JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

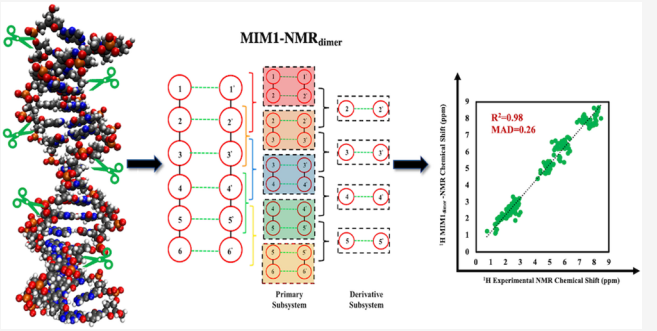
Accurate and Cost-Effective NMR Chemical Shift Predictions for Nucleic Acids Using a Molecules-in-Molecules Fragmentation-Based Method

Sruthy K. Chandy and Krishnan Raghavachari*

Cite This: J. Chem. Theory Comput. 2023, 19, 544−561 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We have developed, implemented, and assessed an efficient protocol for the prediction of NMR chemical shifts of large nucleic acids using our molecules-in-molecules (MIM) fragment-based quantum chemical approach. To assess the performance of our approach, MIM-NMR calculations are calibrated on a test set of three nucleic acids, where the structure is derived from solution-phase NMR studies. For DNA systems with multiple conformers, the one-layer MIM method with trimer fragments ($MIM1_{trimer}$) is benchmarked to get the lowest energy structure, with an average error of only 0.80 kcal/mol with respect to unfragmented full molecule calculations. The $MIMI\text{-}NMR_{dimer}$ calibration with respect to unfragmented full molecule calculations shows a mean absolute deviation (MAD) of 0.06 and 0.11 ppm, respectively, for ^{1}H and ^{13}C nuclei, but the performance with respect to experimental NMR chemical shifts is comparable to the more expensive MIM1-NMR and MIM2-NMR methods with trimer subsystems. To compare with the experimental chemical shifts, a standard protocol is derived using DNA systems with Protein Data Bank (PDB) IDs 1SY8, 1K2K, and 1KR8. The effect of structural minimizations is employed using a hybrid mechanics/semiempirical approach and used for computations in solution with implicit and explicit−implicit solvation models in our $MIM1\text{-}NMR_{dimer}$ methodology. To demonstrate the applicability of our protocol, we tested it on seven nucleic acids, including structures with nonstandard residues, heteroatom substitutions (F and B atoms), and side chain mutations with a size ranging from ∼300 to 1100 atoms. The major improvement for predicted $MIM1\text{-}NMR_{dimer}$ calculations is obtained from structural minimizations and implicit solvation effects. A significant improvement with the explicit−implicit solvation model is observed only for two smaller nucleic acid systems (1KR8 and 7NBK), where the expensive first solvation shell is replaced by the microsolvation model, in which a single water molecule is added for each solvent-exposed amino and imino protons, along with the implicit solvation. Overall, our target accuracy of ∼0.2−0.3 ppm for ^{1}H and ∼2−3 ppm for ^{13}C has been achieved for large nucleic acids. The proposed MIM-NMR approach is accurate and cost-effective (linear scaling with system size), and it can aid in the structural assignments of a wide range of complex biomolecules.



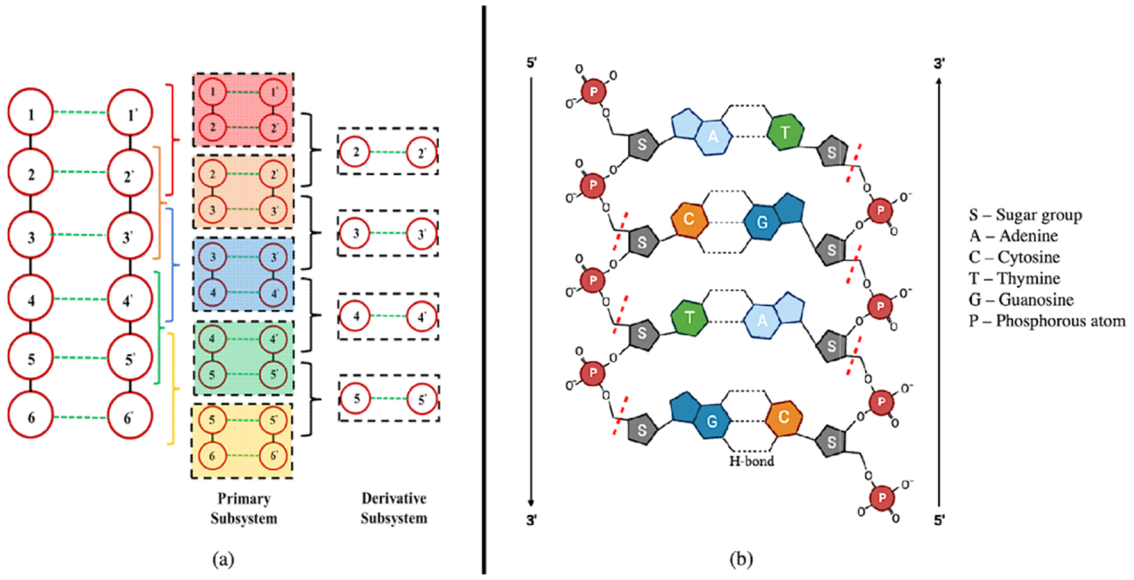
1. INTRODUCTION

Nuclear magnetic resonance (NMR) chemical shifts are heavily employed for the structural elucidation of proteins and nucleic acids. However, chemical shifts are influenced by, and are sensitive to, several important factors, such as the local structures surrounding the nuclei of interest, conformational flexibility of the biomolecule, and the solvation environment. Due to the complex nature of such interactions, predicting accurate chemical shifts is crucial for peak assignments and subsequent structure validation for biological macromolecular systems.[1−4] In particular, a large number of NMR experiments have been conducted on proteins to understand their underlying structural parameters that have subsequently been used to fit empirical formulae as well as to construct various protein prediction models. In contrast, due to the scarcity of high-quality three-dimensional (3D) structures, NMR chemical shift predictions for nucleic acids have lagged behind those for proteins. For nucleic acids, the employment of NMR chemical shifts can be even more advantageous due to the complications in NMR experiments and subsequent assignments. However, due to the high degree of flexibility in nucleic acids, particularly in RNA, it is highly challenging for NMR experiments to generate a complete description of interproton nuclear overhauser effect (NOE)-derived distance restraints that are needed for their structural characterization.[5,6] While

Received: September 27, 2022
Published: January 11, 2023

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

Scheme 1. Cartoon Representing the MIM-NMR Method for ds-DNA Type Nucleic Acids[a]



[a]Circles represent nucleotides that are connected by bonds (bold black lines), dotted bonds in green color represent the intramolecular hydrogen-bonding interactions. Primary subsystems are color coded as open brackets, which represent each dimer unit. Overlapping primary fragments are shown as derivative subsystems. (b) DNA structure with red dotted line through the 4′C−5′C bonds that are cut to make the monomers.

empirical approaches, such as SHIFTS,[7] NUCHEMICS,[8] PPM,[9] and RAMSEY,[10] are used in the field of nucleic acid chemical shift predictions, they have significant limitations. Most of these methods rely on experimental data from a small number of high-quality structures and they use empirical or semiempirical equations to account for the effects of non-neighboring residues. These empirical methods are not well adapted to dealing with noncanonical structures, and are insensitive to structural changes in nucleic acids, leading to inaccurate predictions.[11] Furthermore, even for the experimentally assigned nucleic acids, chemical shifts of non-proton nuclei still limit the reliability of empirical methods targeting these nuclei.[5]

Most currently available quantum mechanics (QM)-based techniques have a key drawback in that they become too expensive as the system size grows. Because full quantum chemical computations for large proteins are currently not possible, a majority of previous investigations have relied solely on localized truncated structural models to determine the NMR chemical shifts. As an alternative strategy, fragmentation-based hybrid approaches have recently emerged as highly efficient tools to achieve asymptotic linear scaling for QM computations. Larger molecules are divided into smaller pieces, and the wavefunction, energy, and other energy derivatives (*i.e.*, spectroscopic characteristics) of each fragment are computed using quantum mechanical calculations. The results of the fragments are then assembled appropriately to derive the properties of the entire large molecule. Fragmentation strategies rely on the chemical localization of macromolecular systems, presuming that atoms far away from the region of interest have only a minor effect on the local region of a macromolecule.

Several fragmentation methods employing a QM/molecular mechanics (MM) framework,[11,12] and a range of density functional theory (DFT)-based methods like adjustable density matrix assembler (ADMA), fragment molecular orbital (FMO) method, combined fragmentation method (CFM), generalized energy-based fragmentation (GEBF), and systematic molecular fragmentation analysis (SMFA) have been developed by different groups to compute the NMR chemical shifts of various macromolecular systems.[13−20] Almost all of these fragmentation methods are tested and benchmarked on either proteins, peptides, or molecular crystals. Only a few studies are on nucleic acids such as a recent study using electrostatically embedded generalized molecular fractionation with conjugate caps (EE-GMFCC) scheme for the excited-state properties of fluorophore RNA systems. Among the few studies specifically on NMR chemical shift predictions, QM/MM-based AFNMR and DFT-based ADMA methods have been used; however, both these methods showed only modestly successful performance with mean deviations of ∼0.4−0.6 ppm for ^{1}H NMR chemical shifts, nearly twice as large as the ∼0.2−0.3 ppm attained for peptide systems.[5,11,21]

In prior work on proteins, we found that the SHIFTX2 and AF-QM/MM methods failed to predict the chemical shifts for labile protons and nonstandard residues.[22] In contrast, our QM-based molecules-in-molecules (MIM) fragmentation method can overcome most of the limitations in empirical methods or other DFT-based fragmentation methods by employing more efficient fragmentation strategies, effective structural minimizations on the relevant conformers, and by accurately describing the solvation environment. Interestingly, quantum chemical calculations on small test systems like mononucleotides, trinucleotides, and base pairs showed that sugar ring puckering, torsional effects of rings, hydrogen bonding between base pairs, and ring current effects from base stacking have profound influence on chemical shifts.[8,23−26] The presence of such nonbonded interactions in nucleic acid systems demands a reasonably large fragment size for accurate MIM predictions.

In this work, we report our results on NMR chemical shift predictions for a diverse set of nucleic acid systems based on our MIM fragmentation-based method.[27−30] For medium-to-large systems, the MIM method has previously demonstrated

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

excellent performance in a variety of spectroscopic studies, including infrared, Raman, vibrational circular dichroism, and Raman optical activity spectra.[31−33] In addition, our previous work using multilayer MIM helped us to accurately predict the NMR chemical shifts of large proteins.[22,34] Because nuclear shielding is a local property, using high-level QM methods on smaller fragment subsystems makes MIM an accurate and cost-effective method for predicting the NMR chemical shifts of large biomolecules. In this study, we report an efficient MIM-NMR protocol for nucleic acids by carefully calibrating fragmentation size and combining various levels of theory. Furthermore, we also performed MIM energy evaluations on conformers, as well as the effect of structural minimization on the computed NMR spectra. Additionally, we present a cost-effective approach to incorporate solvation effects for the accurate prediction of NMR chemical shifts.

2. METHODS

2.1. Molecules-in-Molecules (MIM) Method. All MIM and MIM-NMR calculations were performed using an external Perl module and the Gaussian16 program suite.[19] The details about the working principles of our MIM fragment-based approach, different fragmentation schemes, and capabilities of our method have been described in previous publications.[30−37] Therefore, only a brief and relevant discussion will be given here. In MIM, initial non-overlapping fragments, called "monomers" in this work, are formed by cutting single bonds between heavy (nonhydrogen) atoms. In the case of nucleic acids, we keep all of the partial double bonds intact along with the rings from the nucleobase, pentose sugar rings, glycosidic bonds, and phosphate groups (preserve phosphodiester bonds). Parent fragments are formed by cutting the 4′C−5′C bond of the deoxyribose sugar from each side of the DNA ladder by keeping the base pairs and phosphate groups intact as shown in Scheme 1. It is important to note that hydrogen-bonded nucleotide pairs are considered as monomers and terminal 4′C−5′C bonds are kept intact. Additionally, for DNA structures with local unpaired nucleobases, the mononucleotide is considered as a monomer along with other paired nucleotides for the MIM protocol. Neighboring monomers are combined to form primary and derivative subsystems (*vide infra*) to capture the interactions between the monomers.

