Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2020, **11**, 3180

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 29th December 2019
Accepted 2nd March 2020

DOI: 10.1039/c9sc06561j

rsc.li/chemical-science

Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data†

Jie Li,[ab] Kochise C. Bennett,[ab] Yuchen Liu,[ab] Michael V. Martin[c] and Teresa Head-Gordon *[abcd]

Here we report a new machine learning algorithm for protein chemical shift prediction that outperforms existing chemical shift calculators on realistic data that is not heavily curated, nor eliminates test predictions *ad hoc*. Our UCBShift predictor implements two modules: a transfer prediction module that employs both sequence and structural alignment to select reference candidates for experimental chemical shift replication, and a redesigned machine learning module based on random forest regression which utilizes more, and more carefully curated, feature extracted data. When combined together, this new predictor achieves state-of-the-art accuracy for predicting chemical shifts on a randomly selected dataset without careful curation, with root-mean-square errors of 0.31 ppm for amide hydrogens, 0.19 ppm for Hα, 0.84 ppm for C′, 0.81 ppm for Cα, 1.00 ppm for Cβ, and 1.81 ppm for N. When similar sequences or structurally related proteins are available, UCBShift shows superior native state selection from misfolded decoy sets compared to SPARTA+ and SHIFTX2, and even without homology we exceed current prediction accuracy of all other popular chemical shift predictors.

Introduction

Chemical shifts are a readily obtainable NMR observable that can be measured with high accuracy for proteins, and sensitively probe the local electronic environment that can yield quantitative information about protein secondary structure,[1–3] estimation of backbone torsion angles,[4,5] or measuring the exposure of the amino acid residues to solvent.[6] But in order to take full advantage of these high quality NMR measurements, there is a necessary reliance on a computational model that can relate the experimentally measured NMR shifts to structure with high accuracy. Existing methods for chemical shift prediction rely on extensive experimental databases together with useful heuristics to rapidly, but non-rigorously, simulate protein chemical shifts. As of yet, quantum mechanical methods which would in principle provide more rigor to chemical shift prediction are still in progress.[7]

The heuristic chemical shift back-calculators are formulated as approximate analytical models such as shAIC,[8] PPM,[9] and PPM_One,[10] empirical alignment-based methods such as SHIFTY[11] and SPARTA,[12,13] 3D representations for machine learning of chemical shifts in solid state NMR for small molecules,[14,15] and feature-based methods including SHIFTCALC,[16] SHIFTX,[17] PROSHIFT,[18] CamShift,[19] and SPARTA+;[20] in the case of SHIFTX2 [21] both alignment and features are utilized. Some of the most successful alignment-based methods rely on the fact that proteins with similar sequences will also share similar structures which lead to their exhibiting similar chemical shifts. This idea was first exploited by the program SHIFTY[11] which "transferred" the chemical shifts from known sequences in the database to the query sequence based on a global sequence alignment. Higher accuracy was achieved in the formulation of SHIFTY+ by replacing the global sequence alignment with a local sequence alignment, and is included in the most recent chemical shift prediction program SHIFTX2.[21] Alignment-based methods in general yield predictions with higher accuracy when a good sequence homologue is found in the database, and the constant increase in the number of sequences and associated chemical shifts deposited into the Biomolecular Magnetic Resonance Bank (BMRB),[22] suggests that a similar sequence to the query sequence will continue to increase steadily.

On the other hand, methods that are based on sequence alignments will by definition fail if sequence similarity between the query sequence with any sequence in the database is too low, as well as the possibility that similar sequences can adopt very different structural folds.[23] For query sequences with low

[a]*Pitzer Center for Theoretical Chemistry, University of California, Berkeley, CA 94720, USA. E-mail: thg@berkeley.edu*

[b]*Department of Chemistry, University of California, Berkeley, CA 94720, USA*

[c]*Department of Bioengineering, University of California, Berkeley, CA 94720, USA*

[d]*Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA*

† Electronic supplementary information (ESI) available: The exclusion of erroneous chemical shifts assignments in the test dataset and complete compilation of the test dataset. Evaluation of SPARTA+, SHIFTX2 and UCBShift performance on test datasets. Performance analysis of UCBShift-Y in comparison with SHIFTY+ and learning curves for the random forest models. See DOI: 10.1039/c9sc06561j

sequence identity, the analytical or feature extraction methods predict secondary chemical shifts (*i.e.* from a random coil reference[24]) by providing data input formulated as hypersurfaces of structural data attributes such as backbone ϕ, ψ angles and hydrogen bonding derived from X-ray structures or calculated from quantum mechanics, or physical data observables such as ring currents,[25,26] electric field effects,[27] or Lipari–Szabo order parameters,[28] that can be generated from easily parsed computational models. These feature extracted data are then used to establish empirical hypersurfaces such as used in SHIFTS[7,29] and CamShift,[19] or to train a machine learning algorithm in the cases of PROSHIFT,[18] SPARTA+,[20] and the SHIFTX+ component of SHIFTX2.[21]

In the evaluation of chemical shifts calculators like SPARTA+ and SHIFTX2, extensive effort has been put into making the test dataset as clean as possible. However, for chemical shifts of real-world proteins, large deviations from predicted values are typically considered as "outliers" and removed from the test dataset post-prediction.[20] The definition of outliers can be arbitrary and results in higher test performance than operating on a real-world data set that may not necessarily be plagued by experimental errors. In this work we examine the current performance of feature extracted methods represented by SPARTA+, as well as the combination of sequence alignment and feature extracted method as implemented in SHIFTX2 on a randomly selected test dataset with minimal data filtering. First we assess the performance in terms of root mean square error (RMSE) with respect to experimental chemical shifts for SPARTA+ and SHIFTX2 when evaluated on a fully independent set of test proteins with high-resolution (<2.4 Å) X-ray structures. We use chemical shift data deposited in the BMRB, in which protein chemical shifts have been re-referenced with respect to high-resolution X-ray structures using RefDB developed by Wishart and co-workers.[30] We also assess SPARTA+ and SHIFTX2 performance when eliminating putative outliers as determined by test data set filtering using PANAV,[13] or removing test proteins with >30% sequence identity to the training set, which are dataset preparation steps that provide a more fair comparison to the two standard chemical shift calculators.

Furthermore, we show that higher accuracy for chemical shifts can be achieved with an enhanced hybrid algorithm, UCBShift, that makes predictions using machine learning on a more extended set of extracted features and transferring experimental chemical shifts from a database by utilizing both sequence and structural alignments. Although we find that we can realize better RMSEs if we also filter the different aspects of the test data, the resulting UCBShift chemical shift prediction method on all of the data including outliers yields RMSEs that will be at the level of 0.31, 0.19, 0.84, 0.81, 1.00 and 1.81 ppm for H, Hα, C$'$, Cα, Cβ and N respectively when evaluated on any independently generated test sequence.

The improved chemical shift performance of UCBShift can be utilized in several predictive contexts such as detection of erroneous chemical shift assignments and errors in reference shifts, to refine single folded structures[31] or refinement of ensembles such as we have done in our Experimental Inferential Structure Determination Method (EISD)[32,33] for folded and intrinsically disordered proteins (IDPs). To illustrate the usefulness of the UCBShift method, we consider the discrimination among alternative folds or selection of native structures among structural "decoys". We determine that UCBShift is able to identify the native structure of two different proteins that comprise two different decoy data sets with certainty by examining the correlation between predicted and experimental chemical shifts, with significant improvement over SPARTA+ and SHIFTX2 prediction methods when sequence or structural homology is available.

