Title

A curated rotamer library for common post-translational modifications of proteins.

Permalink

https://escholarship.org/uc/item/4fs236m4

Journal

Computer applications in the biosciences : CABIOS, 40(7)

Authors

Zhang, Oufan
Naik, Shubhankar
Liu, Zi
et al.

Publication Date

2024-07-01

DOI

10.1093/bioinformatics/btae444

Peer reviewed

Bioinformatics, 2024, **40(7)**, btae444
https://doi.org/10.1093/bioinformatics/btae444
Advance Access Publication Date: 12 July 2024
Original Paper

Structural bioinformatics

A curated rotamer library for common post-translational modifications of proteins

Oufan Zhang[1], Shubhankar A. Naik[2], Zi Hao Liu[3,4], Julie Forman-Kay[3,4], Teresa Head-Gordon[1,2,5,6,]*

[1]Kenneth S. Pitzer Center for Theoretical Chemistry, University of California, Berkeley, CA 94720, United States
[2]Department of Chemistry, University of California, Berkeley, CA 94720, United States
[3]Molecular Medicine Program, Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[4]Department of Biochemistry, University of Toronto, Toronto, ON M5S 1A8, Canada
[5]Department of Bioengineering, University of California, Berkeley, CA 94720, United States
[6]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, United States

*Corresponding author. Kenneth S. Pitzer Center for Theoretical Chemistry, University of California, Berkeley, CA 94720, United States.
E-mail: thg@berkeley.edu (T.H-G.)

Associate Editor: Arne Elofsson

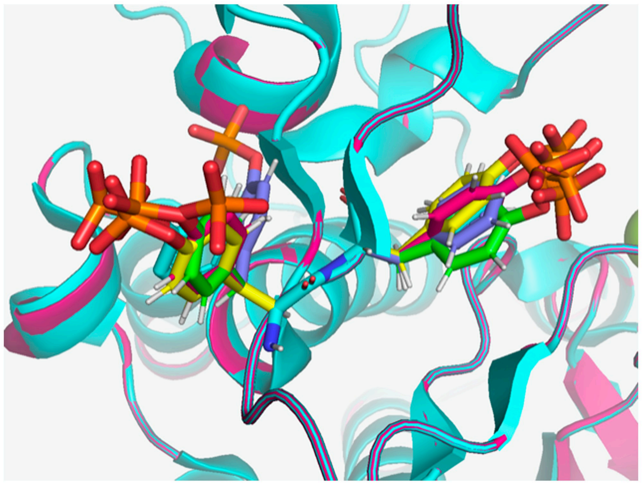
Abstract

Motivation: Sidechain rotamer libraries of the common amino acids of a protein are useful for folded protein structure determination and for generating ensembles of intrinsically disordered proteins (IDPs). However, much of protein function is modulated beyond the translated sequence through the introduction of post-translational modifications (PTMs).

Results: In this work, we have provided a curated set of side chain rotamers for the most common PTMs derived from the RCSB PDB database, including phosphorylated, methylated, and acetylated sidechains. Our rotamer libraries improve upon existing methods such as SIDEpro, Rosetta, and AlphaFold3 in predicting the experimental structures for PTMs in folded proteins. In addition, we showcase our PTM libraries in full use by generating ensembles with the Monte Carlo Side Chain Entropy (MCSCE) for folded proteins, and combining MCSCE with the Local Disordered Region Sampling algorithms within IDPConformerGenerator for proteins with intrinsically disordered regions.

Availability and implementation: The codes for dihedral angle computations and library creation are available at https://github.com/THGLab/ptm_sc.git.

Graphical Abstract



Received: 26 April 2024; **Revised:** 6 June 2024; **Editorial Decision:** 11 June 2024; **Accepted:** 11 July 2024
Published by Oxford University Press 2024.

1 Introduction

Post-translational modifications (PTMs) refer to the chemical modifications that are made to the amino acids of a protein after translation to enable the cell to regulate its function. The importance of PTMs cannot be understated given that at least 80% of mammalian proteins are modulated by PTMs, influencing essentially all biological processes, including cell signaling and metabolic pathways, transcriptional regulation, and DNA repair. In addition, dysregulation of PTMs is implicated in the development and progression of many diseases including cancer (Ramazi and Zahiri 2021) and aberrant phosphorylation of tau is implicated in Alzheimer's diseases. In spite of the fact that there are hundreds of PTMs known to occur in biology (Huang *et al.* 2019), the available experimental structural data, e.g. from X-ray crystallography, cryo-electron microscopy (Cryo-EM), or NMR measurements, remain sparse across the full space of chemical modifications compared to unmodified proteins.

