Title

UCBShift 2.0: Bridging the Gap from Backbone to Side Chain Protein Chemical Shift Prediction for Protein Structures.

Permalink

https://escholarship.org/uc/item/5w44v1bp

Journal

Journal of the American Chemical Society, 146(46)

Authors

Ptaszek, Aleksandra
Li, Jie
Konrat, Robert
et al.

Publication Date

2024-11-20

DOI

10.1021/jacs.4c10474

Peer reviewed

Published in final edited form as:
J Am Chem Soc. 2024 November 20; 146(46): 31733–31745. doi:10.1021/jacs.4c10474.

UCBShift 2.0: Bridging the Gap from Backbone to Side Chain Protein Chemical Shift Prediction for Protein Structures

Aleksandra L. Ptaszek[†,‡], **Jie Li**[¶,‡], **Robert Konrat**[†], **Gerald Platzer**[§], **Teresa Head-Gordon**[*,¶,‖]

[†]Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030-Vienna, Austria

[‡]A.L.P. and J.L. contributed equally to this paper

[¶]Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley CA 94720, USA

[§]MAG-LAB GmbH, Karl-Farkas-Gasse 22, 1030- Vienna, Austria

[‖]Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA

Abstract

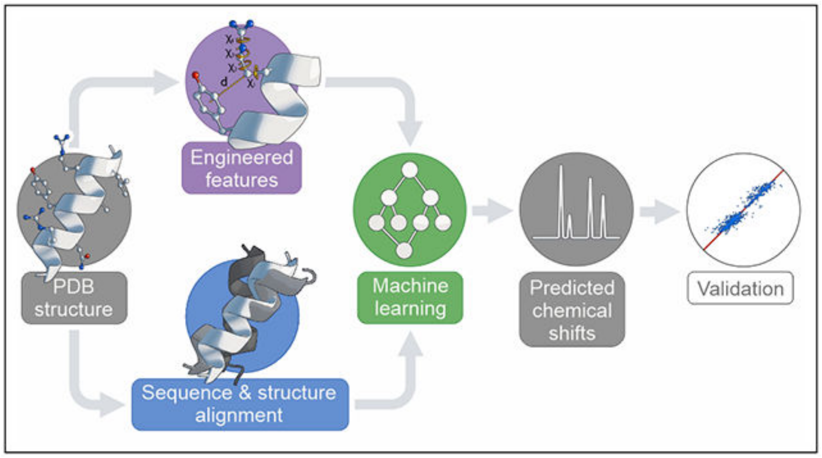
Chemical shifts are a readily obtainable NMR observable that can be measured with high accuracy, and because they are sensitive to conformational averages and local molecular environment, they yield detailed information about protein structure in solution. To predict chemical shifts of protein structures, we introduced the UCBShift method that uniquely fuses a transfer prediction module, which employs sequence and structure alignments to select reference chemical shifts from an experimental database, with a machine learning model that uses carefully curated and physics-inspired features derived from X-ray crystal structures, to predict backbone chemical shifts for proteins. In this work we extend the UCBShift 1.0 method to side chain chemical shift prediction to perform whole protein analysis, that when validated against well-defined test data shows higher accuracy and better reliability compared to the popular SHIFTX2 method. With the greater abundance of cleaned protein shift-structure data, and modularity of the general UCBShift algorithms, users can gain insight into different features important for residue-specific stabilizing interactions for protein backbone and side chain chemical shift prediction. We suggest several backward and forward applications of UCBShift 2.0 that can help validate AlphaFold structures and probe protein dynamics.

Graphical Abstract

[*] thg@berkeley.edu .

Code Availability
The code is provided through GitHub repository link: https://github.com/THGLab/CSpred/tree/SideChain

Supporting Information
Contents include details of proteins and chemical shifts used in training and test sets, and more analysis of UCBShift and SHIFTX2 performance on various benchmarks.



Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a primary experimental tool for characterizing dynamics and the solution structure of biomolecules. NMR chemical shifts for organic systems containing ^{1}H, ^{13}C, and ^{15}N nuclei can provide detailed descriptions of the structure of drug molecules,[1,2] proteins and their complexes,[3-5] and disordered protein states.[6-8] NMR chemical shifts in particular are sensitive not only to changes in local structure, but particularly to conformational changes that depend on sequence context and peptide length, solvent exposure or protein hydrogen-bonding environments, and even vibrational averaging. While quantum chemical calculations of magnetic properties are often powerful, the computational requirements associated with high level quantum mechanical (QM) chemical shift predictions are far too demanding for “on the fly” evaluation[9], although QM calculated chemical shifts have been used in machine learning training for chemical shift predictions.[1,2,10,11]

Therefore it is common practice to develop a database-trained expert system that can produce “predicted” chemical shifts by eliminating QM calculations altogether and instead predict the experimental observable directly. These heuristic calculators have been primarily focused on backbone chemical shifts, and must be trained to account for how NMR observables depend not only on backbone dihedral angles, but other features such as bond angles, deviations from planarity of the peptide bond, and hydrogen-bonding with the surrounding protein or solvent. Meiler and co-workers[12] and later Shen and Bax[13] showed that Artificial Neural Networks (ANNs) are well suited to utilize such features for protein backbone chemical shift prediction. The single-layer feed-forward network developed and packaged as SPARTA+[13] is a popular such example, with other feature-based methods including SHIFTCALC, [14] SHIFTX,[15] PROSHIFT[12], CamShift[16], and PPM[17,18] also showing comparable quality predictions. It is also worth distinguishing the SHIFTX+ component of SHIFTX2[19] as it uses not only backbone geometric features but also residue biological similarity properties like block substitution matrix (BLOSUM) numbers to predict chemical shifts using either ANNs or Bagging and Boosting ensemble models. Even so, these models still suffer from inaccuracies - presumably because the features drawn from

X-ray crystal structures are incomplete or unrepresentative- such that a pragmatic solution is to substitute known experimental chemical shift values of another protein that has high sequence homology to the target protein of interest. For example, SHIFTX2 also takes advantage of existing databases through the SHIFTY+ component that introduces an alignment-and-transfer technique to fully exploit sequence homology in order to make more accurate predictions.

We recently introduced the UCBShift 1.0 method[20] that offers several advancements on these early foundational models for chemical shift prediction of backbone atoms. The first improvement is to modernize the machine learning to utilize a random forest regression model with a greatly expanded X-ray data set and feature extraction and transformations, and is referred to as the UCBShift-X predictor. We also qualitatively change the nature of homology to not only include high sequence homology but also to introduce proteins with low sequence but high structural homology, that comprises the UCBShift-Y module. A final random forest regression step combines the two modules to make chemical shift predictions if homology to the target protein is available, otherwise shift predictions are made with only the UCBShift-X module. The mean absolute error (MAE) of UCBShift 1.0 for backbone atoms of proteins is 0.31 ppm for amide hydrogens, 0.19 ppm for H_α, 0.84 ppm for C', 0.81 ppm for C_α, 1.00 ppm for C_β, and 1.81 ppm for N.

In this work we seek to improve accuracy and robustness, as well as to gain insight, for chemical shift calculations for the carbon, hydrogen, and nitrogen atoms of side chains of aqueous proteins by retraining the UCBShift 1.0 model over an expanded data set. Most existing chemical shift predictors like SPARTA+ are restricted to backbone atoms but the SHIFTX2 algorithm also considers side chains and is a comparative method we will consider in this work. Because both models extract engineered features from high quality protein X-ray crystal structures that are insensitive to alternate conformations of a protein in the thermalized ensemble, we test under what circumstances the −X modules provide good discriminative power for side chain C, H, and N chemical shifts in different atomic environments. The primary importance of sequence and structural homology based alignment in the −Y component is designed to overcome the limitations of extracting static features from crystal structures that undoubtedly miss the underlying dynamics that is inherent in the substitution with an experimental chemical shift from solution-based NMR.

The resulting UCBShift-2.0 algorithm achieves significantly lower mean absolute error (MAE) and lower root-mean-square-error (RMSE) compared to SHIFTX2 (which does not predict on some of the atom types at all) when evaluated on an independently generated test dataset for side chains, with average MAEs of 0.80 ppm for C, 0.16 ppm for H, and 0.99 ppm for N chemical shifts. Overall the performance enhancement of UCBShift-2.0 arises from having more training data relative to the original SHIFTX2 study,[19] and more capacity and better exploitation of features and homology in the overall machine learning model. More specifically, UCBShift-2.0 can better predict the observed variations in chemical shift values for each amino acid, whereas SHIFTX2 mostly classifies shift predictions by residue type with little variation from random coil values. Furthermore we find that while the Y-module is responsible for ~50-60% of the performance enhancement of UCBShift-2.0 over SHIFTX2, side chain features such as ring currents and χ_2 rotamer states found through the X-module are also critically important for sidechain H and C chemical shifts, which are dominated by C-H groups. By providing both source code and cleaned data sets available from an open source github, we recommend that UCBShift-1.0 and UCBShift-2.0, and their UCBShift-X and UCBShift-Y modules, can each offer a current state-of-the-art side chain chemical shift prediction algorithm that can be tailored to desired applications for protein backbone and side chain chemical shift prediction.

