A Very Deep Graph Convolutional Network for $^{13}$C NMR Chemical Shift Calculation with Density Functional Theory Level Performance for Structure Assignment

Wen-Jing Ai,†a Jing Li,‡b Dongsheng Cao, a Shao Liu, b Yi-Yun Yuan, a Yan Li, a Gui-Shan Tan, b Kang-Ping Xu, a Xia Yu, a Fenghua Kang, a Zhen-Xing Zou, a and Wen-Xuan Wang* a

aXiangya School of Pharmaceutical Sciences, Central South University, Changsha, Hunan 410013, PR China. E-mail: wenxuanwang@tom.com

bDepartment of Pharmacy in Xiangya Hospital, Central South University, Changsha, Hunan 410013, PR China.

† These authors contributed equally to this work.

Abstract

Nuclear magnetic resonance (NMR) chemical shift calculation is a powerful tool for structural elucidation, and has been extensively employed in both synthetic and natural product chemistry. However, density functional theory (DFT) NMR chemical shift calculations are usually time-consuming, while fast data-driven methods often lack reliability, making it challenging to apply them to computationally intensive tasks. Herein, we have constructed a 54-layer deep graph convolutional network for $^{13}$C NMR chemical shift calculation, which achieved high accuracy with low time-cost, and performed competitively with DFT NMR chemical shift calculations on structure assignment benchmarks. Our model utilizes a semi-empirical method, GFN2-xTB, and is compatible with a broad variety of organic systems, including those composed of hundreds of atoms or elements ranging from H to Rn. We used this model to resolve the controversial J/K rings junction problem of maitotoxin, which is the largest whole molecule assigned by NMR calculation to date. This model has been developed into a user-friendly software, providing a useful tool for routine rapid structure validation and assignment, as well as a new approach to elucidate the large structures that were previously unsuitable for NMR calculation.

Modification notes, 04/06/2023:
1. Due to a bug in our process scripts, the test set A was not complete in the former version, and the training sets contained several compounds in the test set A. We have re-organized the data sets, and re-trained the weights of model.
2. We added more structures with carbons labelled by STS $^{13}$C NMR calculation.
3. Some errors of experimental data and structures have been modified.
4. The error of KDE in Fig. 5D has been revised.
**Introduction**

Correctly assigning the structures of new synthetic or natural organic compounds is crucial for further chemical and biologic research. Density functional theory (DFT) NMR chemical shift calculation, particularly for $^{13}$C NMR, plays an essential role in validating and elucidating structures, providing high accuracy and reliability in numerous practices.$^{1-6}$ However, these accurate and reliable DFT-based methodologies are often time-consuming and computationally intensive, making it challenging to apply them to the systems with hundreds of heavy atoms or many low-energy conformers.

On the other hand, the fast empirical or data-driven approaches, including empirical additivity rules, similarity-to-database analysis, or machine learning models,$^{7-20}$ have not been widely proven for elucidating complex structures, such as novel natural products with multiple chiral centers.$^{7,8,21-23}$ For instance, CASCADE,$^{8}$ the recent state-of-the-art NMR calculator with DFT level accuracy, still has weakness in distinguishing stereoisomers.$^{22}$ Additionally, because of the limitation of methods (e. g. MMFF94) or algorithms used to calculated molecular features, they cannot support or provide correct results for some organic structures. For example, CASCADE is not compatible with molecules consisting of elements other than C, H, N, O, S, P, F, and Cl, or having formal/net charges (e. g. nitro compounds, amine oxide, anions, cations, zwitterion, etc.), as well as the structures cannot be properly calculated by MMFF94.$^8$

![Fig. 1. The originally proposed (1a) and revised (1b) structures of MTX, and two synthesized moieties (2a and 3a).](image)

