PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 27678

Improving IDP theoretical chemical shift accuracy and efficiency through a combined MD/ADMA/DFT and machine learning approach†

Michael J. Bakker, [a] Arnošt Mládek,[a] Hugo Semrád,[ab] Vojtěch Zapletal[a] and Jana Pavlíková Přecechtělová *[a]

This work extends the multi-scale computational scheme for the quantum mechanics (QM) calculations of Nuclear Magnetic Resonance (NMR) chemical shifts (CSs) in proteins that lack a well-defined 3D structure. The scheme couples the sampling of an intrinsically disordered protein (IDP) by classical molecular dynamics (MD) with protein fragmentation using the adjustable density matrix assembler (ADMA) and density functional theory (DFT) calculations. In contrast to our early investigation on IDPs (Pavlíková Přecechtělová *et al.*, *J. Chem. Theory Comput.*, 2019, **15**, 5642–5658) and the state-of-the art NMR calculations for structured proteins, a partial re-optimization was implemented on the raw MD geometries in vibrational normal mode coordinates to enhance the accuracy of the MD/ADMA/DFT computational scheme. In addition, machine-learning based cluster analysis was performed on the scheme to explore its potential in producing protein structure ensembles (CLUSTER ensembles) that yield accurate CSs at a reduced computational cost. The performance of the cluster-based calculations is validated against results obtained with conventional structural ensembles consisting of MD snapshots extracted from the MD trajectory at regular time intervals (REGULAR ensembles). CS calculations performed with the refined MD/ADMA/DFT framework employed the 6-311++G(d,p) basis set that outperformed IGLO-III calculations with the same density functional approximation (B3LYP) and both explicit and implicit solvation. The partial geometry optimization did not universally improve the agreement of computed CSs with the experiment but substantially decreased errors associated with the ensemble averaging. A CLUSTER ensemble with 50 structures yielded ensemble averages close to those obtained with a REGULAR ensemble consisting of 500 MD frames. The cluster based calculations thus required only a fraction of the computational time.

Received 17th April 2022,
Accepted 9th October 2022

DOI: 10.1039/d2cp01638a

rsc.li/pccp

1 Introduction

In recent years, interest in IDPs has increased among scientists due to their association with many incurable maladies.[1] Alzheimer's has been linked to TAU proteins and α-synuclein possibly contributing to neurodegeneration.[2,3] The aggregation of amyloid proteins in fibrillar aggregates are key events in the propagation of Parkinson's. Characterizing the conformational dynamics involved in these disordered proteins is essential to understanding their functions.[4] Human tyrosine hydroxylase 1 (hTH1) is an IDP regulated by two phosphorylation sites (S19 and S40).[5] Phosphorylation plays a significant role in the function of many disordered proteins as observed in recent investigations.[6–8] Understanding the changes in the phosphorylated IDPs will make great strides toward understanding the functions of these proteins.

Structural characterization of IDPs has been the focus of many scientists in the field, and traditionally, X-ray crystallography has been the method used to understand the 3D density distribution of electrons in proteins. Unfortunately, large and unstable proteins (such as IDPs) are notoriously difficult to crystallize, thus alternative techniques must be implemented such as NMR spectroscopy.[9]

To appropriately simulate the flexibility inherent in biomolecules, theoretical methods usually employ MD trajectories to generate a conformational ensemble.[10–12] There are several advantages in carrying out a computational simulation that experiments cannot provide. An increased computational capacity in the wake of the technology era means such tools have

[a] *Faculty of Pharmacy in Hradec Králové, Charles University, Akademika Heyrovského 1203/8, 500 05 Hradec Králové, Czech Republic. E-mail: precechj@faf.cuni.cz; Tel: +420 495 067 488*

[b] *Department of Chemistry, Faculty of Science, Masaryk University, Kotlářská 267/2, 611 37 Brno, Czech Republic*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2cp01638a

never been more efficient. In lieu of expensive high-end equipment needed for experimental investigations, computations also complement experimental results well, providing insight into quantum influences and atom-level molecular dynamics. A quantum level of theory is particularly useful in these systems as it better represents and accounts for polarization and charges. These influences can help to interpret phosphorylation changes, which by the central dogma of molecular biology helps to understand the protein's purpose and prospective defects.

There are complications with incorporating a quantum level of theory, predominantly the system's size. The emergence and advancement of fragmentation techniques[13–18] throughout the past decade has greatly facilitated biomolecular NMR calculations and made them more readily available than ever. The techniques provide a recipe for the automated construction of molecular clusters that represent fragments of the amino acid sequence in the proteins of interest. This process is facilitated using software making it possible to generate molecular clusters quite expeditiously. This is in large contrast to earlier studies[19–22] that typically designed model systems to represent the desired part(s) of a bio-molecule *de novo* for each system of interest and tailored them to the purpose of a given study. This involved a great deal of effort and time investment in order to first test the prospective model systems, and second to program a tool that constructs the models from the biomolecule Cartesian coordinates.[21] This was especially relevant when multiple biomolecule structures were subject to computations, *i.e.* studies that combine QM calculations with MD simulations.[19,20,23]

Fragmentation techniques are incredibly resource efficient.[24] They provide a quick workaround which allows computations exponentially faster. One such method – the ADMA[25] – shows great promise. ADMA operates by separating the protein molecules into fragments[14] (see a more detailed explanation in the Methodology section). ADMA has already been employed successfully to approximate the effects of distant parts of a given macromolecule in the QM calculations of each fragment when tested upon small oligopeptides.[26] Additionally, it was shown to be capable of computing various molecular properties as well as electron densities relatively accurately.[27] An alternative fragmentation technique for NMR calculations, the automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM), has been proposed by He *et al.*[28–31] The ADMA and AF-QM/MM techniques have become paramount in the field of NMR calculations but other fragmentation techniques exist as well including the fragment molecular orbital method,[32,33] combined fragmentation method,[34] generalized energy-based fragmentation,[35] systematic molecular fragmentation analysis,[36] and molecules-in-molecules fragmentation-based method.[18]

The first applications of fragmentation techniques to structured proteins[13,14] and nucleic acids[37,38] have already been observed. In the context of NMR property calculations, proof-of-concepts were provided by comparing the fragment-based and full system calculations. For NMR CSs, dependence on the level of theory, basis set[39] and solvent models[39] was studied along with the effect of conformational sampling by both classical[40] and *ab initio*[41] MD.

The protein fragmentation has only recently been employed[42] for NMR CS calculations in IDPs. The fragmentation by ADMA was applied to frames generated by a classical MD and the NMR chemical shifts were computed through DFT calculations. The study demonstrates that the accuracy of the MD/ADMA/DFT approach depends on multiple factors, including the quality of MD geometries, the size of structural ensembles employed for DFT calculations and the CS referencing.

Further improvements of fragment-based NMR calculations for IDPs are desired due to their potential in assisting the interpretation of experimental NMR data.[9] IDPs are highly flexible biomolecules and the measured NMR data thus correspond to a structural ensemble rather than a singular structure.[43] The search for a set of structures that matches the experimental CSs is typically carried out through an iterative procedure. Algorithms were devised[44,45] that provide an initial guess of structures. For the proposed structural ensemble, CSs are predicted by fast computer-assisted tools and compared to the experiment. The procedure runs in a loop until self-consistency is reached between the proposed structural ensemble and the experimental NMR data set.[46] So far, the computer-assisted tools have been primarily based on the combination of empirical or QM CS hypersurfaces, semi-classical calculations, neural network models, machine learning, and sequence homology. The main disadvantage of all of these tools is the fact that they are trained exclusively for structured proteins with standard residues only. Efficient and accurate QM calculations of NMR CSs promise more versatility, as they can be applied to an arbitrary system including IDPs with post-translationally modified amino acids such as phosphorylated serine, threonine and tyrosine.

Fragmentation techniques significantly speed up NMR calculations of biomolecules as they facilitate the replacement of a large system by a series of small computationally tractable molecules. Even still, MD/fragmentation/DFT calculations remain demanding. High computational costs stem from the need to include hundreds[23,40,41,47] of MD-generated protein structures to achieve accurate results. In a conventional approach, MD snapshots are progressively added to the structural ensemble with a constant time interval.[21] Since classical MD simulations of IDPs require a long time scale, the number of snapshots rapidly increases. A potential solution to this problem could be the application of a machine learning tool, cluster analysis.[48] Using this tool, large sets of data can be evaluated and ''clustered'' based on their comparative relative properties. The applicability of this tool has already been shown in bioinformatics investigations. It stands to reason that given the immense size of the trajectory at hand, cluster analysis can be employed to seek out and assess repeating or significant geometries from a trajectory. Cluster analysis relies on grouping a set of objects in such a way that the individual data points of the groups are similar based on a variable or property, *e.g.* the root-mean square deviation

(RMSD). Upon completion, a cluster of PDB files is generated that is sorted by the size of the clusters.

While the efficiency of MD/DFT calculations can potentially be improved by cluster analysis, the accuracy is contingent on the quality of MD geometries.[42] Classical MD simulations typically used in MD/DFT studies give only approximate bond lengths and angles, yet a partial molecular geometry re-optimization is usually avoided[21,40,42] for several reasons. First, as the size of the structural ensemble and that of the molecular clusters grows, calculations of ensemble-averaged CSs become intractable when both geometry optimization and NMR calculations are performed. Second, re-optimization by conventional methods becomes impractical,[49] especially for complex molecular clusters as previously described.[42]

2 Objectives

This work pursues the design of an improved MD/ADMA/DFT approach for the calculation of NMR CSs in IDPs. We investigate two principal areas of potential improvement. First, we refine the computational scheme for ^{1}H, ^{13}C, ^{15}N, and ^{31}P NMR calculations in IDPs by implementing a partial geometry re-optimization through the sparsely used normal mode optimization (NMO) technique. Second, we seek to increase the efficiency of the computational scheme by employing cluster analysis in the construction of structural ensembles for fragment-based DFT calculations of CSs. Cluster analysis has not been applied in the context of MD/QM calculations before. We analyze its performance by comparing the ensemble-averaged CSs obtained from structural ensembles devised using the cluster analysis and the conventional approach, respectively. For both methods, we examine the variations of CSs within the structural ensembles, compare the ensemble averages with one another as well as with the experiment while inspecting the standard deviations of the mean at the same time. The extent of agreement or disagreement with the experiment for CSs computed with different model chemistries is explored through the application of multiple referencing schemes. To pursue the objectives detailed above, the doubly phosphorylated disordered part of the hTH1 protein was modelled. CSs calculations were carried out on its two phosphorylated serine sites, pS19 and pS40.