Methods and Models

Preparation of enhanced datasets

The training set of the original UCBShift-1.0[20] was constructed by consolidating the training and testing sets SPARTA+ and SHIFTX+ into one comprehensive dataset. For predicting protein side chain chemical shifts, we enriched this dataset with chemical shifts available in the BMRB[21]. Chemical shift data obtained from the BMRB were re-referenced to high-resolution X-ray structures using Re-referenced Protein Chemical shift Database (RefDB)[22]. The final number of experimental chemical shifts for training and testing are provided in Table 1, and broken down further in Supplementary Table S1 and Supplementary Table S2. The PDB IDs with corresponding BMRBs of the 851 protein structures are provided in Supplementary Table S3. The test set consists of 200 structures that were prepared in the previous release of the UCBShift-1.0, with details provided in Supplementary Table S4.[20]

The structures from training and test sets were downloaded from RCSB and protonated using the PDB2PQR software.[23,24] We chose this algorithm because it allows for Propka-based protonation of ionizable residues based on the pH value. The PDB2PQR protonation was preceded by the optimization of adjustable protons (OH, SH, NH_3^+, Met-CH_3) and ambiguous Asn, Gln and His sidechain orientation with the REDUCE software.[25] Zhang et al. point out the existence of assignment errors present in the BMRB database.[22] The training and test sets filtered out outliers that were 6 and 12 ppm bigger or smaller than the average value for the particular residue for hydrogens and carbon atoms, respectively. This high threshold allowed us to filter out only missassigned shifts following the original UCBShift-1.0 idea of the ”real world” data.[20,26]

Harsch et al. have noted substantial misassignments of asparagine and glutamine side chain amide hydrogens within the BMRB.[27,28] According to their analysis, when the chemical shift differences between the two hydrogens are $\geqslant$0.40 ppm for Asn and $\geqslant$0.42 ppm for Gln, the resonance signal at the higher chemical shift (referred to as ”downfield shifted”) should be assigned to HD21 for Asn and HE21 for Gln, while the signal at the lower chemical shift (”upfield shifted”) should be assigned to HD22 and HE22, respectively. We corrected the assignments in our training and test sets based on this rule. Specifically, we swapped the assignments of HD21/HD22 or HE21/HE22 when $\delta(\mathrm{HD21}) < \delta(\mathrm{HD22})$ or $\delta(\mathrm{HE21}) < \delta(\mathrm{HE22})$, and the difference between them was $\geqslant$0.40 ppm for Asn or $\geqslant$0.42 ppm for Gln. The resulting distributions of chemical shifts before and after correction are shown in Supplementary Figure S1.

Machine learning workflow

The overall design of the UCBShift-1.0 and -2.0 chemical shift prediction algorithm is based on a random forest machine learning model, and consists of two modules, UCBShift-Y and UCBShift-X as shown in Figure 1. We used an Extra Tree Regressor and Random Forest Regressor as implemented in the scikit-learn package.[29] The hyperparameters were initially optimized with TPOT[30] using 3-fold cross-validation on the training set, then fine-tuned with a temporal validation set of 50 randomly selected structures. From here on out we refer to UCBShift-2.0 as UCBShift unless otherwise noted.

The UCBShift-Y component (blue path in Figure 1) transfers experimental chemical shifts from a reference database to a query protein based on sequence and structural similarity. The idea is similar to the SHIFTY+ module of SHIFTX2 predictor that utilizes the transfer of experimental chemical shifts from the protein database if the sequence is identical or closely matches the sequence of the query protein. UCBShift-Y also takes advantage of structural similarity in that it filters out mismatching chemical shifts of proteins with high sequence similarity but significantly different structure or when there is poor sequence alignment but significant structural similarity.

The UCBShift-X prediction algorithm (purple path in Figure 1) uses a feature vector and employs an extra tree regressor (R0) followed by a random forest regressor (R1).[31,32] The second Random Forest regressor (R2) incorporates the feature vector, the R1 regressor, and the secondary shift output from UCBShift-Y, along with additional scores and coverage metrics that indicate the quality of the alignments. The final chemical shift prediction is generated by either R1 (if UCBShift-Y predictions are unavailable) or R2 (if UCBShift-Y predictions are available). In the last step of the chemical shifts prediction, the random coil values are added back to the secondary shift predictions. Every type of atom chemical shift is trained individually. Some important details of the two components are described here for the readers benefit, but additional detail can be found in reference [[20]].

UCBShift-Y module.—The algorithm begins by aligning the target sequence with all sequences available in the RefDB database[22] and selecting significant matches using the BLAST algorithm.[33] Subsequently, the PDB files of identified sequences undergo structural alignment with the query protein via the mTM-align algorithm.[34] Only the alignments with a TM score exceeding 0.8 and an RMSD with a target below 1.75 Å are retained. Afterwards, the Needleman–Wunsch algorithm is used to determine the optimal alignment of each of the PDB sequences with the RefDB sequence.[35] Supplementary Figure S2 shows the calculated TM scores between each test sequence and the best-aligned sequence in the UCBShift and ShiftX2 training sets. Even for cases exhibiting full sequence homology with the training set, the chemical shifts sourced from the BMRB database do not coincide between the training and test sets, since they correspond to different types of atoms.

In case of identical residues, the chemical shifts from RefDB are directly assigned to the query protein. Otherwise, the target shift for atom A in residue I is calculated from the matching residue J according to the equation:

$$\delta_{\mathrm{I,A}} = \delta_{\mathrm{rc,I,A}} + (\delta_{\mathrm{J,A}} - \delta_{\mathrm{rc,J,A}}) \tag{1}$$

where $\delta_{\mathrm{J,A}}$ is the chemical shift of the matching residue, and δ_{rc}, and $\delta_{\mathrm{rc,I/J,A}}$ are the random coil chemical shifts in residues I and J for atom type A. When there is more than one aligned structure the chemical shift is calculated as the weighted average with weights w_{I} given by:

$$w_{\mathrm{I}} = \mathrm{e}^{5(S_{\mathrm{NA}} \times S_{\mathrm{TM}}) + B_{\mathrm{IJ}}} \times \mathbb{1}(B_{\mathrm{IJ}} \geq 0) \tag{2}$$

where $S_{\mathrm{NA}} = S_{blast} / max(S_{blast})$ is the sequence alignment blast score normalized by the maximum blast score; S_{TM} is the structure alignment TM score; B_{IJ} represents the substitution score between the amino acid at position I in the target sequence and the amino acid at position J in the matching sequence, based on the BLOSUM62 matrix.[36] Weights are set to zero when the substitution scores are negative, usually indicating the substitutions in the target sequence are dissimilar residues.

UCBShift-X module.—UCBShift-X is a decision tree-based machine learning module, which extracts structural features from a PDB file or property calculations that depend on the coordinates for each atom type. These residue- and atom type-specific features extracted from the protonated PDB structures include the following:

- BLOSUM62 numbers: Substitution scores derived from the BLOSUM62 matrix, indicating the probability of replacing the query residue with any other amino acid.[37]
- Backbone dihedral Angles: Sine and cosine values of the ϕ and ψ torsion angles of the query residue, as well as for the preceding and following residues.
- Side-Chain Dihedral Angles: Binary indicators for the existence of χ_1, χ_2, χ_3, χ_4 and χ_5 side-chain dihedral angles, and their corresponding sine and cosine values. They are considered for the query residue and the adjacent residues.
- Hydrogen Bonds: For every side chain hydrogen, hydrogen bond's features are described by five numbers: existence (boolean), distance between donor-acceptor pairs, cosine values of the angles at the donor hydrogen and acceptor atom, and hydrogen bond energy calculated via the DSSP model.[38] A hydrogen bond acceptor can be any oxygen or nitrogen atom in the protein.

  In addition, for each atom type, the backbone hydrogen bond descriptors related to the query residue are considered. These include the hydrogen bond between the amide hydrogen and carboxyl oxygen, and between the Cα hydrogen and a carboxyl group. For the query residue all five hydrogen bond descriptors are included for: (1) amide hydrogen (2) carboxyl oxygen, (3) α hydrogen and additionally (4) carboxyl oxygen features for the previous residue and (5) the amide hydrogen features for the next residue. This gives 25 backbone hydrogen bonding features for each atom type.

- Polynomial Transformations: Squared and inverse polynomial transformations of hydrogen bond distances and squares of dihedral angle cosine values, which are included in several empirical chemical shift calculation formulas.[39,40]
- S^2 order Parameters: NMR S^2 order parameters of N-H bond calculated using the contact model.[41]
- Accessible Surface Area (ASA): Absolute and relative accessible surface areas determined by the DSSP program.
- Secondary Structure: One-hot encoded secondary structure representation in 8 categories from the DSSP program.
- B Factor: Average B factor of the residue, extracted from the PDB file.
- Half Surface Exposure (HSE): A measure of the residue’s exposure in the protein structure.[42]
- Hydrophobicity: Hydrophobicity values from the Wimley–White whole residue hydrophobicity scales.[43]
- Ring Current Effect: Calculated using the Haigh–Mallion model, including the ring current for the specific atom type in the training model.[44]
- pH value of the NMR experiment.
- Electric Field: Electric field effect calculated using the formula:

$$\delta_{EF} = \sum_{i} \frac{q_i \epsilon \cos \theta_i}{d_i^2} \tag{3}$$

  for hydrogen and nitrogen atoms. In the equation, q_i is the charge of the interacting atom i, θ is the i-H-X angle (where X is a heavy atom bound to the target hydrogen) and d_i is the distance between the hydrogen i and the target hydrogen. Interacting atom charges are derived from the Amber force field parameters[45].

- Protonation Indicator: This binary feature indicates the presence (1) or absence (0) of atoms corresponding to ionizable residues, as determined by protonation state prediction software. This indicator relies on whether specific atoms, associated with protonation states, are present in the protonated PDB structure of the protein, reflecting significant shifts due to protonation changes.

Results

Table 2 displays the MAE, root mean square error (RMSE), Pearson’s correlation coefficients (R), and improvement factors computed using the UCBShift and SHIFTX2 models for chemical shift prediction by side chain atom type. The data reveals that UCBShift consistently outperforms SHIFTX2 across all side chain atom types, on average by ~0.28 ppm and ~0.03 ppm for carbons and hydrogens, respectively. Nitrogen chemical shifts are predicted by UCBShift, unlike SHIFTX2, with an average MAE of 0.99 ppm. The more detailed results, which distinguish between individual amino acids, also demonstrate improvement over SHIFTX2 in all types of nuclei (Table S5). It is also important to note that we utilized a “test mode” criteria for UCBShift predictions that excludes sequences with more than 99% similarity to the query sequence, while SHIFTX2 testing data may include 100% similarity cases. Overall UCBShift is a more consistent performing model for side chain chemical shift prediction despite this small handicap.

Protein side chains exhibit greater flexibility compared to the backbone dihedral angles, which are usually determined by the secondary structure. This higher flexibility leads to the presence of multiple rotameric states that makes accurate prediction of side chain chemical shifts more challenging.[46-48] In Figure 2 we provide a more detailed comparison of chemical shift accuracy using UCBShift versus SHIFTX2 illustrated with the CG and CD1 side chain atoms, which possess the largest amount of testing data as indicated in Table 1 and is present in many of the amino acid side chains. As seen in Figure 2a, the SHIFTX2 predictions exhibit a tendency to cluster around the amino acid random coil values, rather than accurately representing the experimental spread of chemical shift values that vary by each type of amino acid and ranges of environment. The UCBShift model, however, appears to better account for the shift variations within each amino acid type, providing predictions that are more finely tuned to the actual spread of observed shifts of CG as seen in Figure 2b.