There are challenging tasks remain unsolved due to the limitations of DFT or data-driven approaches. An example is the unsolved controversy over the stereochemistry at the J/K ring junction of the largest known, non-polymeric natural product, maitotoxin (1, MTX) (Fig. 1).$^{24-27}$
Albeit the moiety of GHIJK (2a) and GHIJKLMNO (3a) ring systems had been successfully synthesized, supporting the originally proposed structure (1a) based on comparison between their $^{13}$C NMR chemical shift data, the possibility of $1b$ cannot be fully excluded. Assigning stereochemistry to single diastereoisomer with one set of experimental data is supposed to be the exact application scenario of DFT NMR chemical calculation, for which sophisticated statistic approaches, including DP4, DP5, DP4+, and $P_{rel}$, have been designed. Due to the computational cost, however, DFT level $^{13}$C NMR calculations were only performed for truncated GHIJKLM domain, and the strict calculations for whole molecules of the possible structures have not been achieved yet. Therefore, a potential solution for this predicament is to improve the performance of data-driven approaches to match that of DFT level methods.

It is well-known that feed-forward neural networks with sufficient depth and width are universal function approximators, which has been widely verified in the numerous practices of successful architectures, such as graph convolutional networks (GCN), recurrent neural networks, convolutional neural networks, and transformers. Given this understanding, it is plausible that a sophisticated neural network model can accurately approximate the process of converting chemical environments of atoms into chemical shifts.

In this paper, we introduce a very deep GCN model for $^{13}$C NMR chemical shifts calculation that combines efficiency of machine learning methods and performance of DFT level methods. We will elaborate on our work from the five aspects: data collection, feature engineering, details of model, performance in benchmarks, and application in elucidating J/K rings junction of MTX.

**Results and discussion**

**Data collection**

While DFT is based on fundamental principles of physics, data-driven approaches highly rely on large amounts of correct experimental data. However, flaws of $^{13}$C NMR chemical shift data, such as mislabelling and wrong or ambiguous structures, are pervasive in literature and databases, making it challenging to collect sufficient correct data. To address this issue, Paton’s group cleaned the NMR8K dataset to generate Exp5K by DFT $^{13}$C NMR calculation, which they generously made available for open access. These precious data have been included in our dataset. However, without further correction, their DFT $^{13}$C NMR calculation method fails for a considerable number of chemical environments (e.g. heavy atom bearing carbons), potentially excluding some correctly assigned data. Therefore, we have collected and fully labelled another over 4,000 compounds from accessible literature or AIST database. These data from literature were validated by GIAO DFT calculation on oB97x-D4/TZVP//r$^2$scan-3c level of theory, or manual checking on their 1D/2D NMR data, and any suspicious structures or data were excluded. Additionally, the energetically favourable structures of compounds with tautomerism were also determined by DFT free energy calculation on B3LYP/TZVP level of theory. As a result, we obtained over 9000 compounds with over 110,000 carbons labelled by experimental $^{13}$C NMR chemical shifts. In addition, the elements of these compounds included H, B, C, N, O, F, P, S, Cl, Br, Si, Ge, Se, Sn, and I, covering most organic compounds.

For most polychiral compounds, however, only few diastereoisomers have experimental data. Since determining stereochemistry is a major anticipated capacity of our model, the incompleteness of the
experimental data may hinder its ability to analyse the general rules of stereochemical effects. Fortunately, chiral centres are predominantly sp³ hybridized carbons. Their chemical shifts can be accurately calculated with the sorted training set (STS) protocol, which has a mean absolute error (MAE) less than 0.9 ppm, enabling us to label polychiral compounds with near experimental level quality. Therefore, we artificially truncated the polychiral moieties from the collected compounds to compose a set of their diastereoisomers, and labelled them and their low-energy conformers with chemical shift values calculated by STS protocol.

In summary, the dataset consisted of both the experimental and calculated data, with over 1,080,000 labelled carbons.

**Feature engineering**

Since deep learning model can learn high-level features from data and create new features by itself, we used highly simplified descriptor to generate 3D molecular graph, on the premise of keeping the integrity of molecular information (Fig. 2).

To address stereochemical problems, obtaining accurate 3D molecular geometry is inevitable, particularly for some unique structures. To date, GFN2-xTB is the most accurate semiempirical methods with element-specific parameters available up to radon (Z = 86) and has the highest compatibility with most organic structures. Although slightly more time-consuming than force field methods for small organic molecules, GFN2-xTB provides robustness for as many organic compounds as possible, making it the optimal choice for 3D geometry optimization and molecular property calculation.