3 Methodology

Molecular dynamics simulation

A classical MD trajectory was simulated for a 53-amino acid fragment representing the unstructured part of the hTH1 regulatory domain. The ESI of ref. 42 includes the initial structure used for the MD simulation. A 100 ns trajectory of the doubly-phosphorylated hTH1 fragment with phosphorylated serines pS19 and pS40 was already obtained in our previous investigation.[42] Herein, the length of the simulation was extended to 1 μs to achieve better sampling of hTH1 over time. The MD simulation was carried out in the Gromacs[50–52] simulation package using the protein Amber99SB-ILDN force field[53,54] and phosphoserine parameters (charge −2) obtained from the work of Homeyer *et al.*[55] The solvation of hTH1 employed a rhombic dodecahedral box and the TIP4P-D water[56] model. A minimum distance of 4 nm was used between the box walls and solute. The system's charge was neutralized by adding Na^+ and Cl^- ions; the concentration of the salt was adjusted to the physiological concentration of 150 mM. The simulations were performed under periodic boundary conditions. More details about the MD setup can be found in a previous work.[42]

Ensemble selection

Two principal methods were applied to the MD trajectory in order to select an ensemble of MD frames for subsequent NMR calculations. First, snapshots were extracted at regular time intervals of 2 ns, which led to a total of 500 snapshots (REGULAR ensemble). Second, in order to reduce the number of calculations, the MD trajectory was subjected to cluster analysis using Gromacs, based on the *k*-nearest neighbor machine learning algorithm.[57] This algorithm relies on the assumption that similar things exist in proximity (Fig. 1).

Snapshots of the MD simulation were clustered around the RMSD values. The choice of a cutoff value is highly dependent on the model and the criteria involved. For the calculations presented in this paper, an RMSD cutoff of 0.45 Å was selected as it represents >30% of the total ensemble in the 50 most populous clusters and >90% in the 500 most populous clusters (see Table 1). Provided in the discussion below are the results of cluster analyses performed with an RMSD cutoff ranging from 0.25 Å to 0.45 Å. RMSD is defined as

$$\mathrm{RMSD}(\mathbf{v},\mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i}^{n} ||v_i - w_i||^2}, \qquad (1)$$

where **v** and **w** are two sets of Cartesian coordinates and n is the number of atoms in the system of interest. Based on the cluster analysis, 50 MD snapshots (CLUSTER ensemble) representing

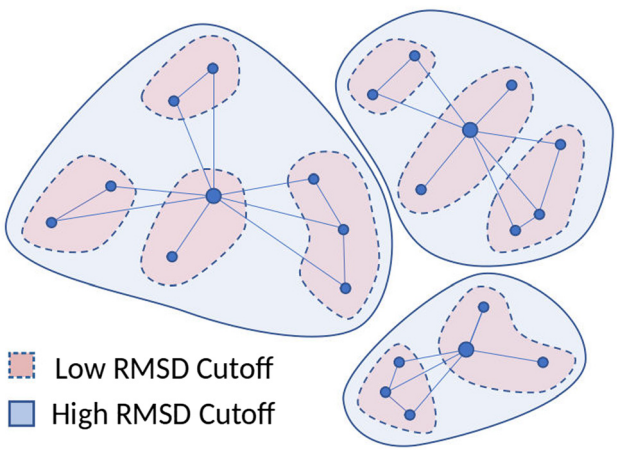

Fig. 1 Graphical representation of the influence of selecting an RMSD cutoff on the cluster analysis output. A higher cutoff value will result in larger clusters, but over representation can be a problem if the individual points in the cluster do not share enough properties.

Table 1 Overview of the cluster analysis. Columns a, c and e give the number of frames encompassed by the 50, 100, and 500 most populous clusters, respectively, and columns b, d and f show the percent representation of these ensembles of the total trajectory (1 μs)

RMSD cutoff/Å	a 50	b % 50	c 100	d % 100	e 500	f % 500
0.25	1373	9.15	2218	14.79	6023	40.15
0.27	1654	11.03	2630	17.53	6963	46.42
0.29	1968	13.12	3061	20.41	7868	52.45
0.31	2256	15.04	3499	23.33	8844	58.96
0.33	2570	17.13	3944	26.29	9760	65.06
0.35	2905	19.37	4184	29.22	10 677	71.18
0.37	3273	21.82	4384	32.39	11 522	76.81
0.39	3580	23.87	5315	35.43	12 385	82.56
0.41	4004	26.69	5853	39.02	13 140	87.59
0.43	4372	29.14	6335	42.23	13 750	91.66
0.45	4760	31.73	6872	45.81	14 226	94.83

the 50 most populous clusters were selected for the NMR calculations.

Protein fragmentation

All snapshots from the two generated structural ensembles (REGULAR and CLUSTER, see above) were subjected to protein fragmentation using ADMA[14] performed with a Python suite of codes. Details of the fragmentation procedure were described previously.[14,40,42] In short, the protein is split into a set of fragments. Two fragments per amino acid are generated, one for the backbone and one for the side chain. The only exceptions to this rule are alanine, glycine, and proline that are represented by one fragment only. In order to account for the effects of the chemical environment, the fragments are expanded with the surrounding protein parts, water molecules and ions. We employ a distance cutoff to determine the size of the surroundings included in the calculation. Bonds broken by the cutoff are saturated with protons. An example of a molecular cluster generated by ADMA is shown in Fig. 2.

Since we focus on NMR calculations of phosphorylated sites in hTH1, we only employ the protein fragments that are

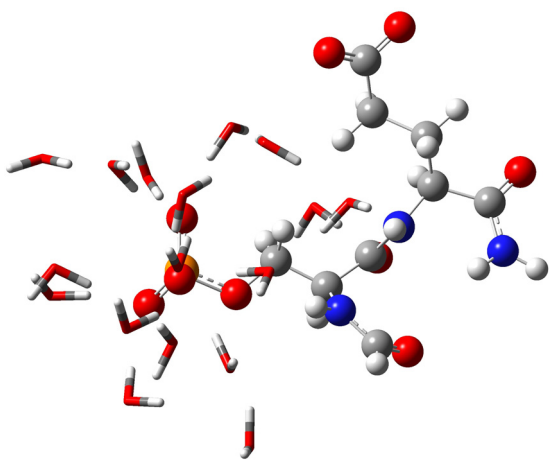

Fig. 2 Example of a molecular cluster produced by the ADMA fragmentation procedure for a pSer residue of hTH1.

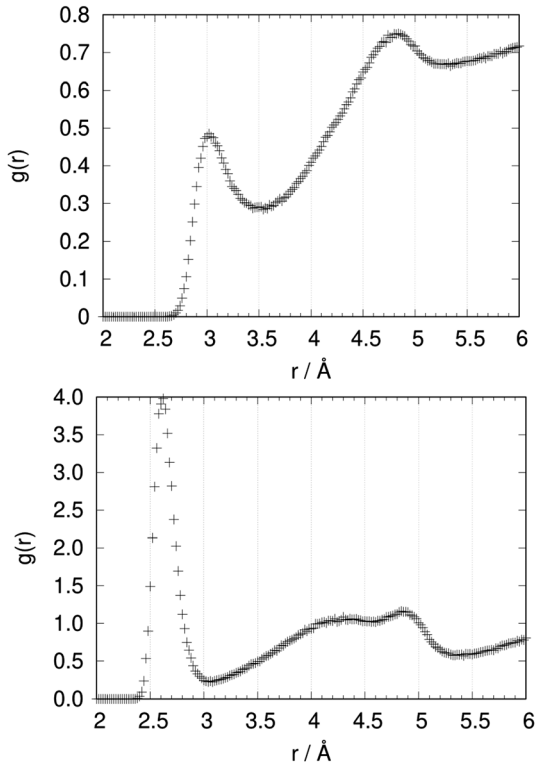

Fig. 3 Radial distribution for the distance of water molecule oxygen from the backbone nitrogen (top) and phosphate oxygen (bottom).

generated for the two phosphoserines (pS). The fragmentation produces one fragment for the backbone and one fragment for the side chain of pS, respectively. Thus, two fragments were made each for pS19 and pS40. The number of geometries employed in further calculations was therefore (N snapshots in the ensemble) × (4 fragments). Prior to the geometry optimization and/or NMR calculation, fragments are embedded into the surroundings of neighboring macromolecule moieties, water molecules and ions. The surroundings within the radius (r) from atoms of the original fragment are included. The radius value was chosen such that the first solvation shell of both the amide and phosphate groups falls within the radius. An inspection of the radial distribution function for the distance between the water and amide protons shows that the end of the first solvation shell is at 3.5 Å (Fig. 3a). At the same time, the boundary between the first and second solvation shell of the phosphate group lies at ~3 Å (Fig. 3b). A radius of r = 3.5 Å for the explicit treatment of the protein surroundings was selected. This results in molecular clusters consisting of up to ~130 atoms.

Geometry optimization

Geometries of molecular clusters constructed from the protein fragments were partially optimized in vibrational normal mode coordinates using the Qgrad program.[49,58] Only low-frequency (<300 cm^{-1}) normal modes corresponding to changes of torsion angles were frozen during the optimization. The partial optimization in normal modes ensures that only bond lengths

and valence angles are re-optimized while the overall geometry of the molecular cluster is preserved. The optimization employs the B3LYP/6-31G(d,p)[59–66] level of theory.

NMR calculations

CS calculations were performed for both optimized and non-optimized molecular geometries using the Gaussian16[67] implementation of the GIAO formalism within the Coupled Perturbed DFT method.[68–72] The calculations employed the B3LYP[59–61] density functional. The choice of the DFT approximation is further commented on in the discussion. For the CLUSTER ensemble calculations, the B3LYP functional was combined with the 6-311++G(d,p)[66,73–76] as well as IGLO-III[77–81] basis set. The REGULAR ensemble calculations used the former basis set only. The B3LYP/6-311++G(d,p) level of theory offers a good compromise between accuracy and computational costs. It previously provided adequate performance within the MD/ADMA/DFT framework,[42] or more broadly, the fragmentation/DFT scheme[13,29,30,47] for the computation of CSs in proteins. IGLO-III was tested as it belongs to the most common basis sets[21,22,77,82] applied in computational NMR spectroscopy. Preliminary calculations also employed the Jensen's pCS-3[83] basis set. However, the calculation revealed high computational costs and convergence difficulties. It was therefore concluded that the application of pCS-3 is not computationally tractable within the MD/DFT framework, where typically large numbers of structures have to be involved in the ensemble averaging of CSs. In order to account for the solvent effects, the explicit solvent within the first solvation shell (see above) was used. The conductor-like implicit solvent model developed within the framework of the polarizable continuum model (CPCM)[84,85] based on the self-consistent reaction field (SCRF) is placed on top of the explicit solvent.