In this study, we have compared the performance of one-layer (MIM1) and two-layer (MIM2) approaches to compute the relevant properties of the molecule. In MIM2, two fragmentation parameters and two levels of theory are used. The primary subsystems formed with a small fragmentation parameter (r) are calculated with both high and low levels of theory, and those with a large parameter (R, full molecule in this study) accounting for long-range interactions are calculated only at a low level of theory. With the smaller fragmentation parameter (r), the primary subsystems are formed by combining two or three of the adjacent monomers resulting in dimer or trimer subsystems. The dimer primary subsystems are ideal for the NMR calculations for nucleic acids since their size is ideal to capture the base pair stacking and intramolecular hydrogen-bonding interactions using a high level of theory. Since the primary subsystems are formed by starting from each of the monomers, there are overlapping parts that need to be accounted for. To account for the overcounting of the overlapping parts, derivative subsystems are formed using the inclusion−exclusion principle. Conveniently, as shown below (*vide infra*), the rigid construct of DNA and the localized property of NMR chemical shifts help to accurately predict the NMR chemical shifts using a single layer of MIM protocol without adding the secondary long-range interactions. The truncated bonds in the subsystems are saturated with link-hydrogen atoms. MIM1 and MIM2 energies can be written like the standard ONIOM extrapolation expression, as shown in eq 1 and eq 2.

$$E^{\mathrm{MIM1}} = E^{r}_{\mathrm{high}} \tag{1}$$

$$E^{\mathrm{MIM2}} = E^{r}_{\mathrm{high}} - E^{r}_{\mathrm{low}} + E^{R}_{\mathrm{low}}\ (r \ll R) \tag{2}$$

Here (r) and (R) represent generalizations of the "model system" and "real system," respectively, as in the standard ONIOM calculations. Thus, E^{r}_{high}, E^{r}_{low}, and E^{R}_{low}, represent the generalized E_{mh}, E_{ml}, and E_{rl} in the ONIOM energy expression. As has been described previously, the energy summation for the high and low levels of theory is carried out, according to the inclusion−exclusion principle, taking into consideration the appropriate signs of the energy terms involving the different primary and derivative subsystems.[30−34]

2.2. NMR Calculations. For the NMR-gauge-including atomic orbital (GIAO) method, 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.

$$\sigma^{N}_{ij} = \left[\frac{\partial^2 E}{\partial B_i \partial m_{N_j}}\right]_{B=0} \tag{3}$$

where σ^{N}_{ij} is the ijth component of the shielding tensor, B_i is the ith component of the external magnetic field, and m_{N_j} is the jth component of magnetic moment of the nucleus N.

In MIM2, isotropic shielding tensor for all of the atoms is calculated using a general expression

$$\sigma^{N}_{ij} = \left[\frac{\partial^2 E_{\mathrm{total}}}{\partial B_i \partial m_{N_j}}\right]_{B=0} = \frac{\partial^2 E_{\mathrm{rl}}}{\partial B_i \partial m_{N_j}} - \frac{\partial^2 E_{\mathrm{ml}}}{\partial B_i \partial m_{N_j}} + \frac{\partial^2 E_{\mathrm{mh}}}{\partial B_i \partial m_{N_j}} \tag{4}$$

The atomic NMR shielding constant is one-third of the sum of the trace of the atomic shielding tensors from eq 3. 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 tetramethylsilane (TMS) as the reference. For ^{15}N, the NH_3 molecule is taken as the reference.

$$\delta_i = \sigma_{\mathrm{ref}} - \sigma_i \tag{5}$$

For the initial calibration calculations, the two-layer MIM (MIM2) model was compared with the one-layer MIM (MIM1) model for energy evaluations as well as NMR chemical shift predictions. For the MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G] method, *m*PW1PW91/6-311G(d,p) is used in high layer and *m*PW1PW91/6-31G in a low layer, whereas for the MIM1[*m*PW1PW91/6-311G-(d,p)] method, *m*PW1PW91/6-311G(d,p) is used in a single layer. 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. Note that *m*PW1PW91 has previously been shown to give quite accurate results for the calculations of the NMR chemical shift predictions of small

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

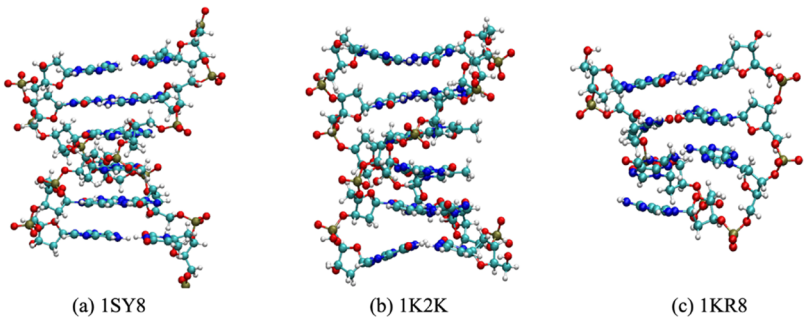


Figure 1. Structures of 1SY8, 1K2K, 1KR8 nucleic acid systems employed for benchmarking studies.

molecules, natural products, and our previous study on proteins.[22,38] It is important to note that nucleic acids have polar functional groups, which may give an overall charge to the system. Since electrostatic interactions in gas-phase calculations are significantly overestimated, charged residues are neutralized, as customarily done in several previous studies.[34,39]

2.3. Solvation Models. The solvent environment for the MIM-NMR calculation is incorporated using implicit, and explicit−implicit solvent models. For implicit solvation, the SMD-SCRF[40] implicit solvation model is used.

$$E_{\text{Total}}^{\text{Implicit}} = E_{\text{mh}}^{\text{Implicit}} \tag{6}$$

Recent theoretical studies suggest that the local solvent environment near sensitive amino groups is important in prediction of accurate chemical shifts.[22,39] It has been shown that adding a small number of direct hydrogen-bonded water molecules is sufficient to accurately predict the amino proton chemical shifts. In our explicit−implicit solvation model, the explicit solvent molecules are added to the spines of the DNA molecule near the exposed amino groups. The well-packed rigid structure of DNA molecules and the local nature of NMR chemical shifts can be used to our advantage while considering the MIM fragmentation strategy. The assumption here is that the dimer primary subsystems with the explicit−implicit solvation model can accurately model the local intramolecular hydrogen-bonding interactions as well as intermolecular explicit interactions with the solvent (*vide infra*) at the high level of theory. Our results show that contributions from the long-range interactions are negligible and the one-layer MIM model with the explicit−implicit solvation model substantially lowers the computational cost of performing the full molecule calculations. In our explicit−implicit solvation model, the short-range hydrogen-bonding interactions are captured by including one explicit water molecule per amine proton, and other solvation effects are captured using the SMD implicit solvation model. Because the turns and twists formed by such interactions cannot accommodate an explicit water molecule, amine groups with intramolecular hydrogen-bonding interactions are excluded from the addition of explicit solvent molecules. This avoids adding a random number of explicit water molecules, which would necessitate proper equilibration and careful sampling of solvent molecules, potentially leading to significant increases in computational cost. In contrast, our approach is systematic while keeping computational costs to a minimum. Finally, all of the explicit water molecules are geometry optimized using the PM6D3H4 semiempirical method using Molecular Orbital PACkage (MOPAC) while freezing the rest of the DNA molecule to preserve its conformation.[41]

3. DETERMINATION OF FRAGMENTATION PROTOCOLS FOR MIM-NMR OF NUCLEIC ACIDS

In this section, we calibrate the performance of different fragmentation schemes in MIM for the calculation of the energy and NMR chemical shifts of the DNA molecules. First, we calibrate the performance of MIM for obtaining accurate energy ordering of multiple conformers of nucleic acids, and then, we calibrate one-layer and two-layer MIM-NMR methods on the lowest energy conformer for the prediction of NMR chemical shifts. To calibrate the MIM methods, we have obtained the structure and experimental chemical shifts of three DNA molecules 1SY8 (BMRB ID 6186),[42] 1K2K (BMRB ID 5339),[43] and 1KR8 (BMRB ID 5282)[44] from the Protein Data Bank (PDB), as shown in Figure 1. 1SY8 and 1K2K DNA molecules are used for regulation of gene expressions and intercalation site for several anticancer drugs, whereas 1KR8 is an extraordinarily stable mini-hairpin d(GCGAAGC) heptamer. 1SY8 and 1KR8 DNA molecules have 10 and 14 conformers each submitted to the Protein Data Bank, 1SY8 being a DNA duplex with d(TGATCA)2 sequence, and 1KR8, a hairpin structure with d(GCGAAGC) sequence. The third system, 1K2K, with d(CGTACG)2 has one submitted conformer in the PDB.

In these calibrations, we first used multiple conformers of the 1SY8 DNA duplex with 12 residues and 6 base pairs for the calibration of MIM-based conformational energies. Both 1SY8 and 1K2K DNA duplexes were then used for calibration of the MIM-NMR method starting from gas-phase calculations at experimentally determined geometries. Subsequently, we optimize the MIM-NMR protocol for 1SY8 and 1K2K structures along with the heptamer hairpin structure of 1KR8 by incorporating structural minimizations and solvation effects for the prediction of NMR chemical shifts by comparing it with the reported experimental values. All three nucleic acids considered for developing the MIM protocol have polar functional groups and side chains, which contribute to overall charge to the systems. To minimize errors resulting from the overestimation of the strengths of electrostatic interactions in the calculations, charged residues are neutralized and this approach has been used previously as a reasonable approximation for proteins.[22,34,39]

3.1. Calibration of MIM Energy *vs* Full Energy for Multiple Conformers of the 1SY8 System. 1SY8 DNA molecule, with 12 residues and 390 atoms, has 10 conformers. To check the performance of the MIM method for conformational energy differences, we evaluated the single-

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

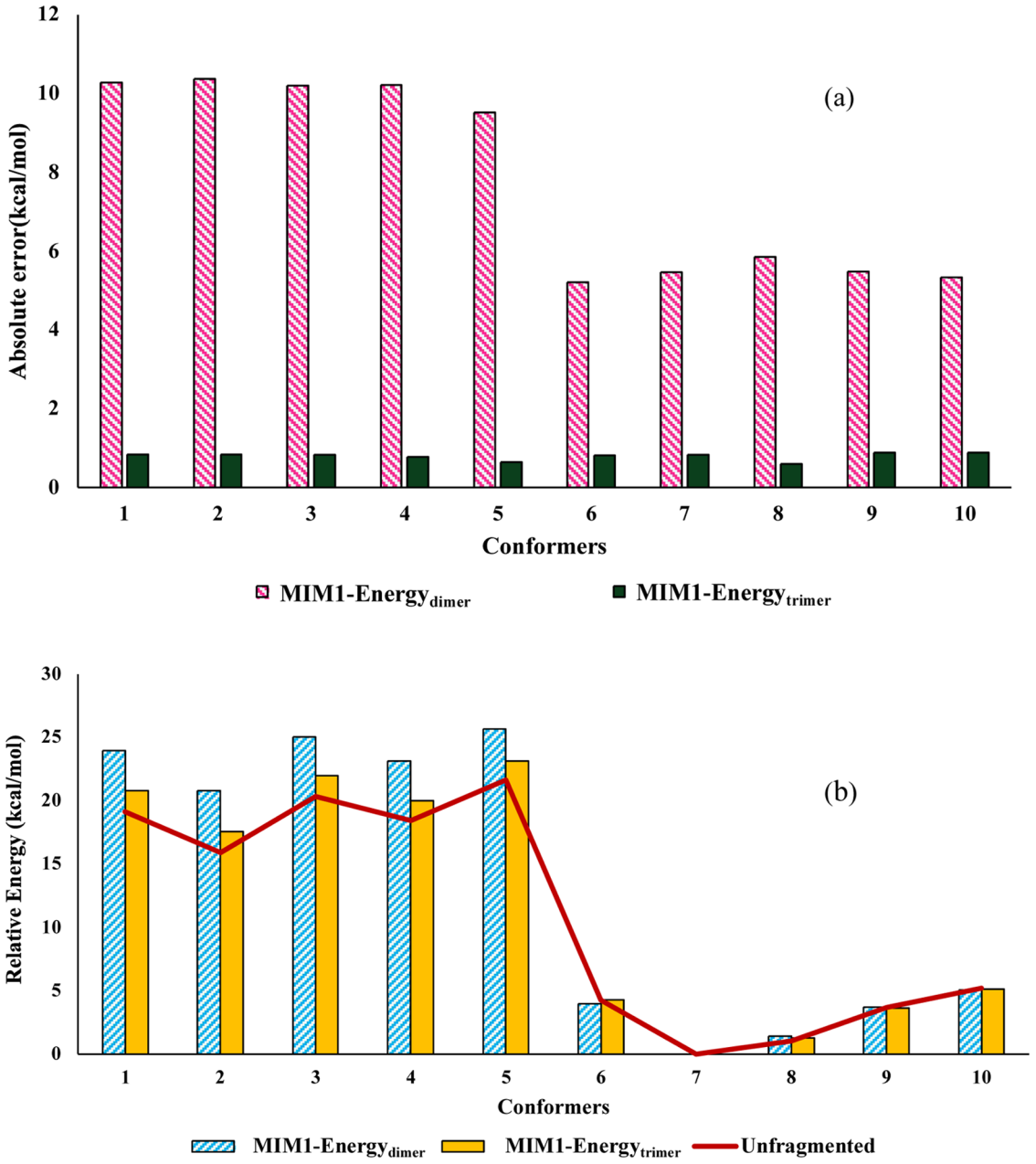


Figure 2. (a) Cluster plot to show the absolute error for calculated energies of 1SY8 DNA duplex using MIM1[CAM-B3LYP-D3BJ/6-31+G(d)] compared to unfragmented calculations. Magenta pattern-filled column and green-filled column for MIM1 dimer and trimer fragmentation energies. (b) Cluster plot for the relative energies of the 10 conformers of 1SY8 DNA duplex with respect to the lowest energy conformer calculated using the CAMB3LYP-D3BJ/6-31+G(d) method. Blue pattern-filled column and yellow-filled column for MIM1 dimer and trimer relative energies, and the red line plot shows the unfragmented relative energies of the conformers.

