Journal of Magnetic Resonance 368 (2024) 107795

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Chemical shift prediction in ^{13}C NMR spectroscopy using ensembles of message passing neural networks (MPNNs)

D. Williamson [b], S. Ponte [b], I. Iglesias [b], N. Tonge [b], C. Cobas [b], E.K. Kemsley [a,*]

[a] University of East Anglia, Norwich Research Park, NR7 6TJ, United Kingdom
[b] Mestrelab Research SL, C/Feliciano Barrera, 9B-Bajo, 15706 Santiago de Compostela, Spain

A B S T R A C T

This study reports a deep learning approach that utilises message passing neural networks (MPNNs) for predicting chemical shifts in ^{13}C NMR spectra of small molecules. MPNNs were trained on two distinct datasets: one with approximately 4000 labelled structures and another with over 40,000. To reduce stochastic variation, an ensemble framework was implemented, which is simple to deploy on multiple nodes of a High-Performance Computing facility.

The results emphasise the critical role of training set size and diversity. While prediction performance was comparable on test sets drawn from each dataset, the ensemble trained on the larger dataset retained its accuracy when these sets were crossed over, and when applied to a further collection of approximately 12,000 previously unseen structures introduced after all development work had been completed. In contrast, the ensemble trained on the smaller dataset showed a notable decline in generalisation ability. This difference is attributed to the greater diversity of atomic environments captured in the larger dataset.

The larger dataset also enabled more robust modelling of various error properties, providing a quantitative foundation for spectral assignment and verification. This was achieved in two ways. First, a clear relationship was observed between prediction errors and the frequency of different node feature vectors in the training data, allowing error estimates to be associated with individual nodes based on their type. These estimates can be used as weights in a modified cityblock distance metric when assigning observed to predicted shifts. Second, the mean absolute prediction error calculated at the structure level is well-fitted by a Gaussian kernel cumulative distribution. This enabled a probabilistic assessment of whether the predicted shifts and assigned observations are consistent with originating from the same molecular structure.

1. Introduction

Small organic molecules are fundamental building blocks in synthetic chemistry and biochemistry. Despite their small size, the number of possible molecular structures arising from different combinations of atoms, bonds, and functional groups is immense. Navigating this vast search space to discover, identify, or design molecules with desired properties requires the integration of experimental techniques with innovative computational methods [1].

High-resolution nuclear magnetic resonance (NMR) spectroscopy is the gold standard for elucidating molecular structure. However, accurate prediction of chemical shifts, invaluable for automated structural assignment and verification, remains a significant challenge. Manual interpretation of NMR spectra is time-consuming, labour-intensive, and prone to human error. Modern AI methods have clear potential to mitigate these issues and are the focus of the present study.

Molecular graphs serve as basic representations of molecules. Until recently, information was typically extracted from such graphs by analysis of the properties ('features') associated with the atoms ('nodes'). Whilst successful up to a point, a limitation of this approach is its failure to directly exploit the atom connectivity information given by the graph adjacency matrix.

The emergence of message passing neural networks (MPNNs) marked a significant advancement [2,3]. A form of convolutional neural network, MPNNs provide a deep learning architecture for analysing collections of intact graphs, that utilises the adjacency matrix in tandem with the feature set, allowing modelling of intricate interactions between nodes. In the cheminformatics context, this yields more refined descriptions of atomic environments, potentially improving models of relationships between structure and experimental observations. The approach has rapidly gained traction, and a variety of applications of MPNNs have been reported [4–6].

Recent studies have shown that graph neural networks can improve the prediction of NMR chemical shifts beyond the capability of conventional approaches [7–9]. In the present work, we disclose an MPNN approach for predicting high-resolution ^{13}C NMR chemical shifts from molecular structures. The framework employs ensembles of MPNNs trained on collections of assigned NMR spectra. By exploiting the power of MPNNs coupled with ensemble learning, we aim to achieve more accurate and reliable chemical shift predictions, with the eventual goal

* Corresponding author.
E-mail address: k.kemsley@uea.ac.uk (E.K. Kemsley).

https://doi.org/10.1016/j.jmr.2024.107795
Received 10 September 2024; Received in revised form 15 October 2024; Accepted 25 October 2024
Available online 28 October 2024
1090-7807/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

of improving automated NMR spectral interpretation, assignment and verification.

2. Material and methods

2.1. Computational facilities

Code development to prepare the training data and establish the MPNN framework was carried out on a Dell XPS PC (11th Gen Intel(R) Core (TM) i9-11900H @ 2.50 GHz, 32 Gb RAM). Subsequent ensemble training was carried out on a High-Performance Computing (HPC) cluster equipped with GPU arrays (28 nodes, includes Nvidia Quadro RTX6000s).

2.2. Software

Coding on the Dell XPS PC was carried out in Matlab R2023a Update 5 (The Mathworks, Cambridge) making use of its Deep Learning toolbox. For molecular graph creation and feature set extraction, calls were made from Matlab to a Python 3.10 environment installed with RDKit version 2023.03.3 [10]. MPNN training carried out on the cluster used Matlab R2022b Update 1. Ensemble learning carried out on the cluster used a Python 3.10 virtual environment.

2.3. Datasets

Three datasets were used in the present study. 'Dataset 1' comprises a small subset of records taken from the 'nmrshiftdb2' database [11]. This was used to develop the AI framework and to demonstrate the approach. The motivation for using publicly available data was to enable others to easily reproduce our work. 'Dataset 2' is an extensive collection taken from a commercially sensitive, proprietary database held by Mestrelab Research SL (Santiago de Compostela, Spain). Results are presented from this collection that highlight the advantage of using a much larger training set. Finally, a further proprietary collection was made available to the authors once all model development was complete, from which 'Dataset 3' was extracted to serve as an independent test set for predictive models obtained from the other two sets. Summaries of compositional aspects of each dataset are given in Supp Fig. S1.

2.3.1. Dataset 1: Extracted from nmrshiftdb

A collection of 3703 records ('mol files') were extracted from the 'nmrshiftdb2' database. Records were selected by the following criteria (see Supp Fig. S2 for additional details):

(a) Complete assignments – chemical shift values were present for all active nuclei.
(b) No organometallics – structures containing any metallic elements were excluded.
(c) Small molecules only – structures with 100 or more heavy atoms were excluded.
(d) Data integrity – records were discarded due to miscellaneous errors (e.g. RDKit import failures, typos, e.g. implausible chemical shift values).