Methods

Preparation of training and new testing datasets

A high-quality database of protein structures and associated accurately referenced chemical shifts are crucial for composing a machine learning approach that can make reliable predictions of the chemical shifts, and for faithfully comparing the performance of existing alternative approaches such as SPARTA+ and SHIFTX2. Several publicly available data sources, including the SPARTA+ training set and the training and testing set for SHIFTX+, were combined into a single training dataset that captures the structure and chemical shift relationship. Since all of these data were used in the development of the original SPARTA+ and SHIFTX2 methods, it ensures that corrections for chemical shift reference values were included in our dataset as well.

Unlike previous incarnations of these data sets, which stripped out all presence of crystal waters and ligands, we generated a data set that retained the small molecules in the crystal structures. Our hypotheses is that for crystal waters especially, they often are highly conserved and functional, and are likely highly populated even in solution NMR experiments.[34,35] Any reported hydrogen atoms in the Research Collaboratory for Structural Bioinformatics protein databank (RCSB or PDB) structures[36] were removed and a systematic approach for adding a complete set of hydrogens used the program REDUCE[60] to keep consistency in the structural data used across all approaches. To ensure more robust training, for each atom type, residues with chemical shifts deviating from the average by 5 standard deviations and residues that DSSP[37] failed to generate secondary structure predictions were removed, which accounted for the removal of 40, 5, 18, 147 and 1 training examples for H, Hα, Cα, Cβ and N shifts, respectively. When stereochemically inequivalent Hα were present, their shifts were averaged. In the creation of data for each of the individual atom types, any residues that do not have recorded chemical shifts in the database were eliminated.

Before excluding redundant chains from the database, there were altogether 852 proteins in the training dataset. Duplicate chains were identified and excluded from our dataset: two chains are regarded as duplicates if the sequences and their structures are exactly the same, or eliminating the shorter sequence if it is a sub-sequence of a longer sequence (which is kept). However, 32 chains in the SPARTA+ dataset were retained because although they had identical sequences, they were found to have different structures and thus different chemical shifts.

After excluding the duplicate chains by this prescription, the number of protein chains in the training dataset decreased to 647. The filtering of the training dataset based on RCI-S^2 [38] in principle excludes flexible residues whose chemical shifts are harder to predict. We did not exclude training data based on RCI-S^2 as was done in some other chemical shift predictors, because a complete training set that covers different prediction difficulties is crucial for obtaining reliable performance for real-world applications. Table 1 reports the total number of training data examples for each of the different atom types. In addition, the RefDB database,[30] which is a database for re-referenced protein chemical shifts assignments extracted and corrected from BMRB, was also compiled for the alignment-based chemical shifts prediction.

Since the training dataset in Table 1 covers all of the data from SPARTA+ and SHIFTX2, a separate test dataset needs to be prepared for a fair comparison of all of the chemical shift programs. Therefore, 200 proteins with high-resolution (<2.4 Å) X-ray structures and with chemical shifts available in the RefDB were selected at random to form a separate test set that do not share the exact same sequence as training structures. These structures were downloaded from RCSB and again hydrogens were added with the REDUCE algorithm.[60] Erroneous chemical shifts assignments were removed from this test dataset, which include 9 (H, Cβ, and N) chemical shifts that were significantly offset from the random coil average; 8 Cβ chemical shifts from cysteines that show strong disagreement with their expected chemical shifts under their oxidation state in the crystal structures, and all C′ chemical shifts from 3 proteins that are anti-correlated with predictions from SPARTA+, SHIFTX2 and UCBShift (see ESI† for details). It is essential to remove these chemical shifts from the test set because of evidence for the existence of experimental or recording errors in these data; but no further processing was done on the test set so that it is a good representation of a "real-world" application.

A more carefully "curated" test dataset based on these 200 proteins was also prepared, which additionally exclude paramagmetic proteins, some Hα chemical shifts that have calculated ring current effect exceeding 1.5 ppm, "outliers" detected by the PANAV program,[13] and chemical shifts corrected by PANAV that are different from their original values by more than 0.3 ppm for hydrogens, 1.0 ppm for carbons and 1.5 ppm for nitrogen. These additional test filters are similar to the procedures used by SPARTA+ and SHIFTX2 in preparing their test datasets.[20,21] A complete list of the 200 testing proteins are given in the ESI (Table S3†). As is inevitable, some of these 200 proteins share high sequence identity with some of the training data, so we also generate test datasets after filtering out proteins with >30% sequence identity to yield a low-homology test set (LH-Test) with 100 test proteins (Table 1).

Machine learning for chemical shift prediction

The new UCBShift chemical shift prediction program is composed of two sub-modules: the transfer prediction module (UCBShift-Y) that utilizes sequence and structural alignments to "transfer" the experimental chemical shift value to the query example, and a machine learning module (UCBShift-X) that learns the mapping between the feature extracted data to the experimental chemical shift in the training data. The overall structure of the algorithm is depicted in Fig. 1.

UCBShift-Y module. UCBShift-Y is similar in spirit to the SHIFTY+ component of SHIFTX2, in that the experimental chemical shift for a given atom type in a given residue can be transferred to the query protein when the sequence of the protein in the database is highly similar or even identical to the sequence of the query protein. However, instead of relying on the sequence alignment alone, we have developed an algorithm that relies on both sequence alignment and structural alignment, which allows for proteins that are highly related in structure but remotely related in sequence to be utilized. The use of structural alignments also prevents proteins that have high sequence similarity but low structural homology to mislead an algorithm to erroneous chemical shifts transfers.

For the UCBShift-Y module, a query sequence is first aligned with all sequences in the RefDB database using the local BLAST algorithm,[39] and the PDB files for all sequences generating significant matches are further aligned with the query PDB structure using the mTM-align algorithm.[40] The alignments were further filtered to only keep those alignments that have TM score greater than 0.8 and an RMSD with the query protein structure that is smaller than 1.75 Å. For each of the aligned PDB sequences, its best alignment with the RefDB sequence is determined using the Needleman–Wunsch alignment.[41] If the residues are exactly the same, the shifts from RefDB are directly

Table 1 Total number of training and testing examples for chemical shift prediction for each atom type. The training set is comprised of the combination of the SPARTA+ training set and the training and testing set for SHIFTX+, and removing all redundant chains. We have developed a new test set comprised of 200 high-resolution proteins with chemical shifts available from RefDB; the test data eliminates duplicate chains, and residues with no deposited chemical shift values. The LH-Test set refers to the subset of the total set of test proteins with only low sequence homology to other proteins such that sequence or structural homology cannot be exploited. We also created two curated test sets which additionally exclude paramagmetic proteins, some Hα chemical shifts that have calculated ring current effect exceeding 1.5 ppm, and "outliers" detected by the PANAV program (13). Further information is provided in Methods and ESI

	# of PDBs	H	Hα	C′	Cα	Cβ	N
Train	647	72 894	56 149	58 228	79 611	70 621	74 896
Test	200	19 120	11 727	8231	13 140	10 139	15 374
Test (curated)	200	18 494	11 240	7861	12 533	9883	14 610
LH-Test	100	8634	4979	3332	5685	4278	6576
LH-Test (curated)	100	8606	4950	3331	5251	4201	6480

