nature computational science

Brief Communication

https://doi.org/10.1038/s43588-025-00783-z

Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts

Received: 14 August 2024

Accepted: 26 February 2025

Published online: 28 March 2025

Check for updates

Fanjie Xu[1,2], Wentao Guo[2,3], Feng Wang[1], Lin Yao[2], Hongshuai Wang[2], Fujie Tang[4,5,6] ✉, Zhifeng Gao[2] ✉, Linfeng Zhang[2,7], Weinan E[7,8,9], Zhong-Qun Tian[1,5] & Jun Cheng[1,5,6] ✉

The study of structure–spectrum relationships is essential for spectral interpretation, impacting structural elucidation and material design. Predicting spectra from molecular structures is challenging due to their complex relationships. Here we introduce NMRNet, a deep learning framework using the SE(3) Transformer for atomic environment modeling, following a pretraining and fine-tuning paradigm. To support the evaluation of nuclear magnetic resonance chemical shift prediction models, we have established a comprehensive benchmark based on previous research and databases, covering diverse chemical systems. Applying NMRNet to these benchmark datasets, we achieve competitive performance in both liquid-state and solid-state nuclear magnetic resonance datasets, demonstrating its robustness and practical utility in real-world scenarios. Our work helps to advance deep learning applications in analytical and structural chemistry.

Spectroscopy is an essential tool for elucidating molecular structures and dynamics, with nuclear magnetic resonance (NMR) being a particularly powerful technique in chemistry, biology and materials science[1,2]. Accurate prediction of NMR chemical shifts aids in spectrum interpretation, structure revision and configuration determination[3–5]. However, traditional methods for NMR chemical shift prediction often struggle to balance accuracy and efficiency, particularly for complex molecular architectures[6–9].

Recent advances in deep learning have provided promising solutions for enhancing NMR predictions. For liquid-state NMR, public datasets such as nmrshiftdb2 (ref. 10) and QM9-NMR[11] have facilitated the development of machine learning models, with approaches like graph convolutional network and equivariant message passing neural network showing improved accuracy over traditional methods[12,13]. While liquid-state models often neglect intermolecular interactions, primarily solute–solvent interactions, recent studies have integrated molecular dynamics and calculations to overcome this limitation[14]. For solid-state NMR, where models must account for periodic boundary conditions (PBC), machine learning models such as ShiftML and NN-NMR have demonstrated efficiency and precision by leveraging computational datasets derived from density functional theory (DFT)[15–18].

Despite these advances, existing frameworks are typically tailored to either liquid-state or solid-state NMR, limiting their generalizability. To address this, we introduce NMRNet, a unified framework leveraging a pretraining and fine-tuning paradigm. By pretraining on extensive structure data with a shared SE(3) Transformer architecture, NMRNet

[1]State Key Laboratory of Physical Chemistry of Solid Surface, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China. [2]DP Technology, Beijing, China. [3]Department of Chemistry, University of California, Davis, CA, USA. [4]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen, China. [5]Laboratory of AI for Electrochemistry, Tan Kah Kee Innovation Laboratory, Xiamen, China. [6]Institute of Artificial Intelligence, Xiamen University, Xiamen, China. [7]AI for Science Institute, Beijing, China. [8]Center for Machine Learning Research, Peking University, Beijing, China. [9]School of Mathematical Sciences, Peking University, Beijing, China. ✉e-mail: tangfujie@xmu.edu.cn; gaozf@dp.tech; chengjun@xmu.edu.cn

effectively models atomic environments for both liquid-state and solid-state systems. We further curated a standardized benchmark dataset, nmrshiftdb2-2024, through extensive cleaning and validation of experimental data, providing a valuable resource to advance the prediction of NMR chemical shifts. Our results demonstrate that NMRNet achieves impressive performance across diverse systems, including natural products, large organic molecules and solid materials. While challenges remain in accurately modeling complex solvation interactions and chiral systems, NMRNet offers substantial potential for spectral interpretation and structural analysis, paving the way for future advancements in materials design and chemical research.

Results

Overview of NMRNet framework

The NMRNet framework integrates four modules: data preparation, pretraining, fine tuning and inference. The overall workflow is illustrated in Fig. 1. Leveraging the SE(3) Transformer architecture from Uni-Mol[19], NMRNet adapts atomic-level representations to accommodate liquid, solid and gaseous states. To ensure accurate predictions across diverse scenarios, we utilized extensive structural databases for pretraining and built NMR datasets for fine tuning.

In the data preparation module, three-dimensional (3D) structural information is extracted to represent the local chemical environment. For liquid-state NMR, individual molecular structures are used, while solid-state NMR uses PBC with a cutoff radius of 6 Å to define atomic environments.

During pretraining, we adapted Uni-Mol to generate robust representations of atomic environments, collecting over 4.8 million structures. This large-scale pretraining mitigates the limitations of scarce experimental NMR data and improves model generality. For fine tuning, NMRNet supports both element-specific and multielement predictions, enabling more precise structure elucidation. The model demonstrates exceptional performance in chemical shift prediction, as validated by metrics that include the mean absolute error (MAE), the root mean square error (RMSE) and the coefficient of determination (R^2) across multiple benchmark datasets.

In the inference stage, NMRNet transcends numerical predictions by facilitating critical applications such as peak assignment and conformation determination, providing insights into structure–spectrum correlations. A web-based tool further enhances accessibility, enabling streamlined chemical shift predictions for the research community.

Fine-tuning NMRNet with liquid-state NMR data

As mentioned, we revisit the latest data from nmrshiftdb2, manually screen and correct erroneous entries and, ultimately, create a more comprehensive and reliable dataset, nmrshiftdb2-2024 to address the limitations of the nmrshiftdb2 version developed in the year 2018 (ref. 12) (hereafter called nmrshiftdb2-2018). The dataset developed here features a greater number of atoms, a wider range of elements and more complex structures (see Supplementary Fig. 3a–c for details). Although the nmrshiftdb2-2024 dataset is more challenging, the final prediction error is lower compared with nmrshiftdb2-2018 (Supplementary Fig. 3d), indicating that the model maintains excellent performance even in more complex systems.

As illustrated in Extended Data Fig. 1a–f, there is a strong correlation between NMRNet's predicted results for ^{1}H, ^{11}B, ^{13}C, ^{15}N, ^{17}O and ^{19}F in nmrshiftdb2-2024 and the corresponding experimental values. The prediction errors (MAE) are 0.181 ppm for ^{1}H and 1.098 ppm for ^{13}C, which are close to the intrinsic experimental errors previously reported[12] (0.09 ppm for ^{1}H and 0.51 ppm for ^{13}C). NMRNet also shows a reliable predictive capability for ^{11}B, ^{15}N and ^{17}O, with R^2 values greater than 0.85, despite these elements being represented by fewer than 230 molecules in the training set (Supplementary Table 2). To demonstrate the effectiveness of pretraining, we compare the fine-tuned NMRNet results on nmrshiftdb2-2024 using models with (marked as 'w/ pre-training') and without (marked as 'w/o pre-training') pretraining. As shown in Extended Data Fig. 1g, prediction errors substantially decrease when pretraining is included. In addition, the pretrained NMRNet shows excellent accuracy in predicting both all elements simultaneously and predicting a single element during fine tuning. As shown in Extended Data Fig. 1h, the model's performance improves as the training set size increases, regardless of whether pretraining is used. Still, the model with pretraining outperforms the one without, especially when the training set is small. Even when the full training set is used (sampling ratio of 100%), the model with pretraining outperforms with lower MAE. This emphasizes the benefit of leveraging large-scale self-supervised pretraining when fine-tuning data is limited.

Moreover, to highlight the advantages of our strategy, we compare it with previous studies[12,13,20,21] on the nmrshiftdb2-2018 (ref. 10) (Extended Data Fig. 1i) and QM9-NMR[11] datasets (Extended Data Fig. 1j). NMRNet shows a further improvement in the prediction accuracy of ^{1}H and ^{13}C elements compared with the previous best methods on the nmrshiftdb2-2018 dataset. Since chemical shifts in QM9-NMR can be calculated by subtracting chemical shielding from reference values, we set shielding values as training targets. The training/test splits for QM9-NMR follow the settings established by DetaNet[13]. NMRNet consistently outperforms across all six environments, with substantial improvements in MAE for ^{1}H (from 0.054 ppm to 0.020 ppm) and ^{13}C (from 0.520 ppm to 0.262 ppm). This highlights NMRNet's ability to accurately model diverse solvent environments.

