Transferable Multi-level Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multi-task Learning

Ziteng Liu[†], Liqiang Lin[§], Qingqing Jia[†], Zheng Cheng[†], Yanyan Jiang[§], Yanwen Guo[*,§], Jing Ma*[†,‡]*

[†]Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, P. R. China

[§]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, P. R. China

[‡]Jiangsu Key Laboratory of Advanced Organic Materials, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

*Address correspondence to: majing@nju.edu.cn (Prof. Jing Ma, ORCID: 0000-0001-5848-9775), ywguo@nju.edu.cn (Prof. Yanwen Guo)

Abstract

The development of efficient models for predicting specific properties through machine learning is of great importance for the innovation of chemistry and material science. However, predicting global electronic structure properties like frontier molecular orbital HOMO and LUMO energy levels and their HOMO-LUMO gaps from the small-sized molecule data to larger molecules remains a challenge. Here we develop a multi-level attention neural network, named DeepMoleNet, to enable chemical interpretable insights being fused into multi-task learning through (1) weighting contributions from various atoms and (2) taking the atom-centered symmetry functions (ACSFs) as the teacher descriptor. The efficient prediction of 12 properties including dipole moment, HOMO, and Gibbs free energy within chemical accuracy is achieved by using multiple benchmarks, both at the equilibrium and non-equilibrium geometries, including up to 110,000 records of data in QM9, 400,000 records in MD17 and 280,000 records in ANI-1ccx for random split evaluation. The good transferability for predicting larger molecules outside the training set is demonstrated in both equilibrium QM9 and Alchemy datasets at density functional theory (DFT) level. Additional tests on non-equilibrium molecular conformations

from DFT-based MD17 dataset and ANI-1ccx dataset with coupled cluster accuracy as well as the public test sets of singlet fission molecules, biomolecules, long oligomers, and protein with up to 140 atoms show reasonable predictions for thermodynamics and electronic structure properties. The proposed multi-level attention neural network is applicable to high-throughput screening of numerous chemical species in both equilibrium and non-equilibrium molecular spaces to accelerate rational designs of drug-like molecules, material candidates, and chemical reactions.

1. Introduction

Chemistry is indispensable in human daily life as well as the research and development of clothing, drugs, and materials, etc. Nowadays, the powerful quantum chemical calculations in combination with big databases and artificial intelligence are changing the painstaking "try and error" works to rational discovery of novel molecules and materials with the desired properties. [1-8] Computational cost of quantum chemistry calculation increases rapidly as the sizes of systems increase. To facilitate the discovery process, quantum chemistry calculations based on density functional theory (DFT) have been widely used in various chemical systems with the computational scaling of $O(N_b^3)$, where N_b is the number of basis sets. The gold standard CCSD(T)/CBS is even more expensive with the cost of $O(N_b^7)$. Development of lower and even linear scaling methods has aroused great interest in the past decade. [9-16] In spite of these advances in the quantum chemical methods, the quick prediction of various electronic structure properties is highly desired in high-throughput searching of large chemical spaces with all possible combinations of functional groups, towards the material or drug design. [4,17, 18]

Many machine learning methods have been introduced in quantum chemical study for the rapid predictions of atomic forces, molecular energy, and electronic structure properties, which are of great importance for the construction of accurate force fields and complicated potential energy surfaces as well as rational design of various materials and drug-like candidates. [1, 19-37] If the data-driven model is trained properly, it could remarkably reduce computational costs but with similar accuracy to quantum chemical calculation. High-throughput computational screening hence

2

becomes possible, outputting the properties of millions of compounds with broad applications in a fast and accurate way. For example, functions of novel optical materials and electronic devices are correlated with the descriptors of dipole moment, polarizability, energy levels of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), and so on. [38, 39] Drug-like molecules can be screened with cohesive Gibbs free energy and other descriptors. [8] The term of electronic spatial extent, <R^2>, may be useful for the design of batteries. [40] Heat capacity is an important descriptor in predicting the properties of ionic liquids and many thermal materials. [41, 42] However, it is still a big challenge to accurately predict some electronic structure properties such as the HOMO and LUMO energy levels, and furthermore, the HOMO-LUMO energy gap, of much larger molecules that beyond the training set with good transferability. [43-45]

To improve the prediction accuracy of HOMO and LUMO energy levels, we propose in this paper an efficient multi-level attention neural network, named as DeepMoleNet, for the molecular systems. Our target is to establish an implicit relationship between the molecular structural information and 12 electronic structure properties including dipole moment, polarizability, HOMO, LUMO, HOMO-LUMO gap, zero point vibration energy (ZPVE), electronic spatial extent (<R^2>), internal energy at zero and room temperature (U_0, U), enthalpy (H), free energy (G), and heat capacity (C_v). It may be illustrative to draw an analogy between the language learning and property prediction from quantum chemical datasets. In the machine learning and translation of a sentence in a certain kind of language, the meaning of the translated word is associated with specific words and phrases in the context in neural machine translation. For a certain molecule in chemistry realm, every atom node is affected by its chemical environment adjacent to it with different weights, like that an oxygen atom always behaviors differently in tetrahydrofuran and tetrahydro-2H-pyran; nitrogen atom in pyridine *versus* pyrimidine. The weights to address the different impacts from 'environment' atoms on the 'center' atom are called 'attention'. In the present work, we apply multi-level attention in every message passing step which would gradually capture the influence of different atomic nodes at each 'time step', T,

and the attention weight varies as the node representation changes in a dynamic manner. Such a multi-level attention strategy differs significantly from some other attention algorithms, in which attention is used once at the defined step. [46-48] To help the neural network learn a richer representation from atomic structural information, we use the multi-task learning (MTL), combining different known information of input data sample as the related auxiliary tasks to improve the predictive power of our main targets. [49-51] In addition to the pursuit of 12 quantum chemistry properties as the learning targets, the atom-centered symmetry functions (ACSFs) [52] descriptor is selected as the auxiliary prediction targets. ACSFs have been widely applied as the input in many chemistry applications such as predictions of organic reactions, phase transition, surface catalysis, etc. [53-58] In the proposed DeepMoleNet model, we obtain the final node feature after T step multi-level attention cycle to predict the auxiliary ACSFs task. In other words, chemistry knowledge underlying the ACSFs descriptor is combined with multi-task learning to improve the generalizability and transferability of the deep learning model. As shown in Figure 1, we train with small molecules in QM9[59] (70,000 record with the number of atoms in the molecule, N_{atom} <19) and Alchemy[44] (72,000 with N_{atom} <22) datasets, respectively, to predict the 12 quantum chemistry properties of larger molecules with up to $N_{atom} = 29$ in QM9 and $N_{atom} = 38$ in Alchemy. The present DeepMoleNet model could simultaneously predict 12 properties with better performance, accuracy, and robustness compared with other single-target training models. We even get better results by using the early epoch of multi-targets model as model initialization to train the single-target training model. Furthermore, this multi-task model is transferable to larger molecules with broad application scopes of drug-like molecules, peptides, macrocyclic molecules, oligomers, protein, and singlet fission molecules than those $N_{atom} \leq 29$ used for training in the QM9 dataset. This indicates the potential of our model in predicting the electronic structure properties of complex quantum chemical systems with satisfactory generalizability and transferability.
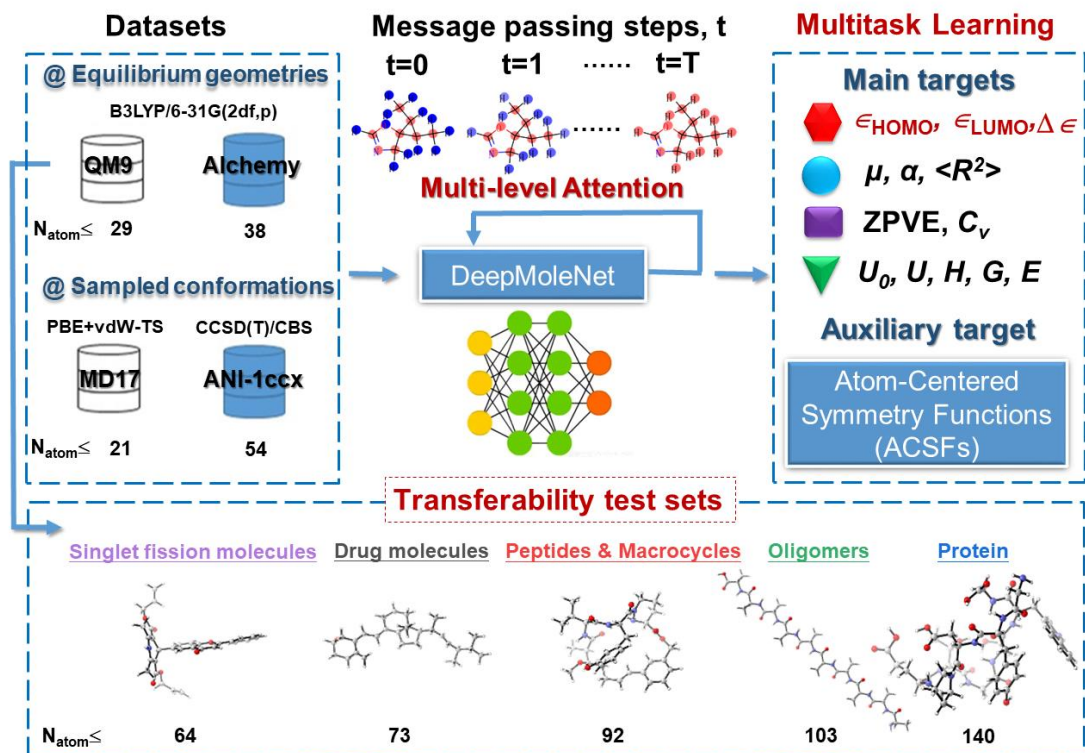
Figure 1. Illustration of the multi-level attention neural network, DeepMoleNet, which was subjected to extensive tests on quantum chemistry datasets at both equilibrium and non-equilibrium conformations as well as multiple public test sets with different applications. Each dataset is split into 2 different groups, the small sized training group (for examples, QM9: $N_{atom} \leq 18$; Alchemy: $N_{atom} \leq 22$) and the large test group (for QM9: $19 \leq N_{atom} \leq 29$; Alchemy: $23 \leq N_{atom} \leq 38$). Multi-level attention is applied to the message passing phase, and then after T steps, the auxiliary task is achieved. Finally, after the readout phase, the main tasks are done to test the transferability of the model.