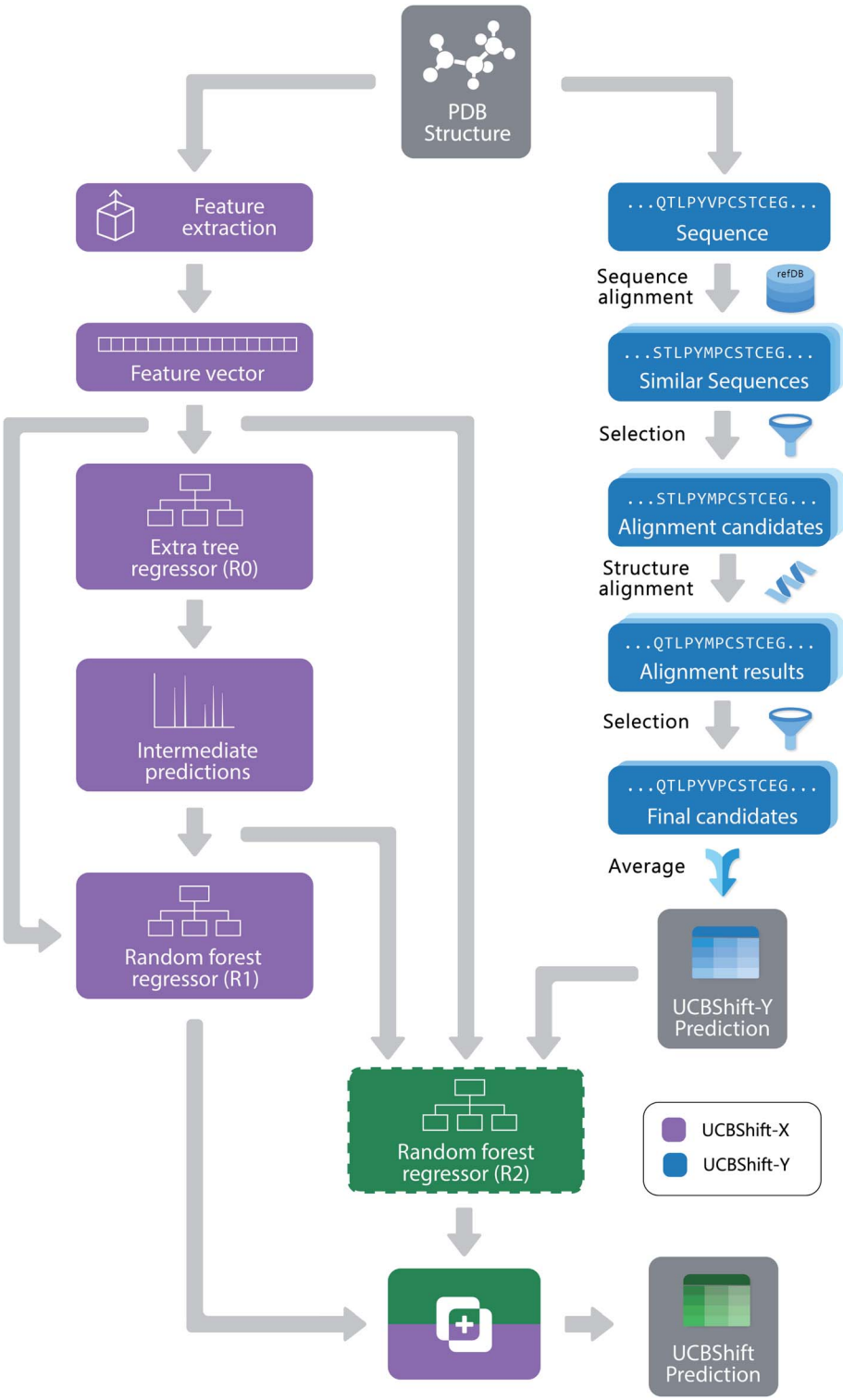

Fig. 1 The overall design of the UCBShift chemical shift prediction algorithm. It combines both a transfer prediction module that relies on both sequence and structural alignments, and a machine learning module that trains a tree regression model on augmented feature extracted data.

transferred to the target; otherwise, the secondary chemical shifts from RefDB are transferred to account for the different chemical shift reference states for different amino acids. To be more specific, the target shift for atom A and residue I is calculated from the matching residue J:

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

where $\delta_{rc,I,A}$ and $\delta_{rc,J,A}$ are the random coil shifts for atom type A in residue I and J, respectively, and $\delta_{J,A}$ is the chemical shift for the matching residue in the database.

When multiple significant structural alignments exist for a given residue, the secondary shifts from these references are averaged with an exponential weighting w_I,

$$w_I = e^{5(S_{NA} \times S_{TM}) + B_{IJ}} \times 1(B_{IJ} \geq 0) \quad (2)$$

given by the normalized sequence alignment score, $S_{NA} = S_{blast}/\max(S_{blast})$, where S_{blast} is the blast score of the matching sequence, and $\max(S_{blast})$ is the maximum blast score from all blast hits; the structure alignment TM score S_{TM} is the pairwise TM score between the query structure and the aligned structure, and B_{IJ} is the substitution likelihood between the residue in the query sequence and the residue in the matching sequence using the BLOSUM62 substitution matrix.[42] Weights with negative substitution scores are set to zero.

UCBShift-X module. The UCBShift-X module requires the formulation of feature extracted data of a given atom type in a query residue, and the ability to map the feature extracted data to the chemical shift value during the training. Similar to the SPARTA+ program or for the SHIFTX+ component of SHIFTX2, we have developed residue-specific features for the query residue and the previous and next residues to the query residue, but we have included more features and polynomial transformations of the features to improve prediction. The feature extracted data generated from the PDB structures of individual residues include:

• 20 numbers representing the score for substituting the residue to any other amino acid, and taken from the BLOSUM62 substitution matrix[42]

• Sine and cosine values of the ϕ and ψ dihedral angles at the residue. Taking the sine and cosine values of the dihedral angles prevents the discontinuity when the dihedral angle goes from +180° to −180°. For the undefined dihedral angles, for example the ϕ angle of a residue at the N-terminus and the ψ angle of a residue at the C-terminus, both the sine and cosine values were set to zero.

• A binary number indicating whether χ_1 or χ_2 dihedral angles for the side chain exists (existence indicator), and the sine and cosine values of these angles when they are defined for the same reasons described for backbone dihedral angles.

• Existence indicators and geometric descriptors for the hydrogen bond between the amide hydrogen and carboxyl oxygen, and between the Cα hydrogen and a carboxyl group (so called α-hydrogen bonds). For each position in the query residue that hydrogen bonds can form, a group of five numbers describe the properties of the hydrogen bond: a boolean number indicating its existence, the distance between the closest hydrogen bond donor–acceptor pair, the cosine values for the angles at the donor hydrogen atom and at the acceptor atom, and the energy of the hydrogen bond calculated with the DSSP model.[37] For the query residue, all hydrogen descriptors for amide hydrogen, carboxyl oxygen and α hydrogen are included, but only the carboxyl oxygen features are included for the previous residue, and the amide hydrogen features for the next residue. These add up to 25 hydrogen bond descriptors for any given residue.

• S^2 order parameters calculated by the contact model[28]

• Absolute and relative accessible surface area produced by the DSSP program.

• Hydrophobicity of the residue by the Wimley–White whole residue hydrophobicity scales.[43]

• Ring current effect calculated by the Haigh–Mallion model.[25,26] For each training model for a specific atom type, the ring current for that atom type is included, while the ring currents for other atom types are excluded from the feature set.

• The one-hot representation of the secondary structure of the residue produced by DSSP program (composed of eight categories)

• Average B factor of the residue extracted from the PDB file.

