Journal of Cheminformatics

RESEARCH

Open Access

GT-NMR: a novel graph transformer-based approach for accurate prediction of NMR chemical shifts

Haochen Chen[1], Tao Liang[1], Kai Tan[1], Anan Wu[1*] and Xin Lu[1*]

Abstract

In this work, inspired by the graph transformer, we presented an improved protocol, termed GT-NMR, which integrates 2D molecular graph representation with Transformer architecture, for accurate yet efficient prediction of NMR chemical shifts. The effectiveness of the GT-NMR was thoroughly examined with the standard nmrshiftdb2 dataset, 37 natural products and structural elucidation of 11 pairs of natural products. Systematical analysis affirms that GT-NMR outperforms traditional graph-based methods in all aspects, achieving state-of-the-art performance, with the mean absolute error of 0.158 and 1.189 ppm in predicting ^{1}H and ^{13}C NMR chemical shifts, respectively, for the standard nmrshiftdb2 dataset. Further scrutiny of its practical applications indicates that GT-NMR's efficacy is closely tied to molecular complexity, as quantified by the size-normalized spatial score (nSPS). For relatively simple molecules (nSPS < = 27.71), GT-NMR performs comparably to the best density functional while its effectiveness significantly diminishes with complex molecules characterized by higher nSPS values (nSPS > = 38.42). This trend is consistent across other graph-based NMR chemical shift prediction methods as well. Therefore, while employing GT-NMR or other graph-based methods for the rapid and routine prediction of NMR chemical shifts, it is advisable to utilize nSPS to assess their suitability. The source codes and trained model of GT-NMR are publicly available at GitHub.

Scientific contribution

GT-NMR, which combines the 2D molecular graph representation with the Transformer architecture, was implemented for the first time to predict atom-level NMR chemical shifts, achieving state-of-the-art performance. More importantly, the reliability of the GT-NMR and graph-based methods was assessed for the first time in terms of molecular complexity, as quantified by the size-normalized spacial score (nSPS). Systematical scrutiny demonstrated that GT-NMR offer a valuable way for routine application in structural screening and elucidation of relatively simple molecules.

Keywords NMR chemical shifts, Machine learning, Graph transformer, Transformer, Graph neural network, Molecular complexity

Introduction

Over the past few decades, nuclear magnetic resonance (NMR) spectroscopy has become one of the most powerful tools for the structural elucidation of organic compounds [1, 2]. However, interpreting NMR spectra is often complex and heavily reliant on the expertise of individual scientists. Consequently, misinterpretations of NMR spectra, which subsequently result in incorrect structural assignments, have persisted as a recurring

*Correspondence:
Anan Wu
ananwu@xmu.edu.cn
Xin Lu
xinlu@xmu.edu.cn
[1] Fujian Provincial Key Laboratory for Theoretical and Computational Chemistry, Departmental of Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, People's Republic of China

challenge [3–5]. In the absence of crystal structures, total synthesis has been unequivocally established as a key tool for structure elucidation. However, identifying a structure via total synthesis is a difficult task, presenting significant challenges and costs, especially for molecules with complex structure [3]. Quantum chemical calculations, which serve as an invaluable complement to experiments, can offer sufficient accuracy to discriminate the correct structure from a set of putative structures [6–10]. Despite their utility, these calculations often require considerable computational resources and time, particularly for complex molecules [7, 8]. Hence, there is a pressing need to develop methods capable of accurately and efficiently predicting NMR chemical shifts, particularly when rapid screening of structures is desired.

It has been generally accepted that the local molecular environment around a nucleus determines its chemical shifts. Thus, the empirical Hierarchical Organization of Spherical Environments (HOSE codes) [11], by systematically encoding the local and extended environment in a spherical manner, has achieved great success in the rapid prediction of NMR chemical shifts. In recent years, with the support of big data, deep learning (DL) techniques have demonstrated superior efficiency and accuracy compared to traditional empirical methods, and have achieved considerable successes in predicting various properties of molecules [12–17]. By encoding the local and extended environments of a nucleus into vector representations, Meiler et al. developed an artificial neural network with a mean deviation as low as 1.8 ppm for the prediction of ^{13}C chemical shifts [18]. However, a more natural and intuitive representation of a molecule in chemistry is a two-dimensional molecular graph, where atoms and bonds are treated as nodes and edges, respectively. Consequently, graph neural networks (GNNs), which have demonstrated substantial improvement in predicting various chemical properties of molecules [19–24], can be readily applied in the prediction of NMR chemical shifts. Jonas and Kuhn [25] reported the first instance of utilizing GNN to predict the ^{1}H and ^{13}C NMR chemical shifts of organic molecules, demonstrating that GNNs can indeed improve the prediction of NMR chemical shifts beyond the capability of conventional approaches. Kwon et al [26] applied a message passing neural network [19] (MPNN) with enhanced molecular graph representation and achieved better prediction performance for the ^{1}H and ^{13}C NMR chemical shifts of small molecules. Recently, they developed a scalable GNN (SG) with sparse graph representations and more effective messaging techniques, achieving the state-of-the-art (SOTA) performance with mean absolute error (MAE) of 0.216 and 1.271 ppm for ^{1}H and ^{13}C NMR chemical shifts, respectively [27].

Although chemical shifts are primarily considered as local property, long-range effects due to substitutions or structural changes can still have non-negligible impacts, particularly in conjugated systems [27–29]. Hence, accurate predictions of NMR chemical shifts require considerations of both local chemical environments and long-range effects. It is known that popular MPNNs are primarily designed to capture local relationship in graph structures, and their ability to capture long-range correlations is not as effective as their ability to handle local interactions [30, 31]. Moreover, traditional MPNNs inherit the limitations of the 1-Weisfeiler-Lehman (1-WL) test [32, 33], implying that they may struggle with graph structures that 1-WL test cannot differentiate, such as cycles with different size or certain types of tree structures.