As another example, we identify three conformational clusters of the experimental CD1 chemical shift for the isoleucine residue based on the χ_2 dihedral angle as seen in Figure 2c-e, in which the conformation with $\chi_2 = -60^\circ$ has been reported to correspond to the chemical shift value of ~11.4 ppm.[47] However, the broad experimental CD1 chemical shift distribution in each of the $\chi_2 = -60^\circ$ as well as $\chi_2 = -180^\circ$ clusters suggests that many conformational states can be present in solution, Figure 2c. We observe that the UCBShift predictor better captures the large spread of chemical shifts within each geometrical cluster, while SHIFTX2 does not predict the same spread of CD1 shifts within clusters observed in the experimental data. UCBShift additionally reveals a more favourable correlation between the predicted and experimental chemical shifts for the CD1 shifts of isoleucine, Figure 2f,g. Similar trends are observed for other atom types such as CG2 and including nitrogen which is only predicted by UCBShift, as summarized in Figures S3-S10.

The bar plots depicted in Figure 3 summarize the enhancement of UCBShift shift prediction for carbon and hydrogen side chain atoms in comparison to SHIFTX2 across various amino acid types. The overall RMSE using UCBShift systematically improves over SHIFTX2 for all amino acids. Histidine demonstrates a particularly pronounced enhancement in carbon chemical shifts prediction, which we attribute to the effective handling of protonation states by UCBShift.[26] Tyrosine also displays a notable decrease in errors. Looking at the details in Table S5, we identify the most significant discrepancies for CG and CZ. Reported experimental data for these nuclei is very limited (Table S1), for reasons discussed in later section. The improved performance of UCBShift for cases with little training data can be attributed to the extensive training set employed during its development, consisting of 851 structures compared to the 197 structures incorporated in SHIFTX2.

Next we consider why there are performance enhancements for UCBShift relative to SHIFTX2. Table S6 makes clear that the X-module of UCBShift shows small performance enhancements relative to SHIFTX2 over most of the C and H side chain chemical shift test data, although there are some exceptions as well. However, the more notable component of success for UCBShift appears to be in the Y-module predictions as reported in Table S6. Hence when any type of homology is available, along with an assigned experimental chemical shift, it improves average chemical shift MAES by 0.5 ppm for side chain carbons and 0.2 ppm for side chain hydrogens compared to the standalone UCBShift-X component.

We believe there are several contributing factors to this overall improvement through the Y-module. First is that we have more training data than the much earlier SHIFTX2 study, and in particular more sequence/structural homology data, that can be exploited better by the more sophisticated R2 regressor (see Figure 1). But to test this possibility, we also retrained the UCBShift model on the original SHIFTX2 data set as seen in Supplementary Table S7. This demonstrates that the UCBShift algorithm is not solely benefitting from an expanded data set but a better algorithm as well. In addition, it appears that, on average, feature extraction using static crystal structures can’t represent the solution NMR experiment, in which the substituted experimental chemical shift better represents the thermal fluctuations and time averaging over variable chemical environments not available in the X-module. Finally, given that ~75% of the data has available homology, UCBShift is trained to rely more heavily on this component of the machine learning algorithm.

This is not to say that feature extraction data is not important, since the R2 regressor takes into account not only the alignment, but direct feature data as well as the extra tree regressor R0 (see Figure 1). In Table S6 we see that the R2 regressor performs better than UCBShift-Y in regards MAE for CG, HB, HB2/HB3, HD21/22, HE21/22 atom types, and often eliminating outliers for many atoms by reducing the RMSE. The enhancement of the R2 regressor for hydrogens HD21/22, HE21/22 is especially noteworthy, as the underperformance of module Y may stem from the previously discussed misassignments of these hydrogens in asparagine and glutamine. We also constructed a low-homology test set of 59 proteins with a sequence identity of 50% or less to those in the training set. Table S8 shows that UCBShift outperforms SHIFTX2 on this more restrictive low homology test set that also demonstrates that the X-module is critically important as well.

Finally, we examine the various features used in the UCBShift model for predicting chemical shifts for each atom type in backbone and side chains, summarized in Tables S9-S11. Unsurprisingly, the UCBShift-Y prediction dominates the R2 regressor, with influence of features through the unoptimized node splits of the R0 regressor, for reasons explained above. Hence we analyze the test examples for the features that are important for the unoptimized node splits of the R0 regressor, and then corrected by the optimized R1 random forest regressor and displayed graphically in Figure 4.

Our previous investigation using UCBShift-1.0 for backbone atoms revealed that for carbon and nitrogen chemical shifts the backbone dihedral angle features are most important (Tables S9 and S11.). Additionally for backbone carbons the secondary structure classification dominates the R0/R1 regressor as seen in Figure 4 and Table S9. For backbone hydrogen atoms the same geometric features are important as well, but with the added component of untransformed and transformed hydrogen-bonding together with the stability and conservation of those structural features as determined by the BLOSUM number (Figure 4 and Table S10). These dominant features are perhaps unsurprising but reassuring, and support the essential role that backbone chemical shifts provide in protein structure determination.[49]

Moreover, various studies have explored the utility of side chain chemical shifts in elucidating protein conformation and dynamics.[47,48,50] It is notable that analysis of R0/R1 reveals the predominant relevance of various forms of side-chain dihedral angles for carbon, nitrogen, and hydrogen chemical shifts, BLOSUM numbers, accessible surfaces, and various forms of hydrogen bond features as shown in Figure 4. Interestingly, B factors, S2 order parameters, hydrophobicity, pH values, and electric fields play more negligible roles for chemical shift predictions for any side chain C, H, or N atom.

Perhaps the most distinguishing feature of side chain hydrogen chemical shift predictions with UCBShift is the importance of ring currents as seen in Figure 4. When a hydrogen is surrounded by an aromatic ring, it undergoes a significant shielding due to the electric current induced by an external magnetic field. Hence hydrogen chemical shifts are highly sensitive reporters of weak $CH\cdots\pi$ interactions that play important role in the stabilization of protein structures[51-54] and protein-ligand complexes[52,55,56]. For most of the hydrogen types including carbon bound hydrogens HB, HB2/3, HG2/3/12/13, HD2/3, and HE21 the ring current effect is a dominating feature with importance greater than 20% (Table S10). Interestingly, the carbon-bonded HA hydrogen reveals a significantly lower, 5.3%, ring current feature importance. The observation can be attributed to the analysis of protein PDB structures, which indicates that HA atoms, compared to the side-chain hydrogens, exhibit the least participation in $CH\cdots\pi$ interactions.[57]

Thus we further analyze UCBShift on a test subset containing hydrogen atoms that reveal notable shielding as a result of the ring current impact stemming from $CH\cdots\pi$ interactions in Figure 5. From the testing set, we select 20 chemical shifts below 1 ppm of HB atoms in Ile and Val, where the distance from the hydrogen to the centre of the aromatic ring $\leqslant 3.7$, Table S12. Figure 5 demonstrates that UCBShift surpasses SHIFTX2 in handling these extreme examples, and among all the sub-models, UCBShift-X exhibits the highest level of performance. In order to assess the impact of the ring current feature on the prediction outcome, we conduct a prediction with the UCBShift-X model while deactivating the ring current feature. Analysis of the last two rows of the heatmap in Figure 5 shows that the ring current feature decreases chemical shift values throughout the dataset. Additionally, it is observed that ShiftX+ and SHIFTX2 exhibit a lower accuracy in predicting the chemical shifts of hydrogens located in the most shielded positions on the left side. In contrast, the consistency of UBShift-X and UCBShift performance remains relatively stable throughout the dataset.

Turning to the hydrogen types including N-bound hydrogens, eg. backbone amide hydrogen, HD21/22 (Asn), HE1 (Trp), HE (Arg) we observe that H-bond features may overcome the ring current effect, Table S10. This is caused by the ability of N-bound hydrogens to act as effective hydrogen bond donors. Strong hydrogen bonds in the form of $\mathrm{N\text{-}H\cdots N\,/\,O}$ lead to a pronounced deshielding effect resulting in a large downfield shift of the interacting hydrogen.[53] Due to the large difference in chemical shift between strong, weak and non-hydrogen-bonded NH hydrogens the chemical shift range of these nuclei is broad and hence more difficult to properly assign.[27,28] For those hydrogens the prediction is usually worse, RMSE ~0.5 pmm, than for carbon-bound hydrogens with only marginal improvement relative to SHIFTX2 as seen in Table S5. Additionally, there is highly limited experimental chemical shifts data for N-bound side chain hydrogens (Tables S1 and S2). Interestingly, the relevance of hydrogen bond features extends to the chemical shifts of carbon-bound hydrogens, which can be attributed to the occurrence of weak $\mathrm{CH\ \cdots\ O\,/\,N}$ hydrogen bonds in protein structures.[58]

Among ionizable residues, histidine is particularly noteworthy because its protonation state can vary within physiological pH ranges. Our analysis indicates that the protonation feature impacts the R0 regression for CD2 and HE1 atom types by 3.1% and 2.2% respectively, Tables S9 and S10. This is in line with literature revealing a significant variation in histidine CD2 and HE1 chemical shifts upon ionization.[26] Overall, both, various protonation states of the histidine as well as its ability to form hydrogen bonds lead to a broader distribution of reported chemical shifts, resulting in a higher prediction error for this residue type (Figure 3, Table S5).[59]

