LEGOLAS: a Machine Learning method for rapid and accurate predictions of protein NMR chemical shifts.

Mikayla Y. Darrows[1], Dimuthu Kodituwakku[1], Jinze Xue[1], Ignacio Pickering[1], Nicholas S. Terrel[1], Adrian E. Roitberg*

1. Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States

* To whom correspondence must be addressed. roitberg@ufl.edu

Abstract

This work introduces LEGOLAS, a fully open source TorchANI-based neural network model designed to predict NMR chemical shifts for protein backbone atoms. LEGOLAS has been designed to be fast, and without loss of accuracy, as our model is able to predict backbone chemical shifts with root-mean-square errors of 2.69 ppm for N, 0.95 ppm for Cα, 1.40 ppm for Cβ, 1.06 ppm for C', 0.52 ppm for amide protons, and 0.29 ppm for Hα. The program predicts chemical shifts at least one order of magnitude faster than the widely utilized SHIFTX2 model. This breakthrough allows us to predict NMR chemical shifts for a very large number of input structures, such as frames from a molecular dynamics trajectory. In our simulation of the protein BBL from *E. coli*, we observe that averaging the chemical shift predictions for a set of frames of an MD trajectory substantially improves the agreement with experiment with respect of using a single frame of the dynamics. We also show that LEGOLAS can be successfully applied to the problem of recognizing the native states of a protein among a set of decoys.

Introduction

NMR has become a routine method to study proteins in experimental settings, as it is highly sensitive to subtle environmental changes, and requires less preparation and is less dependent on sample conditions than X-ray and cryo-EM methods.[1] Namely, NMR chemical shifts (δ) provide a direct and accurate method of determining a variety of protein structural features, such as regions of secondary and post-secondary structure and conformational changes as a result of folding or ligand binding.[2] Although useful, computation of NMR chemical shifts and their trends in proteins has presented challenges. Quantum Mechanical (QM) methods are capable of computing chemical shifts; however, their application to large systems is limited by the rapid increase in computational cost with system size.[3] Fragmentation approaches have been developed to compute chemical shifts of biomolecules, though such quantum mechanics/molecular mechanics techniques are still computationally expensive.[4] Molecular dynamics (MD) methods are commonly used to study large biomolecules such as proteins, and these methods are computationally inexpensive enough to study a dynamic system over time. This efficiency stems from approximating atoms as single points in space, but this same simplification limits their ability to compute detailed atomic properties like chemical shifts.

Machine learning methods can close the gap between computational speed and the accuracy of atomic property predictions. There are several models available currently that can do this by evading the need for expensive QM calculations; instead, they rely on training to experimental data.[3] Models such as SHIFTX2[5], SPARTA+[6], UCBShift[7], NMR GNN[8], and PROSHIFT[9] are often used to support experimental chemical shift research for proteins. Specifically, SHIFTX2 integrates a sequence-based prediction model, SHIFTY+ with a structure-based neural network model, SHIFTX+, resulting in considerable accuracy for native protein states. By construction, it can falter when applied to non-native protein states due to the inability of the SHIFTY+ module to differentiate between folded and non-native states. The success of SHIFTY+ is also limited to whether the query protein has an aligned sequence match in its database. Consequently, for analyzing non-native structures or for sequences with no matched data, SHIFTX2's accuracy is contingent on its SHIFTX+ component. This has been recognized by work that has utilized SHIFTX2, where they note that although they use the SHIFTX2 program, the reported accuracy of SHIFTX+ is more reliable when considering their applications[10,11]. Additionally, the lack of numerical differentiability in both SHIFTX2 and SHIFTX+ prevents their use in applications requiring gradient-based optimization, such as refining molecular structures.[5] SHIFTX2, SHIFTX+, and SPARTA+ are much slower than our implementations[5], limiting many potential practical applications. This can become considerably detrimental when needing to compute chemical shifts for a large dataset of protein structures or frames of an MD run.

This work presents LEGOLAS (neura**L** n**E**twork en**G**ine f**O**r ca**L**culating chemic**A**l **S**hifts), a neural network model that predicts protein backbone NMR chemical shifts. LEGOLAS is implemented in TorchANI[12], a Pytorch-based[13] environment that is designed to be used in the training and inference of ANAKIN-ME (ANI) deep learning models.[14-17] This interface was used because it is light weight, user-friendly, cross platform, and easy to read and modify.[12] TorchANI contains a core library including the atomic environment vector (AEV) computer, amongst other utilities.[12] The AEV computer allows us to encode molecular structure as vectors using highly-transferable modified Behler and Parrinello[18] symmetry functions.[12] These symmetry functions are continuous and differentiable[14], and are well-supported by PyTorch.[12] This feature allows our program to be end-to-end differentiable, making it adaptable to gradient methods.[12] This is essential for applications like refining molecular structure using biased molecular dynamics.[19-21] TorchANI also allows us to complete fast training and inference on modern NVIDIA GPUs, along with computing CUDA-accelerated AEVs (CUAEVs) for efficient calculation and storage of AEVs.[12]

Our model was trained on the SHIFTX2[5] dataset of protein structural information paired with experimental ^1H, ^{13}C, and ^{15}N chemical shifts. Several chemical shift predictors have been trained and tested using this dataset, which allows us to easily benchmark the performance of our model.[22] We evaluate our model performance and compare with the above models using accuracy and timing metrics on the SHIFTX2 test set.[22] Our assessment reveals that LEGOLAS not only demonstrates significantly faster prediction speeds than other models, but also maintains exceptional accuracy. This makes LEGOLAS highly practical for applications requiring the analysis of an exceptionally large number of structures.

We further assessed LEGOLAS by evaluating its capability to discern a native structure from a set of decoys.[23,24] LEGOLAS can reliably identify the native structure from diverse protein datasets, further affirming its reliability in structural prediction and underscoring its promising utility for precise structure identification from a pool of potential candidates.