point energies of all conformers using the CAM-B3LYP-D3BJ/6-31+G(d) method and compared with MIM1$_{dimer}$ (MIM1 with dimer subsystems) and MIM1$_{trimer}$ (MIM1 with trimer subsystems) energies. The calculated mean absolute error (MAE) of energies using MIM is shown in Figure 2a and Table S1. MAE of the MIM1$_{dimer}$ and MIM1$_{trimer}$ methods with respect to the unfragmented (full molecule) calculation gives an average MAE of 7.79 and 0.80 kcal/mol, respectively. As expected, the absolute energies are better reproduced by the trimer model. However, the relative energy ordering of conformers obtained using both dimer and trimer MIM1 fragmentation is the same as in the unfragmented full molecule, with conformer 7 (conf-7) being the lowest energy structure as shown in Figure 2b. Since the errors obtained from the MIM1$_{trimer}$ method are very small (less than 1 kcal/mol), it is an excellent method to get the energy ordering of closely lying conformers instead of doing the expensive full molecule calculation. The MIM1$_{dimer}$ model can also be used to estimate the conformational energies without significant loss in accuracy though we use the more rigorous MIM1$_{trimer}$ in this study. In our previous MIM-NMR studies for proteins, Boltzmann averaging of conformers was used to get the relevant conformers for the NMR chemical shift evaluations.[22] Here, a detailed evaluation of Boltzmann contributions of each conformer from different methods is shown in the Supporting Information (SI) (Tables S2−S4 of the SI). The evaluation of

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

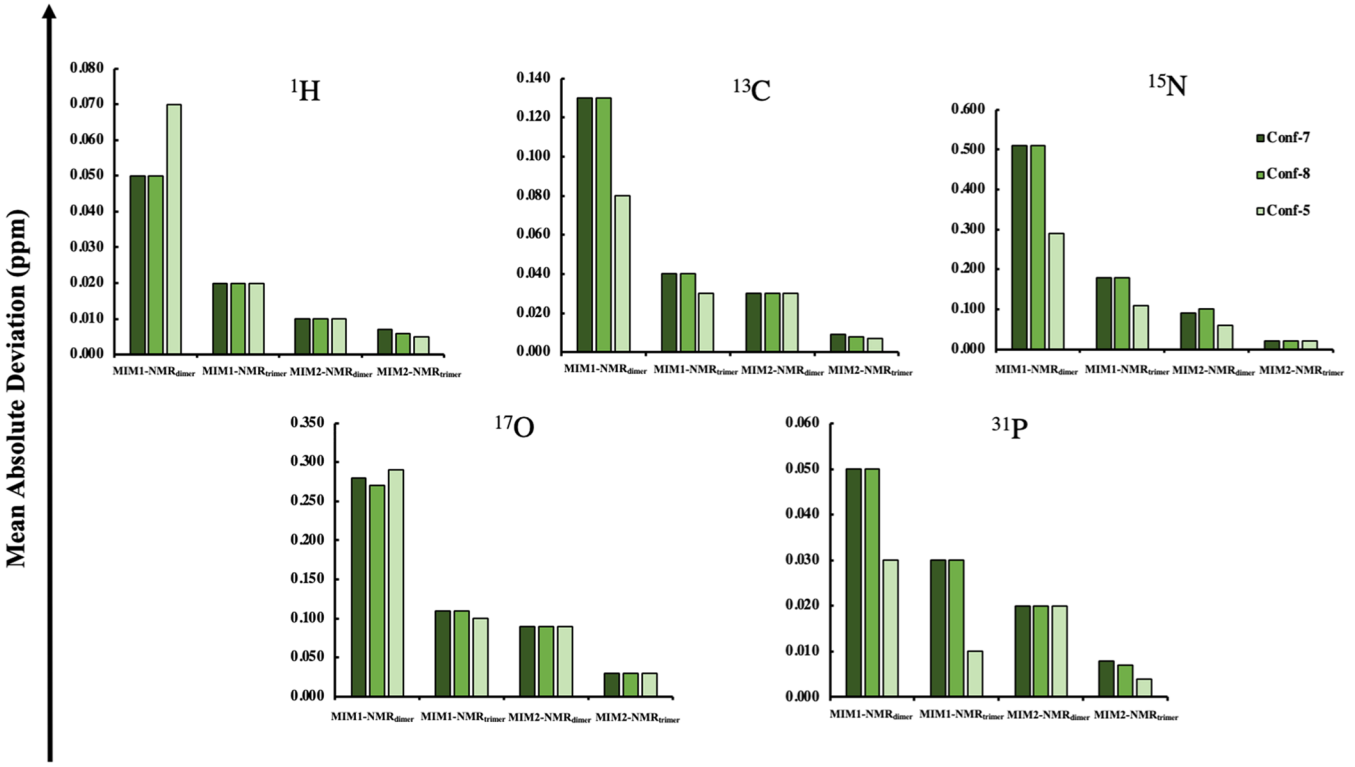


Figure 3. Mean absolute deviation (MAD) for calculated chemical shifts at MIM1[*m*PW1PW91/6-311G(d,p)] and MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G] for dimer and trimer primary subsystems, compared to NMR chemical shift calculated for the full, unfragmented conformers 7, 8, and 5 of DNA duplex 1SY8.

the Boltzmann contributions using the MIM1$_{trimer}$ method shows that conformers 7 and 8 have 89.2 and 10.5% contributions, respectively. MIM-NMR studies were then performed on structures with Boltzmann contributions to understand their importance on chemical shift predictions.

3.2. Calibration of MIM-NMR *vs* Full Molecule NMR Calculations for 1SY8 and 1K2K DNA Duplex Structures. After obtaining the energy ordering of conformers for 1SY8 duplex, we applied the one-layer and two-layer MIM-NMR protocols to evaluate the chemical shifts. We selected the two lowest energy conformers, conformer 7 (conf-7) and conformer 8 (conf-8), along with the highest energy conformer, conformer 5 (conf-5), to compare the NMR chemical shifts of the full molecule *vs* fragmentation calculations in the gas phase. The MIM fragments for DNA molecules were generated by cutting the carbon−carbon single bond in the phosphate backbone while preserving the nucleic acid−base pairing interactions as shown in Scheme 1 (*vide supra*). The NMR calculations are carried out using the *m*PW1PW91/6-311G(d,p) method for MIM1 (and full molecule calculations for calibration), whereas [*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G] methodology is used for MIM2-NMR calculations. The mean absolute deviation (MAD) in isotropic magnetic shielding tensors using MIM1 and MIM2 methods was compared with the full molecule calculations and are shown in Figure 3 and Table S5 of the Supporting Information. From the figure, it is evident that for all three conformers, as the fragment size increases, the mean absolute deviation becomes smaller. For proton (^{1}H) and carbon (^{13}C), the average MAD error of 0.06 and 0.11 ppm for MIM1$_{dimer}$ improved to an average error of 0.02 and 0.04 ppm for MIM1-NMR$_{trimer}$ calculations. Similarly, MIM2-NMR$_{dimer}$ error of 0.01 and 0.03 ppm for proton and carbon improved to 0.006 and 0.008 ppm, respectively, for MIM2$_{trimer}$ calculations. The errors from both MIM2-NMR$_{dimer}$ and MIM2-NMR$_{trimer}$ methodologies show lower errors compared to MIM1-NMR chemical shift evaluations.

Like 1SY8, calibrations are performed on the 1K2K DNA molecule with 12 residues and 388 atoms. The only submitted conformer of 1K2K in the PDB is used for the MIM-NMR calibrations. The MAD values with respect to full calculations for MIM1 and MIM2-NMR evaluations of dimer and trimer subsystems are shown in Table S6 of the Supporting Information. In general, MIM-NMR$_{trimer}$ gives less MAD values compared to MIM-NMR$_{dimer}$, irrespective of one-layer or two-layer models. When the one-layer and two-layer models are compared, MIM2 outperforms the MIM1 model with the lowest MAD values for MIM2-NMR$_{trimer}$ with 0.006 ppm for ^{1}H and 0.01 ppm for ^{13}C.

For both 1SY8 and 1K2K systems, going from dimer to trimer for both MIM1 and MIM2 fragmentation methods showed significant improvement in the calculated deviations with respect to unfragmented calculations. This evaluation shows that the size of the fragment and the addition of the second layer improve the MAD values with respect to the unfragmented full molecule calculations.

3.3. Calibration of Gas-Phase MIM-NMR *vs* Experimental ^{1}H NMR Chemical Shifts for 1SY8 and 1K2K DNA Duplexes. MAD values for calculated NMR chemical shifts with respect to experiments for conformers 7, 8, and 5 of 1SY8 system using different fragmentation schemes are shown in Table 1. It is interesting to note that the MAD values for all conformers with respect to the experimental NMR chemical shifts stayed the same around ∼0.40 ppm with an R^2 value of

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

Table 1. Mean Absolute Deviation of ^{1}H Gas-Phase NMR Chemical Shift Using MIM1[*m*PW1PW91/6-311G(d,p)] and MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G] for Dimer and Trimer Primary Subsystems of Conformers 7, 8, and 5 of 1SY8 DNA Duplex, with Respect to Experimental NMR Chemical Shifts[a]

MIM_{gas}-NMR model	conformers	dimer (MAD/R^2)	trimer (MAD/R^2)	unfragmented (MAD/R^2)
^{1}H MIM1-NMR	conf-7	0.39/0.92	0.40/0.91	0.40/0.91
	conf-8	0.38/0.92	0.39/0.91	0.40/0.91
	conf-5	0.41/0.91	0.42/0.91	0.42/0.91
^{1}H MIM2-NMR	conf-7	0.41/0.91	0.41/0.91	
	conf-8	0.40/0.91	0.40/0.91	
	conf-5	0.42/0.91	0.42/0.91	

[a]MAD values are given in the units of ppm.

0.91, irrespective of the size of the fragment considered or the number of layers used in the MIM fragmentation scheme. The MAD values of the highest energy conformer conf-5 are marginally higher than those of conf-7 and conf-8, whereas the MAD values for both conf 7 and 8 with respect to the experiments stayed the same. Even though the two-layer method outperformed the one-layer method for calibration calculations, a comparison with the experimental calculations shows that the $MIM1_{dimer}$-NMR method can predict the gas-phase NMR experimental chemical shifts with similar accuracy as MIM2 or unfragmented NMR calculations, as shown in Table 1. Comparable errors from the two lowest and the highest energy conformers of the 1SY8 DNA molecule suggest that Boltzmann evaluation may not be needed to predict the NMR chemical shift of DNA molecules, suggesting that the lowest energy conformer can be used as the representative structure for further optimization of NMR chemical shift predictions.

For 1K2K DNA duplex, MAD values for experimental NMR chemical shifts with respect to one-layer and two-layer MIM models with dimer and trimer are shown in Table 2. Again, a

Table 2. Mean Absolute Deviation of ^{1}H Gas-Phase NMR Chemical Shift Using MIM1[*m*PW1PW91/6-311G(d,p)] and MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G] for Dimer and Trimer Primary Subsystems of 1K2K DNA Duplex, with Respect to Experimental NMR Chemical Shifts[a]

MIM_{gas}-NMR model	dimer (MAD/R^2)	trimer (MAD/R^2)	full (MAD/R^2)
^{1}H MIM1-NMR	0.50/0.96	0.50/0.96	0.51/0.96
^{1}H MIM2-NMR	0.51/0.96	0.51/0.96	

[a]MAD values are given in the units of ppm.

cheaper $MIM1_{dimer}$-NMR calculation gives a MAD value of 0.50 ppm with an R^2 of 0.96 with respect to the experiments, whereas a second layer or larger fragmentation method did not show any improvement in the MAD values. This evaluation further strengthens our observations from the 1SY8 molecule that MIM2-NMR methods and $MIM1\text{-}NMR_{trimer}$ methods do not improve the gas-phase chemical shift predictions, compared to much cheaper $MIM1\text{-}NMR_{dimer}$ calculations.

Broadly, two factors play a key role in the success of the $MIM1_{dimer}$-NMR methodology: (1) the double-helical structure of nucleic acids and (2) NMR being a localized property. A dimer fragment consists of two rungs of the DNA ladder with two base pairs and associated backbones. Moreover, the modest size of dimer fragments is enough to capture the major nonbonded interactions like intermolecular hydrogen bonding between the base pairs and the stacking interaction between different rungs of the DNA ladder. This evaluation shows that the $MIM1_{dimer}$-NMR calculation is enough to capture the local information, compared to expensive multilayer fragmentation methods. Therefore, we chose to select the lowest energy conformer of 1SY8 system, conf-7, in this example, to further evaluate the NMR chemical shifts of DNA molecules using a $MIM1\text{-}NMR_{dimer}$ model with the *m*PW1PW91/6-311G(d,p) method.