The UCBShift-2.0 algorithm covers all side chain carbon atom types, but there is still a shortage of experimentally determined chemical shifts for 5 types of hydrogen atoms, specifically the hydroxyl group hydrogen (HH) of tyrosine and the terminal NH hydrogen atoms (HH11/12/21/22) of the guanidinium group in arginine. There is also insufficient experimental data for five nitrogen atom types, namely NH1/2 and NE of the guanidinium group in arginine, NZ of lysine, and ND1 of histidine. In addition, there is a limited amount of experimental chemical shift data and therefore relatively poor prediction performance for O/S- and some of the N-bound hydrogens (HG of Ser and Cys, HG1 of Thr, HZ of Lys, and HD1 and HE2 of His), Tables S1 and S5. Acidic hydrogens are known to be particularly challenging to measure due to their rapid exchange with water.[26,60-63] Direct observation of these nuclei is limited to situations where they are protected from rapid solvent exchange, for example when buried within a folded protein structure (due to hydrogen bonds or steric hindrance) and/or under conditions of low temperature or low pH. The HH11/12/21/22 hydrogen peaks of arginine are even more difficult to detect as they not only undergo proton exchange but are also affected by the NE-CZ bond rotation resulting in a single broad signal for the four hydrogens.[64] Additionally, the detection of O- and S-bound hydrogens (HG of Ser and Cys, HG1 of Thr, HH of Tyr), even under the rare slow exchange conditions, is hampered by the inapplicability of conventional 2D NMR methods.[65-68] In their discussion, Takeda and co-authors pointed out that NMR signals of labile protons can often be misinterpreted due to the above-mentioned challenges.[65]

When it comes to carbon side chain chemical shifts, there is a noticeable shortage of experimental chemical shifts available for quaternary carbon atoms in the dataset (CG chemical shifts of Phe, His, Trp, Tyr, CZ of Arg, Tyr, and CD2, CE2 of Trp), Table S1. The assignment of those nuclei is experimentally more challenging due to a lack of a directly attached hydrogen and requires more complex long-range NMR experiments explaining the sparse data in the BMRB.[69] This is for most of the quaternary carbon atom types reflected in the somewhat subpar predictive performance shown in Table S5.

Discussion and Conclusions

In this work we build upon the UCBShift-1.0 model for backbone atoms by extending it to side chain chemical shift predictions. We have used numerical and categorical features derived from high quality but static crystal structures that are greatly expanded in terms of training and test data, and we have redesigned a structure based alignment module to directly transfer the chemical shifts from the experimental database to the query protein when not only sequence but when the structure homology is sufficiently high, all incorporated in a random forest regression model. Overall UCBShift-1.0 and -2.0 stands as a modernized and state-of-the-art predictor of both backbone and side chain carbon, hydrogen, and nitrogen chemical shifts.

UCBShift yields a far superior side chain chemical shift prediction than SHIFTX2 through small improvements to its X-module, but mostly through its Y-module which is largely responsible for raising overall UCBShift performance within and across all amino acids and their various atom types. This is because as NMR data accumulates over time, 75% of folded proteins have a substitutable chemical shift from another protein with high sequence and/or structural homology, and the experimental chemical shifts contain dynamical and ensemble averaging over environments that is not easily extracted from static crystal structures. But UCBShift-Y module still does not outperform the R2 regressor in all cases, because the engineered feature set also help predict the environmental variations of observed chemical shift values within each amino acid type. In particular, the engineered features of ring currents are especially important for hydrogen chemical shifts of nearly all side chains, especially for $\mathrm{CH}\cdots\pi$ interactions, and the greater environmental variations of observed C chemical shift values for each amino acid depend on a dominant $\chi 2$ rotamer. For extreme cases and/or underrepresented examples such as low-frequency $\mathrm{CH}\cdots\pi$ interactions, the UCBShift-X module outperforms UCBShift itself, due to interference by the UCBShift-Y where the transferred shift is unhelpful. Hence while UCBShift-1.0 adnd -2.0 can be used as a black-box algorithm for chemical shift prediction, we encourage users to take advantage of all modular UCBShift-X and UCBShift-Y components that can be tailored to desired applications for protein backbone and side chain chemical shift prediction.

Encouraged by the performance of our prediction tool we anticipate several valuable applications with significant impact on biomolecular NMR spectroscopy. Specifically, we suggest backward and forward applications of UCBShift 2.0. The most obvious backward application would be a 3D structure-based rigorous validation of experimental NMR signal assignments either obtained via automated or manual approaches. The example of Asn and Gln side-chain signals described in our study suggests valuable applications for these notoriously difficult amino acids. It is interesting to note that a related correction tool was developed in the past for correcting erroneously annotated Asn/Gln rotamers in experimental X-ray structures.[70]

The introduction of AlphaFold2 (AF2) has clearly revolutionized structural biology.[71] Despite the high level of accuracy of AF2 models there is still the need for independent model evaluation and several experimental methods (SAXS, X-ray crystallography, cryo-EM or NMR) have been proposed for this task.[72] The availability of side-chain information in UCBShift-2.0 will clearly boost the relevance of NMR spectroscopy in this endeavor. The necessary prerequisite for NMR-based evaluation of AF2 models is the availability of at least a backbone resonance assignment. With the advancement of machine learning techniques in biomolecular NMR spectroscopy this bottleneck no longer exists. For example, the ARTificial Intelligence for NMR Applications method (ARTINA)[73] and the NMRtist webserver[74] were introduced for visual spectrum analysis and automated signal assignment starting from raw experimental NMR data. More recently, a novel time-optimized deep learning approach for protein NMR assignment was introduced that employs AlphaFold and chemical shift prediction.[75] Importantly, in this approach the previous version of UCBShift was employed to provide backbone chemical shift data. Clearly, the now available side-chain information will lead to a further improvement of this powerful assignment approach.

While applications to automated NMR signal assignment and validation of AF2 predicted structural models are clearly important and highly relevant, we strongly believe that potentially transformative forward applications will be in the study of protein conformational dynamics. It is common and undisputed knowledge that NMR provides unique insight into the structural dynamics of proteins in solution and numerous NMR spin relaxation experiments have been designed over the years.[76] Importantly, the observed solution chemical shifts in proteins are the result of subtle conformational averaging process and therefore probe the entirety of the proteins' dynamics. An interesting exploitation of this intricate relationship was recently introduced with an approach to assess structural dynamics along the protein backbone via a method called Accuracy of NMR Structures Using the so-called NMR random coil index (RCI)[77] and Rigidity (ANSURR).[78] It calculates the local rigidity of a protein structure[79] and compares it with the local rigidity as measured using a version of the random coil index (RCI) based on backbone NMR chemical shifts. While this method was merely introduced to assess the accuracy of solution structures an extension of this conceptual thinking to side-chain chemical shift data could radically change the way protein dynamics are probed by NMR spectroscopy.

Specifically, there is a well established dependence of ^{13}C side-chain chemical shifts on dihedral angles as a tool for conformational analysis in proteins,[47] and simple relations exist to relate Ile ($^{13}C^{\delta_1}$)[48] and Leu ($^{13}C^{\delta_1,\delta_2}$)[50,80] methyl ^{13}C chemical shifts into χ_2 rotamer populations for these residues. Analogous relationships and procedures exist for obtaining χ_1 torsion angle distributions for Val residues in proteins on the basis of measured $^{13}C^{\gamma_1,\gamma_2}$ shifts exclusively. Importantly, such chemical shifts, when available through relaxation dispersion NMR measurements, can even be used to infer 'invisible' excited protein states[81-83] that are only sparsely and transiently populated. Analogous extensions to other amino acids can be envisaged.

Moreover, NMR spin relaxation measurements obtained for methyl group carbons revealed that, for example, extracted order parameters correlate with conformational averaging along the amino acid side-chain and thus again providing support for the expected correlation between chemical shift and conformational dynamics. While ^{13}C chemical shifts primarily report on side-chain conformational averaging, ^{1}H chemical shifts will be exquisitely sensitive to subtle structural arrangements of aromatic ring systems within a protein. We have recently demonstrated how this information can be used to extract binding poses in protein-ligand complexes with significant ramifications in structure-based drug design programs.[84,85] Most importantly, QM calculations of ligand ^{1}H chemical shifts for the protein-bound state allowed for refinement of the solution structures of protein-ligand (drug) complexes.[84] Developments (exploiting machine learning techniques) are ongoing to improve QM calculations and make them applicable also to larger biomolecular systems.

To conclude, we anticipate that the improved chemical shift prediction tool presented here, and the explosive growth of machine learning to NMR structural biology[2,11,86-88] to generate protein conformational ensembles will offer exciting possibilities that cannot be analyzed using more conventional tools of structural biology. A better quantitative understanding of the intricate relationship between protein structure and NMR chemical shift will be highly valuable in this endeavour.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

A. L. Ptaszek was funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria and received further support from the DosChem doctoral school program faculty of Chemistry, University of Vienna. UC Berkeley investigators thank the National Institutes of Health (2R01GM127627-05) for funding of this work.

References

(1). Paruzzo FM; Hofstetter A; Musil F; De S; Ceriotti M; Emsley L Chemical shifts in molecular solids by machine learning. Nature communications 2018, 9, 4501–4501.

(2). Liu S; Li J; Bennett KC; Ganoe B; Stauch T; Head-Gordon M; Hexemer A; Ushizima D; Head-Gordon T Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. J Phys Chem Lett 2019, 10, 4558–4565. [PubMed: 31305081]

(3). Case DA Molecular dynamics and NMR spin relaxation in proteins. Acc Chem Res 2002, 35, 325–31. [PubMed: 12069616]

(4). Case DA Chemical shifts in biomolecules. Curr. Opin. Struct. Bio 2013, 23, 172–6. [PubMed: 23422068]

(5). Christensen A; Linnet T; Borg M; Boomsma W; Lindorff-Larsen K; Hamelryck T; Jensen J Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics. PLoS ONE 2013, e84123. [PubMed: 24391900]

(6). Dyson HJ; Wright PE Unfolded Proteins and Protein Folding Studied by NMR. Chem. Rev 2004, 104, 3607–3622. [PubMed: 15303830]

(7). Ball KA; Wemmer DE; Head-Gordon T Comparison of structure determination methods for intrinsically disordered amyloid-beta peptides. J Phys Chem B 2014, 118, 6405–16. [PubMed: 24410358]

(8). Bhowmick A; Brookes DH; Yost SR; Dyson HJ; Forman-Kay JD; Gunter D; Head-Gordon M; Hura GL; Pande VS; Wemmer DE; Wright PE; Head-Gordon T Finding Our Way in the Dark Proteome. J Am Chem Soc 2016, 138, 9730–42. [PubMed: 27387657]

(9). Swails J; Zhu T; He X; Case DA AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. Journal of biomolecular NMR 2015, 63, 125–139. [PubMed: 26232926]

(10). Haghighatlari M; Li J; Heidar-Zadeh F; Liu Y; Guan X; Head-Gordon T Learning to Make Chemical Predictions: the Interplay of Feature Representation, Data, and Machine Learning Methods. Chem 2020, 6, 1527–1542. [PubMed: 32695924]

(11). Li J; Liang J; Wang Z; Ptaszek AL; Liu X; Ganoe B; Head-Gordon M; Head-Gordon T Highly Accurate Prediction of NMR Chemical Shifts from Low-Level Quantum Mechanics Calculations Using Machine Learning. J Chem Theory Comput 2024, 20, 2152–2166. [PubMed: 38331423]

(12). Meiler J; Baker D Coupled prediction of protein secondary and tertiary structure. Proceedings of the National Academy of Sciences 2003, 100, 12105–12110.