NMR referencing

The chemical shielding σ is converted to the δ-scale using various referencing schemes. ^{1}H and ^{13}C CSs are referenced to tetramethylsilane (TMS) computed at the same level of theory as the atom of interest (X)

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{TMS}}^{\mathrm{calc}} - \sigma_{\mathrm{X}}. \quad (2)$$

Three different referencing schemes were used for the ^{15}N CS calculations: Nref1 references the computed chemical shieldings to the absolute ^{15}N chemical shielding of liquid ammonia employing the calculated ^{15}N chemical shielding of CH_3NH_2[86] used as a secondary standard (the multi-standard approach)[87]

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{CH_3NH_2(gas)}}^{\mathrm{calc}} - \sigma_{\mathrm{X}}^{\mathrm{calc}} + (\sigma_{\mathrm{NH_3(liq)}}^{\mathrm{exp}} - \sigma_{\mathrm{CH_3NH_2(gas)}}^{\mathrm{exp}}), \quad (3)$$

where $\sigma_{\mathrm{CH_3NH_2(gas)}}^{\mathrm{calc}}$ is the chemical shielding of methylamine computed in the gas phase, $\sigma_{\mathrm{NH_3(liq)}}^{\mathrm{exp}} = 244.6$ ppm is the absolute ^{15}N chemical shielding of liquid ammonia at 25 °C,[88] and $\sigma_{\mathrm{CH_3NH_2(gas)}}^{\mathrm{exp}} = 249.5$ ppm is the experimental ^{15}N chemical shielding of methylamine.[89]

Nref2 employs the absolute ^{15}N chemical shielding of liquid ammonia at 25 °C[41,82,88]

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{NH_3(liq)}}^{\mathrm{exp}} - \sigma_{\mathrm{X}}^{\mathrm{calc}}, \quad (4)$$

while Nref3 uses the ^{15}N chemical shielding of NH_3 calculated at the same level of theory as the molecule of interest[14,29,39,40]

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{NH_3(gas)}}^{\mathrm{calc}} - \sigma_{\mathrm{X}}^{\mathrm{calc}}. \quad (5)$$

Similarly, three referencing schemes were also applied for ^{31}P NMR CS calculations. Pref1 references the ^{31}P chemical shifts to 85% H_3PO_4 using the secondary standard PH_3 as proposed by van Wüllen[90]

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{PH_3(gas)}}^{\mathrm{calc}} - \sigma_{\mathrm{X}}^{\mathrm{calc}} + (\sigma_{\mathrm{H_3PO_4(85\%\,solution)}}^{\mathrm{exp}} - \sigma_{\mathrm{PH_3(gas)}}^{\mathrm{exp}}), \quad (6)$$

where $\sigma_{\mathrm{PH_3(gas)}}^{\mathrm{calc}}$ is the chemical shielding of ^{31}P in PH_3 calculated at the same level of theory as the parent molecules constructed from the protein structure, $\sigma_{\mathrm{H_3PO_4(85\%\,solution)}}^{\mathrm{exp}}$ is the absolute experimental chemical shielding of the 85% H_3PO_4 (328.4 ppm)[91] and $\sigma_{\mathrm{PH_3(gas)}}^{\mathrm{exp}}$ is the absolute experimental chemical shielding of PH_3 (594.5 ppm).[91] Pref2 for ^{31}P employs the absolute experimental chemical shielding of the 85% H_3PO_4 (328.4 ppm)[91] as the standard reference

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{85\%H_3PO_4(liq)}}^{\mathrm{exp}} - \sigma_{\mathrm{X}}^{\mathrm{calc}} \quad (7)$$

and Pref3 references the ^{31}P CSs to the chemical shielding of ^{31}P in H_3PO_4 calculated at the same level of theory

$$\delta_{\mathrm{X}}^{\mathrm{calc}} = \sigma_{\mathrm{H_3PO_4(gas)}}^{\mathrm{calc}} - \sigma_{\mathrm{X}}^{\mathrm{calc}}. \quad (8)$$

Ensemble averaging of CSs

The results in this work are reported as ensemble-averaged CSs. For the REGULAR ensemble, the statistical average is calculated as

$$\overline{x} = \sum_{i=1}^{N} x_i / N, \quad (9)$$

where x_i is the value of the CS for a given atom type in the i-th frame of the ensemble while $N = 500$ is the ensemble size. The statistical distribution of the CSs within the ensemble is then expressed as the standard deviation of the sample mean $s_{\overline{x}}$ defined as

$$s_{\overline{x}} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2}. \quad (10)$$

For the CLUSTER ensemble, we employ the following formula for the weighted average

$$\overline{x}_{\mathrm{w}} = \frac{\sum_{i}^{N} w_i x_i}{\sum_{i}^{N} w_i} \quad (11)$$

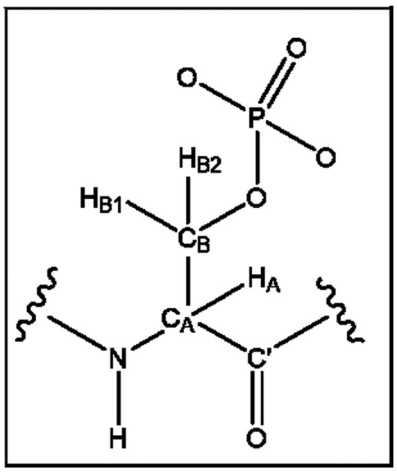

Fig. 4 Labeling of atoms in the phosphorylated serine. For naming of the oxygens, Og is used for the serine and the phosphate oxygens are O1P, O2P and O3P.

where w_i is the weight of the corresponding cluster in the MD trajectory and $N = 50$. The formula for the standard deviation of the mean then reads

$$s_{\bar{x}_w} = \sqrt{\frac{\sum_{i=1}^{N} w_i (x_i - \bar{x}_w)^2}{\frac{M-1}{M}\sum_{i}^{N} w_i}} \quad (12)$$

where M is the number of nonzero weights and N is the number of observations. We take into consideration that the ensembles employed for the statistical averaging are much smaller than the true population of the IDP in question. We therefore report the average CSs as the 95% confidence intervals (CI) given by the formula

$$\mathrm{CI} = \bar{x} \pm z s_{\bar{x}} \quad (13)$$

or

$$\mathrm{CI} = \bar{x} \pm z s_{\bar{x}_w}, \quad (14)$$

respectively. Eqn (13) and (14) use the z-value of 1.96 from the standard normal z-table. The term $zs_{\bar{x}}$ is called the maximum error of estimate (MEE). We compare the computed ensemble-averaged CSs with the experimental CSs (see the ESI of ref. 5 for ^{1}H, ^{13}C, and ^{15}N CSs and ref. 92 for ^{31}P CSs). CSs are calculated for HA, HB1, HB2, H^{N}, CA, CB, C′, N, and P atoms (Fig. 4).

4 Results and discussion

Molecular dynamics

The dynamic behavior of the hTH1 regulatory domain has been studied in detail elsewhere (*e.g.* ref. 93). Here we examine the trajectory properties most relevant to the objectives outlined. First, we inspect the time dependence of the ϕ and ψ angles (Fig. 5) for the 500 frames selected to represent the continuous trajectory. The torsion angles fluctuate rapidly within the equilibrated state, which demonstrates the flexibility of the

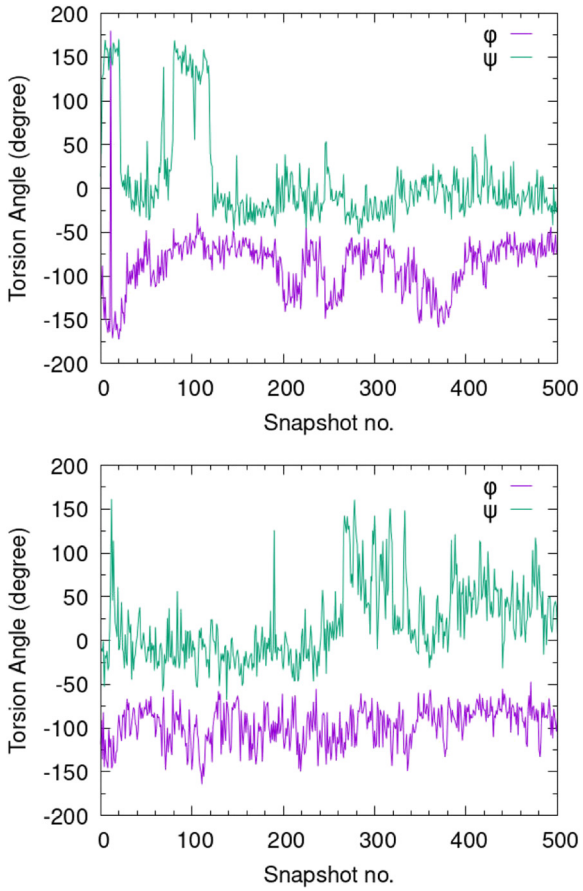

Fig. 5 Time dependence of ϕ(C′(i − 1)–N(i)–CA(i)–C′(i)) and ψ(N(i)–CA(i)–C(i)–N(i + 1)) angles in (a) pS19 and (b) pS40.

system. The plot implies that no specific conformation prevails in the trajectory and that no dominant conformational switches occur. To further support this evidence, we plot the time dependence of the radius of gyration (R_g), see Fig. 6. The initial value of R_g corresponds to an extended state from which the simulation was started. R_g then fluctuates between ~70.6 Å and ~71.5 Å. The fluctuations indicate that the protein is not collapsed in one or a few conformational states and that the conformational space of the protein is sampled as much as possible. This can be even better demonstrated by the analysis of secondary structure propensities. We apply the Define

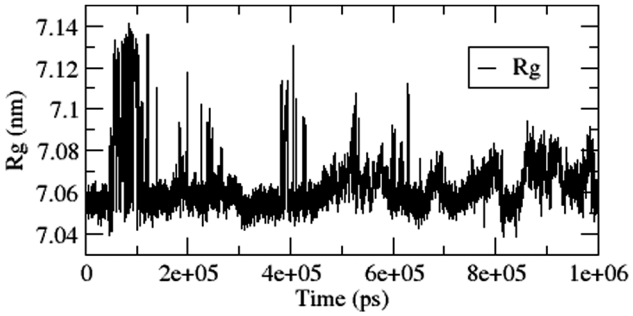

Fig. 6 Time dependence of R_g over the 1 μs MD trajectory of hTH1.