2. Method

Shown in Figure 2 is an illustrative flowchart for the implementation of the proposed DeepMoleNet model, which has three steps, i.e., input, message passing, and readout. In this section, we will introduce details of the learning method, databases, and error analysis of our predicted results.
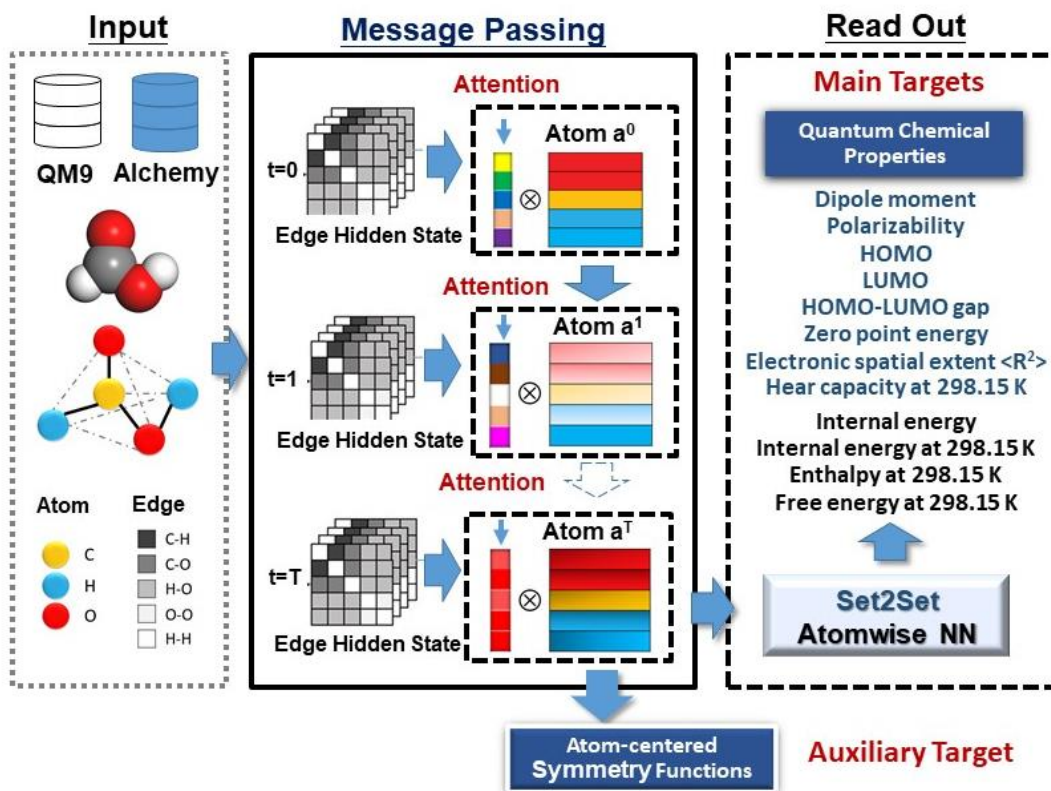
Figure 2. Implementation steps of DeepMoleNet. Representation of each node is generated at first through multi-linear perceptron, and subsequently, the attention coefficient is initialized to make Hadamard product with each node. Then the node hidden states are aggregated with edge hidden states. Hidden states of LSTM are used for multi-level attention generation. After T steps, ACSFs are predicted by each node representation. Finally, each node hidden state is used in set2set and atomwise neural network to predict the 12 quantum properties.

2.1 DeepMoleNet architecture

Input. In our DeepMoleNet model, each molecule (labeled as *M*) in QM9 and Alchemy datasets is input as a set of atom types (also called nodes), a_i, and atom pairs (edge types), b_{ij}, as shown in Figure 2. The selected features of nodes (a_i) and edges (b_{ij}) inputs are given in Table 1. For nodes, we use atom types, chirality, atomic number, hybridization, etc., which were generated by using RDkit code[60]. Gasteiger partial charge, as embedded in RDKit, was used for calculating atomic charges. The selected edges include descriptors of bond type, same ring, topological path length, shortest path bonds, graph distance, extended distance, geometric distance, in which the position information is implicitly encoded in atom pair features through Gaussian expansion [61] in equation (1).

$$F_{\text{Gaussian expansion}} = e^{\frac{-(r-r_0)^2}{\sigma^2}} \tag{1}$$

where r is the distance between 2 nodes. The Gaussian functions are centered at 20 locations with the peak center parameters, r_0, which are linearly placed between 0 and 4. We set the peak width parameter $\sigma = 0.5$.

In this work, we aim to design a neural network algorithm to map the relation between molecule, $M\,(a_i, b_{ij})$, and their 12 quantum chemical properties, P, through f: $\{M_i\}| \rightarrow P_{i=1}^{12}$. It is noticed that extended-connectivity fingerprints (ECFP4) [61] were applied as inputs for quantum chemical property predictions.

In Table 1, we also list the input features of other machine learning methods for quantum chemical applications. The ACSFs were widely used in the construction of high-dimensional neural network potential-energy surfaces. The kernel-based machine learning of molecular properties is realized by transforming fingerprints and representations non-linearly with kernel functions. [25, 34, 62-66] There are also some descriptors, such as the smooth overlap of atomic positions (SOAP) [67] method, the bag of bonds[25, 68] approach and Fourier series of atomic radial distribution functions[69]. However, deep learning[70] can directly learn from low-level molecular structure information (e.g., atom types and bond types), and then gradually extract high-level representation through deep multiple neural network layers to predict targets. [43, 61, 71-86] Since we can generally achieve acceptable performance in deep learning by 'focusing' on our target tasks, the information of ACSFs could be excluded from the input, instead we set ACSFs as one of the targets in the present DeepMoleNet model. Such ACSFs information could come from the training signals of the related tasks in the multi-task learning (MTL), which has been successfully used in many fields ranging from natural language processing and speech recognition, computer vision, and drug discovery. [49-51] MTL is able to improve generalization by parallel training tasks with the shared representations between the related tasks. [87]

Table 1. Comparisons between various neural network methods

Section		DeepMoleNet this work	SchNet[77]	enn-s2s[73]	MGCN[74]	DimeNet[43]	AIMNet
Input	Node a_i	Atom type, Chirality, Atomic number, Acceptor, Donor, Aromatic, Hybridization, Number of Hydrogens, Forcefield charge, Valence, van der Waals radius, Node degree	Element Embedding	Atom type, Atomic number, Acceptor, Donor, Aromatic, Hybridization, Number of Hydrogens	Element embedding	Element embedding	atom-centered environment vectors (AEVs)
	Edge b_{ij}	Bond type, Same ring, Topological path length, Shortest path bonds, Graph distance, Extended distance, Geometric distance	Radial basis functions	Bond type, distance	Radial basis functions	Fourier-bessel basis functions	
Message passing		Multi-level attention	Continuously filter convolutional layers based interaction	Gated graph neural networks	Multi-level interaction (atom-wise, atom-pair, triple-wise atom interaction, etc.)	Directional message passing (message passed through angle for triple-wise atom interaction)	Atomic interactions in molecules
Readout		Atomwise layer Neural Network and Set2Set	Atomwise layer Neural Network	Set2Set	Atomwise layer Neural Network	Elemental wise layer Neural Network	Atomwise layer Neural Network

Message passing neural networks with multi-level attention. After the input section, DeepMoleNet will subsequently run message passing and readout processes. Message passing neural networks (MPNN) [73] have been used to learn features from molecular graphs through abstracting the commonalities between several existing

neural models for graph structured data. It is useful to solve molecular systems on molecular graphs M with node and edge features. Several kinds of graph neural networks have been developed with the emphasis laid on the dynamically updating the node features in response to its neighboring node features and edge information. When performing message passing, every node is connected to all its neighbors. In addition, the edges between two nodes act as a modulator to transmit these messages. We can exemplify the message passing by energy prediction of a molecule, which consists of several atoms, and there are various interactions among these atoms. If the two atoms are too far away from each other, their atomic interaction may be very weak and the edges in the graph would reduce messages; but if two atoms are close to each other, the edges in the graph would enlarge the messages due to the significant atom-pair interaction.

We can give a mathematic expression for the message passing phase, in which node features x_v are fed into a node network to obtain a hidden feature of dimension d, and edge features e_{vw} are fed into an edge network to obtain a $d{\times}d$ matrix, respectively. Then hidden states of each node are aggregated with message function M_t to generate the node hidden message, m_v^{t+1}, where $N(v)$ denotes the neighbors of node v, as shown in equation (2).

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \qquad (2)$$

Here, M_t is defined as follows.

$$M_t(h_v^t, h_w^t, e_{vw}) = A(e_{vw})h_w \qquad (3),$$

where $A(e_{vw})$ is the edge network.

The message function is followed by the update function. The update process can be described by analogy of molecular system. The chemically active atoms would strongly affect the surrounding atom; in return, the adjacent active atoms are more likely to surround an active atom.

In the present work, we extend the MPNN framework with the multi-level 'attention mechanism' in the update function of the message passing stage to capture

the long ranged interaction information in quantum chemistry calculation data. Attention is similar to the way of observing an object. Our visual system tends to pay attention to some parts of the image selectively, while ignoring other irrelevant information. Therefore, attention mechanisms assign different weights to different parts of the inputs, allowing the extraction of more critical and important information with modest computational costs. In such a way, attention enables the model to distinguish different input information and explain what the model has learned, so as to give an interpretable view for the 'black box' deep learning systems.

The update process is presented as follows.

$$X_v^{t+1}, q_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{4}$$

When there is no attention mechanism applied, X_v^{t+1}, the output of U_t, goes back to h_v^{t+1} in the update function of the traditional MPNN[73]. The q_v^{t+1} term in equation (4) is the hidden state of time-dependent update function, similar to that used in Long Short Term Memory Network (LSTM)[88]. LSTM is used to update the node state for further aggregation in an order free manner. As shown in equation (5), the sigmoid activation function is used to obtain the attention score, Z, which indicates the importance of this atom in the molecule.

$$Z = \sigma(MLP(q_v^{t+1}) \cdot X_v^{t+1}) \tag{5}$$

Then, h_v^{t+1} is the weighted information of X_v^{t+1} with the broadcast of the attention score Z, as shown in equation (6).

$$h_v^{t+1} = X_v^{t+1} \otimes Z \tag{6}$$

where $\otimes$ is the Hadamard operator.

We utilize information extracted from message passing phase and the LSTM hidden state to generate attention vector through multiplying this vector by the coefficient generated by the attention mechanism, and feed it to the LSTM block. Therefore, the whole model could be viewed as a model with the embedded pooling layer MPNN

and attention mechanism.

Node features, x_v, and edge features, e_{vw}, run with both message function and update function for T time steps in this message passing phase. After the multi-level attentional node message aggregation, the final node representation at T step is used to predict the ACSFs and the outputs of readout function are used for the main target prediction of 12 quantum chemistry properties.

Read Out. In the present work, two readout functions, Set-to-set (Set2Set) and atom-wise summation neural networks, are used to simultaneously predict the 12 targets in a neural network. In the readout phase, a graph level representation is calculated with defined readout function R according to the following equation.

$$\hat{y} = R(\{h_v^T, |v \in M). \tag{7}$$

where, h_v^T means the final node representation of message passing in the T step, v is the node of the molecular graph, M. The selection of two kinds of readout functions is judged by the intensive or extensive nature of the 12 predicted molecular properties. Among them, HOMO, LUMO, HOMO-LUMO gap, dipole moment, and heat capacity belong to the intensive properties, for which the Set2Set is used in the multi-task training progress to make predictions. In contrast, the energy-related properties, G, H, U, and U_0, are extensive properties, whose prediction is completed by using the atom-wise summation neural networks. The remaining three kinds of properties, i.e., polarizability, $<R^2>$, and ZPVE are predicted by using the Set2Set readout for convergence consideration.

Set-to-set (Set2Set) framework has been used in enn-s2s[73] model, which produces graph-level embedded representation with global attention, instead of summing up the final node states. The atom-wise summation neural networks are applied with atom-wise layers for each node separately to the node hidden representations h_i with shared weights W^l and biases b^l on layer l as follows.

$$h_v^{T,l+1} = W^l h_v^{T,l} + b^l \tag{8}$$

where the $h_v^{T,l}$ is the final T step output of the message passing phase at layer *l*. The atom-wise neural network produces an output on each node representation for each node, and then predicts the target properties by simply summing up or averaging the final node states. In the DeepMoleNet, the message passing steps is set as 6 by trial and error. The learning rate is 1×10^{-5}. Adam[89] is used as the optimization model, and the batch size is 32. Our model is run in 400 epochs, in which 20 epochs are used for warm-up. More details are shown in Table S1.