(13). Shen Y; Bax A SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. Journal of biomolecular NMR 2010, 48, 13–22. [PubMed: 20628786]

(14). Iwadate M; Asakura T; Williamson MP Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database. Journal of biomolecular NMR 1999, 13, 199–211. [PubMed: 10212983]

(15). Neal S; Nip AM; Zhang H; Wishart DS Rapid and accurate calculation of protein 1 H, 13 C and 15 N chemical shifts. Journal of biomolecular NMR 2003, 26, 215–240. [PubMed: 12766419]

(16). Kohlhoff KJ; Robustelli P; Cavalli A; Salvatella X; Vendruscolo M Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. Journal of the American Chemical Society 2009, 131, 13894–13895. [PubMed: 19739624]

(17). Li D-W; Brüschweiler R PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. Journal of biomolecular NMR 2012, 54, 257–265. [PubMed: 22972619]

(18). Li D; Brüschweiler R PPM_One: a static protein structure based chemical shift predictor. Journal of biomolecular NMR 2015, 62, 403–409. [PubMed: 26091586]

(19). Han B; Liu Y; Ginzinger SW; Wishart DS SHIFTX2: significantly improved protein chemical shift prediction. Journal of biomolecular NMR 2011, 50, 43–57. [PubMed: 21448735]

(20). Li J; Bennett KC; Liu Y; Martin MV; Head-Gordon T Accurate prediction of chemical shifts for aqueous protein structure on “Real World” data. Chemical science 2020, 11, 3180–3191. [PubMed: 34122823]

(21). Ulrich EL et al. BioMagResBank. Nucleic Acids Research 2007, 36, D402–D408. [PubMed: 17984079]

(22). Zhang H; Neal S; Wishart DS RefDB: a database of uniformly referenced protein chemical shifts. Journal of biomolecular NMR 2003, 25, 173–195. [PubMed: 12652131]

(23). Dolinsky TJ; Nielsen JE; McCammon JA; Baker NA PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic acids research 2004, 32, W665–W667. [PubMed: 15215472]

(24). Dolinsky TJ; Czodrowski P; Li H; Nielsen JE; Jensen JH; Klebe G; Baker NA PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic acids research 2007, 35, W522–W525. [PubMed: 17488841]

(25). Word JM; Lovell SC; Richardson JS; Richardson DC Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of molecular biology 1999, 285, 1735–1747. [PubMed: 9917408]

(26). Platzer G; Okon M; McIntosh LP pH-dependent random coil 1 H, 13 C, and 15 N chemical shifts of the ionizable amino acids: a guide for protein p K a measurements. Journal of biomolecular NMR 2014, 60, 109–129. [PubMed: 25239571]

(27). Harsch T; Dasch C; Donaubauer H; Baskaran K; Kremer W; Kalbitzer H Stereospecific assignment of the asparagine and glutamine side chain amide protons in random-coil peptides by combination of molecular dynamic simulations with relaxation matrix calculations. Applied Magnetic Resonance 2013, 44, 319–331.

(28). Harsch T; Schneider P; Kieninger B; Donaubauer H; Kalbitzer HR Stereospecific assignment of the asparagine and glutamine sidechain amide protons in proteins from chemical shift analysis. Journal of biomolecular NMR 2017, 67, 157–164. [PubMed: 28197852]

(29). Pedregosa F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.

(30). Olson RS; Urbanowicz RJ; Andrews PC; Lavender NA; Kidd LC; Moore JH Automating biomedical data science through tree-based pipeline optimization. Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30–April 1, 2016, Proceedings, Part I 19. 2016; pp 123–137.

(31). Geurts P; Ernst D; Wehenkel L Extremely randomized trees. Machine learning 2006, 63, 3–42.

(32). Breiman L. Random forests. Machine learning 2001, 45, 5–32.

(33). Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ Basic local alignment search tool. Journal of molecular biology 1990, 215, 403–410. [PubMed: 2231712]

(34). Dong R; Peng Z; Zhang Y; Yang J mTM-align: an algorithm for fast and accurate multiple protein structure alignment. Bioinformatics 2017, 34, 1719–1725.

(35). Needleman SB; Wunsch CD A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 1970, 48, 443–453. [PubMed: 5420325]

(36). Henikoff S; Henikoff JG Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 1992, 89, 10915–10919.

(37). Henikoff S; Henikoff JG Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 1992, 89, 10915–10919.

(38). Kabsch W; Sander C Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules 1983, 22, 2577–2637.

(39). Wishart D; Sykes B; Richards F Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. Journal of Molecular Biology 1991, 222, 311–333. [PubMed: 1960729]

(40). Wagner G; Pardi A; Wuethrich K Hydrogen bond length and proton NMR chemical shifts in proteins. Journal of the American Chemical Society 1983, 105, 5948–5949.

(41). Zhang F; Bruschweiler R Contact Model for the Prediction of NMR N-H Order Parameters in Globular Proteins. Journal of the American Chemical Society 2002, 124, 12654–12655. [PubMed: 12392400]

(42). Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins: Structure, Function, and Bioinformatics 2005, 59, 38–48.

(43). Wimley WC; White SH Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nature structural biology 1996, 3, 842–848. [PubMed: 8836100]

(44). Haigh C; Mallion R Ring current theories in nuclear magnetic resonance. Progress in Nuclear Magnetic Resonance Spectroscopy 1979, 13, 303–344.

(45). Case DA; Aktulga HM; Belfon K; Ben-Shalom I; Brozell SR; Cerutti DS; Cheatham III TE; Cruzeiro VWD; Darden TA; Duke RE; others Amber 2021; University of California, San Francisco, 2021.

(46). MacArthur MW; Thornton JM Protein side-chain conformation: a systematic variation of χ1 mean values with resolution–a consequence of multiple rotameric states? Acta Crystallographica Section D: Biological Crystallography 1999, 55, 994–1004. [PubMed: 10216296]

(47). London RE; Wingad BD; Mueller GA Dependence of amino acid side chain 13C shifts on dihedral angle: application to conformational analysis. Journal of the American Chemical Society 2008, 130, 11097–11105. [PubMed: 18652454]

(48). Hansen DF; Neudecker P; Kay LE Determination of isoleucine side-chain conformations in ground and excited states of proteins from chemical shifts. Journal of the American Chemical Society 2010, 132, 7589–7591. [PubMed: 20465253]

(49). Cavalli A; Salvatella X; Dobson CM; Vendruscolo M Protein structure determination from NMR chemical shifts. Proceedings of the National Academy of Sciences 2007, 104, 9615–9620.

(50). Mulder FA Leucine Side-Chain Conformation and Dynamics in Proteins from 13C NMR Chemical Shifts. ChemBioChem 2009, 10, 1477–1479. [PubMed: 19466705]

(51). Brandl M; Weiss MS; Jabs A; Suhnel J; Hilgenfeld R CH−π-interactions in proteins. Journal of molecular biology 2001, 307, 357–377. [PubMed: 11243825]

(52). Platzer G; Mayer M; Beier A; Brüschweiler S; Fuchs JE; Engelhardt H; Geist L; Bader G; Schörghuber J; Lichtenecker R; others PI by NMR: probing CH−π interactions in protein–ligand complexes by NMR spectroscopy. Angewandte Chemie 2020, 132, 14971–14978.

(53). Scheiner S. Assessment of the presence and strength of H-bonds by means of corrected NMR. Molecules 2016, 21, 1426. [PubMed: 27801801]

(54). Carter-Fenk K; Liu M; Pujal L; Loipersberger M; Tsanai M; Vernon RM; Forman-Kay JD; Head-Gordon M; Heidar-Zadeh F; Head-Gordon T The Energetic Origins of Pi-Pi Contacts in Proteins. J Am Chem Soc 2023, 145, 24836–51. [PubMed: 37917924]

(55). Höfurthner T; Toscano G; Kontaxis G; Beier A; Mayer M; Geist L; McConnell DB; Weinstabl H; Lichtenecker R; Konrat R Synthesis of a 13C-methylene-labeled isoleucine precursor as a useful tool for studying protein side-chain interactions and dynamics. Journal of Biomolecular NMR 2024, 78, 1–8. [PubMed: 37816933]

(56). Toscano G; Höfurthner T; Nagl B; Beier A; Mayer M; Geist L; McConnell DB; Weinstabl H; Konrat R; Lichtenecker RJ 13Cβ-Valine and 13Cγ-Leucine Methine Labeling To Probe Protein Ligand Interaction. ChemBioChem 2024, 25, e202300762. [PubMed: 38294275]

(57). Kumar M; Balaji PV CH… pi interactions in proteins: prevalence, pattern of occurrence, residue propensities, location, and contribution to protein stability. Journal of molecular modeling 2014, 20, 1–14.