We conclude in this section that conformational energy differences can be reliably determined with both $MIM1_{dimer}$ and $MIM1_{trimer}$ models, though the latter has much smaller absolute energy errors. Since single-point energy evaluations are quite efficient, we chose to use the $MIM1_{trimer}$ model to find the lowest energy conformer. For the more expensive MIM-NMR calculations, the $MIM1_{dimer}$ model is adequate and gives results very similar to those from more expensive $MIM1_{trimer}$ or MIM2 (dimer or trimer) or even unfragmented calculations for the prediction of NMR chemical shifts.

4. OPTIMIZATION OF MIM1-NMR COMPUTATIONAL PROTOCOLS USING CONSTRAINED MINIMIZATIONS AND SOLVATION EFFECTS FOR 1SY8, 1K2K AND 1KR8 SYSTEMS

The calibrations carried out thus far for both 1SYS and 1K2K systems used experimentally determined geometries along with gas-phase MIM-NMR calculations. They showed that the errors from fragmentation are small and that the computed results using MIM models are close to those using unfragmented models. However, any error coming from the computational models (*e.g.*, density functional, solvation model) is still present. The calculated ^{1}H chemical shift errors for 1SYS and 1K2K, 0.39 and 0.50 ppm, are outside the target range of 0.2−0.3 ppm though they are not too far off. However, for 1KR8, a small nucleic acid hairpin system of 7 residues (228 atoms) with three paired and an unpaired nucleotide, the errors are significantly larger (MAD 0.78 ppm, *vide infra*). For the 1KR8 system, the unpaired mononucleotide is considered as a monomer for the fragmentation protocol, along with three paired nucleotides counting to a total of four monomers for the MIM protocol (*vide supra*). The evaluation of 14 conformers of 1KR8 using the MIM1:CAM-B3LYP-D3BJ/6-31+G(d) method gave conformer 3 (conf 3) as the lowest energy structure and used for further NMR calculations. The $MIM1_{trimer}$ energies for 1KR8 conformers are shown in Table S7 of the supporting information. It is interesting to note that the energy error in $MIM1_{trimer}$ with respect to full unfragmented calculation is just 0.11 kcal/mol. However, as noted above, the MAD in the ^{1}H NMR chemical shifts with the $MIM1_{dimer}$ protocol is significantly larger (0.78 ppm). The same MAD value with respect to experiments is obtained for the $MIM1_{dimer}$, $MIM1_{trimer}$, and the unfragmented calculations as shown in Table S8, strengthening the fact that the larger error is not coming from the MIM fragmentation protocol. More sophisticated computational models that go beyond the gas-phase NMR calculations are clearly needed.

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

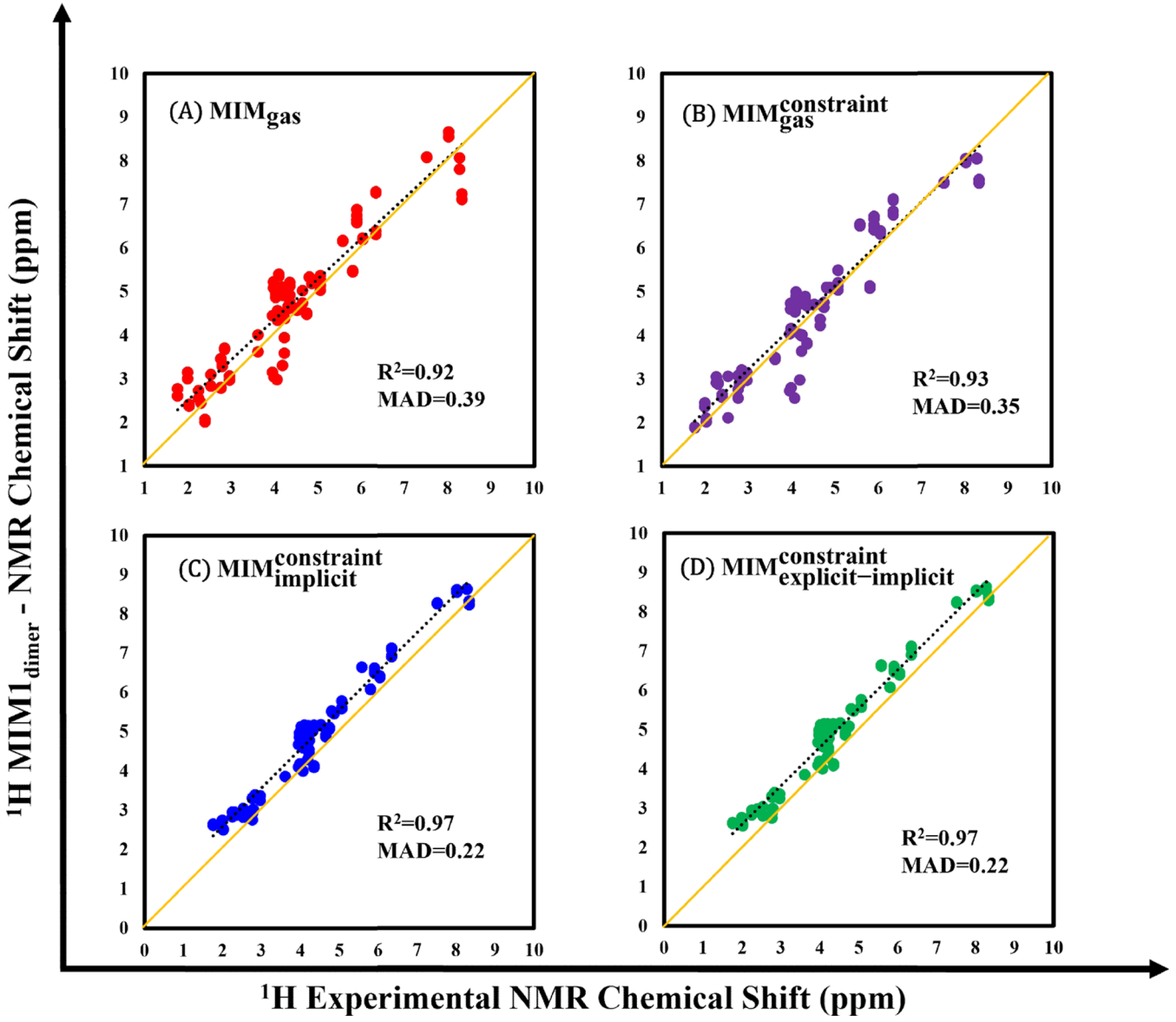


Figure 4. Comparison of experimental ^{1}H NMR of 1SY8 with MIM1$_{dimer}$ ^{1}H NMR calculated at the MIM1[*m*PW1PW91/6-311G(d,p)] level in MIM$_{gas}$, MIM$_{gas}^{constraint}$, MIM$_{imp}^{constraint}$, and MIM$_{explicit-implicit}^{constraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

The major challenge of all NMR prediction models is to accurately obtain the chemical shifts of protons, which are involved in hydrogen-bonded systems. Hydrogen-bonded backbone, base stacking, and base pairing protons have an important role in maintaining the structure of nucleic acids.[45] Therefore, we use the chemical shift of protons to optimize our MIM-NMR protocol by minimizing the MAD values of calculated NMR chemical shifts from the experiments. Based on the analysis in the previous section, the lowest energy conformer from MIM1$_{trimer}$ [CAM-B3LYP-D3BJ/6-31+G(d)] and chemical shifts from MIM1$_{dimer}$ [*m*PW1PW91/6-311G-(d,p)] are used as a good starting point in this subsection.

Further improvement of gas-phase NMR chemical shifts can be obtained by constrained minimization of the geometry using a combination of molecular mechanics and semiempirical methods along with the addition of implicit and explicit solvation models, which are evaluated using four different MIM-NMR models as shown in Figures 4−8. These four MIM-NMR models correspond to (A) MIM$_{gas}$, (B) MIM$_{gas}^{constraint}$ (C) MIM$_{implicit}^{constraint}$, and (D) MIM$_{implicit-explicit}^{constraint}$. A detailed description of each of the above models is given and discussed below.

4.1. Gas-Phase MIM1-NMR Calculations. Figure 4 shows the comparison of ^{1}H MIM1$_{gas}$ calculations with the experimental chemical shifts for the 1SY8 system, which spans around 0−9 ppm in the NMR spectra, while Figures 5 and 6 show similar comparisons for 1K2K and 1KR8 systems with chemical shift spans between 0 and 14 ppm. For 1SY8 and 1K2K systems, the linearly fitted plot gives the MAD value of 0.39 and 0.50 ppm with R^2 values of 0.92 and 0.96, respectively. Compared to 1SY8 and 1K2K systems, 1KR8 gave the largest MAD of 0.78 ppm from the experiment with an R^2 value of 0.81 for ^{1}H MIM gas-phase calculations. Along with ^{1}H chemical shifts for 1KR8, ^{13}C and ^{15}N experimental NMR chemical shifts are compared with the MIM$_{gas}$ calculations and shown in Figures 7 and 8. Here, it is

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

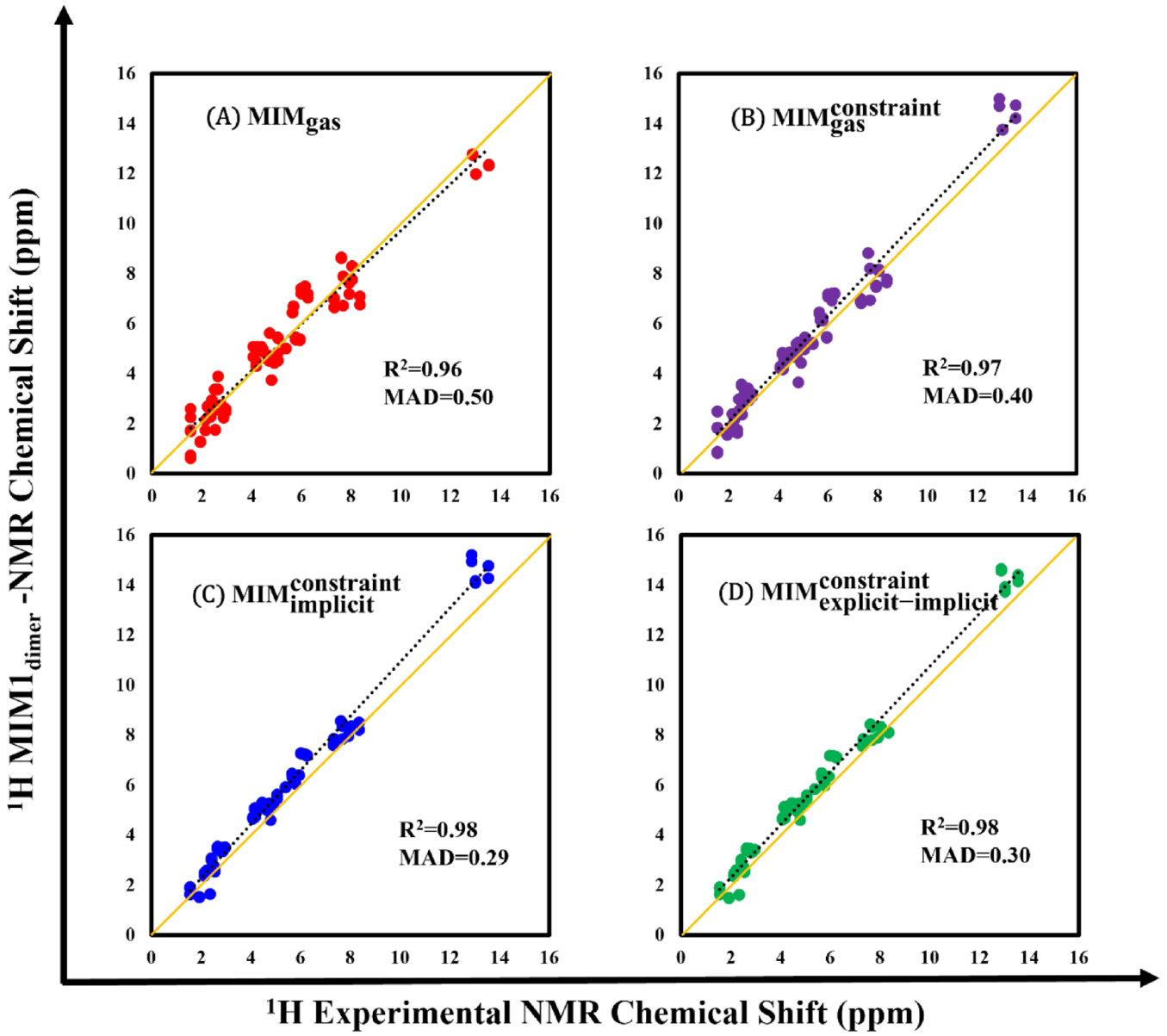


Figure 5. Comparison of experimental ^{1}H NMR of 1K2K with MIM1$_{dimer}$ ^{1}H NMR calculated at the MIM1[*m*PW1PW91/6-311G(d,p)] level in MIM$_{gas}$, MIM$_{gas}^{constraint}$, MIM$_{imp}^{constraint}$, and MIM$_{explicit-implicit}^{constraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

important to note that these results are from the gas-phase structures directly obtained from the Protein Data Bank without further minimization. To obtain a good starting structure for NMR calculations, we performed constrained minimizations on DNA structures without disturbing the complex intramolecular hydrogen-bonding network that holds the shape of the molecule (*vide infra*).

