JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

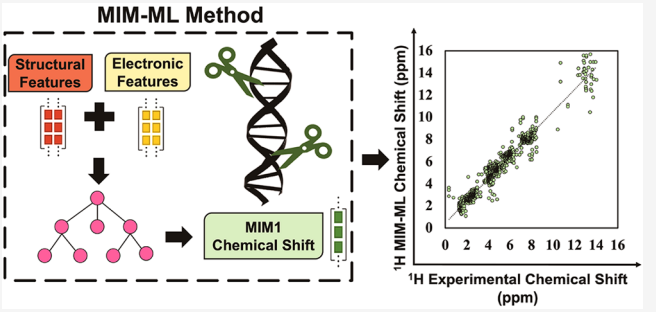
MIM-ML: A Novel Quantum Chemical Fragment-Based Random Forest Model for Accurate Prediction of NMR Chemical Shifts of Nucleic Acids

Sruthy K. Chandy* and Krishnan Raghavachari*

Cite This: J. Chem. Theory Comput. 2023, 19, 6632–6642 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: We developed a random forest machine learning (ML) model for the prediction of ^{1}H and ^{13}C NMR chemical shifts of nucleic acids. Our ML model is trained entirely on reproducing computed chemical shifts obtained previously on 10 nucleic acids using a Molecules-in-Molecules (MIM) fragment-based density functional theory (DFT) protocol including microsolvation effects. Our ML model includes structural descriptors as well as electronic descriptors from an inexpensive low-level semiempirical calculation (GFN2-xTB) and trained on a relatively small number of DFT chemical shifts (2080 ^{1}H chemical shifts and 1780 ^{13}C chemical shifts on the 10 nucleic acids). The ML model is then used to make chemical shift predictions on 8 new nucleic acids ranging in size from 600 to 900 atoms and compared directly to experimental data. Though no experimental data was used in the training, the performance of our model is excellent (mean absolute deviation of 0.34 ppm for ^{1}H chemical shifts and 2.52 ppm for ^{13}C chemical shifts for the test set), despite having some nonstandard structures. A simple analysis suggests that both structural and electronic descriptors are critical for achieving reliable predictions. This is the first attempt to combine ML from fragment-based DFT calculations to predict experimental chemical shifts accurately, making the MIM-ML model a valuable tool for NMR predictions of nucleic acids.



1. INTRODUCTION

NMR spectroscopic investigations play an integral role in the determination of the structural properties of proteins and nucleic acids. Due to the complex nature of the analysis of experimental data needed to derive reliable structural assignments, it is imperative to develop complementary theoretical and computational methods to assist the prediction of NMR chemical shifts of biomolecules.

Proteins, with a larger database of experimental NMR chemical shifts (as of May 2023, around 12,000 solution NMR-resolved structures in PDB),[2] have been the subject of numerous prediction studies.[3−9] In the early 2000s, empirical NMR predictors were developed, which subsequently paved the way for machine learning-based (ML) predictors, which are trained on high quality experimental data of proteins. Among these empirical predictors, there are simple machine learning models such as SPARTA+[10] with a single layer feed-forward network, SHIFTX2/SHIFTX+,[11] which leverages sequence homology, PPM_One[12] utilizing an artificial neural network model, and Graph NMR[13] utilizing graph neural networks. More recently, UCBShift[14] involving decision tree ensemble models, which employ both sequence and structural alignment, demonstrated better performance in predicting chemical shifts of aqueous protein structures relative to earlier predictors.[15]

NMR chemical shift studies play a pivotal role in the characterization of nucleic acids, despite their historical focus lagging behind that of proteins. These studies are particularly valuable for nucleic acids due to the inherent complexities associated with NMR experiments and the subsequent assignment of chemical shifts. Notably, the dynamic nature of nucleic acids, especially in the case of RNA, presents a significant challenge in generating a comprehensive description of interproton NOE-derived distance restraints required for their structural determination.[16,17] In contrast with proteins, the number of nucleic acid structures with experimental chemical shift data deposited in PDB is relatively small (less than 2000 structures solved using solution NMR).[2] As a result, fewer empirical methods have been explored for predicting the

Received: May 26, 2023
Published: September 13, 2023

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

chemical shifts of nucleic acids.[16] The methods by Altona et al.,[18] Barton et al.,[19] and DSHIFT[20] by Lam et al. have used empirical approaches based on a central nucleotide and its neighbors. Other empirical methods such as SHIFTS[21] and NUCHEMICS[22] along with the method by Sahakyan et al.[23] and Cromsigt et al.[22] use empirical equations to model ring current and electrostatic effects. The ML-based RAMSEY[24] method by Frank et al. uses a random forest approach with 3D structural descriptors including torsions and hydrogen bonding. Later a support vector regression method trained on sequences of RNA systems was introduced by Brown et al.[25]

Although empirical methods offer quick predictions based on experimental nucleic acid data, they have limitations. These methods heavily rely on a limited set of high-quality nucleic acid structures. Since they are based on empirical or semiempirical equations, they may not be able to handle noncanonical nucleic acid structures. Also, since they are mainly trained on nonexchangeable protons, they perform poorly for solvent-exchangeable protons, imino protons, and nonproton nuclei. Additionally, these empirical techniques can also prove insensitive to structural variations in nucleic acids, potentially leading to inaccuracies in predictions.[16]

Density functional theory (DFT) methods have shown accurate NMR chemical shift predictions of a wide variety of molecules.[26−31] As a result, ML models for the prediction of DFT-quality chemical shifts for small and medium sized organic molecules have been explored extensively.[32−38] These ML techniques leverage the localized nature of the chemical shielding tensor to make predictions. Interestingly, SHIFTML model based on local atomic environments has been used to predict chemical shifts of molecular solids within DFT accuracy and has demonstrated good performance in predicting experimental chemical shifts.[39]

However, due to the steep computational cost, it is nearly impossible to do DFT-based GIAO NMR calculations on large biological systems with thousands of atoms. Therefore, ML models on predicting DFT-quality chemical shifts for large biological systems such as proteins and nucleic acids are not explored. In this context, fragment-based methods have shown success in calculating chemical shifts of large biomolecules using DFT or other first-principles methods.[40−46] Among the methods based on *ab initio* techniques, the QM/MM fragment-based AFNMR[47] and ADMA[16] methods have shown modest performance with errors of around 0.4−0.6 ppm for ^{1}H chemical shifts of DNA and RNA systems.

To overcome the deficiencies of existing methods and obtain better performance, we have developed the QM-based Molecules-in-Molecules (MIM) fragmentation method for the accurate prediction of experimental NMR chemical shifts for both backbone and side-chain ^{1}H and ^{13}C in proteins as well as nucleic acids.[48−50] Our DFT-based MIM-NMR has emerged as an accurate predictive tool for NMR chemical shifts of nucleic acids with a mean absolute deviation from experiment of ∼0.3 ppm for ^{1}H and ∼2−3 ppm for ^{13}C.[49] In this study, we leveraged the simplicity and accuracy of our MIM-NMR model to develop and train a machine learning model (MIM-ML) with minimal loss of accuracy to make faster predictions that are applicable to larger systems.

The main objective of this work is to train and develop an AI/ML model for predicting chemical shifts of DFT/experimental quality. In contrast to existing machine learning (ML) models that are trained on experimental data sets, which can be prone to errors and noise, our MIM-ML model was exclusively trained on DFT-calculated chemical shifts derived from MIM-NMR of 10 nucleic acid systems from our previous study. Our approach employed a random forest[51] (RF) architecture, which is well-suited for establishing relationships between chemical shifts and molecular structures. It excels in handling multidimensional data sets and modeling nonlinear relationships. Moreover, the random forest approach is robust against overfitting and computationally efficient due to its algorithmic simplicity.[24] Most importantly, we incorporated both structure-based and electronic descriptors for training our ML model. We utilized this model to predict the chemical shifts of eight new nucleic acids with sizes ranging from 600 to 900 atoms and compared them directly with experimental data to obtain excellent results with very little loss of accuracy compared to the training set.