**Fig. 2** The edge features (EF) of four types of edges connecting nodes (atoms) i–m.

$$\text{EF}(e_{ij}) = \{WBO, \frac{1}{r^2}, 0, 0, 0, 0, WBO, \frac{1}{r^2}\}$$
$$\text{EF}(e_{ij}) = \{0, 0, \alpha + 1, 0, 0, 0, 0, 0\}$$
$$\text{EF}(e_{iii}, e_{jjj}, \text{etc.}) = \{0, 4, 0, 0, 0, 0, 0, 0\} \text{ (self connection)}$$
$$\text{EF}(e_{imm}) = \{0, 0, 0, \theta_i + 1, \theta_i + 1, \theta_j + 1, \theta_j + 1, 0, 0\}$$

Herein, we used GFN2-xTB to optimize geometries and calculate coordination number (CN),
atomic partial charge (q), and Wiberg bond order (WBO), with 3 kcal/mol energy window to remove high-energy conformers for flexible compounds. To avoid the loss of important conformers resulting from errors of energy calculation, we equally weighted all the conformers within this window, instead of using Boltzmann weighting.

In our molecular graphs, atoms are nodes, with features consisting of Z, CN, q, squared q, and Z multiplied by q, as shown in Formula (1). The last two features were expanded quadratic terms that increased the expressivity of model.

The edges and edge features are extracted from molecular internal coordinate, namely bond-length (r, in Å), bond-angle (α, in radian), and dihedral angle (θ, in radian), as well as WBO. For atoms separated by one, two or three bonds (WBO > 0.3), a pair of atoms were assigned with an undirected edge with nine edge features. For atoms directly bonded, the first two and the last two edge features were WBO, the reciprocal of r², WBO divided by r², and squared WBO, respectively, and the other features were set into zero. The last two edge features are also feature expansion terms. For the atoms separated by two bonds, the third edge feature is the value of α + 1 of the two bonds, and the other features are set into zero. Edge features calculated from bond orders, bond-lengths, and bond-angles do not vary significantly between conformers of flexible compounds. Therefore, these edge features of flexible compounds can be directly averaged from all low-energy conformers. However, the dihedral angles involving rotatable bonds are the major differences between conformers for the atoms separated by three bonds. If the dihedral angles of conformers have a large deviation (e. g. 0 for one conformer, and π for another one), the averaged intermediate value may not represent the dihedral angles of any conformers. Thus, we divide the range of dihedral angle [0, π] into four intervals, corresponding to the staggered conformation of ethane, namely [0, π/6], (π/6, π/2], (π/2, 5π/6], and (5π/6, π] (Fig. 2). We use the four edge features (the fourth to seventh) to record θ values, which corresponded one-to-one to the four intervals. If θ fell within one of the four intervals, the value of θ + 1 was recorded in the corresponding feature (Fig. 2). For flexible compounds, non-zero values of these four features were averaged respectively. Having more than one non-zero values in these four features was also a character of flexible moieties.

The detail of model

In the previous studies on NMR chemical shifts calculation, GCN models have been widely used, due to the natural graph structure of organic molecules. However, stacking many GCN layers leads to vanishing gradient problem, and causes that features of graph vertices converging to the same value. Consequently, most state-of-the-art GCNs were shallow. The numbers of GCN layers in the recent NMR predicting models developed by Paton’s group, Choi’s group, Scheidt’s group, and White's group were three, five, three, and four, respectively. The limitation of depth in these models may cause a significant loss of their power.

Herein, we have developed a GCN model named as GFN2NMR, which overcame the training difficulty and over-smoothing problem associated with much deeper GCN models, using residual connections and deep adaptive strategy. Our optimized model comprised fifteen-four GCN layers, which were organized into eighteen stacked residual blocks with short-cut connections (Fig. 3). Each residual block consisted of three stacked GCN units with a short-cut connection. The massage passing and propagation unit was a NNConv module that used an attention-augmented multilayer perceptron (MLP) to map the edge features. The node features output from the 1st, 4th,
The architecture of GFN2NMR.