Fine-tuning NMRNet with solid-state NMR data

We tested four strategies (S1–S4; Methods) to enhance the model's understanding of solid-state NMR, fine tuning it on the ShiftML1 dataset[15] for performance comparison. The error distributions of the model predictions under these four strategies are shown in Extended Data Fig. 2f and Supplementary Fig. 4. For inputs involving the unit cell, using the global distance matrix (S2) provides more accurate descriptions than the intracell distance matrix (S1). Moreover, the S3 of setting the cutoff radius offers a more comprehensive depiction of the chemical environment and is expected to further improve with crystal pretraining. Therefore, we adopt S4, which resulted in a high correlation between its predicted results and DFT calculations for the four elements ^{1}H, ^{13}C, ^{15}F and ^{17}O in the ShiftML1 dataset (Extended Data Fig. 2a–d). Similar performance was observed on the ShiftML2 dataset[22] (Supplementary Table 3). It is important to note that the training set is sampled using farthest point sampling, while the test set is randomly selected. This sampling method creates an extrapolated test set that includes complex chemical environments not found in the training set[15,22,23]. The rigor of this approach is illustrated by the distribution of chemical shifts in the training and test sets (see Extended Data Fig. 2e and Supplementary Fig. 5 for ShiftML1 and ShiftML2, respectively). Moreover, Extended Data Fig. 2g clearly illustrates the substantial performance improvements of NMRNet using strategy 4 over previous methods[15,23], especially for ^{15}N and ^{17}O.

In the previous work, Cheng et al. calculated the dynamic chemical shifts of P2-type cathode materials and then used SOAP to extract local structural descriptors, which were subsequently used by a neural network to predict chemical shifts[18]. NMRNet successfully reduces the prediction error (RMSE) of the chemical shifts of ^{23}Na for P2-type $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$ from 125 ppm in previous work to 48 ppm (Extended Data Fig. 2h), demonstrating a substantial performance improvement brought by our method.

Applications of the fine-tuned NMRNet

Generalization of the NMRNet. Our first experiment to test the generality of NMRNet (all elements, fine tuned on the nmrshiftdb2-2024 dataset) focuses on the task of chemical shift prediction for five nerve agents, which is crucial for evaluating the proposed procedures and preparing for future threats from new variants[24]. The experimental

https://doi.org/10.1038/s43588-025-00783-z

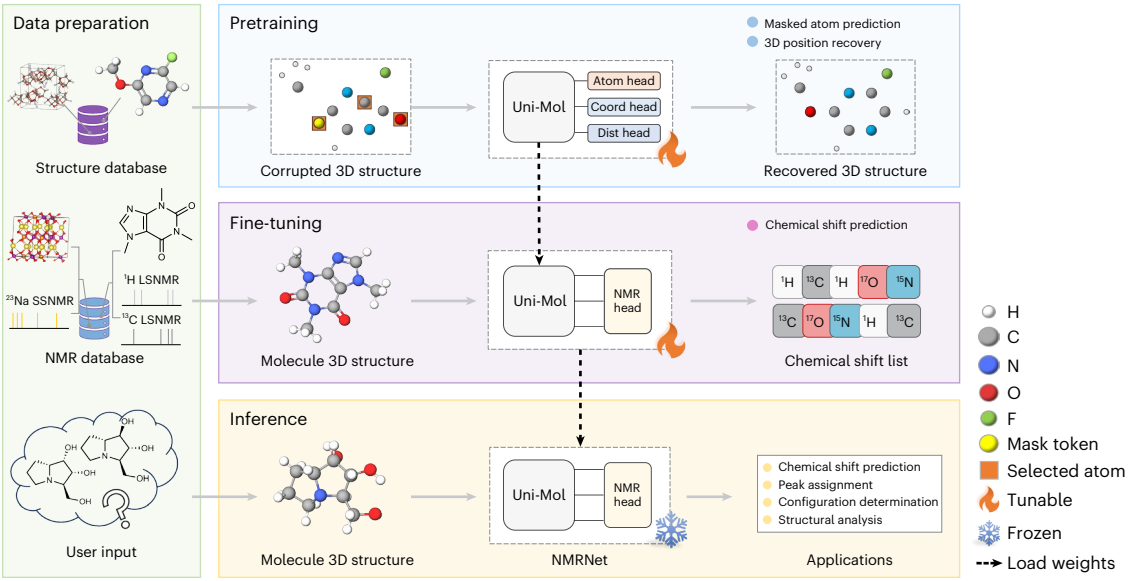


Fig. 1 | A schematic diagram of the NMRNet framework. Left: data preparation, providing structure and NMR data. LSNMR, liquid-state NMR; SSNMR, solid-state NMR. Right top: pretraining, using pure structural information for self-supervised tasks, including masked atom prediction and 3D position recovery. Coord, coordinate; Dist, distance. Right middle: fine tuning, for the supervised NMR chemical shift prediction. Right bottom: inference, where the fine-tuned NMRNet model parameters are frozen and applied to various tasks.

results demonstrate that the prediction accuracy of NMRNet is comparable with DFT methods (Supplementary Fig. 8), without any preliminary knowledge or computational requirements for the nerve agents.

Since the five nerve agent molecules have relatively simple structures, we next tested the NMRNet (all elements), fine tuned on the nmrshiftdb2-2018 dataset (the maximum number of atoms is 64), on a more challenging scenario. We select all molecules containing more than 70 atoms from the nmrshiftdb2-2024 dataset to serve as a more challenge test set. As shown in Fig. 2a, the prediction error shows no substantial changes with the increase of the number of the atoms. Specifically, as shown in Supplementary Fig. 9, when using all molecules with more than 70 atoms as the test set, the R^2 values between the predicted and experimental results are 0.955 for ^{1}H and 0.996 for ^{13}C. Although there is a slight decrease in predictive accuracy compared with the nmrshiftdb2-2018 test set (0.971 for ^{1}H and 0.998 for ^{13}C), the predictions still show a very high correlation with the experimental values, indicating minimal overfitting. When all molecules containing over 100 atoms are used as the test set, the R^2 values are 0.954 for ^{1}H and 0.997 for ^{13}C, further demonstrating NMRNet's notable generalization and extrapolation robustness. To further challenge NMRNet and assess its performance in extreme cases, we extract two molecules with more than 150 atoms (Fig. 2b) and predict their ^{1}H and ^{13}C NMR spectra (Fig. 2c). Despite the complexity of these molecules, the performance of NMRNet remains satisfactory.

Peak assignment. The peak assignment requires matching NMR signals and the atomic environment. Specifically for this task, we develop a module where the experimental chemical shifts can be assigned to the atoms corresponding to the predicted values. The results for the five nerve agent molecules are summarized in Supplementary Tables 4–13. For the assignment of the ^{13}C NMR, NMRNet achieved an accuracy of 94%. Due to the smaller values and closer differences in experimental chemical shifts between different atoms, this task is more challenging in the ^{1}H NMR. The results show that NMRNet only achieved an accuracy of 72% for the assignment of the ^{1}H NMR. There is still substantial room for improvement in the ^{1}H NMR peak assignment task, which involves complex structural identification and relies on multiple spectroscopic techniques.

Configuration determination. Identifying configuration and stereochemistry in NMR interpretation is challenging due to the similarity of chemical environments in isomers. By incorporating 3D molecular representations, NMRNet addresses the limitations of one-dimensional and two-dimensional molecular data, solving this task without requiring extensive chemical knowledge. To evaluate NMRNet's performance, we selected six isomer cases (Extended Data Fig. 3) with comprehensive data from previous studies[4,25–27]. We determined the structure by comparing experimental and predicted chemical shifts, identifying the configuration with the lowest root mean square deviation (RMSD).