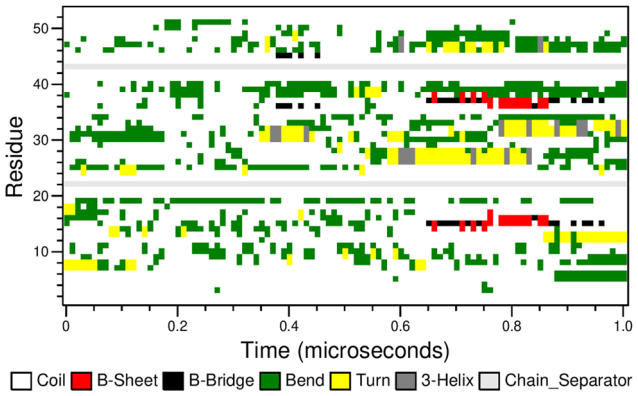

Fig. 7 DSSP analysis from a 1 μs MD trajectory of hTH1.

Secondary Structure of Proteins (DSSP) algorithm which tracks the intra-backbone hydrogen bonds of a protein to assign secondary structures and relay this information over the course of the trajectory. From the DSSP scan (Fig. 7), we were able to determine that very little stable secondary structures exist within the fragment of the protein. Small β-sheets form towards the 0.8 μs mark, but they soon dissipate. Interestingly, the appearance of this secondary structure is noted in the residue quite close to the phosphorylation site, a possible indicator of its role in post-phosphorylation. To inspect how the sampling of the conformational space evolves beyond the 1 μs, we have extended the trajectory up to 2 μs and ran the DSSP calculation again. The outcome of the DSSP analysis for the second microsecond of the simulation time is shown in Fig. S5 (ESI†). It reveals that although new structures do emerge, their lifetime is rather short as none of them survives longer than 0.1 μs. The extended simulation does not show any dramatic changes to the protein behavior.

Cluster analysis

The cluster analysis technique was originally introduced as a method to improve efficiency, *i.e.* to reduce the computational costs. If calculations can be done on a smaller set of frames and still provide similar levels of accuracy, the method is justified. Of the 15 individual cluster analysis runs (Table 1) the clusters obtained from a cut-off value of 0.45 Å were selected. This cluster analysis allows for the representation of approximately one third of the total trajectory. As there is no definitive answer of how to cut a dendogram (cluster analysis is essentially an exploratory approach to facilitating data analysis) the interpretation of an RMSD cut-off is entirely context dependent. Further investigations may attempt a variety of cut-offs and test the values using a silhouette plot or computing the cophenetic correlation, although this preliminary value was computed to describe the efficacy for cluster analysis as a tool to facilitate data analysis. To demonstrate the ergodicity expressed in each cluster ensemble, Ramachandran plots were generated for the two phosphorylated fragments, pS19/pS40 (Fig. 8a–c), and compared to those generated by 500 uniformly selected frames (Fig. 8d). Made evident by Fig. 8d, there are two primary regions

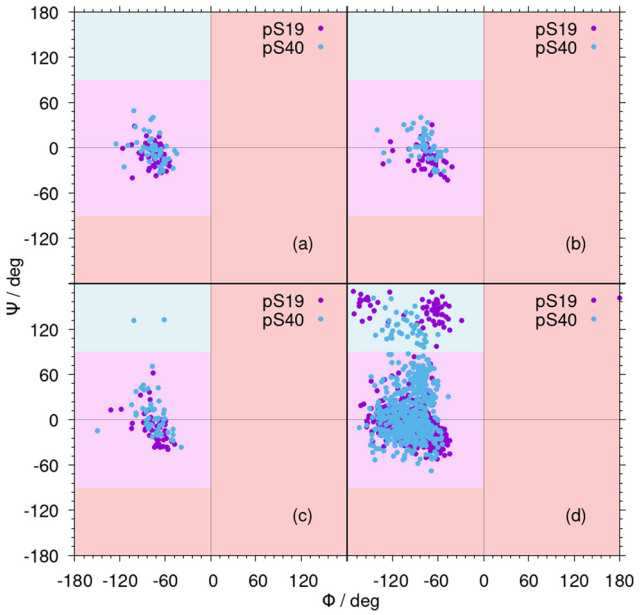

Fig. 8 Ramachandran plots demonstrating the conformational ensembles obtained from running a cluster analysis with RMSD cut-off values at (a) 0.25 Å, (b) 0.35 Å, (c) 0.45 Å and (d) 500 frames selected at regular intervals.

in which this IDP exists in, and by expanding the cluster analysis to include both the magenta and blue regions, it provides better agreement between the clusters and a complete conformational ensemble.

Choice of the DFT functional

The B3LYP functional employed here is a common choice for NMR chemical shift calculations in organic molecules[94,95] as well as proteins[39,41,82,96] although it is known to suffer from various problems.[97–100] Of particular concern is the failure of the functional to appropriately describe van der Waals dispersion interactions.[101] The problem is severe but it is a deficiency that plagues not only B3LYP but virtually all (semi)local[95,102–104] as well as global hybrid functionals.[105] Empirical dispersion corrections[106,107] are available for B3LYP but they do not contribute to the NMR chemical shielding tensor.[18,95]

Dispersion effects can alternatively be factored into the calculations through the use of DFT approximations that include the non-covalent interactions by construction. Double hybrid functionals[108] belong to this category but they are computationally expensive.[109] Implicitly, *i.e.* through fitting to non-local data sets, dispersion interactions are incorporated in the Minnesota functionals.[101,110,111] The performance of two Minnesota M06-family functionals has been tested within the AF-QM/MM fragmentation scheme for protein NMR calculations.[112] As mentioned above, the scheme is an alternative to the ADMA fragmentation employed here. He *et al.*[112] computed the H^N protein CSs using M062X, M06L as well as using B3LYP, B3PW91, mPW1PW91, OB98, and OPBE. The best results were produced by OPBE but the overall performance of all the inspected functionals was very similar. RMSE with

respect to the experimental values ranged from 0.47 ppm (OPBE), through 0.48 (M062X), 0.49 (B3LYP, OB98), 0.50 (M06L, mPW1PW91) to 0.53 ppm (B3PW91). That is, M06-functionals were not significantly better than the other functionals and the RMSE for B3LYP differed only by 0.01 ppm from that for M062X and M06L. It can be expected that the same functionals would perform equivalently if they were combined with ADMA. Explicitly, B3LYP, mPW91PW91, HCTH and VSXC functionals have been previously tested[39] within the ADMA scheme. The mPW1PW91 functional in conjunction with a valence triple-zeta basis set was identified as the best model chemistry while B3LYP was a close runner-up for all nuclei using the same basis set.

It has been suggested[113] recently that the GGAs, meta-GGAs and hybrid-GGAs usually used for NMR calculations should be replaced with long-range corrected DFT approximations. Although the long-range corrections can potentially improve the results, they will also increase the computational costs. The higher demands are not prohibitive, but it is an aspect that becomes increasingly important in combined MD/DFT calculations within which hundreds of calculations must be averaged out.

While in principle many potentially more accurate (than B3LYP) DFT approximations exist, their application in multi-scale MD/DFT calculation may not be fruitful. It has been shown many times before[21,40,47,112,114,115] that the main errors do not stem from the model chemistry. Instead, effects of conformational averaging and solvent within the first solvation shell are the source of major inaccuracies.[41] The relatively lesser influence of the DFT functional on the results is partially a consequence of error cancellations facilitated by the choice of a suitable NMR reference and the referencing scheme as we have thoroughly discussed in our previous publication.[42] Based on all the considerations explained above, we chose B3LYP as a compromise between computational costs and accuracy.

Basis set effect

Table 2 shows that 6-311++G(d,p) and IGLO-III produce almost identical results for the CSs of all protons, which is also

Table 2 Ensemble averaged CSs (in ppm) calculated for the CLUSTER ensemble and REGULAR ensemble, and performed without geometry optimization and with geometry optimization

		CLUSTER[a]			REGULAR[b]	
		NMR[c]	NMR[c]	NMR[c]	OPT[d]	
Res.	Atom	6-311++G(d,p)	IGLO-III	6-311G++(d,p)	6-311++G(d,p)	Exp.[k]
pS19	H^N	7.64 ± 0.23	7.56 ± 0.24	7.26 ± 0.31	8.22 ± 0.28	8.859
	HA	4.85 ± 0.09	4.7 ± 0.11	4.77 ± 0.23	4.75 ± 0.09	4.316
	HB1	4.25 ± 0.08	4.1 ± 0.12	4.21 ± 0.22	4.24 ± 0.06	3.905
	HB2	3.9 ± 0.07	3.81 ± 0.09	3.92 ± 0.21	4.07 ± 0.06	3.905
C		176.21 ± 1.33	157.88 ± 1.40	176.21 ± 2.10	183.21 ± 0.53	174.3
CA		64.8 ± 1.04	53.78 ± 1.01	64.18 ± 1.41	64.34 ± 0.46	66.8
CB		65.13 ± 0.68	53.34 ± 0.82	66.86 ± 1.24	68.29 ± 0.45	59.1
N^e		118.89 ± 1.85	109.31 ± 1.86	119.28 ± 2.62	123.08 ± 1.30	119.7
N^f		132.73 ± 1.85	123.14 ± 1.86	133.12 ± 2.62	136.92 ± 1.30	119.7
N^g		146.16 ± 1.85	136.58 ± 1.86	146.55 ± 2.62	150.36 ± 1.30	119.7
P^h		−34.53 ± 0.98	−87.25 ± 1.21	−33.87 ± 0.38	0.81 ± 0.17	3.76
P^i		−3.39 ± 0.98	−59.74 ± 1.21	−2.73 ± 0.38	31.95 ± 0.17	3.76
P^j		−40.45 ± 0.98	−96.8 ± 1.21	−39.79 ± 0.38	−5.11 ± 0.17	3.76
pS40	H^N	7.06 ± 0.16	6.98 ± 0.16	6.97 ± 0.29	7.96 ± 0.22	8.841
	HA	4.66 ± 0.09	4.64 ± 0.09	4.84 ± 0.23	4.74 ± 0.07	4.238
	HB1	4.22 ± 0.07	4.13 ± 0.1	4.05 ± 0.23	4.81 ± 0.06	3.904
	HB2	3.94 ± 0.11	3.84 ± 0.14	3.67 ± 0.26	4.60 ± 0.06	3.904
C′		177.12 ± 1.24	159.09 ± 1.18	178.22 ± 1.68	182.22 ± 0.53	174.4
CA		62.91 ± 0.88	52.34 ± 0.92	63.76 ± 1.48	63.12 ± 0.49	66.3
CB		64.29 ± 0.85	52.59 ± 0.87	65.07 ± 1.19	70.14 ± 0.36	59.8
N^e		115.40 ± 1.99	106.24 ± 2.11	118.77 ± 2.32	123.80 ± 1.41	117.2
N^f		129.24 ± 1.99	120.08 ± 2.11	132.61 ± 2.32	137.63 ± 1.41	117.2
N^g		142.68 ± 1.99	133.52 ± 2.11	146.05 ± 2.32	151.07 ± 1.41	117.2
P^h		−33.5 ± 1.04	−86.62 ± 1.16	−31.39 ± 0.70	0.45 ± 0.23	4.18
P^i		−2.36 ± 1.04	−59.12 ± 1.16	−0.24 ± 0.70	31.60 ± 0.23	4.18
P^j		−39.42 ± 1.04	−96.18 ± 1.16	−37.30 ± 0.70	−5.46 ± 0.23	4.18