• Half-sphere exposure of the residue[44]

• Polynomial transformations of some of the residue-specific features, such as the hydrogen bond distances (d_{HB}), by including d_{HB}^2, d_{HB}^{-1}, d_{HB}^{-2}, d_{HB}^{-3}, and the squares of the cosine values of the dihedral angles are also included as additional features. These polynomial quantities have been found to be correlated with secondary chemical shifts, and have occurred in several empirical formulas for calculating chemical shifts.[3,45]

Unlike SPARTA+ and SHIFTX+, we have developed a pipeline with an extra tree regressor[46] followed by random forest regressor[47] as the machine learning based predictor shown in Fig. 1. Both the extra tree regressor and random forest regressor are ensembles of tree regressors that split the data using a subset of the features, and make ensemble-based predictions *via* a majority vote. However, extra tree regressors split the nodes in each tree randomly by selecting an optimal cut-point from uniformly distributed cut-points in the range of the feature, while the random forest regressors calculate the locally optimal cut in a feature by comparing the information entropy difference before and after the split. The random forest regressor learns based on the predictions from the first tree regressor and all the other input features, which can be regarded as a variant of the boosting algorithm,[48] since it learns from the mistakes the first predictor makes. The pipeline was first optimized using the TPOT tool[49] with 3-fold cross validation on the training set, and all the parameters were fine-tuned using a temporal validation dataset with 50 structures randomly selected from the training set. Because tree-based ensemble models are robust to the inclusion of irrelevant features,[50] feature selection was not performed. A more detailed analysis of the feature importance will be given in the Results.

Algorithmically, two separate random forest (RF) regressors are trained. The first RF regressor (R_1) only accepts features extracted from the structure and the prediction from the extra tree regressor, and the second RF regressor (R_2) additionally takes the secondary shift output from UCBShift-Y, together with additional scores and coverage indicating the quality of the alignments, and is trained using only a subset of the training data for which UCBShift-Y is able to make a prediction. Based on the availability of UCBShift-Y predictions, the final prediction of the whole algorithm is generated either by R_1 (when no UCBShift-Y predictions are available) or R_2 (when UCBShift-Y is able to make predictions). Finally, the random coil reference values are added back to the prediction to complete the total

chemical shift prediction, *i.e.* the predictions are calculated with

where f_{R_0} represents the first-level extra tree regressor, f_{R_1} and f_{R_2} are the two second-level random forest regressors, X are all the features extracted from the structure, S are the predictions from

SHIFTX2 is seen to outperform SPARTA+ for chemical shift RMSE for all atom types on the curated dataset when there is high sequence homology for which it was designed, and it performs comparably to SPARTA+ on the curated data for target sequences with low sequence similarity to the training data.

$$\delta_{\text{pred}} = \begin{cases} f_{R_1}\left(X, f_{R_0}(X)\right) + \delta_{\text{RC}} & \text{when UCBShift} - \text{Y generates no prediction} \\ f_{R_2}\left(X, f_{R_0}(X), S\right) + \delta_{\text{RC}} & \text{when UCBShift} - \text{Y generates predictions} \end{cases} \quad (3)$$

UCBShift-Y and the identity scores, and δ_{RC} is the random coil chemical shift for the given residue.

Results

The performance of SPARTA+, SHIFTX2, and UCBShift are evaluated across the newly created test dataset of 200 proteins (test) and the subset of 100 low sequence homology with respect to the training set (LH-Test), each of which is uncurated or curated as described in Methods (Table 2). The mean average errors (MAEs) and correlation coefficients (R^2) are available in Table S4.† In general, the performance of SPARTA+ is even across both the curated test and curated LH-Test datasets. The average RMSE error for SPARTA+ (and for all chemical shift predictors) on the uncurated Test and uncurated LH-Test datasets increases further, in which we provide the minimum error and the maximum error for each protein for SPARTA+ in graphical form in Fig. S2 and S3.†

However, we find that the actual performance on curated data set is less accurate than the reported performance of the SHIFTX2 method.[21] One possible explanation is that a sequence similarity analysis revealed that out of the original 61 testing proteins of SHIFTX2, 4 proteins had 100% sequence alignment with a protein in the training dataset, sometimes under different identification numbers (Table S5†). This problem of training data leakage into the testing data of the original SHIFTX2 method could be a non-trivial source of the better performance of SHIFTX2 reported in the literature. The protein-specific average RMSE error and the scatter plots for the SHIFTX2 predicted and experimental shifts are also given in Fig. S2 and S3† on the uncurated test dataset.

By comparison we find that filtering of the test set for outliers that disagree with the predictions, the elimination of paramagnetic proteins, and removing test shifts for hydrogen due to potentially inaccurate and large ring currents effects has more limited effect on prediction performance. To illustrate, the

Table 2 Test set RMSE between predicted and experimental chemical shifts of SPARTA+, SHIFTX2, and UCBShift of relevant atom types found in proteins. We compare the performance of SPARTA+[a], SHIFTX2[b], and UCBShift across an independently generated uncurated test dataset of 200 proteins that do not share the same sequence as the training set (Test) and a subset of 100 proteins with <30% sequence identity to the training set (LH-Test). We also compare the 3 methods against curated test data that removes "outliers" according to SHIFTX2 and SPARTA+ standards. Uncertainties are calculated from 50 random draws of 75% of the test data. All in units of ppm

Dataset	Test			LH-Test		
Atom type	SPARTA+	SHIFTX2	UCBShift	SPARTA+	SHIFTX2	UCBShift
H	0.51 ± 0.003	0.44 ± 0.003	**0.31 ± 0.003**	0.49 ± 0.004	0.49 ± 0.003	**0.45 ± 0.004**
Hα	0.27 ± 0.002	0.23 ± 0.003	**0.19 ± 0.002**	0.27 ± 0.003	**0.26 ± 0.003**	**0.26 ± 0.003**
C′	1.25 ± 0.01	1.16 ± 0.01	**0.84 ± 0.01**	1.16 ± 0.01	1.20 ± 0.01	**1.14 ± 0.01**
Cα	1.16 ± 0.01	1.05 ± 0.01	**0.81 ± 0.01**	1.13 ± 0.02	1.15 ± 0.01	**1.09 ± 0.01**
Cβ	1.35 ± 0.02	1.27 ± 0.03	**1.00 ± 0.03**	1.36 ± 0.05	1.37 ± 0.06	**1.34 ± 0.05**
N	2.72 ± 0.02	2.40 ± 0.02	**1.81 ± 0.02**	2.73 ± 0.02	2.73 ± 0.03	**2.61 ± 0.02**
Dataset	Test (curated)			LH-Test (curated)		
Atom type	SPARTA+	SHIFTX2	UCBShift	SPARTA+	SHIFTX2	UCBShift
H	0.49 ± 0.002	0.42 ± 0.002	**0.30 ± 0.002**	0.48 ± 0.003	0.47 ± 0.003	**0.43 ± 0.003**
Hα	0.26 ± 0.002	0.22 ± 0.002	**0.18 ± 0.002**	0.26 ± 0.003	0.25 ± 0.002	**0.24 ± 0.003**
C′	1.15 ± 0.009	1.06 ± 0.009	**0.77 ± 0.008**	1.16 ± 0.01	1.19 ± 0.01	**1.13 ± 0.01**
Cα	1.09 ± 0.008	0.98 ± 0.009	**0.76 ± 0.009**	1.08 ± 0.01	1.10 ± 0.01	**1.04 ± 0.01**
Cβ	1.17 ± 0.009	1.09 ± 0.02	**0.82 ± 0.01**	1.15 ± 0.01	1.15 ± 0.01	**1.12 ± 0.02**
N	2.59 ± 0.02	2.25 ± 0.02	**1.71 ± 0.01**	2.67 ± 0.02	2.66 ± 0.02	**2.55 ± 0.02**

[a] Reported SPARTA+ values from ref. 20: 0.49 ppm for H, 0.25 ppm for Hα, 1.09 ppm for C′, 0.94 ppm for Cα, 1.14 ppm for Cβ, and 2.45 ppm for N. [b] Reported SHIFTX2 values from ref. 51: 0.17 ppm for H, 0.12 ppm for Hα, 0.53 ppm for C′, 0.44 ppm for Cα, 0.52 ppm for Cβ, and 1.12 ppm for N.

distributions of absolute errors from SPARTA+ for paramagnetic proteins and diamagnetic proteins in the Test dataset are shown in Fig. S4.† The error distributions are not that different for H, Hα, Cβ, and N, and while paramagnetic proteins show higher prediction errors than diamagnetic proteins for the C′ and Cα data types, they are not egregious errors.