2. METHODS

2.1. MIM1-NMR Pipeline for Calculated Chemical Shifts. The MIM-NMR calculated chemical shifts are taken from our previous study on 10 nucleic acids.[49] We used "dimer" primary subsystems that incorporate stacking interactions between adjacent units, defined by the fragmentation parameter (r), calculated with the selected high level of theory (*vide infra*). The derivative subsystems to account for the overcounting are then the corresponding "monomer" units. Single-layer MIM (MIM1) energies are obtained from a summation of the subsystem energies and can be written like the standard ONIOM extrapolation, as shown in eq 1.

$$E^{\mathrm{MIM1}} = E^{r}_{\mathrm{high}} = \sum \text{dimers} - \sum \text{overlapping monomers} \tag{1}$$

For the NMR-GIAO method, the isotropic shielding tensor, σ^{N} for atom N, is given as the second derivative of the electronic energy, E, with respect to the external magnetic field B and the nuclear magnetic moment m_N.

$$\sigma^{N}_{ij} = \left[\frac{\partial^2 E}{\partial B_i \partial m_{N_j}}\right]_{B=0} \tag{2}$$

σ^{N}_{ij} is the ij^{th} component of the shielding tensor, B_i is the i^{th} component of the external magnetic field, and m_{N_j} is the j^{th} component of magnetic moment of the nucleus N.

The atomic NMR shielding constant is one-third of the sum of the trace of the atomic shielding tensors from eq 2. σ_i, which is the isotropic chemical shift, is subtracted from the corresponding standard reference value (σ_{ref}) to yield the chemical shift of each atomic species. For ^{1}H and ^{13}C, the chemical shift is calculated using tetramethyl silane (TMS) as the reference. For ^{15}N, the NH_3 molecule is taken as the references.

$$\delta_i = \sigma_{\mathrm{ref}} - \sigma_i \tag{3}$$

In our MIM1-NMR model, we use the *m*PW1PW91/6-311G(d,p) method to calculate the shielding constants. The scaling factors determined for ^{1}H, ^{13}C, and ^{15}N nuclei using the *m*PW1PW91/6-311G(d,p) method are 31.86, 190.33, and 273.38 ppm, respectively. All MIM-NMR calculations were performed using an external perl module and the Gaussian16 program suite.[52]

For all of the nucleic acids, we performed structure minimization using AMBER10: EHT force fields by constraining all the heavy atoms and allowing only the protons to move.[53,54] Additionally, we incorporated the solvation effects in our MIM1-NMR method using an explicit-implicit solvation

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

model ($MIM^{constraint}_{explicit-implicit}$), which is herein termed the $MIM1_{microsolavation}$ model. In our microsolvation model, the short-range hydrogen-bonding interactions are captured by including *one explicit water molecule per amine proton*, and the remaining solvation effects are captured using an appropriate implicit solvation model. For implicit solvation, we use the widely used SMD-SCRF continuum solvation model.[55] The assumption here is that the dimer primary subsystems with explicit-implicit solvation can accurately model the local intramolecular hydrogen bonding interactions as well as intermolecular explicit interactions with the solvent at a high level of theory. Before the chemical shift calculation, the explicit water molecules are geometry-optimized using the PM6D3H4 semiempirical method using MOPAC (Molecular Orbital PACkage), while freezing the rest of the DNA molecule to preserve its conformation.[56]

The $MIM1_{microsolavtion}$-NMR model is efficient and was applied recently to a test set of 10 nucleic acids, including some nonstandard systems containing B and F atoms, to achieve a target performance of 0.3 ppm for ^{1}H and 2−3 ppm for ^{13}C chemical shifts, similar in magnitude to the values obtained previously on a variety of protein systems.[48,49] We train our ML model on this set of nucleic acids, as described in the next section.

For the MIM-ML model, the input structural minimization protocol is similar to the procedure in the $MIM1_{microsolavtion}$-NMR model, with only the exception of not including the explicit solvent molecules. We expect our MIM-ML model to capture any additional interactions present in the $MIM1_{microsolavtion}$-NMR model from appropriate structural and electronic descriptors (*vide infra*).

2.2. Descriptors. The choice of descriptors can greatly impact the accuracy of any ML model, as well as its interpretability. It is therefore important to carefully consider the specific problem being addressed and to select descriptors that are relevant as well as informative. Since chemical shift is a highly localized property, it is logical to include the local *structural* and *chemical* environment of the target atom to design the input feature vector.[57] Thus, we chose appropriate structural and electronic descriptors to represent the features of the data and to provide information about the nuclei type being analyzed to build a reliable random forest (RF) model. More precisely, we designed separate RF models for ^{1}H and ^{13}C incorporating such atomic features to predict their chemical shifts.

2.2.1. Structural Descriptors. We extracted the structural descriptors from the representative structure of the nucleic acid deposited in the Protein Data Bank using the open source "Biopython library". The first step is to parse the PDB files using the PDBParser module, which transforms the information stored in the files into a structured format that can be easily processed. The data obtained through parsing include details about individual residues, atoms, and their respective properties. The extracted data are encapsulated within a Structure object, which serves as a container that holds essential information about the molecular structure of the nucleic acid. Through the utilization of various functions and methods provided by the Biopython library, relevant information such as atom types and nucleotide residues is extracted from the Structure object.

To incorporate the local environmental effects surrounding specific nuclei within the nucleic acid structure, a NeighborSearch object is employed. This object facilitates the identification of neighboring atoms and residues in the proximity of the target nuclei. Specifically, the search method of the NeighborSearch object is utilized to determine the nearest neighbors of nuclei such as ^{1}H and ^{13}C. For ^{1}H nuclei, the two closest neighboring atoms are identified, while for ^{13}C nuclei, the four closest neighbors are ascertained. Additionally, the search method is also applied to locate the two closest neighboring residues for each residue in the nucleic acid structure. These neighboring residues correspond to the adjacent bases in the single chain of the structure.

Overall, this protocol allowed us to extract important structural descriptors such as (1) atom type, (2) residue type, (3) neighbor atom type, and (4) neighbor residue type from the PDB file and analyze the local environment of the nucleic acid structure. In order to prepare the extracted structural features for analysis using the MIM-ML model, a one-hot encoding representation is employed. This encoding scheme transforms categorical data, such as atom types and residue types, into binary vectors, where each category is represented as a unique binary pattern. The resulting one-hot encoded feature vectors are subsequently used as inputs for the MIM-ML model.

2.2.2. Electronic Descriptors. To obtain electronic features, we used the highly efficient and well-calibrated GFN2-xTB semiempirical QM method developed by Grimme's group to calculate the electronic properties of the nucleic acids.[58] We performed single point calculations using GFN2-xTB in conjunction with a GBSA (Generalized Born with Solvent Accessibility) implicit solvation model on a hybrid molecular mechanics/semiempirical constraint-optimized structure. Here, GBSA[59] models the solvent as a continuous dielectric medium and accounts for the electrostatic interactions between the nucleic acid and solvent molecules. The hybrid method combines features of both molecular mechanics and semiempirical methods to optimize the geometry. As mentioned in the method section (*vide supra*), we first used AMBER10: EHT force fields to minimize the representative PDB structure of the nucleic acid, constraining all heavy atoms and allowing only protons to move. We then refined the geometry using PM6D3H4 with similar constraints. This two-step process provides a reliable structure for NMR calculations by optimizing bond lengths, angles, and torsional angles as well as removing any steric clashes or bad contacts that may be present. Overall, GFN2-xTB with GBSA solvation is a computationally cost-effective model to obtain the electronic descriptors for different atom types in nucleic acid systems.