In principle, computational models can fill the gap for modeling the protein structural changes introduced by PTMs, although most structure-based prediction algorithms are primarily designed for the canonical amino acids. For example, while the original AlphaFold2 (Jumper *et al.* 2021) and RosettaFold (Baek *et al.* 2021) have revolutionized protein structure prediction for unmodified sequences, they did not handle PTMs. However, these algorithms are also largely limited to prediction of single stable protein conformations, and do not provide insights into protein flexibility (Lane 2023, Wayment-Steele *et al.* 2024, Wolff *et al.* 2023), including the most flexible class with intrinsic disorder (Uversky *et al.* 2008, Wright and Dyson 2015, Bhowmick *et al.* 2016, Lazar *et al.* 2021, Ghafouri *et al.* 2024). PTMs modulate protein energy landscapes, which can lead to changes in conformations and in their dynamic interconversions. Protein flexibility also leads to the fluctuations of side chain packing arrangements that often have a direct functional role (Fraser *et al.* 2011, Moorman *et al.* 2012, Fenwick *et al.* 2014, Richard 2019, Welborn and Head-Gordon 2019). A large number of high-quality physical algorithms (Liang *et al.* 2011, Bhowmick and Head-Gordon 2015, Ollikainen *et al.* 2015, Jumper *et al.* 2018, Huang *et al.* 2020, Dicks and Wales 2022) and machine learning approaches (Nagata *et al.* 2012, Misiura *et al.* 2022, McPartlon and Xu 2023) exist to perform side chain repacking for the canonical amino acids for folded proteins, disordered proteins when they undergo binding-upon-folding, or when they form dynamical complexes.

However, the ability to model side chain ensembles with PTMs are currently quite limited. There have been a small number of studies for computational modeling of PTMs (Petrovskiy *et al.* 2023); software packages such as Rosetta (Leaver-Fay *et al.* 2011, Alford *et al.* 2017), FoldX (Schymkowitz *et al.* 2005), and SIDEpro (Nagata *et al.* 2012) have the ability to model PTMs using rotamer libraries (Renfrew *et al.* 2014). A recent method named GlycoSHIELD exclusively focuses on the modeling of glycosylation by grafting glycan conformer candidates sampled from molecular dynamics trajectories (Tsai *et al.* 2024). Given that experimental PTM structural data have continued to accumulate since many of these methods have been developed, it is worth creating new side chain rotamer libraries containing PTMs. Furthermore, since some rotamer libraries such as SIDEpro only characterized the effect of the PTM modifications on side chain torsion angles, it is also valuable to account for the rotamer distribution shifts of the backbone dihedral angles.

In this study, we have undertaken the creation of both backbone-dependent (BD) and backbone-independent (BI) rotamer libraries to comprehensively investigate the influence of PTMs on sidechain conformational ensembles. The decision to generate both types of libraries arises from the goal of capturing nuanced details of sidechain variability within the local structural context provided by the protein's backbone, as well as to understand more general patterns of sidechain conformations across diverse protein structures where data sparsity is less of a concern. We compare our rotamer libraries against SIDEpro (Nagata *et al.* 2012) and Rosetta (Leaver-Fay *et al.* 2011, Alford *et al.* 2017), and show that the overall trend across various metrics show systematic improvements against folded proteins and intrinsically disordered protein (IDP) compared to these standard methods.

2 Materials and methods

2.1 Structural data for PTMs of amino acids

To generate structured data for PTM-modified amino acids to construct our datasets, we first accessed the PTM Structural Database (PTM-SD) (Craveur *et al.* 2014). Each data entry in the PTM-SD corresponds to a distinct PTM on a specific amino acid residue. Within the PTM-SD, an inquiry targeting these amino acids and their associated modifications facilitated the retrieval of RCSB Protein Data Bank (PDB) identification codes for each modified residue.

Utilizing the obtained PDB identification codes, we conducted searches within the PDB database (Berman *et al.* 2000) to acquire structural data for the corresponding modified amino acids. A notable observation was that a subset of proteins frequently exhibited high sequence similarity and identity. Such a high degree of redundancy would introduce biases that could impact the interpretation of patterns and relationships in the rotamer data. To address this, the independent NCBI BLAST tool (Altschul *et al.* 1997) was employed to detect identical sequences. Through the alignment of FASTA sequences from all proteins within the dataset, a threshold for sequence similarity of 90% or above was established, resulting in the clustering of protein structures. The 90% cutoff, while only qualitative, is based on the idea that 90% sequence similarity would correspond to ∼1 Å RMSD in the backbone of a generated structure from the similar template, which is enough backbone variation to give new side chain packings, i.e. a notable difference in rotamer states. Subsequently, within each group, the structure having the highest resolution was selected for further analysis. Furthermore, PDB files with incomplete structural data were excluded.

