THE JOURNAL OF PHYSICAL CHEMISTRY Letters

pubs.acs.org/JPCL Letter

Toward Accurate Predictions of Atomic Properties via Quantum Mechanics Descriptors Augmented Graph Convolutional Neural Network: Application of This Novel Approach in NMR Chemical Shifts Predictions

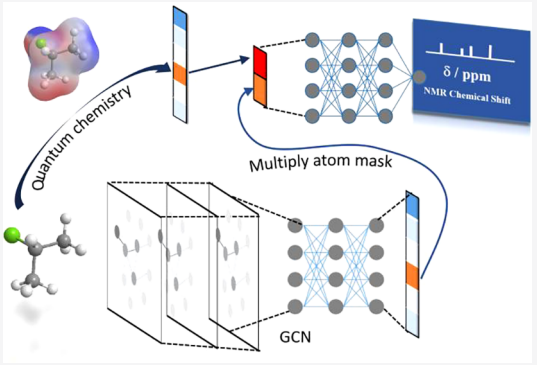
Peng Gao, Jie Zhang,* Yuzhu Sun, and Jianguo Yu

Cite This: J. Phys. Chem. Lett. 2020, 11, 9812–9818

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: In this study, an augmented Graph Convolutional Network (GCN) with quantum mechanics (QM) descriptors was reported for its accurate predictions of NMR chemical shifts with respect to experimental values. The prediction errors of $^{13}C/^{1}H$ NMR chemical shifts can be as small as 2.14/0.11 ppm. There are two crucial characteristics for this modified GCN: in one aspect, such a novel neural network could efficiently extract the overall molecule structure information; in another aspect, it could accurately solve the chemical environment of the target atom. As there exists an imperfect linear regression between the experimental NMR chemical shifts (δ) and the density functional theory (DFT) calculated isotropic shielding constants (σ), the inclusion of QM descriptors within GCN can largely improve its performance. Moreover, few-shot learning also becomes feasible with these descriptors. The success of this novel GCN in chemical shifts predictions also indicates its potential applicability for other computational studies.

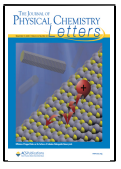


In recent years, increasingly more deep learning approaches have been applied in the field of computational chemistry; with them, many of the routine calculations have been largely facilitated.[1−12] For instance, graph based neural networks have been proved useful for chemical properties predictions.[13,14] And moreover, some other novel data-driven approaches were also applied to augment the performance of traditional density functional theory (DFT) calculations.[15] The background reason lies in the fact that the common DFT methods can provide an overall estimations numerically, through electronic structure calculations, while the deep neural networks could further refine these calculation results via applying related descriptors to build more straightforward correlations with the prediction goals. Currently, accurate predictions of atomic properties have drawn considerable attention, as such predictions could provide valuable chemical insights for researchers. In the current stage, lots of promising graph based tools are rapidly growing in popularity for molecular properties predictions, and a 3-dimensional (3D) molecule structure can be seen as a graph unit, and therefore, related information associated with the predicted properties can be directly extracted.[13,14,16] However, proper ways of extending these emerging approaches to atomic studies still remain to be explored; the largest challenge lies in the solution of the atomic environment information inside molecule.

DFT or other quantum mechanics (QM) based packages are helpful to provide useful calculations, which are associated with atomic properties, and with the inclusion of these calculated QM descriptors in neural networks, high-accuracy predictions of atomic properties may become available. And moreover, a higher numerical correlation between the included QM descriptor and the predicted quantity may even make the few-shot learning feasible. However, to fully realize this within the framework of graph based neural networks, which are powerful to differentiate molecules, some development work is needed to optimize the architecture of this kind of neural work.

NMR is a useful tool in dealing with challenging structural assignments for chemical scientists, and high-accuracy yet efficient predictions of NMR chemical shifts with respect to experimental values via affordable computations will be of great importance. For a target atom, the value of its NMR chemical shift is largely dependent on its local chemical environment. In