4.2. Constrained Minimizations. To assess the effect of geometry minimizations on the accuracy of calculated NMR chemical shift predictions, constrained minimizations were performed in two steps. Initially, the lowest energy structures were minimized using AMBER10:EHT forcefields by constraining all of the heavy atoms and allowing the protons to move.[46,47] In the next step, the PM6D3H4 semiempirical method is used for constrained minimization of all of the amino and imino groups in nucleobases, using MOPAC,[41] while freezing the rest of the DNA molecule to preserve the conformation.

As shown in Figures 4−8B, MIM gas-phase constrained optimizations resulted in a significant improvement in calculated NMR chemical shifts, compared to unoptimized structures for 1SY8, 1K2K, and 1KR8 systems. The MIM$_{gas}^{constraint}$ method shows that the deviation from the experimental chemical shift improved to 0.35 ppm with an R^2 value of 0.93 for the 1SY8 molecule, whereas the MAD value improved to 0.40 ppm with an R^2 value of 0.97 for the 1K2K system. Finally, the MIM constrained gas-phase structure of the 1KR8 system showed significant improvement in MAD values of ^{1}H (0.62 ppm), ^{13}C (2.58 ppm), and ^{15}N (6.44 ppm).

4.3. Implicit and Explicit−Implicit Solvation Model for MIM1-NMR Calculations. Along with geometry minimizations, modeling an accurate solvation model is important in capturing the noncovalent interactions between the nucleic acid and the solvent water molecules. The NMR chemical shifts of labile protons are largely affected by the solvation

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

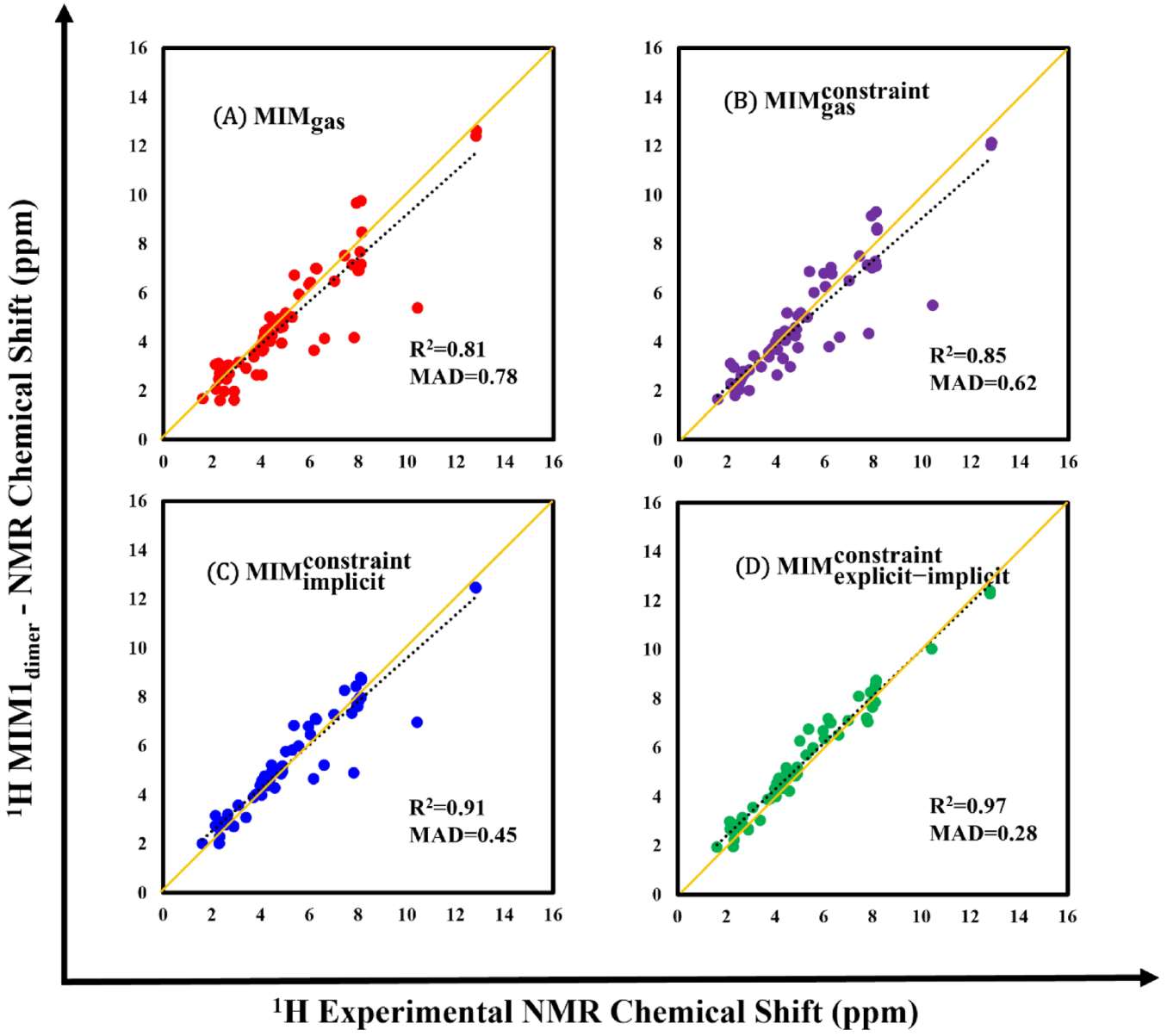


Figure 6. Comparison of experimental ^{1}H NMR of 1KR8 with MIM1$_{dimer}$ ^{1}H NMR calculated at the MIM1[*m*PW1PW91/6-311G(d,p)] level in MIM$_{gas}$, MIM$^{constraint}_{gas}$, MIM$^{constraint}_{imp}$, and MIM$^{constraint}_{explicit-implicit}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

environment, and we used the SMD implicit solvation model and explicit−implicit solvation environment to capture the missing solvation effects in the gas-phase MIM calculations. As shown in Figures 4C and 5C, incorporating implicit solvation calculations on constrained minimized structures substantially improved the deviations and helped bring down the MAD values to 0.22 ppm for 1SY8 and 0.29 ppm for 1K2K nucleic acid systems, well within our target accuracy of ∼0.2−0.3 ppm for ^{1}H. However, while we saw a significant improvement from the implicit solvation model in the MIM$^{constraint}_{implicit}$ deviations for the 1KR8 system, the MAD value of 0.45 ppm for ^{1}H is outside our target accuracy. In our previous study for proteins using the MIM-NMR method, a microsolvation model of adding a few water molecules near the highly hydrophilic exposed amino groups and capturing the rest of the solvent interactions using the implicit model remarkably improved the predicted chemical shifts.[22] To explore this, for the 1KR8 system, after adding six explicit water molecules, PM6D3H4 optimizations were carried out for the exposed hydrogen-bonding amino protons along with the explicit water molecules by fixing the heavy atoms and other protons. This dramatically improved the MAD values for the 1K2K system and reduced the error to 0.28, 2.43, and 3.15 ppm, respectively, for ^{1}H, ^{13}C, and ^{15}N nucleic acids. Being a small nucleic acid unit, most of the amino groups are exposed in the DNA molecule and implicit solvation is not sufficient to fully capture the solvation effects here. As shown in Figures 6−8D, MIM$^{constraint}_{explicit-implicit}$ calculations showed a huge improvement from gas-phase calculations for ^{1}H, ^{13}C, and ^{15}N NMR chemical shifts, and we achieved the target accuracy (∼0.2−0.3 ppm for ^{1}H, ∼2.0−3.0 ppm for ^{13}C, and ∼3 ppm for ^{15}N) for all three nuclei with reported experimental NMR chemical shifts.

MIM$^{constraint}_{explicit-implicit}$ calculations were also performed on 1SY8 and 1K2K systems and, interestingly, no improvement is

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

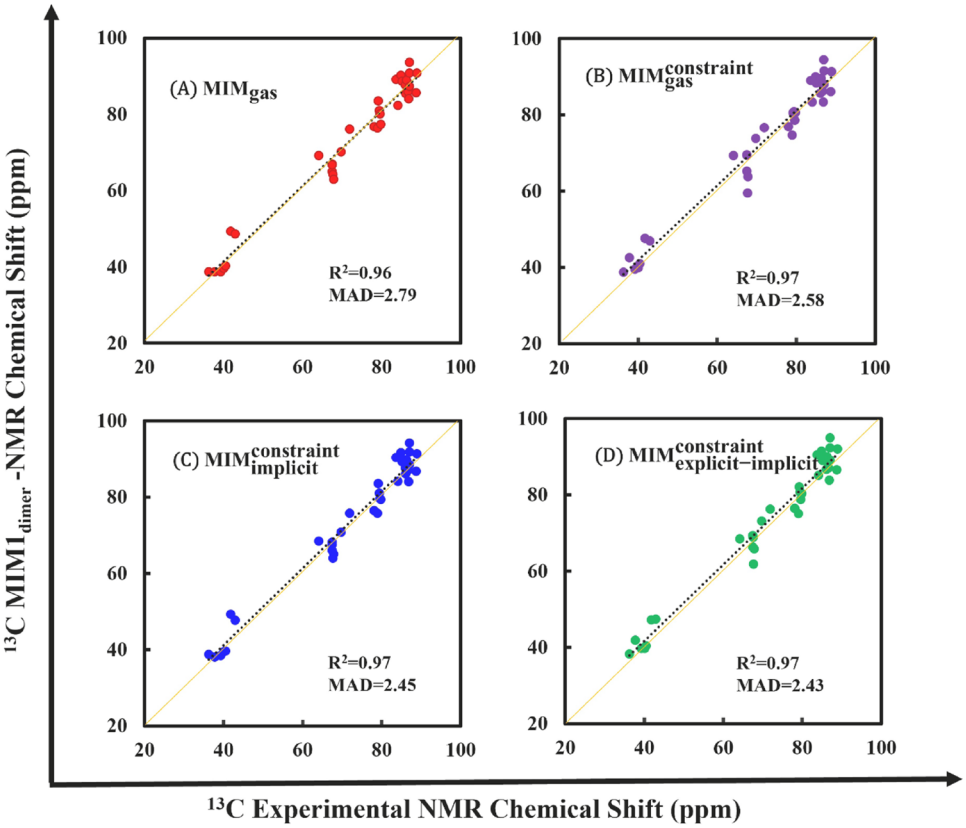


Figure 7. Comparison of the experimental ^{13}C NMR spectrum of 1KR8 with neutral residues using MIM_{gas}, $MIM_{gas}^{constraint}$, $MIM_{imp}^{constraint}$, and $MIM_{explicit-implicit}^{constraint}$ models at the MIM1[mPW1PW91/6-311G(d,p)] level. The MIM ^{13}C NMR chemical shifts are depicted in reference to tetramethylsilane (TMS).

observed compared to $MIM_{implicit}^{constraint}$ predictions. Additionally, $MIM_{implicit}^{constraint}$ calculations on the highest energy conformer of 1SY8 (conf-5) gave nearly the same results as conf-7, and no improvement is observed with the $MIM_{explicit-implicit}^{constraint}$ model. In this context, we note that for both 1SY8 and 1K2K nucleic acid molecules, we already achieved the target accuracy of 0.2−0.3 ppm with implicit solvation on the constraint-minimized structure, while for the 1KR8 system, major improvement came from the added explicit solvent molecules using the microsolvation model. Thus, we have adopted the more complete $MIM_{explicit-implicit}^{constraint}$ method as the standard protocol for DNA-type nucleic acid systems and assessed the performance of the MIM1-NMR_{dimer} method in the next section.

5. RESULTS AND DISCUSSION: APPLICATION OF MIM1-NMR PROTOCOL FOR THE PREDICTION OF NMR CHEMICAL SHIFTS

Now that we have established the protocol for NMR chemical shift predictions for nucleic acids, we have applied it to a wide variety of nucleic acid systems shown in Figure 9.

These include a DNA duplex with PDB ID: 2N5P (structure a), five nonstandard DNA systems with PDB IDs: 6XAH, 2LIB, 1N2W, 2LFX, and 7NBK (structures b−f), and one DNA/RNA hybrid system with PDB ID: 2LAR (structure g). Table 3 shows MIM_{gas}, $MIM_{gas}^{constraint}$, $MIM_{implicit}^{constraint}$, and $MIM_{explicit-implicit}^{constraint}$ evaluations of the chemical shifts of ^{1}H, ^{13}C, and ^{15}N nuclei with respect to experiments for all seven nucleic acids tested here. For comparison, the corresponding values for the three nucleic acids used for training the MIM1-NMR_{dimer} protocol are also listed in Table 3. Figures S1−S8 of the supporting information show the improvement in MAD and correlation coefficients of the MIM1-NMR_{dimer} method with respect to experiments for all seven nucleic acids with ∼300−1100 atoms, going from unoptimized gas-phase calculations to calculations in explicit−implicit solvation environment on constraint-minimized DNA structures (*vide infra*). Figure 10 illustrates the performance for all seven DNA systems with our best model, *viz.* $MIM_{explicit-implicit}^{constraint}$.