Other neural network models such as SHIFTX+ have been further put to the test beyond their traditional role in predicting chemical shifts for individual structures. Several studies suggest that dynamically averaged NMR chemical shifts obtained through molecular dynamics (MD) simulations show substantial improvements when compared to individual structures[10,25,26], even highlighting a noteworthy alignment with experimental chemical shifts.[10] Yet, it is crucial to highlight that not every method designed for predicting chemical shifts is suitable for MD datasets. For effective application in MD, a chemical shift predictor must be sensitive to structural changes and fast enough to efficiently compute chemical shifts for thousands to tens of thousands of frames. LEGOLAS's rapid calculation capabilities position it as an ideal candidate for the application of a chemical shift predictor within this context. The integration of a chemical shift predictor and MD opens a new avenue of possibilities, enabling a deeper understanding of the intricate behaviors and interactions within biomolecules.

Methods

Model

LEGOLAS is implemented using PyTorch[13], which allows us to encode our data and employ our model using custom functions. PyTorch is compatible with CUDA[27], which allows us to utilize NVIDIA GPUs throughout training and inference. Input data for our model consists of Protein Data Bank (PDB)[28] files that are then encoded as local descriptors (AEVs) in TorchANI, with one AEV computed per backbone atom. We augmented our model by including amino acid type information as a vector embedding. This embedding was appended to the input AEV, providing the model with valuable information about the specific characteristics of each amino acid. The encoding involves utilizing continuous and differentiable symmetry functions to encapsulate information from spatially nearby atoms, which is further detailed in our previous work with ANI.[14] This process consists of two components: a radial part containing interatomic distance information within a cutoff of 5.1 Å, as shown in *Figure 1a*, and an angular part containing interatomic angular information within a cutoff of 3.5 Å. The AEVs incorporate details about all atoms within these cutoffs, irrespective of their binding to the selected atom. In this case, the AEV does not contain any details about atoms beyond these cutoffs, meaning that atoms outside these limits won't impact the predicted chemical shift of this atom.

Three layers were used for each network (two hidden, one output), and all applied linear transformation and ELU[29] activation functions. The size of the input layer was 570, where AEV size was 560 and embedding size was 10. The size of the first hidden layer was 256, the second was 64, and the output layer was 1. The output provides chemical shifts in ppm. The final Chemical shifts are an average of the outputs from 5 independent models for enhanced accuracy. The structure of LEGOLAS is summarized using *Figure 1b*.

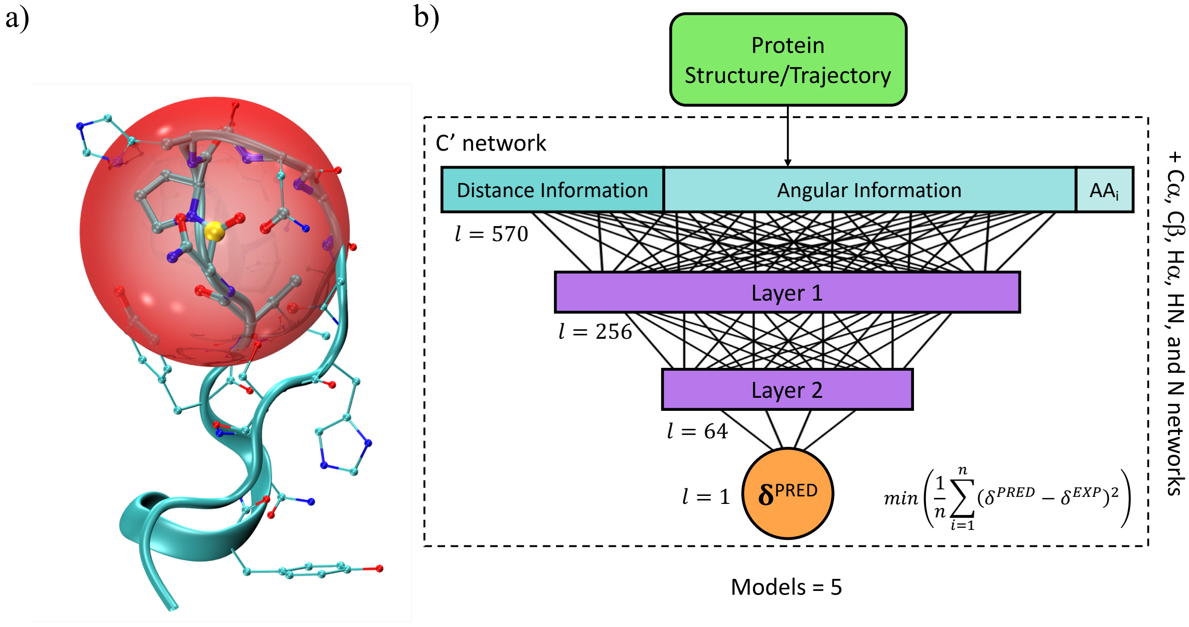
a)
b)
Protein Structure/Trajectory
C' network
Distance Information
Angular Information
AA_i
$l = 570$
Layer 1
$l = 256$
Layer 2
$l = 64$
$l = 1$
δ^{PRED}
$min\left(\frac{1}{n}\sum_{i=1}^{n}(\delta^{PRED} - \delta^{EXP})^2\right)$
+ Cα, Cβ, Hα, HN, and N networks
Models = 5

Figure 1. a) Example of distance information included in a Ca's AEV within a radial cutoff of 5.1 A. b) LEGOLAS architecture. The structure files (PDB or trajectory) are input to 6 networks corresponding to backbone atom type. Molecular structure is encoded as AEVs, which contains interatomic distance information, interatomic angular information, and amino acid type (AA_i). AEVs are computed for each atom and are of length 570. Using ELUs, AEVs are passed through two hidden layers, the first having 256 neurons and the second having 64 neurons, and finally to the output layer containing predicted chemical shift (δ^{PRED}). This is completed for 5 models trained independently.