In addition, DeepMoleNet is featured as the combination of feature engineering and the deep learning approach within the framework of multi-task learning (MTL) theory. Feature engineering could gain knowledge from dataset, so we introduce an auxiliary task to predict ACSFs at the end of the message passing phase. Outputs of the readout function are used for the main targets of 12 properties prediction with the predicting loss, *L*, defined as follows.

$$L = L_{main\ targets} + \lambda L_{ACSFs} \tag{9},$$

where $L_{main\ targets}$ represents the loss of 12 predicted molecular properties (main targets), as shown in equation (10). L_{ACSFs} is the loss of the predicted ACSFs values, shown in equation (10). Here, λ is the loss weight of L_{ACSFs}. It is set to be 16 by trial and error.

$$L_{main\ targets} = \frac{\sum_P^{8\ properties\ in\ Set2Set} \sum_i^{all} \left| P_i^{DeepMoleNet} - P_i^{DFT} \right|}{N_{molecules}}$$

$$+ \rho \frac{\sum_P^{4\ properties\ in\ Atom-wise\ NN} \sum_i^{all} \left| P_i^{DeepMoleNet} - P_i^{DFT} \right|}{N_{molecules}}$$

$$\tag{10},$$

where the first term is set for the readout 8 properties, including dipole, polarizability, HOMO, LUMO, HOMO-LUMO gap, <R2>, ZPVE, and heat capacity. For those properties using Set2Set, they are normalized, and for extensive properties, the weight is 0.5. In the second term of equation (10), ρ=0.5, which is the weight for all 4

energy-related properties, U_0, U, H, and G, in Atom-wise NN.

$$L_{ACSFs} = \frac{\Sigma_i^{all}\left|P_i^{DeepMoleNet} - P_i^{ACSFs}\right|}{N_{molecules}} \tag{11}$$

During the message passing phase, the hidden state of node is aggregated by the influence of its neighbors. However, the spatial information is lost in this way of topological information aggregation. There are many efforts devoted to MTL. One way is to combine Coulomb matrix with neural network [90]. Another is using atom-centered environment vectors (AEVs) as input like AIMNet[91]. Instead of using AEVs as input attributes in the deep learning approach like AIMNet[91], DeepMoleNet sets the ACSFs descriptors as the predicting targets at the end of message passing phase to help every node get the spatial information fused with the learned representations. For spatially sensitive properties such as molecular dipole moment, the auxiliary ACSFs prediction tasks may be helpful to fuse spatial information with the representation learned in the data-driven way. In subsection 3.4, we will further elaborate the performance of using ACSFs as both descriptor and prediction task.

To summarize, our model has features of (1) introducing more detailed description of atom and atom pair information in the input section; (2) using multi-level attention for node message aggregation; (3) employing the auxiliary target, ACSFs, to help the graph convolution model obtain better learned representation; and (4) adopting both set2set readout function and atomwise neural network in simultaneously predicting the 12 quantum properties, which are the main targets of this work.

2.2 Comparison with other methods

It is useful to make comparison between various deep learning methods for molecular property predictions in Table 1. In this work, the inclusion of the complete edge feature vector (bond type, spatial distance) and treating hydrogen atoms as explicit nodes in the graph are found to be crucial to get good predictions for a number of targets in multi-task learning. The enn-s2s[73] selected less descriptors of node and bond

features in message passing. The difference between the Set2Set[73] and the multi-level attention we used lies in the node information aggregation. We made attention on local chemical environment of each atom to extract the final node representation, while Set2Set[73] further made attention with the final node representation to produce the final global graph representation. In addition, SchNet[77] used element embedding for node feature and radial basis functions to incorporate the continuously filter convolutional layers in message passing and atom-wise layers for final prediction. MGCN[74] developed the multi-level interaction (atom-wise, atom-pair, triple-wise atom interaction, etc.) in message passing; and DimeNet[43] used directional message of angle to pass information of the surrounding atoms with Fourier-bessel basis functions. AIMNet[91] used node embedding from both atomic coordinates and atomic numbers to form the atom-centered environment vectors, then obtained atomic feature vectors (AFVs) in the iterative message passing for downstream tasks (Table 1).

2.3 Datasets

In the application of DeepMoleNet, we employed multiple public datasets including QM9[59], Alchemy[44], MD17[92], ANI-1ccx[93, 94] quantum chemistry computation results, as shown in Table 2. Some typical molecules with potential applications in drug and material design were also studied using DeepMoleNet.

Table 2. Dataset details and error analysis of the predicted data

Dataset	Computation level	Tasks	Data Size[a]	Rec-split	Rec-metric
DFT data @ equilibrium geometries					
QM9[59]	B3LYP/6-31G(2df,p)	12	133,885/129,428	Random	MAE, std_{MAE}
Alchemy[44]	B3LYP/6-31G(2df,p)	12	202,579/183,051	Random	MAE
DFT or CCSD data @ sampled conformations					
MD17[92]	PBE+vdW-TS	1	3,611,115/560,000	Random	MAE
ANI-1ccx	CCSD(T)/CBS	1	489,571/332,196	Random	MAE, RMSE
Transferability test sets					
Drug-like molecules[95]	All the test molecules	2	24/17		MAE, Δ_{MAE}
MPCONF196[96]	were re-optimized	2	192/11		MAE, Δ_{MAE}
Singlet fission molecules[37]	with	2	262/9		MAE, Δ_{MAE}
oligomers[37]	B3LYP/6-31G(2df,p)	2	12		MAE, Δ_{MAE}
protein[37]		2	2/1		MAE, Δ_{MAE}

[a] The amount of data in the public dataset is given before the slash, and the number of data used in this work is shown after the slash.

2.3.1 Chemical equilibrium datasets

QM9. The QM9[59] dataset contains about 130,000 organic molecules composed of 9 heavy atoms including C, O, N, and F, etc. within the GDB-17 database. Various molecular properties were calculated at the theoretical level of DFT/B3LYP/6-31G(2df,p). The collected molecules in QM9 cover a wide range including (hetero-) alkane, amide, amine, alcohol, epoxy, ether, ester, chloride, aliphatic, and aromatic groups. In this work, the 110,000 QM9 data are randomly selected as the training set, 10,000 data for validation, and the rest for test set, with the molecular size, N_{atom}, varying from 3 to 29 atoms (Figure S1).

Alchemy. Alchemy[44] includes 202,579 molecules with a maximum of 14 heavy atoms (including C, O, N, F, S, Cl, etc.) and $N_{atom} = 11 \sim 38$ sampled from GDB MedChem dataset. All the 12 properties were obtained with Python-based Simulations of Chemistry Framework (PySCF) [97]. All geometries were calculated with the density fitting approximation for electron repulsion integrals using the B3LYP/6-31G(2df,p). Unlike the QM9 dataset, the auxiliary basis cc-pVDZ-jkfit was used in density fitting to build the Coulomb matrix and the HF exchange matrix. The meta-Lowdin population analysis was employed to obtain the atomic charges. It should be mentioned that some of the quantum chemistry results obtained from Alchemy and QM9 datasets are somewhat different from each other[44] due to the different biased generation routines used in these two datasets despite the same computation level of B3LYP/6-31G(2df,p) was used. We find that both Alchemy and QM9 datasets give nearly identical results for Gibbs free energy with the pearson correlation coefficient of 1.0. However, the pearson correlation coefficient is just about 0.63 for the predicted dipole moments and 0.85 for HOMO results in Alchemy with the QM9-trained DeepMoleNet model (Table S2).

2.3.2 Non-equilibrium conformation datasets

MD17. The MD17[92] contains eight organic molecules with up to 21 atoms composed of heavy atoms like C, N, O, F. For each molecule, ab initio molecular dynamics simulation was performed to obtain the energy and forces using PBE +

vdW-TS electronic structure method. At each time step, the energy and forces together with its coordination were recorded. Here, we focus on the relative conformational energies of a certain molecule in potential energy surface (PES).

ANI-1ccx. ANN-1ccx [93] is a high-quality and diverse data set which contains 500,000 molecules in both wide chemical and conformer spaces in the benchmark calculations at CCSD(T)/CBS level for isomerization energies, reaction energies, molecular torsion profiles, and energies and forces at non-equilibrium geometries with molecule size N_{atom} ranging from 2 to 54. The ANI-1ccx dataset is intelligently selected 10% sub-sample of the ANI-1x dataset, in which the molecular conformations were derived from 57,000 distinct molecular configurations containing the C, H, N and O elements computed with DFT through an active learning algorithm with four kinds of sampling methods, namely, molecular dynamics simulations, normal mode analysis, dimer sampling, and torsion sampling.

2.3.3 Other test sets

Some typical molecules, which are of great importance in drug or material design, were selected as external data set for the transferability test (Figure S2) based on the model trained with 110,000 QM9 data.

Drug-like molecules data set [37]. There are 24 drug molecules like aspirin, abacavir, vitamins, and drug-like molecules in this data set. All of them are natural products with pharmacological activity. They all contain rings with rotatable bonds and certain amount of hydrogen bond donors and acceptors in the spatial structure. The N_{atom} of these molecules ranges from 14 to 73. In this work, 17 molecules, including vitamin B3, vitamin C, aspirin, amphetamine, vitamin B5, abiraterone, cocaine, amitriptyline, testosterone, vitamin A1, vitamin B12, progesterone, abacavir, glucosepane, cholic acid, vitamin D3, and vitamin D2 were randomly selected for the test.

MPCONF196 [96]. This dataset with the data size of 196 selects 13 acyclic and cyclic model peptides and several macrocyclic compounds, with 15 or 16 conformers (in both high- and low-energy regions) for each compound, whose conformation

energies were computed by DFT(-D3) and CCSD(T)/CBS methods. The selected macrocycles are also collected in the Cambridge Structural Database (CSD) and are further denoted by their CSD codes. In this work, 5 peptides (FGG, GGF, WG, WGG, and GFA) and 6 macrocyclic molecules (with CSD codes of POXTRD, CAMVES, COHVAW, CHPSAR, Gpd_A, and Gpd_B) were randomly selected from the **MPCONF196** dataset for the transferability test of the proposed **DeepMoleNet** method in prediction of Gibbs free energy and HOMO properties.

Singlet fission molecules data set [95]. This dataset includes 262 singlet fission molecule candidates. The dataset was screened out through a procedure of exploiting quantum chemical calculations of excitation energies, which were calibrated against experimental data. The candidate molecules were stored into different chemical families, enabling the design of further singlet fission materials using the hits as lead compounds for further exploration. For simplified, the selected 9 test molecules are also named by their CSD codes, AFZPYM, OXTPTZ, BATWUO, FEFLEK, EVAWON, LULLOT04, DIFQEP, TOSYAD, and QELNOK, in our transferability test of the Gibbs free energy and some other properties.

Oligomers data set [37]. The originally reported data set selects 5 classes of oligomers, i.e., polyethylene (PE, n=28), polyacetylene (PA, n=15), polylactic acid (PLA, n=10), and alanine peptide (ala, n=10), quaternary ammonium polysulphon (bQAPS, n=3). Here, we adopt 4 kinds of oligomers with different polymerization degrees, including polyethylene (PE, n=3, 7, 14), polyacetylene (PA, n=6, 8, 15), polylactic acid (PLA, 3, 5, 10), and alanine peptide (ala, n=2, 4, 10), as our test sets.

2.4 Error analysis of predicted results

DeepMoleNet features regression tasks among 12 different electronic properties at DFT level and energy prediction at CCSD(T)/CBS calculation level. Similar to other publications, MAE and RMSE metrics were calculated for evaluating the prediction error. To further reflect the average error compared to the standard deviation of each target, we report the standardized MAE, std.$_{MAE,}$ of different properties. For each target, the std. $_{MAE}$ is defined as follows.

$$\text{std.}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Predict_m^i - DFT_m^i|}{\sigma_m} \qquad (12),$$

where $Predict_m^i$ and DFT_m^i are the machine learning predicted and actual DFT values for the property m (e.g., the dipole moment) of the i-th molecule, respectively. The term σ_m is the standard deviation from DFT result of property m. N is the number of data for the test set. In some cases, for example, to reflect the relative deviation of HOMO prediction, we also calculate the relative MAE error, Δ_{MAE}, which comes from the division of MAE by the real DFT value, as shown below.

$$\Delta_{\text{MAE}} = \left|\frac{Predict_m^i - DFT_m^i}{DFT_m^i}\right| \qquad (13).$$

3. Results and Discussion
3.1. Learning QM9/Alchemy molecular properties

In the first step, the 110,000 QM9 data are randomly selected as the training set, 10,000 data for validation and the rest for test set, with the molecular size, N_{atom}, varying from 3 to 29 atoms (Figure S1). The transferability test is then done on the randomly split 60,000 molecules in Alchemy, which contains the computational data of the 12 electronic structure properties (dipole moment, polarizability, HOMO, LUMO, HOMO-LUMO gap, ZPVE, $<R^2>$, U_0, U, H, G, and C_v) at the same B3LYP/6-31G(2df,p) level but with different calculation workflow from those data in QM9. As mentioned above, it is not practical to use QM9 trained model to directly predict the corresponding Alchemy properties. [44] Instead, we used the Pearson correlation coefficient to test the relative trends between the 12 properties for Alchemy molecules predicted by the model trained with QM9 and the real values reported in Alchemy data set (Table S2).