Benefitting from the success of Transformers across various fields, extensions of Transformers for graph representation, namely Graph Transformers (GTs), have achieved significant success [33–37]. A primary motivation for adopting GTs is to address limitations associated with classic MPNNs. By integrating the transformer architecture into graph-based learning, GTs can inherently capture long-range correlations between nodes through the self-attention mechanism while also effectively describing local relationships as in traditional MPNNs. This makes them particularly suited for tasks where understanding the overall structure of the graph is crucial. To date, GTs have been extensively used in graph-based tasks [38–40], such as graph-level regression/classification, node-level classifications and link predictions. However, node-level regression has not yet been implemented to the best of our knowledge. In an effort to rapidly predict ^{1}H and ^{13}C NMR chemical shifts, we adopted GT to predict atom-level NMR chemical shifts by learning from the data, denoted as GT-NMR. The effectiveness of the GT-NMR model was thoroughly examined in the blind tests, compared with the density functional theory (DFT) calculations. More importantly, the reliability of the GT-NMR and GCN was assessed for the first time in terms of molecular complexity, represented by the size-normalized spacial score [41] (nSPS). This assertion concerning molecular complexity can be extended to other graph-based method, providing a valuable metric for evaluating their reliability in complex molecular systems.

Methods

Dataset

The nmrshiftdb2 used by Jonas and Kuhn [25] was employed in this work and can be accessed via https://jcheminf.biomedcentral.com/articles/https://doi.org/10.1186/s13321-019-0374-3. Molecules with more than 64

Table 1 Data Statistics

Data set	N_{mol}			N_{nonH}		N_{bond}	
	Train. set	Val. set	Test set	Range	Avg.	Range	Avg.
^{1}H NMR	10,252	1277	1277	[1, 46]	17.01	[0,48]	17.95
^{13}C NMR	21,516	2694	2695	[1, 44]	14.18	[0,48]	14.81

N_{mol} number of molecules, N_{nonH} number of heavy atoms per molecule, N_{bond} number of bonds without C-H per molecule

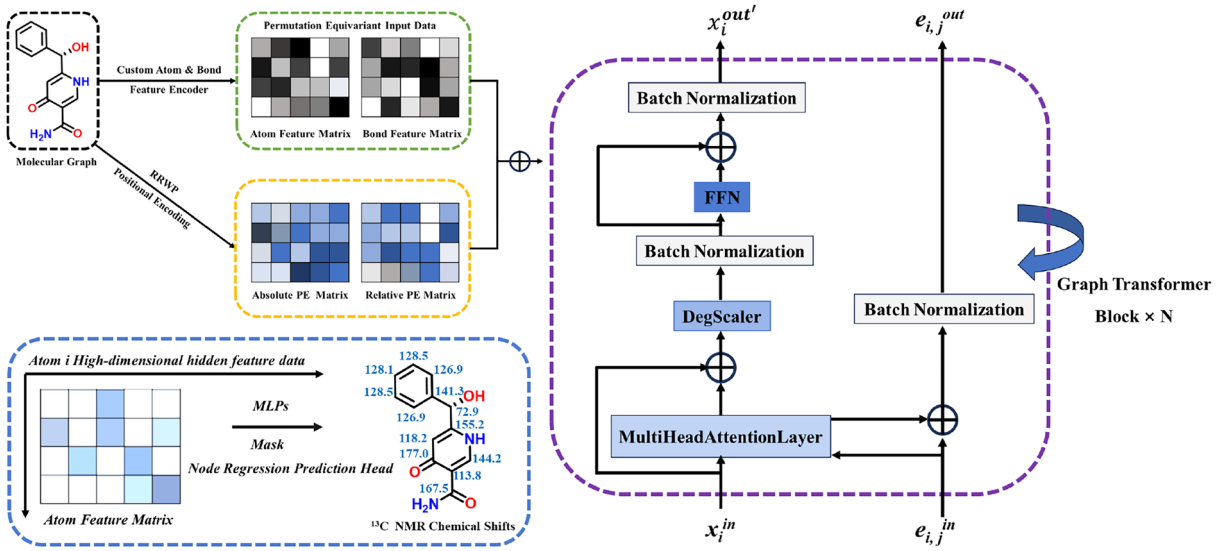

Fig. 1 Model architecture of GT-NMR. It contains permutation-equivariant graph encoding (feature encoder in green box and positional encoding in yellow box), graph transformer block (purple box) and node regression prediction head (blue box)

atoms were excluded, and the remaining molecules were identified with annotated ^{1}H or ^{13}C chemical shifts, containing only the elements of H, C, O, N, P, S, F, and Cl. In total, this dataset consists 12,806 molecules annotated with ^{1}H chemical shifts and 26,905 molecules annotated with ^{13}C chemical shifts. For model training, we used the same training set as Jonas and Kuhn [25]. The original test set was randomly split into validation and test sets in a 1:1 ratio. Consequently, the model was trained on the training set, selected against the validation set, and finally evaluated on the test set to assess the its effectiveness. The statistics of the data sets are listed in Table 1.

GT-NMR

GT-NMR is a graph-transformer-based architecture designed for node-level regression task. It enables end-to-end learning from graph representation of organic molecules to predict the ^{1}H and ^{13}C chemical shifts. For the ^{13}C chemical shifts, labels are directly associated with the corresponding carbon atoms, while for the ^{1}H chemical shifts, we used the same implicit treatment [26, 27] as adopted in the literature. Specifically, characteristics of the carbon atom attached to the hydrogen atom were used to train and make predictions of both ^{1}H and ^{13}C chemical shifts. Correspondingly, labels of hydrogen atoms on methylene were averaged. GT-NMR comprises four main modules as shown in Fig. 1: feature encoder, positional encoding, graph transformer block and node regression prediction head.

Feature encoder

Each molecule is represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the set of atoms (nodes) and bonds (edges), respectively. The node vectors $\boldsymbol{x_i} \in \mathcal{V}$ and edge vectors $\boldsymbol{e_{i,j}} \in \mathcal{E}$ are associated with heavy atoms and their all possible pairs in the molecule since the implicit model is employed. To construct the node vectors, atom features used by Jonas and Kuhn [25] were adopted. Edge_index instead of traditional adjacency matrices was employed to represent bond features as edge_index is a preferred choice for implementing GNN in PyTorch Geometric [42](PyG), particularly when dealing with complex and large-scale graph data. These features were calculated using RDKit [43], and details can be found

in the supplementary information (see Tables S1 and S2). The atom and bond encoders (Fig. 1, block in green box) then encode these vectors $\boldsymbol{x_i}$ and $\boldsymbol{e_{i,j}}$, in which each integer feature is embedded independently and aggregated to form the input vectors $\boldsymbol{x_i} \in \mathbb{R}^d$ and $\boldsymbol{e_{i,j}} \in \mathbb{R}^d$, where d represents the dimensionality of the embedding space.