Using Python RDKit utilities, a molecular graph was generated from each record, comprising an adjacency matrix to describes the node connectivities and a 52-element feature vector associated with each heavy atom in the structure. These matrix-based representations organise the information in a format suitable for input into the graph neural network. A descriptive list of the features is provided in Supp Fig. S3, and an illustration of these concepts in Supp Fig. S4. The node features were selected through both chemical and mathematical considerations. Specifically, any features that were linear combinations (or very nearly so) of others were excluded to prevent them from compromising the training process. This produced a feature set from the graph collection with a matrix rank of 52.

The collection was randomly subdivided into two sets comprising respectively 2778 and 925 unique structures. The latter were reserved for use as independent test or 'holdout' items. The feature sets and sparse adjacency matrices for the training and test sets are provided in the Supp Data ('Dataset1_Features_Training_Set.xlsx', 'Dataset1_Adjacencies_Training_Set.xlsx', 'Dataset1_Features_Test_Set.xlsx' and 'Dataset1_Adjacencies_Test_Set.xlsx'). These files also contain associated metadata, along with the target variable to be modelled, the ^{13}C ppm values.

2.3.2. Dataset 2: Mestrelab proprietary collection

Dataset 2 comprised a collection of 48,920 records containing structures with associated ^{13}C chemical shift assignments. These were prepared using the same data cleaning criteria as outlined for Dataset 1 above (except for requiring an analogous record for ^{1}H; the proprietary databases contain information for ^{13}C only).

As for Dataset 1, Python RDKit utilities were used to extract a molecular graph and features for each record. The full set was likewise filtered to contain only linearly independent features; these numbered 60, reflecting the larger collection size which contains greater diversity at the node level. The collection was also partitioned into training and test sets, comprising respectively 46,945 and 1975 unique structures.

2.3.3. Dataset 3: additional, post-development proprietary collection

A further proprietary database was made available to the authors only after the completion of all model development. It contained a diverse range of molecular structures, each with associated ^{13}C chemical shift assignments. Dataset 3 was extracted from this database, and comprised records for 11,780 unique structures, none of which were present in either Dataset 1 or 2. This allowed Dataset 3 to function as an independent test set, providing a robust challenge for the predictive models developed from both other datasets. The graph and feature sets for Dataset 3 were prepared as described for Dataset 2.

3. Results and discussion

3.1. Dataset 1

3.1.1. Training an MPNN for chemical shift prediction

The training set was used to generate an MPNN through the following steps:

(i) Load the target values, feature set and adjacency data.
(ii) Randomly split the data into internal training and validation partitions; scale variables using the training items, retaining scaling parameters.
(iii) Define the network architecture, comprising message passing and regression layers, and initialise their weights.
(iv) Using only the training partition, train the MPNN via Adam optimisation [12], which involves adjusting the weights after each pass through the network to minimise errors in predicting chemical shifts. Regularly evaluate the prediction error ('loss') on the validation partition, to monitor convergence and prevent overfitting.
(v) Save the trained model weights and other parameters for the current MPNN.

The network architecture defined in step (iii) was informed by literature reports of similar applications [8,13] and is illustrated in Fig. 1. The inputs are the concatenated adjacency and feature set matrices from the training partition. The network outputs are predicted chemical shifts, which are compared to known target values during training to compute the model error.

The architecture includes 4 MPNN layers, which iteratively update the feature set, incrementing the receptive field around each node via

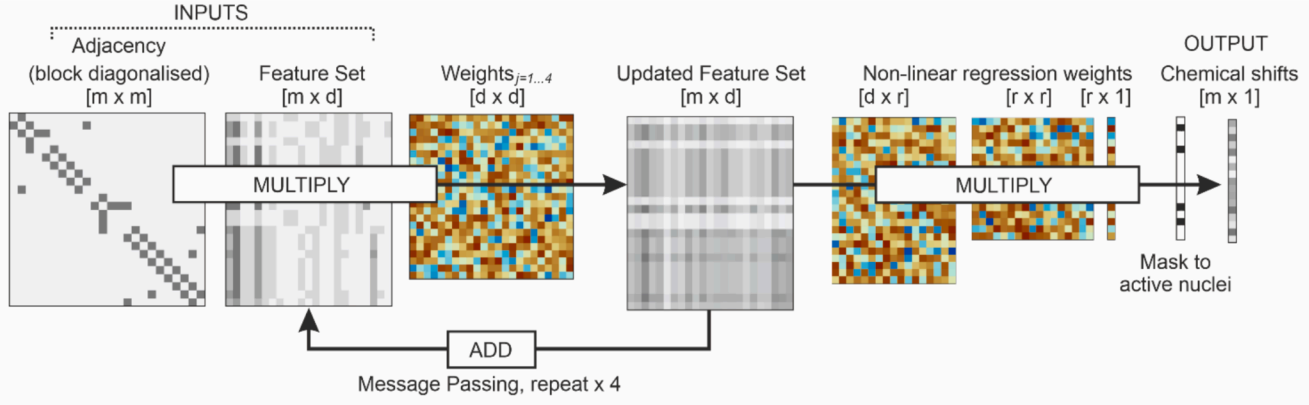


Fig. 1. Schematic showing the deep learning architecture used in the present work. Coloured heatmaps represent matrices of learnable weights, updated after each pass through the network. Grayscale heatmaps represent the input variables (adjacency matrix, feature set), the 'updated feature set' from the message passing layers, and the output/target variable (chemical shifts). A binary mask restricts evaluation of the prediction error to active nuclei only. ReLU transfer functions are applied to the outputs of each layer, except in the [d × r] regression layer, where a tanh function introduces non-linearity.

the connectivity information on each pass. Initial exploratory work showed that increasing the number of MPNN layers up to 4 gave a clear advantage, after which there was no additional gain in model performance.

The MPNN cycles are followed by regression layers. Non-linearity is introduced into the first of these by application of a tanh (hyperbolic tangent) 'transfer' function to the layer output. ReLu (Rectified Linear Unit) activation was used for all other layers.