We find that UCBShift outperforms SPARTA+ and the SHIFTX2 algorithm for chemical shift prediction RMSE when tested on the uncurated test data set, the more carefully curated test data, and regardless of the level of sequence homology. The protein-specific average RMSE error and the scatter plots for the UCBShift predicted and experimental shifts are also given in Fig. S2 and S3† on the uncurated Test dataset. Therefore UCBShift is more accurate for real-world applications, where the types of proteins may be more diverse than the test sets for SPARTA+ and SHIFTX2. In order to understand the improved performance of UCBShift in particular, we analyze the components of the algorithm including the UCBShift-X and UCBShift-Y modules, as well as the importance of the extracted features, utilizing the full test set of 200 proteins in more detail below.

Analysis of UCBShift-Y module

The major difference of our transfer prediction module (UCBShift-Y) in comparison with SHIFTY or SHIFTY+ is the inclusion of a structural alignment to select reference sequences for transfer of the chemical shift value. There is a trade-off between the coverage UCBShift-Y can achieve and the average accuracy of the prediction, requiring tuning the thresholds for accepting an imperfectly aligned protein as reference. Empirically we chose a relatively permissive threshold for sequence alignment to enable sequences that do not have much similarity with the query protein proceed to the next step in case it generates a good structure alignment. The thresholds for the TM score and RMSD in structure alignment were optimized during training to ensure the reference structures are close enough to the query structure. A TM score threshold of 0.8 was selected because NMR chemical shifts are sensitive to local structures, and only a well-aligned structure provides reliable chemical shift references.

We hypothesized that a structure-based alignment followed by sequence alignment would be more reliable since it would (1) allow for transferring shifts from structurally homologous proteins with low sequence identity, while also (2) ensuring that the transferred chemical shift values are not from a protein that has high sequence similarity but low structural homology with the query protein. This is confirmed in Fig. 2 which plots the difference of the RMSE on amide hydrogen chemical shift prediction between our UCBShift-Y and SHIFTY+ as a function of sequence identity. Plots for other atom types are given in Fig. S5.† Here the sequence identity is defined as the ratio of the number of matched residues to either the length of the query sequence or the length of the matched sequence, whichever is longer. Furthermore, the UCBShift results are reported with a specifically designed "test mode" which will not utilize sequences with more than 99% identity with the query sequence for making the prediction; this practice ensures the testing performance is a more realistic reflection of the actual performance when operating on input data which is not included in the search database. It is evident that on average predictions on query sequences with low sequence similarity but high structural homology are greatly improved with UCBShift-Y.

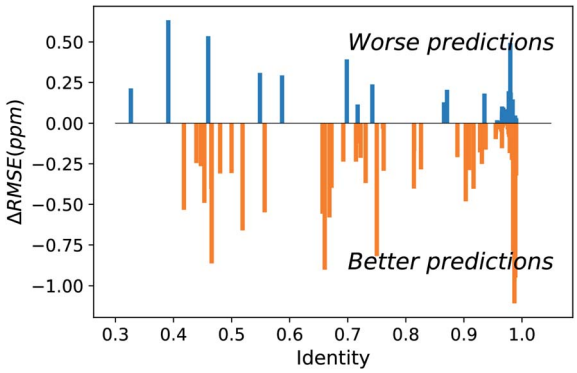

Fig. 2 Difference between UCBShift-Y and SHIFTY+ for protein specific RMSEs for amide hydrogens as a function of sequence identity. The presence of more negative values indicates UCBShift-Y is making better predictions than SHIFTY+ across the range of sequence identity. The better RMSE even at low sequence identity arises from finding a structural homolog.

A particularly interesting example is the prediction for adenylate kinase (PDB ID: 4AKE)[52] and its mutant (PDB ID: 1E4V),[53] both of which are identical but for a single substitution of a valine for a glycine residue at position 10 (Fig. 3). Even with such high sequence identity, these two proteins adopt quite different tertiary structures with a backbone RMSD of 7.08 Å as can be seen from the overlay of their two structures in Fig. 3a. Hence while the experimental chemical shifts for these two proteins have a root-mean-square difference (RMSD) of 0.38 ppm for amide hydrogen shifts overall, the maximum H chemical shift difference is much larger at 1.34 ppm and is reflected in the surprisingly lower correlation (R-value = 0.86) between the amide hydrogen shifts for two proteins given the high sequence similarity (Fig. 3b). Therefore when using SHIFTY or SHIFTY+ for the 1E4V query sequence, the best sequence match will be 4AKE, thus increasing the chemical shift prediction error due to the huge structural deviation between the two proteins. Instead when using our UCBShift-Y module it selects two alternative proteins, 1AKE and 2CDN, which share an average sequence similarity of only 67% with the query protein. The correlation between the predicted 1E4V amide hydrogen shifts with UCBShift-Y which chooses references based on structural alignment and the experimental values are given in Fig. 3c, raising the R-value to 0.94 and lowering the RMSE to 0.25 ppm.

Fig. S6† summarizes the results of UCBShift-Y *vs.* SHIFTY+ for chemical shift prediction for all atom types, validating that the structural alignments successfully found better reference proteins for the query protein which improved the overall prediction quality. In comparison with SHIFTY+, all atom types other than carboxyl carbon are improved in accuracy; although