(58). Derewenda ZS CH groups as donors in hydrogen bonds: a historical overview and occurrence in proteins and nucleic acids. International Journal of Molecular Sciences 2023, 24, 13165. [PubMed: 37685972]

(59). Hass MA; Hansen DF; Christensen HE; Led JJ; Kay LE Characterization of conformational exchange of a histidine side chain: Protonation, rotamerization, and tautomerization of His61 in plastocyanin from Anabaena variabilis. Journal of the American Chemical Society 2008, 130, 8460–8470. [PubMed: 18540585]

(60). Liepinsh E; Otting G Proton exchange rates from amino acid side chains—implications for image contrast. Magnetic resonance in medicine 1996, 35, 30–42. [PubMed: 8771020]

(61). Kougentakis CM; Grasso EM; Robinson AC; Caro JA; Schlessman JL; Majumdar A; García-Moreno E B Anomalous properties of Lys residues buried in the hydrophobic interior of a protein revealed with 15N-detect NMR spectroscopy. The Journal of Physical Chemistry Letters 2018, 9, 383–387. [PubMed: 29266956]

(62). Iwahara J; Jung Y-S; Clore GM Heteronuclear NMR spectroscopy for lysine NH3 groups in proteins: unique effect of water exchange on 15N transverse relaxation. Journal of the American Chemical Society 2007, 129, 2971–2980. [PubMed: 17300195]

(63). Nguyen D; Chen C; Pettitt BM; Iwahara J NMR methods for characterizing the basic side chains of proteins: electrostatic interactions, hydrogen bonds, and conformational dynamics. Methods in enzymology 2019, 615, 285–332. [PubMed: 30638532]

(64). Henry GD; Sykes BD Determination of the rotational dynamics and pH dependence of the hydrogen exchange rates of the arginine guanidino group using NMR spectroscopy. Journal of Biomolecular NMR 1995, 6, 59–66. [PubMed: 22911578]

(65). Takeda M; Jee J; Ono AM; Terauchi T; Kainosho M Hydrogen exchange study on the hydroxyl groups of serine and threonine residues in proteins and structure refinement using NOE restraints with polar side-chain groups. Journal of the American Chemical Society 2011, 133, 17420–17427. [PubMed: 21955241]

(66). Chen J; Yadav NN; Stait-Gardner T; Gupta A; Price WS; Zheng G Thiol-water proton exchange of glutathione, cysteine, and N-acetylcysteine: Implications for CEST MRI. NMR in Biomedicine 2020, 33, e4188. [PubMed: 31793114]

(67). Takeda M; Jee J; Terauchi T; Kainosho M Detection of the sulfhydryl groups in proteins with slow hydrogen exchange rates and determination of their proton/deuteron fractionation factors using the deuterium-induced effects on the 13Cβ NMR signals. Journal of the American Chemical Society 2010, 132, 6254–6260. [PubMed: 20384326]

(68). Nordstrand K; Åslund F; Meunier S; Holmgren A; Otting G; Berndt KD Direct NMR observation of the Cys-14 thiol proton of reduced Escherichia coli glutaredoxin-3 supports the presence of an active site thiol-thiolate hydrogen bond. FEBS letters 1999, 449, 196–200. [PubMed: 10338131]

(69). Davis DG Proton NMR detection of long-range heteronuclear multiquantum coherences in proteins: the complete assignment of the quaternary aromatic carbon-13 chemical shifts in lysozyme. Journal of the American Chemical Society 1989, 111, 5466–5468.

(70). Weichenberger CX; Sippl MJ NQ-Flipper: recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures. Nucleic acids research 2007, 35, W403–W406. [PubMed: 17478502]

(71). Jumper J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021, 596, 583–589. [PubMed: 34265844]

(72). Agarwal V; McShan AC The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. Nature Chemical Biology 2024, 1–10. [PubMed: 38123656]

(73). Klukowski P; Riek R; Güntert P Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. Nature Communications 2022, 13, 6151.

(74). Klukowski P; Riek R; Güntert P NMRtist: an online platform for automated biomolecular NMR spectra analysis. Bioinformatics 2023, 39, btad066. [PubMed: 36723167]

(75). Klukowski P; Riek R; Guntert P Time-optimized protein NMR assignment with an integrative deep learning approach using AlphaFold and chemical shift prediction. Science Advances 2023, 9, eadi9323. [PubMed: 37992167]

(76). Sekhar A; Kay LE An NMR view of protein dynamics in health and disease. Annual review of biophysics 2019, 48, 297–319.

(77). Berjanskii MV; Wishart DS Application of the random coil index to studying protein flexibility. Journal of biomolecular NMR 2008, 40, 31–48. [PubMed: 17985196]

(78). Fowler NJ; Sljoka A; Williamson MP A method for validating the accuracy of NMR protein structures. Nature Communications 2020, 11, 6321.

(79). Jacobs DJ; Rader AJ; Kuhn LA; Thorpe MF Protein flexibility predictions using graph theory. Proteins: Structure, Function, and Bioinformatics 2001, 44, 150–165.

(80). Hansen DF; Neudecker P; Vallurupalli P; Mulder FA; Kay LE Determination of Leu side-chain conformations in excited protein states by NMR relaxation dispersion. Journal of the American Chemical Society 2010, 132, 42–43. [PubMed: 20000605]

(81). Hansen DF; Vallurupalli P; Kay LE Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. Journal of biomolecular NMR 2008, 41, 113–120. [PubMed: 18574698]

(82). Vallurupalli P; Hansen DF; Kay LE Structures of invisible, excited protein states by relaxation dispersion NMR spectroscopy. Proceedings of the National Academy of Sciences 2008, 105, 11766–11771.

(83). Korzhnev DM; Religa TL; Banachewicz W; Fersht AR; Kay LE A transient and low-populated protein-folding intermediate at atomic resolution. Science 2010, 329, 1312–1316. [PubMed: 20829478]

(84). Platzer G; Ptaszek AL; Böttcher J; Fuchs JE; Geist L; Braun D; McConnell DB; Konrat R; Sánchez-Murcia PA; Mayer M Ligand 1H NMR Chemical Shifts as Accurate Reporters for Protein-Ligand Binding Interfaces in Solution. ChemPhysChem 2024, 25, e202300636. [PubMed: 37955910]

(85). Beier A; Platzer G; Höfurthner T; Ptaszek AL; Lichtenecker RJ; Geist L; Fuchs JE; McConnell DB; Mayer M; Konrat R Probing Protein–Ligand Methyl-π Interaction Geometries through

Chemical Shift Measurements of Selectively Labeled Methyl Groups. Journal of Medicinal Chemistry 2024, 67, 13187–13196. [PubMed: 39069741]

(86). Kuhn S. Applications of machine learning and artificial intelligence in NMR. Magnetic Resonance in Chemistry 2022, 60, 1019–1020. [PubMed: 36225085]

(87). Zhang O; Haghighatlari M; Li J; Liu ZH; Namini A; Teixeira JMC; Forman-Kay JD; Head-Gordon T Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. The Journal of Chemical Physics 2023, 158, 174113. [PubMed: 37144719]

(88). Cortés I; Cuadrado C; Hernández Daranas A; Sarotti AM Machine learning in computational NMR-aided structural elucidation. Frontiers in Natural Products 2023, 2

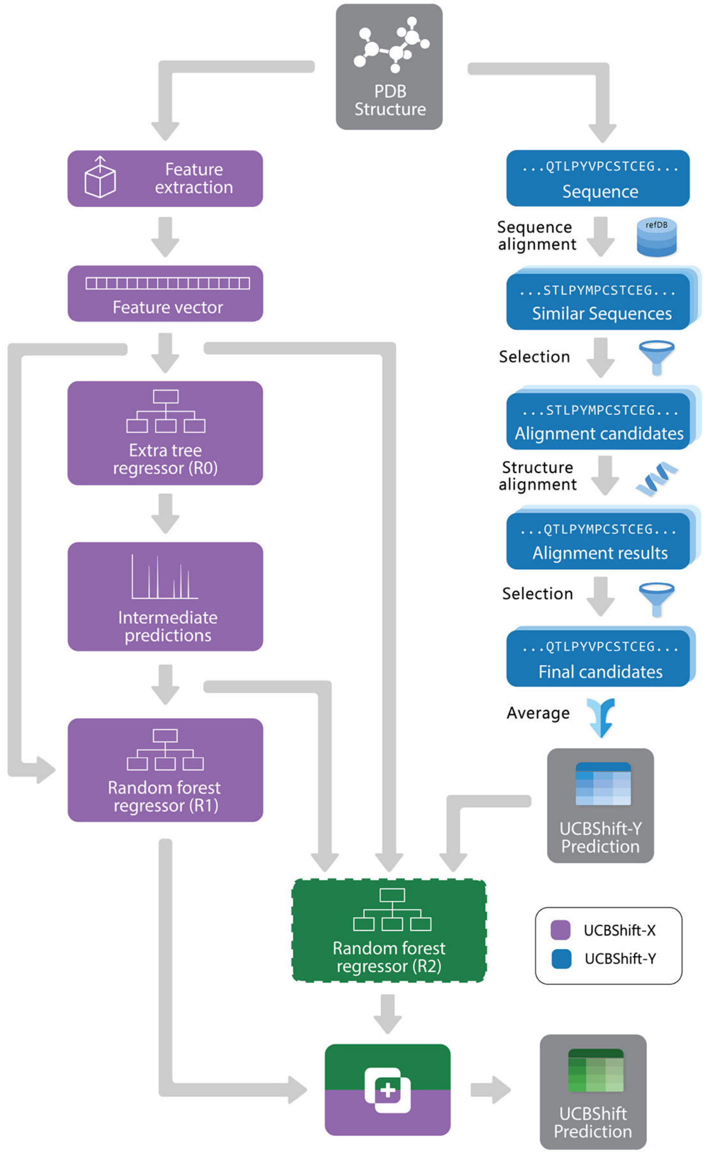


Figure 1: The overall design of the original UCBShift-1.0 chemical shift prediction algorithm used for UCBShift-2.0.

The general UCBShift algorithm combines both a transfer prediction module that relies on both sequence and structural alignments, and a machine learning module that trains a tree regression model on augmented feature extracted data. The feature vector has been augmented for side chains as explained in the text. Reproduced from Ref.[20] with permission from the Royal Society of Chemistry.