A Matlab function ('MPNN_Train_Function.m') is supplied in the Supp Data for carrying out these steps on the Dataset 1 training set. On the high-specification Dell PC, convergence of the MPNN is typically achieved in less than 40 min. Using the randomisation seeds at steps (ii)

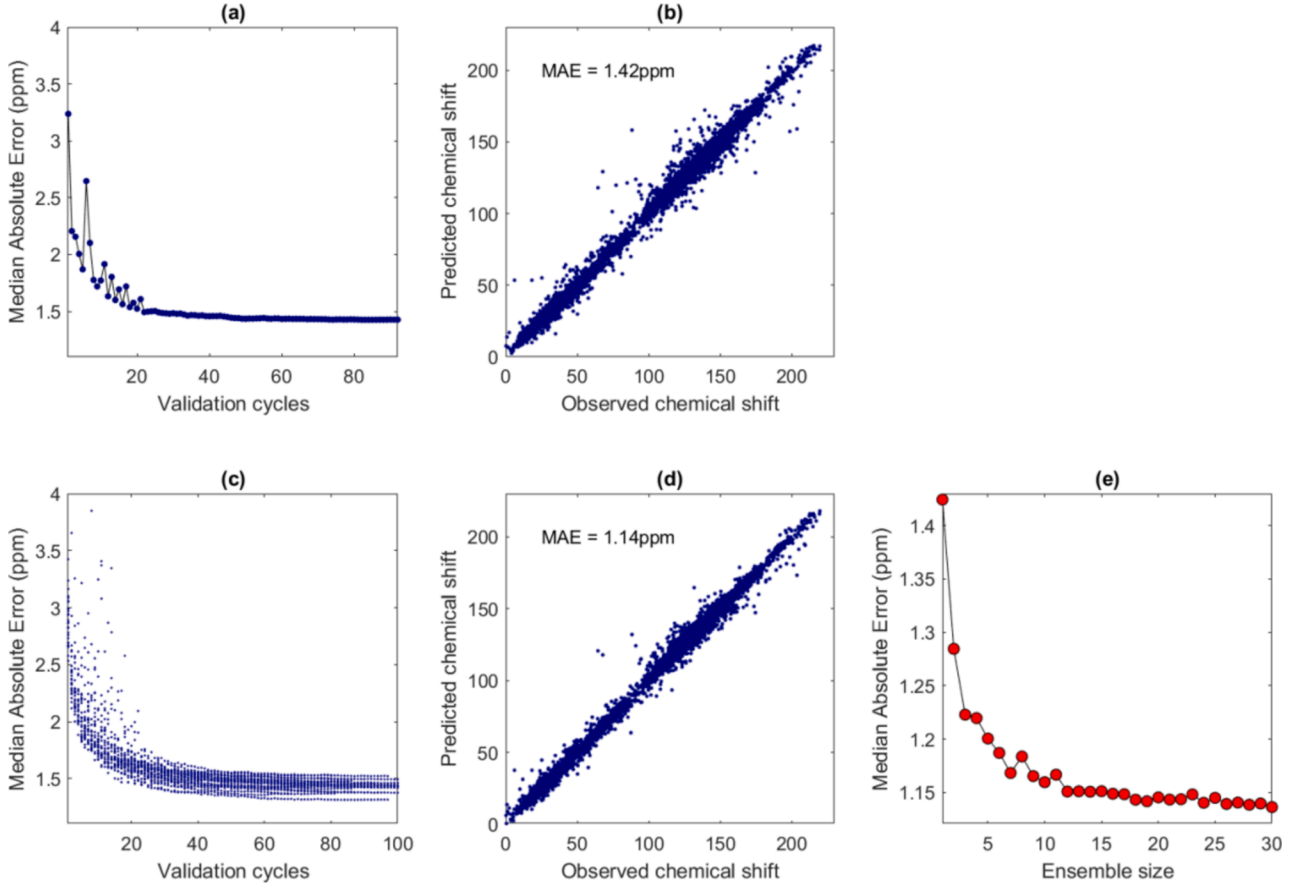


Fig. 2. (a) Validation median absolute error (MAE) vs. the number of completed validation cycles for typical MPNN training, illustrating initial instability and subsequent stabilisation as the termination criterion is approached. (b) Predicted vs. observed ^{13}C chemical shifts from the MPNN applied to Dataset 1 test set. The MAE = 1.42 ppm, similar to the validation error for the MPNN and indicating good generalisation to these unseen structures. (c) Validation MAE vs. validation cycles for an ensemble of 30 MPNNs, highlighting variability due to random partitioning and weight initialisation. (d) Predictions vs. observed ^{13}C chemical shifts from the ensemble applied to the Dataset 1 test set. The MAE = 1.14 ppm, demonstrating improved accuracy with ensemble learning. (e) MAE vs. number of MPNNs in the ensemble.

and (iii) as given in the script, the results illustrated in Fig. 2(a) and (b) can be reproduced. Fig. 2(a) shows the median absolute error (MAE) in chemical shift prediction calculated from the internal validation subset as a function of the validation intervals. This is typical gradient descent behaviour: after an initial phase of instability, the network settles, and the objective function descends relatively smoothly until the termination criterion is achieved. Here this occurs after 91 validation cycles, at which point the MAE is 1.42 ppm.

The network parameters saved in step (v) can be used to make ^{13}C chemical shift predictions for previously unseen structures. As inputs, this requires only the molecular graph and node feature set, which are calculable via RDKit for any valid structure. In the present work, the test set outlined above is used to serve in this way. As well as being unseen items, these benefit from labelled chemical shift information that enables assessment of the MPNN's predictive performance. The 925 structures in the test set contain a total of 12,513 nodes, of which 9315 correspond to Carbon atoms. Each has an observed, assigned ^{13}C chemical shift value, and these are plotted in Fig. 2(b) against the predictions made by the MPNN. The prediction MAE is 1.42 ppm, which corresponds well to the internal validation error in panel (a) and indicates that the MPNN is able to generalise to the unseen structures in the test set. A Matlab script for carrying out this step is provided ('MPNN_Apply_To_Test.m') as well as the parameter set as used to generate the results in Fig. 2 ('MPNN_13C_params_20240829155238.mat').

3.1.2. *Improving prediction performance using an ensemble of MPNNs*

The predictive performance of neural networks is strongly influenced by certain training hyperparameters. In this study, these factors were determined through a combination of literature review and exploratory searches of the hyperparameter space. Major sources of stochastic variability are the validation partitioning and the initialisation of the network weights. This suggests an ensemble learning approach may offer an advantage.