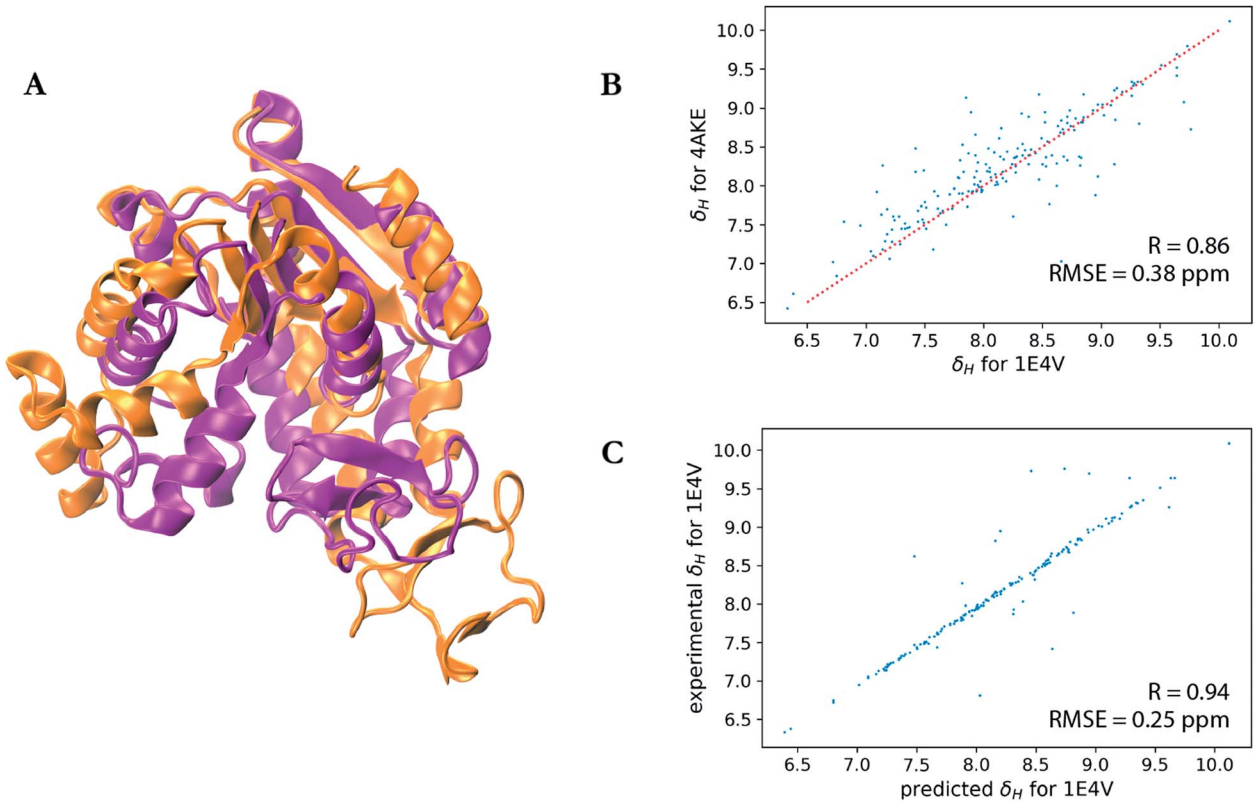

Fig. 3 Analysis of the transfer prediction module for UCBShift which uses sequence and structural alignment. (a) Structural alignment of adenylate kinase (4AKE, orange) and the mutant G10V of adenylate kinase (1E4V, purple). (b) Correlation between experimental chemical shifts of the amide hydrogen for 4AKE and 1E4V. (c) Correlation between predicted amide hydrogen chemical shifts using UCBShift with experimental values. In this case structural alignments instead of sequence alignments were used for selecting references for the transfer prediction.

predictions for the carboxyl carbons are at the same level of accuracy as SHIFTY+, the failure to improve this atom type with UCBShift-Y is likely due to the lower number of chemical shifts available for transfer prediction for this atom type. Finally we note that our UCBShift-Y can be used as a standalone chemical shift predictor when sequence and structural alignments exist and have available experimental chemical shifts.

Analysis of the UCBShift-X module

The connection between features extracted from PDB files and the secondary chemical shifts was explored using several machine learning methods, including neural networks with a single hidden layer, deep fully-connected neural networks, residual neural networks, convolutional neural networks, recurrent neural networks, as well as tree-based ensemble models. The more complex and deeper neural networks performed well on the training dataset and validation dataset, however their performance on the test data was found to be no better than SPARTA+ or SHIFTX+, likely indicating that more feature extracted data is needed and/or due to problems with data representation, to fully exploit the potential of these methods. Thus the tree-based ensemble models stood out as a more competitive machine learning predictor for chemical shifts with limited data. Even so, the learning curves for the random forest models show that the cross-validation error steadily decreases as the number of training examples increases (Fig. S7†), suggesting even better predictions can be achieved if more training data were available.

The RMSE of the pipeline with extra tree regressor and random forest regressor but without inputs from UCBShift-Y (R_1) between the predicted chemical shifts and the observed shifts is summarized in Table 3 and named UCBShift-X. It is found to be statistically better for all the atom types when compared with SPARTA+, or the SHIFTX+ component of SHIFTX2, which also use no sequence and/or structural

Table 3 RMSE for the individual elements of transfer prediction (UCBShift-Y) and machine learning module (UCBShift-X) on the test dataset. The standalone UCBShift-Y prediction when sequence and structural alignments exist and have available experimental chemical shifts. The chemical shift prediction of the machine learning module (UCBShift-X) which is trained independent of any transfer prediction. The prediction of the R_2 module with input from UCBShift-Y module, and the combined R_1 and R_2 modules that defines the UCBShift calculator

UCBShift components	H	Hα	C′	Cα	Cβ	N
UCBShift-X (R_1)	0.44	0.25	1.17	1.08	1.28	2.49
UCBShift-Y	0.21	0.17	0.64	0.57	0.67	1.25
ML with UCBShift-Y input (R_2)	0.19	0.15	0.66	0.57	0.70	1.23
UCBShift (utilizing both R_1 and R_2)	0.31	0.19	0.84	0.81	1.00	1.81

alignments. The overall performance of the UCBShift-X machine learning module is promising, and it also can be used as a reliable standalone predictor for chemical shifts, especially when no faithful alignment is found using UCBShift-Y.

If we consider using the R_2 module (which is trained using only a subset of the training data for which UCBShift-Y is able to make a prediction), the errors of some atom types further decrease (Table 3). Interestingly, the averaged RMSE from R_2 for H, Hα and N is even smaller than the average RMSE of UCBShift-Y, indicating that the second ML module is doing better than just combining the results from UCBShift-Y and from the first level machine learning module R_0 for these atom types. But given the uncertainties in sequence and structural alignments or the lack of chemical shift data for UCBShift-Y, both the R_1 and R_2 machine learning modules are utilized to yield the final UCBShift algorithm and results for chemical shifts as given in Table 3 for all the six atom types.

Analysis of the data representation

A further test is done to analyze the contributions of different features extracted from the structural PDB files to the R_0, R_1, and R_2 pipelines that define the UCBShift algorithm (Fig. 1). Relative feature importance is calculated as the total decrease in node impurity weighted by the probability of reaching a node decided with that feature, and averaged over all the trees in the ensemble.[54] The results are analyzed on the predictions for amide hydrogen as a working example, and are given in Table 4. For the R_0 module we find that the most predictive features are the backbone dihedral angles, the secondary structure, BLOSUM numbers, hydrogen bond features, and the ring current effect which are included in SPARTA+ and SHIFTX2. However, the polynominal transformations of the structural data and half surface exposure are unique in our feature set, and they have very high importance among all the features for the R_0 component. Not surprisingly, the R_0 input is nearly half of the important features for R_1, but the backbone dihedral angles and the polynomial transformations account for an additional ~25% of the important extracted features.

The UCBShift-Y prediction as well as the prediction from R_0 are the dominant factors for the R_2 model; this result indicates the network is indeed trying to differentiate situations when UCBShift-X predictions are more reliable and when they are not so accurate in comparison to R_0 predictions, as well as based on the other structure-derived features. Therefore, using machine learning to combine the predictions from feature-based prediction and alignment-based prediction is a better strategy than doing a weighted average of the two predictions. Finally features such as hydrophobicity and pH values, and to some extent *B*-factors and S^2 order parameters, seem to play a minor role in predictive capacity of the ML module.

Table 4 Importance of different input features into the R_0, R_1, and R_2 pipelines of the machine learning modules

Feature categories	R_0	R_1	R_2
Backbone dihedral angles	0.22	0.11	0.04
Transformed features	0.23	0.11	0.04
Secondary structure	0.17	0.005	0.001
BLOSUM numbers	0.11	0.02	0.005
Hydrogen bond	0.11	0.06	0.02
Half surface exposure	0.05	0.06	0.008
Ring current	0.04	0.03	0.005
Sidechain dihedral angles	0.03	0.03	0.005
Atomic surface area	0.02	0.01	0.002
B Factor	0.008	0.01	0.002
S^2 order parameters	0.006	0.01	0.002
Hydrophobicity	0.002	0.001	0.0002
pH values	0.001	0.002	0.001
Prediction from R_0	N/A	0.53	0.19
UCBShift-Y prediction	N/A	N/A	0.67
UCBShift-Y metrics	N/A	N/A	0.01

Application of UCBShift to protein structure discrimination

One practical application of an accurate protein chemical shift calculator is to detect native structures based on the correlation between predicted and experimental chemical shifts.[55] To illustrate UCBShift's applicability to determine the native structure of a protein, we obtained two decoy datasets that have a range of altered and misfolded structures as measured by the α-carbon root mean square deviation (RMSD) with the native state. The average correlation coefficients for each structure with available experimental chemical shifts of H, Hα and N from BMRB 4429 (1CTF) and BMRB 4811 (1HFZ) are plotted over RMSDs to their native structures in Fig. 4.