5.1. DNA Duplex with PDB ID: 2N5P (BMRB: 25724).

Our first example, 2N5P, is a synthetically constructed oligonucleotide DNA, used as a control for universal nucleobase studies for applications, such as templates for PCR primers, randomized sequencing, and DNA-based devices.[48] We used the only conformer of the DNA duplex submitted to the PDB, which is composed of 570 atoms, containing 18 residues with $d(GAGCTCCAT)_2$ sequence for MIM1-NMR_{dimer} calculations. Following our protocol, we added 14 explicit water molecules forming hydrogen bonding with the DNA structure along with the implicit SMD solvation model. From Table 3 and Figure S1 (in SI), it is evident that going from gas-phase MIM1-NMR_{dimer} calculations with a MAD value of 0.39 ppm (R^2 = 0.94) to the gas-phase constraint-minimized structure, we observe a slight improvement in the calculated values, with a MAD of 0.32 ppm (R^2 = 0.96). Significant improvement was observed when using the implicit solvation calculations, bringing down the MAD values to 0.19 ppm (R^2 = 0.98). However, there is no further improvement observed for explicit solvation $MIM_{explicit-implicit}^{constraint}$

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

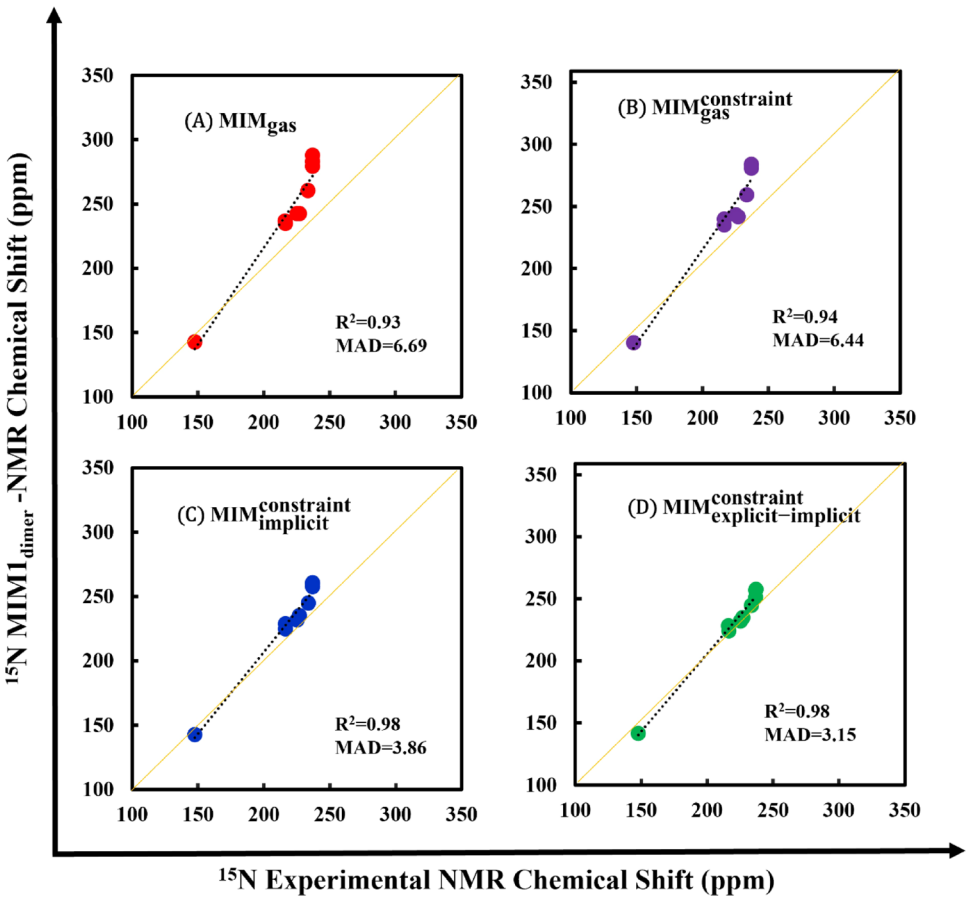


Figure 8. Comparison of the experimental ^{15}N NMR spectrum of the 1KR8 molecule with neutral residues using MIM_{gas}, $MIM_{gas}^{constraint}$, $MIM_{imp}^{constraint}$, and $MIM_{explicit-implicit}^{constraint}$ models at the MIM1[*m*PW1PW91/6-311G(d,p)] level. The MIM ^{15}N NMR chemical shifts are depicted in reference to NH_3.

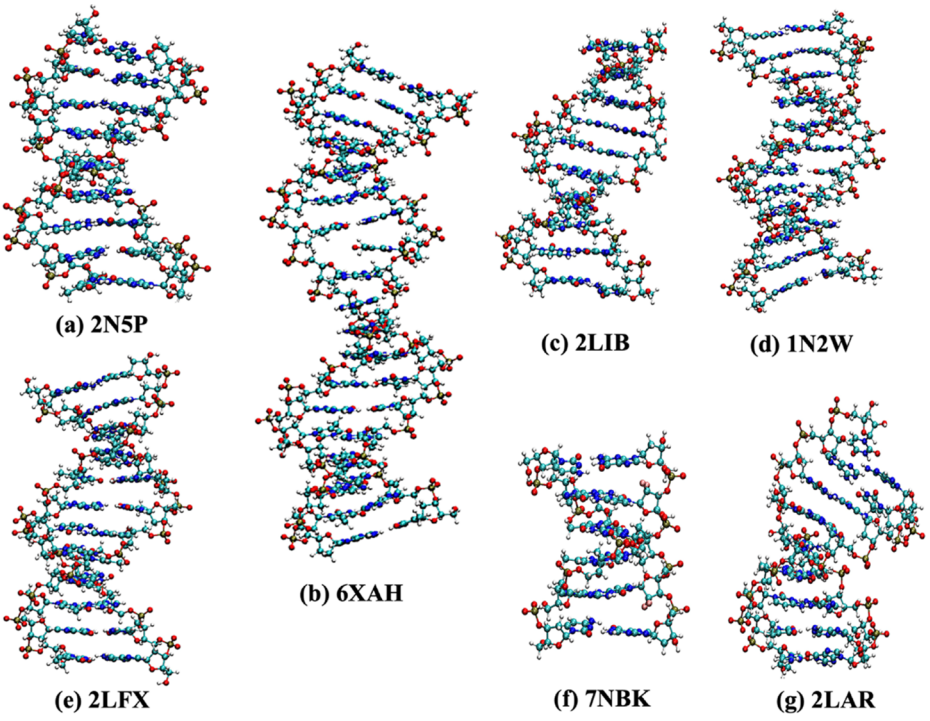


Figure 9. Molecules used to compare the MIM method with experimental chemical shifts. (a) 2N5P with 571 atoms, (b) 6XAH with 1130 atoms, (c) 2LIB with 649 atoms, (d) 1N2W with 789 atoms, (e) 2LFX with 783 atoms, (f) 7NBK with 378 atoms, and (g) 2LAR with 594 atoms.

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

Table 3. Structural Information, MAD (in ppm) Values between the Experiment and MIM-NMR$_{dimer}$ Predictions for Nine Nucleic Acid Systems Used in This Study

no.	PDB entry	description	N atom/residues	nuclei type	MIM_{gas} (MAD/R^2)	$MIM_{gas}^{constraint}$(MAD/R^2)	$MIM_{implicit}^{constraint}$ (MAD/R^2)	$MIM_{explicit-implicit}^{constraint}$ (MAD/R^2)
1	1SY8	$d(TGATCA)_2$	390/12	^1H	0.39/0.92	0.35/.93	0.22/0.97	0.22/0.97
2	1K2K	$d(CGTACG)_2$	388/12	^1H	0.50/0.96	0.40/0.97	0.29/0.98	0.30/0.98
3	1KR8	d(GCGAAGC) hairpin	228/7	^1H	0.78/0.81	0.62/0.85	0.45/0.91	0.28/0.97
				^{13}C	2.79/0.96	2.58/0.97	2.45/0.97	2.43/0.97
				^{15}N	6.69/0.93	6.44/0.94	3.86/0.98	3.15/0.98
4	2N5P	DNA duplex with no substitutions	571/18	^1H	0.39/0.94	0.32/0.96	0.19/0.98	0.20/0.98
5	6XAH	modified nucleobases with abasic site	1130/36	^1H	0.39/0.95	0.40/0.95	0.26/0.98	0.26/0.98
6	2LIB	DNA duplex with one a anomeric adenosine nonstandard residue	649/20	^1H	0.43/0.96	0.34/0.98	0.21/0.99	0.22/0.99
				^{13}C	2.32/0.99	2.57/0.99	1.93/0.99	1.89/0.99
7	1N2W	modified DNA duplex with deoxyguanosine residues	789/24	^1H	0.61/0.89	0.58/0.90	0.44/0.94	0.37/0.96
8	2LFX	modified DNA duplex with deoxyguanosine residues	783/24	^1H	0.33/0.96	0.31/0.96	0.22/0.98	0.22/0.98
9	7NBK	fluorine-substituted nonstandard residues	378/12	^1H	0.81/0.74	0.81/0.74	0.65/0.81	0.47/0.91
10	2LAR	DNA/RNA hybrid with boron-substituted nonstandard residue	594/18	^1H	0.46/0.96	0.41/.96	0.33/0.97	0.30/0.98
				^{13}C	2.98/0.98	2.88/0.98	2.47/0.99	2.38/0.99

calculations relative to $MIM_{implicit}^{constraint}$ calculations. Perhaps, this is not surprising since the implicit solvation already yielded results that exceeded our performance measures (0.2−0.3 ppm). These MIM evaluations show that including the implicit solvation is critical in obtaining accurate NMR chemical shift predictions on the optimized structure.

5.2. Stable Interstrand DNA Cross-Link Involving the deoxyAdenosine (dA) Amino Group and an Abasic Site with PDB ID: 6XAH (BMRB: 30759). 6XAH is a stable DNA with interstrand cross-link (ICL) between the nucleophilic N^6 amino group of deoxyAdenosine (dA) and the abasic site (AP). Abasic sites in DNA are formed by spontaneous depurination of canonical and chemically modified nucleobases, which results in the most common form of DNA damage in Homo sapiens.[49] 6XAH has 10 submitted conformers in the PDB with 36 residues and 1130 atoms, which make it the largest DNA system in our study for the prediction of NMR chemical shifts using the proposed MIM protocol. From the MIM energy method shown in Table S9, the lowest energy conformer, conf 3, is used for MIM-NMR calculations. The fifth row of Table 3 and panels A−D of Figure S2 (in SI) show the MAD value for the ^1H chemical shifts with respect to the experiments using MIM_{gas}, $MIM_{gas}^{constraint}$, $MIM_{implicit}^{constraint}$, and $MIM_{explicit-implicit}^{constraint}$ models. The MIM_{gas} method gave a MAD value of 0.39 with an R^2 value of 0.95. Adding the constraint minimizations alone did not improve the MAD value but the $MIM_{implicit}^{constraint}$ model improved the MAD value to 0.26 ppm (R^2 = 0.98). As in the case of 2N5P, there is no further improvement observed from the $MIM_{explicit-implicit}^{constraint}$ calculations relative to $MIM_{implicit}^{constraint}$ calculations. This analysis shows that for a large DNA structure, where traditional unfragmented DFT methods would be too expensive, our accurate $MIM_{implicit}^{constraint}$ protocol gave outstanding performance with a MAD value of 0.26 ppm, exceeding our target accuracy of 0.30 ppm.

5.3. DNA Duplex with Nonstandard α-Anomeric Adenosine (αA) Base with PDB ID: 2LIB (BMRB: 17887). We chose a DNA duplex, with a 5′CαAG-3′ core (PDB entry 2LIB, BMRB entry 17887), which is used in a DNA sequence context to modulate detection and repair of DNA damage. The nonstandard α-anomeric adenosine (αA) base of 2LIB is intrahelical in a reverse Watson−Crick orientation and forms a weak base pair with thymine of the opposite strand. Because of the core structure, 2LIB has a significantly reduced local structural perturbation in the backbone, stacking, tilt, roll and twists, resulting in a straighter DNA with a narrow minor groove, with 20 residues and 640 atoms.[50] For testing our MIM-NMR protocol, we used the only conformer of 2LIB submitted to the PDB. Panels A−D of Figure S3 (in SI) and Table 3 depict the theoretical−experimental correlations calculated using MIM_{gas}, $MIM_{gas}^{constraint}$, $MIM_{implicit}^{constraint}$, and $MIM_{explicit-implicit}^{constraint}$ models for ^1H and ^{13}C chemical shifts. The MAD values for the entire range of ^1H are (A) 0.43, (B) 0.34 (C) 0.21, and (D) 0.22 ppm, with R^2 values of (A) 0.96 (B) 0.98 (C) 0.99, and (D) 0.99, respectively. Figure S4A−D and Table 3 show the ^{13}C correlation graphs with MAD values of (A) 2.32, (B) 2.57, (C) 1.93, and (D) 1.89 ppm with an R^2 value of 0.99 for all computational models. ^1H NMR chemical shift for the 2LIB structure did not show any further improvement in the MAD values from the $MIM_{implicit}^{constraint}$ model, whereas ^{13}C NMR showed a slight improvement when microsolvation with 12 explicit water molecules is introduced.