To create an ensemble, steps (ii)–(v) are repeated up to the desired number of MPNNs. Whilst this is feasible on a laptop, it is more practical on the GPU nodes of the high-performance computer cluster, where training time reduces to <10 min per MPNN. (Note that the provided Matlab script requires adapting to exploit GPU functionality.)

A 30-MPNN ensemble was created from the training collection. The validation error as a function of training cycles for all the MPNNs is shown in Fig. 2(c) and gives an impression of the variability that arises from the random partitioning and weight initialisation steps. Chemical shift predictions were calculated by averaging the values obtained from application of all MPNNs in the ensemble to the Dataset 1 test set. These are shown plotted against the observed values in Fig. 2(d). The MAE is 1.14 ppm: this is a substantial improvement compared to typical MAEs obtained from single MPNNs. It also compares well with values reported in the literature [8] and with results from a commercially available tool (illustrated in Supp Fig. S5). Fig. 2(e) illustrates the cumulative effect of ensemble learning by plotting the MAE against increasing numbers of MPNNs used in the pooled output. The choice of 30 MPNNs for the final ensemble size was pragmatic, as diminishing returns were obtained beyond this value.

The prediction errors are distributed symmetrically about a median value of −0.04 ppm (Fig. 3(a)). The distribution is fat-tailed compared with a normal distribution but can be well fitted by a Gaussian kernel probability density function (pdf). An interesting association is found between the error magnitude and the frequency count of different atom environments in the training set; these are represented by the 52-element node feature vectors, of which there are 5369 unique types. Fig. 3(b) plots the errors against the corresponding node type's count. Unrepresented nodes typically exhibit twice the error as those with the highest representation in the training set. Further, the relationship between the median error magnitude at each frequency count can be well-modelled by a decaying exponential, as shown in Fig. 3(c).

3.2. *Dataset 2*

An ensemble of MPNNs was trained using the procedure outlined above, with small adjustments to the network architecture to accommodate the larger feature set width. Summary outcomes for Dataset 2 training (analogous to Fig. 2(c)–(e)) are given in Supp Fig. S6. Running on a single GPU node of the HPC facility, each MPNN took ~3 h to reach

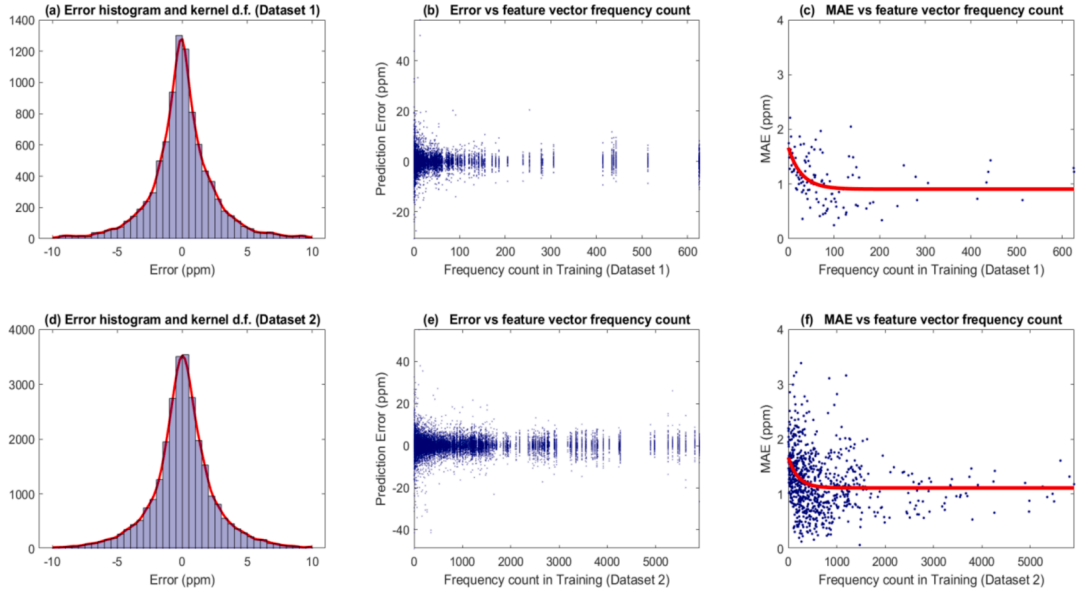


Fig. 3. (a) Histogram showing the ensemble prediction errors (n = 9315) for Dataset 1 test set. These are symmetrically distributed about a median of −0.04 ppm and well-fitted by a Gaussian kernel pdf (red line). (b) The errors plotted against the frequency count of their corresponding node feature vectors in the Dataset 1 training set. (c) Median absolute errors plotted vs. the node frequency count. The decaying exponential (red line) shows the fit of the MAE as a function of node representation. (d) Histogram showing the ensemble prediction errors (n = 39,892) for Dataset 2 test set. The median is 0.04 ppm, and the fitted pdf is shown by the red line. (e) The errors and (f) the median absolute errors plotted vs. the node frequency count for Dataset 2, with the red line again indicating an exponential fit of the MAE as a function of node representation.

the termination criterion, which typically occurs at ~120 iterations. Note that training time does not scale linearly with the number of structures but rather the size of the block diagonal adjacency matrix (see Supp Fig. S4) which in effect defines a single combined training graph. An ensemble size of 30 was sufficient for the prediction error from the Dataset 2 test set to stabilise, at an MAE of 1.17 ppm. This is comparable to the prediction performance obtained for Dataset 1.

3.2.1. Error and generalisation ability dependence on training set diversity

The histogram of the prediction errors and their relationship to the node frequency counts are shown in Fig. 3(d), (e) and (f). The correspondence with the upper panels from Dataset 1 is striking. The errors are again characterised by a fat-tailed, symmetric distribution, here with a median value of 0.04 ppm, and there is an obvious relationship between error magnitude and node type count in the training set.

An important difference, however, is that the Dataset 2 training set contains 31,414 unique atom environments, approximately six-fold that of Dataset 1. This is reflected in the greater point densities with respect to the x-axis of Fig. 3(e) and (f) compared with (b) and (c). Taking advantage of the greater number of values available, the behaviour of additional percentiles including Q1 and Q3 as a function of the node frequency was also estimated; these are illustrated in Supp Fig. S7. These percentile curves provide a means of associating an error estimate with any prediction, based upon its feature vector's representation in the training set.