Positional encoding

It has been demonstrated that the integration of relative random walk probabilities (RRWP) into Transformers when applied to graph is crucial, which can lead to improvements in performance across a range of tasks and applications involving complex graph structures [37]. In this work, the RRWP-based initial positional encoding [37] (PE) was employed.

Let $A \in \mathbb{R}^{n\times n}$ represents the adjacency matrix (generated from edge_index in PyG) of the graph$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n nodes, and D denotes the diagonal degree matrix of $\mathcal{G}$ The probability of reaching one node from another through a random walk can be defined by the matrix (**M**), $\boldsymbol{M}:=\boldsymbol{D}^{-1}\boldsymbol{A}$, in which $\boldsymbol{M}_{i,j}$ represents the probability that a random walk moves from node i to j in one step of a simple random walk. Hence, the initial edge PE for each pair of nodes $i, j \in \mathcal{V}$ can be defined as:

$$\mathcal{P}_{i,j} = \left[\mathbf{I}, \mathbf{M}, \mathbf{M}^2, \mathbf{M}^3, \ldots\ldots \mathbf{M}^{\mathrm{K}-1}\right]_{i,j} \in \mathbb{R}^K \tag{1}$$

where, **I** is the identity matrix, and the hyperparameter $K \in \mathbb{N}$ controls the maximum length of the random walks under consideration. For any node $i \in \mathcal{V}$, the diagonal element $\mathcal{P}_{i,i}$ can be utilized as an initial node encoding. The tensor **P** can be updated by an elementwise MLP:$\mathbb{R}^K \rightarrow \mathbb{R}^D$ to get new relative PEs.

Graph transformer block

A similar transformer layer proposed by Ma et al. [37]. was adopted in this work. The attention computation is defined as follows:

$$\hat{\boldsymbol{e}}_{\boldsymbol{i,j}} = \sigma\left\{\rho\left[\left(\boldsymbol{W_Q x_i} + \boldsymbol{W_K x_j}\right) \odot \boldsymbol{W_{Ew} e_{i,j}}\right] + \boldsymbol{W_{Eb} e_{i,j}}\right\} \in \mathbb{R}^{d'} \tag{2}$$

$$\alpha_{i,j} = \mathrm{Softmax}_{j\in\mathcal{V}}\left(\boldsymbol{W_A}\hat{\boldsymbol{e}}_{\boldsymbol{i,j}}\right) \in \mathbb{R} \tag{3}$$

$$\hat{\boldsymbol{x}}_i = \sum_{j\in\mathcal{V}} \alpha_{i,j}\left(\boldsymbol{W_V x_j} + \boldsymbol{W_{Ev}}\hat{\boldsymbol{e}}_{\boldsymbol{i,j}}\right) \in \mathbb{R}^d \tag{4}$$

$$\rho(\boldsymbol{x}) = \left\{[\mathrm{ReLU}(\boldsymbol{x})]^{\frac{1}{2}} - [\mathrm{ReLU}(-\boldsymbol{x})]^{\frac{1}{2}}\right\} \tag{5}$$

where σ is the activation function. In this work, the signed-square-root activation function was adopted; $\boldsymbol{W_Q}$,$\boldsymbol{W_k}$,$\boldsymbol{W_{Ew}}$,$\boldsymbol{W_{Eb}} \in \mathbb{R}^{d'\times d}$, $\boldsymbol{W_A} \in \mathbb{R}^{1\times d'}$ and $\boldsymbol{W_v}$,$\boldsymbol{W_{Ev}} \in \mathbb{R}^{d\times d'}$ are learnable parameters; ⊙ is the element-wise product; and ρ(x) is the signed-square-root, which stabilizes training by reducing the magnitude of large inputs.

With multiple heads (N_h heads) attention mechanism, the output is defined as:

$$\boldsymbol{x}_i^{out} = \sum_{h=1}^{N_h} \boldsymbol{W_O^h}\hat{\boldsymbol{x}}_i^h \in \mathbb{R}^d \tag{6}$$

$$\boldsymbol{e}_{i,j}^{out} = \sum_{h=1}^{N_h} \boldsymbol{W_{Eo}^h}\hat{\boldsymbol{e}}_{i,j}^h \in \mathbb{R}^d \tag{7}$$

where $\boldsymbol{W_O^h}$, $\boldsymbol{W_{Eo}^h}$ represent the weight matrices for each head $h \in \{1,2\ldots\ldots N_h\}$ and are also learnable parameters. To produce efficient aggregation and maintain degree information of node representations, an adaptive degree-scaler was applied to the attention mechanism as follows:

$$\boldsymbol{x}_i^{out'} := \boldsymbol{x}_i^{out} \odot \boldsymbol{\theta}_1 + \left[log(1+d_i)\boldsymbol{x}_i^{out} \odot \boldsymbol{\theta}_2\right] \in \mathbb{R}^d \tag{8}$$

where d_i is the degree of node i; $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \mathbb{R}^d$ are learnable parameters. Outputs $\boldsymbol{x}_i^{out'}$ and $\boldsymbol{e}_{i,j}^{out}$ are then subjected to 1-dimernsiontal Batch Normalization (Fig. 1, block in purple box) to ensure gradient stabilization. Subsequently, a feed-forward network is applied to map $\boldsymbol{x}_i^{out'}$ from $\mathbb{R}^d$ to $\mathbb{R}^{2d}$ and then back to $\mathbb{R}^d$. This process aims to enhance the model's expressivity in higher-dimensional space. The final outputs of graph transformer block $\boldsymbol{x}_i^{out'}$ and $\boldsymbol{e}_{i,j}^{out}$ are then fed into the node regression prediction head block for the prediction of NMR chemical shifts.

Node regression prediction head

Once the final node feature representations $\boldsymbol{x}_i^{out'}$ are generated through the graph transformer block, they are transformed via a three-layer MLP $\boldsymbol{x}_i^{out'} \in \mathbb{R}^d \rightarrow \mathbb{R}^{d/2} \rightarrow \mathbb{R}^{d/4} \rightarrow \mathbb{R}^1$ see Fig. 1, block in blue box) to yield the desired NMR chemical shifts. Note that a mask was applied to filter out irrelevant atoms, e.g. non-carbon atoms, defined as follows:

$$mask = \begin{cases} \boldsymbol{x_i} \text{ is True if } y_i \text{ is present and } i \text{ is a carbon atom else False (Train Stage)} \\ \boldsymbol{x_i} \text{ is True if } i \text{ is a carbon atom else False (Inference Stage)} \end{cases} \tag{9}$$

Table 2 Comparisons of baseline models and GT-NMR for the predictions of ^{1}H and ^{13}C chemical shifts on nmrshiftdb2-subset.