Electronic descriptors that we calculated using GFN2-xTB for individual atoms are

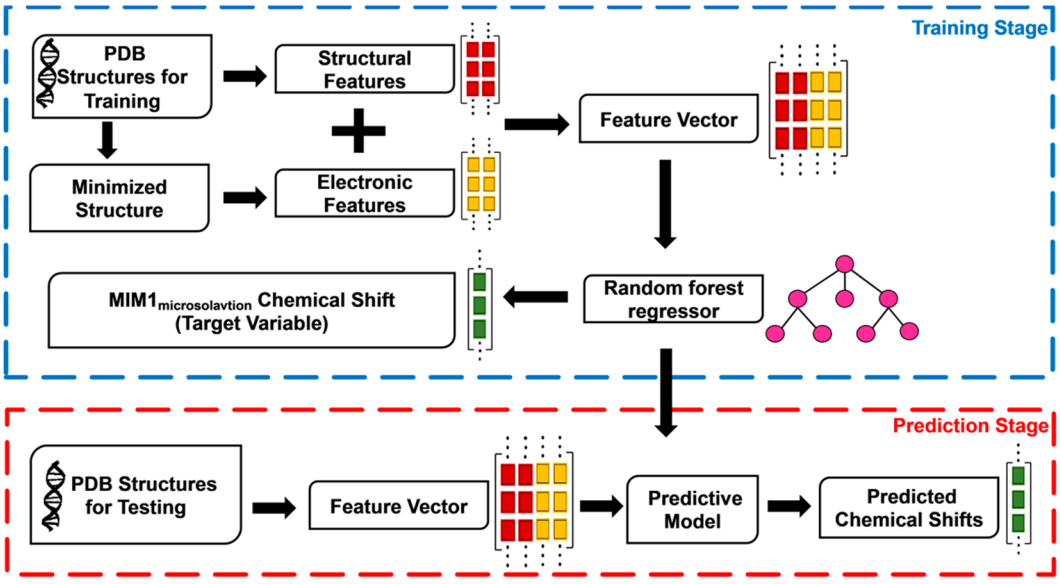
(1) Atomic partial charges (q): The atomic partial charges are taken from a Mulliken population analysis in GFN2-xTB and are solved self-consistently.[58]

(2) Covalent coordination number (CovCN): This term is the element-specific parameter obtained from diagonal elements of the *Extended Hückel Type* Hamiltonian (EHT) matrix and is a crucial ingredient to describe covalent bonds in tight-binding methods. Covalent coordination number employs electronegativity and is obtained by approximately matching Wiberg bond orders that describe the electron density between pairs of atoms in a molecule.[60]

(3) Born radius: The Born radius is a theoretical parameter used in the generalized Born (GB) solvation model to account for the solute−solvent electrostatic interactions. In the GB model, the Born radius is used to calculate the

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

Scheme 1. Schematics of MIM-ML Model[a]



[a]The input to RF regressor includes a set of structural and electronic features. Illustration of the training and prediction stages in MIM-ML architecture.

solvation energy by considering the effective volume of the solute and the dielectric properties of the solvent.[61]

(4) H-bonding parameter: This parameter is used in the noncovalent interaction term in the xTB Hamiltonian, which measures the strength of hydrogen bonds between atoms and solvation environment.[60]

2.3. Random Forest Architecture. Random forest (RF) is a powerful ensemble learning algorithm that combines the predictions of multiple decision trees to generate more accurate predictions.[51] The RF algorithm involves the following steps: First, a subset of features is randomly selected from the data set. Then, a decision tree is built by using the selected features. This process is repeated multiple times to build a collection of decision trees. Finally, the predictions of all of the decision trees are aggregated to make the final prediction. The benefit of using the RF model is that it can handle high-dimensional data sets with many different features and is less prone to overfitting compared to other machine learning models.[15,24]

In the context of chemical shift prediction, the proposed regression model takes the form below.

$$\delta_i = f(\vec{x}_i)$$

Where δ_i is the estimated chemical shift for the i^{th} nucleus, which is calculated from isotropic chemical shift subtracted from the corresponding standard reference. f is a nonlinear regression function and $\vec{x}_i$ is a feature vector whose components encode the variables of the regression model. These variables correspond to computable properties in each nucleus' environment, including atomic charges, solvation effects, structural parameters, etc. Essentially, these properties are the features used in the RF algorithm to make the chemical shift prediction. Scheme 1 shows the general scheme for the MIM-ML model.

During the training stage, we feed an input feature vector consisting of both structural and electronic descriptors into a random forest regressor from scikit-learn. This model is used to predict the MIM-calculated chemical shifts in a microsolvation solvation environment. Notably, we utilize the structural information along with xTB-derived features in the GBSA solvation environment to predict QM-derived chemical shift values. As mentioned earlier, MIM chemical shift values are calculated using SMD implicit solvation with a few explicit solvent molecules near the amino and imino protons. Essentially, by using the combination of structural and electronic descriptors, we are trying to predict highly accurate MIM chemical shifts for both ^{1}H and ^{13}C nuclei in a complex solvation environment. Once the regressor is trained, the model is used to predict the chemical shifts for a new set of PDB structures, which were not used in the training phase. The performance of this MIM-ML model is evaluated by using the mean absolute deviation (MAD) of the predicted ML chemical shift with respect to experimental chemical shifts of the test systems.

To understand the relative importance of each input feature in determining the target variable, we employed the "feature_importances" attribute of the training model from scikit-learn. Feature importance is a measure of how much each feature contributes to the accuracy of the random forest model. The importance of each feature is calculated based on the decrease in the impurity of the decision tree nodes in which the feature is weighted by the probability of reaching that node.

Random Forest regression models have previously been shown to provide accurate predictions of chemical properties for a variety of molecules and materials.[14,24,62,63] However, it is important to note that the accuracy of RF models can be affected by the quality and size of the training data set, the choice of features used, and the hyperparameters of the model.

3. RESULTS AND DISCUSSION

3.1. Data Extraction. In our previous publication, we used the $\text{MIM}_{\text{microsolvation}}$ (formerly referred to as $\text{MIM}^{\text{constraint}}_{\text{explicit-implicit}}$) method to calculate the NMR chemical shifts of 10 nucleic acids with PDB ID's 1SY8,[64] 1K2K,[65] 1KR8,[66] 2N5P,[67] 6XAH,[68] 2LIB,[69] 1N2W,[70] 2LFX,[71] 7NBK,[72] and 2LAR.[73] The calculated chemical shifts for all ^{1}H and ^{13}C nuclei of these nucleic acids were compared to experimentally available shifts, and the results are depicted in Figure 1.

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

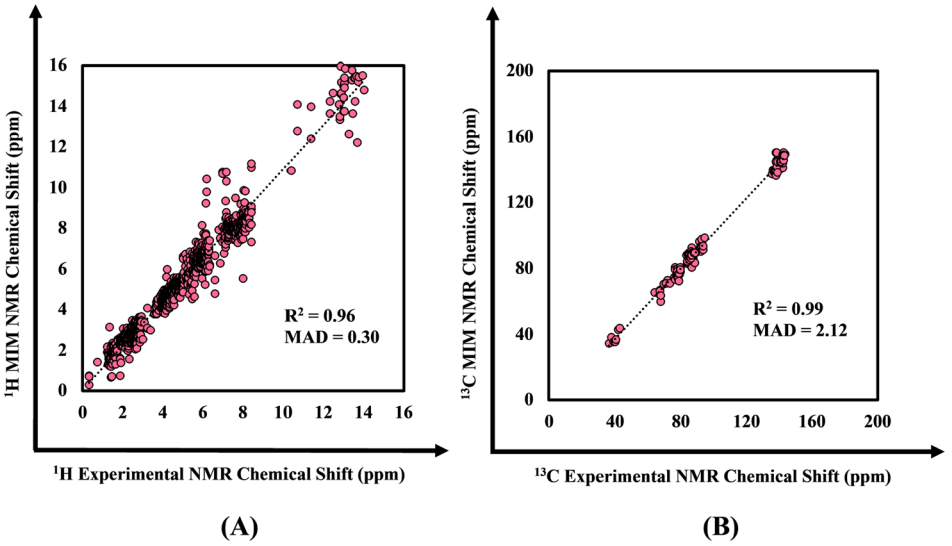


Figure 1. (A) Experimental chemical shifts plotted against $MIM_{microsolvation}$ NMR chemical shifts for 1282 ^{1}H's and (B) experimental chemical shifts plotted against $MIM_{microsolvation}$ NMR chemical shifts for 128 ^{13}C's for PDB ID's 1SY8, 1K2K, 1KR8, 2N5P, 6XAH, 2LIB, 1N2W, 2LFX, 7NBK, and 2LAR from our earlier work on MIM-NMR predictions for nucleic acids.