The greater node diversity of Dataset 2 compared to Dataset 1 also impacts their respective ensembles' generalisation abilities. This can be examined by carrying out crossover validation, that is, prediction on the unseen structures from each other's test sets. A further, demanding challenge to both ensembles is provided by the large number of structures in Dataset 3, none of which are represented in either of the other datasets; these 11,780 structures yield 244,294 chemical shift predictions. A summary of the MAEs obtained from application of both ensembles to each of the test sets is given in Fig. 4. The Dataset 2 ensemble achieves better MAEs throughout, remaining remarkably consistent at <1.18 ppm irrespective of the test set. In contrast, the Dataset 1 ensemble performs significantly worse on the larger test collections. The probable explanation for this is Dataset 2's demonstrably broader coverage of 'chemical space' – the vast landscape of possible structures which numbers in the billions even for small, drug-like molecules [14,15]. This finding reinforces the importance of a large and diverse corpus of labelled data in training complex, heavily parameterised models, when broad generalisation ability is the primary requirement. For completeness, error histograms for the crossover and Dataset 3 predictions made by both ensembles are given in Supp Fig. S8, along with analogues of Fig. 3(e) and (f) obtained from applying the Dataset 2 ensemble to Dataset 3.

3.2.2. Structure-level prediction errors

An aggregated error from all predictions for a molecule can be calculated and is a useful statistic. Fig. 5(a) shows the histogram for the mean of the absolute errors obtained from Dataset 2 test structures containing $N = 10$ carbon atoms, fitted with a Gaussian kernel pdf. Fig. 5 (b) shows the corresponding empirical and fitted cumulative distribution function (cdf). Strictly speaking, the distribution of this statistic varies with N, the number of prediction errors used in its calculation. However, per the central limit theorem, it is found that there is comparatively little change when $N > 10$, as shown in Fig. 5(c) and (d).

Fig. 6 presents an illustrative comparison between the observed and predicted chemical shifts obtained from applying the Dataset 2 ensemble to a representative Dataset 1 test item, cyclopropyl-phenylmethanone. Although this compound contains 10 carbons, due to molecular symmetries there are only 7 unique predicted shift values. In this case, there were also only 7 unique observed shifts, but more generally, it is common for there to be fewer observed than predicted values, especially in crowded spectra with weakly resolved peaks.

As this is a test set item, observed shifts are available, thus the mean of the absolute errors in prediction can be calculated and is found to be 1.54 ppm. From the cdfs of Fig. 5, it is seen that predictions with comparable levels of accuracy are obtained from almost 50 % of structures with $N \geq 10$. This provides a straightforward route to a probabilistic score (p-value) for assessing whether the sample of prediction errors obtained for a structure could plausibly be drawn from the error distribution of Fig. 3(d); this idea is explored further in the following section.

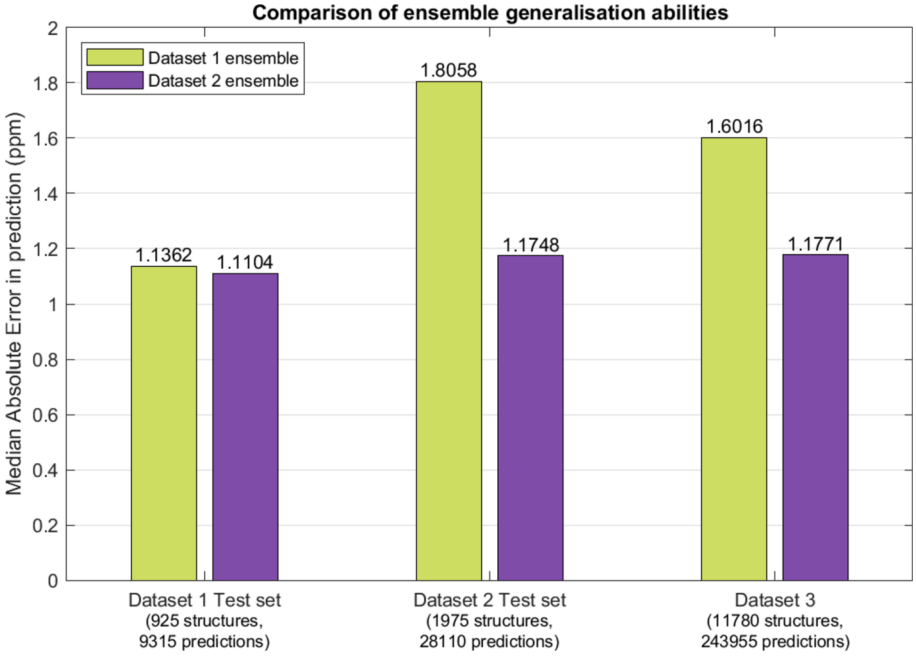


Fig. 4. A summary of the Dataset 1 and Dataset 2 ensembles' abilities to generalise to unseen structures. The Dataset 2 ensemble gives the best performance on all test sets, but this is most notable for the larger test sets.