In the structure revision task, we successfully revise the structures of 1-benzoylpyrrolidine, TIC10 and nevirapine among pairs of close isomers using only ^{13}C NMR data (Extended Data Fig. 3, top, and Supplementary Tables 14–16). Moving to the more challenging task of chiral isomer identification, specifically *R*/*S* stereochemistry determination (Extended Data Fig. 3, bottom, and Supplementary Tables 17–23), NMRNet identifies three out of four isomers of (3*R*)-3-hydroxy-2,4,6-trimethylheptanoic using ^{13}C NMR data alone. Another example is (3*R*)-5-phenyl-2-propyloxolan-3-ol, where all four possible isomers are successfully identified when both ^{1}H and ^{13}C NMR data are considered. However, when only ^{1}H or ^{13}C data are used, only two isomers are correctly assigned, highlighting the importance of combining both ^{1}H and ^{13}C NMR data, which aligns with chemists' experimental approaches. In the case of hyacinthacines, a compound with eight possible isomers (seven synthesized and one yet to be synthesized), five of the eight isomers are correctly assigned using either ^{13}C predicted chemical shifts, even when the atoms in the diastereomers share very similar chemical environments. When both ^{1}H and ^{13}C NMR results are combined, the correct isomer's RMSD decreases, demonstrating that integrating multiple NMR datasets facilitates more accurate structure assignment. This approach underscores the critical role of comprehensive NMR analysis in stereochemistry determination, offering a robust pathway for accurate molecular identification.

Moving forward, it will be crucial to collect and standardize more NMR data on chiral isomers to enhance deep learning methods, with the aim of training models that are more sensitive to these subtle differences. Moreover, combining deep learning identification methods

https://doi.org/10.1038/s43588-025-00783-z

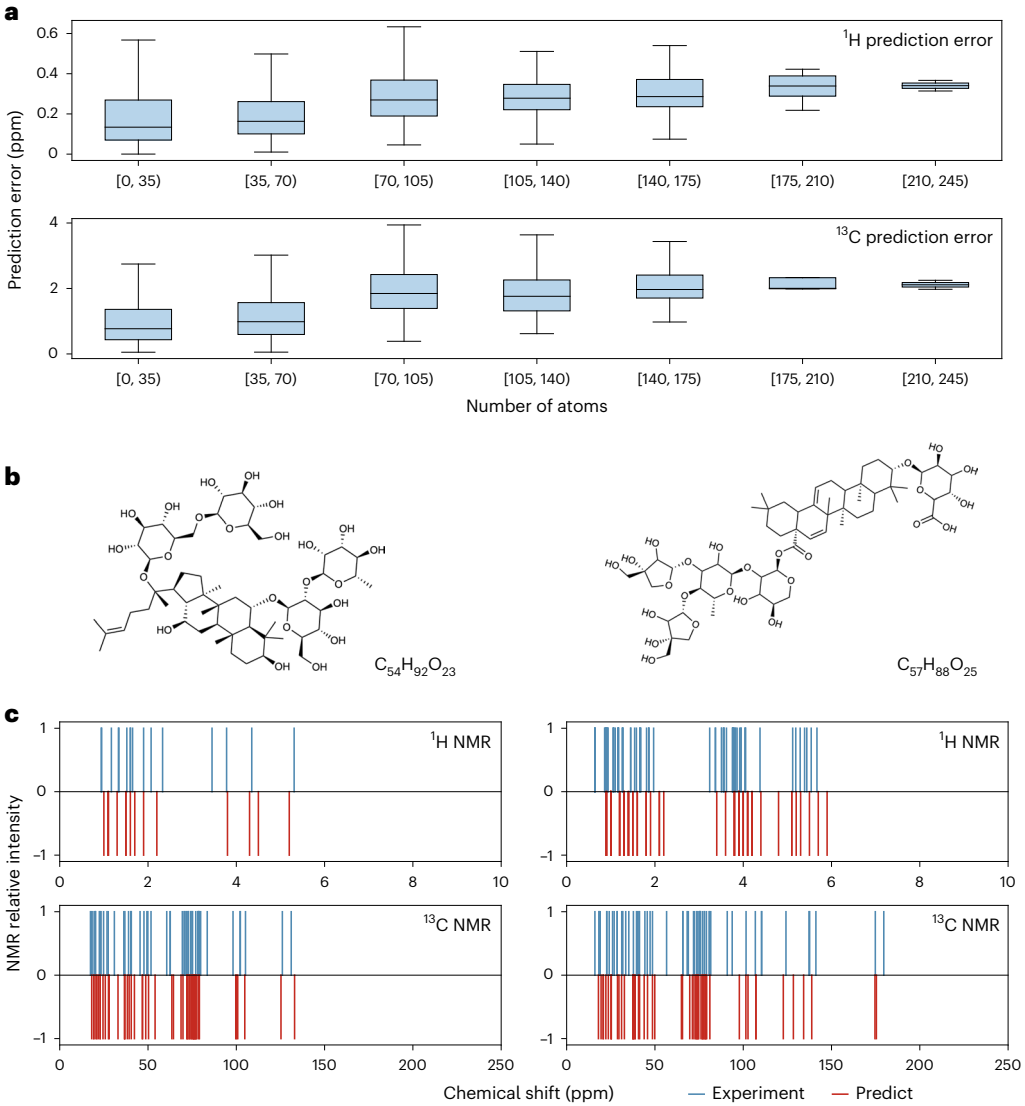


Fig. 2 | Evaluation of the generalization capability of NMRNet. a, The prediction errors on the test set from nmrshiftdb2-2018 (molecules with ≤64 atoms) and an additional test set from nmrshiftdb2-2024 (molecules with ≥70 atoms), grouped by number of atoms in molecules. The samples represent individual atoms with labeled chemical shifts; each data point corresponds to the absolute error between predicted and actual shifts. The top plot shows ^{1}H errors with sample sizes (*n*): 1572, 980, 1218, 191, 70, 29 and 2 (from left to right). Bottom: the ^{13}C errors with *n*: 4273, 1105, 760, 114, 62, 7 and 2 (from left to right). All box plots show the median (central line) and the 25th–75th percentile range (box), and the whiskers extend to the minimum and maximum values within 1.5 times the interquartile range from the box edges. **b**, Two molecules to demonstrate the performance of NMRNet. Left: Yesanchinoside E ($C_{54}H_{92}O_{23}$, nmrshiftdb2 ID: 20173355). Right: chiococcasaponin I ($C_{57}H_{88}O_{25}$, nmrshiftdb2 ID: 20253108). **c**, A comparison of predicted (red) and experimental (blue) chemical shifts for the two molecules in **b**. ^{1}H NMR (top) and ^{13}C NMR (bottom). The peak intensities are all normalized to 1.

with other characterization techniques, such as circular dichroism, could further streamline this complex task.

Correlation between NMR and the local environment. Since chemical shifts reflect the atomic local environment, we validate that NMRNet can accurately represent these environments during both the pretraining and fine-tuning phases (with solid-state NMR data). This is explored by visualizing the local structural representations and their relationship with chemical shifts for all Na^{+} in P2-type $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$ using t-distributed stochastic neighbor embedding. The pretrained model can preliminarily distinguish structures with different shifts, while fine tuning enhances this distinction, indicating improved representation of local environments (Extended Data Fig. 4a,b).

To further explore atomic interactions, we visualized the multi-head attention mechanism of the transformer model using a 64-head attention approach. Focusing on a local environment in P2-type $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$, we analyzed interactions between the central Na^{+} and surrounding atoms. Strong interactions were observed between Na^{+} and Mn^{4+} ions, influencing the electrostatic potential and electronic environment, while fewer interactions were found with Mg^{2+} and O^{2-}. This compound belongs to the P2-type layered oxide family, where Na^{+} ions bridge layers of $Mg_{1/3}Mn_{2/3}$, with substantial electronic contributions from O_{2p} and Mn^{3d} orbitals[28]. Our attention map model successfully captures these electronic signatures.

This study provides a tool to analyze atomic-level interactions without prior chemical knowledge, enhancing the interpretation of NMR spectra and aiding in structure–property relationship determination in complex materials.

Discussion. While NMRNet shows promise in various tasks, including structural assignment and stereochemistry determination, certain limitations remain. For example, the model performance can vary

when applied to very complex or highly specific systems. The reliance on available datasets may also limit its applicability to scenarios where data are scarce or heterogeneous.

Future work should focus on incorporating additional experimental factors such as solvent and temperature. These factors could help improve the model's predictive accuracy and extend its utility to a wider range of chemical environments. Furthermore, further validation through real-world experimental data and larger-scale studies would provide a more comprehensive assessment of NMRNet's generalization ability.