Method	δ(^{13}C)		δ(^{1}H)	
	MAE	RMSE	MAE	RMSE
HOSE[1]	2.85	–	0.33	–
GCN[1]	1.43	–	0.28	–
FCG[2]	1.355 ± 0.022	–	0.224 ± 0.002	–
Weakly-Supe. MPNN[3]	1.552 ± 0.056	–	0.243 ± 0.003	–
SG-only[4]	1.286 ± 0.010	2.273 ± 0.017	0.224 ± 0.002	0.509 ± 0.008
SG-IMP[4]	1.293 ± 0.007	2.266 ± 0.009	0.224 ± 0.002	0.508 ± 0.012
SG-IR[4]	1.282 ± 0.007	2.285 ± 0.043	0.215 ± 0.002	0.487 ± 0.005
SG-proposed[4]	1.271 ± 0.008	2.232 ± 0.018	0.216 ± 0.001	0.484 ± 0.005
GT-NMR(this work)	**1.189**	**2.206**	**0.158**	**0.293**

[1] . taken from ref. [25]; [2]. taken from ref. [26]; [3]. taken from ref. [45]; [4]. taken from ref. [27]

MAE mean absolute error, *RMSE* root mean squared error. Method with the best performance is highlighted in bold. Units in ppm

Training details

In GT-NMR, the l_1-loss is used as the training objective $\mathcal{J}$. More precisely, the targets are generated as follows:

$$\mathcal{J} = \mathcal{L}\left(y_i^{Pred}, y_i^{True}\right) \tag{10}$$

$$\mathcal{L}(MAE) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{Pred} - y_i^{True}\right| \tag{11}$$

where y_i^{Pred} and y_i^{True} are predicted and true values for atom i, respectively; n represents the effective number of labels in a batch; and $\mathcal{L}$ is the loss function used for regression, mean absolute error (MAE) in GT-NMR.

The current GT-NMR was built on PyTorch and PyG[42]. GraphGym[44] was adopted for experiment management and parameter searching. The source codes and trained model can be found at https://github.com/Anan-Wu-XMU/GT-NMR/.

Results and discussion

Model performance in general

To facilitate the evaluation of various methods, we selected the method proposed by Jonas and Kuhn as a baseline. For the Hose codes and GCN, results on the entire test set of nmrshiftdb2 were taken from reference [25]. The fully connected graph (FCG) [26], weakly-supervised MPNN [45] and SG series [27] results were also included for comparison. Note that only half of the original test set of nmrshiftdb2 was used to evaluate the performance of GT-NMR, while the other half was used as validation test. Table 2 presents the comparison results of the prediction accuracy of ^{1}H and ^{13}C chemical shifts. The best results for each dataset are highlighted in bold.

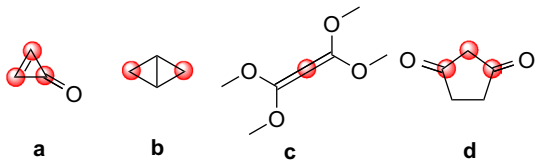

Fig. 2 Molecules for which performances of GT-NMR are unsatisfactory. Carbon atoms with error large than 10 ppm are highlighted with red circle.**a: cyclopropenone (MAE 22.03 ppm). b: bicyclo[1,1,0]butane (MAE 16.65 ppm). c: tetramethoxyallene(MAE 13.58 ppm). d: cyclopentane-1,3-dione (MAE 24.39 ppm)**

It is apparently that the GT-NMR method proposed in this work significantly outperformed other methods on both datasets. For the ^{13}C chemical shifts prediction, GT-NMR yielded the lowest MAE and RMSE of 1.189 and 2.206 ppm, respectively, followed by SG-proposed method, with an MAE and RMSE of 1.261 and 2.232 ppm, respectively. For the ^{1}H chemical shifts prediction, GT-NMR also achieved the lowest statistical metrices, establishing SOTA performance on both datasets. This indicates that the GT-NMR method can predict the NMR spectrum of new molecules more accurately than other methods.

Large prediction errors mainly occurred for molecules with uncommon chemical structure or those highly underrepresented in the calibration. For instance, large prediction errors were observed for cyclopropenone (a, MAE: 22.03 ppm) and bicycle [1,1,0]butane (b, MAE: 16.65 ppm), both of which have exotic structures as shown in Fig. 2. Additionally, the tetramethoxyallene skeleton (**c** in Fig. 2) was completely absent from the training set, resulting in poor performance with an MAE of 13.58 ppm. The large prediction error for cyclopentane-1,3-dione (CPD, d in Fig. 2, MAE: 24.39 ppm) was

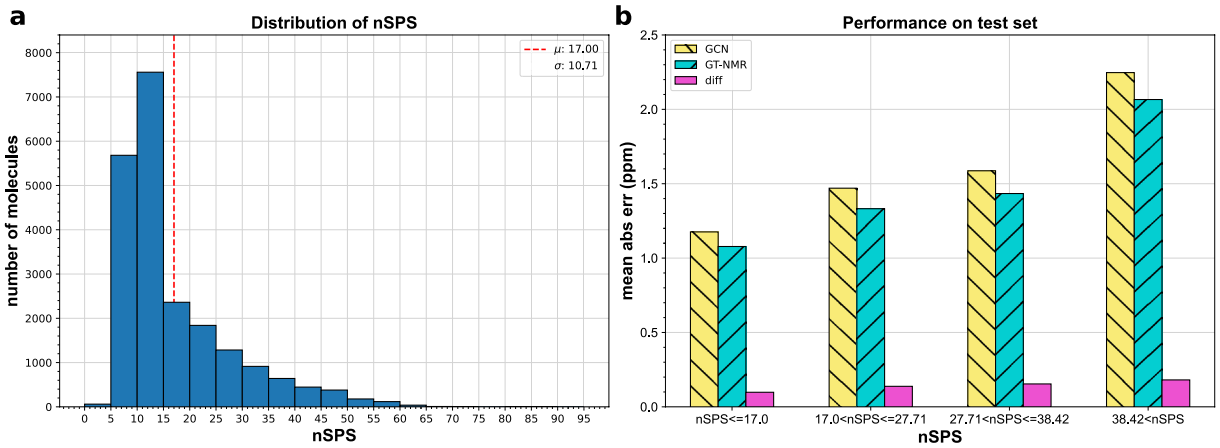

Fig. 3 Statistic on nSPS, where the larger the nSPS value, the more complex the molecule. **a** Distribution of nSPS values in the training set, with average (μ) and standard deviation(σ) of 17.00 and 10.71, respectively. **b** Prediction errors of GCN (Yellow) and GT-NMR (Cyan), and the differences between GCN and GT-NMR (Purple) with respect to nSPS in the test set. GCN was retained with the same dataset and settings as those in reference [25]