[a] CLUSTER ensemble. [b] REGULAR ensemble, and performed. [c] Without geometry optimization. [d] With geometry optimization. [e] ^{15}N CSs were referenced using Nref1 (eqn (3)). [f] ^{15}N CSs were referenced using Nref2 (eqn (4)). [g] ^{15}N CSs were referenced using Nref3 (eqn (5)). [h] ^{31}P CSs were referenced using Pref1 (eqn (6)). [i] ^{31}P CSs were referenced using Pref2 (eqn (7)). [j] ^{31}P CSs were referenced using Pref3 (eqn (8)). [k] Experimental values of ^{1}H, ^{13}C, and ^{15}N CSs are taken from the ESI of ref. 5, ^{31}P CSs are taken from ref. 92.

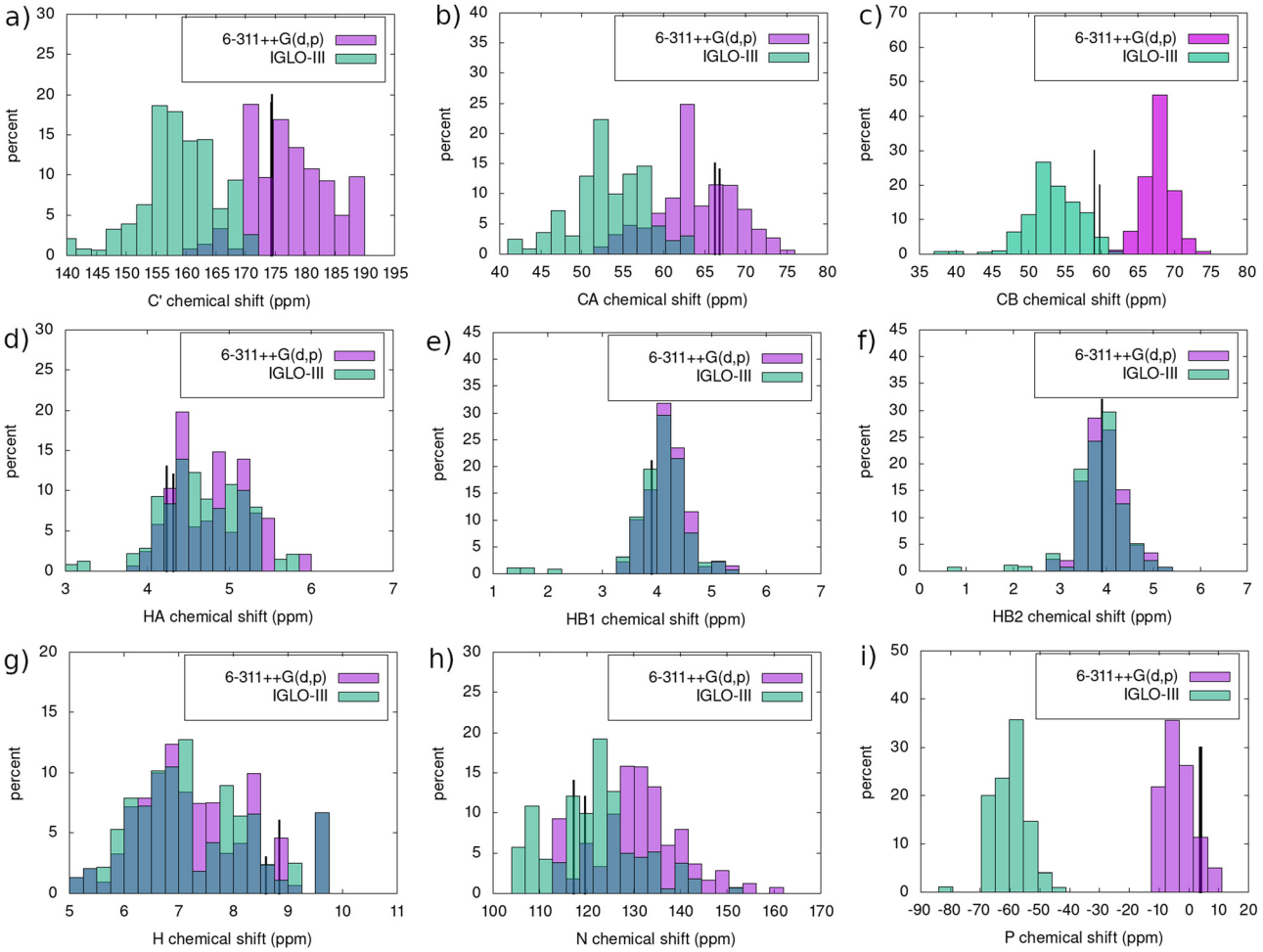

Fig. 9 Comparison of CS histograms of both pS19 and pS40 computed with B3LYP/6-311++G(d,p) and B3LYP/IGLO-III for atoms: (a) C′, (b) CA, (c) CB, (d) HA, (e) HB1, (f) HB2, (g) H^N, (h) NRef1, and (i) PRef1 atoms. For reference, black lines are included to represent the experimentally obtained CSs.

demonstrated by overlapping CS histograms (Fig. 9(d)–(g)). The maximum ^{1}H CS difference between the two basis sets amounts to 0.15 ppm. The basis set has a much more pronounced impact on the ^{13}C CSs. For almost all carbon atoms calculated, the Pople basis set performed notably better in terms of a quantitative agreement with the experiment as can be seen in Table 2 and Fig. 9(a)–(c). Curiously, some calculations performed with 6-311++G(d,p) produce divergent results. For instance, the weighted averages computed with the Pople basis set for C′ and CB (176.21 ppm and 65.13 ppm) in pS19 are overestimated compared to the observed experimental values (174.3 ppm and 59.1 ppm) while the CB CS from 6-311++G(d,p) (64.8 ppm) is underestimated relative to the experiment (66.8 ppm). The CS histograms obtained with the IGLO-III basis are shifted toward smaller CS values for all carbon atom types (Fig. 9(a)–(c)). The CS of ^{15}N largely depends on the choice of the NMR reference (Table 2). The ^{15}N CS in pS19 calculated using Nref1 and 6-311++G(d,p) is less than 1 ppm away from the experimentally obtained data. For IGLO-III, the best agreement with the experiment is achieved when Nref2 is applied. Nref1 previously gave the best results for CS calculations with B3LYP/6-311++G(d,p) model chemistry.[42] The same level of theory best performs in ^{31}P CS calculations when employed along with Pref2 (Table 2 and ref. 42). However, all referencing schemes for ^{31}P yield calculated ^{31}P CSs far from the experiment when the IGLO-III basis set is used. We can thus conclude that B3LYP calculations with 6-311++G(d) exhibit a superior performance compared to calculations with IGLO-III.

Effect of the geometry optimization

We chose NMO as optimization in Cartesian, internal or mixed coordinates is very troublesome when applied to solute–solvent molecular clusters. The reasons for the difficulties were previously explained by Bour *et al.*[49] and we also discussed them in our previous work.[42] NMO, on the contrary, has been shown to perform very well for weakly-bonded systems.[58] For instance, it helped tremendously to accurately predict the vibrational properties of various systems[121–123] with hydrogen bonds including the hydrated phosphate group in nucleic acids.[124] We recall that the normal mode as well as Cartesian coordinates form a complete basis in the same linear vector space.[58] As a result, the two coordinate sets mostly exhibit a

Table 3 Comparison of MD and NMO-optimized bond lengths (in Å)

Bond	MD	NMO	Experiment
Backbone			
CA–HA	1.090	1.095	1.090[a]
CA–C′	1.530	1.534	1.525 ± 0.021[b]
C′–O	1.233	1.242	1.231 ± 0.020[b]
C′–N	n.a.	n.a.	1.329 ± 0.014[b]
N–H	1.010	1.021	1.023 ± 0.006[c]
N–CA	1.483	1.459	1.458 ± 0.019[b]
Side chain			
CA–CB	1.527	1.526	1.530 ± 0.020[b]
CB–HB1	1.090	1.094	1.090[a]
CB–HB2	1.090	1.096	1.090[a]
CB–OG	1.409	1.416	1.433 ± 0.012[d]
P–OG	1.593	1.673	1.621 ± 0.009[d]
P–O1P	1.466	1.536	1.531 ± 0.002[e]
P–O2P	1.462	1.536	1.531 ± 0.002[e]
P–O3P	1.473	1.536	1.531 ± 0.002[e]

[a] Ref. 116. [b] Ref. 117. [c] Ref. 118. [d] Ref. 119. [e] Ref. 120.

similar behavior. In the case of flexible molecules and molecular clusters including hydrogen bonds, NMO proved to perform better than the conventional optimization in Cartesian coordinates.[58]

NMO increases the CA–HA and CB–HB1/HB2 bond lengths by no more than 0.006 Å (Table 3). CSs of aliphatic protons therefore barely change (Fig. 10(d)–(f)) upon optimization. On the contrary, the CSs of the amide protons are affected notably (Fig. 10(g)). For pS19, the H^N CS increases from 7.26 ± 0.31 ppm to 8.22 ± 0.28 ppm and thus better agrees with the experimental value of 8.859 ppm (Table 2). The H^N CS change is likely largely caused by the shortening of the $H^N\cdots H_w$ hydrogen bond. It has been recognized previously that the H^N CS linearly increases as the hydrogen bond becomes smaller.[40] The CA CSs seem very little influenced by the optimization (Table 2). This suggests that the local environment of CA is well-parameterized by the force field. This can be illustrated by CA–HA, CA–CB, and CA–C′ bond lengths that change by no more than 0.005 Å (Table 3) due to NMO. Just the CA–N bond length undergoes a larger modification by 0.024 Å. In contrast to CA, the CS of CB changes significantly by up to ∼7 ppm. NMO corrects the local geometry of the neighboring phosphate group to a large extent (see below), which results in a notable downfield shift of CB.