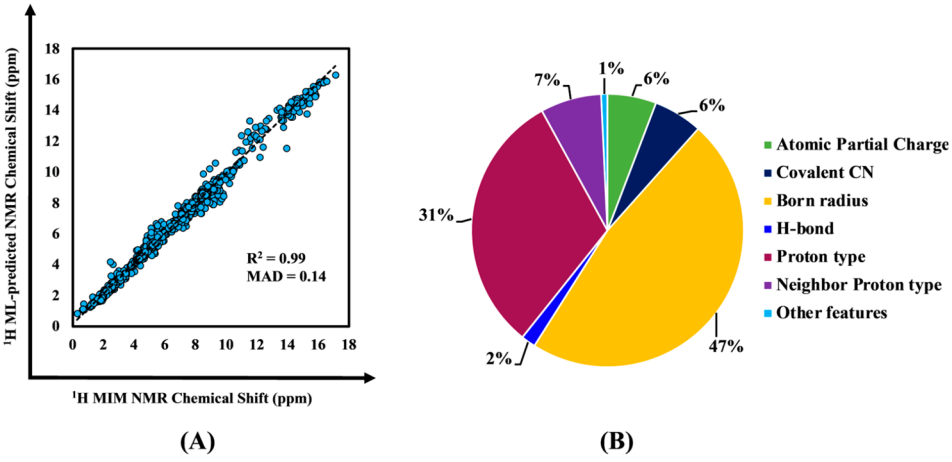


Figure 2. (A) Comparison of $MIM^{constraint}_{explicit-implicit}$ NMR chemical shifts plotted against MIM/ML predicted chemical shifts for 2080 ^{1}H's. (B) Percentage of feature importance score for all of the features used to train the MIM/ML model.

Plotting all 1282 experimental ^{1}H chemical shifts gave an R^2 of 0.96 with a MAD value of 0.30 ppm, whereas a rather small set of 128 available experimental ^{13}C chemical shifts gave an R^2 of 0.99 with a MAD value of 2.12 ppm. The lower MAD values and high correlation between calculated chemical shifts and experimental chemical shifts shows that the chemical shifts obtained from the $MIM_{microsolvation}$ model can be used as the target with a larger data set of 2080 chemical shift values for ^{1}H and 1780 chemical shift values for ^{13}C. In this study, we utilized structure and electronic descriptors as features to train random forest models separately for ^{1}H and ^{13}C predictions.

3.2. Random Forest Regression for ^{1}H Chemical Shifts.

3.2.1. Training on ^{1}H MIM-NMR Chemical Shifts of Nucleic acids. A general scheme for the random forest regressor pipeline for the MIM-NMR chemical shift prediction is shown in Scheme 1. The regressor is trained using atomic feature vectors that combine electronic and structural features, as explained in the method section (*vide supra*), to make predictions. The model is trained using all available MIM calculated data, and the hyperparameters n_estimators, max_depth, min_sample_split, and min_sample_leaf are tuned using a grid search. The best combination of hyperparameters, identified as n_estimators = 500, max_depth = 30, min_sample_split = 2, and min_sample_leaf = 1, is then used for training the RF regression model. The correlation plot, displayed in Figure 2A, shows a high degree of agreement with an R^2 of 0.99 and MAD of 0.14 ppm, between MIM calculated NMR chemical shifts and those predicted by the random forest regressor. To understand the importance of the electronic features, we trained the model with just the structural features accompanied by a hyperparameter grid search. The best performing model including only the structural features gave a MAD of 0.37 ppm between the MIM and the RF chemical shifts with an R^2 of 0.95, which confirms that the electronic features play a critical role in the predictive power of the random forest model.

Next, we compare the predicted MIM-ML chemical shifts directly to available experimental NMR chemical shifts for the training set, giving MAD values of 0.35 ppm for ^{1}H and 2.17 ppm for ^{13}C as depicted in Figure 3A,B. There is only a very small falloff in the performance of the RF model relative to the

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

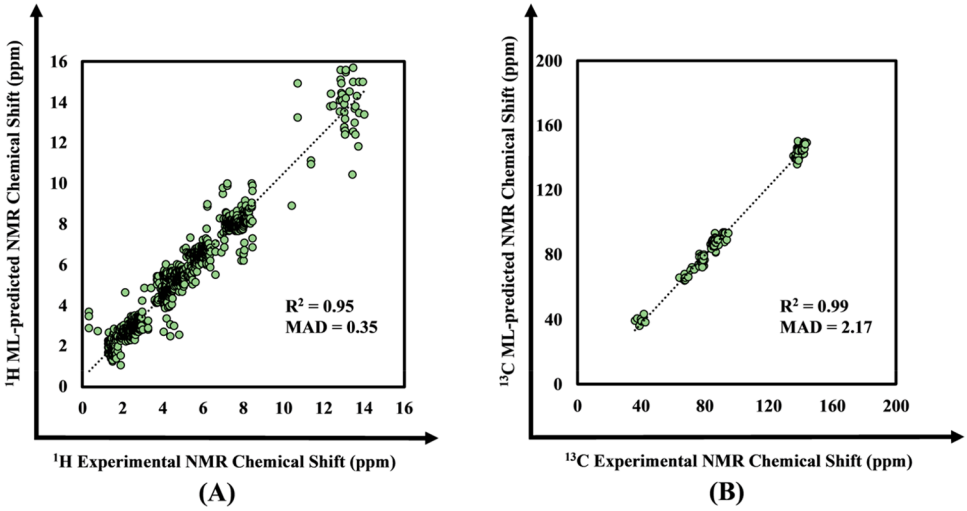


Figure 3. (A) Experimental chemical shifts plotted against MIM-ML predicted NMR chemical shifts for 1282 ^{1}H's and (B) experimental chemical shifts plotted against MIM-ML predicted NMR chemical shifts for 128 ^{13}C's for PDB ID's 1SY8, 1K2K, 1KR8, 2N5P, 6XAH, 2LIB, 1N2W, 2LFX, 7NBK, and 2LAR from our earlier work on MIM-NMR predictions for nucleic acids.

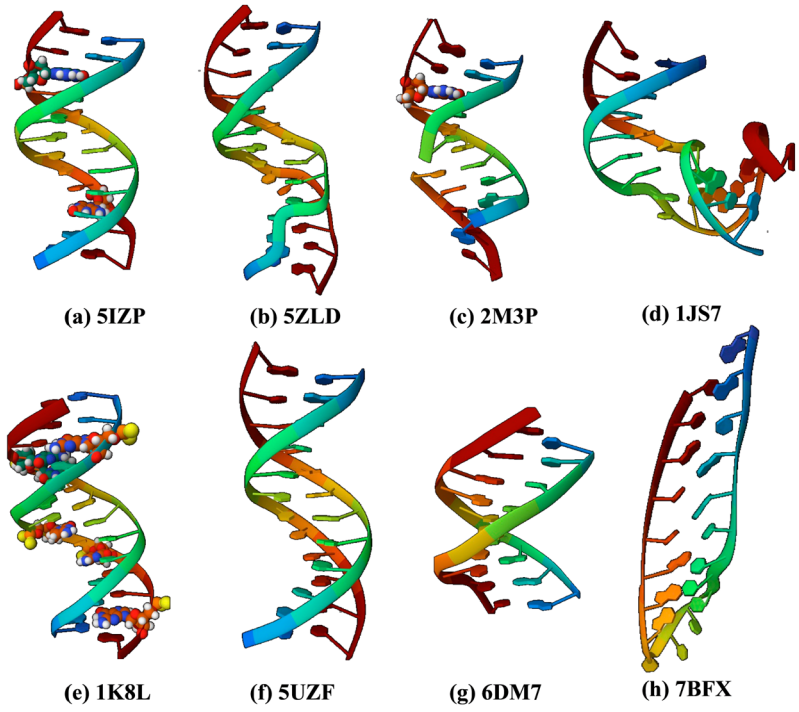


Figure 4. Nucleic acid structures used to compare MIM-ML predicted NMR chemical shifts to experimental ones: (a) 5IZP with 760 atoms, (b) 5ZLD with 886 atoms, (c) 2M3P with 684 atoms, (d) 1JS7 with 915 atoms, (e) 1K8L with 886 atoms, (f) 5UZF with 757 atoms, (g) 6DM7 with 565 atoms, and (h) 7BFX with 627 atoms. Images are created using Mol*.[1]

$MIM_{microsolvation}$ model (MAD of 0.30 ppm for ^{1}H and 2.12 ppm for ^{13}C with respect to experiments). We note in this context that these deviations are comparable to the best chemical shift predictions obtained for nucleic acids. The larger spread of deviations around 12−14 ppm for ^{1}H chemical shifts typically represents imino protons that are involved in different types of hydrogen bonding in the base alignment. Overall, our MIM-ML model accurately predicts experimental chemical shifts without prior knowledge of any experimental NMR chemical shifts in the fitting.