To ensure the model's generalizability, we randomly divided overall dataset into training set (90%) and test set (10%). The model was trained for 2000 epochs with a learning rate of 0.001, and both training and test losses converged (Fig. 4). The resulting best root mean square (RMS) values for training and test were approximately 1.17 and 1.31 ppm, respectively. Subsequently, we trained the model with all data, utilizing a decaying learning rate (from 0.001 to 0.00001), and achieved the best RMS for training of 1.15 ppm, before conducted benchmarks on the external test sets mentioned below. We also confirmed that none of the compounds in the external test sets were included in the training data.

Comparison with state-of-the-art data-driven methods
To assess the performance of GFN2NMR, we selected the recent state-of-the-art methods, namely online version of CASCADE and Ensemble NMR Prediction consisting of Mestrelab and Modgraph (Neural Network + HOSE), for comparison.

To mitigate potential researcher’s bias and prevent erroneous data, we used all the natural products analyzed by X-ray diffraction and published on the journals *Organic Chemistry Frontiers, Organic Letters*, and *Chinese Chemical Letters* in 2022 to compose the benchmark set (test set A), except the ones not suitable for online CASCADE (due to non-zero net charges or too many heavy atoms) (Fig. S1). Most of these compounds had previously undescribed skeletons, and were not likely included in the training sets or database of the two state-of-the-art methods. We evaluated the performance of each method using several metrics, namely coefficient of determination ($R^2$), MAE, RMS, max error, and kernel density estimation (KDE). Our results indicated that GFN2NMR outperformed the other two methods significantly (Fig. 5).

![Fig. 5. The linear relationship, $R^2$, MAE, RMS, max error, and KDE (D) between experimental chemical shifts and those predicted by GFN2NMR (A), CASCADE (B), and Ensemble NMR Prediction (C).](image)

**Comparison with DFT-based methods**

Firstly, we tested GFN2NMR on the compound set used to test regular GIAO DFT $^{13}$C NMR chemical shift calculation method with global scaling correction (GSC) and STS protocol. The accuracy of GFN2NMR ($R^2 = 0.998$, MAE = 1.65 ppm, RMS = 2.38 ppm) fell between the regular GSC method and STS protocol (Table 1).

Furthermore, structure assignment is a critical application of DFT $^{13}$C NMR chemical shift
calculation. Although Paton’s group has achieved some positive results with CASCADE in discriminating between constitutional isomers and diastereomers, to our knowledge, no data-driven method had ever been publicly benchmarked for structure assignment on the same baseline as DFT $^{13}$C NMR chemical shift calculation methods, prior to this study. We selected the thirteen pairs of incorrect/revised structures as the test set for constitutional isomers (test set B),$^{29}$ and twenty-five polychiral compounds as the test set for diastereomers (test set C),$^{47}$ which were used by Goodman’s group to assess the performance of DFT-based methods.

For test set B, GFN2NMR can assign all the thirteen revised structures correctly, while the DFT calculation only correctly assigned eleven structures (on $\text{mPW1PW91/6-311G*/B3LYP/6-31G*}$ level with $\text{M06-2X/def2-TZVP}$ calculated single-point energy) (Table 1). For the assignment of diastereomers in test set C, GFN2NMR can correctly identify 72% of the structures in this test set, which fell in the range of the correct rates of DFT-based methods (56%–84%) (Table 1).

Additionally, DFT-based methods without further correction have large errors in calculating heavy-atom-bearing carbons,$^{1–6,31,38}$ which caused low confidence in the assignment of those structures (e.g. $\text{S9a/S9b}$ and $\text{S13a/S13b}$ in test set B$^{29}$), while GFN2NMR had no such problem.