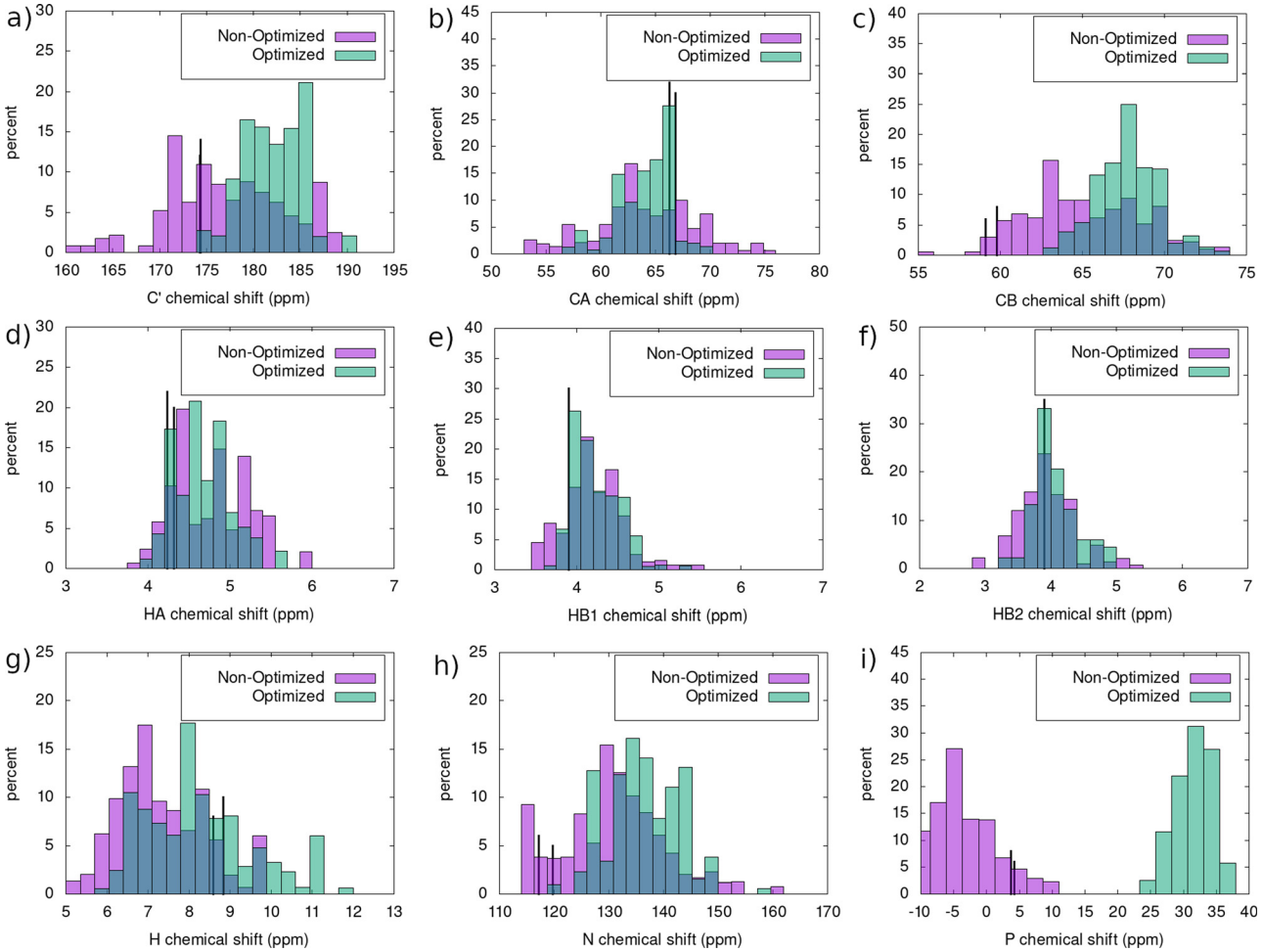

Fig. 10 Histograms of CSs of the combined pS19 and pS40 fragments obtained for C′ (a), CA (b), CB (c), HA (d), HB1 (e), HB2 (f), H^N (g), NRef1 (h), and PRef1 (i) using the non-optimized and optimized geometries of the protein fragments. For reference, black lines are included to represent the experimentally obtained CSs.

The ensemble averaged CS of C′ grows from 176.21 ppm to 183.21 ppm (pS19) and from 178.22 ppm to 182.22 ppm (pS40). It thus deviates substantially from the experimental values of 174.3 ppm and 174.4 ppm, respectively as seen in Fig. 10(a). We attribute the deviation observed after the geometry optimization to the rectification of the carbonyl double bond length and the CO$\cdots$H–O_w hydrogen bond geometry. The ^{15}N CSs obtained for the optimized molecular geometries are larger by 4–5 ppm compared to the CSs obtained from calculations with raw MD geometries.

The most profound effect has been observed for the ^{31}P nuclei in the phosphoserine side chains, where the optimization shifts the CS histograms by more than 30 ppm (Fig. 10(i)). This is caused by NMO that alters the geometry of the solvated phosphate group. For instance, the P–OG bond length increases by 0.08 Å upon optimization. The P–O1P/O2P/O3P bond length increases by up to 0.074 Å (Table 3). The MD force field also provides P–O$\cdots$$H_w$ hydrogen bond lengths that are both significantly shorter as well as longer than the optimized values.[21] Table 2 demonstrates that a quantitative agreement with the experiment is for ^{31}P CSs achieved with a different referencing scheme when geometry optimization is employed. While Pref2 performed best for NMR calculations with non-optimized geometries, Pref1 is the superior reference for calculations based on optimized protein fragments. This conclusion is in line with findings of our previous study[42] that show how various referencing schemes benefit from systematic error cancellations.

The inclusion of the geometry optimization affects the compensation of various systematic errors. This is nicely illustrated by the fact that a different referencing scheme for ^{31}P CSs has to be used after the optimization to achieve the best quantitative agreement with the experiment. The extent to which geometry optimization influences the efficiency of error compensation differs among the atom types. It is probably the reason why we do not observe a universal improvement of computed CSs in comparison with experimental data for all atoms when NMO is introduced. Nevertheless, we recall that the present work only assesses a quantitative agreement with the experiment. We assume that geometry optimization would likely improve a qualitative agreement of computed and experimental CS trends. However, the validation of qualitative trends requires the computation of CSs for the full protein sequence. This task is computationally extremely demanding and is thus beyond the scope of the current paper. To sum up, geometry optimization improves the MEEs for all CSs among both the pS19 and pS40 fragments.

Effect of the ensemble type

The small size CLUSTER ensemble provides CS estimates that are very close to the ensemble averages obtained with the much larger REGULAR ensemble, see column no. 3 and 5 of Table 2. The average ^{1}H CSs differ by no more than 0.4 ppm between the ensembles. The largest deviation observed for the ^{13}C CS amounts to 1.7 ppm. The ^{15}N CSs are almost identical for pS19 while a difference of $\sim$4 ppm is found for pS40. ^{31}P CSs differ

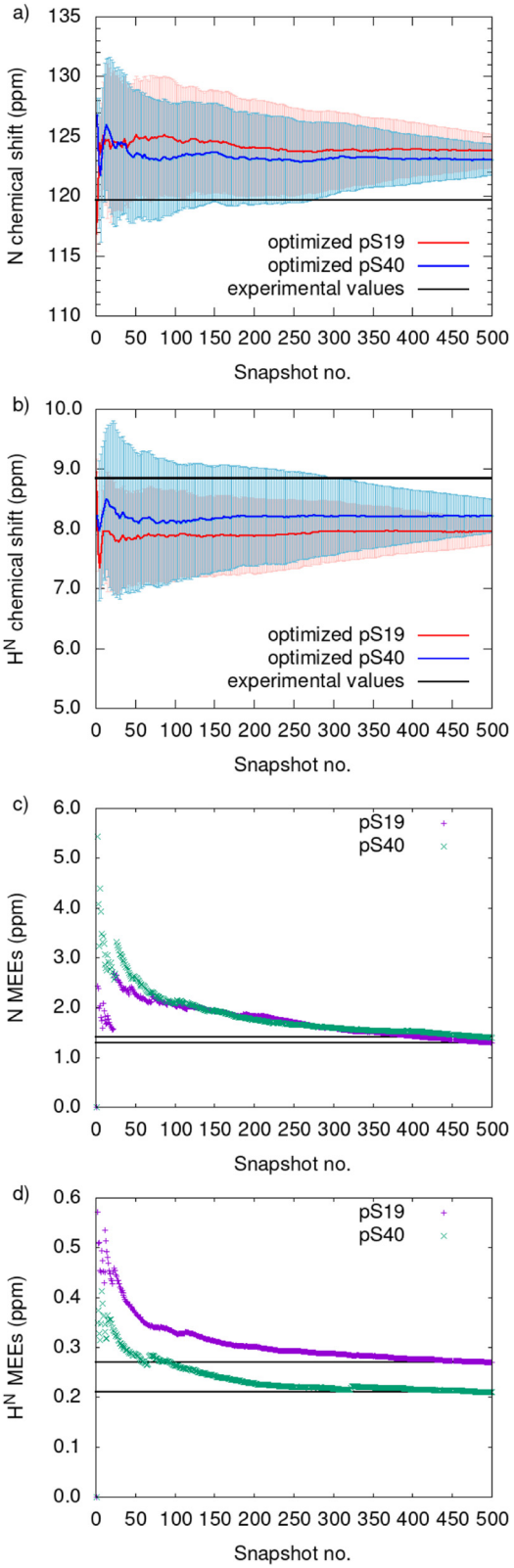

Fig. 11 Dependence of the average (a) N and (b) H^N CSs and the corresponding MEEs (represented by error bars) on the number of frames in the REGULAR ensemble computed with geometry optimization. Plots (c and d) show the decrease of MEEs for the N and H^N CSs, respectively, observed for the optimized molecular geometries.

by at most ~0.7 ppm (pS19) and ~2 ppm (pS40) between the two ensembles. The MMEs computed for the CLUSTER ensemble are in many cases smaller than for the REGULAR ensemble, due in part to an increased representation from the cluster analysis. Fig. S1 (ESI†) reveals that histograms calculated with the REGULAR ensemble typically display a broader distribution of CSs while the CLUSTER ensemble leads to narrow profiles with much higher percentage occurrences (Fig. S1(d)–(i), ESI†). The REGULAR ensemble gives an MEE of 2.10 ppm for the C′ atom in pS19 while the CLUSTER ensemble leads to an MEE of 1.33 ppm only. MEE calculations for CA in pS19 yield the value 1.04 ppm (CLUSTER) *vs.* 1.41 ppm (REGULAR). The MEE of the N atoms in pS19 increases from 1.85 ppm to 2.62 ppm when replacing the CLUSTER ensemble by the REGULAR ensemble. The MEEs can further decrease by performing geometry optimization. Fig. 11 displays the running averages for the N and H^N CSs while running averages for other atom types are shown in Fig. S2 and S3 (ESI†). The running average as well as its MEE is compared for the non-optimized and optimized REGULAR ensemble. The running averages converge sufficiently within the 500 frames set and the MEEs decrease with time. The MEE for the optimized REGULAR ensemble is smaller than that for the non-optimized REGULAR ensemble (Table 2), *e.g.* the MEE for the average ^{15}N CS decreases from 2.62 ppm to 1.30 ppm (pS19) and from 2.32 ppm to 1.41 ppm (pS40) upon optimization. Similarly, the MEE drops from 0.38 ppm to 0.17 ppm for P and from 2.10 ppm to 0.53 ppm for C′ in pS19.