Further, the feature importance analysis of the input features is presented in Figure 2B. The results indicate that the Born radius feature has the highest contribution toward the prediction task, accounting for 47% of the importance score. The proton type feature is the second most significant, with a contribution of 31%. On the other hand, the remaining features such as H-bond information, covalent coordination number, neighboring atoms, residue type, and neighboring residue type have a combined score of 22%. These findings suggest that the Born radius and proton type features are crucial in predicting the target variable.

3.2.2. Predicting Experimental ^{1}H NMR Chemical Shift Using the MIM-ML Regressor Model. Once the model is trained on MIM-NMR chemical shifts, we used it directly to predict NMR chemical shifts for a test set of eight new nucleic acids with

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

Table 1. Structural Information, MAD (in ppm), and R^2 Values between Experimental and MIM/ML Chemical Shifts Predictions for Eight Nucleic Acid Systems Used in This Study

No.	PDB Entry	BMRB	Description	N atom/ residues	Nuclei type	MBIAIL chemical shifts (MAD/R^2)
1	5IZP	30044	DNA dodecamer with 8-oxoguanine at 10th position	760/24	^{1}H	0.26/0.97
2	5ZLD	36174	DNA duplex in *Homo sapiens*	886/28	^{1}H	0.33/0.97
3	2M3P	18973	DNA containing a cluster of 8-oxo- guanine and abasic site lesion	684/22	^{1}H	0.41/0.95
4	1TS7	5134	dAAUAA DNA bulge	915/29	^{1}H	0.34/0.97
5	IK8L	5716	Dithioate backbone modified duplex aptamneers	886/28	^{1}H	0.25/0.98
6	SUZF	30254	DNA duplexes containing N1- methylated adenine (m1 A) lesion	757/24	^{1}H	0.38/0.97
					^{13}C	2.19/0.99
7	6DM7	30473	Sp1 transcription factor duplex for *Homo sapiens*	565/18	^{1}H	0.38/0.96
					^{13}C	2.14/0.99
8	7BFX	34588	Deoxyxylose nucleic acid hairpin	627/20	^{1}H	0.34/0.95
					^{13}C	3.24/0.98

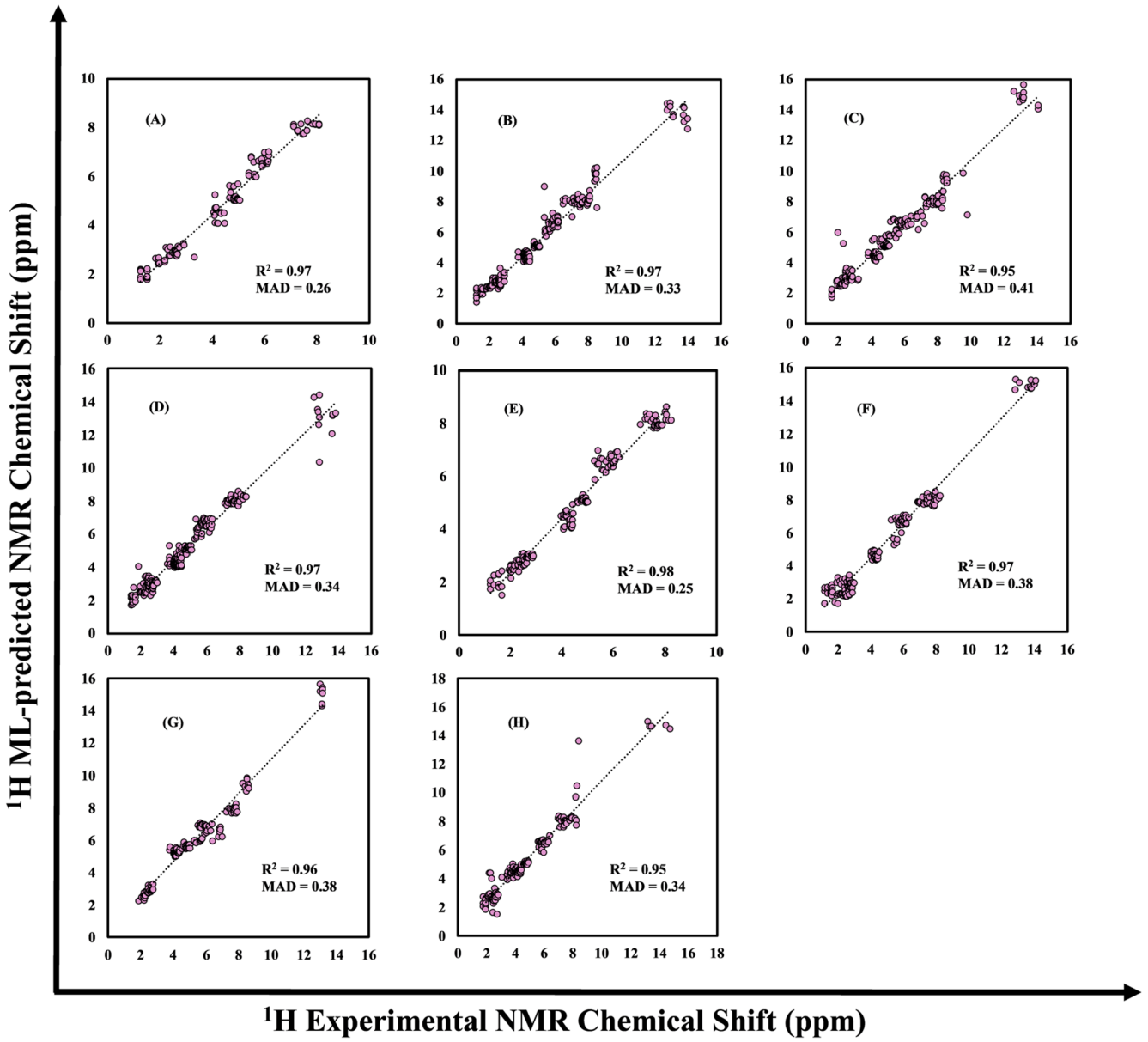


Figure 5. Comparison of experimental ^{1}H NMR chemical shifts of (a) 5IZP, (b) 5ZLD, (c) 2M3P, (d) 1JS7, (e) 1K8L, (f) 5UZF, (g) 6DM7, and (h) 7BFX with MIM-ML predicted chemical shifts.

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

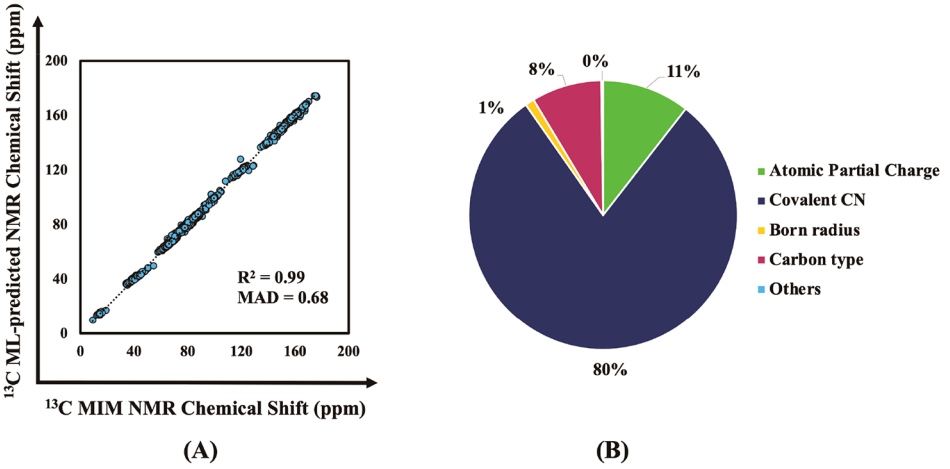


Figure 6. (A) Comparison of $MIM^{constraint}_{explicit-implicit}$ NMR chemical shifts plotted against MIM-ML predicted chemical shifts for 1780 ^{13}C's. (B) Percentage of feature importance score for all the features used to train the MIM/ML model.