unexpected. However, a previous study [46] has pointed out that CPD is a mixture of rapidly interconverting tautomers, complicating the interpretation of its ^{13}C NMR spectra. Consequently, a possible reference error may occur. To calibrate the experimental ^{13}C chemical shifts of CPD, a benchmark calculation at FPA-M [7] level of theory was conducted. This method has been shown to yield chemical shifts with the accuracy of CCSD(T) complete basis set limit [7, 8]. As expected, large deviations were observed between the experimental values and the calculated ^{13}C chemical shifts (see Table S3). Given the high accuracy of the FPA-M method and experimental values of similar molecules, we believe that there is indeed a reference error in the ^{13}C chemical shifts of CPD. Taking the FPA-M results as reference, the MAE of GT-NMR for CPD is dramatically reduced to a reasonable value of 3.25 ppm.

Model analysis

As mentioned above, traditional MPNNs may struggle with complex graph structures, and the primary motivation for adopting GTs is to address limitations associated with classic MPNNs. To assess the complexity of molecules, a recent proposed scoring metric, the size-normalized spacial score (nSPS) [41] was employed, where the larger the nSPS, the more complex the molecule. This metric has been shown to effectively reflect the chemist's intuitive assessment of molecular complexity and is applicable to both natural products and synthetic compounds [41]. Figure 3 present statistics associated with nSPS.

As seen in Fig. 3a, the nSPS distribution in the training set is highly biased toward simple molecules, with majority of molecules (66.57%) having an nSPS less than the average value (17.00). This implies that trained models may not perform well for complex molecules. This is evidenced by the increased prediction errors with increasing nSPS in the test set. As illustrated in Fig. 3b, both GCN (Yellow) and GT-NMR (Cyan) performed well for simple molecules (nSPS < =17.00), with MAEs of 1.176 and 1.078 ppm, respectively. However, when applied to complex molecules (nSPS > 38.42), the prediction errors increased significantly, nearly doubling those for simple molecules, with MAEs of 2.247 (GCN) and 2.066 (GT-NMR) ppm, respectively. Such large errors mainly attributed to two factors. Firstly, the labelled data for complex molecules are highly underrepresented in the calibration, as shown in Fig. 3a, resulting in poor performance of the trained models on complex molecules. Secondly, complex molecules often consist of multiple chiral centers (see Figure S1), while neither the annotated data nor the atomic features include stereochemical information. Consequently, it is not surprising that neither GCN nor GT-NMR performed satisfactorily on complex molecules.

Encouragingly, GT-NMR outperformed traditional GCN method throughout the entire test set, with the performance gap (Purple in Fig. 3b) between these two methods becoming more pronounced as molecular complexity increases. For example, when nSPS < =17.0, the difference in the predicted MAE between GT-NMR and GCN is 0.098 ppm, increasing to 0.138 ppm when 17.0 < nSPS < =27.71, and finally reaching 0.181 ppm when nSPS > 38.42. This result affirms the improved performance of GT-NMR in dealing with complex graph structures compared to traditional graph-based

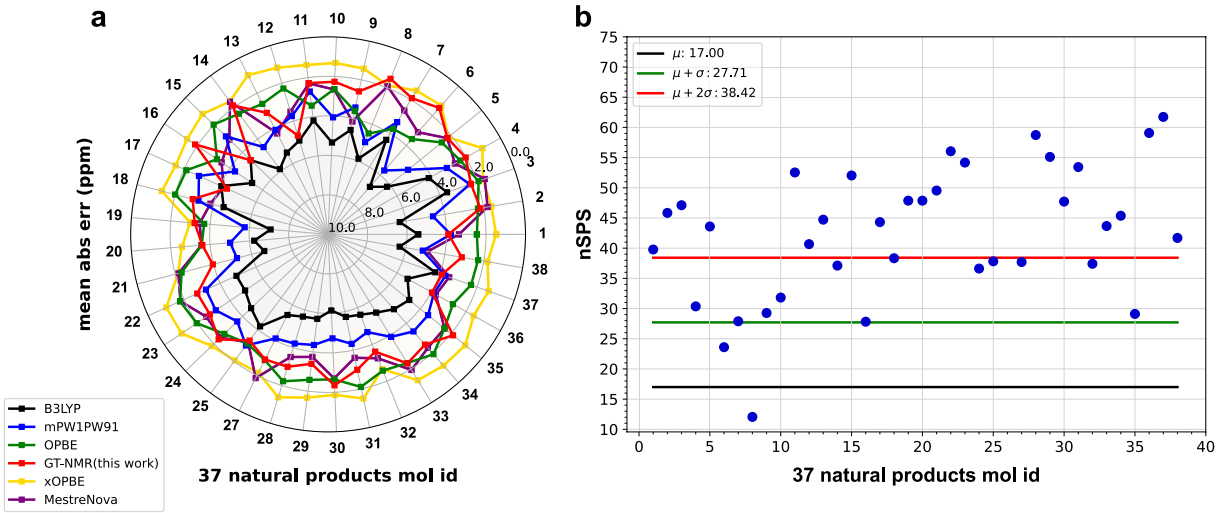

Fig. 4 Results of the selected 37 natural products. Detailed results can be found in Table S4 and the corresponding structures are presented in Figure S2. **a** Radar plot comparing the mean absolute errors (MAEs) of ^{13}C chemical shifts predicted by various methods. Each spoke represents a different natural product. The proximity to the center of the plot indicates the magnitude of the error for each method, with closer to the center indicating less accuracy. xOPBE (Gold); OPBE (Green); mPW1PW91(Blue); B3LYP(Black); MestreNova (Purple); GT-NMR (Red); **b** nSPS values of the 37 natural products.μ: the average nSPS in the training set; σ: the standard deviation of nSPS in the training set

methods. It should be noted that stereochemical effects and geometry-specific effects were not considered with the standard nmrshiftdb2. Hence, GT-NMR, built upon this dataset, will be only effective for determination and verification of constitution, but is inherently incapable of assigning relative configurations. Further improvement of GT-NMR's performance on complex molecules requires not only increasing the annotated data of complex molecules, but also incorporating stereochemical information into the atomic features and bond features. These enhancements will be left for further investigation in future studies.