5 Conclusions

Through the analysis of the data, an overall general preference for the Pople, 6-311++G(d,p) was found over the IGLO-III basis set. Each of the aliphatic H atoms' CS averages had a deviation from the experiment of around 0.5 ppm and an MEE of about 0.2 ppm. The C atoms had much greater agreement with the non-optimized Pople, deviating by only 2 ppm in pS19, and about 3 ppm in pS40 among CA and C′. Regardless of the referencing scheme, 6-311++G(d,p) performed paramount among the P CSs, with Pref2 giving the most accuracy. The precision of 6-311++G(d,p) is noticeably better as observed in the MEEs with minimal exceptions.

Interestingly, the results from the geometry optimization produced varied degrees of success. The H^N CSs improved approximately by 1 ppm upon optimization in both p19 and p40, deviating from the experimental value by between 0.6 and 0.9 ppm. Among the HA atoms there was a minimal change in the accuracy between the non-optimized and optimized values (increased deviation of 0.02 ppm in pS19 and a decrease of 0.1 ppm in pS40 from the experiment). The remaining non-polar H CSs (HB1 and HB2) showed minute changes for pS19, but a loss of accuracy by about 0.8 ppm to 1 ppm upon optimization in pS40. There was an improvement in the MEEs calculated for each of the H atoms, a trend which applies to nearly all atoms computed. Among the carbon atoms, an obvious favoritism appears for the non-optimized calculations. The CSs for C′ atoms depreciated in accuracy by approximately 5–7 ppm. CA CSs were the only exception to this, showing small or no improvement between the non-optimized and optimized results. Once more, regardless of the referencing scheme, the N CSs show a preference for non-optimized by approximately 4–5 ppm with Nref1. Finally, the P CSs, being the most sensitive to geometry changes, deviated nearly 30 ppm between the optimized and non-optimized CSs in both pS19 and pS40. The conclusion from this investigation is that while optimization may improve or depreciate the accuracy of the sample, nearly across the board, all values of MEEs showed a reduction.

The success of the cluster analysis can be influenced by the adjustment of specific parameters, as seen in the Ramachandran plots. We can demonstrate that these parameters play a role in appropriately representing the conformational phase space in a trajectory. We have also validated the cluster analysis method through CS computations. Among all the atoms in pS19, there was an improvement in precision even including the particularly sensitive P CSs. Not only did the data heavily support its ability to create ensembles that preserve accuracy and relative precision, but it also drastically reduces the computational cost of the combined MD/ADMA/DFT framework. In fact, due to the increased representation of frames in the cluster analysis, the MEEs differ greatly between the values obtained from CLUSTER and REGULAR ensembles.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Dr Thomas E. Exner (Eberhard Karls Universität Tübingen, Tübingen, Germany) for providing the Python suite of codes for the ADMA workflow procedure and the related technical assistance. We are also grateful to Dr Valery Andrushchenko (Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czech Republic) for helping us to set up and implement the normal mode geometry optimization procedure. The research was financed by the Czech Science Foundation Grant 19-14886Y. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ ID: 90140 and e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. The research was implemented in the MetaCentrum and IT4I supercomputing facilities.

Notes and references

1 V. Vacic and L. M. Iakoucheva, *Mol. BioSyst.*, 2012, **8**, 27–32.
2 H. He, Y. Liu, Y. Sun and F. Ding, *J. Chem. Inf. Model.*, 2021, **61**, 2916–2925.

3 T. J. Fiolek, C. L. Magyar, T. J. Wall, S. B. Davies, M. V. Campbell, C. J. Savich, J. J. Tepe and R. A. Mosey, *Bioorg. Med. Chem. Lett.*, 2021, **36**, 127821.
4 M. Hashemi and Y. L. Lyubchenko, *Methods*, 2022, **197**, 89–96.
5 P. Louša, H. Nedozrálová, E. Župa, J. Nováček and J. Hritz, *Biophys. Chem.*, 2017, **223**, 25–29.
6 T. Wei, H. Liu, B. Chu, P. Blasco, Z. Liu, R. Tian, X. Li and X. Li, *Cell Chem. Biol.*, 2021, **28**, 722–732.
7 J. C. Dunlap and J. J. Loros, *Mol. Cell*, 2018, **69**, 165–168.
8 P. Kulkarni, M. K. Jolly, D. Jia, S. M. Mooney, A. Bhargava, L. T. Kagohara, Y. Chen, P. Hao, Y. He, R. W. Veltri, A. Grishaev, K. Weninger, H. Levine and J. Orban, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E2644–E2653.
9 S. Kosol, S. Contreras-Martos, C. Cedeño and P. Tompa, *Molecules*, 2013, **18**, 10802–10828.
10 A. Watson, C. Simmermaker, E. Aung, S. Do, S. Hackbusch and A. H. Franz, *Carbohydr. Res.*, 2021, **503**, 108296.
11 C. Reinknecht, A. Riga, J. Rivera and D. A. Snyder, *Molecules*, 2021, **26**, 1420–3049.
12 A. Philips and J. Autschbach, *J. Chem. Theory Comput.*, 2020, **16**, 5835–5844.
13 X. He, T. Zhu, X. Wang, J. Liu and J. Z. H. Zhang, *Acc. Chem. Res.*, 2014, **47**, 2748–2757.
14 A. Frank, I. Onila, H. M. Möller and T. E. Exner, *Proteins*, 2011, **79**, 2189–2202.
15 J. Swails, T. Zhu, X. He and D. A. Case, *J. Biomol. NMR*, 2015, **63**, 125–139.
16 K. J. Jose and K. Raghavachari, *J. Chem. Theory Comput.*, 2017, **13**, 1147–1158.
17 X. Jin, T. Zhu, J. Z. H. Zhang and X. He, *Front. Chem.*, 2018, **6**, 150.
18 S. K. Chandy, B. Thapa and K. Raghavachari, *Phys. Chem. Chem. Phys.*, 2020, **22**, 27781–27799.
19 C. Scheurer, N. R. Skrynnikov, S. F. Lienin, S. K. Straus, R. Brüschweiler and R. R. Ernst, *J. Am. Chem. Soc.*, 1999, **121**, 4242–4251.
20 D. A. Case, C. Scheurer and R. Brüschweiler, *J. Am. Chem. Soc.*, 2000, **122**, 10390–10397.
21 J. Přecechtělová, P. Novák, M. L. Munzarová, M. Kaupp and V. Sklenář, *J. Am. Chem. Soc.*, 2010, **132**, 17139–17148.
22 L. Benda, B. Schneider and V. Sychrovský, *J. Phys. Chem. A*, 2011, **115**, 2385–2395.
23 J. Přecechtělová, M. L. Munzarová, J. Vaara, J. Novotný, M. Dračínský and V. Sklenář, *J. Chem. Theory Comput.*, 2013, **9**, 1641–1656.
24 M. S. Gordon, D. G. Fedorov, S. R. Pruitt and L. V. Slipchenko, *Chem. Rev.*, 2012, **112**, 632–672.
25 T. E. Exner and P. G. Mezey, *J. Comput. Chem.*, 2003, **24**, 1980–1986.
26 T. E. Exner and P. G. Mezey, *J. Phys. Chem. A*, 2004, **108**, 4301–4309.
27 P. G. Mezey, *J. Comput. Methods Sci. Eng.*, 2001, **1**, 99–105.
28 X. He, B. Wang and K. M. Merz, *J. Phys. Chem. B*, 2009, **113**, 10380–10388.
29 T. Zhu, X. He and J. Z. H. Zhang, *Phys. Chem. Chem. Phys.*, 2012, **14**, 7837–7845.
30 T. Zhu, J. Z. H. Zhang and X. He, *J. Chem. Theory Comput.*, 2013, **9**, 2104–2114.
31 T. Zhu, J. Z. H. Zhang and X. He, Quantum Calculation of Protein NMR Chemical Shifts Based on the Automated Fragmentation Method, in *Advance in Structural Bioinformatics. Advances in Experimental Medicine and Biology*, ed. D. Wei, Q. Xu, T. Zhao and H. Dai, Springer, Dordrecht, 2015, vol. 827.
32 Q. Gao, S. Yokojima, T. Kohno, T. Ishida, D. G. Fedorov, K. Kitaura, M. Fujihira and S. Nakamura, *Chem. Phys. Lett.*, 2007, **445**, 331–339.
33 Q. Gao, S. Yokojima, D. G. Fedorov, K. Kitaura, M. Sakurai and S. Nakamura, *J. Chem. Theory Comput.*, 2010, **6**, 1428–1444.
34 H.-J. Tan and R. P. A. Bettens, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7541–7547.
35 D. Zhao, R. Song, W. Li, J. Ma, H. Dong and S. Li, *J. Chem. Theory Comput.*, 2017, **13**, 5231–5239.
36 R. Kobayashi, R. D. Amos, D. M. Reid and M. A. Collins, *J. Phys. Chem. A*, 2018, **122**, 9135–9141.
37 A. Victora, H. M. Möller and T. E. Exner, *Nucl. Acids Res.*, 2014, **42**, e173.
38 D. A. Case, *Curr. Opin. Struc. Biol.*, 2013, **23**, 172–176.
39 A. Frank, H. M. Möller and T. E. Exner, *J. Chem. Theory Comput.*, 2012, **8**, 1480–1492.
40 T. E. Exner, A. Frank, I. Onila and H. M. Möller, *J. Chem. Theory Comput.*, 2012, **8**, 4818–4827.
41 M. Dračínský, H. M. Möller and T. E. Exner, *J. Chem. Theory Comput.*, 2013, **9**, 3806–3815.
42 J. Pavlíková Přecechtělová, A. Mládek, V. Zapletal and J. Hritz, *J. Chem. Theory Comput.*, 2019, **15**, 5642–5658.
43 R. Schneider, J. Rong Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. R. Jensen and M. Blackledge, *Mol. BioSyst.*, 2012, **8**, 58–68.
44 C. K. Fisher and C. M. Stultz, *Curr. Opin. Struct. Biol.*, 2011, **21**, 426–431.
45 M. Bonomi, G. T. Heller, C. Camilloni and M. Vendruscolo, *Curr. Opin. Struct. Biol.*, 2017, **42**, 106–116.
46 F. M. Ytreberg, W. Borcherds, H. Wu and G. W. Daughdrill, *Intrinsically Disord. Proteins*, 2015, **3**, e984565.
47 J. Fukal, O. Páv, M. Buděšínský, J. Šebera and V. Sychrovský, *Phys. Chem. Chem. Phys.*, 2017, **19**, 31830–31841.
48 T. M. Abramyan, J. A. Snyder, A. A. Thyparambil, S. J. Stuart and R. A. Latour, *J. Comput. Chem.*, 2016, **37**, 1973–1982.
49 P. Bouř and T. A. Keiderling, *J. Chem. Phys.*, 2002, **117**, 4126–4132.
50 H. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.
51 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
52 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.
53 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950–1958.
54 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.