Our model can predict 12 molecular properties with satisfactory accuracy, as shown in Table 3 (for single-task) and Tables S3 (multi-task). Most of MAEs of our simultaneous predictions of 12 properties relative to DFT results are slightly lower than those of other multi-target models trained on MoleculeNet [84] and multi-target model of DimeNet (Table S3). Among all the 12 properties, the present model works very well for the energy-related properties, U_0, U, H, and G, with MAEs of 7.7 meV, 7.8 meV, 7.8 meV, and 8.6 meV, respectively. The performance of the prediction of

electronic structure properties, such as HOMO (MAE: 23.9 meV), LUMO (MAE: 22.7 meV), and HOMO-LUMO gap (MAE: 33.2 meV), is also satisfactory. For each target, we also calculated the mean standardized MAE, std. $_{MAE}$. As shown in Table S4 and Table S5, the calculated std. $_{MAE}$ values of the present model are about 1%, approaching the chemical accuracy. It is still a big challenge to achieve the balanced and accurate predictions for all those 12 targets simultaneously. In fact, the MAEs of our multi-target model (Table S3) is even smaller than those of other single-target models for the predictions of dipole, HOMO, HOMO-LUMO gap, U_0, U, H, and G (Table 3), which is perhaps the first report for the multi-task model to have such a performance to the best of our knowledge. The single-target prediction accuracy of 12 targets using our DeepMoleNet model is the best with small MAEs of HOMO (21.8 meV), LUMO (18.5 meV), and HOMO-LUMO gap (32.1 meV). To illustrate the statistic performance, the error bars of dipole moment, HOMO energy level, and free energy in single-target prediction and those 12 properties in multi-target task on QM9 are calculated in Table S6 and Table S7, respectively.

Table 3. MAEs of DeepMoleNet single-target training and other single-target training methods using QM9 dataset.

Property	Unit	SchNet[77]	enn-s2s[73]	MEGNET[61]	Cormorant[71]	MGCN[74]	DimeNet[43]	**DeepMoleNet**
μ	D	0.033	0.03	0.05	0.13	0.056	0.0286	**0.0178±0.0000**
α	a_0^3	0.235	0.318	0.081	0.092	**0.03**	0.0469	0.0475
ϵ_{HOMO}	meV	41	43	43	36	42.1	27.8	**21.9±0.1**
ϵ_{LUMO}	meV	34	37	44	36	57.4	19.7	**18.5**
Δ_ϵ	meV	63	69	66	60	64.2	34.8 [a]	**32.1**
$<R^2>$	a_0^2	**0.073**	1.8	0.302	0.673	0.11	0.331	0.115
ZPVE	meV	1.7	1.5	1.43	1.43	**1.12**	1.29	1.22
U_0	meV	14	19	12	28	12.9	8.02	**6.1**
U	meV	19	19	13	21	14.4	7.89	**6.1**
H	meV	14	17	12	21	14.6	8.11	**6.1**
G	meV	14	19	12	20	16.2	8.98	**7.1±0.0**
C_v	cal /mol K	0.033	0.040	0.029	0.031	0.038	0.025	**0.0241**
std. $_{MAE}$	%	1.76	-	1.37	2.14	-	1.05	**0.84**

[a] The Δ_ϵ is predicted simply by taking $\epsilon_{LUMO} - \epsilon_{HOMO}$ in DimeNet. [43]
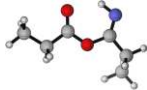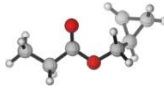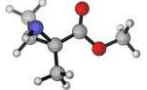
We apply the multi-target DeepMoleNet model, which was trained on 110,000 QM9 data, to predict the 12 properties of the molecules in the whole QM9 dataset. The prediction MAEs of the energy-related properties, U_0, U, H, and G, for all 130,000 molecules are just about 2.5-4.0 meV (0.058-0.092 kcal mol^{-1}). However, predictions of HOMO, LUMO, and HOMO-LUMO gap are relatively more difficult with MAEs of about 6.3-9.3 meV (Figure S3).

The difficulty in predicting HOMO and LUMO energy levels probably lies in the complexity of modulation of frontier molecular orbitals and their energy levels by different molecular topology and substitution groups. As shown in Figure S4 and Table S8, there is no evident correlation between the substituent groups and the frontier orbitals for the substituted furans at α-position (also called 2-position in this work). According to our chemical intuition, the electron-donating or accepting ability of the substituted groups may be closely related to the upshift or down shift of the HOMO and LUMO energy levels. For this purpose, we calculated the group charge, ρ_{sub}, of substituent, X= H, NH$_2$, OCH$_3$, CH$_3$, OCHO by summation of the atomic charges (Table S9) in the substituted group. However, different population analysis methods, Hirshfeld[98] and Mulliken[99-101] population, give different group charges, as shown in Figure S4. Accordingly, these two different the population analysis methods exhibit different relationship between the group charges and the frontier molecular orbital energy levels. Therefore, the factors that affecting the HOMO and LUMO energy levels are rather complicated, leading to the difficulty in the accurate prediction of frontier molecular orbitals. More examples, with or without the molecular rings, are given in Table 4. It can be seen that our method is able to accurately predict molecular properties for various systems with different topology and molecular sizes (Table 4 and Table S10).

We further tested model predictivity with model trained on 110,000 data in QM9 to make predictions on molecules in Alchemy. As mentioned above, we used the Pearson correlation coefficient to make an indirect comparison between our predicted results on the randomly sampled 60,000 molecules and the original values given in Alchemy, as shown in Table S2. Good correlations were obtained for most of the

predicted molecular properties, except for the dipole moment and polarizability properties.

Table 4. The randomly selected test sets of QM9 dataset and their absolute errors of the predicted properties with the proposed deep learning method

absolute error					
	CCC(=N)OC(=O)CC	CCC(=O)OCC1CC1	COC(=O)C1(C)NC1C	OC1CCCC1OC=N	CC1COC2C=CCC12
μ(D)	0.03	0.02	0.03	0.02	0.01
α(a_0^3)	0.07	0.00	0.06	0.02	0.00
ϵ_{HOMO} (meV)	22.8	4.9	4.5	13.5	2.2
ϵ_{LUMO} (meV)	6.2	14.5	29.6	16.8	15.8
Δ_ϵ (meV)	29.0	9.5	22.4	30.0	13.7
U (meV)	2.0	4.9	3.0	8.8	18.5
G (meV)	0.0	3.9	2.9	12.7	17.5
C_v (cal/mol K)	0.02	0.01	0.02	0.01	0.01

3.2. Learning non-equilibrium conformation energies

In most cases, especially when the molecules aggregate in the condensed phase or take place chemical reactions at room or even higher temperature, molecules are not always staying in their lowest energy states with the equilibrium geometry. As shown in Table 5, we adopt eight organic molecules, i.e., aspirin, benzene, ethanol, malonaldehyde, naphthalene, salicylic acid, tolunene, and uracil, with up to $N_{atom} = 21$ collected in the MD17[92] dataset to test the predicting ability of DeepMoleNet toward conformational energies in the non-equilibrium molecular configuration space. For each molecule, a training, validation, and test split scheme is set as 50,000, 10,000, and 10,000 molecules, respectively. A satisfactory prediction was achieved by using our DeepMoleNet model with MAE values of less than 0.08 kcal/mol for the selected molecules in MD17[71, 92] dataset. The MAE values of other models using the same training numbers of 50,000 points are also listed in Table 5. It should be mentioned that different models were trained with different training set numbers as well as different approaches to get the energies. The SchNet[77] was trained on energies, and

both energies and forces with 50k for training and 1k for validation, respectively. Both DTNN[78] and DeepMoleNet were only trained with energies on 50,000 points with 1k for DTNN as validation set and 10k for DeepMoleNet, respectively. Some models were trained only on a small number of configurations. For example, sGDML[71, 92] was trained with 1,000 point (validation was part of the training set) on forces. But some other model like the DeepMD[102] used a much larger training set of 95,000 points (validation was part of the training set) for the training on both energies and forces. In other words, one can get low error with 1k points used for sGDML which is comparable to errors obtained with other models trained on 50k+ points.

We choose one molecule, aspirin, of the MD17 collected molecules, to draw the PES of the AIMD sampled conformers, which were characterized by two flexible dihedral angles, ϕ, and φ, shown in Figure 3. The AIMD potential energy surface is well reproduced by DeepMoleNet predictions, indicating the applicability of deep learning method to predict reaction barriers of chemical reactions. We carried out three independent calculations with the MAE values of 0.662, 0.665, 0.661, respectively. The MAE with the standard error of aspirin is thus presented as 0.07 ± 0.00 (Table 5). However, DeepMoleNet was trained only with energies, without force. Thus, the present model is not able to perform dynamics simulations. Further work is desired to extend the DeepMoleNet to the force predictions of non-equilibrium configurations.

Table 5. MAE values of various machine learning methods using the different training numbers points for the conformation energies (in units of kcal/mol) of the selected organic molecules in MD17 dataset.

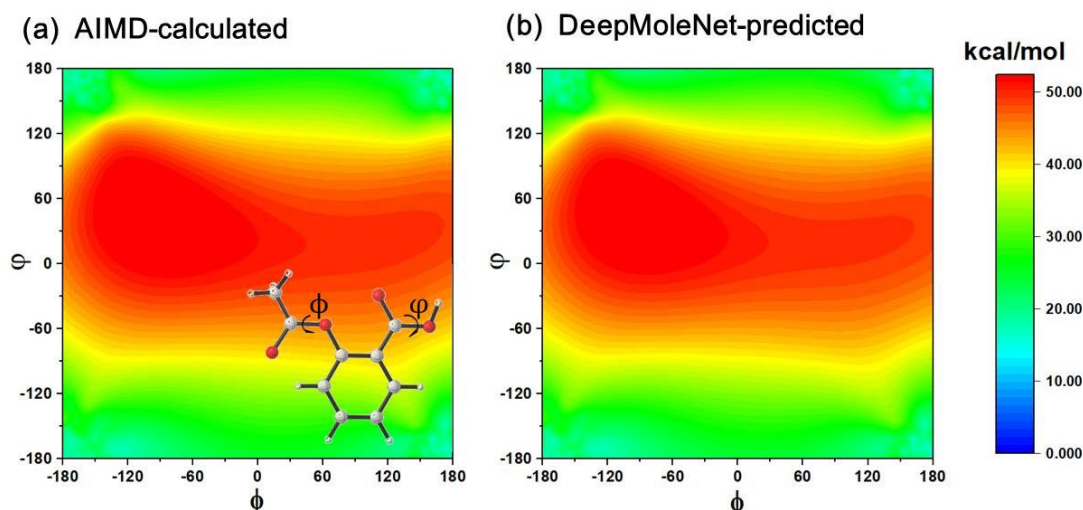| Test species (N_{atom})	sGDML[71, 92, 103] trained with force	SchNet[77] trained with force and energy	DeepMD[102] trained with force and energy	SchNet[77] trained with energy	DTNN[78] trained with energy	DeepMoleNet trained with energy
Training size/validation size	1k	50k/1k	95k	50k/1k	50k/1k	50k/10k
Aspirin (21)	0.19	0.12	0.20	0.25	-	**0.07 ± 0.00**
Benzene (12)	0.10	0.07	0.07	0.08	0.04	**0.02**
Ethanol (9)	0.07	0.05	0.06	0.07	-	**0.02**
Malonaldehyde (9)	0.10	0.08	0.09	0.13	0.19	**0.03**
Naphthalene (18)	0.12	0.10	0.10	0.20	-	**0.08**
Salicylic Acid (15)	0.12	0.11	0.11	0.25	0.41	**0.07**
Tolunene (15)	0.10	0.09	0.09	0.16	0.18	**0.05**
Uracil (12)	0.11	0.10	0.09	0.14	-	**0.06**

Figure 3. (a) AIMD calculated versus (b) DeepMoleNet predicted potential energy surfaces of aspirin, whose conformations are characterized by two flexible dihedral angles, ϕ, and φ, shown in inset.