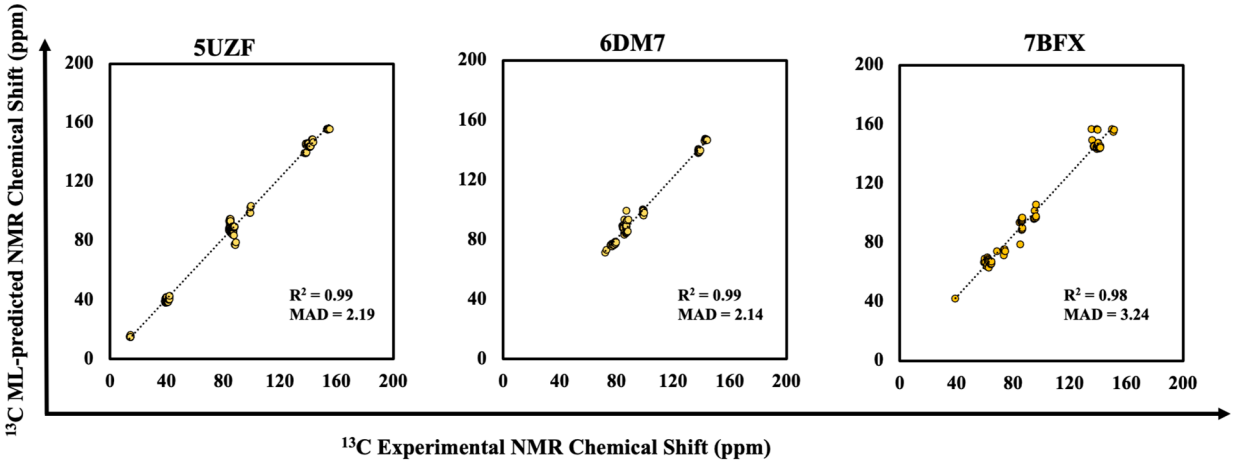


Figure 7. Comparison of experimental ^{13}C NMR chemical shifts of 5UZF, 6DM7, and 7BFX nucleic acids with MIM-ML predicted chemical shifts.

sizes ranging from 600 to 900 atoms (Figure 4). A short description about these nucleic acids with PDB ID's 5IZP,[74] 5ZLD,[75] 2M3P,[76] 1JS7,[77] 1K8L,[78] 5UZF,[79] 6DM7,[2] and 7BFX[80] can be found in Table 1.

Among the nucleic acid structures studied, 2M3P with a missing residue and a modified nonstandard residue showed slightly higher MAD value (0.41 ppm) with respect to experiments. However, other nucleic acids with nonstandard residues including 5IZP and 1K8L showed an excellent agreement with the experiments with a lower MAD value of 0.26 and 0.25 ppm, respectively. As shown in Figure 5 and Table 1, the MIM-ML predicted ^{1}H chemical shifts of the eight nucleic acids gave an average MAD value of 0.34 ppm with an average R^2 of 0.96. Importantly, as shown in Figure 3A, when the predicted chemical shifts from the training model are compared to the corresponding experiments, the MAD value was calculated to be 0.35 ppm with an R^2 of 0.95, which is very similar to the observed MAD values for the new systems in the test set. This shows that our MIM-ML model which employs both structural and electronic features is excellent for prediction on new data for ^{1}H chemical shifts.

3.3. Random Forest Regression for ^{13}C. *3.3.1. Training on ^{13}C MIM-NMR Chemical Shifts of Nucleic acids.* As in the case of ^{1}H training, we used both structural and electronic features described in the Methods section for the training of ^{13}C MIM calculated chemical shifts. Except for the hydrogen bonding parameter, we used all of the other features used in the ^{1}H training model. The model is trained using MIM calculated data and hyperparameters tuned using a grid search. The best combination of hyperparameters, identified as n_estimators = 500, max_depth = 30, min_sample_split = 2, and min_sample_leaf = 1, is used for training the RF regression model. The RF model is trained on 1780 MIM-NMR calculated chemical shifts, and the comparison of MIM chemical shifts with MIM-ML predictions showed a MAD of 0.68 ppm and an R^2 value of 0.99 (Figure 6A). To investigate the importance of electronic descriptors, we used a prediction model just with the structural features and obtained an MAD value of 1.62 ppm for MIM values with respect to MIM-ML values. This analysis validates the importance of electronic features in chemical shift predictions for ^{13}C nuclei using a RF regressor model.

Further, a feature importance analysis is performed on the input features for the MIM-ML model as shown in Figure 6B. It is interesting to note that >90% of the feature importance score is covered by electronic features, i.e., the covalent coordination number covers 80% of the feature importance followed by the atomic partial charge with 11% and finally a small slice of 1% by

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

the Born radius. The only structural feature involved in the prediction architecture is the carbon atom type feature with a contribution of 8%. Other structural features including neighboring atom type, residue type, and neighboring residues seems to have no contribution in the RF predictor for ^{13}C. It is evident from the feature importance analysis that electronic features are important for the prediction of chemical shifts.

3.3.2. Predicting Experimental ^{13}C NMR Chemical Shifts Using MIM-ML Random Forest Regressor Model. After completing the training of our MIM-ML model, we evaluated its performance on a test set consisting of three nucleic acids for which ^{13}C chemical shifts are available. These nucleic acids are identified by their PDB IDs: 5UZF, 6DM7, and 7BFX. Table 1 and Figure 7 provide a detailed description of these nucleic acids including their predicted ^{13}C chemical shifts. Our MIM-ML model accurately predicted the ^{13}C chemical shifts of these nucleic acids with an average MAD value of 2.52 ppm and an average R^2 value of 0.99. It is noteworthy that the 7BFX DNA structure, with a hairpin geometry, exhibited a slightly larger deviation of 3.24 ppm for the MAD value, while the 5UZF and 6DM7 structures showed relatively low deviations, with MAD values of 2.19 and 2.14 ppm, respectively.

Furthermore, in Figure 3B, we compare the predicted chemical shifts from the training model to experimental values, resulting in a MAD value of 2.17 ppm and an R^2 value of 0.99. These values are similar to the observed MAD values for the 5UZF and 6DM7 structures in the test set, but slightly smaller than that for the 7BFX structure with hairpin geometry. Nonetheless, our results demonstrate the ability of our MIM-ML model to accurately predict ^{13}C chemical shifts for nucleic acids of varying sizes and geometries.

4. CONCLUSIONS

In this study, the MIM-ML model, combining machine learning and the MIM fragmentation methodology, has emerged as a reliable and accurate tool for predicting NMR chemical shifts of nucleic acids. Our ML model is trained solely on reproducing previously computed chemical shifts obtained through a Molecules-in-Molecules (MIM) fragment-based density functional theory (DFT) protocol, incorporating microsolvation effects.

By incorporating both structural and electronic descriptors of the local atomic environments, derived from an inexpensive low-level semiempirical calculation (GFN2-xTB), our ML model demonstrates excellent performance in predicting chemical shifts, despite having been trained on a relatively small data set of DFT chemical shifts for 10 nucleic acids. Furthermore, the model's predictive accuracy is confirmed by its comparison with experimental data for eight additional nucleic acids of varying sizes (ranging from 600 to 900 atoms) with a mean absolute deviation (MAD) of approximately 0.30 ppm for ^{1}H and 2−3 ppm for ^{13}C. Further, our MIM-ML model offers several advantages, including the ability to be applied to nonstandard structures and predict both sugar−phosphate backbone and nucleotide chain shifts for ^{1}H and ^{13}C.