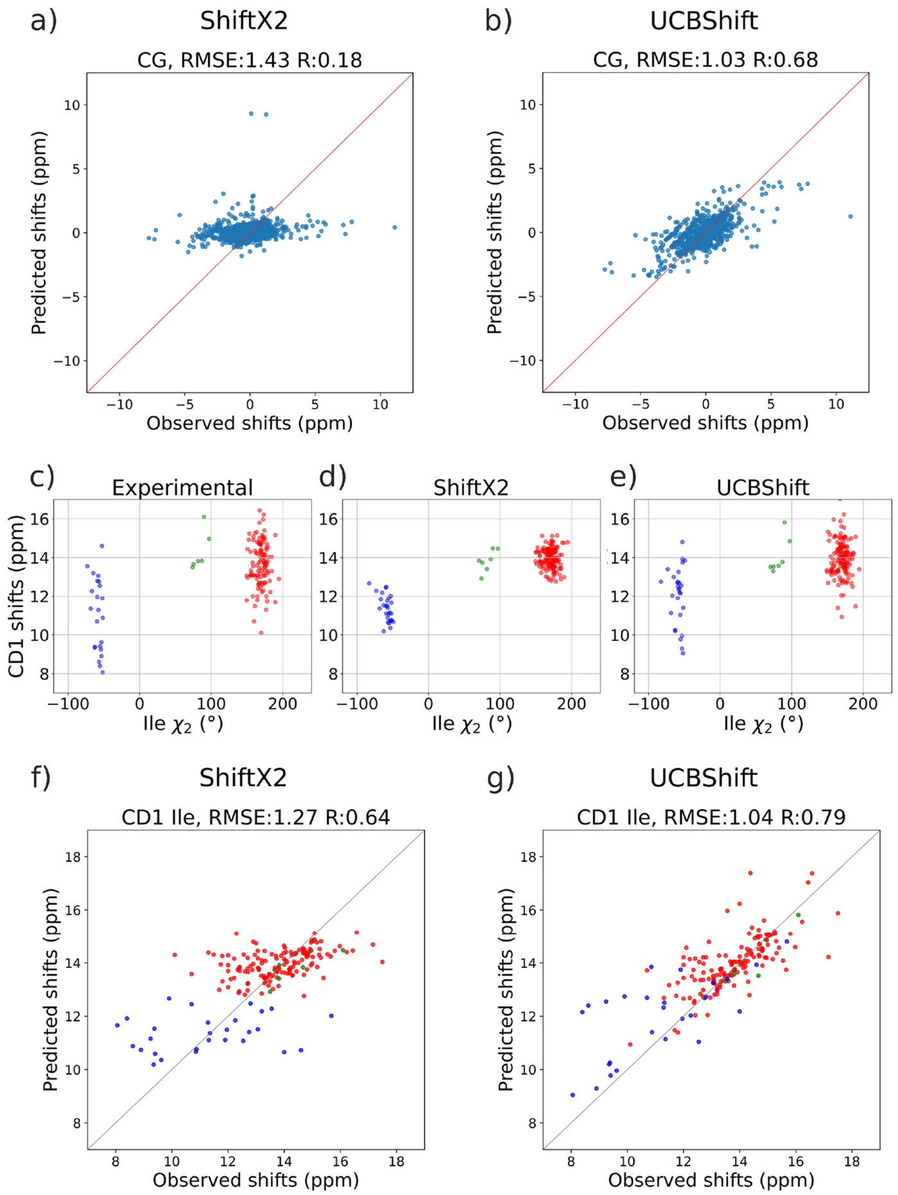


Figure 2: Predicted CG and CD1 chemical shifts by UCBShift and SHIFTX2 compared to experiment.

Correlation between experimental and predicted CG chemical shifts subtracted from the random coil references for (a) SHIFTX2 and (b) UCBShift. (c-e) CD1 chemical shifts of isoleucine plotted against the χ_2 dihedral angle. Correlation between experimental and predicted CD1 chemical shifts of isoleucine for (f) SHIFTX2 and (g) UCBShift. In (c-g), three conformational clusters are marked with different colours.

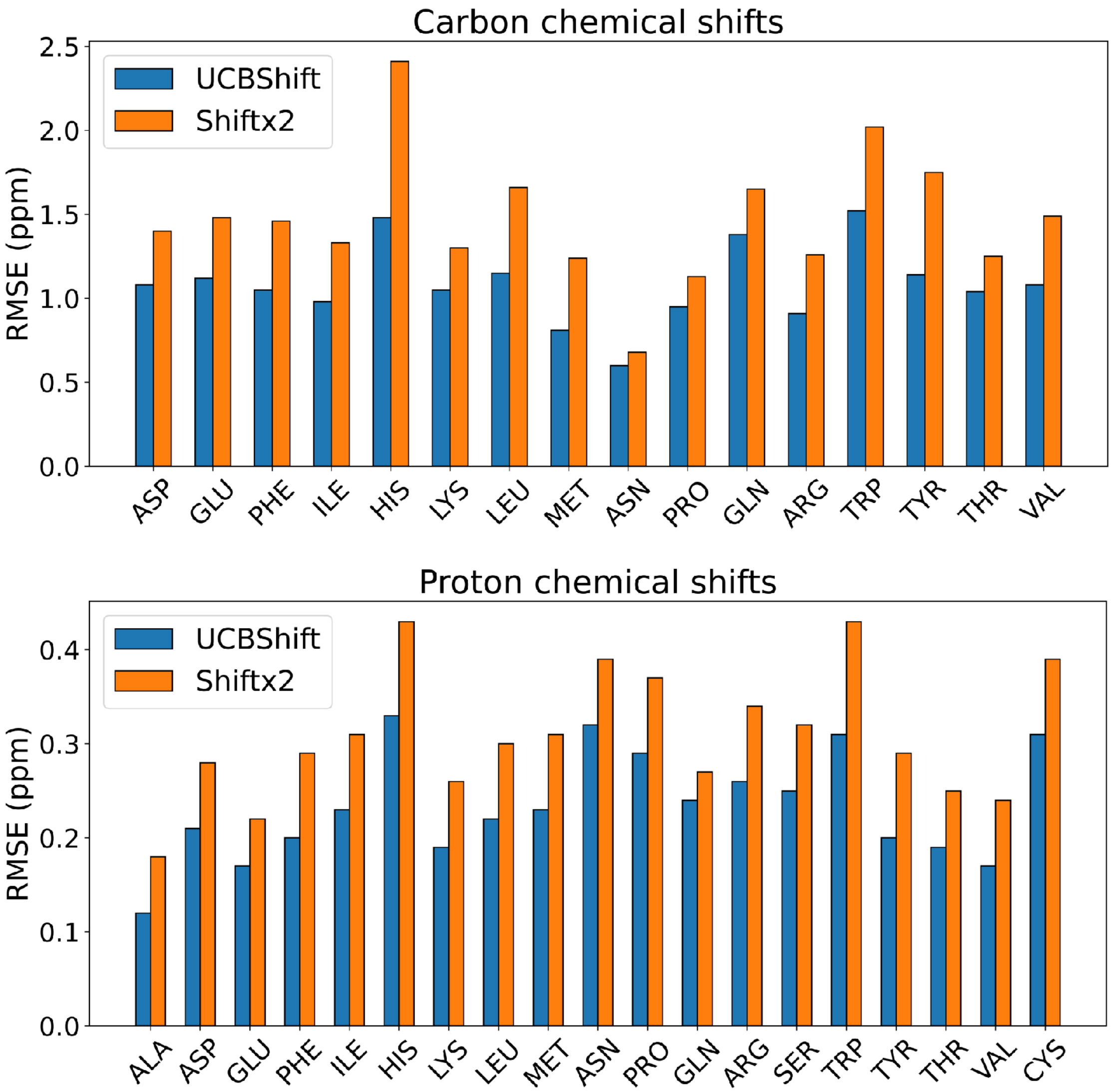


Figure 3:

The root mean square error (RMSE) of carbon and hydrogen chemical shifts for every amino acid predicted by UCBShift and SHIFTX2 compared to experiment.

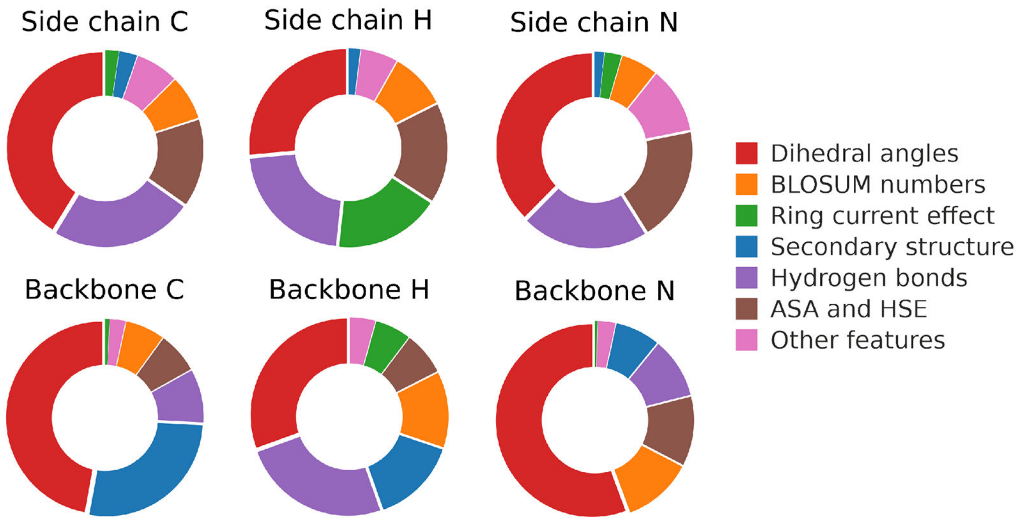
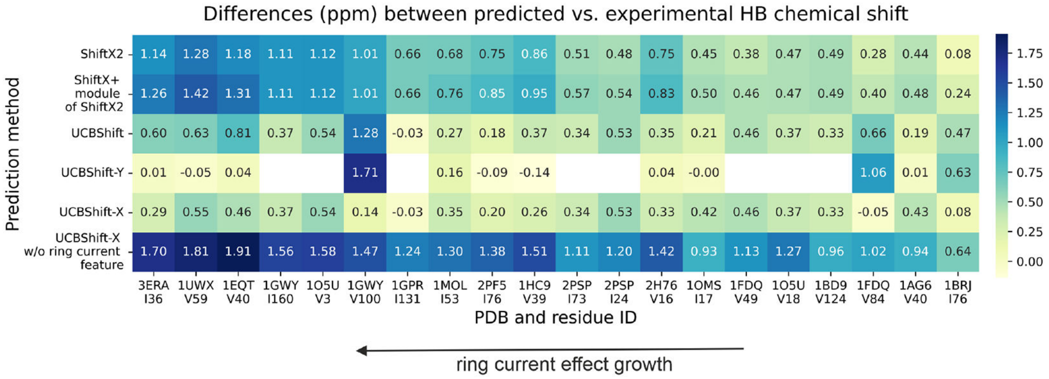


Figure 4: Relative importance of all the input features analyzed from the UCBShift model for side chain and backbone C, H and N atom types.