Comparison with ab initio methods

To further assess the accuracy of GT-NMR in practical applications, we selected 37 natural products from a previous study [9] and compared the results with those obtained using DFT methods (see the supplementary information for computational details). Detailed results can be found in Table S4, and the corresponding structures are presented in Figure S2. Figure 4 shows a radar plot comparing MAEs of ^{13}C chemical shifts predicted by various methods (Fig. 4a) and nSPS values of the 37 natural products (Fig. 4b). Predictions by MestReNova [47] were also included for comparison. Each spoke on the radar plot represents a different natural product, with proximity to the center indicating the magnitude of the error (MAE) for each method, where closer to the center indicates less accuracy.

As shown in Fig. 4a, xOPBE functional (Gold) significantly outperforms the other methods. It consistently exhibits lower errors across nearly all molecules, demonstrating its enhanced accuracy in predicting ^{13}C chemical shifts. On the other hand, GT-NMR (Red) is highly dependent on the molecular complexity. For relatively simple molecules (nSPS < =27.71, points below the green line in Fig. 4b), the prediction accuracy of GT-NMR is comparable to that of the xOPBE functional. For instance, compound 8 and 6, with nSPS values of 12.05 and 23.6, respectively, have their ^{13}C chemical shifts accurately predicted by GT-NMR with MAE of 1.512 and 1.474 ppm (see Table S4), while the corresponding values predicted by the xOPBE functional are 1.878 and 1.313 ppm. For complex molecules (nSPS > 38.42, points above the red line in Fig. 4b), the performance of GT-NMR is clearly inferior to xOPBE functional, and even to the OPBE functional, with MAEs typically higher than 2.0 ppm (see Figures S3 and S5).

Nonetheless, the overall performance of GT-NMR, with an average MAE of 2.867 ppm for the 37 natural products, is superior to the commonly used B3LYP (5.284 ppm) and mPW1PW91 (3.982 ppm) functionals, and comparable to the OPBE functional (2.646 ppm) and MestreNova (2.950 ppm). Hence, GT-NMR can be applied with high confidence in rapid structural screening of relatively simple molecules (nSPS < =27.71). However, for complex molecules (nSPS > 38.42), the more accurate xOPBE functional is preferred, although GT-NMR is more efficient. Notably, predicting the ^{13}C

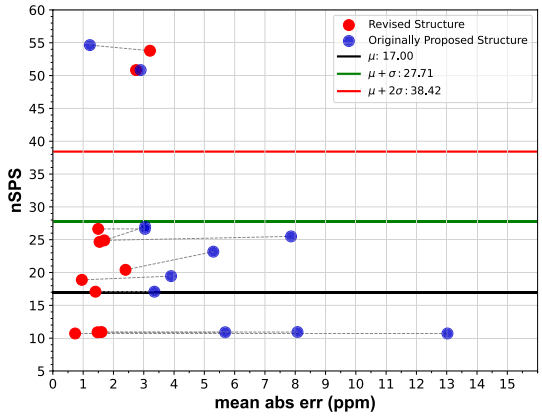

Fig. 5 GT-NMR results for the 11 natural product pairs. The blue points represent the originally proposed structures and the red points represent the revised structures. Each pair is connected with a dashed line

chemical shifts of a new molecule using GT-NMR takes an average of 0.025 s on a single RTX 4090 GPU.

Structure elucidation

As the purpose of this work is to establish an accurate yet efficient model to predict NMR chemical shifts for rapid structure screening, we further examined 11 natural products [48] that had been incorrectly assigned along with their revised structures using total synthesis, to illustrate the capabilities of GT-NMR. Detailed results can be found in Table S5 and S6, and the structures are presented in Figure S4.

Figure 5 presents a comparative analysis of prediction errors for two different molecular structures (originally proposed and revised) across 11 natural products pairs. The scatter plot differentiates between the originally proposed structure (blue points) and the revised structure (red points) based on their MAEs against nSPS values. The data points are connected with dashed lines to highlight comparisons. It is evidently that for relatively simple molecules (nSPS < =27.71), the revised structures (red points) consistently show lower MAEs compared to the originally proposed structures (blue points), indicating that GT-NMR can successfully discriminate correct structures from incorrect ones. However, as nSPS values increase (nSPS > 38.42), it becomes challenging for GT-NMR to determine the correct structure, or it may even give the incorrect assignment, as shown in Fig. 5.

These examples further strengthen the argument that while the GT-NMR method developed in this work exhibits consistent reliability for relatively simple molecules (nSPS < =27.71), it becomes less reliable for complex molecules (nSPS > 38.42). The nSPS values can be effectively utilized to assess the reliability of the GT-NMR model. Hence, it is necessary to combine the GT-NMR with nSPS values for rapid structure screening.

Conclusion

With the advent of deep learning, graph neural network-based methods have been extensively used in predicting NMR chemical shifts, and have achieved considerable success. However, traditional graph neural network-based methods have inherent limitations, making them challenging to handle complex graph structures. How to rationally apply these graph neural network-based methods in chemical shifts prediction appears to be underestimated and overlooked. In this work, inspired by the graph transformer, we presented an improved method, denoted as GT-NMR, which combines molecular graph representation with Transformers for accurate yet efficient prediction of NMR chemical shifts. The effectiveness of GT-NMR was thoroughly examined with the standard nmrshiftdb2, 37 natural products and structural elucidation of 11 pairs natural products. More importantly, the reliability of the GT-NMR was assessed for the first time in terms of molecular complexity, represented by the size-normalized spacial score (nSPS).