Received: August 30, 2020
Accepted: October 19, 2020
Published: November 5, 2020



https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812–9818

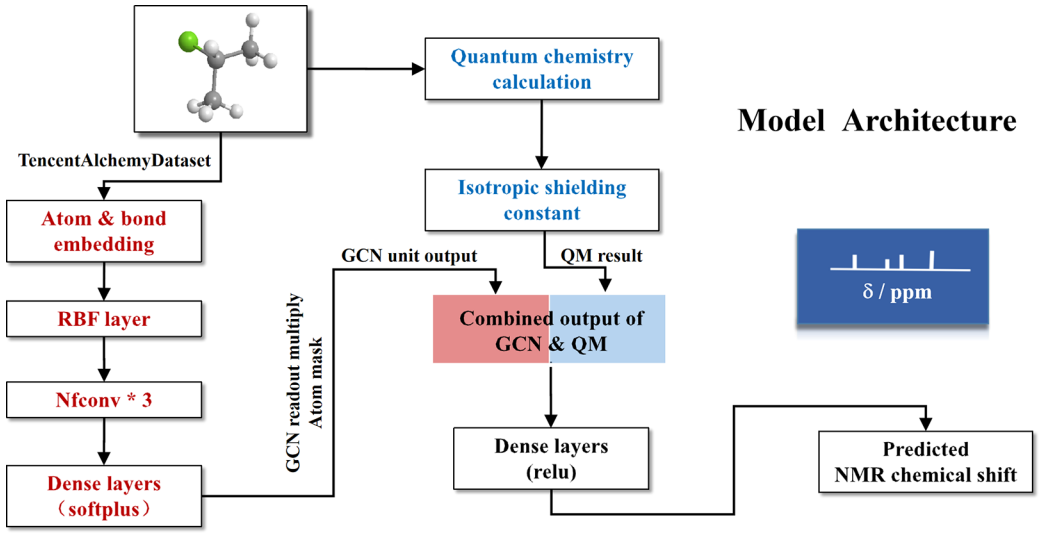


Figure 1. Illustration of the modified GCN architecture, designed for NMR chemical shifts (δ) predictions, with inclusion of a QM descriptor: DFT calculated isotropic shielding constant (σ).

the current stage, with the introduction of the gauge-including atomic orbital (GIAO) method,[17] the calculation of the isotropic shielding constant with DFT has become feasible. However, the accuracy of this routine DFT/GIAO approach is usually limited by many factors, including the conformational averaging, relativistic effects, and so on.[18] Tantillo and others first applied linear regression (LR) scaling factors to numerically correct the calculated isotropic shielding constants with respect to experimental NMR chemical shift values, and they realized higher accuracy.[18−22] The imperfect linear regression between these two quantities has incurred the interests of following researchers to apply increasingly more advanced data-driven approaches, some of which are even suitable for few-shot learning. In our previous study, we had successfully applied the deep neural network (DNN) to solve the molecular environment and improve the performance of DFT/GIAO approach in $^{13}C/^{1}H$ NMR chemical shifts predictions.[15] However, it is worth noting that without efficient extraction of 3D molecular structure information, the application of pure numerical chemical environment descriptors is not sufficient for exact solution of overall information, and the final accuracy may be limited for general use. Currently, some researchers also tried to apply message passing or other neural networks, which merely depend on the effective representations of molecules, to conduct NMR chemical shifts predictions; the most obvious advantage of this approach lies in the fact that it could deal with a large amount of molecular data within a short time period, while DFT calculations may be comparably expensive.[23,24] However, without the inclusion of QM descriptors, in one aspect, the few-shot learning may not be feasible, and in another aspect, considering the fact such kinds of neural networks usually depend on large data sets for training, the developed models may be less sensitive to the structures beyond their training set, and their transferrability among independent assignments may be further limited.

Architecture of the Designed GCN. Graph Convolutional Network (GCN) is a promising tool for computational chemists to conduct molecular properties predictions.[13,14] The advantage of this neural network lies in the fact that it could efficiently extract the useful information from molecular structures, without applying complex descriptors. However, to successfully transfer such a novel tool for atomic properties predictions, in one aspect, the neural network's architecture should be designed with respect to practical purpose,[25,26] and in another aspect, some related atomic descriptors should be included to refine the model. Both of these are hurdles for computational chemists.

In this report, we proposed a novel graph convolutional neural network (GCN, Figure 1) with implementation of SchNet[13] for atomic properties predictions, and via the inclusion of DFT calculated descriptor, few-shot learning becomes feasible. The general workflow of this proposed GCN model can be outlined as below. Before input to the model, the molecules were transformed into edges and nodes using the *TencentAlchemyData set* within the DGL library.[27,28] Each molecular graph can be seen as an ensemble of nodes and edges. The node features correspond to atomic descriptors including atom type, number of attached hydrogen atoms, number of valencies, proton donor indicator, hybridization, and aromaticity indicator. These kinds of descriptors were presented by a one-hot encoded vector. The edge represents bonds or connections between atom pairs, and the corresponding descriptors include bond type and length. In this study, we applied the fully connected molecular graph to the present molecules, indicating that all the connections between every two atoms were created and recorded for analysis. After embedding, a radial basis function (RBF) layer was applied to construct a distance tensor for the atomic representations recording. Its input is a value of distance between nodes (atoms), and its output is a radial basis vector. Given the fact that intramolecular atoms do not lie on a regular grid, the implicit neighboring environment is difficult to represent. To handle this, three crucial continuous-filter convolution layers were implemented to optimize the interatomic interactions by changing the distance tensor. The continuous evolution of the *i*th atom's representation can be expressed with the following equation:

$$a_i^{l+1} = \sum_{j=0}^{N} a_j^l \circ \omega^l(d_{ij}) \quad (1)$$

https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812−9818

where ω^l plays a filter-generating role in mapping the atom representations to the corresponding values of the filter bank and ° indicates the element-wise multiplication. On the basis of the study by K. T. Schütt et al. focusing on the SchNet model,[13] we learned that within the framework of this modified GCN, interatomic rational invariance can also be realized via directly applying the computed pairwise distance d_{ij}, instead of relative positions of atoms. Thus, the continuous evolution of atomic representations can be conducted more efficiently. The equation expressed in the form of the Gaussian function, g_k, which is shown below, was used as conditions to restrict the continuous evolution of the filter values and, therefore, to further improve the optimization performance of this neural network.

$$g_k(d_{ij}) = \exp(-\alpha(d_{ij} - \mu_k)^2) \tag{2}$$

where μ_k indicates a certain value between zero and the distance of a complete cutoff. Actually, the accuracy of the calculations of radial basis vectors within the RBF layer can be controlled by the number of the applied Gaussian functions, and the set of hyper parameter α. In this study, the value of α is set to 0.1.[13] To note, with the feature of the interatomic rotational invariance, as well as the convolution linearity, such a neural network is also able to keep each atom-wise vector quantity unchanged across different times of periodic repetitions. That is to say, such a novel neural network is also consistent with the periodic boundary conditions (PBCs) for calculations of periodic systems.[13]

Within the framework of this neural network, any given property Q can be actually obtained from the summation over all the atomic contributions. As the initial embedding order is consistent with the initial atom order, the obtained interaction by the pairwise layer can be transferred to each atom independently, and thus with the invariance of index, the quality of the predicted property Q' can be guaranteed as the summation of atomic contributions. The accuracy of the prediction can be controlled with a function of squared loss:

$$L(Q, Q') = (Q - Q')^2 \tag{3}$$

However, to further enable this neural network to accurately predict atomic properties, we need to modify its architecture, one extra atom-wise mark ($[0...1_i...0]$) was applied in the readout stage (ReS) to include extra yet related atomic information, to make the applied GCN focus on the target ith atom. The dense layers are designed to include the potential QM descriptors.

Inside the molecule, the chemical shift (CS) of a given atomic nucleus i is due to the nuclear shielding effect of an external magnetic field. Such an effect was caused by the induced field, which was formed via the circulation of electrons surrounding this nucleus. The strength of the induced field, B_1, is therefore proportional to that of the external field B_0 and can be expressed as

$$B_1 = B_0(\mathbf{1} - \sigma_i) \tag{4}$$

where $\mathbf{1}$ represents the unit matrix and σ_i indicates the nuclear shielding tensor. For regular NMR measurement, B_0 is usually the uniformed field along z-axis, therefore, $\sigma_i \approx \sigma_{izz}$. And the NMR frequency of the ith nucleus, υ_i, can be written as

$$\upsilon_i = (\gamma_i/2\pi)B_1 = (\gamma_i/2\pi)B_0(\mathbf{1} - \sigma_i) \tag{5}$$

where γ_i indicates the gyromagnetic ratio of the ith nucleus and is usually taken as a constant. The calculation of σ_i can be realized via the DFT/GIAO approach; however, in some cases, the accuracy of QM calculations may be impacted by many factors.[18] The value of the chemical shift, δ_i (in ppm), can be transferred via the equation

$$\delta_i = 10^6(\upsilon_i - \upsilon_0)/\upsilon_0 \tag{6}$$

where υ_0 indicates the resonance frequency of a referenced nucleus.[29−31]

On the basis of the discussion above, we realized that there does exist an imperfect linear regression correlation between the experimental CSs (δ_i) and the DFT computed isotropic shielding constants (σ_i).[18,19] However, to note, in some cases of complex bonding environment, large nonlinear deviations of the computed σ_i may also happen.[17,18] In many studies, a numerical approximation of the δ_i can be made via the following equation, which uses scaling factors for correction.[18−20]

$$\delta_i = \frac{\text{intercept} - \sigma_i}{-\text{slope}} \tag{7}$$

To reasonably incorporate this important chemical information into the framework of the proposed GCN, a separate processing is needed. With our proposed approach, the structural information obtained by GCN and the QM descriptor σ can be combined in a dense layer with separate weight functions. The addition of former is able to overcome the nonlinear deviations. The prediction for the CS of the ith atom can be made via the equation below. With such a embedded layer, the prediction accuracy can be largely improved, and few-shot learning also becomes feasible.

$$CS_i = f(W_{\text{ReS}}*\text{ReS}[i] + W_{\text{QM}}*\sigma_i) \tag{8}$$

With these modifications, we enabled the GCN to learn new chemical knowledge via utilizing the novel QM descriptor, σ, and its performance in atomic properties predictions can be enhanced, indicating the high applicability of modern data-driven tools in computational chemistry. More technical details of this developed package and data sets can be found on https://github.com/jeah-z/NMR-GCN.