Table 1 The performance of GNF2NMR on the reported dataset, using DFT methods for comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy of calculation for 21 molecules in ref. 31</strong></td>
<td></td>
</tr>
<tr>
<td>GFN2NMR</td>
<td>$R^2 = 0.9980$, MAE = 1.65 ppm, RMS = 2.38 ppm</td>
</tr>
<tr>
<td>B3LYP/6-31G**//B3LYP/6-31G** (GSC)$^{31}$</td>
<td>$R^2 = 0.9973$, MAE = 2.15 ppm, RMS = 2.79 ppm</td>
</tr>
<tr>
<td>STS protocol$^{31}$</td>
<td>$R^2 = 0.9994$, MAE = 1.03 ppm, RMS = 1.35 ppm</td>
</tr>
<tr>
<td><strong>Assigning 13 pair of wrongly-proposed/revised molecules in ref. 29, by DP4 evaluation</strong></td>
<td></td>
</tr>
<tr>
<td>GFN2NMR</td>
<td>Correct = 13, Wrong = 0</td>
</tr>
<tr>
<td>$\text{mPW1PW91/6-311G*/B3LYP/6-31G*}$ (M06-2X/def2-TZVP single-point energy)$^{29}$</td>
<td>Correct = 11, Wrong = 2</td>
</tr>
<tr>
<td><strong>Correct rate of assigning 25 compounds from their diastereomers in ref. 47, by DP4 evaluation</strong></td>
<td></td>
</tr>
<tr>
<td>GFN2NMR</td>
<td>72%</td>
</tr>
<tr>
<td>B3LYP/6-311G**//B3LYP/6-31G**$^{47}$</td>
<td>60%</td>
</tr>
<tr>
<td>$\text{mPW1PW91/6-311G*/B3LYP/6-31G**}$</td>
<td>72%</td>
</tr>
<tr>
<td>$\text{M06-2X/6-311G*/B3LYP/6-31G**}$</td>
<td>72%</td>
</tr>
<tr>
<td>$\text{mPW1PW91/6-311G*/B3LYP/6-31G**}$ (M06-2X/6-31G**)</td>
<td>84%</td>
</tr>
</tbody>
</table>
Applying GFN2NMR on the J/K ring junction problem of MTX

Due to the high flexibility of structures 1a/1b and 3a/3b, we used the reported conformational analysis results for MTX to constrain the initial structures in conformational search, except for the conformation between K and L rings in structures 1b and 3b, as their C-56 are at axial position instead. However, the obtained global minimal conformers (1b-1) of 1b are tangled with intramolecular H-bonding, which may not form in the solvents with strong H-bond donors/acceptors (pyridine : methanol = 1 : 1). Thus, we redid the conformational search of 1b with constrained dihedral angle between K/L ring, to avoid intramolecular H-bonding, resulting in another set of unfolded global minimal conformers (1b-2). The conformational searches for structures 2a/2b were performed regularly. We then calculated the 13C NMR chemical shifts based on the screened conformational clusters of 1a/1b-1/1b-2, 2a/2b and 3a/3b by GFN2NMR (Tables S1–3). Notably, the 13C NMR calculation for 1a/1b(1b-1/1b-2) consumed nearly the same time as single-point calculation by GFN2-xtB (approximately 3.97 s per conformer) on a personal computer equipped with I5-12600KF.

Fig. 6. The errors of calculated chemical shifts of carbons closely relevant to J/K rings (at positions 48–55) of structures 1a and 1b (1b-1/1b-2).

Table 2 The MAE (in ppm), RMS (in ppm), and DP4 probability of the GFN2NMR-calculated 13C NMR chemical shifts of structures 1a/1b (1b-1/1b-2), 2a/2b, and 3a/3b, fitting to the experimental data of MTX, 2a, and 3a, respectively.

<table>
<thead>
<tr>
<th>Structure</th>
<th>MAE</th>
<th>RMS</th>
<th>DP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>1.79</td>
<td>2.34</td>
<td>100%</td>
</tr>
</tbody>
</table>
We assessed the calculated data with the experimental $^{13}$C NMR chemical shifts of MTX, 2a and 3a, respectively, via MAE, RMS and DP4 analysis. The results showed that GFN2NMR can accurately calculate the $^{13}$C NMR chemical shifts of structures 2a and 3a, and correctly assign them with 100% DP4 probability (Table 2). For MTX, the calculated data of originally proposed structure 1a showed significantly lower error, particularly for the carbons on J/K rings (Fig. 6). The DP4 analysis also yielded 100% confidence for the assignment of MTX as structure 1a. Therefore, we can confidently assign the configuration of J/K rings in MTX as the original proposed structure (1a).