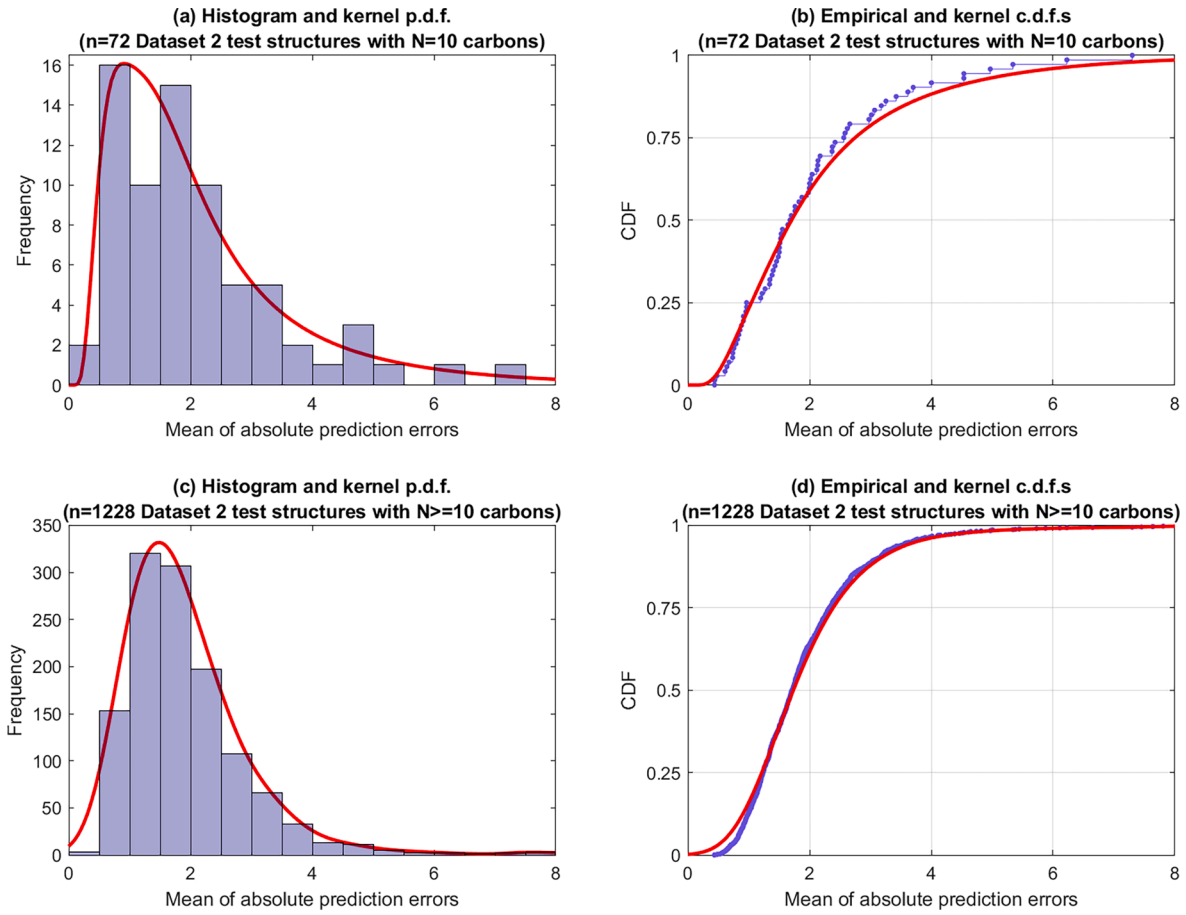
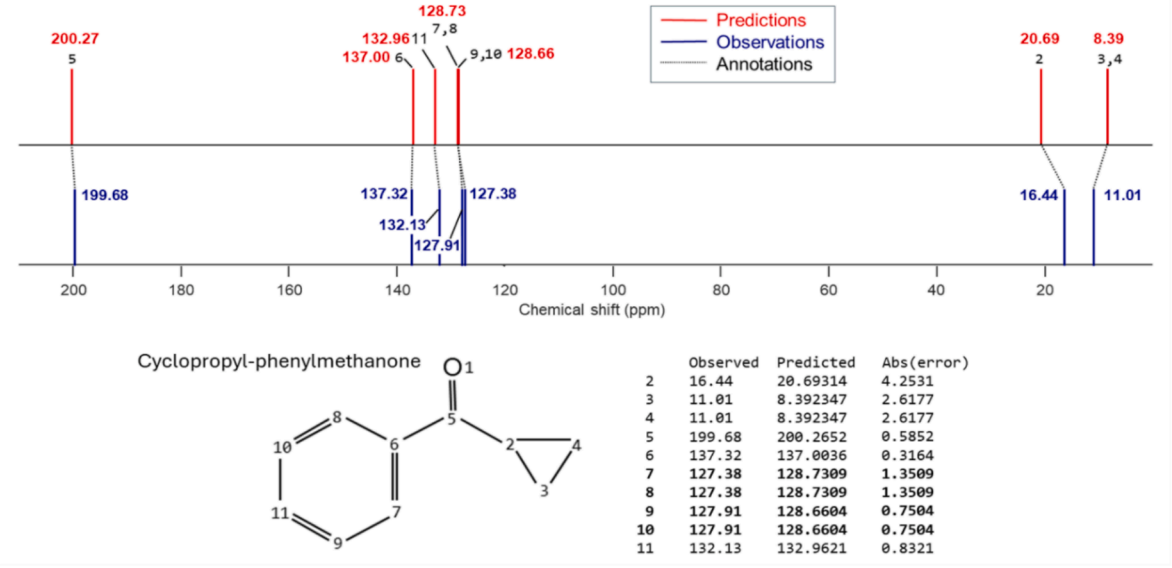


Fig. 5. (a) Histogram showing the structure-level mean of the absolute prediction errors for Dataset 2 test structures that contain N = 10 carbon atoms. The fitted Gaussian kernel pdf is marked (red line). (b) The corresponding cdf (red line) with the empirical cdf shown for comparison (blue). Panels (c) and (d) show analogous figures for structures with N ≥ 10.



	Observed	Predicted	Abs(error)
2	16.44	20.69314	4.2531
3	11.01	8.392347	2.6177
4	11.01	8.392347	2.6177
5	199.68	200.2652	0.5852
6	137.32	137.0036	0.3164
7	127.38	128.7309	1.3509
8	127.38	128.7309	1.3509
9	127.91	128.6604	0.7504
10	127.91	128.6604	0.7504
11	132.13	132.9621	0.8321

Fig. 6. Shows the observed and predicted chemical shifts obtained from an unseen structure, cyclopropyl-phenylmethanone, using the Dataset 2 ensemble. This is a representative item from the Dataset 1 test set. Observed and predicted shifts are indicated on the chemical shift scale in blue or red respectively, and are also tabulated along with the absolute error for each carbon nucleus as indicated alongside the structure diagram. The (known) assignment of observed to predicted values is indicated by black dotted lines. For nuclei 7/8 and 9/10 these lines are crossed, meaning that simple rank-matching assignment would be incorrect.

3.2.3. Using the predictions: Assignment and verification

Chemical shift prediction is the foundation of both spectral assignment and structural verification [16,17]. Predictions are generally used in tandem with additional information such as molecular formulas and 2-D NMR experiments. However, in combination with advanced computational tools, single nucleus 1-D spectra can be effective for assignment, verification and even elucidation [18].