Although challenges remain, this work represents progress in artificial-intelligence-assisted spectral analysis. The framework developed here has the potential to advance the interpretation of NMR spectra and deepen our understanding of molecular structures, while also serving as a foundation for future developments in the field.

Methods

Data preparation process

For both the pretraining and fine-tuning stages, extraction of the list of atom types and their corresponding 3D coordinates from the original dataset is essential. This process can be facilitated by RDKit[29], Atomic Simulation Environment (ASE)[30] or Python Materials Genomics (pymatgen)[31] to extract 3D information from various types of molecular source.

The previous nmrshiftdb2-2018 dataset utilizes data published by Chio et al.[20], which retains the original training/test set division by Kuhn et al.[12]. To build nmrshiftdb2-2024 dataset (introduced by this work), we extract all valid data from nmrshiftdb2 (ref. 10). Then, we use EmbedMolecule and MMFFOptimizeMolecule options of Rdkit to generate 3D conformations of the molecules. For duplicate NMR data records, the median of multiple experiments was used. In addition, InChIKeys were generated using RDKit as the unique ID for each molecule, to ensure that there are no duplicate molecules in the training and test sets. The dataset underwent screening and cleaning to remove recording errors. While the QM9-NMR dataset follows the latest research settings, we could not locate the specific division of NMR data according to DetaNet. Consequently, we reconstructed the original training/test set division based on data reported in the Supplementary Information. For pretraining and solid-state NMR, we utilized the 3D structures from their original datasets, using ASE or pymatgen to read and analyze the local environment of each atom. The ShiftML1 (ref. 15) and ShiftML2 (ref. 22) datasets adhere to their original divisions, while the NN-NMR[18] dataset follows the setup previously established by our research group.

For the liquid state, the chemical environment is based on a single molecule. In contrast, for situations requiring PBC, typically in solid state, the chemical environment of each atom is defined within the unit cell, encompassing all atoms within a specified cutoff radius. In this study, the cutoff radius is set to 6 Å (the convergence check with different cutoff radius is shown in Supplementary Fig. 12), which adequately covers the local environment for most of the atoms. The specific processing of these two representations is shown in Supplementary Fig. 1. After extracting the 3D information, it is further converted into a list of atomic types and a matrix representing the pairwise distances between atoms, serving as input for the model.

The processing steps for each dataset are described below.

Aflow. The entire dataset was downloaded from the Aflow database[32] at https://aflowlib.org/, restricted to unit cells with fewer than 500 atoms for pretraining.

CSD. The entire dataset was downloaded from the Cambridge Structural Database (CSD) database[33] at https://www.ccdc.cam.ac.uk/, restricted to the 30 most common elements and unit cells with fewer than 500 atoms for pretraining.

Materials project. The entire dataset was downloaded from the Materials Project database[34] at https://legacy.materialsproject.org/, restricted to unit cells with fewer than 500 atoms for pretraining.

nmrshiftdb2-2018. The processed data were obtained via Choi et al.[20] at https://github.com/seokhokang/nmr_mpnn/tree/master and follow the original training/test set split[12].

nmrshiftdb2-2024. The data were obtained via the nmrshiftdb2 database[10] at https://nmrshiftdb.nmr.uni-koeln.de/ in .sd file format. Using a script, fields beginning with NMREDATA_1D were selected. The peaks not assigned (probably impurity peaks), empty spectra data and data where the atomic type corresponding to the atomic number does not match its NMR spectrum type were removed. Moreover, data that could not generate 3D conformations using RDKit were excluded. The data outside the specified ranges were filtered: ^{1}H (<0 ppm or >15 ppm), ^{13}C (<−50 ppm or >250 ppm), ^{15}N (<−100 ppm), ^{17}O (>490 ppm) and ^{19}F (<−250 ppm or >100 ppm). The data for ^{31}P and ^{33}S spectra were also filtered. The .sd files also contained cases where the atomic number of an equivalent atom was recorded on the same atom; these were manually corrected and reported to the database administrator. Finally, the median values from multiple repeated experiments for the same molecule were saved, resulting in new .sd files and processed files. The training and test sets were randomly split (0.8:0.2) using InChIKey to ensure molecules in the test set were not in the training set.

QM9-NMR. The entire dataset was downloaded from the QM9-NMR database[11] at https://moldis-group.github.io/qm9nmr/. Because DetaNet[13] did not explicitly display their data split, we finally restored their dataset's training/test split based on the original data provided in DetaNet's supplementary information.

ShiftML1. The entire ShiftML1 dataset was downloaded from the supplementary information of ref. 15, following the original training/test set split. Similar to the MR-3D-DenseNet[23], the data outside the specified ranges were filtered: ^{1}H (<0 ppm or >40 ppm), ^{13}C (<−100 ppm), ^{14}N (<−350 ppm), ^{16}O (<−450 ppm) from the training set, with no additional processing on the test set.

ShiftML2. The entire ShiftML2 dataset[22] was downloaded from their supplementary information at ref. 35, following the original training/test set split. The data outside the specified ranges were filtered: ^{35}Cl (<−150 ppm or >950 ppm), ^{23}Na (>560 ppm), ^{43}Ca (<1,050 ppm or >1,162 ppm) from the training set, with no additional processing on the test set.

NN-NMR. The entire NN-NMR dataset[18] was downloaded via GitHub at https://github.com/chenggroup/nmr, and the original settings of P2-type $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$ were maintained, with the training and test sets randomly split (0.8:0.2).

NMRNet framework

The molecular domain is introduced by Uni-Mol[19], a representation framework with a SE(3) Transformer architecture that has been extensively used for molecular downstream tasks[19,36,37]. Specifically for NMRNet, we adopted and modified Uni-Mol to describe its local environment using atomic representation. Notably, NMRNet applies Uni-Mol to atomic-level tasks, extending its capabilities to solids as well as gaseous and liquid states. To ensure accurate predictions across these scenarios, we collect extensive structural databases for pretraining and NMR datasets for fine tuning, as summarized in Supplementary Table 1.

Pretraining. For the liquid-state NMR, we directly utilized the pretrained weights obtained from previous work by Uni-Mol on a

large-scale molecular dataset[19]. For solid-state NMR, the first strategy (S1, Uni-Mol's original representation strategy) utilizes pretrained weights from previous work[19] and incorporates the unit cell with an intracell distance matrix. The second strategy (S2) uses the same pretrained weights as S1 but uses the unit cell with a global distance matrix. In the third strategy (S3), we maintain the pretrained weight setup but introduce a cutoff radius of 6 Å to define the local environment. Finally, the fourth strategy (S4) was similar to S3, but the pretraining weights are adapted to the cutoff format using a large-scale crystal database. The representation strategy based on single molecules by Uni-Mol was no longer applicable. Therefore, we adopted a representation based on cutoff radius.

Currently, several studies have attempted to mix data from different scales and chemical systems for pretraining, thereby constructing a unified atomic model across scales[38,39]; however, the problem of unified modeling and training has not yet been fully resolved. In this work, our pretraining database comprises Aflow[32], CSD[33] and Materials Project[34]. The CSD primarily consists of organic crystals, while the other two databases mostly include inorganic crystals. Given the substantial differences in chemical environments between organic and inorganic crystals and the varying structural sources of the different datasets, we found that mixed pretraining does not yield better results than pretraining on separated chemical systems (Supplementary Fig. 13). Therefore, we conducted separate pretraining for organic and inorganic crystals (see Supplementary Table 24 for detailed hyperparameter configurations). The pretraining was conducted on eight NVIDIA A100 graphics processing units (GPUs), each equipped with 80 GB of GPU memory.

In the crystal pretraining phase, we used two tasks similar to the mask token tasks in Bert[40] and Uni-Mol[19]: masked atom type prediction and 3D coordinate reconstruction. For each atom's local environment, we replaced the center atom and 15% of randomly selected atoms with (mask) tokens and added ±1 Å noise to their coordinates. A key distinction is that the model only needs to be dedicated to extracting the local environment of the center atom during this phase. Consequently, the additional heads only need to reconstruct the center atom, rather than addressing all masked atoms.