Training

Our model was trained and tested using a dataset of protein structural information paired with experimental ^{1}H, ^{13}C, and ^{15}N chemical shifts provided by SHIFTX2[5]. The SHIFTX2 dataset consists of a training set of 197 proteins and a test set of 61 proteins. 6 different networks were trained, one for each backbone atom type: C', Cα, Cβ, Hα, HN, N. For each backbone atom type, we applied z-score normalization by first shifting the values using the experimental average specific to its residue type and then dividing by the standard deviation for that residue type. After prediction, the resulting chemical shifts were denormalized and shifted back to their original scale in ppm. To ensure robustness and reliability, we adopted a model ensembling approach, generating variations in training/validation splits without replacement. Specifically, the SHIFTX2 training set was partitioned into five segments, with four used for training and one reserved for validation in each iteration. This 80:20 split was systematically repeated, guaranteeing that every molecule appeared in the validation set at least once.

We employed the mean squared error as our loss function as the following:

$$min\left(\frac{1}{n}\sum_{i=1}^{n}(\delta^{PRED}-\delta^{EXP})^{2}\right)$$

The AdamW optimizer, with a learning rate of 0.001, was chosen to prevent overfitting and improve the generalization performance of the model. Additionally, we employed an early stopping mechanism during training. The "ReduceLRonPlateau" strategy was configured with a factor of 0.5, a patience parameter of 100 epochs, and a threshold of 0.001. The best-performing model, determined by the lowest Root Mean Squared Deviation (RMSD) on the validation set, was saved for subsequent analyses.

Decoy Test

We studied 631 structures of monomer A of the ribosomal protein L7/L12 from E.coli (PDB:1CTF) from the 4-state-reduced decoy data set[30] provided by the Decoy 'R' Us database.[31] This dataset can be found under the "multiple" decoy sets, which include proteins where a range of conformations with different RMSDs to the experimental structure are present. Included in this set are the 631 PDBs and their corresponding Ca RMSDs ranging from 0 to 10 Å from the native structure. Hydrogens were added to all structure files using the *Reduce*[32] program in AmberTools23.[33]

The decoy test follows the procedure presented in UCBShift.[7] Experimental ^{1}H and ^{15}N chemical shift assignments for the 1CTF protein were obtained from BMRB.[34] Correlation coefficients between experimental chemical shifts and chemical shifts predicted by LEGOLAS were computed for each structure. These were averaged over HN, Hα, and N, then compared to Cα RMSD to the native structure.

Molecular Dynamics

The structural information for NaF-BBL (2CYU) was obtained from the PDB database.[35] The AMBER ff19SB force field was used for the protein.[36] Simulations were conducted utilizing the AMBER20 suite.[37] The protein was solvated in a cubic box with a buffer of 10 Å using TIP3P water molecules and neutralized by adding counter ions. The SHAKE[38] algorithm was employed to enforce constraints on bonds involving hydrogen atoms during the simulations. Additionally, hydrogen mass repartitioning was implemented, allowing the utilization of a time step of 4 fs with SHAKE.[39] For the treatment of long-range interactions, Particle Mesh Ewald was utilized, and non-bonded interactions were calculated with an 8 Å cutoff.[40]

The system underwent an initial minimization of 500 steps of steepest descent and 500 steps of conjugate gradient, while subject to restraints of 500 kcal mol^{-1} Å^{-2} applied from residue 2 to 40. Subsequently, an additional minimization phase comprising 1000 steps of steepest descent and 9000 steps of conjugate gradient, during which the restraints were removed. The minimized structure was then gradually heated at a constant volume to 303.8 K with backbone restraints of 10 kcal mol^{-1} Å^{-2} over a period of 4 ns. Then, it underwent additional relaxation at constant volume for 400 ps without restraints. Constant pressure and temperature simulations were performed using Langevin thermostat with a friction coefficient of 2 ps^{-1} and Monte Carlo barostat for 4 μs.[41,42] Subsequently, constant volume and temperature simulations for 19 μs. For analysis, a trajectory comprising 23,469 frames was generated by extracting every 50th frame excluding the first microsecond of the production run. A water shell was created by selecting the closest 180 water molecules around the protein, and the remaining water molecules for the selected 23,469 frames were deleted using CPPTRAJ.[43]

LEGOLAS is employed to derive chemical shifts for individual frames and to directly calculate the average chemical shifts for each atom throughout an entire MD simulation. The input for LEGOLAS consists of MD simulation result files: the parameter/topology file, and the trajectory file. The output file generated by LEGOLAS provides the average and standard deviation of chemical shifts across the 5 models for each atom across every frame of the simulation.

Results and Discussion

Performance on SHIFTX2 test set

The correlation of LEGOLAS predictions with the true reference values are mapped in Figure 2, where the predicted chemical shift (ppm) versus experimental chemical shift (ppm) were plotted for the test set of 61 proteins, separated by backbone atom type. The root-mean-squared error (RMSE) of each backbone atom type was computed to be 2.69 ppm for N, 0.95 ppm for Cα, 1.40 ppm for Cβ, 1.06 ppm for C’, 0.52 ppm for amide protons, and 0.29 ppm for Hα.