In assignment, experimentally observed shifts are matched with predicted shifts for each active nucleus. In verification, the annotation is assessed qualitatively or by some numerical score to confirm a compound's identity. Observed shifts may also be compared with predictions from more than one structure, to identify the most likely candidate.

In the interests of conciseness, the discussion here will be confined to the special case where the number of experimentally observed shifts equals the number of unique predictions made for a proposed structure. In such scenarios, annotation can proceed by sorting the observed and predicted chemical shifts by value and matching up these vectors elementwise. A natural score for such an assignment is the cityblock distance, which is the *sum* of the absolute differences between the two vectors, but there are some caveats. First, although the cityblock distance is minimised by this assignment approach, there will generally be multiple tied solutions with no way of distinguishing between them. Further, assignments made in this way may not necessarily be correct: returning to the illustration in Fig. 6, it will be seen that matching by ranked values would not correspond to the *a priori* known assignment marked on the figure (evident from the crossed annotation lines for nuclei 7,8 and 9,10).

In the present work, an alternative approach is implemented in which observations are assigned to predictions to minimise a *modified* cityblock metric that incorporates a cost function. The weights are provided by error estimates for the predictions, extracted from the fitted curves for error versus node representation described in section 3.2.1 above. The effect of this modification is that predictions with larger estimated errors can be matched to more distant observations with comparatively less penalty. These workings are detailed using toy examples in Supp Figs. S9 and S10. A corresponding real-world example is the compound used in Fig. 6: the modified method is found to give an assignment that exactly matches the known annotation shown in the figure.

Once obtained, the assigned observations and predictions can be used to calculate the *conventional* cityblock distance. Dividing this value by N gives the mean absolute error for the structure, a statistic whose cdf is characterised in Fig. 5, that can provide a probabilistic score by which to assess the assignment. This approach is demonstrated for a collection of 16 structures taken from the Dataset 1 test collection. These were selected by the mutual Jaccard similarity of their SMILES strings. All contained either 10 or 11 carbon nuclei and yielded exactly 10 unique predictions and observed shifts to be assigned. For larger structures and numbers of predictions, combinatorial explosion means a metaheuristic method is required to find the optimal assignment. However, for $N \leq 10$, an exhaustive search of the solution space is feasible, as in the present example.

Fig. 7 shows heatmaps of cross-verification matrices containing different forms of verification score for each possible pairwise assignment of observed to predicted shifts for the 16 items; structure diagrams for each of these compounds are given in Supp Fig. S11. In Fig. 7(a), each row contains the mean cityblock distance between observations from 'actual' structures (per the y-axis label) and predictions from 'proposed' structures (per the x-axis label). The minimum value of each row occurs on the diagonal, which shows that, for the actual structures under consideration, the predictions from the same proposed structure in all cases yielded the lowest cityblock assignment score.

Fig. 7(b) maps the cityblock values into p-values via application of the cdf of Fig. 5. Although this is a monotonic transformation, the heatmap is visually quite different and is effective at conveying certain information. The p-values indicate whether the absolute errors between observed and predicted values calculated from an assignment are plausible given the expected error distribution for correctly proposed structures. Less credible assignments are flagged by p-values less than some desired threshold (typically 0.05). In the present case, the diagonal contains values between 0.28 and 0.91. This indicates that the predictions from proposed structures are all consistent with observations from the same structure. Further, almost all the off-diagonal elements are close to 0. Just six of the p-values exceed 0.05, and these arise from pairwise assignments amongst three compounds, with SMILES strings Nc1ccc2ccccc2c1, Oc1ccc2ccccc2c1 and Oc1ccc2cc(Br)ccc2c1. However, it is notable that even for these similar structures, the p-values on the

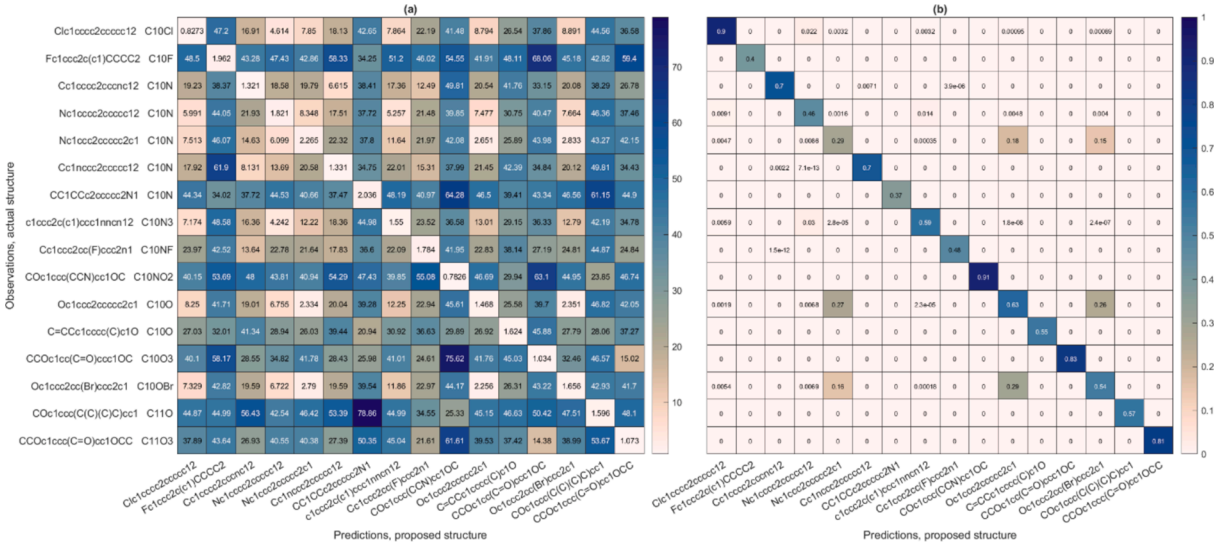


Fig. 7. (a) Heatmap showing the mean cityblock distance between observations from 'actual' structures (y-axis) and predictions from 'proposed' structures (x-axis). For information, the molecular formula is also included in the y-axis labels. The minimum value in each row occurs on the diagonal, indicating that predictions from the same proposed structure yield the lowest cityblock assignment score for the actual structures. (b) Heatmap showing the cityblock values transformed into p-values via the cdf of Fig. 4(b). The diagonal values range from 0.28 to 0.91, indicating consistency between predictions and observations from the same structure. Six p-values exceed 0.05; these arise from pairwise assignments among three similar compounds.

diagonal are the highest values in each row, demonstrating high verification specificity for this collection of items.