Importance values from the extra tree regressor (R0) were scaled in proportion to the *"Prediction from R0"* contribution in the random forest regressor (R1), Tables S9-S11. These scaled values were then added to the corresponding importance values from the R1 regressor, giving the overall R0/R1 importance of features. Feature importance was calculated as the mean decrease in impurity across all trees, using the built-in feature importance method from Scikit-learn.[29]



	3ERA I36	1UWX V59	1EQT V40	1GWY I160	1O5U V3	1GWY V100	1GPR I131	1MOL I53	2PF5 I76	1HC9 V39	2PSP I73	2PSP I24	2H76 V16	1OMS I17	1FDQ V49	1O5U V18	1BD9 V124	1FDQ V84	1AG6 V40	1BRJ I76
ShiftX2	1.14	1.28	1.18	1.11	1.12	1.01	0.66	0.68	0.75	0.86	0.51	0.48	0.75	0.45	0.38	0.47	0.49	0.28	0.44	0.08
ShiftX+ module of ShiftX2	1.26	1.42	1.31	1.11	1.12	1.01	0.66	0.76	0.85	0.95	0.57	0.54	0.83	0.50	0.46	0.47	0.49	0.40	0.48	0.24
UCBShift	0.60	0.63	0.81	0.37	0.54	1.28	-0.03	0.27	0.18	0.37	0.34	0.53	0.35	0.21	0.46	0.37	0.33	0.66	0.19	0.47
UCBShift-Y	0.01	-0.05	0.04			1.71		0.16	-0.09	-0.14			0.04	-0.00				1.06	0.01	0.63
UCBShift-X	0.29	0.55	0.46	0.37	0.54	0.14	-0.03	0.35	0.20	0.26	0.34	0.53	0.33	0.42	0.46	0.37	0.33	-0.05	0.43	0.08
UCBShift-X w/o ring current feature	1.70	1.81	1.91	1.56	1.58	1.47	1.24	1.30	1.38	1.51	1.11	1.20	1.42	0.93	1.13	1.27	0.96	1.02	0.94	0.64

Figure 5: Heatmap presenting differences between predicted and experimental HB chemical shifts of Ile and Val from the subset including hydrogens involved in $CH\cdots\pi$ interactions. UCBShift-X with turned off ring current effect feature, UCBShift-X, UCBShift-Y, UCBShift, ShiftX+ module of SHIFTX2 and SHIFTX2 prediction models are considered. Chemical shifts are in ascending order from left to right (from the strongest to the weakest ring current effect). The average absolute error (ppm) across this data set is 0.78 for SHIFTX2, 0.84 for ShiftX+, 0.52 for UCBShift, 0.61 for UCBShift-Y and 0.36 for UCBShift-X.

Table 1:
Number of training and testing set shift-structure examples for every type of atom.

Atom	Train	Test	Atom	Train	Test	Atom	Train	Test
CG	14995	1289	HB2	17423	4554	HG1	1234	493
CD1	4590	621	HB3	16259	4318	HD21	722	184
CG2	4300	731	HG2	9667	2684	HD22	720	185
CD	3467	515	HB	6442	1820	HE	604	162
CG1	2919	472	HD2	6104	1705	HE21	595	164
CD2	2735	367	HG3	5901	1416	HE22	594	162
CE	1649	256	HD1	4203	1234	HZ	558	164
CE1	1090	114	HD3	2911	758	HZ2	242	61
CE2	600	49	HE2	2612	645	HH2	222	59
CZ	428	55	HG	2076	450	HZ3	211	56
CZ2	176	16	HE1	2045	481			
CH2	167	13	HE3	1409	345	ND2	672	120
CE3	156	17	HG12	1373	290	NE2	562	126
CZ3	151	15	HG13	1239	274	NE1	287	47

Table 2:

Root mean square error (RMSE), mean absolute error (MAE) and Pearson’s correlation coefficients (R) for UCBShift and SHIFTX2.

We found that 25% and 3.5% of the testing structures are only predicted by UCBShift-X and ShiftX+ algorithms, respectively, in which no alignment is possible. For R coefficients calculations, chemical shifts were subtracted from the random coil references. Uncertainties were computed based on 50 random samples from 75% of the test data. All units in ppm.

	UCBShift			SHIFTX2			Improvement factors	
Atom	RMSE	MAE	R	RMSE	MAE	R	ΔRMSE	ΔMAE
CG	1.028(2)	0.635(1)	0.68	1.429(3)	0.909(2)	0.18	0.401	0.274
CD1	1.1253(3)	0.719(2)	0.74	1.530(4)	1.087(2)	0.44	0.405	0.368
CG2	1.0037(2)	0.671(1)	0.69	1.311(2)	0.916(1)	0.35	0.307	0.245
CD	1.2102(4)	0.745(2)	0.60	1.445(4)	0.901(3)	0.29	0.235	0.156
CG1	1.0935(4)	0.715(2)	0.69	1.373(4)	0.954(3)	0.44	0.280	0.239
CD2	1.2031(2)	0.846(2)	0.77	1.697(5)	1.252(3)	0.40	0.494	0.406
CE	0.9221(4)	0.584(2)	0.47	1.026(3)	0.671(2)	0.26	0.104	0.087
CE1	0.935(4)	0.637(3)	0.66	1.32(1)	0.889(5)	0.06	0.385	0.252
CE2	0.9975(4)	0.760(5)	0.53	1.128(1)	0.878(4)	0.23	0.131	0.118
CZ	1.5504(2)	0.97(1)	0.64	2.43(2)	1.64(1)	−0.25	0.88	0.67
CZ2	0.9999(8)	0.73(1)	0.60	-	-	-	-	-
CH2	1.5763(2)	1.02(2)	−0.09	-	-	-	-	-
CE3	1.3219(1)	1.08(1)	0.52	-	-	-	-	-
CZ3	1.9102(4)	1.11(2)	0.55	-	-	-	-	-
HB2	0.2257(3)	0.1228(1)	0.77	0.2632(3)	0.1479(1)	0.68	0.0375	0.0251
HB3	0.2278(3)	0.1265(1)	0.77	0.2882(4)	0.1720(2)	0.61	0.0604	0.0455
HG2	0.182(3)	0.0964(1)	0.80	0.2169(3)	0.1266(1)	0.68	0.0349	0.0302
HB	0.1696(5)	0.0906(2)	0.86	0.2115(4)	0.1156(2)	0.78	0.0419	0.0250
HD2	0.2181(1)	0.1116(3)	0.77	0.264(1)	0.1436(2)	0.64	0.046	0.0320
HG3	0.2069(5)	0.1153(2)	0.76	0.2457(5)	0.1394(2)	0.64	0.0388	0.0241
HD1*	0.2095(1)	0.1036(3)	0.82	0.247(1)	0.1385(3)	0.64	0.049	0.0374
HD3	0.2842(1)	0.1421(4)	0.69	0.315(1)	0.1577(4)	0.60	0.031	0.0156
HE2*	0.1668(1)	0.0931(3)	0.82	0.226(1)	0.1269(3)	0.64	0.060	0.0343
HG*	0.2504(1)	0.1430(4)	0.71	0.262(1)	0.1514(5)	0.66	0.015	0.0107
HE1	0.2688(1)	0.1416(5)	0.77	0.322(1)	0.1595(5)	0.66	0.053	0.0179
HE3	0.1698(1)	0.1016(3)	0.82	0.252(1)	0.1293(5)	0.55	0.082	0.0277
HG12	0.2782(1)	0.1965(1)	0.72	0.327(1)	0.228(1)	0.59	0.049	0.032
HG13	0.2932(1)	0.2048(1)	0.76	0.357(1)	0.250(1)	0.61	0.064	0.045
HG1*	0.2911(2)	0.1254(5)	0.58	0.215(1)	0.1221(3)	0.69	0.029	0.0207
HD21	0.402(2)	0.255(1)	0.53	0.498(2)	0.301(1)	0.24	0.096	0.046
HD22	0.317(1)	0.210(1)	0.44	0.361(1)	0.237(1)	0.23	0.044	0.027
HE	0.4575(4)	0.215(1)	0.77	0.526(4)	0.246(1)	0.69	0.069	0.031

	UCBShift			SHIFTX2			Improvement factors	
Atom	**RMSE**	**MAE**	**R**	**RMSE**	**MAE**	**R**	**ΔRMSE**	**ΔMAE**
HE21	0.412(3)	0.231(1)	0.35	0.428(4)	0.215(1)	0.41	0.016	−0.016
HE22	0.255(1)	0.1664(5)	0.47	0.291(1)	0.180(1)	0.35	0.036	0.014
HZ	0.3541(4)	0.172(1)	0.72	0.40(1)	0.179(1)	0.62	0.05	0.007
HZ2	0.2393(1)	0.161(1)	0.67	-	-	-	-	-
HH2	0.2216(1)	0.169(1)	0.62	-	-	-	-	-
HZ3	0.3701(3)	0.251(2)	0.53	-	-	-	-	-
ND2	1.7795(1)	1.02(1)	0.79	-	-	-	-	-
NE2	1.71(2)	0.88(1)	0.73	-	-	-	-	-
NE1	1.4729(1)	1.07(1)	0.77	-	-	-	-	-

*RMSE and MAE excluding atom types which ShiftX2 does not predict. RMSE for HD1, HE2, HG, HG1: 0.198(1), 0.166(1), 0.247(1), 0.186(1), respectively. MAE for HD1, HE2, HG, HG1: 0.1011(3), 0.0926(3), 0.1407(4), 0.1014(3).