**Discussion**

One of the key advantages of GFN2NMR is the deep architecture, that ensures the wide receptive field on the molecular graph and the extraction of high-level features of graph data. Notably, stereochemical information encoded in edge features can be directly perceived by every GCN layer, which probably reinforces the learning on stereochemical effects. Furthermore, GFN2-xTB, a semi-empirical method utilized in this model, provides high-quality geometric data, and introduces non-empirical features that may also contribute to its enhanced performance.

Like all the other data-driven approaches, however, GFN2NMR is still limited by the data used for training. Although we made efforts to collect the data of chemically diverse compounds and GFN2-xTB supports a broad range of elements, the results for some unique chemical systems still need further verification. In addition, GFN2NMR cannot deal with conformational energy accurately. For these cases sensitive to conformational population (e. g. nivario152), the users should always be cautious, and combine other methods including NOE or J-based analysis to draw correct conclusions.

**Experimental**

**Computation details.** CREST53 was used to search the conformational space of compounds on GFNFF54 level of theory, followed by optimization on GFN2-xTB41 level in gas phase, with 3 kcal/mol energy window to remove high energy conformers. The features of conformers in the energy window were equivalently weighted. All the DFT calculations were performed by Gaussian 1655 or ORCA 5.056 software packages. In the benchmark for data-driven methods, CASCADE was an online version in February, 2023 (http://nova.chem.colostate.edu/cascade/predict), and the Ensemble NMR Prediction was provided by MestReNova (version 14.3.2, released in 03/04/2023). In the benchmark for comparison with DFT methods, the GFN2NMR calculated $^{13}$C NMR chemical shift values were scaling corrected, and the DP4 probability of each candidate structure was calculated via the equation for DP4 in literature,28 where $\sigma$ was set to 2.22 based on the error

<table>
<thead>
<tr>
<th></th>
<th>1b-1</th>
<th>1b-2</th>
<th>2a</th>
<th>2b</th>
<th>3a</th>
<th>3b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b</td>
<td>1.99</td>
<td>2.60</td>
<td>0%</td>
<td>1.91</td>
<td>2.58</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>2a</td>
<td>1.43</td>
<td>1.65</td>
<td>100%</td>
<td>2.59</td>
<td>3.54</td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>1.91</td>
<td>2.77</td>
<td>100%</td>
<td>2.45</td>
<td>3.48</td>
</tr>
</tbody>
</table>
distribution of GFN2NMR performed in the external test sets.

**Conclusions**

In summary, we present a very deep GCN model (GFN2NMR) for $^{13}$C NMR chemical shift calculation that unprecedentedly offers both low time-cost and DFT-level performance in structure assignment. Based on GFN2-xTB, this model is compatible with a broad variety of organic structures. Therefore, it can be an ideal tool for assigning large or flexible organic molecules, and rapidly validating proposed structures. We demonstrated the utility of our model by resolving the J/K ring junction of MTX and its synthetic moieties. Our results illustrated that the power of NMR chemical shift calculation can extend to much larger systems than previously feasible using this model. Furthermore, we believe that GFN2NMR has the potential to be applied to the structure determination of bio-macromolecules, which would open up a fascinating avenue for future research.

**Author Contributions**

W.-J. A. and J. L. contributed equally in data collection, computation tasks, data process, results organization, and draft writing. D. C. contributed in the development and test of model. S. L. and G.-S. T. contributed in the supervision of work. Y.-Y. Y. and Y. L. contributed in data collection. K.-P. X., X. Y., F. K., and Z.-X. Z. contributed in the test/debug of software. W.-X. W. proposed the original idea, supervised this work, provided computational resources and most funding, developed the model, composed the codes of GFN2NMR, and wrote the major part of original draft. All authors discussed and commented on the manuscript.

**Conflicts of interest**

The authors declare no conflict of interest.

**Acknowledgements**

The work was financially supported by National Natural Science Foundation of China (No. 82173713, 81903513), China Postdoctoral Science Foundation (No. 2020M682599), Natural Science Foundation of Hunan Province (No. 2021JJ41000, 2021JJ40812), and Key Research and Development Program of Hunan Province (No. 2022SK2031).

**Notes and references**

37 T. Lin, Y. Wang, X. Liu and X. Qiu, *AI Open*, 2022, 3, 111.
39 SDBSWeb: https://sdbss.db.aist.go.jp