By leveraging machine learning, the model eliminates the need for expensive QM calculations while delivering QM and experimental quality results. This innovative combination of fragment-based QM calculations and machine learning represents a significant step forward in NMR prediction. The success of this approach opens possibilities for studying the properties of other large biomolecules and accurately predicting their chemical shifts or other important electronic properties. It provides a practical application of machine learning by overcoming the lack of quality experimental chemical shift data for nucleic acids and highlights the potential of theory-to-theory learning and predicting experiments when reliable experimental data is scarce. In summary, our approach has the potential to greatly enhance our understanding of the properties and behavior of complex biomolecules.

■ ASSOCIATED CONTENT

Data Availability Statement

Additional information and sample MIM-ML code may be accessed at https://github.com/schandy2211/MIM_ML.git.

■ AUTHOR INFORMATION

Corresponding Authors

Sruthy K. Chandy − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0002-1061-647X; Email: schandy@iu.edu

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00563

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from the National Science Foundation Grant CHE-2102583 at Indiana University. We thank Sarah Maier for fruitful discussions. The Big Red 3 supercomputing facility at Indiana University was used for most of the calculations in this study.

■ REFERENCES

(1) Sehnal, D.; Bittrich, S.; Deshpande, M.; Svobodová, R.; Berka, K.; Bazgier, V.; Velankar, S.; Burley, S. K.; Koča, J.; Rose, A. S. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **2021**, *49* (W1), W431−W437.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(3) Xu, X. P.; Case, D. A. Automated prediction of 15N, 13Cα, 13Cβ and 13C′ chemical shifts in proteins using a density functional database. *J. Biomol. NMR* **2001**, *21*, 321.

(4) Neal, S.; Nip, A.; Zhang, H.; Wishart, D. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR* **2003**, *26*, 215.

(5) Meiler, J. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **2003**, *26* (1), 25−37.

(6) Shen, Y.; Bax, A. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR* **2007**, *38*, 289.

(7) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *J. Am. Chem. Soc.* **2009**, *131* (39), 13894−13895.

(8) Cheung, M. S.; Maguire, M. L.; Stevens, T. J.; Broadhurst, R. W. DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *J. Magn. Reson.* **2010**, *202*, 223.

(9) Martin, O. A.; Vila, J. A.; Scheraga, H. A. CheShift-2: Graphic validation of protein structures. *Bioinformatics* **2012**, *28*, 1538.
(10) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13.
(11) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **2011**, *50*, 43.
(12) Li, D.-W.; Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *Journal of Biomolecular NMR* **2012**, *54* (3), 257−265.
(13) Yang, Z.; Chakraborty, M.; White, A. D. Predicting chemical shifts with graph neural networks. *Chemical Science* **2021**, *12* (32), 10802−10809.
(14) Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T. Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data. *Chemical Science* **2020**, *11* (12), 3180−3191.
(15) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem.* **2020**, *6* (7), 1527−1542.
(16) Victora, A.; Möller, H. M.; Exner, T. E. Accurate ab initio prediction of NMR chemical shifts of nucleic acids and nucleic acids/protein complexes. *Nucleic Acids Res.* **2014**, *42*, No. e173.
(17) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. *journal of physical chemistry. B* **2013**, *117* (7), 2045−2052.
(18) Altona, C.; Faber, D. H.; Hoekzema, A. J. A. W. Double-helical DNA 1H chemical shifts: an accurate and balanced predictive empirical scheme. *Magn. Reson. Chem.* **2000**, *38* (2), 95−107.
(19) Barton, S.; Heng, X.; Johnson, B. A.; Summers, M. F. Database proton NMR chemical shifts for RNA signal assignment and validation. *Journal of Biomolecular NMR* **2013**, *55* (1), 33−46.
(20) Lam, S. L. DSHIFT: a web server for predicting DNA chemical shifts. *Nucleic Acids Res.* **2007**, *35* (suppl_2), W713−W717.
(21) Dejaegere, A.; Bryce, R. A.; Case, D. A. An Empirical Analysis of Proton Chemical Shifts in Nucleic Acids. In *Modeling NMR Chemical Shifts*; ACS Symposium Series, Vol. *732*; American Chemical Society, 1999; pp 194−206.
(22) Cromsigt, J. A. M. T. C.; Hilbers, C. W.; Wijmenga, S. S. Prediction of proton chemical shifts in RNA − Their use in structure refinement and validation. *Journal of Biomolecular NMR* **2001**, *21* (1), 11−29.
(23) Sahakyan, A. B.; Vendruscolo, M. Analysis of the Contributions of Ring Current and Electric Field Effects to the Chemical Shifts of RNA Bases. *J. Phys. Chem. B* **2013**, *117* (7), 1989−1998.
(24) Frank, A. T.; Bae, S.-H.; Stelzer, A. C. Prediction of RNA 1H and 13C Chemical Shifts: A Structure Based Approach. *J. Phys. Chem. B* **2013**, *117* (43), 13497−13506.
(25) Brown, J. D.; Summers, M. F.; Johnson, B. A. Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *J. Biomol NMR* **2015**, *63* (1), 39−52.
(26) Beran, G. J. O. Calculating nuclear magnetic resonance chemical shifts from density functional theory: A primer. *eMagRes.* **2019**, *8* (3), 215.
(27) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational Prediction of 1H and 13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012**, *112* (3), 1839−1862.
(28) Charpentier, T. The PAW/GIPAW approach for computing NMR parameters: a new dimension added to NMR study of solids. *Solid State Nucl. Magn. Reson.* **2011**, *40*, 1.
(29) Willoughby, P. H.; Jansma, M. J.; Hoye, T. R. A guide to small-molecule structure assignment through computation of (1H and 13C) NMR chemical shifts. *Nat. Protoc.* **2014**, *9*, 643.
(30) Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G. Structure validation of natural products by quantum-mechanical GIAO calculations of 13C NMR chemical shifts. *Chemistry* **2002**, *8*, 3233.
(31) Kutateladze, A. G.; Reddy, D. S. High-Throughput in Silico Structure Validation and Revision of Halogenated Natural Products Is Enabled by Parametric Corrections to DFT-Computed 13C NMR Chemical Shifts and Spin-Spin Coupling Constants. *J. Org. Chem.* **2017**, *82*, 3368.
(32) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309.
(33) Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P. IMPRESSION − prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508.
(34) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501.
(35) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.
(36) Salager, E.; Day, G. M.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L. Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution 1H Solid-State NMR Spectroscopy. *J. Am. Chem. Soc.* **2010**, *132*, 2564.
(37) Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T. Multi-resolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *J. Phys. Chem. Lett.* **2019**, *10*, 4558.
(38) Das, S.; Edison, A. S.; Merz, K. M., Jr. Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal. Chem.* **2020**, *92* (15), 10412−10419.
(39) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9* (1), 4501.
(40) Exner, T. E.; Frank, A.; Onila, I.; Möller, H. M. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 3. Conformational Sampling and Explicit Solvents Model. *J. Chem. Theory Comput.* **2012**, *8*, 4818.
(41) Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S. Ab initio NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment. *Chem. Phys. Lett.* **2007**, *445*, 331.
(42) Gao, Q.; Yokojima, S.; Fedorov, D. G.; Kitaura, K.; Sakurai, M.; Nakamura, S. Fragment-Molecular-Orbital-Method-Based ab Initio NMR Chemical-Shift Calculations for Large Molecular Systems. *J. Chem. Theory Comput.* **2010**, *6*, 1428.
(43) Hartman, J.; Monaco, S.; Schatschneider, B.; Beran, G. Fragment-based 13-C nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to planewave methods. *J. Chem. Phys.* **2015**, *143* (10), 102809.
(44) Lee, A. M.; Bettens, R. P. A. First Principles NMR Calculations by Fragmentation. *J. Phys. Chem. A* **2007**, *111*, 5111.
(45) Zhao, D.; Song, R.; Li, W.; Ma, J.; Dong, H.; Li, S. Accurate Prediction of NMR Chemical Shifts in Macromolecular and Condensed-Phase Systems with the Generalized Energy-Based Fragmentation Method. *J. Chem. Theory Comput.* **2017**, *13* (11), 5231−5239.
(46) He, X.; Wang, B.; Merz, K. M. Protein NMR Chemical Shift Calculations Based on the Automated Fragmentation QM/MM Approach. *J. Phys. Chem. B* **2009**, *113*, 10380.
(47) Swails, J.; Zhu, T.; He, X.; Case, D. A. AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J. Biomol. NMR* **2015**, *63* (2), 125−139.
(48) Chandy, S. K.; Thapa, B.; Raghavachari, K. Accurate and cost-effective NMR chemical shift predictions for proteins using a molecules-in-molecules fragmentation-based method. *Phys. Chem. Chem. Phys.* **2020**, *22* (47), 27781−27799.
(49) Chandy, S. K.; Raghavachari, K. Accurate and Cost-Effective NMR Chemical Shift Predictions for Nucleic Acids Using a Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Theory Comput.* **2023**, *19* (2), 544−561.