All the DFT calculations were conducted with Gaussian 09.[32] In general, there are two main steps for DFT calculations: first, geometry optimizations of the target molecules need to be conducted in the gas phase to find the minimum point on the potential energy surface, and these optimized structures have to be further verified via vibrational frequency calculations. Second, the isotropic shielding constants can be calculated by GIAO approach, and during this step, the SMD implicit solvent model[33] was applied to further improve the calculation accuracy. We believe that the choice of different levels of theory in DFT part may impact the final performance of the developed model, and on the basis of previous studies,[19,21,22,34] we recommend applying M062X/6-31+G(d,p) for geometry optimization and mPW1PW91/6-311+G(2d,p) or PBE0/6-311+G(2d,p) for NMR GIAO calculation.

Overall Performance of the Designed Neural Network. We applied this developed GCN in predictions of ^{13}C, ^{1}H, ^{15}N, ^{31}P, and ^{19}F NMR chemical shifts. And its performance was summarized in Table 1 and Figure 2. We can see that such a GCN+DFT model has largely improved the prediction accuracy, compared to LR+DFT (using scaling factors for

https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812−9818

Table 1. Prediction Errors (in ppm) for ^{13}C, ^{1}H, ^{15}N, ^{31}P, and ^{19}F NMR Chemical Shifts with Respect to Experimental Values

NMR[a]	MAE test[d]	MAE test[e]	MAE test[f]
^{13}C[b]	2.14	2.86	3.10
^{1}H[b]	0.11	0.19	0.14
^{15}N[b]	4.38	N/A	N/A
^{31}P[c]	7.34	N/A	N/A
^{19}F[c]	4.62	N/A	N/A

[a]There are 476 ^{13}C, 217 ^{1}H, 67 ^{15}N, 35 ^{31}P, and 45 ^{19}F chemical shifts in the original data set; experimental data were taken from refs 15, 19, 21, and 35. 85% of the original data was applied as the training set, and the remaining 15% as the test set. [b]The level of theory for DFT calculation: M062X/6-31+G(d,p) for geometry optimization and mPW1PW91/6-311+G(2d,p) for NMR GIAO calculation. [c]The level of theory for DFT calculation: B3lyp/cc-pVDZ for geometry optimization and B3lyp/cc-pVDZ for NMR GIAO calculation. [d]The predictions were made using the GCN+DFT model. [e]The predictions were made using the DNN+DFT model. [f]The predictions were made using linear scaling factors.[19]

correction) and GCN+DFT approaches.[15,19] Different from our previous study focusing on the development of the DNN +DFT model,[15] which utilizes numerical RDKit descriptors to solve chemical environment, this novel GCN+DFT model is more dependent on the 3D structure to extract the overall molecule information; thus, its accuracy was proved to be higher, and its transferrability among different classes of compounds can be reasonably expected. While for linear regression corrections, we admit its effectiveness on most regular structural assignments, for some challenging bonding environments, large deviations will appear; more details will be discussed in the next section. And, moreover, the inclusion of the QM descriptor makes the modified GCN available and accurate for few-shot learning, with lower probability of overfit issue; such an architecture will also be preferred over other routine GCN models, which require a large amount of molecules for training, in atomic properties predictions. GCN could efficiently refine the DFT calculated isotropic shielding constants (σ) with respect to experimental chemical shifts (δ), by graphical features. Another advantage of this proposed approach can be indicated by the fact that all kinds of NMR chemical shifts can be obtained conveniently in just one set of calculation. We believe in the future such a promising methodology can also be applied to predict other atomic properties beyond NMR chemical shifts.