Regarding to the nmrshiftdb2 database, the GT-NMR method achieved state-of-the-art performance with mean absolute error of 0.158 and 1.189 ppm in the prediction of ^{1}H and ^{13}C NMR chemical shifts, respectively, which are clearly superior to the performance of existing methods (GCN, MPNN, weakly-super MPNN, and SG). Detailed analysis affirms the improved performance of the GT-NMR in dealing with complex graph structures compared to traditional graph-based methods. In the practical application of predicting ^{13}C Chemical Shifts of 37 natural products, the GT-NMR method also demonstrated its reliability. It outperforms the commonly used mPW1PW91 and B3LYP functionals, and is only inferior to the xOPBE, a specialized functional for accurate prediction of ^{13}C Chemical Shifts. Close inspections revealed that the performance of GT-NMR is highly dependent on molecular complexity. For relatively simple molecules (nSPS < =27.71), GT-NMR yields reliable results and can be applied with high confidence. However, for complex molecules represented by high nSPS values, its performance is significantly degraded due to a lack of stereochemical information and sufficient representative data. This argument was further strengthened by the subsequent structural elucidation of 11 natural product pairs. In combination with nSPS, we believe that GT-NMR will be a valuable tool for routine application in structural screening and elucidation of relatively simple molecules. Future work will focus on the rational construction of

datasets that include more complex molecules and incorporation stereochemical information into the model to improve its applicability domain.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00927-9.

Supplementary material 1. Fig. 1. Model architecture of GT-NMR. Fig. 2. Molecules for which performances of GT-NMR are unsatisfactory. Fig. 3. Statistic on nSPS, where the larger the nSPS value, the more complex the molecule. Fig. 4. Results of the selected 37 natural products. Fig. 5. GT-NMR results for the 11 natural product pairs.

Acknowledgements
The authors thank the Department of Chemistry of Xiamen University and Fujian Provincial Key Laboratory for Theoretical and Computational Chemistry for use of their facilities and services.

Author contributions
A.A. Wu and X. Lu conceptualized the research. H.C. Chen conducted the research and constructed the model. H.C. Chen, T. Liang and K. Tan drafted the manuscript. All authors commented and revised the manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Nos. 92161117 and 22373079) and the Natural Science Foundation of Fujian Province of China (Grant No. 2021J01020).

Availability of data and materials
The source codes and trained model are publicly available in the Github repository https://github.com/AnanWu-XMU/GT-NMR. Furthermore, the training and test data used for this study can be downloaded from https://github.com/AnanWu-XMU/GT-NMR/tree/main/datasets.

Declarations

Competing interests
The authors declare no competing interests.

Received: 18 June 2024 Accepted: 8 November 2024
Published online: 26 November 2024

References

1. Breton RC, Reynolds WF (2013) Using NMR to identify and characterize natural products. Nat Prod Rep 30:501. https://doi.org/10.1039/c2np20104f
2. Sanders JK, Hunter BK (1993) Modern NMR Spectroscopy. Oxford University Press, Oxford, UK
3. Nicolaou KC, Snyder SA (2005) Chasing molecules that were never there: misassigned natural products and the role of chemical synthesis in modern structure elucidation. Angew Chem Int Ed 44:1012–1044. https://doi.org/10.1002/anie.200460864
4. Suyama TL, Gerwick WH, McPhail KL (2011) Survey of marine natural product structure revisions: a synergy of spectroscopy and chemical synthesis. Bioorg Med Chem 19:6675–6701. https://doi.org/10.1016/j.bmc.2011.06.011
5. Chhetri BK, Lavoie S, Sweeney-Jones AM, Kubanek J (2018) Recent trends in the structural revision of natural products. Nat Prod Rep 35:514–531. https://doi.org/10.1039/C8NP00011E
6. Sarotti AM, Pellegrinet SC (2009) A Multi-standard approach for GIAO ^{13}C NMR Calculations. J Org Chem 74:7254–7260. https://doi.org/10.1021/jo901234h
7. Sun M, Zhang IY, Wu A, Xu X (2013) Accurate prediction of nuclear magnetic resonance shielding constants: towards the accuracy of CCSD(T) complete basis set limit. J Chem Phys 138:124113. https://doi.org/10.1063/1.4796485
8. Wang K, Sun M, Cui D et al (2018) Accurate prediction of nuclear magnetic resonance shielding constants: an extension of the focal-point analysis method for magnetic parameter calculations (FPA-M) with improved efficiency. J Chem Phys 149:184101. https://doi.org/10.1063/1.5041979
9. Zhang J, Ye Q, Yin C et al (2020) xOPBE: a specialized functional for accurate prediction of ^{13}C chemical shifts. J Phys Chem A 124:5824–5831. https://doi.org/10.1021/acs.jpca.0c02873
10. Wu A, Ye Q, Zhuang X et al (2023) Elucidating structures of complex organic compounds using a machine learning model based on the ^{13}C NMR chemical shifts. Precis Chem 1:57–68. https://doi.org/10.1021/prechem.3c00005
11. Bremser W (1978) Hose — a novel substructure code. Analytica Chimica Acta 103:355–365. https://doi.org/10.1016/S0003-2670(01)83100-7
12. He J, You H, Sandström E et al (2021) Molecular optimization by capturing chemist's intuition using deep neural networks. J Cheminform 13:26. https://doi.org/10.1186/s13321-021-00497-0
13. Howarth A, Ermanis K, Goodman JM (2020) DP4-AI automated NMR data analysis: straight from spectrometer to structure. Chem Sci 11:4351–4359. https://doi.org/10.1039/D0SC00442A
14. Sturm N, Mayr A, Le Van T et al (2020) Industry-scale application and evaluation of deep learning for drug target prediction. J Cheminform 12:26. https://doi.org/10.1186/s13321-020-00428-5
15. He J, Nittinger E, Tyrchan C et al (2022) Transformer-based molecular optimization beyond matched molecular pairs. J Cheminform 14:18. https://doi.org/10.1186/s13321-022-00599-3
16. Martinez-Mayorga K, Rosas-Jiménez JG, Gonzalez-Ponce K et al (2024) The pursuit of accurate predictive models of the bioactivity of small molecules. Chem Sci 15:1938–1952. https://doi.org/10.1039/D3SC05534E
17. Kotlyarov R, Papachristos K, Wood GPF, Goodman JM (2024) Leveraging language model multitasking to predict C-H borylation selectivity. J Chem Inf Model 64:4286–4297. https://doi.org/10.1021/acs.jcim.4c00137
18. Meiler J, Meusinger R, Will M (2000) Fast determination of ^{13}C NMR chemical shifts using artificial neural networks. J Chem Inf Comput Sci 40:1169–1176. https://doi.org/10.1021/ci000021c
19. J Gilmer, SS Schoenholz, PF Riley, et al (2017) Neural message passing for Quantum chemistry. In: Proceedings of the 34th international conference on machine learning - 70. JMLR.org, Sydney, NSW, Australia
20. Jiang D, Sun H, Wang J et al (2022) Out-of-the-box deep learning prediction of quantum-mechanical partial charges by graph representation and transfer learning. Brief Bioinform. https://doi.org/10.1093/bib/bbab597
21. Jiang D, Wu Z, Hsieh C-Y et al (2021) Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. J Cheminform 13:12. https://doi.org/10.1186/s13321-020-00479-8
22. Walter M, Webb SJ, Gillet VJ (2024) Interpreting neural network models for toxicity prediction by extracting learned chemical features. J Chem Inf Model 64:3670–3688. https://doi.org/10.1021/acs.jcim.4c00127
23. Zhao Q, Anstine DM, Isayev O, Savoie BM (2023) Δ^2 machine learning for reaction property prediction. Chem Sci 14:13392–13401. https://doi.org/10.1039/D3SC02408C
24. Zou Z, Zhang Y, Liang L et al (2023) A deep learning model for predicting selected organic molecular spectra. Nat Comput Sci 3:957–964. https://doi.org/10.1038/s43588-023-00550-y
25. Jonas E, Kuhn S (2019) Rapid prediction of NMR spectral properties with quantified uncertainty. J Cheminform 11:50. https://doi.org/10.1186/s13321-019-0374-3
26. Kwon Y, Lee D, Choi Y-S et al (2020) Neural message passing for NMR chemical shift prediction. J Chem Inf Model 60:2024–2030. https://doi.org/10.1021/acs.jcim.0c00195
27. Han J, Kang H, Kang S et al (2022) Scalable graph neural network for NMR chemical shift prediction. Phys Chem Chem Phys 24:26870–26878. https://doi.org/10.1039/D2CP04542G
28. Neuvonen K, Fülöp F, Neuvonen H et al (2001) Substituent influences on the stability of the ring and chain tautomers in 1,3- O, N -heterocyclic systems: characterization by ^{13}C NMR chemical shifts, pm3 charge densities,