Supplementary Tables S1 and S2 provide the curated PDB dataset for PTM-containing amino acids, further delineated by the refined resolution ranges for structures including each modified residue type, and the number of PDB files available in each resolution category. We found PDB files for phosphorylated serine (SEP), phosphorylated threonine (TPO), phosphorylated tyrosine (PTR), methylated arginine (AGM), mono-methylated lysine (MLZ), di-methylated lysine (MLY), tri-methylated lysine (M3L), oxidized methionine (OMT), and acetylated lysine (ALY) (see Supplementary Table S3). However, we chose to perform our analysis and rotamer generation to create PTM rotamer libraries only on modified amino acids having sufficient data by defining a resolution cutoff of 3.5 Å and lower, and requiring a minimum of 40 total PDB files. Based on these criteria, we only developed rotamer libraries for SEP, TPO, PTR, M3L, and ALY.

2.2 Statistical analysis for PTM rotamer libraries

Given that the modified amino acids exhibit a non-uniform distribution of backbone phi (ϕ) and psi (ψ), and the relative sparsity of the observations of side chain rotamer chi (χ_i) for specific backbone dihedral angles, we use the adaptive kernel density estimation of Shapovalov and Dunbrack to estimate the rotamer probability density functions (PDFs) (Shapovalov and Dunbrack 2011). This method determines the width of the kernel based on the local density of data points, such that in denser regions narrower kernels are applied to capture more local variations while broader kernels create smoother distributions in regions where rotamer occupancy is sparse.

We employ the von Mises kernel (Mardia and Zemroch 1975) which is well-suited to periodic data such as dihedral angles. The Nadaraya–Watson kernel regression model (Nadaraya 1964, Watson 1964) considers the influence of neighboring data points in a weighted manner, producing smoothed estimates of the mean and standard deviation values for χ angles within each ϕ/ψ bin. This smoothing helps mitigate the impact of noise and fluctuations in the data, providing a more robust and reliable characterization of the relationships between the backbone ϕ/ψ and sidechain χ dihedral angles. Bayes' rule is applied to these rotamer PDFs to acquire the rotamer probabilities.

2.3 BD PTM rotamer libraries

In the construction of the BD-rotamer library, we discretized the backbone space into bins, each spanning a 30° interval. This level of granularity allows for a detailed representation of the local backbone geometry and facilitates the accurate prediction of sidechain conformations. The 30° bin size was chosen to balance computational efficiency with the need to capture subtle variations in sidechain orientations influenced by the local backbone structure. Regarding discretized bins which lack occupancy in the rotamer library, the probability and χ angle means and standard deviation are estimated using the BI probability distributions.

In the process of estimating rotamer means and standard deviations and discretizing angle bins for a rotamer sidechain library, the specified angle ranges for trans [T, (120°, 180°) or (−180°, −120°)], gauche [G+, (0°, 120°)], and gauche− [G−, (−120°, 0°)] conformers are defined for sp3 carbons (Supplementary Fig. S1). Given a sp3–sp3 hybridized bond, the degrees of freedom of the dihedral angles have probability density distributions that contain three distinct and symmetric peaks that occur at 60°, −60°, and −180/180° and that align with the conformers mentioned above. For ALY χ_5, the rotamer bins are instead defined for gauche [G+, (30°, 90°) or (−150°, −90°)], gauche− [G−, (90°, 150°) or (−90°, −30°)], and trans [T, (−30°, 30°) or (−180°, −135°) or (135°, 180°)] given the sp3–sp2 hybridized bond. For PTR χ_2, the rotamer bins are defined for gauche [G, (45°, 135°) or (−135°, −45°)] and trans [T, (−45°, 45°) or (−180°, −135°) or (135°, 180°)] to account for the symmetry of a benzene ring. When discretizing angle bins, they are aligned with these conformer-specific ranges, ensuring that each bin corresponds to the appropriate conformer type. The estimation of rotamer means and standard deviations is then performed within these conformer-specific ranges.