Considering the fact that, for a target atom, the value of its NMR chemical shift is largely dependent on its chemical environment; in some cases, like bonding with heavy atoms or with higher bond order, simple numerical corrections may not be sufficient. The background reason lies in the fact that such kinds of chemical environments may cause the asymmetric distribution of the electron density, and the calculated σ may nonlinearly deviate from the value of the NMR chemical shift.[18] To overcome these challenges, more advanced tools for refining predictions are needed. To further demonstrate the advantages of the proposed GCN+DFT model over the LR +DFT approach, which directly depends on scaling factors to conduct predictions, we compared their performance on several molecules, the structures of which can be found in Table 2. The prediction errors for ^{13}C NMR chemical shifts of the carbon connected with halides are hard to correct by linear scaling factors, but the modified GCN is good at this kind of prediction (Molecules 1 and 2). In addition, for the carbon atoms with double or triple bonding environment, the performance of our GCN+DFT model was also proved to be reliable (Molecules 3−5). It was worth noting that if the diversity of the training molecules is extended, the prediction accuracy of this GCN+DFT model can be further improved.

Application of This Developed GCN+DFT Model in Chemical Research. (1) General Structural Assignment. To efficiently differentiate the molecules with complex functional groups, the application of $^{13}C/^{1}H$ NMR spectra is sometimes insufficient, especially for the ones that consist of heterocycles. Here, we applied our modified GCN+DFT model for ^{13}C, ^{1}H, and ^{19}F NMR chemical shift predictions on a set of typical molecules, the structures of which are illustrated in Figure 3. To note, all these molecules are not included in the original training or test data sets. And, all the collected experimental data were measured in chloroform or DMSO, the most commonly applied solvents. The prediction results were presented in Tables 3−6. And we can find that all the prediction errors are within a reasonable range, in most cases, less than 1 MAE (mean absolute error).

However, we need to point out that, for some molecules shown in Figure 3, the larger errors in their ^{13}C NMR chemical shift predictions are mainly caused by relativistic effects instead of the model, if the target carbon atoms are bonded with fluorine or other halogen atoms.[34] As discussed above, the application of deep learning approaches is one of the most useful ways to reduce the errors rising from this effect.

(2) Structural Assignment for Tautomers. We also applied this novel approach for the structural assignment of a tautomeric

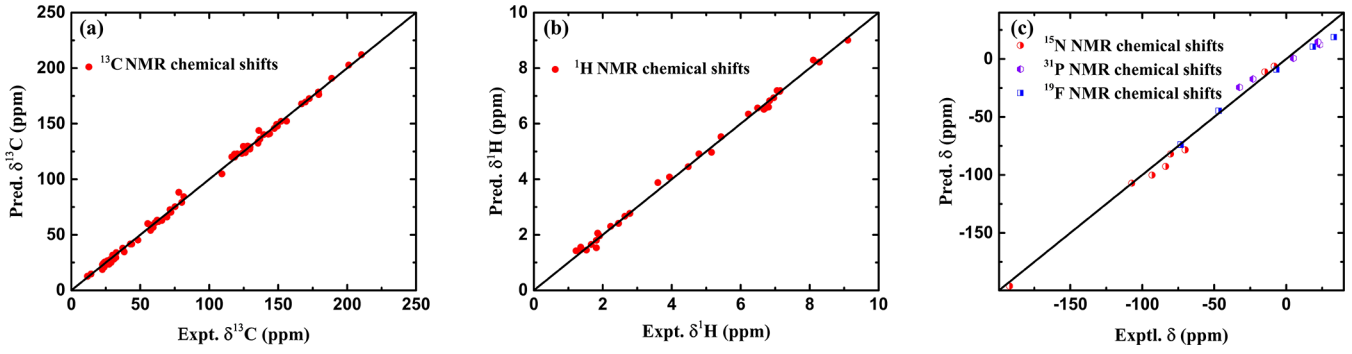


Figure 2. Comparison between the predicted and experimental NMR chemical shifts: (a) ^{13}C NMR chemical shifts; (b) ^{1}H NMR chemical shifts; (c) ^{15}N, ^{31}P, and ^{19}F NMR chemical shifts.

https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812−9818

Table 2. Predicted and Experimental ^{13}C NMR Chemical Shifts (in ppm) for Some Selected Atoms

Molecule	SMILES[a]	Structure	Exp.	Error.[b]	Error.[c]
1	C1CCC(CC1)Cl	$ClCH_2CH_2CH(CH_2Cl)Cl$	60.17	5.75	0.76
2	C(Cl)Cl	CH_2Cl_2	53.70	12.63	3.05
3	C(=CCl)(C)C	$(CH_3)_2C{=}CHCl$	134.91	9.86	2.15
4	CC(C)C#N	$(CH_3)_2CHC{\equiv}N$	123.83	6.05	0.76
5	C(#N)C	$CH_3C{\equiv}N$	117.40	6.50	0.33

[a]SMILES: simplified molecular input line entry system. [b]The prediction errors of NMR chemical shifts via the LR+DFT approach, the scaling factors were obtained from ref 19. Slope: −1.0446. Intercept: 186.7246. [c]The prediction errors of NMR chemical shifts via the trained GCN+DFT model.