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642

(50) Jose, K. V. J.; Raghavachari, K. Fragment-Based Approach for the Evaluation of NMR Chemical Shifts for Large Biomolecules Incorporating the Effects of the Solvent Environment. *J. Chem. Theory Comput.* **2017**, *13* (3), 1147−1158.
(51) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5−32.
(52) *Gaussian 16*, Rev. C.01; Gaussian, Inc.: Wallingford, CT, 2016.
(53) Gerber, P. R.; Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9* (3), 251−268.
(54) Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *J. Chem. Theory Comput.* **2014**, *10* (10), 4515−4534.
(55) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161.
(56) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design* **1990**, *4*, 1−103.
(57) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. *J. Chem. Theory Comput.* **2021**, *17* (2), 826−840.
(58) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652−1671.
(59) Cramer, C. J.; Truhlar, D. G. General parameterized SCF model for free energies of solvation in aqueous solution. *J. Am. Chem. Soc.* **1991**, *113* (22), 8305−8311.
(60) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **2021**, *11* (2), No. e1493.
(61) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104* (15), 3712−3720.
(62) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine learning for predicting product distributions in catalytic regioselective reactions. *Phys. Chem. Chem. Phys.* **2018**, *20* (27), 18311−18318.
(63) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (3), 1339−1345.
(64) Barthwal, R.; Awasthi, P.; Monica; Kaur, M.; Sharma, U.; Srivastava, N.; Barthwal, S. K.; Govil, G. Structure of DNA sequence d-TGATCA by two-dimensional nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J. Struct. Biol.* **2004**, *148* (1), 34−50.
(65) Lam, S. L.; Ip, L. N. Low Temperature Solution Structures and Base Pair Stacking of Double Helical d(CGTACG)2. *J. Biomol. Struct. Dyn.* **2002**, *19* (5), 907−917.
(66) Padrta, P.; Stefl, R.; Králík, L.; Zídek, L.; Sklenár, V. Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J. Biomol NMR* **2002**, *24* (1), 1−14.
(67) Spring-Connell, A. M.; Evich, M. G.; Debelak, H.; Seela, F.; Germann, M. W. Using NMR and molecular dynamics to link structure and dynamics effects of the universal base 8-aza, 7-deaza, N8 linked adenosine analog. *Nucleic Acids Res.* **2016**, *44* (18), 8576−8587.
(68) Kellum, A. H., Jr.; Qiu, D. Y.; Voehler, M. W.; Martin, W.; Gates, K. S.; Stone, M. P. Structure of a Stable Interstrand DNA Cross-Link Involving a β-N-Glycosyl Linkage Between an N6-dA Amino Group and an Abasic Site. *Biochemistry* **2021**, *60* (1), 41−52.
(69) Johnson, C. N.; Spring, A. M.; Desai, S.; Cunningham, R. P.; Germann, M. W. DNA Sequence Context Conceals α-Anomeric Lesions. *J. Mol. Biol.* **2012**, *416* (3), 425−437.
(70) Thiviyanathan, V.; Somasunderam, A.; Hazra, T. K.; Mitra, S.; Gorenstein, D. G. Solution Structure of a DNA Duplex Containing 8-Hydroxy-2′-Deoxyguanosine Opposite Deoxyguanosine. *J. Mol. Biol.* **2003**, *325* (3), 433−442.
(71) Huang, Y. J.; Brock, K. P.; Ishida, Y.; Swapna, G. V. T.; Inouye, M.; Marks, D. S.; Sander, C.; Montelione, G. T. Combining Evolutionary Covariance and NMR Data for Protein Structure Determination. *Methods Enzymol.* **2019**, *614*, 363−392.
(72) Cabrero, C.; Martín-Pintado, N.; Mazzini, S.; Gargallo, R.; Eritja, R.; Aviñó, A.; González, C. Structural Effects of Incorporation of 2'-Deoxy-2'2'-Difluorodeoxycytidine (Gemcitabine) in A- and B-Form Duplexes. *Chem. Eur. J.* **2021**, *27* (26), 7351−7355.
(73) Johnson, C. N.; Spring, A. M.; Sergueev, D.; Shaw, B. R.; Germann, M. W. Structural basis of the RNase H1 activity on stereo regular borano phosphonate DNA/RNA hybrids. *Biochemistry* **2011**, *50* (19), 3903−3912.
(74) Hoppins, J. J.; Gruber, D. R.; Miears, H. L.; Kiryutin, A. S.; Kasymov, R. D.; Petrova, D. V.; Endutkin, A. V.; Popov, A. V.; Yurkovskaya, A. V.; Fedechkin, S. O.; et al. 8-Oxoguanine Affects DNA Backbone Conformation in the EcoRI Recognition Site and Inhibits Its Cleavage by the Enzyme. *PLoS One* **2016**, *11* (10), No. e0164424.
(75) Ganguly, S.; Murugan, N. A.; Ghosh, D.; Narayanaswamy, N.; Govindaraju, T.; Basu, G. DNA Minor Groove-Induced cis−trans Isomerization of a Near-Infrared Fluorescent Probe. *Biochemistry* **2021**, *60* (26), 2084−2097.
(76) Zálešák, J.; Lourdin, M.; Krejči, L.; Constant, J.-F.; Jourdan, M. Structure and Dynamics of DNA Duplexes Containing a Cluster of Mutagenic 8-Oxoguanine and Abasic Site Lesions. *J. Mol. Biol.* **2014**, *426* (7), 1524−1538.
(77) Gollmick, F. A.; Lorenz, M.; Dornberger, U.; von Langen, J.; Diekmann, S.; Fritzsche, H. Solution structure of dAATAA and dAAUAA DNA bulges. *Nucleic Acids Res.* **2002**, *30* (12), 2669−2677.
(78) Volk, D. E.; Yang, X.; Fennewald, S. M.; King, D. J.; Bassett, S. E.; Venkitachalam, S.; Herzog, N.; Luxon, B. A.; Gorenstein, D. G. Solution structure and design of dithiophosphate backbone aptamers targeting transcription factor NF-κB. *Bioorganic Chemistry* **2002**, *30* (6), 396−419.
(79) Sathyamoorthy, B.; Shi, H.; Zhou, H.; Xue, Y.; Rangadurai, A.; Merriman, D. K.; Al-Hashimi, H. M. Insights into Watson−Crick/Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m1A. *Nucleic Acids Res.* **2017**, *45* (9), 5586−5601.
(80) Mattelaer, C.-A.; Maiti, M.; Smets, L.; Maiti, M.; Schepers, G.; Mattelaer, H.-P.; Rosemeyer, H.; Herdewijn, P.; Lescrinier, E. Stable Hairpin Structures Formed by Xylose-Based Nucleic Acids. *ChemBioChem.* **2021**, *22* (9), 1638−1645.

https://doi.org/10.1021/acs.jctc.3c00563
J. Chem. Theory Comput. 2023, 19, 6632−6642