5.4. DNA Structures with Modified Deoxyguanosine Residues (PDB IDs 1N2W and 2LFX with BMRB: 5385 and 17786). 1N2W, containing a deoxyguanosine residue with hydroxylated reactive oxygen species at the C-8 position, is used to study the oxidative damage of DNA and its role in cancer and aging.[51] 2LFX is a deoxyguanosine-modified DNA, which is used in mutagenicity and genotoxicity studies to understand hydroxyl radical-damaged nucleic acids.[52] Both 1N2W and 2LFX nucleic acids have 24 residues each with 789 atoms and 783 atoms, respectively. The seventh and eighth rows of Table 3 and Figures S5 and S6 of the SI show the improvements in the MAD values for both these modified DNA systems. For 1N2W, the MAD value for the gas-phase structure is 0.61 ppm, which is almost double compared to the 2LFX system with a MAD value of 0.33 ppm. This appeared to indicate that 1N2W gas-phase structure has more room for improvement compared to the 2LFX system. However, the constrained minimization of the gas-phase structure offered only a modest improvement in the deviation for the 1N2W

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

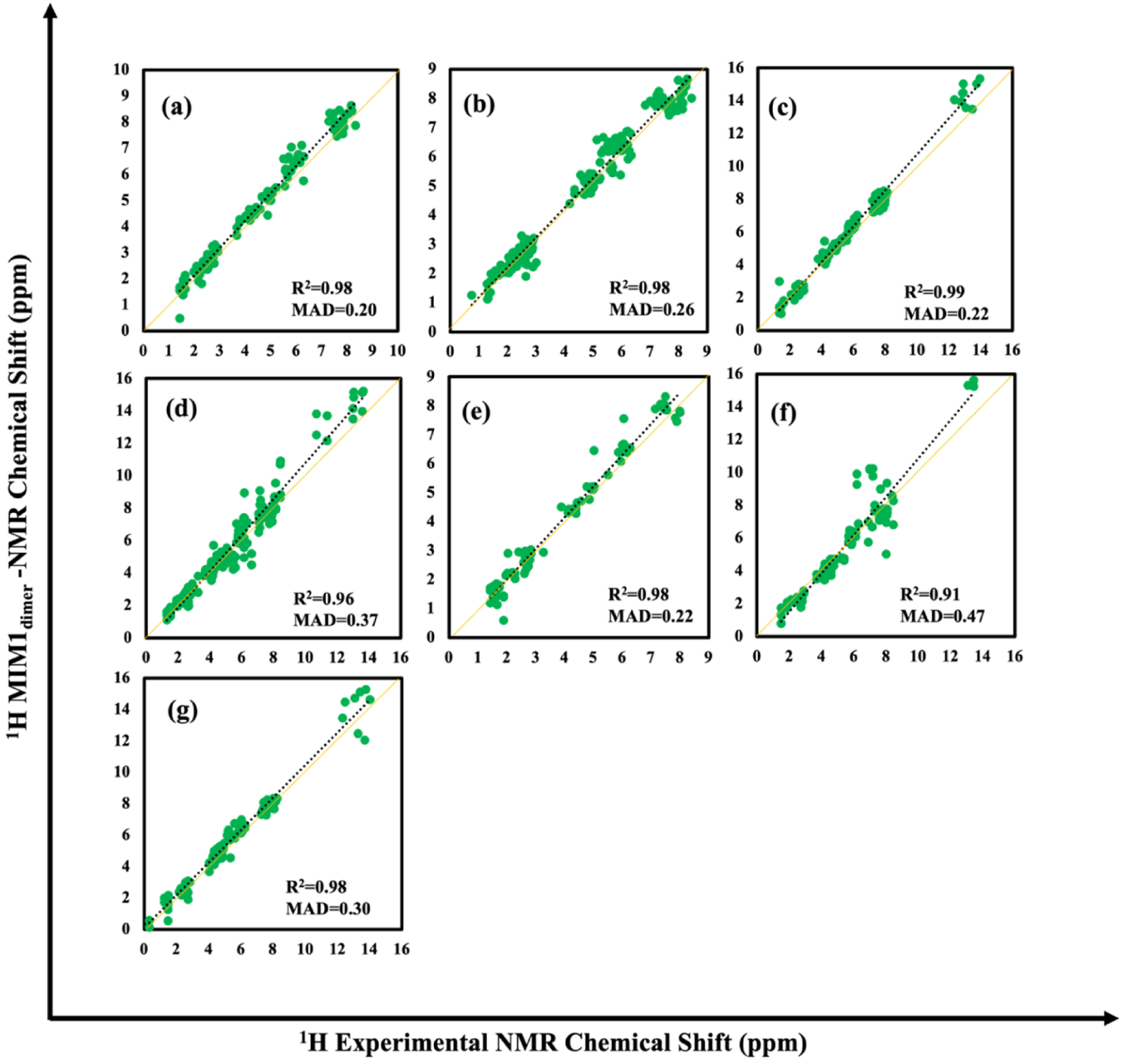


Figure 10. Comparison of experimental ^{1}H NMR of (a) 2N5P, (b) 6XAH, (c) 2LIB, (d) 1N2W, (e) 2LFX, (f) 7NBK, and (g) 2LAR with $MIM1_{dimer}$ ^{1}H NMR calculated at the MIM1[*m*PW1PW91/6-311G(d,p)] level using the $MIM^{constraint}_{explicit-implicit}$ model. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

system (0.61−0.58 ppm) or 2LFX system (0.33−0.31 ppm). For both 1N2W and 2LFX structures, implicit solvation ($MIM^{constraint}_{implicit}$ calculation) yielded significant improvement to the MAD values and brought them down to 0.44 and 0.22 ppm, respectively. The addition of explicit solvation along with the implicit solvation model showed a further modest improvement for the 1N2W system (0.37 ppm) while the deviation remained unchanged for the 2LFX system. Overall, SMD implicit solvation on constraint-minimized structures yielded the largest improvement for calculated NMR chemical shifts, though explicit solvation brings the agreement closer to the target value for 1N2W.

5.5. DNA Duplex with Fluorine-Substituted Sugar with PDB ID: 7NBK (BMRB: 34594). 2′-Deoxy-2′2′-difluorodeoxycytidine (gemcitabine) substituted B form DNA: DNA duplex 7NBK is a synthetic DNA containing fluorine-substituted sugar to enhance the stability of the DNA structure by providing charge polarizations from the high electronegativity of F.[53] 7NBK has 10 submitted conformers in PDB, and the energy ordering of conformers using the MIM energy method as well as with unfragmented calculations is shown in Table S10 (in the SI). Based on this, we concentrated on the lowest energy conformer, conformer 2 for our MIM1-NMR studies. In this case, as shown in Table 3 and Figure S7 of the SI, the error in ^{1}H NMR calculated using the MIM protocol decreased by 42%, going from unrefined gas-phase calculations (MAD = 0.81 ppm) to the constraint-minimized structure in solution phase with explicit−implicit solvation (MAD = 0.47 ppm), along with a substantial improvement in R^2 values (0.74 to 0.91). In particular, similar to the 1KR8

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

system from calibration studies, adding the explicit−implicit solvation model compared to the implicit solvation model showed a significant improvement in calculated MAD values of ^{1}H, lowering the error from 0.67 ppm with an R^2 value of 0.81 to 0.47 ppm with an R^2 value of 0.91. In this context, it is important to note that predicting the NMR chemical shift for modified systems is challenging, especially when attached to highly electronegative fluorine atoms. From Figure 10f, it is evident that a few protons in ∼6−8 ppm are responsible for the relatively large MAD value for the $MIM^{constraint}_{explicit-implicit}$ model. When the labile amide protons and α protons, which are close to the F atom-containing monomer (modified monomer, as well as the monomers above and below), are omitted from the analysis, the chemical shift MAD value drops to 0.30 ppm (Figure S8). This clearly shows that the four fluorine atoms present in the 7NBK backbone can drastically affect their neighboring environments.

Even here, the error comes down to 0.47 ppm after applying our MIM-NMR protocol with a simple one-layer MIM model. Thus, lowering the error to 0.47 ppm, while it is somewhat above our target accuracy of 0.3 ppm, is still a major improvement, particularly when considering that empirical treatments cannot be applied to nonstandard residues.

5.6. DNA−RNA Hybrid with Boranophosphate Linkage (PDB ID: 2LAR, BMRB:17535). 2LAR, a DNA/RNA hybrid with boranophosphate linkage, is used to study RNase H1 enzymatic activity.[54] 2LAR is the only system in this study with a uracil base. For 2LAR, an 18-residue system with 594 atoms, the single conformer submitted in the PDB, was used. Figures S9 and S10 show the linearly fitted correlation of MIM1-NMR_{dimer} computed chemical shifts of ^{1}H and ^{13}C with respect to experimental values. As seen in the benchmarking studies and test systems discussed above, panels A−D of Figures S9 and S10 show the NMR results calculated using MIM_{gas}, $MIM^{constraint}_{gas}$, $MIM^{constraint}_{implicit}$and $MIM^{constraint}_{explicit-implicit}$ models for ^{1}H and ^{13}C chemical shifts, respectively. For the ^{1}H NMR chemical shifts calculated using the MIM_{gas} method, we obtained a MAD value of 0.46 ppm with an excellent R^2 (0.96). A slight improvement was observed when the constrained minimized structure was used, with a MAD value of 0.41 ppm (R^2 = 0.96). Adding implicit solvation environment to the gas-phase minimized structure significantly improved the ^{1}H MAD values to 0.33 ppm (R^2 = 0.97) and showed a further slight improvement to 0.30 ppm (R^2 = 0.98) for the $MIM^{constraint}_{explicit-implicit}$ model. For ^{13}C NMR chemical shifts, MAD values for four computational models are (A) 2.98, (B) 2.88, (C) 2.47, and (D) 2.38 ppm with a correlation coefficient of ∼0.99 for each of the models. Interestingly, with a boron heteroatom in the backbone of the 2LAR system, the performance of our MIM-NMR protocol is remarkable with the target accuracy achieved for both ^{1}H and ^{13}C chemical shifts.

6. COMPUTATIONAL SPEEDUP AND EFFICIENCY OF THE MIM-NMR METHOD

For the MIM fragmentation method, there are two key factors to be considered when discussing the computational cost: (a) the actual time taken for the MIM and unfragmented calculations using computational resources for the smaller systems and (b) the scaling of DFT for the unfragmented systems. For standard DFT SCF calculations, the computational cost grows with $O(N^{3-3.5})$ scaling, whereas scaling can be $O(N^4)$ for GIAO-NMR calculations.[55−58] This scaling is a bottleneck in efforts to apply DFT calculations to structures with more than a few hundred atoms, as is typically the case in biomolecular systems like DNA.

As shown in Figure 11, the plot between the total CPU time for MIM1 calculations *vs* the size of the system (*i.e.*, the total number of atoms in the DNA) shows a linear fit with an R^2 value of 0.98. This shows that the scaling of computational cost of our MIM approach with the size of the DNA system is linear. In our MIM fragmentation approach, the number of fragments generated grows linearly with the size of the system but the size of the largest fragment remains constant irrespective of the size of DNA systems, as shown in Table S11. For a small 1K2K DNA system with 338 atoms, we found a speedup of 9×, when comparing the computational cost of the largest fragment to the unfragmented calculation using the *m*PW1PW91/6-311G(d,p) method. However, for a much larger 6XAH system with 1130 atoms (∼3× atom count compared to 1K2K), a dramatic speedup of several 100× is expected, when computational scaling for unfragmented NMR calculation is considered. Moreover, an unfragmented NMR calculation of the full molecule would involve 15728 basis function and is impossible with our existing computational resources. Thus, it is evident that larger DNA systems benefit enormously from the MIM fragmentation method for accurate NMR chemical shift predictions by substantially reducing the computational time.

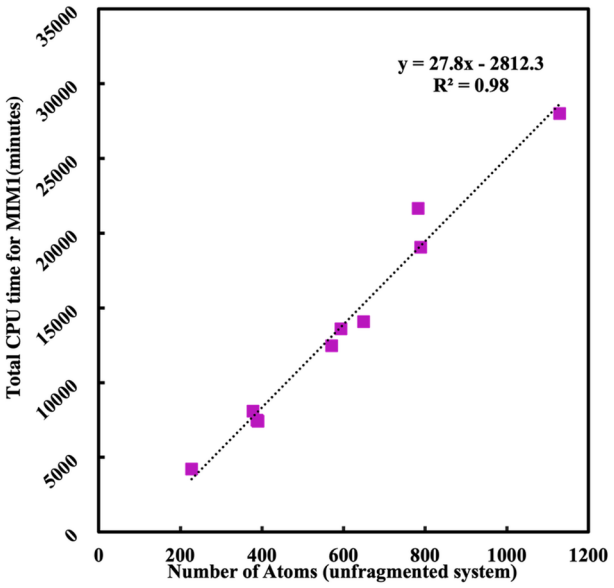


Figure 11. Linear correlation plot between total CPU time for the MIM1 calculations *vs* number of atoms in the unfragmented DNA systems used in this study.