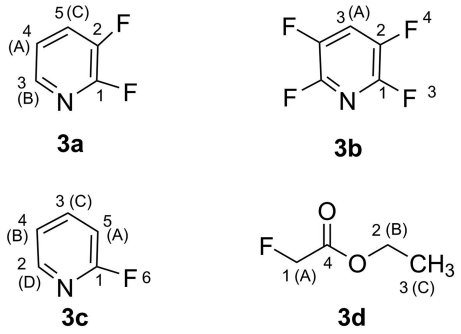


Figure 3. Structures of **3a**−**d**.

Table 3. Predicted and Experimental ^{13}C, ^{1}H, and ^{19}F NMR Chemical Shifts (in ppm) for 2,3-Difluoropyridine (3a)

NMR	pos[a]	exp	pred[b]	error	rel error (%)
^{13}C	C-1	153.43	155.19	1.76	1.1
	C-2	146.88	149.03	2.15	1.5
	C-3	141.78	144.00	2.22	1.6
	C-4	126.70	124.26	−2.44	1.9
	C-5	123.33	131.52	9.19	7.5
^{1}H	H(A)	7.99	7.75	−0.24	3.0
	H(B)	7.58	7.04	−0.54	7.1
	H(C)	7.20	7.41	0.21	2.9
^{19}F	F-6	−149.00	−137.42	11.58	7.8
	F-7	−87.00	−85.92	1.08	1.2

[a]Positions for the atom of interest. [b]The predicted NMR chemical shifts via the trained GCN+DFT model.

Table 4. Predicted and Experimental ^{13}C, ^{1}H, and ^{19}F NMR Chemical Shifts (in ppm) for 2,3,5,6-Tetrafluoropyridine (3b)

NMR	pos[a]	exp	pred[b]	error	rel error (%)
^{13}C	C-1	144.80	150.86	6.06	4.2
	C-2	143.15	149.68	6.53	4.6
	C-3	119.04	121.33	2.29	1.9
^{1}H	H(A)	7.63	7.39	−0.24	3.1
^{19}F	F-3	−93.00	−93.25	−0.25	0.3
	F-4	−141.00	−139.25	1.75	1.2

[a]Positions for the atom of interest. [b]The predicted NMR chemical shifts via the trained GCN+DFT model.

compound, adenine (shown in Figure 4). To efficiently detect the correct structure between **4a** and **4b**, merely applying experimental methods was not sufficient. A consistency between the predicted and experimental NMR chemical shifts is preferred for clarity. For adenine, the most efficient way of structural assignment is to conduct both ^{13}C and ^{15}N NMR

Table 5. Predicted and Experimental ^{13}C, ^{1}H, and ^{19}F NMR Chemical Shifts (in ppm) for 2-Fluoropyridine (3c)

NMR	pos[a]	exp	pred[b]	error	rel error (%)
^{13}C	C-1	168.48	163.80	−4.68	2.8
	C-2	148.10	148.07	−0.03	0.02
	C-3	141.41	142.01	0.60	0.42
	C-4	121.28	120.53	−0.75	0.62
	C-5	108.95	110.88	1.93	1.8
^{1}H	H(A)	8.19	8.01	−0.18	2.2
	H(B)	7.15	7.02	−0.13	1.8
	H(C)	7.77	7.64	−0.13	1.7
	H(D)	6.88	6.78	−0.10	1.5
^{19}F	F-6	−68.00	−69.27	−1.27	1.9

[a]Positions for the atom of interest. [b]The predicted NMR chemical shifts via the trained GCN+DFT model.

Table 6. Predicted and Experimental ^{13}C and ^{1}H NMR Chemical Shifts (in ppm) for Ethylfluoroacetate (3d)

NMR	pos[a]	exp	pred[b]	error	rel error (%)
^{13}C	C-1	74.22	76.38	2.16	2.9
	C-2	61.54	60.71	−0.83	1.3
	C-3	14.18	13.34	−0.84	5.9
	C-4	168.52	170.83	2.31	1.4
^{1}H	H(A)	4.84	4.71	−0.13	2.7
	H(B)	4.28	3.94	−0.34	7.9
	H(C)	1.32	1.34	0.02	1.5

[a]Positions for the atom of interest. [b]The predicted NMR chemical shifts via the trained GCN+DFT model.