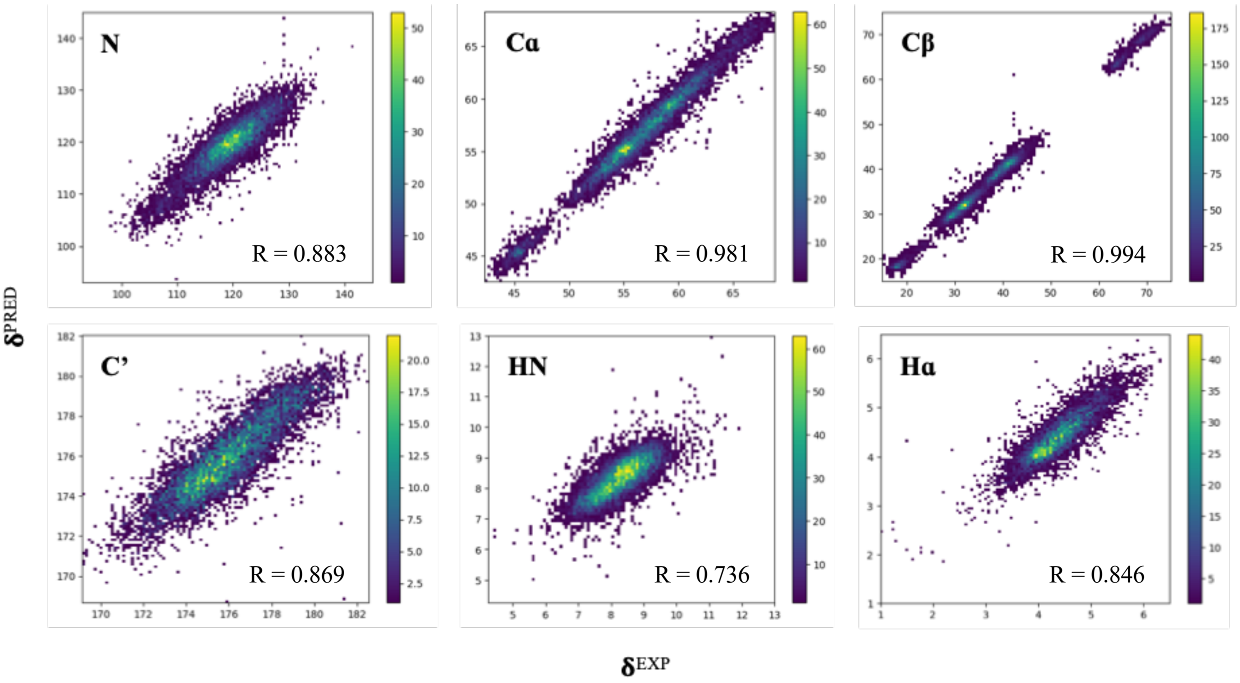

Figure 2. 2D Histograms of the prediction of chemical shift in ppm per backbone atom type over a test set of 61 proteins. Correlation coefficients are indicated for each backbone atom type.

Model performance is shown in Figure 3. A complete breakdown of RMSE values computed for LEGOLAS are given in SI Table S1, including RMSEs for each atom type during validation. As shown in Figure 3, we benchmark model performance by computing correlation coefficients (R) and RMSE per backbone atom type using SHIFTX2 and SHIFTX+, as these models have been trained using the same dataset as LEGOLAS. This allows us to make proper conclusions about our model's performance by eliminating differences in performance due to variations in data.

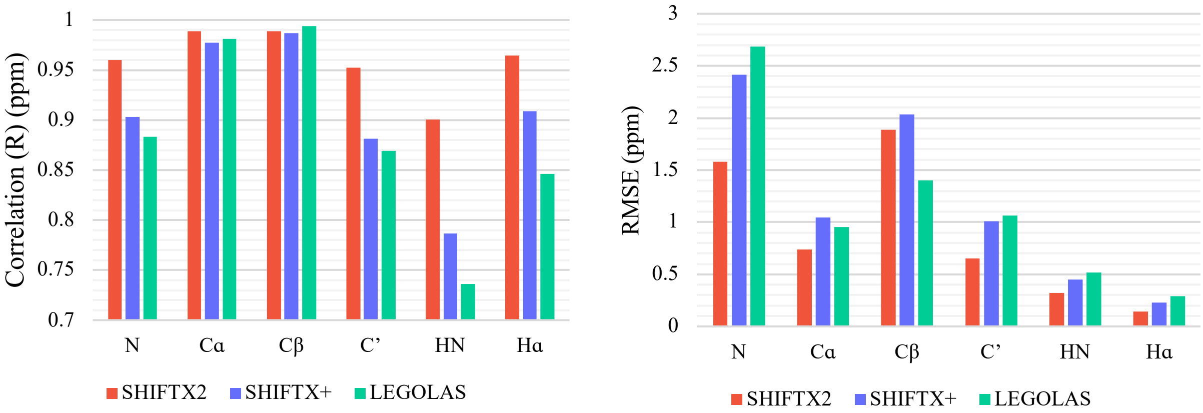

Figure 3. Comparison of prediction accuracy of LEGOLAS to SHIFTX2 and SHIFTX+ on the SHIFTX2 test set. Metrics used include correlation coefficient (R) (left) and RMSE (right).

With the version of SHIFTX2 presently accessible for download, we encountered challenges replicating the performance achieved by SHIFTX2 and SHIFTX+, as reported in the original SHIFTX2 manuscript. This is detailed in SI Table S2, which shows comparisons of RMSE and R values per atom type, encompassing both our computations those reported by reference 5, the original manuscript presenting SHIFTX2. Additional sources report similar inconsistencies when running SHIFTX2.[7,8]

The addition of the SHIFTY+ module leads to a significant improvement in correlation and error on the SHIFTX2 test set, mostly because ~75% of the molecules in the test set have sequence homologues in the RefDB database.[44] As for molecules beyond this test set, there are certain instances in which relying solely on SHIFTX+ is generally more dependable than SHIFTX2, such as if the query protein does not have a matching sequence in the SHIFTX2 database and for proteins that are in a non-native state. Since our applications require studies of these instances, we mainly compare our model's performance alongside other structure-based predictors. Here, we show the comparison of LEGOLAS to SHIFTX+, as it has shown high correlation and low error in comparison to several other models.[22] We find that LEGOLAS and SHIFTX+ perform similarly on this test set, and slight differences vary depending on the atom type. An analysis of RMSE vs residue type for each backbone atom type in the test set shows the presence of high RMSEs for cysteine CB and HN along with proline N as displayed in SI Figure S1. This is likely due to insufficient training and testing data for these residue-specific atom types. We are actively working to improve accuracy, ensuring a more reliable simulation of NMR chemical shifts for our applications.

