−2384.
(22) Liu, T.; Johnson, K. R.; Jansone-Popova, S.; Jiang, D.-E. Advancing Rare-Earth Separation by Machine Learning. *JACS Au* **2022**, *2* (6), 1428−1434.
(23) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th ICML; Proceedings of Machine Learning Research*, 2021.
(24) Arrigoni, M.; Madsen, G. K. H. Evolutionary computing and machine learning for discovering of low-energy defect configurations. *Npj Comput. Mater.* **2021**, *7* (1), 71.
(25) Fung, V.; Hu, G.; Ganesh, P.; Sumpter, B. G. Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* **2021**, *12* (1), 88.
(26) Schütt, O.; Vandevondele, J. Machine Learning Adaptive Basis Sets for Efficient Large Scale Density Functional Theory Simulation. *J. Chem. Theory Comput.* **2018**, *14* (8), 4168−4175.
(27) Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *J. Chem. Inf. Model.* **2021**, *61* (3), 1053−1065.
(28) Masood, H.; Toe, C. Y.; Teoh, W. Y.; Sethu, V.; Amal, R. Machine Learning for Accelerated Discovery of Solar Photocatalysts. *ACS Catal.* **2019**, *9* (12), 11774−11787.
(29) Tsymbalov, E.; Shi, Z.; Dao, M.; Suresh, S.; Li, J.; Shapeev, A. Machine learning for deep elastic strain engineering of semiconductor electronic band structure and effective mass. *Npj Comput. Mater.* **2021**, *7* (1), 76.
(30) Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine Learning for Predicting Electron Transfer Coupling. *J. Phys. Chem. A* **2019**, *123* (36), 7792−7802.
(31) Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.-B.; Minoux, H.; et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* **2020**, *16* (8), 4757−4775.
(32) Gawriljuk, V. O.; Zin, P. P. K.; Puhl, A. C.; Zorn, K. M.; Foil, D. H.; Lane, T. R.; Hurst, B.; Tavella, T. A.; Costa, F. T. M.; Lakshmanane, P.; et al. Machine Learning Models Identify Inhibitors of SARS-CoV-2. *J. Chem. Inf. Model.* **2021**, *61* (9), 4224−4235.
(33) Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine learning property prediction for organic photovoltaic devices. *Npj Comput. Mater.* **2020**, *6* (1), 166.
(34) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Machine Learning: Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra (Adv. Sci. 9/2019). *Adv. Sci.* **2019**, *6* (9), 1970053.
(35) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **2021**, *20* (6), 750−761.
(36) Saidi, W. A.; Shadid, W.; Castelli, I. E. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *Npj Comput. Mater.* **2020**, *6* (1), 36.
(37) Cools-Ceuppens, M.; Dambre, J.; Verstraelen, T. Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential. *J. Chem. Theory Comput.* **2022**, *18* (3), 1672−1691.
(38) Zhang, S.; Liu, Y.; Xie, L. Molecular Mechanics-Driven Graph Neural Network with Multiplex Graph for Molecular Structures. *arXiv* **2020**, *2011*, 07457.
(39) Mezei, P. D.; Von Lilienfeld, O. A. Noncovalent Quantum Machine Learning Corrections to Density Functionals. *J. Chem. Theory Comput.* **2020**, *16* (4), 2647−2653.
(40) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. *J. Chem. Theory Comput.* **2021**, *17* (2), 826−840.
(41) Ye, Z.-R.; Huang, I.-S.; Chan, Y.-T.; Li, Z.-J.; Liao, C.-C.; Tsai, H.-R.; Hsieh, M.-C.; Chang, C.-C.; Tsai, M.-K. Predicting the emission wavelength of organic molecules using a combinatorial QSAR and machine learning approach. *RSC Adv.* **2020**, *10* (40), 23834−23841.
(42) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13* (11), 5255−5264.
(43) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein pKa Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* **2022**, *18* (4), 2673−2686.
(44) Margraf, J. T.; Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. *Nat. Commun.* **2021**, *12* (1), 344.
(45) Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-De-Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *J. Chem. Inf. Model.* **2017**, *57* (1), 11−21.
(46) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *ICML., Proceedings of Machine Learning Research*, 2017.
(47) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.
(48) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K. R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15* (1), 448−455.
(49) Schütt, K. T.; Hessmann, S. S. P.; Gebauer, N. W. A.; Lederer, J.; Gastegger, M. SchNetPack 2.0: A neural network toolbox for atomistic machine learning. *J. Chem. Phys.* **2023**, *158* (14), 144801.
(50) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8* (1), 13890.
(51) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv* **2017**, *1706*, 08566.
(52) Bjørn Jørgensen, P.; Wedel Jacobsen, K.; Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. *arXiv* **2018**, *1806*, 03146.
(53) Gasteiger, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* **2020**, *2003*, 03123.
(54) Gasteiger, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* **2020**, *2011*, 14115.
(55) Choudhary, K.; Decost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *Npj Comput. Mater.* **2021**, *7* (1), 185.
(56) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; De Bin, R.; Balcells, D. Deep learning metal complex properties with natural quantum graphs. *Digit. Discovery* **2023**, *2*, 618.
(57) Duan, C.; Chen, S.; Taylor, M. G.; Liu, F.; Kulik, H. J. Machine learning to tame divergent density functional approximations: a new path to consensus materials design principles. *Chem. Sci.* **2021**, *12* (39), 13021−13036.
(58) Duan, C.; Nandy, A.; Terrones, G. G.; Kastner, D. W.; Kulik, H. J. Active Learning Exploration of Transition-Metal Complexes to Discover Method-Insensitive and Synthetically Accessible Chromophores. *JACS Au* **2023**, *3* (2), 391−401.
(59) Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60* (12), 6135−6146.