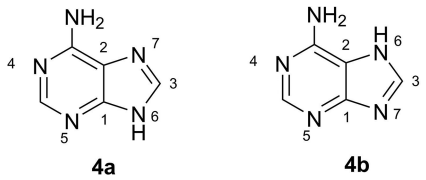


Figure 4. Structures of two different tautomers of adenine (**4a**−**b**).

chemical shift predictions at the same time. The predicted results of these two possible structures by our GCN+DFT model were compared with experimental values. In Table 7, we found that for structure **4b**, the prediction errors of both ^{13}C and ^{15}N NMR chemical shifts are larger; some of them are more than 3 MAEs. For **4a**, the predicted results were in good agreement with experimental values, indicating that it is the correct structure.[21] Our conclusion is consistent with previous studies, focusing on the structural assignment of this compound.[36,37]

https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812−9818

Table 7. Predicted and Experimental ^{13}C and ^{15}N NMR Chemical Shifts (in ppm) for Adenine (4a)

	NMR	pos[a]	exp	pred[b]	error	rel error (%)
4a	^{13}C	C-1	151.36	148.69	−2.67	1.8
		C-2	119.07	117.49	−1.58	1.3
		C-3	140.30	137.58	−2.72	1.9
	^{15}N	N-4	−149.80	−154.68	−4.88	3.3
		N-5	−151.50	−161.12	−9.62	6.3
		N-6	−222.60	−230.79	−8.19	3.7
		N-7	−140.30	−151.43	−11.13	7.9
4b	^{13}C	C-1		158.87	7.51	5.0
		C-2		111.61	−7.46	6.3
		C-3		142.38	2.08	1.5
	^{15}N	N-4		−143.64	6.16	4.1
		N-5		−136.73	14.77	9.7
		N-6		−236.22	−13.62	6.1
		N-7		−140.61	−0.31	0.22

[a]Positions for the atom of interest. [b]The predicted NMR chemical shifts via the trained GCN+DFT model.

In summary, accurate predictions of atomic properties, like different kinds of NMR chemical shifts, have been realized via the proposed GCN+DFT model, which has also been proved suitable for few-shot learning. To note, the prediction performance of this developed model can be potentially improved by increasing the diversity of the molecules contained in the original data set or applications of higher level QM calculations. Moreover, further optimization of the GCN architecture is also of great importance to make it more suitable for cheminformatics studies. The work presented in this study indicates a new yet promising page of computational chemistry with the assistance of advanced artificial intelligence (AI) tools. In the future, we hope more properties can be accurately predicted via the methodology proposed in this study.

AUTHOR INFORMATION

Corresponding Author

Jie Zhang − *Centre of Chemistry and Chemical Biology, Bioland Laboratory (Guangzhou Regenerative Medicine and Health-Guangdong Laboratory), Guangzhou 53000, China; School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0002-5575-303X; Email: j.chang@ecust.edu.cn

Authors

Peng Gao − *School of Chemistry and Molecular Bioscience, University of Wollongong, Wollongong, NSW 2500, Australia;* orcid.org/0000-0002-1290-6972

Yuzhu Sun − *School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China*

Jianguo Yu − *School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c02654

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the NCI system, supported by the Australian Government (Project id: v15) to provide computational resource. And, we also thank the Australian Government, which supported P.G.'s Ph.D. study via offering him an Australian International Postgraduate Award scholarship.

REFERENCES

(1) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(2) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828−12840.

(3) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5*, 755−767.

(4) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.

(5) Meldgaard, S. A.; Kolsbjerg, E. L.; Hammer, B. Machine learning enhanced global optimization by clustering local environments to enable bundled atomic energies. *J. Chem. Phys.* **2018**, *149*, 134104.

(6) Ouyang, R.; Xie, Y.; Jiang, D.-e. Global minimization of gold clusters by combining neural network potentials and the basin-hopping method. *Nanoscale* **2015**, *7*, 14817−14821.

(7) Sørensen, K. H.; Jørgensen, M. S.; Bruix, A.; Hammer, B. Accelerating atomic structure search with cluster regularization. *J. Chem. Phys.* **2018**, *148*, 241734.

(8) Wexler, R. B.; Martirez, J. M. P.; Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 4678−4683.

(9) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844−9853.

(10) Panapitiya, G.; Avendaño-Franco, G.; Ren, P.; Wen, X.; Li, Y.; Lewis, J. P. Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140*, 17508−17514.

(11) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.

(12) Bai, Y.; Wilbraham, L.; Slater, B. J.; Zwijnenburg, M. A.; Sprick, R. S.; Cooper, A. I. Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. *J. Am. Chem. Soc.* **2019**, *141*, 9063−9071.

(13) 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*, 241722.

(14) Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; He, L.Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective. *arXiv1906.11081* **2019**.