Aside from accuracy, we optimized LEGOLAS for speed. LEGOLAS is specifically programmed so that the calculation of AEVs and chemical shifts can be completed using either CPUs or GPUs. Table 1 includes the total time it takes for LEGOLAS as a 5-model ensemble to predict chemical shifts for six backbone chemical shifts in the SHIFTX2 test set containing 61 proteins (10,760 total amino acid residues) when computed on CPUs or GPUs. These inference times are compared to three other models on the same test set. Included are SHIFTX2 and SHIFTX+, along with UCBShift-X, the machine learning module of UCBShift[7]. None of the other chemical shift predictors referenced in this work have any available method to run on GPU. Therefore, timings for all programs were computed using two CPUs. The timings for the models previously mentioned are listed in Table 1 as Total Prediction Time, and without including the time required for loading and reading files, which we call "Raw Prediction Time".

Table 1. Total and raw prediction times for four different protein chemical shift predictors on a test dataset of 61 proteins. Timing of LEGOLAS was computed on an NVIDIA A100 PCIE 40GB GPU and two Intel® Xeon® E5-2637 v4 (3.50GHz) CPUs. SHIFTX2, SHIFTX+, and UCBShift-X on two Intel® Xeon® CPU E5-2637 v4 (3.50GHz), and the speed up of this in comparison to UCBShift-X was applied to these models.

Program	Total Prediction Time (s)	Speed up	Raw Prediction Time (s)	Speed up
UCBShift-X	1534	1x	390	1x
SHIFTX2	89.8	17x	63.5	6x
SHIFTX+	77.6	20x	51.1	8x
LEGOLAS: CPU	32.5	47x	15.2	26x
LEGOLAS: GPU	6.51	236x	3.23	121x

On CPU, UCBShift-X takes the longest to make predictions on the SHIFTX2 test set, while SHIFTX2, SHIFTX+, and LEGOLAS demonstrate similar inference timings. Our model is at a great advantage since it can be fully run on GPUs. On GPU, LEGOLAS is ~14 times faster than the widely used SHIFTX2. Another unique property of our model is its ability to make atom-type specific calculations. In some cases, not all chemical shifts are needed, so only a reduced number of networks would be used, with an increased speedup. When considering raw prediction timings, LEGOLAS on GPU demonstrates a significant advantage over SHIFTX2, boasting a speedup of 20 times. We can attribute this to a slower time opening and reading of files. This provides a notable benefit. Molecular dynamics simulations typically require only two files: a trajectory file containing information about atom positions over time, and a topology file describing molecular structure and atom connectivity. LEGOLAS would be exceptionally fast in such cases, as it would only need to process these two essential files, potentially making it more efficient for molecular dynamics applications compared to models that may require processing additional files or data.

Utilizing LEGOLAS for Protein Structure Determination

Following the procedure described in the UCBShift manuscript[7], LEGOLAS is evaluated on its ability to select the experimental structure from a set of decoy structures, using chemical shift predictions. We obtained a decoy dataset that has a range of altered and misfolded structures as measured by the α-carbon RMSD versus the native state. We predict the chemical shifts for each structure and compute the correlation between the prediction and the experimental chemical shifts. The average correlation coefficients for each structure with experimental chemical shifts of H, Hα and N are plotted versus RMSDs to their native structures in Figure 4. More details and sources are described in Methods. We expect that a good prediction method would have a high correlation coefficient between the predict and measured chemical shifts for the correct experimental structure but exhibit much worse correlation for the decoy structures.

LEGOLAS is evaluated in comparison to UCBShift-X, the structure-based neural network model within UCBShift.[7] The UCBShift framework combines UCBShift-X and a sequence-based identifier known as UCBShift-Y. UCBShift is excluded from the comparison for two primary reasons. First, as LEGOLAS functions as a structure-based identifier, it is appropriately benchmarked against a program with a similar operational approach. Second, UCBShift includes BMRB 4429 (PDB:1RQU) in its RefDB dataset, utilized for UCBShift-Y predictions, resulting in an exact match with 1CTF. This would not be representative of a neural network models' prediction ability, but rather, looking up an exact match.

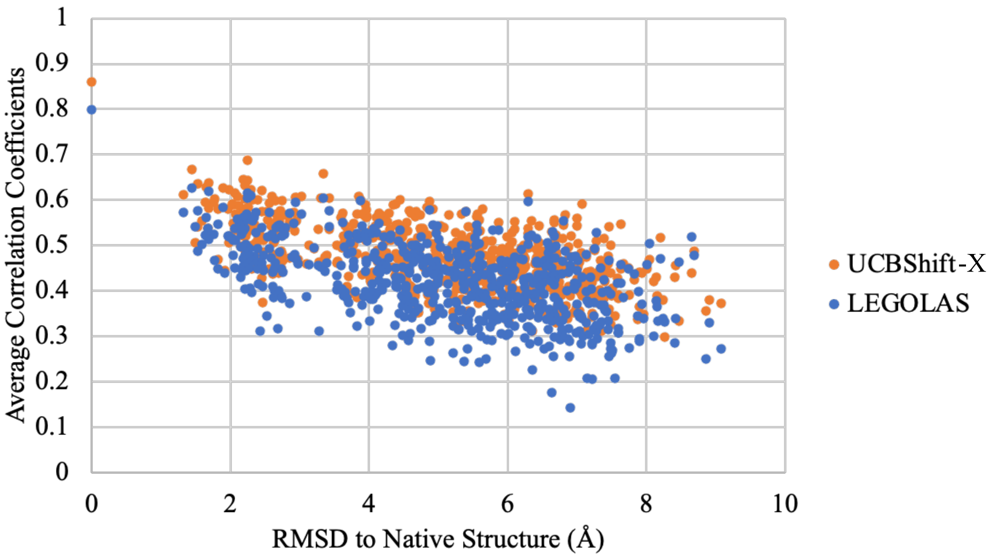

Figure 4. Average correlation coefficients between predicted and experimental H, HA, and N chemical shifts versus Cα RMSD to native structure for PDB 1CTF. Here, the results of LEGOLAS are compared to the results of UCBShift-X, as reported by UCBShift.[7]

In our analysis, LEGOLAS demonstrates a remarkable ability to discern the native structure, yielding a correlation coefficient of 0.80. Notably, there is a discernible trend showing a decrease in correlation as the RMSD increases. However, a more crucial finding is a substantial gap in correlation coefficients between the native structure and all other structures. The difference in correlation between the native structure and all structures is significant, at 0.17. UCBShift-X, on the other hand, assigns the native structure a correlation coefficient of 0.86, slightly surpassing LEGOLAS's prediction. However, UCBShift-X consistently predicts higher coefficients for all structures, as the difference between their native structure and all others is also 0.17. For the purpose of clearly distinguishing the native structure from a set of decoy structures, they have similar performance. We conclude that LEGOLAS can perform well with this test, not only affirming the validity and trustworthiness of its predictions but also demonstrating practical utility in real-life applications.