and isodesmic reactions. J Org Chem 66:4132–4140. https://doi.org/10.1021/jo001114w
29. Neuvonen H, Neuvonen K, Fülöp F (2006) Substituent cross-interaction effects on the electronic character of the CN bridging group in substituted benzylidene anilines − models for molecular cores of mesogenic compounds. A ^{13}C NMR study and comparison with theoretical results. J Org Chem 71:3141–3148. https://doi.org/10.1021/jo0600508
30. Villar S, Priebe C et al (2022) From local to global: spectral-inspired graph neural networks. arXiv. https://doi.org/10.48550/arXiv.2209.12054
31. Oono K, Suzuki T (2020) Graph neural networks exponentially lose expressive power for node classification. arXiv. https://doi.org/10.48550/arXiv.1905.10947
32. Morris C, Ritzert M, Fey M et al (2019) Weisfeiler and leman go neural: higher-order graph neural networks. arXiv. https://doi.org/10.48550/arXiv.1810.02244
33. Ying C, Cai T, Luo S et al (2021) Do transformers really perform badly for graph representation? arXiv. https://doi.org/10.48550/arXiv.2106.05234
34. Shi Y, Zheng S, Ke G et al (2022) Benchmarking graphormer on large-scale molecular modeling datasets. arXiv. https://doi.org/10.48550/arXiv.2203.04810
35. Luo S, Li S, Zheng S et al (2022) Your transformer may not be as powerful as you expect. arXiv. https://doi.org/10.48550/arXiv.2205.13401
36. Zhang B, Luo S, Wang L, He D (2023) Rethinking the expressive power of GNNs via graph biconnectivity. arXiv. https://doi.org/10.48550/arXiv.2301.09505
37. Ma L, Lin C, Lim D et al (2023) Graph inductive biases in transformers without message passing. arxiv. https://doi.org/10.48550/arXiv.2305.17589
38. Dwivedi VP, Joshi CK, Luu AT et al (2022) Benchmarking graph neural networks. arXiv. https://doi.org/10.48550/arXiv.2003.00982
39. Hu W, Fey M, Zitnik M et al (2021) Open graph benchmark: datasets for machine learning on graphs. arXiv. https://doi.org/10.48550/arXiv.2005.00687
40. Hu W, Fey M, Ren H et al (2021) OGB-LSC: a large-scale challenge for machine learning on graphs. arXiv. https://doi.org/10.48550/arXiv.2103.09430
41. Krzyzanowski A, Pahl A, Grigalunas M, Waldmann H (2023) Spacial score—a comprehensive topological indicator for small-molecule complexity. J Med Chem 66:12739–12750. https://doi.org/10.1021/acs.jmedchem.3c00689
42. https://pytorch-geometric.readthedocs.io/,
43. The RDKit: Open-source cheminformatics software; https://www.rdkit.org.
44. You J, Ying R, Leskovec J (2021) Design space for graph neural networks. arXiv. https://doi.org/10.48550/arXiv.2011.08843
45. Kang S, Kwon Y, Lee D, Choi Y-S (2020) Predictive modeling of NMR chemical shifts without using atomic-level annotations. J Chem Inf Model 60:3765–3769. https://doi.org/10.1021/acs.jcim.0c00494
46. Ballatore C, Soper JH, Piscitelli F et al (2011) Cyclopentane-1,3-dione: a novel isostere for the carboxylic acid functional group. Application to the design of potent thromboxane (A2) receptor antagonists. J Med Chem 54:6969–6983. https://doi.org/10.1021/jm200980u
47. Willcott MR (2009) MestReNova. J Am Chem Soc 131:13180–13180. https://doi.org/10.1021/ja906709t
48. Elyashberg M, Tyagarajan S, Mandal M, Buevich AV (2023) Enhancing efficiency of natural product structure revision: leveraging CASE and DFT over total synthesis. Molecules 28:3796. https://doi.org/10.3390/molecules28093796

Publisher's Note