To further test DeepMoleNet at a more accurate theoretical level like CCDS(T)/CBS, the molecules in ANN-1ccx [93, 94] with molecular size, N_{atom}, ranging from 2 to 54 were also studied, with the prediction errors shown in Table 6. DeepMoleNet achieves comparable accuracy in comparison with ANI-1ccx results. Being limited by the computational costs and some really challenging conformations for DeepMoleNet model, we only use 280,000 data (about half of the whole ANI-1ccx dataset) to train our model. The well and poorly predicted systems by our DeepMoleNet model are presented in Figure S5, from which one can find that the MAE values of the most of the acyclic or aromatic cyclic systems are close to 0.0 kcal/mol but the nitrogen-containing five or seven membered ring (5-MR, 7-MR) heterocyclic molecules are not predicted well. Improving the predictive performance of molecules with complex chemical environments will be an important direction in our future works.

Table 6. Prediction errors, in units of kcal/mol, on ANI-1ccx dataset with different ANI models (ANI-1ccx model and ANI-1ccx-R model) and our DeepMoleNet model.

	ANI-1ccx [94,a]	ANI-1ccx-R [94,b]	DeepMoleNet[c]
MAE	**1.8**	2.3	2.3
RMSE	**2.6**	3.3	3.3

[a] training based on ANI-1x 5 M data points and 500,000 ANI-1ccx data points through transfer learning and only considers conformations from 8x model ensemble.

[b] training with ANI-1ccx 500,000 data points and only considers conformations from 8x model ensemble. The data were taken from literature. [94]

[c] training with ANI-1ccx 280,000 data points

3.3. Transferability test to the larger sized molecules

It is very useful to use small data samples for training to predict the data of larger sized molecules. We know that the DFT computation time increases with the scaling of $O(N_b^3)$, while the increase in prediction time for the machine learning model is modest. The scaling of DeepMoleNet is $O(N_{atom}^2)$. The QM9 data set only contains molecules with the size ranging from $3 \leq N_{atom} \leq 29$, so the Alchemy dataset which contains relatively larger molecules (with up to 38 atoms) was introduced for further transferability test. A series of tests were conducted with small molecules in the training dataset while big molecules for tests on these two datasets, respectively. In the first step, we sampled all molecules with $N_{atom} \leq 18$ (about 72,000 data) for training. Then we made tests on molecules with larger N_{atom}, as shown in Figure 4 and Figure S6. The MAE is satisfactory with the training on 18-atoms molecules and testing on 29-atom ones, although the error slightly accumulates with the size increasing. Due to the data sparsity, we could use all 72,000 data for training. Accuracy could be improved if the training data size is big enough. Even though, it is still a big challenge for realizing good transferability.

In Alchemy dataset, molecules with N_{atom} less than 23 (instead of $N_{atom} \leq 18$ in QM9) were used for training to make sure that same percentage of about 50% training samples in each dataset is used as QM9 to make fair comparison of DeepMoleNet. As shown in Figure 5, a good transferability is still held on prediction of free energy G,

HOMO, LUMO, and HOMO-LUMO gap. Predictions of all 12 properties could be found in Table S11 and Table S12. It should be mentioned that the chemical space of the medium sized molecules in Alchemy is not as rich as QM9 so the prediction errors are larger than those test for QM9.
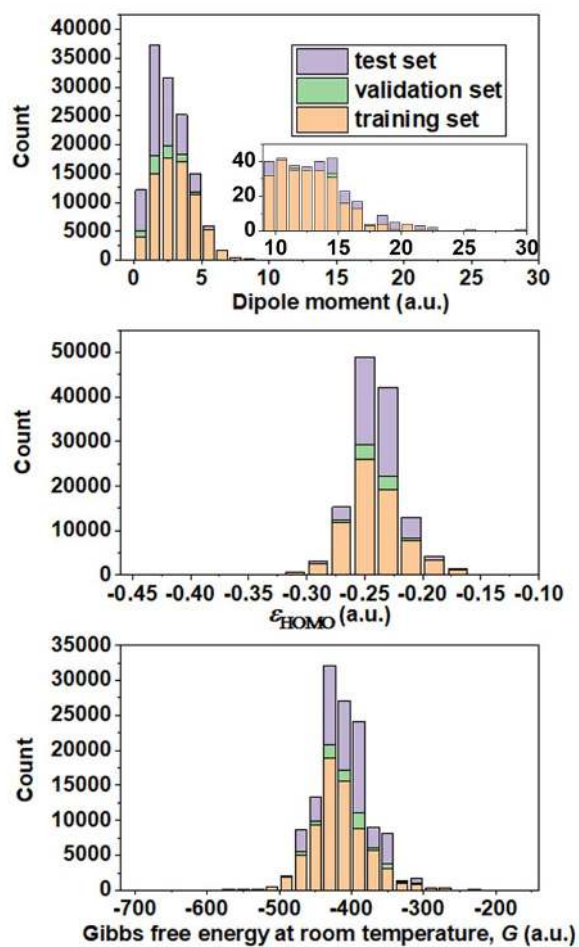

Figure 4. Data distribution of dipole moment, HOMO, and Gibbs free energy, *G*, with training set of 72,367 molecules, validation set of 8,000 molecules, and test data of 49,061 molecules in QM9 on transferability test.
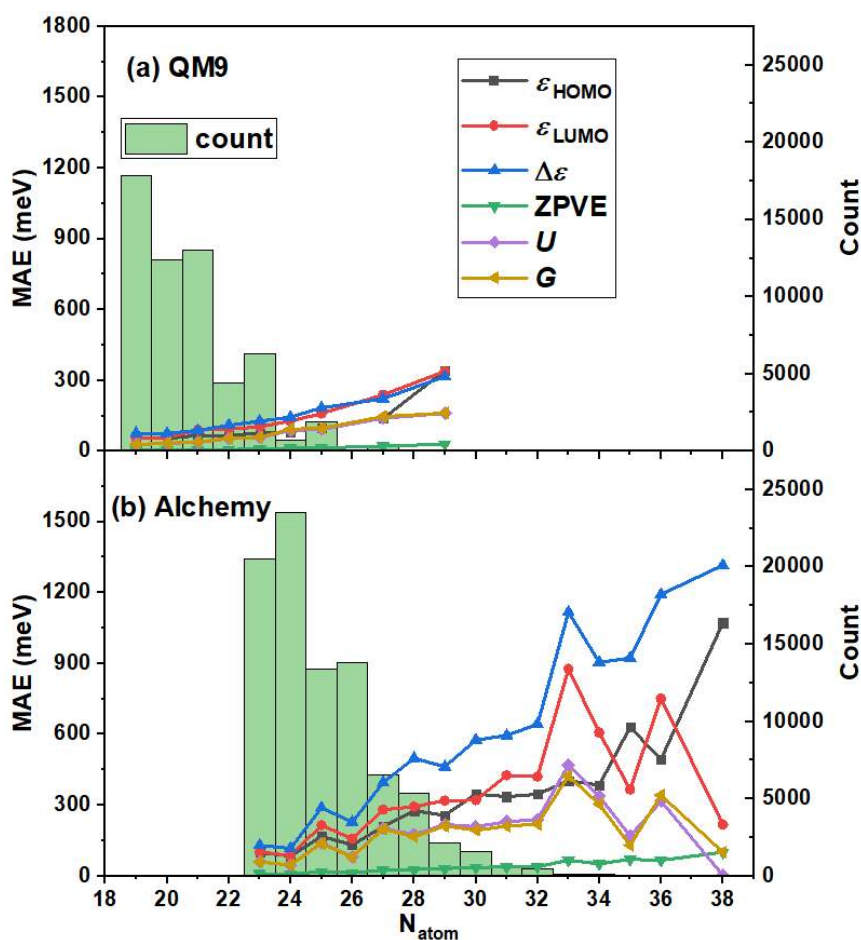
Figure 5. Transferability on QM9 trained with $N_{atom} < 19$ and Alchemy with $N_{atom} < 23$ to predict larger molecules in each dataset, respectively.

In the next step, we carried out more transferability tests on many other molecules including drug molecules, peptides and macrocycles, oligomers, protein, and singlet fission molecules, using deep learning model trained with 110,000 samples of $N_{atom} \leq 29$ in QM9, as shown in Figure 6. All the selected test molecules were optimized under the B3LYP/6-31G(2df,p) level. Among those test molecules, the largest one, chignolin protein has 140 atoms, much larger than the biggest trained molecule which only consists of 29 atoms in the training set. Good correlations are indicated for all the studied chemical systems between the calculated DFT Gibbs free energies and the DeepMoleNet predicted results

(Figure 6, Figure S2, and Figure S7). The MAE for the whole test set is about 0.11 eV/atom.
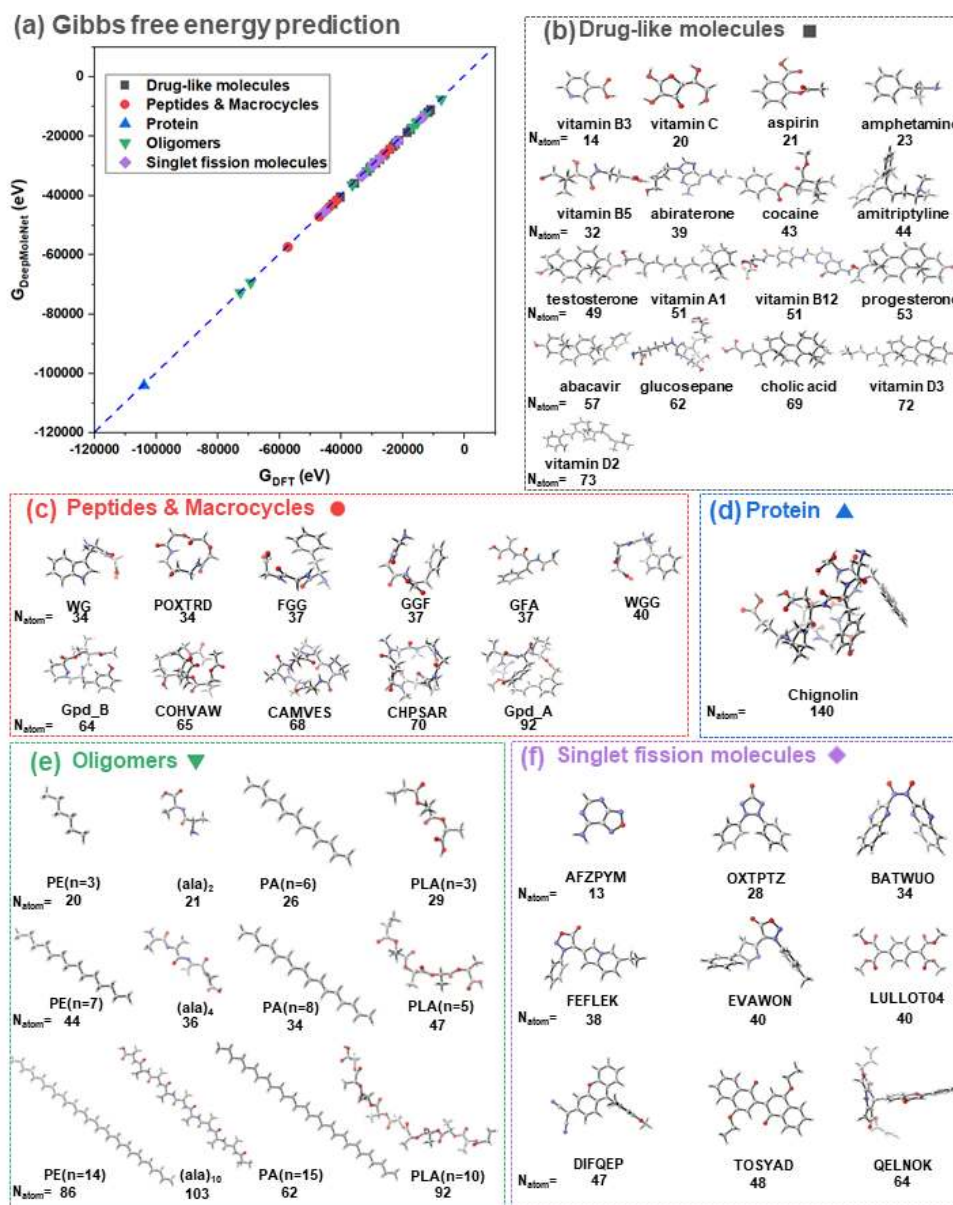