The total time required for this prediction is recorded in Table 2. UCBShift-X is not compatible with GPUs in its current release, so it was run using CPUs. LEGOLAS was run on a GPU.

Table 2. Differences in timing of chemical shift prediction for UCBShift-X and LEGOLAS on a dataset of 631 proteins containing 68 residues each. Timing of UCBShift-X was computed on two Intel® Xeon® CPU E5-2637 v4 (3.50GHz). Timing of LEGOLAS was computed on an NVIDIA A100 PCIE 40GB GPU.

Program	**Time (s)**	**Speed up**
UCBShift-X	4860	1x
LEGOLAS	42.0	116x

From start to finish, LEGOLAS was able to make 1,290,395 chemical shift predictions (631 structures * 409 backbone atom chemical shift predictions per structure * 5 models) in 42 seconds, completing its predictions for the decoy dataset 116 times faster than UCBShift-X. This decoy test emphasizes the crucial role of faster processing speeds when dataset sizes grow. In this instance, with a dataset of 631 proteins, a simulation that would take over an hour using UCBShift-X can be completed in just 42 seconds with LEGOLAS.

Chemical shift predictions in molecular dynamics simulations

An advantage of our network's high-speed inferences is that it can be applied to molecular dynamics simulations. Molecular dynamics simulations can contain up to tens of thousands of frames for one system, and fast computational speeds are necessary for collecting data of this volume.

This is an advantageous application for a chemical shift predictor because averaging the chemical shifts for an atom over a period of time would be more representative of molecular space than taking the chemical shift of that atom in a single frame. This holds especially true for atoms on highly dynamic residues and/or residues that interact with dynamic solvent molecules.

Due to the dynamic nature of protein conformations, each atomic nucleus will experience varying degrees of shielding in simulation. This variability reflects the importance of analyzing the chemical shifts of proteins over time, as a single frame could produce different chemical shifts than the next.

We applied LEGOLAS to a simulation of the 40-residue BBL (PDB: 2CYU) in explicit solvent over 23,469 frames. Our model can compute the chemical shifts for all 6 backbone atoms in each frame of this simulation on an A100 GPU in only 17.3 seconds. The total number of predictions was 235 chemical shifts per frame x 5 models x 23,469 frames = 27,576,075.

From this, a histogram for each atom can be extracted to fully understand the spread of values. An example of this is shown in Figure 5.

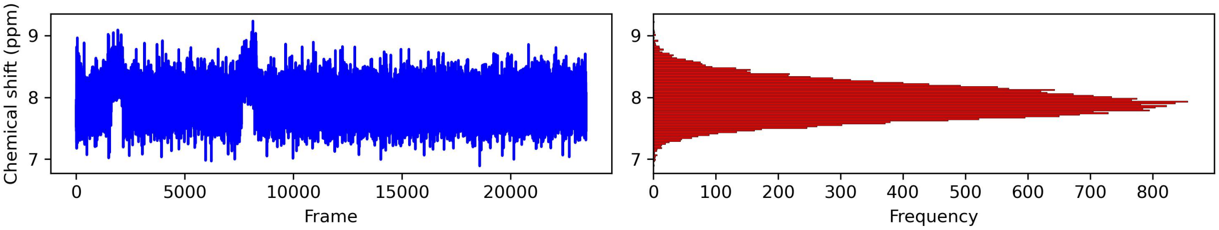

Figure 5. Chemical shift changes for an amide proton on residue 5 of 2CYU[35] over 23,469 frames. These chemical shifts predictions are plotted in a histogram (right), showing the variation of values depending on the protein conformation at each frame.

Figure 5 shows that at any frame of this simulation, the amide proton on residue 5 of 2CYU[35] could have a chemical shift as low as 7 ppm and as high as 9 ppm. This variation emphasizes the potentially significant fluctuations in chemical shifts across backbone atoms under constant volume and temperature (NVT) conditions, therefore underscoring the importance of considering a protein's complete structural dynamics when predicting chemical shifts.

Predicted versus experimental chemical shifts for each fragment as a single structure and as an average over MD simulation are detailed in SI Figure 2. Each graph is split based on three fragments of BBL: residues 5-12, residues 13-30, and residues 31-38, as labeled in Figure 6. Correlation appears to have a direct relationship with rigidity of structure. The first and third fragments have a considerably better correlation than the second fragment as they consist of atoms within larger, more rigid α-helices.

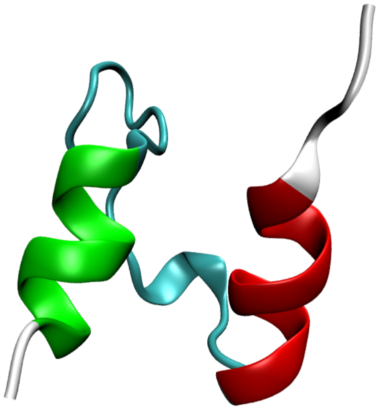

Figure 6. Native Structure of NaF-BBL from *E. coli*. Three fragments were individually analyzed based on secondary structure (residues 5-12: red, residues 13-30: cyan, residues 31-38: green).