A substantial challenge we addressed was the substantial imbalance in the number of local structures centered around different elements within the dataset. This imbalance made it difficult for the model to adequately focus on the local structures of less common elements. For instance, in the processed CSD dataset containing 92 elements, local structures centered around H or C elements alone accounted for over 84% of the total. To mitigate this issue, we implemented a log-weighted resampling approach to balance the distribution of different local structures within the dataset. The results of this resampling, shown in Supplementary Fig. 2b, ensure that the model can effectively learn the chemical environments of various local structures, including those of less common elements,

$$p_i = \frac{\log(x_i)}{\sum_{j=1}^{n} \log(x_j)}, \tag{1}$$

where p_i is the sampling probability of the ith element in each training iteration after resampling, x_i represents the number of samples in the original dataset that belong to the local structure centered around the ith element, and n is the total number of elements in the dataset.

Fine tuning. In the fine-tuning phase, the model is trained to predict the chemical shift of each atom based on the representation of its local environment. For the prediction of liquid-state NMR, the chemical environment is described by a single molecule. Specifically, the model is provided with a list of atom types in the molecule and a distance matrix of pairwise atomic distances to obtain the representation of each atom's local environment. For solid-state NMR prediction, we comparatively evaluated four distinct strategies to assess the model's capability in representing the local crystal environment

$$L = [\ell_1, \ell_2, \ell_3] \in \mathbb{R}^{3\times 3} \tag{2}$$

$$A_{\text{unitcell}} = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{3\times n} \tag{3}$$

$$A_{\text{infinite}} = \left\{a_i' | a_i + k_1\ell_1 + k_2\ell_2 + k_3\ell_3, k_1, k_2, k_3 \in \mathbb{Z}, i \in \mathbb{Z}, 1 \le i \le n\right\} \tag{4}$$

$$A_{\text{i\_cutoff}} = \left\{a_j'' = a_j' \middle| \left\|a_i - a_j'\right\|_2 < r_{\text{cut}}, 1 \le j \le n\right\}, \tag{5}$$

where n is the number of atoms in the unit cell, and r_{cut} is the cutoff radius, which is set to 6 Å in this work. The unit cell is defined by the lattice matrix L and the set of atoms within the unit cell A_{unitcell}, A_{infinite} represents the infinite crystal structure composed of these unit cells, and $A_{\text{i\_cutoff}}$ denotes the local environment of the ith atom within the unit cell, which includes all atoms in the infinite crystal structure that are within the cutoff radius of this atom;

$$D_1 = \left\{D[i,j] = \left\|a_i - a_j\right\|_2 | 1 \le i,j \le n\right\} \tag{6}$$

$$D_2 = \left\{D[i,j] = \min\left(\left\|a_i - a_j'\right\|_2\right) | 1 \le i,j \le n\right\} \tag{7}$$

$$D_3 = D_4 = \left\{D[i,j] = \left\|a_i'' - a_j''\right\|_2 | 1 \le i,j \le |A_{\text{i\_cutoff}}|\right\}, \tag{8}$$

where n represents the number of atoms in the unit cell, $|A_{\text{i\_cutoff}}|$ is the number of atoms in the local environment of the ith atom under the cutoff format, and D_1–D_4 represent the input distance matrices for the four strategies, respectively. Strategies 1 and 2 input atoms from the unit cell, while strategies 3 and 4 input all atoms within the cutoff radius. The first three strategies utilize the pretrained weights obtained from previous molecular datasets, whereas strategy 4 involves pretraining on crystal structures based on strategy 3,

$$X = F_{\text{emb}}(A, D) = [x_{\text{CLS}}, x_1, x_2, \ldots, x_n] \tag{9}$$

$$\delta_i = G_{\text{nmr}}(x_i),\ 1 \le i \le n, \tag{10}$$

where A and D represent the input atom list and their distance matrix, respectively. The classification token (CLS) is a special token in the input atom list that can be used to represent the entire structure, and n denotes the number of atoms in the input structure. X is the set of atom representations, x_{CLS} is the representation of the CLS token, x_i are the individual atom representation and δ_i is the predicted chemical shift. F_{emb} is the network that obtains the local environment representation for each atom, composed of the backbone of Uni-Mol. G_{nmr} is the network that utilizes these representations to predict chemical shifts, consisting of a feedforward neural network with two layers of nonlinear transformations. The fine-tuning process used fivefold cross validation, and the average performance of the five models on the test set is reported. The detailed hyperparameter configurations are provided in Supplementary Table 25. The fine tuning was performed on one NVIDIA V100 GPU with 32 GB of GPU memory.

Evaluation metrics

In this study, we use several metrics to evaluate the performance of chemical shift predictions: the MAE, the RMSE, the R^2, the mean absolute deviation (MAD) and the RMSD. In these formulas, y_i represents the actual observed chemical shift for the ith sample, and $\hat{y}_i$ denotes the predicted chemical shift for the ith sample.

The MAE is calculated as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{11}$$

where n is the number of samples. MAE measures the average magnitude of errors in a set of predictions, without considering their direction.

The RMSE is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{12}$$

where RMSE represents the standard deviation of the residuals (prediction errors).

Before calculating the R^2, we introduce $\bar{y}$, which represents the mean of all actual observed chemical shifts

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{13}$$

The R^2 is then computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{14}$$

where R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The ranges are from 0 to 1, with values closer to 1 indicating a better fit of the model to the data.

For a set of predicted and experimental chemical shifts, we use slightly different notation to emphasize the specific context of chemical shift comparisons. $\delta_{\mathrm{pred},j}$ and $\delta_{\mathrm{exp},j}$ are the predicted and experimental chemical shifts for the jth nucleus, respectively, and m is the total number of chemical shift comparisons. The MAD is calculated as

$$\mathrm{MAD} = \frac{1}{m}\sum_{j=1}^{m}|\delta_{\mathrm{pred},j} - \delta_{\mathrm{exp},j}|. \tag{15}$$

Similarly, the RMSD is given by

$$\mathrm{RMSD} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\delta_{\mathrm{pred},j} - \delta_{\mathrm{exp},j})^2}. \tag{16}$$

Both the MAD and RMSD provide measures of the average difference between predicted and experimental chemical shifts, with the RMSD giving more weight to larger deviations.

To provide a comprehensive evaluation of chemical shift predictions for both carbon and proton nuclei, we calculate a total RMSD that combines the RMSD values for ^{13}C and ^{1}H

$$\text{Total RMSD} = {}^{13}\text{C RMSD} + 10 \times {}^{1}\text{H RMSD}, \tag{17}$$

where ^{13}C RMSD is the RMSD for ^{13}C chemical shifts and ^{1}H RMSD is the RMSD for ^{1}H chemical shifts. The factor of 10 applied to the proton RMSD accounts for the typical difference in scale between carbon and proton chemical shifts, allowing a balanced contribution from both nuclei in the total score.

Modeling and analysis of molecules and materials

RDKit[29] is used for reading or generating molecular SMILES, molecular objects, InChIKeys, two-dimensional structures and 3D conformations. ASE[30] and pymatgen[31] are utilized for reading and processing 3D conformations of crystal structures. The Molecule Recognition application[41] of Bohrium platform is utilized for automated identification and extraction of SMILES of molecular structures from images and scientific literature. MolView[42] is used for the visualization of 3D molecular structures, while VESTA[43] is used for the visualization of 3D crystal structures.

Data availability

Source data for Fig. 2 and Extended Data Figs. 1, 2 and 4 are available with this Brief Communication. All structural datasets used for pretraining are publicly accessible. The Aflow dataset[32] is available at https://aflowlib.org/, the Materials Project dataset[34] is accessible at https://next-gen.materialsproject.org/ and the CSD dataset[33] is accessible at https://www.ccdc.cam.ac.uk/. All processed NMR datasets used for fine tuning are available via Zenodo at https://doi.org/10.5281/zenodo.13317524 (ref. 44).

Code availability

The NMRNet code is available via GitHub at https://github.com/Colin-Jay/NMRNet and via Zenodo at https://doi.org/10.5281/zenodo.14741405 (ref. 45) under an open-source license. The trained model parameters are available via Zenodo at https://doi.org/10.5281/zenodo.13317524 (ref. 44). A demo notebook of NMRNet is available at https://bohrium.dp.tech/notebooks/38356712597, and an online service is available at https://ai4ec.ac.cn/apps/nmrnet and https://bohrium.dp.tech/apps/nmrnet001.