(15) Gao, P.; Zhang, J.; Peng, Q.; Zhang, J.; Glezakou, V.-A. General Protocol for the Accurate Prediction of Molecular $^{13}C/^{1}H$ NMR Chemical Shifts via Machine Learning Augmented DFT. *J. Chem. Inf. Model.* **2020**, *60*, 3746−3754.

(16) Gao, P.; Zhang, J.; Sun, Y.; Yu, J. Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architecture. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23766−23772.

(17) Ditchfield, R. Self-consistent perturbation theory of diamagnetism. *Mol. Phys.* **1974**, *27*, 789−807.

(18) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational Prediction of1H and13C Chemical Shifts: A Useful Tool for Natural

Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012**, *112*, 1839−1862.
(19) Lodewyk, M. W.; Soldi, C.; Jones, P. B.; Olmstead, M. M.; Rita, J.; Shaw, J. T.; Tantillo, D. J. The Correct Structure of Aquatolide—Experimental Validation of a Theoretically-Predicted Structural Revision. *J. Am. Chem. Soc.* **2012**, *134*, 18550−18553.
(20) Xin, D.; Sader, C. A.; Chaudhary, O.; Jones, P.-J.; Wagner, K.; Tautermann, C. S.; Yang, Z.; Busacca, C. A.; Saraceno, R. A.; Fandrick, K. R.; et al. Development of a ^{13}C NMR Chemical Shift Prediction Procedure Using B3LYP/cc-pVDZ and Empirically Derived Systematic Error Correction Terms: A Computational Small Molecule Structure Elucidation Method. *J. Org. Chem.* **2017**, *82*, 5135−5145.
(21) Gao, P.; Wang, X.; Yu, H. Towards an Accurate Prediction of Nitrogen Chemical Shifts by Density Functional Theory and Gauge-Including Atomic Orbital. *Advanced Theory and Simulations* **2019**, *2*, 1800148.
(22) Gao, P.; Wang, X.; Huang, Z.; Yu, H. ^{11}B NMR Chemical Shift Predictions via Density Functional Theory and Gauge-Including Atomic Orbital Approach: Applications to Structural Elucidations of Boron-Containing Molecules. *ACS Omega* **2019**, *4*, 12385−12392.
(23) Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, M.; Kang, S. Neural Message Passing for NMR Chemical Shift Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2024−2030.
(24) Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P. IMPRESSION − prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chemical Science* **2020**, *11*, 508−515.
(25) Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J. Toward More Reliable ^{13}C and ^{1}H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *J. Chem. Inf. Model.* **2008**, *48*, 128−134.
(26) Kiryanov, I. I.; Tulyabaev, A. R.; Mukminov, F. K.; Khalilov, L. M. Neural network for prediction of ^{13}C NMR chemical shifts of fullerene C60 mono-adducts. *J. Chemom.* **2018**, *32*, No. e3037.
(27) Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv1909.01315* **2019**.
(28) 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. *arXiv1906.09427* **2019**.
(29) Pople, J. A.; Bernstein, H. J.; Schneider, W. G. *High-resolution nuclear magnetic resonance*; McGraw-Hill: New York, NY, 1959.
(30) Becker, E. *High Resolution NMR: Theory and Chemical Applications*; Elsevier Science: San Diego, 1999.
(31) Slichter, C. *Principles of Magnetic Resonance*; *Springer Series in Solid-State Sciences*; Springer Berlin Heidelberg, 2013.
(32) Frisch, M. J.; et al. *Gaussian 09*, Revision E.01; Gaussian Inc.: Wallingford CT, 2009.
(33) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(34) Latypov, S. K.; Polyancev, F. M.; Yakhvarov, D. G.; Sinyashin, O. G. Quantum chemical calculations of ^{31}P NMR chemical shifts: scopes and limitations. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6976−6987.
(35) Hans Reich Collection. https://organicchemistrydata.org/hansreich/.
(36) Laxer, A.; Major, D. T.; Gottlieb, H. E.; Fischer, B. ($^{15}N_5$)-Labeled Adenine Derivatives: Synthesis and Studies of Tautomerism by ^{15}N NMR Spectroscopy and Theoretical Calculations. *J. Org. Chem.* **2001**, *66*, 5463−5481.
(37) Xin, D.; Sader, C. A.; Fischer, U.; Wagner, K.; Jones, P.-J.; Xing, M.; Fandrick, K. R.; Gonnella, N. C. Systematic investigation of DFT-GIAO ^{15}N NMR chemical shift prediction using B3LYP/cc-pVDZ: application to studies of regioisomers, tautomers, protonation states and N-oxides. *Org. Biomol. Chem.* **2017**, *15*, 928−936.

https://dx.doi.org/10.1021/acs.jpclett.0c02654
J. Phys. Chem. Lett. 2020, 11, 9812−9818