The decoy dataset for PDB structure 1CTF was obtained from the Decoy 'R' Us database[56] which contains the native structure and 630 structures with a range of 1.3–9.1 Å in the α-carbon RMSD, and the decoy dataset for PDB structure 1HFZ generated using 3DRobot[57] which contains the native structure and 298 structures with a range of 0.4–4.2 Å in the α-carbon RMSD. Using UCBShift-X alone has similar discriminative power for the native state as SPARTA+, and predicted chemical shifts for lower RMSD structures also tend to have better correlation with experimental values using UCBShift-X (Fig. 4a and c). The complete UCBShift method shows greater discriminative power for the native state than found with either SPARTA+ or SHIFTX2, in which we can differentiate between structures within experimental resolutions (<2 Å) against unlikely structures more easily, while still retaining the highest correlation for the (experimental) native structure (Fig. 4b and d).

Discussion and conclusion

Prediction of protein chemical shifts from structure has relied on robust and popular algorithms such as SPARTA+ and SHIFTX2 that represent the 3-dimensional structure by a set of extracted features that are presented to a machine learning algorithm, sometimes supplemented with direct transfer of experimental data taken on related proteins of a given query sequence. In this paper, we tested the performance of SPARTA+ and SHIFTX2 on a large test set of proteins not previously encountered in previous training and test sets, and showed that SPARTA+ performs as reported and evenly across high and low

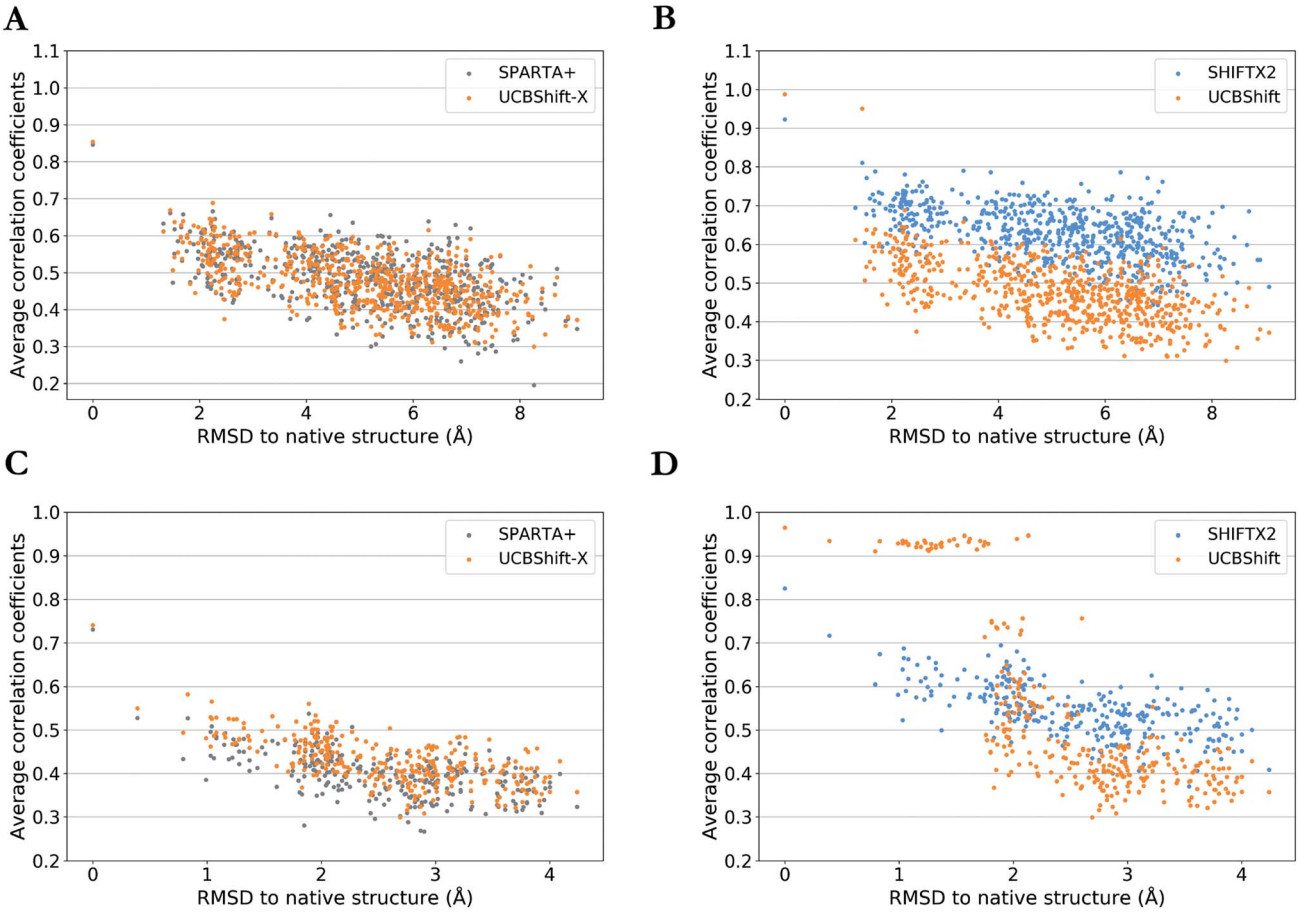

Fig. 4 Average correlation coefficients between predicted chemical shifts of decoyed structures and observed chemical shifts *versus* Cα RMSD between decoyed structures and native structure for PDB 1CTF (a and b) and 1HFZ (c and d). Results are visualized as UCBShift-X compared to SPARTA+ (a and c) and UCBShift compared to SHIFTX2 (b and d).

sequence homology test data, as expected. SHIFTX2 still outperforms SPARTA+ on test sequences with high sequence homology, but not at the same levels expected from the reported RMSE literature values.[51] This test dataset contains "outliers" which may be harder to predict, and hence is a more faithful representation of actual real-world data. We have also developed and tested a new generation algorithm, UCBShift, for solution chemical shift prediction for all relevant protein atom types, and utilizing more small molecular structure information (water and ligands), physically inspired non-linear transformation of features derived from structure, together with a two-level machine learning pipeline that exploits sequence as well as structural alignments to achieve this current state-of-the-art performance. The feature extraction algorithm, the UCBShift prediction program, and all training and testing data can be downloaded from a publicly available github repository https://github.com/THGLab/CSpred.

Although the performance of these algorithms are much better when applied to carefully curated test data, the filtering out of test data risks the inability to distinguish between a poor prediction from a poor experimental chemical shift value. Large outliers would certainly result from the wrong random coil reference for Cβ shifts due to ambiguous cysteine oxidation states, or single whole proteins which exhibit many chemical shift outliers for particular atom types, and should not be considered a failure in algorithmic performance, but a problem of the experimental data. However, further test filtering can start to become arbitrary as we move from deviant to suspicious to acceptable experimental agreement with *the prediction*; one can't have it both ways. Thus in this paper we have provided a realistic range of test performance since scientists use these chemical shift predictors on real-world data that may differ from the original training datasets such that the algorithms do not generalize well-*i.e.* some measure of disagreement with experiment may just simply be prediction error. As such Table 2 provides a more realistic range of test reliability for all three methods.