References

1. Xue, X. et al. Advances in the application of artificial intelligence-based spectral data interpretation: a perspective. *Anal. Chem.* **95**, 13733–13745 (2023).
2. Lu, X.-Y. et al. Deep learning-assisted spectrum–structure correlation: state-of-the-art and perspectives. *Anal. Chem.* **96**, 7959–7975 (2024).
3. Hu, G. & Qiu, M. Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Nat. Prod. Rep.* **40**, 1735–1753 (2023).
4. Smith, S. G. & Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J. Am. Chem. Soc.* **132**, 12946–12959 (2010).
5. Tsai, Y. -H. et al. ML-*J*-DP4: an integrated quantum mechanics–machine learning approach for ultrafast NMR structural elucidation. *Org. Lett.* **24**, 7487–7491 (2022).
6. Jonas, E., Kuhn, S. & Schlörer, N. Prediction of chemical shift in NMR: a review. *Magn. Reson. Chem.* **60**, 1021–1031 (2022).
7. Cortés, I., Cuadrado, C., Hernández Daranas, A. & Sarotti, A. M. Machine learning in computational NMR-aided structural elucidation. *Front. Nat. Prod.* **2**, 1122426 (2023).
8. Gerrard, W. et al. Impression–prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **11**, 508–515 (2020).
9. Yang, Z., Chakraborty, M. & White, A. D. Predicting chemical shifts with graph neural networks. *Chem. Sci.* **12**, 10802–10809 (2021).
10. Kuhn, S. & Schlörer, N. E. 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**, 582–589 (2015).
11. Gupta, A., Chakraborty, S. & Ramakrishnan, R. Revving up ^{13}C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach. Learn. Sci. Technol.* **2**, 035010 (2021).
12. Jonas, E. & Kuhn, S. Rapid prediction of NMR spectral properties with quantified uncertainty. *J. Cheminform.* **11**, 50 (2019).
13. Zou, Z. et al. A deep learning model for predicting selected organic molecular spectra. *Nat. Comput. Sci.* **3**, 957–964 (2023).

14. Atwi, R. et al. An automated framework for high-throughput predictions of NMR chemical shifts within liquid solutions. *Nat. Comput. Sci.* **2**, 112–122 (2022).
15. Paruzzo, F. M. et al. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **9**, 4501 (2018).
16. Lin, M. et al. Unravelling the fast alkali–ion dynamics in paramagnetic battery materials combined with NMR and deep-potential molecular dynamics simulation. *Angew. Chem.* **133**, 12655–12661 (2021).
17. Lin, M., Fu, R., Xiang, Y., Yang, Y. & Cheng, J. Combining NMR and molecular dynamics simulations for revealing the alkali–ion transport in solid-state battery materials. *Curr. Opin. Electrochem.* **35**, 101048 (2022).
18. Lin, M. et al. A machine learning protocol for revealing ion transport mechanisms from dynamic NMR shifts in paramagnetic battery materials. *Chem. Sci.* **13**, 7863–7872 (2022).
19. Zhou, G. et al. Uni-Mol: a universal 3D molecular representation learning framework. In *Proc. International Conference on Learning Representations* (eds Yan, L. et al.) (ICLR, 2023).
20. Kwon, Y., Lee, D., Choi, Y.-S., Kang, M. & Kang, S. Neural message passing for nmr chemical shift prediction. *J. Chem. Inf. Model.* **60**, 2024–2030 (2020).
21. Han, J. et al. Scalable graph neural network for nmr chemical shift prediction. *Phys. Chem. Chem. Phys.* **24**, 26870–26878 (2022).
22. Cordova, M. et al. A machine learning model of chemical shifts for chemically and structurally diverse molecular solids. *J. Phys. Chem. C* **126**, 16710–16720 (2022).
23. Liu, S. et al. Multiresolution 3D-densenet for chemical shift prediction in NMR crystallography. *J. Phys. Chem. Lett.* **10**, 4558–4565 (2019).
24. Jeong, K. et al. Precisely predicting the ^{1}H and ^{13}C NMR chemical shifts in new types of nerve agents and building spectra database. *Sci. Rep.* **12**, 20288 (2022).
25. Gao, P., Zhang, J., Peng, Q., Zhang, J. & Glezakou, V.-A. General protocol for the accurate prediction of molecular ^{13}C/^{1}H nmr chemical shifts via machine learning augmented DFT. *J. Chem. Inf. Model.* **60**, 3746–3754 (2020).
26. Wu, A. et al. Elucidating structures of complex organic compounds using a machine learning model based on the ^{13}C NMR chemical shifts. *Precis. Chem.* **1**, 57–68 (2023).
27. Ai, W.-J. et al. A very deep graph convolutional network for ^{13}C NMR chemical shift calculations with density functional theory level performance for structure assignment. *J. Nat. Prod.* **87**, 743–752 (2024).
28. Vergnet, J., Saubanère, M., Doublet, M.-L. & Tarascon, J.-M. The structural stability of P2-layered Na-based electrodes during anionic redox. *Joule* **4**, 420–434 (2020).
29. Landrum, G. et al. Rdkit. *Zenodo* https://doi.org/10.5281/zenodo.14779836 (2024).
30. Larsen, A. H. et al. The Atomic Simulation Environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
31. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
32. Curtarolo, S. et al. Aflow: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
33. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The cambridge structural database. *Acta Cryst. B* **72**, 171–179 (2016).
34. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL. Mater.* **1**, 011002 (2013).
35. Cordova, M. et al. ShiftML. *Zenodo* https://doi.org/10.5281/zenodo.6782653 (2022).
36. Luo, W. et al. Bridging machine learning and thermodynamics for accurate pK_a prediction. *JACS Au* **4**, 3451–3465 (2024).
37. Yao, L. et al. Node-aligned graph-to-graph: elevating template-free deep learning approaches in single-step retrosynthesis. *JACS Au* **4**, 992–1003 (2024).
38. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* **630**, 493–500 (2024).
39. Zhang, D. et al. DPA-2: a large atomic model as a multi-task learner. *NPJ Comput. Mater.* **10**, 293 (2024).
40. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT* (eds Burstein, J. et al.) (Association for Computational Linguistics, 2019).
41. Fang, X. et al. MolParser: end-to-end visual recognition of molecule structures in the wild. Preprint at https://arxiv.org/abs/2411.11098v2 (2024).
42. Bergwerf, H. Molview: an attempt to get the cloud into chemistry classrooms. *Comm. Comput. Chem. Educ.* **9**, 1–9 (2015).
43. Momma, K. & Izumi, F. Vesta 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).
44. Xu, F. et al. NMRNet dataset. *Zenodo* https://doi.org/10.5281/zenodo.13317524 (2024).
45. Xu, F. NMRNet v1.0.0 code. *Zenodo* https://doi.org/10.5281/zenodo.14741405 (2025).

Acknowledgements

We thank Y. Ren and J. Zhang for their contributions to the design of the manuscript's cover. We thank Y. Tang and J. Qiu for his valuable improvements to the schematic diagram. We are also grateful for the insightful discussions and suggestions from Y. Liu, J. Zou, Y. Zhuang, Y. Jin, F. Fu, W. Luo, G. Zhou and J. Wang. F.T. acknowledges the National Key R&D Program of China (grant no. 2024YFA1210804) and a startup fund at Xiamen University. J.C. acknowledges the National Natural Science Foundation of China (grant nos. 22225302, 92470201, 22021001, 92461312, 21991151, 21991150, 92161113 and 22411560277), the Fundamental Research Funds for the Central Universities (20720220009), Laboratory of AI for Electrochemistry (AI4EC), IKKEM (grant nos. RD2023100101 and RD2022070501).

Author contributions

J.C., F.T. and Z.G. contributed to the design of the work. F.X. and W.G. completed data collection and cleaning. F.X. developed the NMRNet code. F.X., W.G. and Z.G. contributed to the software development. F.X., W.G. and F. T. performed data analysis. All authors participated in the discussion and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-025-00783-z.