Figure 6. (a) Transferability test on Gibbs free energy of a wide range of functional molecules, including (b) drug-like species, (c) peptides and macrocycles, (d) chignolin protein, (e) medium-sized oligomers, and (e) singlet fission molecules based on the model trained with 110,000 QM9 data ($N_{atom} \leq 29$)

The prediction of HOMO remains a big challenge for the test molecules.

Among them, some natural products and PLA oligomers have relatively small prediction errors within 5%, as exemplified in Figure 7. In fact, the relationship between MAE and N_{atom} is not evident for the prediction of electronic structure property such as the HOMO energy level. However, for some π-conjugated systems, such as, semiconducting PA oligomers (n=15), vitamin B12, and singlet fission molecule, TOSYAD, the prediction power of DeepMoleNet is significantly decreased (Figure S8). The difficulty in the prediction of the electron delocalized systems containing the naphthalene or anthracene ring and long π-conjugated chain in PA may be caused by the sparse in the chemical space of delocalized systems in the training set. On the other hand, DeepMoleNet uses the one-hot encoding to represent the molecular path length. In this case, if the molecules have larger sized ring structure than the trained molecules, these systems cannot be accurately described even in the input.

	vitamin B5	POXTRD	WGG	COHVAW
$N_{atom}=$	32	34	40	65
$\varepsilon_{HOMO}=$	-6.56 eV	-6.45 eV	-6.31 eV	-6.99 eV
$\Delta_{MAE}=$	1.23%	2.15%	2.48%	0.83%

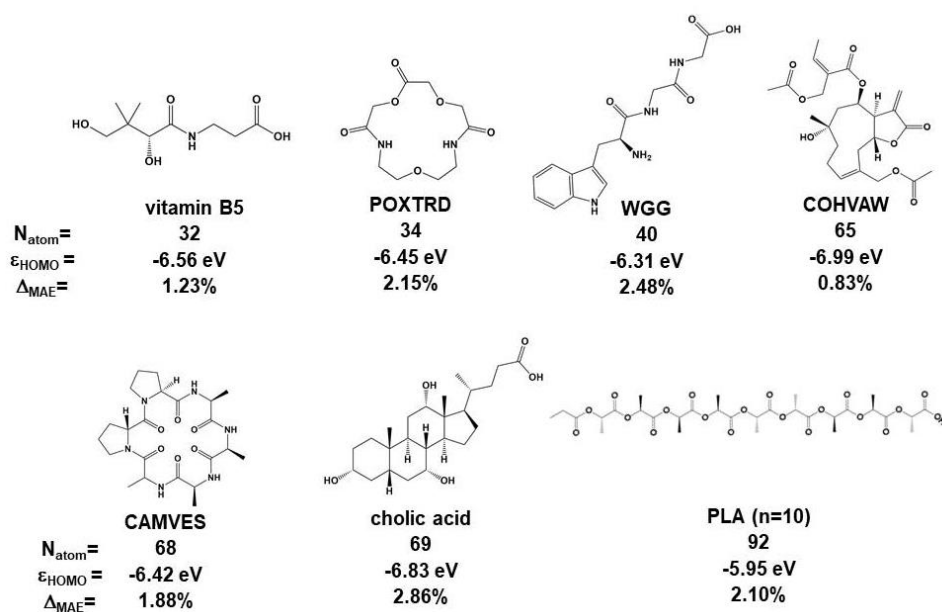|	CAMVES	cholic acid	PLA (n=10)
$N_{atom}=$	68	69	92
$\varepsilon_{HOMO}=$	-6.42 eV	-6.83 eV	-5.95 eV
$\Delta_{MAE}=$	1.88%	2.86%	2.10%

Figure 7. Transferability test on HOMO energy levels of some well predicted molecules based on the model trained with 110,000 QM9 data ($N_{atom} \leq 29$)

3.4. Factors that influence the performance of deep learning model

Steps of message passing. Multi-level attention mechanism has been widely used in many areas including urban computing, computer vision, and natural language processing [104-106]. In this work, the T steps of message passing local attention were performed after each aggregation of different node feature levels, leading to the importance variance after each step. At the different T steps, the node is affected by the neighbors, as exemplified by the α- and β-substituted furans with T=4 and T=6, respectively, in Figure 8. The attention tends to give large weights (labelled in red color) to non-hydrogen heavy atoms at the beginning, and then gradually diffuses to the surrounding H atoms. The weights of the seemingly 'less-important' H atoms gradually increase, and finally climb to 1. These coefficients are learned by the neural network itself. It can be found that the setting of T=6 is sufficient to get satisfactory prediction results.
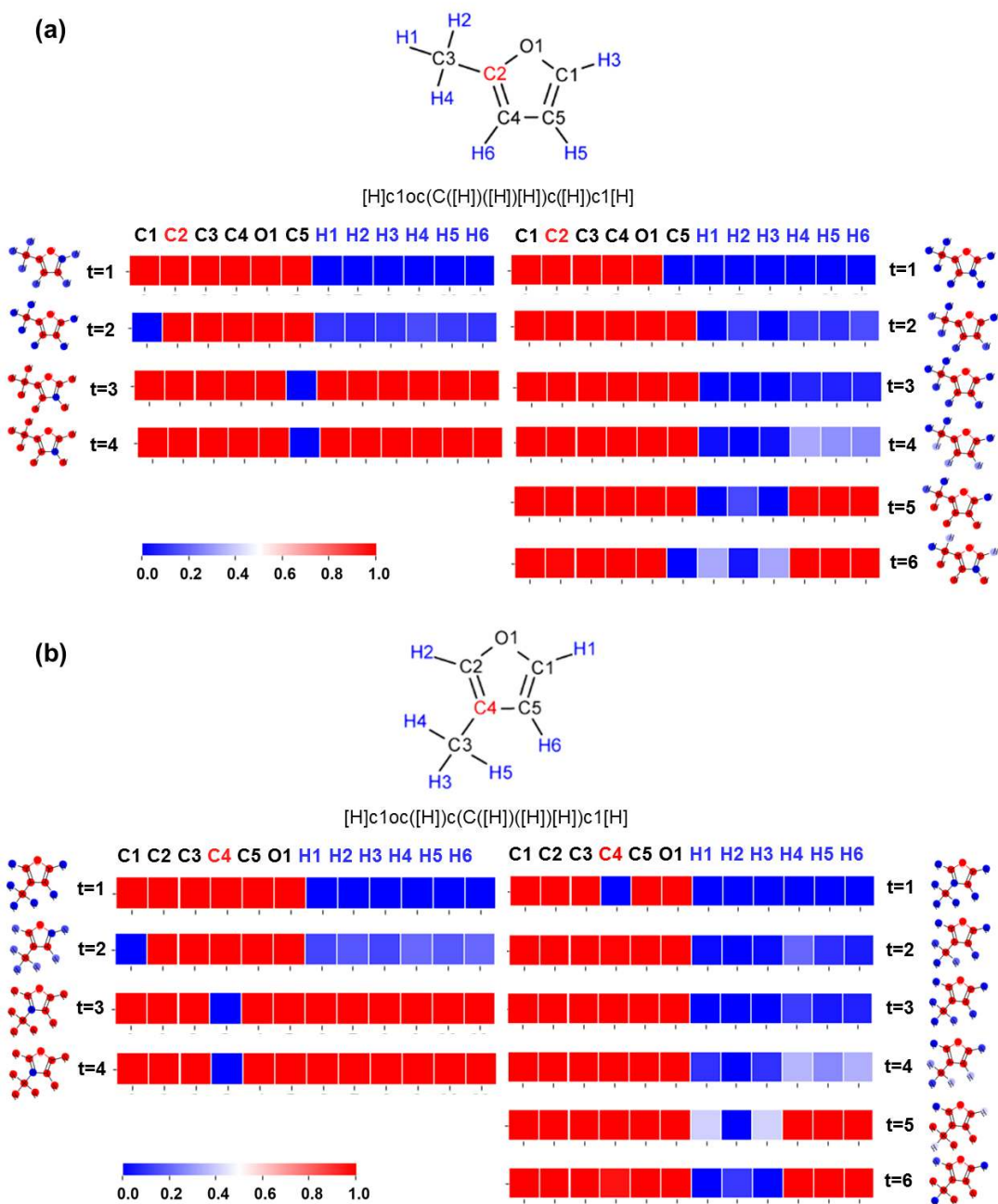
Figure 8. Illustration of multi-level attention mechanism in different message passing phase. The color indicates the relative importance. The red color corresponds to the importance of 1 and the blue color indicates the importance of 0, respectively.

ACSFs as the auxiliary teaching descriptor. For the i-th atom, ACSFs are composed of three types of radial functions, $G_i^{1,z_1}, G_i^{2,z_1}, G_i^{3,z_1}$, as shown in equation (14), equation (15), equation (16), and two types of angular functions, G_i^{4,z_1,z_2} and G_i^{5,z_1,z_2}, in equation (17) and equation (18), respectively. ACSFs of the i-th atom are presented as follows, where the summation for j runs over all atoms with atomic number z_1 in two-body term, and for the three-body term, summations for j and k run over all atoms with atomic number z_1 and z_2, respectively.

$$G_i^{1,z_1} = \sum_j^{z_1} f_c(R_{ij}) \tag{14}$$

$$G_i^{2,z_1} = \sum_j^{z_1} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{15}$$

$$G_i^{3,z_1} = \sum_j^{z_1} \cos(\kappa R_{ij}) f_c(R_{ij}) \tag{16}$$

$$G_i^{4,z_1,z_2} = 2^{1-\zeta} \sum_{j\neq i}^{z_1} \sum_{k\neq i}^{z_2} (1 +$$

$$\lambda \cos \theta_{ijk})^\zeta \cdot e^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)^2} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{17}$$

$$G_i^{5,z_1,z_2} = 2^{1-\zeta} \sum_{j\neq i}^{z_1} \sum_{k\neq i}^{z_2} (1 + \lambda \cos \theta_{ijk})^\zeta \cdot e^{-\eta\left(R_{ij}^2+R_{ik}^2\right)^2} f_c(R_{ij}) f_c(R_{ik}) \tag{18}$$

In the above equations, the R_{ij} is inter-atom distance between atoms i and j. In the three-body angular function, the angle θ_{ijk} between the three atoms (with the i-th atom in the center) is calculated by

$$\theta_{ijk} = acos \frac{R_{ij} \cdot R_{ik}}{R_{ij} \cdot R_{jk}} \tag{19}.$$

The inter-atom distance related term, f_c, is the cutoff symmetry function, which is defined as,

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\pi \frac{R_{ij}}{R_c}\right) + 1\right] & if\ R_{ij} \leq R_c \\ 0 & if\ R_{ij} > R_c \end{cases} \tag{20},$$

where R_c is a cutoff radius, which is set to be 6.0 here.