Although we have realized noticeable improvement over other protein chemical shifts predictors, we believe we are reaching the limit of accuracy by using extracted feature from structures or transfer predictions through alignments. Deep learning may be helpful in the next step since it can operate directly on 3D data representations without the potential bias introduced by features extracted by human experts,[58,59] as we and others have shown recently for chemical prediction in the solid state,[14,15] in which we greatly improved prediction for all

atom types and approached chemical accuracy on par with *ab initio* calculations for hydrogen in particular. The ability to move to 3D representations will be important for QM chemical shift predictions for intrinsically disordered proteins, since feature extracted data will be less available and likely less representative for this class of protein, and the results can be ensemble averaged to provide a prediction that can be compared against solution NMR experimental data for structural ensemble refinement as we have shown for IDPs.[32,33]

Funding

This work was supported by the National Institutes of Health [5U01GM121667]; and the computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy [DE-AC02-05CH11231]. KCB thanks the California Alliance Postdoctoral Fellowship for early support of this work.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Mojtaba Haghighatlari, Brad Ganoe, Tim Stauch, Lars Urban, Shuai Liu, and Martin Head-Gordon for fruitful discussions.

References

1 V. Saudek, A. Pastore, M. A. C. Morelli, R. Frank, H. Gausepohl, T. Gibson, F. Weih and P. Roesch, *Protein Eng., Des. Sel.*, 1990, **4**, 3–10.
2 M. P. Williamson, *Biopolymers*, 1990, **29**, 1423–1431.
3 D. S. Wishart, B. D. Sykes and F. M. Richards, *J. Mol. Biol.*, 1991, **222**, 311–333.
4 S. Spera and A. Bax, *J. Am. Chem. Soc.*, 1991, **113**, 5490–5492.
5 D. S. Wishart and A. M. Nip, *Biochem. Cell Biol.*, 1998, **76**, 153–163.
6 W. F. Vranken and W. Rieping, *BMC Struct. Biol.*, 2009, **9**, 20.
7 D. A. Case, *Curr. Opin. Struct. Biol.*, 2013, **23**, 172–176.
8 J. T. Nielsen, H. R. Eghbalnia and N. C. Nielsen, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2012, **60**, 1–28.
9 D.-W. Li and R. Brüschweiler, *J. Biomol. NMR*, 2012, **54**, 257–265.
10 D. Li and R. Brüschweiler, *J. Biomol. NMR*, 2015, **62**, 403–409.
11 D. S. Wishart, M. S. Watson, R. F. Boyko and B. D. Sykes, *J. Biomol. NMR*, 1997, **10**, 329–336.
12 Y. Shen and A. Bax, *J. Biomol. NMR*, 2007, **38**, 289–302.
13 B. Wang, Y. Wang and D. S. Wishart, *J. Biomol. NMR*, 2010, **47**, 85–99.
14 S. Liu, J. Li, K. C. Bennett, B. Ganoe, T. Stauch, M. Head-Gordon, A. Hexemer, D. Ushizima and T. Head-Gordon, *J. Phys. Chem. Lett.*, 2019, **10**, 4558–4565.
15 F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Nat. Commun.*, 2018, **9**, 4501.
16 M. Iwadate, T. Asakura and M. P. Williamson, *J. Biomol. NMR*, 1999, **13**, 199–211.
17 S. Neal, A. M. Nip, H. Zhang and D. S. Wishart, *J. Biomol. NMR*, 2003, **26**, 215–240.
18 J. Meiler and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 12105–12110.
19 K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella and M. Vendruscolo, *J. Am. Chem. Soc.*, 2009, **131**, 13894–13895.
20 Y. Shen and A. Bax, *J. Biomol. NMR*, 2010, **48**, 13–22.
21 B. Han, Y. Liu, S. W. Ginzinger and D. S. Wishart, *J. Biomol. NMR*, 2011, **50**, 43.
22 E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao and J. L. Markley, *Nucleic Acids Res.*, 2007, **36**, D402–D408.
23 H. J. Dyson and P. E. Wright, *Chem. Rev.*, 2004, **104**, 3607–3622.
24 D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges and B. D. Sykes, *J. Biomol. NMR*, 1995, **5**, 67–81.
25 C. W. Haigh and R. B. Mallion, *Prog. Nucl. Magn. Reson. Spectrosc.*, 1979, **13**, 303–344.
26 D. A. Case, *J. Biomol. NMR*, 1995, **6**, 341–346.
27 A. Buckingham, *Can. J. Chem.*, 1960, **38**, 300–307.
28 F. Zhang and R. Brüschweiler, *J. Am. Chem. Soc.*, 2002, **124**, 12654–12655.
29 D. A. Case, *Acc. Chem. Res.*, 2002, **35**, 325–331.
30 H. Zhang, S. Neal and D. S. Wishart, *J. Biomol. NMR*, 2003, **25**, 173–195.
31 Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker and A. Bax, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 4685–4690.
32 D. H. Brookes and T. Head-Gordon, *J. Am. Chem. Soc.*, 2016, **138**, 4530–4538.
33 J. Lincoff, M. Krzeminski, M. Haghighatlari, J. M. C. Teixeira, G.-N. W. Gomes, C. C. Gradinaru, J. D. Forman-Kay and T. Head-Gordon, Arxiv, 1912.12582 [physics.bio-ph], 2020.
34 R. M. Daniel, J. L. Finney, M. Stoneham and M. Nakasako, *Philos. Trans. R. Soc. London, Ser. B*, 2004, **359**, 1191–1206.
35 A. De Simone, G. G. Dodson, F. Fraternali and A. Zagari, *FEBS Lett.*, 2006, **580**, 2488–2494.
36 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
37 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
38 M. V. Berjanskii and D. S. Wishart, *J. Am. Chem. Soc.*, 2005, **127**, 14970–14971.
39 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
40 R. Dong, Z. Peng, Y. Zhang and J. Yang, *Bioinformatics*, 2017, **34**, 1719–1725.
41 S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, 1970, **48**, 443–453.

42 S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 10915–10919.
43 W. C. Wimley and S. H. White, *Nat. Struct. Biol.*, 1996, **3**, 842–848.
44 T. Hamelryck, *Proteins: Struct., Funct., Bioinf.*, 2005, **59**, 38–48.
45 G. Wagner, A. Pardi and K. Wuethrich, *J. Am. Chem. Soc.*, 1983, **105**, 5948–5949.
46 P. Geurts, D. Ernst and L. Wehenkel, *Machine Learning*, 2006, **63**, 3–42.
47 L. Breiman, *Machine Learning*, 2001, **45**, 5–32.
48 R. E. Schapire, Theoretical views of boosting, in *European conference on computational learning theory*, Springer, Berlin, Heidelberg, 1999, pp. 1–10.
49 R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender and J. H. Moore, Automating biomedical data science through tree-based pipeline optimization, in *European Conference on the Applications of Evolutionary Computation*, Springer, Cham, 2016, pp. 123–137.
50 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning* Springer, New York, 2nd edn, 2008.
51 B. Han, Y. F. Liu, S. W. Ginzinger and D. S. Wishart, *J. Biomol. NMR*, 2011, **50**, 43–57.
52 C. W. Müller, G. J. Schlauderer, J. Reinstein and G. E. Schulz, *Structure*, 1996, **4**, 147–156.
53 C. W. Müller and G. E. Schulz, *Proteins: Struct., Funct., Bioinf.*, 1993, **15**, 42–49.
54 L. Breiman, *Classification and regression trees*, Routledge, 2017.
55 A. S. Christensen, T. E. Linnet, M. Borg, W. Boomsma, K. Lindorff-Larsen, T. Hamelryck and J. H. Jensen, *PLoS One*, 2013, e84123.
56 R. Samudrala and M. Levitt, *Protein Sci.*, 2000, **9**, 1399–1401.
57 H. Deng, Y. Jia and Y. Zhang, *Bioinformatics*, 2015, **32**, 378–387.
58 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
59 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
60 J. M. Word, *et al.*, *J. Mol. Biol.*, 1999, **285**, 1735–1747.