55 N. Homeyer, A. H. C. Horn, H. Lanig and H. Sticht, *J. Mol. Model.*, 2006, **12**, 281–289.
56 S. Piana, A. G. Donchev, P. Robustelli and D. E. Shaw, *J. Phys. Chem. B*, 2015, **119**, 5113–5123.
57 A. Yosipof and H. Senderowitz, *J. Comput. Chem.*, 2015, **36**, 493–506.
58 P. Bouř, *Collect. Czechoslov. Chem. Commun.*, 2005, **70**, 1315–1340.
59 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
60 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
61 B. Miehlich, A. Savin, H. Stoll and H. Preuss, *Chem. Phys. Lett.*, 1989, **157**, 200–206.
62 R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724–728.
63 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.
64 M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *J. Am. Chem. Soc.*, 1982, **104**, 2797–2803.
65 P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222.
66 M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, *J. Chem. Phys.*, 1982, **77**, 3654–3665.
67 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09, Revision A.03*, Gaussian, Inc., Wallingford CT, 2016.
68 F. London, *J. Phys. Radium*, 1937, **8**, 397–409.
69 R. McWeeny, *Phys. Rev.*, 1962, **126**, 1028–1034.
70 R. Ditchfield, *Mol. Phys.*, 1974, **27**, 789–807.
71 K. Wolinski, J. F. Hinton and P. Pulay, *J. Am. Chem. Soc.*, 1990, **112**, 8251–8260.
72 J. R. Cheeseman, G. W. Trucks, T. A. Keith and M. J. Frisch, *J. Chem. Phys.*, 1996, **104**, 5497–5509.
73 T. Clark, J. Chandrasekhar, G. W. Spitznagel and P. V. R. Schleyer, *J. Comput. Chem.*, 1983, **4**, 294–301.
74 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
75 A. D. McLean and G. S. Chandler, *J. Chem. Phys.*, 1980, **72**, 5639–5648.
76 G. W. Spitznagel, T. Clark, P. V. R. Schleyer and W. J. Hehre, *J. Comput. Chem.*, 1987, **8**, 1109–1116.
77 A. E. A. Fouda and N. A. Besley, *Theor. Chem. Acc.*, 2017, **137**, 6.
78 W. Kutzelnigg, U. Fleischer and M. Schindler, in *Deuterium and Shift Calculation*, ed. M. L. Martin and G. J. Martin, Springer, Berlin Heidelberg, 1991, vol. 16 of NMR Basic Principles and Progress, pp. 165–262.
79 D. Feller, *J. Comput. Chem.*, 1996, **17**, 1571–1586.
80 K. L. Schuchardt, B. T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li and T. L. Windus, *J. Chem. Inf. Model.*, 2007, **47**, 1045–1052.
81 B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibsom and T. L. Windus, *J. Chem. Inf. Model.*, 2019, **59**, 4814–4820.
82 J. Vícha, M. Babinský, G. Demo, O. Otrusinová, S. Jansen, B. Pekárová, L. Žídek and M. L. Munzarová, *Proteins*, 2016, **84**, 686–699.
83 F. Jensen, *J. Chem. Theory Comput.*, 2015, **11**, 132–138.
84 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
85 M. Cossi, N. Rega, G. Scalmani and V. Barone, *J. Comput. Chem.*, 2003, **24**, 669–681.
86 L. Cai, D. Fushman and D. S. Kosov, *J. Biomol. NMR*, 2008, **41**, 77–88.
87 A. M. Sarotti and S. C. Pellegrinet, *J. Org. Chem.*, 2009, **74**, 7254–7260.
88 C. J. Jameson, A. K. Jameson, D. Oppusunggu, S. Wille, P. M. Burrell and J. Mason, *J. Chem. Phys.*, 1981, **74**, 81–88.
89 C. J. Cramer, *Essentials Of Computational Chemistry: Theories and Models*, Wiley, 2002, New York, 2nd edn, 1961, p. 347.
90 C. van Wüllen, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2137–2144.
91 C. J. Jameson, A. D. Dios and A. K. Jameson, *Chem. Phys. Lett.*, 1990, **167**, 575–582.
92 J. Hritz, I.-J. L. Byeon, T. Krzysiak, A. Martinez, V. Sklenář and A. M. Gronenborn, *Biophys. J.*, 2014, **107**, 2185–2194.
93 V. Zapletal, A. Mládek, K. Melková, P. Louša, E. Nomilner, Z. Jaseňáková, V. Kubáň, M. Makovická, A. Laníková, L. Žídek and J. Hritz, *Biophys. J.*, 2020, **118**, 1621–1633.
94 M. T. de Oliveira, J. M. A. Alves, A. A. C. Braga, D. J. D. Wilson and C. A. Barboza, *J. Chem. Theory Comput.*, 2021, **17**, 6876–6885.
95 G. J. Beran, *Calculating Nuclear Magnetic Resonance Chemical Shifts from Density Functional Theory: A Primer*, John Wiley Sons, Ltd, 2019, pp. 215–226.
96 T. Zhu, X. He and J. Z. H. Zhang, *Phys. Chem. Chem. Phys.*, 2012, **14**, 7837–7845.
97 E. Torres and G. A. DiLabio, *J. Phys. Chem. Lett.*, 2012, **3**, 1738–1744.
98 H. Kruse, L. Goerigk and S. Grimme, *J. Org. Chem.*, 2012, **77**, 10824–10834.
99 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.
100 L. Goerigk and S. Grimme, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6670–6688.
101 Y. Zhao and D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 157–167.

102 S. Kristyán and P. Pulay, *Chem. Phys. Lett.*, 1994, **229**, 175–180.
103 A. D. Becke, A. A. Arabi and F. O. Kannemann, *Can. J. Chem.*, 2010, **88**, 1057–1062.
104 C.-W. Wang, K. Hui and J.-D. Chai, *J. Chem. Phys.*, 2016, **145**, 204101.
105 Y.-S. Lin, G.-D. Li, S.-P. Mao and J.-D. Chai, *J. Chem. Theory Comput.*, 2013, **9**, 263–272.
106 S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 211–228.
107 L. Goerigk, *Non-Covalent Interactions in Quantum Chemistry and Physics*, ed. A. Otero de la Roza and G. A. DiLabio, Elsevier, 2017, pp. 195–219.
108 G. L. Stoychev, A. A. Auer and F. Neese, *J. Chem. Theory Comput.*, 2018, **14**, 4756–4771.
109 J. Klimeš and A. Michaelides, *J. Chem. Phys.*, 2012, **137**, 120901.
110 L. Goerigk, *J. Phys. Chem. Lett.*, 2015, **6**, 3891–3896.
111 Y. Zhao and D. G. Truhlar, *Chem. Phys. Lett.*, 2011, **502**, 1–13.
112 T. Zhu, J. Z. H. Zhang and X. He, *J. Chem. Theory Comput.*, 2013, **9**, 2104–2114.
113 M. A. Iron, *J. Chem. Theory Comput.*, 2017, **13**, 5798–5819.
114 F. A. A. Mulder and M. Filatov, *Chem. Soc. Rev.*, 2010, **39**, 578–590.
115 C. R. Jacob and L. Visscher, *J. Chem. Phys.*, 2006, **125**, 194104.
116 G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga, *J. Phys. Chem.*, 1992, **96**, 6472–6484.
117 R. A. Engh and R. Huber, *Acta Crystallogr. A*, 1991, **47**, 392–400.
118 L. Yao, B. Vögeli, J. Ying and A. Bax, *J. Am. Chem. Soc.*, 2009, **130**, 16518–16520.
119 B. Schneider, M. Kabeláč and P. Hobza, *J. Am. Chem. Soc.*, 1996, **118**, 12207–12217.
120 I. Persson, M. Trublet and W. Klysubun, *J. Phys. Chem. A*, 2018, **122**, 7413–7420.
121 P. Bouř, J. Hudecová and K. H. Hopmann, *J. Phys. Chem. B*, 2012, **116**, 336–342.
122 X. Li, K. H. Hopmann, J. Hudecová, W. Stensen, J. Novotná, M. Urbanová, J.-S. Svendsen, P. Bouř and K. Ruud, *J. Phys. Chem. A*, 2012, **116**, 2554–2563.
123 K. H. Hopmann, J. Šebestík, J. Novotná, W. Stensen, M. Urbanová, J. Svenson, J. S. Svendsen, P. Bouř and K. Ruud, *J. Org. Chem.*, 2012, **77**, 858–869.
124 V. Andrushchenko, L. Benda, O. Páv, M. Dračínský and P. Bouř, *J. Phys. Chem. B*, 2015, **119**, 10682–10692.