The average chemical shift over all 23,469 frames for each atom obtained from simulation and the chemical shifts obtained using just the last frame of this simulation are compared to the experimental chemical shifts. From analyzing a single frame to analyzing the average over the simulation, we can see an increase in correlation coefficient between predicted and experimental shifts for all 3 fragments, along with a fit line that exhibits a closer alignment with the parity line in all cases, which is showcased in Figure 7 for Hα protons in residues 5-12.

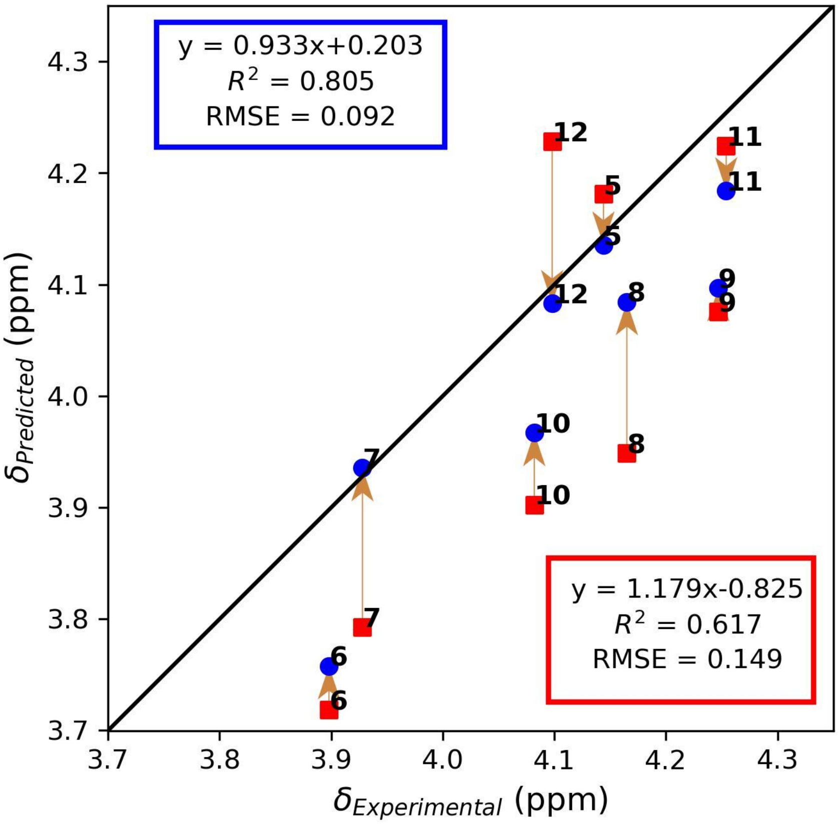

Figure 7. Experimental versus predicted Hα chemical shifts in residues 5-12 of BBL plotted against the parity line (black). Comparisons are made between predictions for a single frame (square, red) and the average over 23,469 frames (dot, blue). Shifts towards the identity line when considering MD instead of a single frame are denoted by arrows.

This indicates a better agreement between the predicted and observed values, signifying a more accurate predictive model. As further shown in SI Figure S2, this is especially apparent in its more rigid segments (residues 5-12 and 31-38) for both α-hydrogen and amide protons. Amide protons exhibit a great improvement in all fragments when taking the average chemical shifts over a simulation in comparison to a single frame, going from having little to no correlation (0.03) for residues 5-12 to having a correlation of 0.89.

Conclusion

This work introduces LEGOLAS, a powerful open-source TorchANI-based model designed for predicting NMR chemical shifts from protein coordinate data. LEGOLAS demonstrates significantly accelerated prediction speeds compared to other published models, while maintaining high accuracy. The increased efficiency is attributed to the simple footprint of the Neural Network and to LEGOLAS being programmed to utilize CUDA on the PyTorch framework, enabling parallel processing on GPUs. It is noteworthy that currently, the other models used for comparisons in this study cannot be executed on a GPU. This enhanced speed not only makes LEGOLAS a valuable tool for rapid and efficient computation of chemical shifts in MD simulations but also positions it for on-the-fly calculations.

Future endeavors will concentrate on enhancing LEGOLAS's accuracy. Although the current accuracy is comparable to that of other structural-based predictors, incorporating additional features during training has the potential to elevate LEGOLAS's predictive capabilities to the next level.

Data Availability

LEGOLAS is freely available on GitHub (https://github.com/roitberg-group/legolas). The SHIFTX2 and decoy datasets are available on GitHub under data/.

Acknowledgements

MYD acknowledges partial funding from the Frontera Computational Science Fellowship NSF Award OAC-1818253. AER acknowledges partial funding from the National Science foundation, through award OAC-2311632. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu. The authors acknowledge University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication. URL: http://researchcomputing.ufl.edu. We gratefully acknowledge Kate Huddleston and Yinuo Yang for their invaluable insights and contributions throughout this project.

Author Contributions

MYD, IP, and AER carried out model development and model analysis. DK carried out generation and analysis of MD data. JX edited the LEGOLAS source code to be run on trajectories and visualize chemical shift changes over frames. AER project lead and supervisor. MYD, DK, NST, and AER contributed to writing the manuscript.

References

(1) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93* (4), 1866–1879.

(2) Unraveling the Meaning of Chemical Shifts in Protein NMR. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2017**, *1865* (11), 1564–1576.

(3) Jonas, E.; Kuhn, S.; Schlörer, N. Prediction of Chemical Shift in NMR: A Review. *Magnetic Resonance in Chemistry* **2021**, *60* (11), 1021–1031.

(4) Case, D. A. Using Quantum Chemistry to Estimate Chemical Shifts in Biomolecules. *Biophysical Chemistry* **2020**, *267*, 106476.

(5) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* **2011**, *50* (1), 43–57.

(6) Shen, Y.; Bax, A. SPARTA+: A Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR* **2010**, *48* (1), 13–22.

(7) Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T. Accurate Prediction of Chemical Shifts for Aqueous Protein Structure on "Real World" Data. *Chem. Sci.* **2020**, *11* (12), 3180–3191.