In the functions of $G_i^{2,z_1}, G_i^{3,z_1}, G_i^{4,z_1,z_2}$ and G_i^{5,z_1,z_2}, η , R_s, κ, ζ , and λ are user-defined parameters. We applied G_i^{2,z_1} and G_i^{4,z_1,z_2} functions in the following calculations. Two $-$ body G_i^{2,z_1} values are calculated by using different (η, R_s)

parameters, (1,1), (1,2), and (1,3).   Three-body G_i^{4,z_1,z_2} is obtained by triplet (η, ζ, λ) parameter sets of (1,1,1), (1,2,1), (1,1,-1), and (1,2,-1).

To figure out which operation is important in improving the accuracy of molecular property predictions, in Table 7 we made comparison between **'level A'**, i.e., DeepMoleNet with ACSFs as learning input descriptors with both multi-level attention and auxiliary learn ACSFs target (not as input), and those cases with multi-level attention but without the auxiliary ACSFs target (**level B**) and without both multi-level attention and the auxiliary ACSFs target (**level C**), respectively. As expected, the knowledge of attention mapping to the atoms in **level A** and **level B** could improve the prediction accuracy for HOMO and LUMO energy levels, relative to those obtained at **level C** without attention.

Specifically, the use of auxiliary target ACSFs in **level A** is crucial to get good predictions for the dipole moment, HOMO, LUMO, HOMO-LUMO gap, U_0, U, H, and G. ACSFs have wide applications in constructing potential energy surfaces (PESs) and properties prediction including catalysis, reactions, phase transition, etc. [53-58] The main idea of ACSFs is to represent a chemical system's geometry with symmetrized invariant functions as input descriptors. To the best of our knowledge, it is the first attempt to set the ACSFs to be one of the prediction targets in this work. Ablation studies are further carried out to find out which information of ACSFs is helpful for the final node hidden states to learn better representation. We make comparisons among different combination of various test cases, including that only applying two-body radial functions (learn G_i^{2,z_1} without using the angular functions G_i^{4,z_1,z_2}), only three-body angular functions (learn G_i^{4,z_1,z_2} without using radial functions G_i^{2,z_1}), and both radial and angular functions (learn both G_i^{2,z_1} and G_i^{4,z_1,z_2}) as the auxiliary targets, respectively. This ablation study indicates that ACSFs information could not be learned by the data-driven manner of the topological message passing. Instead, it can be obtained by predicting ACSFs. Directly using ACSFs as input in the message passing phase obtains the unsatisfactory prediction results. This suggests that in deep learning, ACSFs may be not necessarily needed to function as the input descriptors.

Instead, When ACSFs are taken as one of the prediction targets, a boost in the model performance is observed in Table 7.

One can also find from Table 7 that using either the radial functions (G_i^{2,z_1}) or the angular functions (G_i^{4,z_1,z_2}) in ACSFs could achieve better performance than that without using ACSFs. Especially for the prediction of the dipole moments, the employment of radial or angular functions is crucial to greatly improve the accuracy, implying that the structural information is important for predicting properties sensitive to structural distance. Moreover, the introduction of three-body angular functions (G_i^{4,z_1,z_2}) is able to give better predictions than that only using the two-body radial functions (G_i^{2,z_1}). For the most challenging prediction tasks of the HOMO, LUMO, and HOMO-LUMO gap, the solely use of two-body radial functions is insufficient. Fortunately, the combination of radial and three-body angular functions in **level A** could greatly improve the prediction accuracy for HOMO, LUMO, and HOMO-LUMO gap. This displays the importance of the many-body interactions in learning better node representations, and hence, properly predicting electronic structure properties. It is still not clear why **level A** scheme would work. Further exploration is going on in our laboratory. The error bars of the estimated properties in Table 7 were obtained from three independent calculations, which were listed in Table S7.

Table 7. The MAE values of QM9 predictions made by DeepMoleNet multi-target training with both multi-level attention and auxiliary target ACSFs (**level A**), in comparison between those cases with ACSFS as inputs (**level B**), without multi-level attention and auxiliary target ACSFs (**level C**), and without auxiliary target ACSFs, respectively.

Property	Unit	Without using attention and ACSFs (level C)	Using attention		Using ACSFs as prediction task (level A)[a]		
			without using ACSFs	ACSFs as inputs (level B)	learn G_i^{2,z_1} w/o G_i^{4,z_1,z_2}	learn G_i^{4,z_1,z_2} w/o G_i^{2,z_1}	**DeepMoleNet** G_i^{4,z_1,z_2} & G_i^{2,z_1}
μ	D	0.061	0.0438	0.499	0.0265	0.0256	0.0267±0.0014
α	a_0^3	0.109	0.0802	0.476	0.0731	0.0706	0.0683±0.0002
ϵ_{HOMO}	meV	32.1	25.4	146.4	26.1	24.5	24.0±0.1
ϵ_{LUMO}	meV	31.8	24.7	153.6	24.6	22.9	22.8±0.1
Δ_ϵ	meV	44.5	35.4	204.8	35.7	33.6	33.2±0.0
$<R^2>$	a_0^2	1.86	1.58	6.03	0.91	0.86	0.69±0.01
ZPVE	meV	3.2	4.0	13.1	2.5	2.4	1.9±0.0
U_0	meV	12.8	11.0	65.7	8.6	8.2	7.6±0.1
U	meV	12.9	11.1	66.1	8.6	8.3	7.7±0.1
H	meV	12.9	11.0	66.1	8.7	8.3	7.7±0.1
G	meV	13.4	11.7	65.8	9.7	9.2	8.5±0.1
C_v	cal /mol K	0.0459	0.0376	0.1583	0.0313	0.0299	0.0293±0.0002
std. MAE	%	1.55	1.22	8.24	1.09	1.03	1.01±0.01

[a] We investigated three cases, i.e., only applying two-body radial functions (learn G_i^{2,z_1} without using the angular functions G_i^{4,z_1,z_2}), only three-body angular functions (learn G_i^{4,z_1,z_2} without using radial functions G_i^{2,z_1}), and both radial and angular functions (learn both G_i^{2,z_1} and G_i^{4,z_1,z_2}) as the auxiliary targets, respectively.

As mentioned above, the using of ACSFs has an advantage in representing the local neighboring environment of an atom by using a fingerprint, which is composed of the output of several two-body and three-body functions that can be customized to detect specific structural features. Surprisingly, we found that its prediction error is closely sensitive to the number of atoms, N_{atom}, and the square of the number of atoms, $(N_{atom})^2$, with the Pearson correlation coefficients of 0.93 and 0.87, respectively (Figure S9). In the process of message passing, every added atom will affect other atoms and increase prediction errors through many-body interactions, which should be considered in our future work.

4. Conclusion

In summary, the DeepMoleNet model provides an efficient way to generate node representation that could feel chemical environments with multi-level attention and auxiliary target ACSFs. Source codes[107,108] were available on both https://github.com/Frank-LIU-520/DeepMoleNet and http://106.15.196.160:5659 for academic use. The website of DeepMoleNet and login page are illustrated in Figures S10-S14. DeepMoleNet can accurately predict electronic structure properties. Furthermore, the DeepMoleNet model can be generalized to larger sized molecules than the training molecules. It should be mentioned that QM9 is just used as the dataset to test the graph neural network model performance. We will do more calculations in our future work for making comparisons including confidence intervals, tests for significance, etc.

Indeed, the HOMO/LUMO prediction is still a challenge for the large-sized molecules. DeepMoleNet could be improved in some aspects. The present model needs to calculate the ACSFs for each atom. The combination of ACSFs knowledge with deep multi-level attention is crucial to get good accuracy. In addition, choosing the appropriate chemical descriptors as the auxiliary task requires a deep understanding of the problem. It is also a pity that when the information of atomic nodes is aggregated, the surrounding structure cannot be well touched. What is learned is an isolated node embedding information. Further improvement is anticipated to use less sample data with the goal of reaching similar prediction ability. The multi-level attention neural network will become an efficient tool to accelerate rational designs of functional molecules and chemical reactions.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at the attachment. Details of data distribution, standard deviation MAE, single-target results, multi-target prediction for all the QM9 dataset, multi-target prediction on different molecules, transferability test results for both QM9 and

Alchemy datasets, and instruction for using the source codes of DeepMoleNet.

AUTHOR INFORMATION

Corresponding Author

*E-mail: majing@nju.edu.cn (Prof. Jing Ma);
ywguo@nju.edu.cn (Prof. Yanwen Guo)

ORCID

Jing Ma: 0000-0001-5848-9775

Notes

The authors declare no competing financial interest.

References

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

2.   Chen, P.; Tang, Z.; Zeng, Z.; Hu, X.; Xiao, L.; Liu, Y.; Qian, X.; Deng, C.; Huang, R.; Zhang, J.; Bi, Y.; Lin, R.; Zhou, Y.; Liao, H.; Zhou, D.; Wang, C.; Lin, W., Machine-Learning-Guided Morphology Engineering of Nanoscale Metal-Organic Frameworks. *Matter* **2020**, 2, 1651-1666.

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

4.   Kang, B.; Ceder, G., Battery materials for ultrafast charging and discharging. *Nature* **2009**, 458, 190-3.

5.   Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H., Towards the computational design of solid catalysts. *Nature Chemistry* **2009**, 1, 37-46.

6.   Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F., A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, 6, 1379-1390.

7.   Tkatchenko, A., Machine learning for chemical discovery. *Nat Commun* **2020**, 11, 4125.

8.   Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A., Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* **2019**, 37, 1038-1040.

9.   Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P., The combined fragmentation and systematic molecular fragmentation methods. *Acc Chem Res* **2014**, 47, 2776-85.

10.   Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V., Fragmentation methods: a route to accurate calculations on large systems. *Chem Rev* **2012**, 112, 632-72.

11.   Li, S.; Li, W.; Ma, J., Generalized energy-based fragmentation approach and its applications to macromolecules and molecular aggregates. *Acc Chem Res* **2014**, 47, 2712-20.

12.   Li, S. H.; Ma, J.; Jiang, Y. S., Linear scaling local correlation approach for solving the coupled cluster equations of large systems. *J Comput Chem* **2002**, 23, 237-244.

13.   Schutz, M., A new, fast, semi-direct implementation of linear scaling local coupled cluster theory. *Physical Chemistry Chemical Physics* **2002**, 4, 3941-3947.

14.   Scuseria, G. E.; Ayala, P. Y., Linear scaling coupled cluster and perturbation theories in the atomic orbital basis. *Journal of Chemical Physics* **1999**, 111, 8330-8343.

15.   Yang, J.; Kurashige, Y.; Manby, F. R.; Chan, G. K. L., Tensor factorizations of local second-order Moller-Plesset theory. *Journal of Chemical Physics* **2011**, 134, 044123.

16.   Cheng, Z.; Zhao, D.; Ma, J.; Li, W.; Li, S., An On-the-Fly Approach to Construct Generalized Energy-Based Fragmentation Machine Learning Force Fields of Complex Systems. *J Phys Chem A* **2020**, 5007-5014.

17.   Agrawal, A.; Choudhary, A., Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials* **2016**, 4, 053208.

18.   Calderon, C. E.; Plata, J. J.; Toher, C.; Oses, C.; Levy, O.; Fornari, M.; Natan, A.; Mehl, M. J.; Hart, G.; Nardelli, M. B., The AFLOW standard for high-throughput materials science calculations. *Computational Materials Science* **2015**, 108, 233-238.