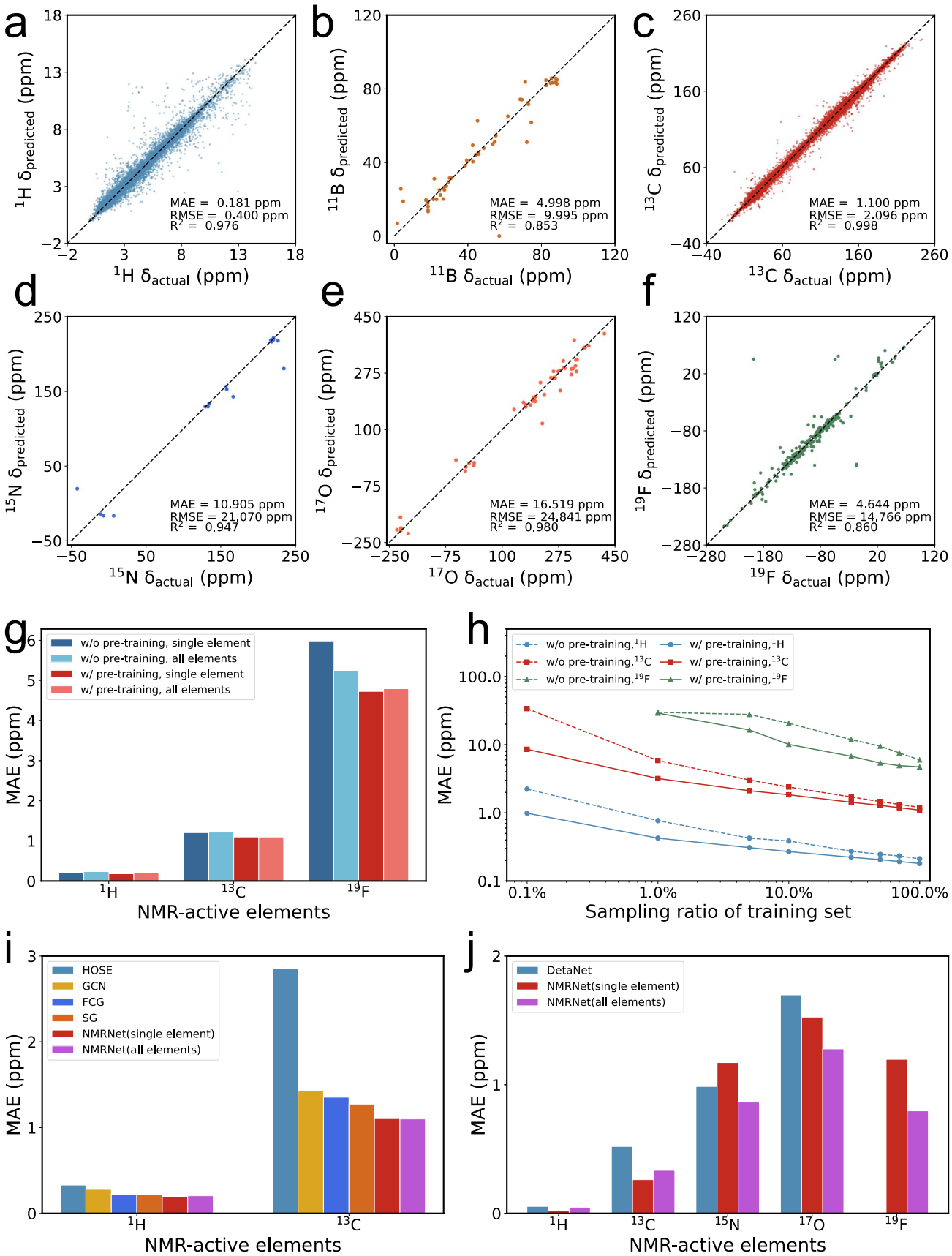
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00783-z.

Correspondence and requests for materials should be addressed to Fujie Tang, Zhifeng Gao or Jun Cheng.

Peer review information *Nature Computational Science* thanks Joshua D. Hartman, Nav Nidhi Rajput and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary

Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

https://doi.org/10.1038/s43588-025-00783-z

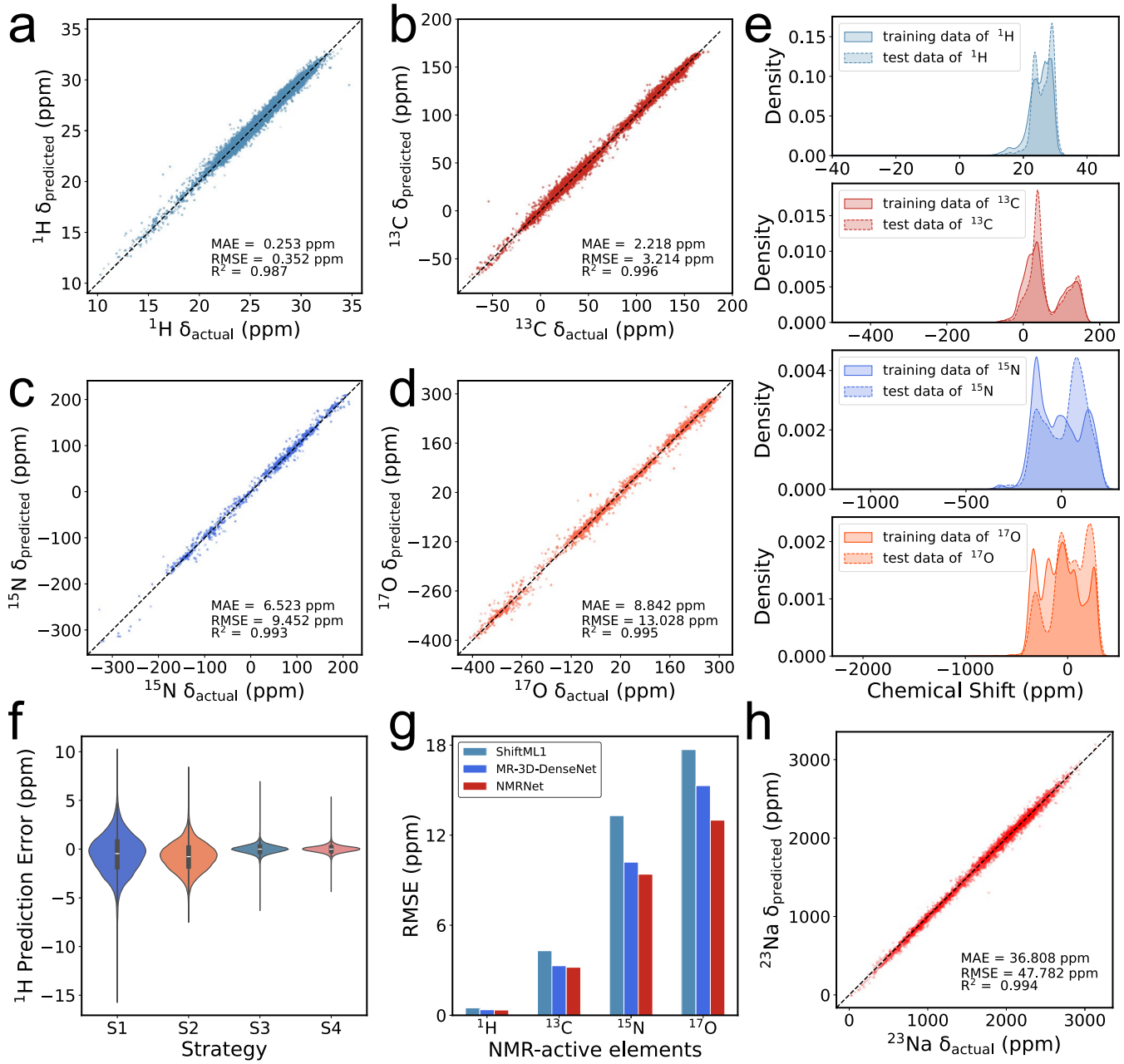


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s43588-025-00783-z

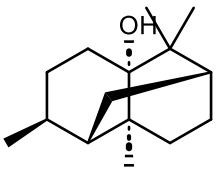
Extended Data Fig. 1 | Performance of NMRNet in liquid-state NMR prediction. NMRNet's correlation scatter plots of predicted versus experimental chemical shifts for (**a**) ^{1}H, (**b**) ^{11}B, (**c**) ^{13}C, (**d**) ^{15}N, (**e**) ^{17}O, and (**f**) ^{19}F in the nmrshiftdb2-2024 dataset. (**g**) Comparison of the prediction error (MAE) for different elements in the nmrshiftdb2-2024 dataset when using (marked as 'w/ pre-training') or not using pre-trained (marked as 'w/o pre-training') weights and when predicting a single element versus all elements simultaneously. (**h**) Comparison of prediction error (MAE) for different elements in the nmrshiftdb2-2024 dataset using different proportions of the training set, noting that the data volume for the ^{19}F element does not support a 0.1% setting. To facilitate the presentation, both the horizontal and vertical axes are scaled logarithmically. Comparison of prediction error (MAE) for different elements in (**i**) the nmrshiftdb2-2018 dataset and (**j**) the QM9NMR dataset against previous studies. Note that DetaNet has not reported results for ^{19}F.