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420

(60) Garrison, A. G.; Heras-Domingo, J.; Kitchin, J. R.; dos Passos Gomes, G.; Ulissi, Z. W.; Blau, S. M. Applying Large Graph Neural Networks to Predict Transition Metal Complex Energies Using the tmQM_wB97MV Data Set. *J. Chem. Inf. Model.* **2023**, *63* (24), 7642−7654.
(61) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure−Property Relationships. *J. Phys. Chem. A* **2017**, *121* (46), 8939−8954.
(62) RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed Oct 18, 2021).
(63) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Chem. Inf.* **2011**, *3* (1), 33.
(64) Calderon, C. E.; Plata, J. J.; Toher, C.; Oses, C.; Levy, O.; Fornari, M.; Natan, A.; Mehl, M. J.; Hart, G.; Buongiorno Nardelli, M.; et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **2015**, *108*, 233−238.
(65) Toher, C.; Oses, C.; Hicks, D.; Gossett, E.; Rose, F.; Nath, P.; Usanmaz, D.; Ford, D. C.; Perim, E.; Calderon, C. E.; et al. The AFLOW Fleet for Materials Discovery. In *Handbook of Materials Modeling: Methods: Theory and Modeling*; Andreoni, W., Yip, S., Eds.; Springer International Publishing, 2018; pp 1−28.
(66) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—a Python library for working with atoms. *J. Condens. Matter Phys.* **2017**, *29* (27), 273002.
(67) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37* (22), 2106−2117.
(68) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145* (16), 161102.
(69) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114* (10), 105503.
(70) Hung, S.-H.; Ye, Z.-R.; Cheng, C.-F.; Chen, B.; Tsai, M.-K. Enhanced Predictions for the Experimental Photophysical Data Using the Featurized Schnet-Bondstep Approach. *J. Chem. Theory Comput.* **2023**, *19* (14), 4559−4567.
(71) Sim, J.; Kim, D.; Kim, B.; Choi, J.; Lee, J. Recent advances in AI-driven protein-ligand interaction predictions. *Curr. Opin. Struct. Biol.* **2025**, *92*, 103020.
(72) Zhao, X.; Wang, B.; Zhou, K.; Wu, J.; Song, K. High-Throughput Prediction of Metal-Embedded Complex Properties with a New GNN-Based Metal Attention Framework. *J. Chem. Inf. Model.* **2025**, *65* (5), 2350−2360.
(73) Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *arXiv* **2018**, *1803*, 02155.
(74) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is All you Need. In *31st Conference on Neural Information Processing Systems*, 2017.
(75) Krenn, M.; Ai, Q.; Barthel, S.; Carson, N.; Frei, A.; Frey, N. C.; Friederich, P.; Gaudin, T.; Gayle, A. A.; Jablonka, K. M.; et al. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3* (10), 100588.
(76) Shoghi, N.; Kolluru, A.; Kitchin, J. R.; Ulissi, Z. W.; Zitnick, C. L.; Wood, B. M. From Molecules to Materials: Pre-training Large Generalizable Models for Atomic Property Prediction. *arXiv* **2023**, *2310*, 16802.

https://doi.org/10.1021/acs.jctc.5c00085
J. Chem. Theory Comput. 2025, 21, 6410−6420