7. CONCLUSIONS AND PERSPECTIVES

The results reported here and in the Supporting Information demonstrate that MIM fragment-based methods can accurately predict the NMR chemical shifts of canonical and non-canonical nucleic acid structures. The fragments are generated by cutting the C−C bonds in the backbone, and a number-based connectivity scheme is used for making the primary subsystems. For DNA systems with multiple conformers, the MIM1-$energy_{trimer}$ method is benchmarked on 10 conformers

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

of the 1SY8 system using the [CAM-B3LYP-D3BJ/6-31+G(d)] method to get the lowest energy structure, with an average error of only 0.80 kcal/mol with respect to unfragmented full molecule calculations. The MIM1-NMR$_{dimer}$ calibration with respect to unfragmented full molecule calculation using the [*m*PW1PW91/6-311G(d,p)] method showed a deviation of 0.06 and 0.11 ppm, respectively, for ^{1}H and ^{13}C nuclei, but the performance with respect to experimental NMR chemical shifts is similar to those from the more expensive trimer MIM1-NMR and MIM2-NMR methods.

Furthermore, we demonstrated that geometry minimizations employing molecular mechanics/semiempirical approaches are effective for obtaining an appropriate starting geometry for doing computations in solution with implicit and explicit−implicit solvation models in our MIM-NMR$_{dimer}$ methodology. One of the major drawbacks of existing force field techniques is the inability to precisely define hydrogen-bonding distances.[59] Clearly, the combination of heavy atom-constrained amber force field minimizations and PM6D3H4 minimizations on the imino and amino protons improved the structure determination of nucleic acids substantially in this study. As a result of our structural investigations, we believe that NMR chemical shifts can be employed as additional useful data for developing enhanced force field methods for better nucleic acid geometry determinations.

For the testing set of seven DNA systems, the ^{1}H MAD target value of 0.30 ppm is achieved overall. Two of the systems showed larger deviations, a slightly larger value of 0.37 ppm for 1N2W, and the largest deviation of 0.47 ppm for the challenging F-substituted 7NBK system. Our overall observations are very similar to performance obtained for proteins from our previous study. The deviations from experiments in the gas-phase calculations can be assigned to the structural flexibility and an inadequate description of the interactions with the solvent. It has been shown that these effects can be better described in proteins by conformational averaging and explicit solvent molecules using molecular dynamics and quantum chemical calculations.[13,22] However, for nucleic acid systems used for this study, the major improvements in the predicted MIM-NMR calculations are obtained from the structural minimizations and the implicit solvation effects. Interestingly, a significant improvement with the explicit−implicit solvation model is observed in two of the smallest nucleic acid systems (1KR8 and 7NBK), which can be attributed to the more exposed amino and imino protons in these systems.

With a few exceptions, there is a high correlation between MIM1-NMR$_{dimer}$ shift predictions and experiments, with coefficients of determination ranging from 0.91 to 0.99 for ^{13}C and ^{1}H for practically all nucleic acids. A satisfactory target accuracy has been attained using our proposed technique with a constraint-minimized structure and an explicit−implicit solvation model: ∼0.2−0.3 ppm for ^{1}H and ∼2.0−3.0 ppm for ^{13}C. More importantly, unlike empirical approaches, our protocol may easily be applied to structures with nonstandard residues, heteroatom substitutions, and side chain mutations. Moreover, our modified one-layer MIM protocol with reasonable-sized fragment calculations bypasses the expensive unfragmented calculations of the full molecule, yields a dramatic speedup for calculations, and shows excellent correlations with experiments. Finally, the current analysis for DNA may not be fully transferable for RNA systems because of their flexible structures. Therefore, developing a new fragmentation scheme for RNA systems is important, and studies are ongoing in our group. Our MIM1-NMR protocol's precision and cost-effectiveness on complex nucleic acid systems bode well for structural predictions of a wide spectrum of large biomolecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00967.

Energy ordering of 10 conformers of 1SY8 DNA duplex; relative energies of all conformers of unfragmented 1SY8 DNA system calculated using CAM-B3LYP-D3BJ/6-31+G(d) level of theory; MIM1$_{dimer}$ and MIM1$_{trimer}$ relative energies of all conformers of 1SY8 DNA system calculated using CAM-B3LYP-D3BJ/6-31+G(d) level of theory; mean absolute deviations; energy ordering of 14 conformers of 1KR8 DNA; energy ordering of 10 conformers of 6XAH DNA; energy ordering of 10 conformers of 7NBK DNA; total CPU timings on all DNA systems; comparison of experimental ^{1}H NMR of 2N5P, 6XAH, 2LIB, 1N2W, 2LFX, and 7NBK with MIM1$_{dimer}$ ^{1}H NMR; comparison of experimental ^{13}C NMR spectrum of 2LIB with neutral residues (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Author

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

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00967

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

(1) Wishart, D. S.; Sykes, B. D.; Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **1991**, *222*, 311−333.

(2) Wishart, D. S.; Watson, M. S.; Boyko, R. F.; Sykes, B. D. Automated 1H and 13C chemical shift prediction using the BioMagResBank. *J. Biomol. NMR* **1997**, *10*, 329.

(3) Wishart, D. S.; Case, D. A. [1]—Use of Chemical Shifts in Macromolecular Structure Determination. In *Methods in Enzymology*; James, T. L.; Dötsch, V.; Schmitz, U., Eds.; Academic Press, 2002; Vol. *338*, pp 3−34.

(4) Li, D. W.; Brüschweiler, R. Certification of Molecular Dynamics Trajectories with NMR Chemical Shifts. *J. Phys. Chem. Lett.* **2010**, *1*, 246.

(5) 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.
(6) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. *J. Phys. Chem. B* **2013**, *117*, 2045−2052.
(7) Dejaegere, A.; Bryce, R. A.; Case, D. A. An Empirical Analysis of Proton Chemical Shifts in Nucleic Acids. *Modeling NMR Chemical Shifts*, ACS Symposium Series; American Chemical Society, 1999; Vol. *732*, pp 194−206.
(8) 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. *J. Biomol. NMR* **2001**, *21*, 11−29.
(9) Li, D.-W.; Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **2012**, *54*, 257−265.
(10) 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*, 13497−13506.
(11) 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*, 125−139.
(12) Cui, Q.; Karplus, M. Molecular Properties from Combined QM/MM Methods. 2. Chemical Shifts in Large Molecules. *J. Phys. Chem. B* **2000**, *104*, 3721.
(13) 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.
(14) 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.
(15) 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.
(16) Hartman, J. D.; 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*, No. 102809.
(17) Lee, A. M.; Bettens, R. P. A. First Principles NMR Calculations by Fragmentation. *J. Phys. Chem. A* **2007**, *111*, 5111.
(18) 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*, 5231−5239.
(19) Frisch, M. J.et al.*Gaussian 16*, M. J.; Gaussian Inc.: Wallingford, CT, 2016.
(20) Unzueta, P. A.; Beran, G. J. O. Polarizable continuum models provide an effective electrostatic embedding model for fragment-based chemical shift prediction in challenging systems. *J. Comput. Chem.* **2020**, *41*, 2251−2265.
(21) Shen, C.; Wang, X.; He, X. Fragment-Based Quantum Mechanical Calculation of Excited-State Properties of Fluorescent RNAs. *Front. Chem.* **2021**, *9*, No. 801062.
(22) 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*, 27781−27799.
(23) Giessner-Prettre, C. Ab-inito Quantum Mechanical Calculations of NMR Chemical Shifts in Nucleic Acids Constituents II Conformational Dependence of the lH and 13C Chemical Shifts in the Ribose. *J. Biomol. Struct. Dyn.* **1985**, *3*, 145−160.
(24) Dejaegere, A. P.; Case, D. A. Density Functional Study of Ribose and Deoxyribose Chemical Shifts. *J. Phys. Chem. A* **1998**, *102*, 5280−5289.
(25) Czernek, J. An ab Initio Study of Hydrogen Bonding Effects on the 15N and 1H Chemical Shielding Tensors in the Watson−Crick Base Pairs. *J. Phys. Chem. A* **2001**, *105*, 1357−1365.
(26) 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*, 1989−1998.
(27) Karadakov, P. B.; Morokuma, K. ONIOM as an efficient tool for calculating NMR chemical shielding constants in large molecules. *Chem. Phys. Lett.* **2000**, *317*, 589.
(28) Hall, K. F.; Vreven, T.; Frisch, M. J.; Bearpark, M. J. Three-Layer ONIOM Studies of the Dark State of Rhodopsin: The Protonation State of Glu181. *J. Mol. Biol.* **2008**, *383*, 106.
(29) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. QM/MM Study of the NMR Spectroscopy of the Retinyl Chromophore in Visual Rhodopsin. *J. Chem. Theory Comput.* **2005**, *1*, 674.
(30) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336.
(31) Jovan Jose, K. V.; Raghavachari, K. Molecules-in-molecules fragment-based method for the evaluation of Raman spectra of large molecules. *Mol. Phys.* **2015**, *113*, 3057.
(32) Jovan Jose, K. V.; Raghavachari, K. Raman Optical Activity Spectra for Large Molecules through Molecules-in-Molecules Fragment-Based Approach. *J. Chem. Theory Comput.* **2016**, *12*, 585.
(33) Jose, K. V. J.; Beckett, D.; Raghavachari, K. Vibrational Circular Dichroism Spectra for Large Molecules through Molecules-in-Molecules Fragment-Based Approach. *J. Chem. Theory Comput.* **2015**, *11*, 4238−4247.
(34) 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*, 1147−1158.
(35) Thapa, B.; Beckett, D.; Jovan Jose, K. V.; Raghavachari, K. Assessment of Fragmentation Strategies for Large Proteins Using the Multilayer Molecules-in-Molecules Approach. *J. Chem. Theory Comput.* **2018**, *14*, 1383−1394.
(36) Thapa, B.; Raghavachari, K. Energy Decomposition Analysis of Protein−Ligand Interactions Using Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Inf. Model.* **2019**, *59*, 3474−3484.
(37) Thapa, B.; Beckett, D.; Erickson, J.; Raghavachari, K. Theoretical Study of Protein−Ligand Interactions Using the Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Theory Comput.* **2018**, *14*, 5143−5155.
(38) 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*, 1839−1862.
(39) Thapa, B.; Raghavachari, K. Accurate pKa Evaluations for Complex Bio-Organic Molecules in Aqueous Media. *J. Chem. Theory Comput.* **2019**, *15*, 6025−6035.
(40) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161.
(41) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(42) 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*, 34−50.
(43) 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*, 907−917.
(44) Padrta, P.; Štefl, R.; Králík, L.; Žídek, L.; Sklenář, V. Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J. Biomol. NMR* **2002**, *24*, 1−14.

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561

(45) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* **2006**, *34*, 564−574.
(46) 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*, 251−268.
(47) 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*, 4515−4534.
(48) 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*, 8576−8587.
(49) 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*, 41−52.
(50) 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*, 425−437.
(51) 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*, 433−442.
(52) Huang, H.; Das, R. S.; Basu, A. K.; Stone, M. P. Structures of (5′S)-8,5′-Cyclo-2′-deoxyguanosine Mismatched with dA or dT. *Chem. Res. Toxicol.* **2012**, *25*, 478−490.
(53) 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*, 7351−7355.
(54) 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*, 3903−3912.
(55) Beran, G. J. O. Calculating Nuclear Magnetic Resonance Chemical Shifts from Density Functional Theory: A Primer. *eMagRes* **2019**, 215−226.
(56) Cole, D. J.; Hine, N. D. M. Applications of large-scale density functional theory in biology. *J. Phys.: Condens. Matter* **2016**, *28*, No. 393001.
(57) Schattenberg, C. J.; Reiter, K.; Weigend, F.; Kaupp, M. An Efficient Coupled-Perturbed Kohn−Sham Implementation of NMR Chemical Shift Computations with Local Hybrid Functionals and Gauge-Including Atomic Orbitals. *J. Chem. Theory Comput.* **2020**, *16*, 931−943.
(58) Stoychev, G. L.; Auer, A. A.; Neese, F. Efficient and Accurate Prediction of Nuclear Magnetic Resonance Shielding Tensors with Double-Hybrid Density Functional Theory. *J. Chem. Theory Comput.* **2018**, *14*, 4756−4771.
(59) Schroeder, K. T.; Skalicky, J.; Greenbaum, N. NMR spectroscopy of RNA duplexes containing pseudouridine in super-cooled water. *RNA* **2005**, *11*, 1012−1016.

https://doi.org/10.1021/acs.jctc.2c00967
J. Chem. Theory Comput. 2023, 19, 544−561