(8) Yang, Z.; Chakraborty, M; White, A.D. Predicting chemical shifts with graph neural networks. *Chem. Sci.* **2021**, *12*, 10802-10809.

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

(10) Robustelli, P.; Stafford, K. A.; Palmer, A. G., 3rd. Interpreting Protein Structural Dynamics from NMR Chemical Shifts. *J. Am. Chem. Soc.* **2012**, *134* (14), 6365–6374.

(11) Pérez-Conesa, S.; Keeler, E. G.; Zhang, D.; Delemotte, L.; McDermott, A. E. Informing NMR Experiments with Molecular Dynamics Simulations to Characterize the Dominant Activated State of the KcsA Ion Channel. *J. Chem. Phys.* **2021**, *154* (16), 165102.

(12) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60* (7), 3408–3415.

(13) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; Devito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Ai, Q.; Steiner, B.; Fang, L. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*; 2019. https://arxiv.org/pdf/1912.01703.pdf.

(14) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203.

(15) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241733.

(16) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Scientific Data* **2020**, *7* (1), 1–10.

(17) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**. https://doi.org/10.1021/acs.jctc.0c00121.

(18) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401.

(19) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics. *Nature* **2005**, *433* (7022), 128–132.

(20) SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19* (1), 109–116.

(21) White, A. D.; Voth, G. A. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *J. Chem. Theory Comput.* **2014**, *10* (8), 3023–3030.

(22) *SHIFTX2*. http://www.shiftx2.ca/ (accessed 2023-12-07).

(23) Chen, J.; Siu, S. W. I. Machine Learning Approaches for Quality Assessment of Protein Structures. *Biomolecules* **2020**, *10* (4). https://doi.org/10.3390/biom10040626.

(24) Mirzaei, S.; Sidi, T.; Keasar, C.; Crivelli, S. Purely Structural Protein Scoring Functions Using Support Vector Machine and Ensemble Learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16* (5), 1515–1523.

(25) Li, D.-W.; Brüschweiler, R. Certification of Molecular Dynamics Trajectories with NMR Chemical Shifts. **2009**. https://doi.org/10.1021/jz9001345.

(26) Markwick, P. R. L.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. Enhanced Conformational Space Sampling Improves the Prediction of Chemical Shifts in Proteins. *J. Am. Chem. Soc.* **2010**, *132* (4), 1220–1221.

(27) (CUDA) NVIDIA, Vingelmann, P., & Fitzek, F. H. P. (2020). *CUDA, release: 10.2.89*. Retrieved from https://developer.nvidia.com/cuda-toolkit

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

(29) Clevert, D.A., Unterthiner, T., Hochreiter, S. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*; 2016. https://arxiv.org/abs/1511.07289

(30) Bahadur, R. P.; Chakrabarti, P. Discriminating the Native Structure from Decoys Using Scoring Functions Based on the Residue Packing in Globular Proteins. *BMC Struct. Biol.* **2009**, *9*, 76.

(31) Samudrala, R.; Levitt, M. Decoys "R" Us: A Database of Incorrect Conformations to Improve Protein Structure Prediction. *Protein Sci.* **2000**, *9* (7), 1399–1401. (accessed 2023-8-14)

(32) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* **1999**, *285* (4), 1735–1747.

(33) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Andrés Cisneros, G.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.;

Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E., III; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M., Jr. AmberTools. *J. Chem. Inf. Model.* **2023**. https://doi.org/10.1021/acs.jcim.3c01153.

(34) Hoch, J. C.; Baskaran, K.; Burr, H.; Chin, J.; Eghbalnia, H. R.; Fujiwara, T.; Gryk, M. R.; Iwata, T.; Kojima, C.; Kurisu, G.; Maziuk, D.; Miyanoiri, Y.; Wedell, J. R.; Wilburn, C.; Yao, H.; Yokochi, M. Biological Magnetic Resonance Data Bank. *Nucleic Acids Res.* **2023**, *51* (D1), D368–D376.

(35) Sadqi, M.; Fushman, D.; Muñoz, V. Atom-by-Atom Analysis of Global Downhill Protein Folding. *Nature* **2006**, *442* (7100), 317–321.

(36) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528–552.

(37) D.A.Case, K.Belfon, I.Y.Ben-Shalom, S.R.Brozell,D.S.Cerutti, T.E.Cheatham, III, V.W.D.Cruzeiro, T.A.Darden, R.E.Duke, G.Giambasu, M.K.Gilson, H.Gohlke, A.W.Goetz, R . Harris, S. Izadi, S.A. Izmailov, K.Kasavajhala, A.Kovalenko, R.Krasny, T.Kurtzman, T.S.Lee, S.LeGrand,P.Li, C.Lin ,J.Liu, T.Luchko, R.Luo, V.Man, K.M.Merz, Y.Miao, O.Mikhailovskii, G.Monard, H.Nguyen, A.Onufriev, F. Pan, S.Pantano, R.Qi, D.R.Roe, A.Roitberg, C.Sagui, S.Schott-Verdugo, J.Shen, C.L.Simmerling, N.R. Skrynnikov, J.Smith, J.Swails, R.C.Walker, J.Wang, L.Wilson, R.M.Wolf, X.Wu, Y.Xiong, Y.Xue, D.M.York and P.A.Kollman (2020), AMBER2020, University of California, San Francisco.

(38) Gonnet, P. P-SHAKE: A Quadratically Convergent SHAKE in. *J. Comput. Phys.* **2007**, *220* (2), 740–750.

(39) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11* (4), 1864–1874.

(40) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An *N*·log(*N*) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(41) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374.

(42) Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular Dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm. *Chem. Phys. Lett.* **2004**, *384* (4-6), 288–294.

(43) Roe, D. R.; Cheatham, T. E., 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

(44) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: A Database of Uniformly Referenced Protein Chemical Shifts. *J. Biomol. NMR* **2003**, *25* (3), 173–195.