https://doi.org/10.1038/s43588-025-00783-z

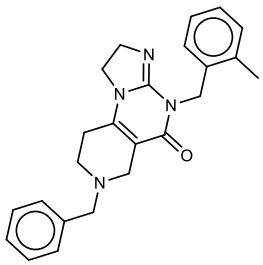


Extended Data Fig. 2 | Performance of NMRNet in solid-state NMR prediction. NMRNet's correlation scatter plots of predicted versus DFT-calculated chemical shifts (chemical shieldings) for (**a**) ^{1}H, (**b**) ^{13}C, (**c**) ^{15}N, and (**d**) ^{17}O in the ShiftML1 dataset. (**e**) Distribution of chemical shifts for four elements in the ShiftML1 dataset. (**f**) The impact of four strategies on the prediction error (MAE) for ^{1}H in the ShiftML1 dataset using NMRNet. Samples represent individual atoms with labeled chemical shifts; each data point corresponds to the absolute error between predicted and actual shifts. The sample sizes (n) is 29,913. S1-S3 utilized pre-trained weights on molecular dataset, differing in their use of the unit cell with intra-cell distance matrix, the unit cell with global distance matrix, and cutoff radius = 6 Å as the local environment for a single atom, respectively. S4 modifies the pre-training in S3 to pre-training with the cutoff format on a large-scale crystal database. (**g**) Comparison of the prediction error (RMSE) for different elements in the ShiftML1 dataset using NMRNet with previous studies. (**h**) NMRNet's correlation scatter plot of predicted versus DFT-calculated chemical shifts (chemical shieldings) for ^{23}Na in the NN-NMR dataset.

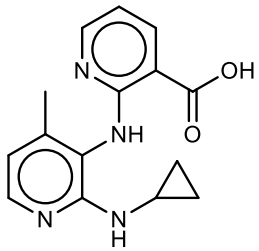
https://doi.org/10.1038/s43588-025-00783-z



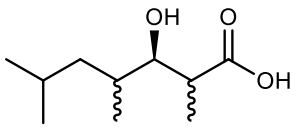
1-Benzoylpyrrolidine
2 configuration isomers



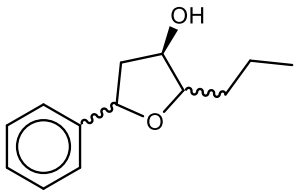
TIC10
3 configuration isomers



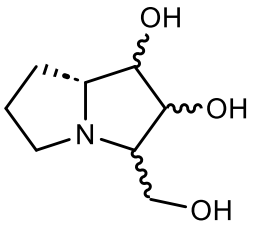
Nevirapine
4 configuration isomers



(3R)-3-hydroxy-2,4,6-trimethylheptanoic
4 R/S isomers



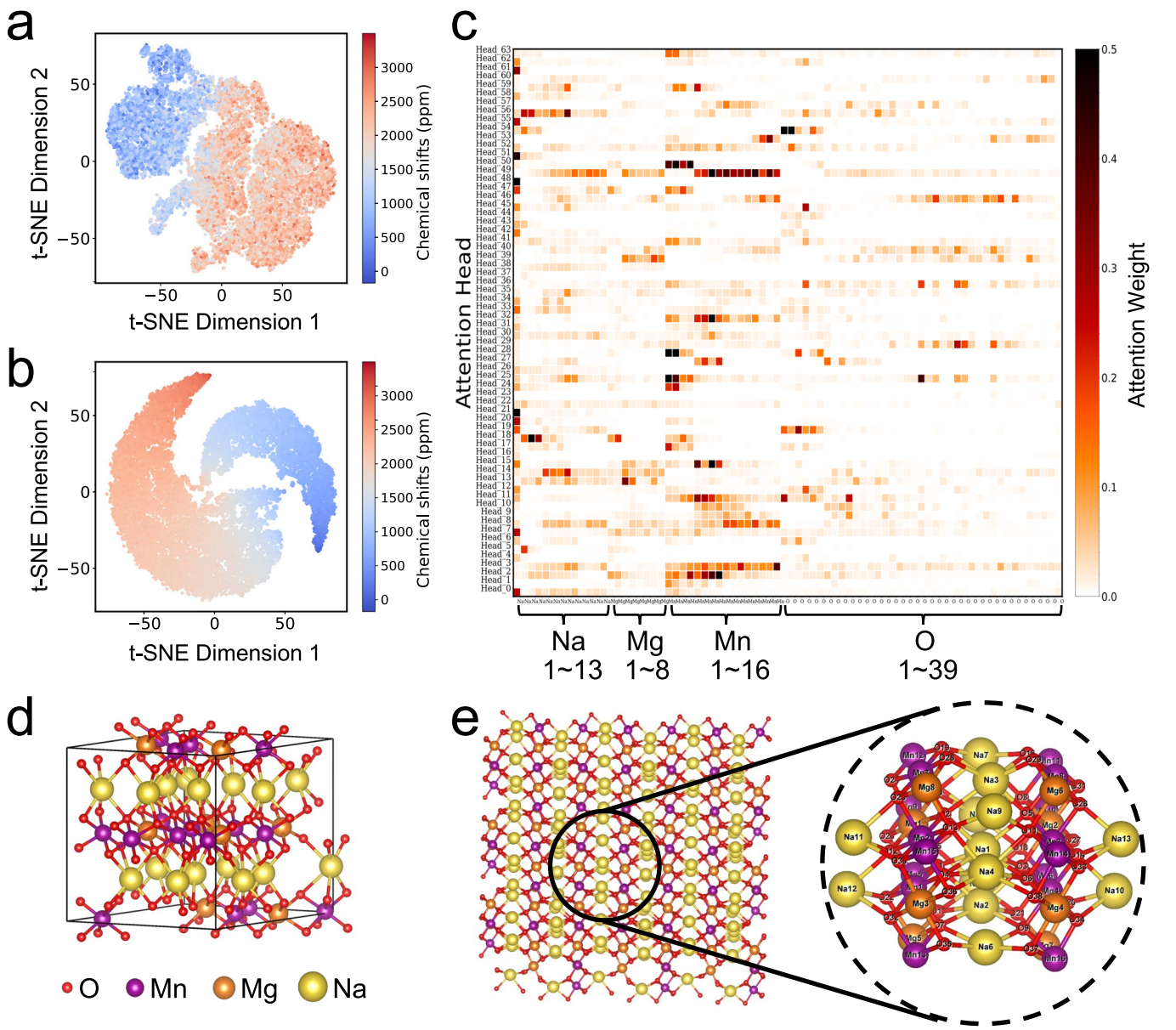
(3R)-5-phenyl-2-propyloxolan-3-ol
4 R/S isomers



Hyacinthacines
8 R/S isomers

Extended Data Fig. 3 | Six examples used in configuration determination. The top three are for the structure revision task, and the bottom three are for the chiral isomer identification task.

https://doi.org/10.1038/s43588-025-00783-z



Extended Data Fig. 4 | Structural representations by NMRNet. Local structural representations of and their relationship with chemical shifts for all Na^+ in P2-type $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$ using t-SNE for the (**a**) pre-trained NMRNet and (**b**) fine-tuned NMRNet. (**c**) Extract the interaction information between each central atom (represented as Na1) and its local environment ($Na_{13}Mg_8Mn_{16}O_{39}$) from the results of the 64-head attention mechanism of the Transformer, each head's results are represented as a separate row, and these results are then concatenated together. Identical elements are arranged in ascending order based on their distances from the central atom. The darker color in the visualization indicates stronger correlations between the central atom and its local environment. (**d**) A unit cell of $Na_{2/3}(Mg_{1/3}Mn_{2/3})O_2$. (**e**) The local environment of Na extracted from the infinite crystal structure corresponding to the unit cell in (**d**).