19.   Artrith, N.; Urban, A.; Ceder, G., Constructing first-principles phase diagrams of amorphous LixSi

using machine-learning-assisted sampling with an evolutionary algorithm. *Journal of Chemical Physics* **2018**, 148, 241711.

20. Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J., Neural-Network Models of Potential-Energy Surfaces. *Journal of Chemical Physics* **1995**, 103, 4129-4137.

21. Chen, W. K.; Liu, X. Y.; Fang, W. H.; Dral, P. O.; Cui, G. L., Deep Learning for Nonadiabatic Excited-State Dynamics. *Journal of Physical Chemistry Letters* **2018**, 9, 6702-6708.

22. Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schutt, K. T.; Muller, K. R., Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, 3, 1603015.

23. Dral, P. O., Quantum Chemistry in the Age of Machine Learning. *Journal of Physical Chemistry Letters* **2020**, 11, 2336-2347.

24. Gastegger, M.; Behler, J.; Marquetand, P., Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science* **2017**, 8, 6924-6935.

25. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *Journal of Physical Chemistry Letters* **2015**, 6, 2326-2331.

26. Hu, D. P.; Xie, Y.; Li, X. S.; Li, L. Y.; Lan, Z. G., Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *Journal of Physical Chemistry Letters* **2018**, 9, 2725-2732.

27. Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R., A universal strategy for the creation of machine learning-based atomistic force fields. *Npj Computational Materials* **2017**, 3, 37.

28. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *Apl Materials* **2013**, 1, 011002.

29. Le, H. M.; Sau, H.; Raff, L. M., Molecular dissociation of hydrogen peroxide (HOOH) on a neural network ab initio potential surface with a new configuration sampling method involving gradient fitting. *Journal of Chemical Physics* **2009**, 131, 014107.

30. Li, Z. W.; Kermode, J. R.; De Vita, A., Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Physical Review Letters* **2015**, 114, 096405.

31. Prudente, F. V.; Acioli, P. H.; Neto, J. J. S., The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of H-3(+). *Journal of Chemical Physics* **1998**, 109, 8801-8808.

32. Richings, G. W.; Habershon, S., Direct Quantum Dynamics Using Grid-Based Wave Function Propagation and Machine-Learned Potential Energy Surfaces. *Journal of Chemical Theory and Computation* **2017**, 13, 4012-4024.

33. Shen, L.; Yang, W. T., Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *Journal of Chemical Theory and Computation* **2018**, 14, 1442-1455.

34. Witkoskie, J. B.; Doren, D. J., Neural network models of potential energy surfaces: Prototypical examples. *Journal of Chemical Theory and Computation* **2005**, 1, 14-23.

35. Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J., The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical Science* **2018**, 9, 2261-2269.

36. Zhang, Y. L.; Zhou, X. Y.; Jiang, B., Bridging the Gap between Direct Dynamics and Globally Accurate Reactive Potential Energy Surfaces Using Neural Networks. *Journal of Physical Chemistry*

Letters **2019**, 10, 1185-1191.

37. Huang, B.; von Lilienfeld, O. A., Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat Chem* **2020**, 12, 945-951.

38. Sahu, H.; Rao, W.; Troisi, A.; Ma, H., Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials* **2018**, 8, 1801032.

39. Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A., Jr.; Ceriotti, M., Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc Natl Acad Sci U S A* **2019**, 116, 3401-3406.

40. Xu, F.; Lee, C. h.; Koo, C. M.; Jung, C., Effect of electronic spatial extents (ESE) of ions on overpotential of lithium ion capacitors. *Electrochimica Acta* **2014**, 115, 234-238.

41. Agne, M. T.; Voorhees, P. W.; Snyder, G. J., Phase Transformation Contributions to Heat Capacity and Impact on Thermal Diffusivity, Thermal Conductivity, and Thermoelectric Performance. *Adv Mater* **2019**, 31, 1902980.

42. Zorębski, E.; Zorębski, M.; Dzida, M.; Goodrich, P.; Jacquemin, J., Isobaric and Isochoric Heat Capacities of Imidazolium-Based and Pyrrolidinium-Based Ionic Liquids as a Function of Temperature: Modeling of Isobaric Heat Capacity. *Industrial & Engineering Chemistry Research* **2017**, 56, 2592-2606.

43. Klicpera, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. In International Conference on Learning Representations, 2020; 2020.

44. Tsubaki, M.; Mizoguchi, T., Fast and Accurate Molecular Property Prediction: Learning Atomic Interactions and Potentials with Neural Networks. *J Phys Chem Lett* **2018**, 9, 5733-5741.

45. Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J. Alchemy: A quantum chemistry dataset for benchmarking ai models. In International Conference on Learning Representations, 2019; 2019.

46. Bahdanau, D.; Cho, K.; Bengio, Y., Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* **2014**.

47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, 2017; 2017; pp 5998-6008.

48. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, 2015; 2015; pp 2048-2057.

49. Argyriou, A.; Evgeniou, T.; Pontil, M. Multi-task feature learning. In Advances in neural information processing systems, 2006; 2006; pp 41-48.

50. Bansal, T.; Belanger, D.; McCallum, A. Ask the gru: Multi-task learning for deep text recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems, 2016; 2016; pp 107-114.

51. Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V., A Survey of Multi-task Learning Methods in Chemoinformatics. *Mol Inform* **2019**, 38, 1800108.

52. Behler, J., Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J Chem Phys* **2011**, 134, 074106.

53. Behler, J.; Martonak, R.; Donadio, D.; Parrinello, M., Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Physical Review Letters* **2008**, 100, 185501.

54. Gastegger, M.; Marquetand, P., High-Dimensional Neural Network Potentials for Organic

Reactions and an Improved Training Algorithm. *Journal of Chemical Theory and Computation* **2015**, 11, 2187-2198.

55.   Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S., Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials* **2018**, 4, 37.

56.   Morawietz, T.; Behler, J., A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van der Waals Corrections. *J Phys Chem A* **2013**, 117, 7356-7366.

57.   Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J., How van der Waals interactions determine the unique properties of water. *P Natl Acad Sci USA* **2016**, 113, 8368-8373.

58.   Natarajan, S. K.; Behler, J., Neural network molecular dynamics simulations of solid-liquid interfaces: water at low-index copper surfaces. *Physical Chemistry Chemical Physics* **2016**, 18, 28704-28725.

59.   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.

60.   Landrum, G., RDKit: Open-source cheminformatics. **2006**.

61.   Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials* **2019**, 31, 3564-3572.

62.   Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, 50, 742-754.

63.   Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. *Physical Review B* **2013**, 87, 184115.

64.   Bartok, A. P.; Payne, M. C.; Kondor, R.; Csanyi, G., Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys Rev Lett* **2010**, 104, 136403.

65.   Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* **2012**, 108, 058301.

66.   Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A., Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J Chem Theory Comput* **2017**, 13, 5255-5264.

67.   Bartok, A. P.; Kondor, R.; Csanyi, G., On representing chemical environments. *Physical Review B* **2013**, 87, 184115.

68.   Ramakrishnan, R.; von Lilienfeld, O. A., Many Molecular Properties from One Kernel in Chemical Space. *Chimia* **2015**, 69, 182-186.

69.   von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A., Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int J Quantum Chem* **2015**, 115, 1084-1093.

70.   LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *nature* **2015**, 521, 436-444.

71.   Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant molecular neural networks. In Advances in Neural Information Processing Systems, 2019; 2019; pp 14510-14519.

72.   Chen, B.; Barzilay, R.; Jaakkola, T., Path-augmented graph transformer network. *arXiv preprint arXiv:1905.12712* **2019**.

73.   Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., Neural message passing for quantum chemistry. 2017. *arXiv preprint arXiv:1704.01212* **2017**.

74.   Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; He, L. Molecular property prediction: A multilevel quantum interactions modeling perspective. In Proceedings of the AAAI Conference on Artificial

Intelligence, 2019; 2019; Vol. 33; pp 1052-1060.

75.   Lubbers, N.; Smith, J. S.; Barros, K., Hierarchical modeling of molecular energies using a deep neural network. *J Chem Phys* **2018**, 148, 241715.

76.   Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S., Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *J Chem Theory Comput* **2018**, 14, 4687-4698.

77.   Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Advances in neural information processing systems, 2017; 2017; pp 991-1001.

78.   Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A., Quantum-chemical insights from deep tensor neural networks. *Nat Commun* **2017**, 8, 13890.

79.   Schutt, K. T.; Gastegger, M.; Tkatchenko, A.; Muller, K. R.; Maurer, R. J., Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat Commun* **2019**, 10, 5024.

80.   Schutt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Muller, K. R., SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J Chem Theory Comput* **2019**, 15, 448-455.

81.   Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R., SchNet - A deep learning architecture for molecules and materials. *J Chem Phys* **2018**, 148, 241722.

82.   Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S., Discovering a Transferable Charge Assignment Model Using Machine Learning. *J Phys Chem Lett* **2018**, 9, 4495-4501.

83.   Unke, O. T.; Meuwly, M., PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J Chem Theory Comput* **2019**, 15, 3678-3693.

84.   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.

85.   Xie, T.; Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys Rev Lett* **2018**, 120, 145301.

86.   Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R., Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* **2019**, 59, 3370-3388.

87.   Caruana, R., Multitask learning. *Machine learning* **1997**, 28, 41-75.

88.   Hochreiter, S.; Schmidhuber, J., Long Short-Term Memory. *Neural Computation* 9, 1735-1780.

89.   Kingma, D. P.; Ba, J., Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.

90.   Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Machine learning of molecular electronic properties in chemical compound space. *New J Phys* **2013**, 15.

91.   Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O., Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science Advances* **2019**, 5, eaav6490.

92.   Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R., Machine learning of accurate energy-conserving molecular force fields. *Science advances* **2017**, 3, e1603015.

93.   Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S., The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for

molecules. *Sci Data* **2020**, 7, 134.

94. Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E., Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **2019**, 10, 2903.

95. Padula, D.; Omar, Ö. H.; Nematiaram, T.; Troisi, A., Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy & Environmental Science* **2019**, 12, 2412-2416.

96. Rezac, J.; Bim, D.; Gutten, O.; Rulisek, L., Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *J Chem Theory Comput* **2018**, 14, 1254-1266.

97. Sun, Q. M.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z. D.; Liu, J. Z.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K. L., PYSCF: the Python-based simulations of chemistry framework. *Wires Comput Mol Sci* **2018**, 8.

98. Hirshfeld, F. L., Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, 44, 129-138.

99. Mulliken, R., Electronic population analysis on LCAO‐MO molecular wave functions. III. Effects of hybridization on overlap and gross AO populations. *The Journal of Chemical Physics* **1955**, 23, 2338-2342.

100. Mulliken, R., Electronic population analysis on LCAO–MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. *The Journal of Chemical Physics* **1955**, 23, 1841-1846.

101. Mulliken, R. S., Electronic population analysis on LCAO–MO molecular wave functions. I. *The Journal of Chemical Physics* **1955**, 23, 1833-1840.

102. Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W., Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys Rev Lett* **2018**, 120, 143001.

103. http://quantum-machine.org/datasets/.

104. Chaudhari, S.; Polatkan, G.; Ramanath, R.; Mithal, V., An attentive survey of attention models. *arXiv preprint arXiv:1904.02874* **2019**.

105. Liang, Y.; Ke, S.; Zhang, J.; Yi, X.; Zheng, Y. Geoman: Multi-level attention networks for geo-sensory time series prediction. In IJCAI, 2018; 2018; pp 3428-3434.

106. Li, X.; Zhao, B.; Lu, X. MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning. In IJCAI, 2017; 2017; pp 2208-2214.

107. Ziteng Liu, L. L., Qingqing Jia, Zheng Cheng, Yanyan Jiang, Yanwen Guo, Jing Ma, DeepMoleNet. http://106.15.196.160:5659   (Released on November 26, 2020).

108. https://github.com/Frank-LIU-520/DeepMoleNet.

TOC