4. Conclusions

This work presented a deep learning approach for predicting NMR chemical shifts of small molecules. Graph convolutional neural networks with four message-passing layers were trained on large numbers of molecular structures, fully annotated with ^{13}C chemical shifts. To reduce stochastic variation, an ensemble framework was employed, which is straightforward to implement across multiple nodes of an HPC cluster.

Two distinct data collections were used in the modelling work, one comprising approximately 4000 labelled structures, the other exceeding 40,000. Separate ensembles were trained from each dataset, and were challenged with multiple test sets containing collections of mutually unseen structures. The results demonstrate the crucial importance of training set size and diversity. While prediction performance was similar for the holdout sets from within each collection, the ensemble trained on the larger dataset maintained its prediction accuracy for all test sets, including a third data collection (11,780 structures, 243,955 chemical shifts) made available to the authors after all development work was complete. In contrast, the performance of the ensemble from the smaller dataset declined notably. This discrepancy is attributed to the greater diversity of atomic environments in the larger dataset: examination of the prediction errors showed these to be clearly linked to the frequency count of different node feature vectors present in the training sets.

Further, the larger dataset allowed for more robust modelling of various error properties. These models provide a quantitative foundation for spectral assignment and verification in two ways. First, an estimated error can be associated with each prediction for a structure using the modelled relationship between error magnitude and node representation. These estimates are used as weights in a modified city-block distance metric during the assignment of observed to predicted shifts, which seeks the annotation that minimises this metric. Second, the mean absolute prediction error at the structure level is well fitted by a Gaussian kernel cdf. This allows for a probabilistic assessment of the assignment, evaluating whether the predicted shifts and assigned observations are consistent with originating from the same molecular structure.

CRediT authorship contribution statement

D. Williamson: Validation, Methodology, Formal analysis, Data curation, Conceptualization. **S. Ponte:** Validation, Software, Resources, Methodology. **I. Iglesias:** Validation, Supervision, Software. **N. Tonge:** Project administration, Funding acquisition, Conceptualization. **C. Cobas:** Writing – review & editing, Methodology, Data curation, Conceptualization. **E.K. Kemsley:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research presented in this paper was carried out on the HPC cluster supported by the Research and Specialist Computing Support service at the University of East Anglia (UEA), Norwich, United Kingdom. E.K.K. appreciates the assistance of Jacob Newman and Sarah Wilson Kemsley (Faculty of Science, UEA) in configuring the HPC coding environments, and also extends thanks to Mestrelab Research SL for supporting this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmr.2024.107795.

Data availability

'Dataset 1' used in this work is a subset of records extracted from the publicly available nmrshiftdb2 database, and is shared here as an Excel file. 'Dataset 2' and 'Dataset 3' are confidential and are not shared.

References

[1] L.C. Blum, J.L. Reymond, 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13, J. Am. Chem. Soc. 131 (25) (2009) 8732–8733.
[2] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, in: Neural Message Passing for Quantum Chemistry. 34th International Conference on Machine Learning; 2017 Aug 06-11; Sydney, Australia, Jmlr-Journal Machine Learning Research, San Diego, 2017 Aug, p. 2017.
[3] T.N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks. arXiv: 160902907, 2017.
[4] M. Ma, X.J. Lei, A dual graph neural network for drug-drug interactions prediction based on molecular structure and interactions, PLoS Comput. Biol. 19 (1) (2023) 20.
[5] C.Y. Liu, Y. Sun, R. Davis, S.T. Cardona, P.Z. Hu, ABT-MPNN: an atom-bond transformer-based message-passing neural network for molecular property prediction, J. Cheminform. 15 (1) (2023) 14.
[6] Z. Wang, C. Wang, S.B. Zhao, Y. Xu, S.G. Hao, C.Y. Hsieh, et al., Heterogeneous relational message passing networks for molecular dynamics simulations, NPJ Comput. Mater. 8 (1) (2022) 9.
[7] E. Jonas, S. Kuhn, Rapid prediction of NMR spectral properties with quantified uncertainty, J. Cheminform. 11 (1) (2019) 7.
[8] Y. Kwon, D. Lee, Y.S. Choi, M. Kang, S. Kang, Neural message passing for NMR chemical shift prediction, J. Chem. Inf. Model. 60 (4) (2020) 2024–2030.
[9] T. Sajed, Z. Sayeeda, B.L. Lee, M. Berjanskii, F. Wang, V. Gautam, et al., Accurate prediction of ^{1}H NMR chemical shifts of small molecules using machine learning, Metabolites 14 (5) (2024) 14.
[10] C. Kuenneth, RDKit, 2023.
[11] S. Kuhn, N.E. Schlörer, Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2-a free in-house NMR database with integrated LIMS for academic service laboratories, Magn. Reson. Chem. 53 (8) (2015) 582–589.
[12] D.P. Kingma, Jimmy Ba, A method for stochastic optimization, arXiv preprint arXiv:14126980, 2014.
[13] The MathWorks I. Node Classification Using Graph Convolutional Network, 2021.
[14] L. Ruddigkeit, R. van Deursen, L.C. Blum, J.L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, J. Chem. Inf. Model. 52 (11) (2012) 2864–2875.
[15] J.L. Reymond, The chemical space project, Accounts Chem. Res. 48 (3) (2015) 722–730.
[16] Y.H. Tsai, M. Amichetti, M.M. Zanardi, R. Grimson, A.H. Daranas, A.M. Sarotti, ML-J-DP4: an integrated quantum mechanics-machine learning approach for ultrafast NMR structural elucidation, Org. Lett. 24 (41) (2022) 7487–7491.
[17] F. Hu, M.S. Chen, G.M. Rotskoff, M.W. Kanan, T.E. Markland, Accurate and efficient structure elucidation from routine one-dimensional NMR spectra using multitask machine learning, arXiv:240808284 [physicschem-ph], 2024.
[18] A. Howarth, J.M. Goodman, The DP5 probability, quantification and visualisation of structural uncertainty in single molecules, Chem. Sci. 13 (12) (2022) 3507–3518.