While canonical amino acids fit well into this discretization strategy, we find that PTMs sometimes do not fit into standard categories. To accommodate non-standard rotamer angle distributions observed for the χ_2 of SEP and TPO, and χ_3 of PTR, we split the probability distributions by their population density with clusters defined using DBSCAN (Ester *et al.* 1996). Furthermore, as the terminal χ angles of phosphorylated SEP, TPO, PTR, and the methylated M3L have a 3-fold symmetry around the torus axis and show broad distributions regardless of the other sidechain angles, we estimated their means and standard deviations by only aligning with the conformer ranges of the angles themselves (Supplementary Fig. S2).

2.4 Creating protein structures using PTM rotamer libraries for folded proteins and IDPs

We use our newly generated PTM libraries by generating side chain ensembles for folded proteins with the Monte-Carlo Side Chain Entropy (MCSCE) (Bhowmick and Head-Gordon 2015). The MCSCE algorithm is also embedded within the Local Disordered Region Sampling (LDRS) (Liu *et al.* 2023) algorithms within IDPConformerGenerator (Teixeira *et al.* 2022) for proteins with intrinsically disordered regions (IDRs). The backbone conformers of disordered regions were aligned and attached to folded domains using the LDRS module in IDPConformerGenerator, discarding conformations with steric clashes. For each disordered region case, we sampled 20 000 backbone conformations from a combination of loop, helices, and β-strands regions; successful and complete sidechain packing statistics are given in Supplementary Table S4.

Candidate sidechain packings were generated with MCSCE (Bhowmick and Head-Gordon 2015) program for the disordered regions of the unmodified sequence using the Dunbrack library (Shapovalov and Dunbrack 2011), and with PTMs using the BD-rotamer and BI-rotamer libraries we developed in this work. We also used MCSCE to generate sidechain packings using the SIDEpro (Nagata *et al.* 2012) PTM rotamer libraries. As the SIDEpro rotamer library only defines the additional sidechain angles for PTMs conditioned on the rotamer ranges of the existing angles in the unmodified amino acid residues, these sidechain angles were sampled according to the probability distribution of the corresponding canonical amino acids. In both cases, the rotamers of the folded domains are unchanged. The Rosetta (Leaver-Fay *et al.* 2011, Alford *et al.* 2017) packed conformers were generated by invoking the fixbb application starting from the same backbone conformers attached to the folded domains, similarly with the folded domains held fixed. As Rosetta's scoring function during conformer packing does not define an explicit clash cutoff as in MCSCE, the Rosetta generated conformers were also filtered with the same clash criteria in MCSCE for comparison.

3 Results

3.1 BD and BI PTM libraries

We first consider a BD-rotamer library for SEP, TPO, PTR, M3L, and ALY, that categorize sidechain conformations based on specific backbone ϕ and ψ dihedral angles. As shown by Dunbrack and Cohen (1997), a BD library for

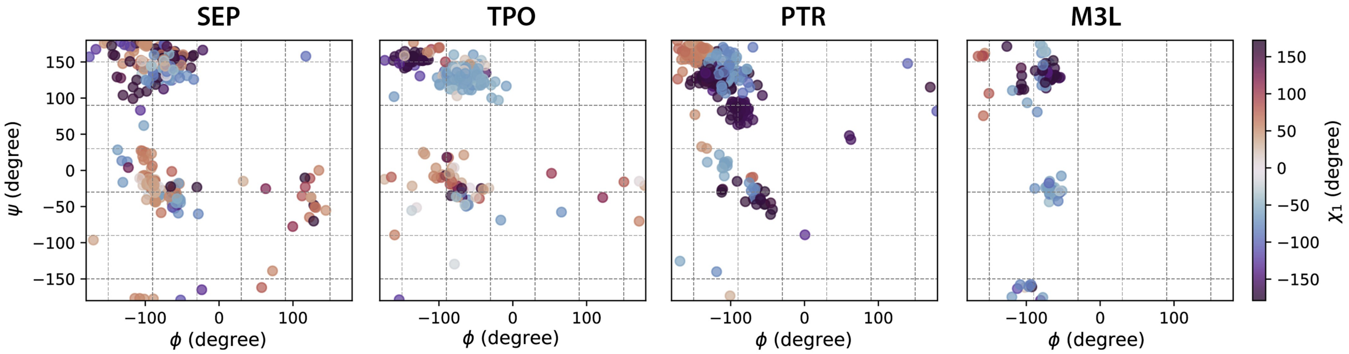


Figure 1. Ramachandran plots color coded by χ_1 angle ranges for PTM-modified amino acids using the backbone-dependent library. Backbone bins in 60° separation are shown in gray dash lines. We consider phosphorylated serine (SEP), phosphorylated threonine (TPO), phosphorylated tyrosine (PTR), and tri-methylated lysine (M3L).

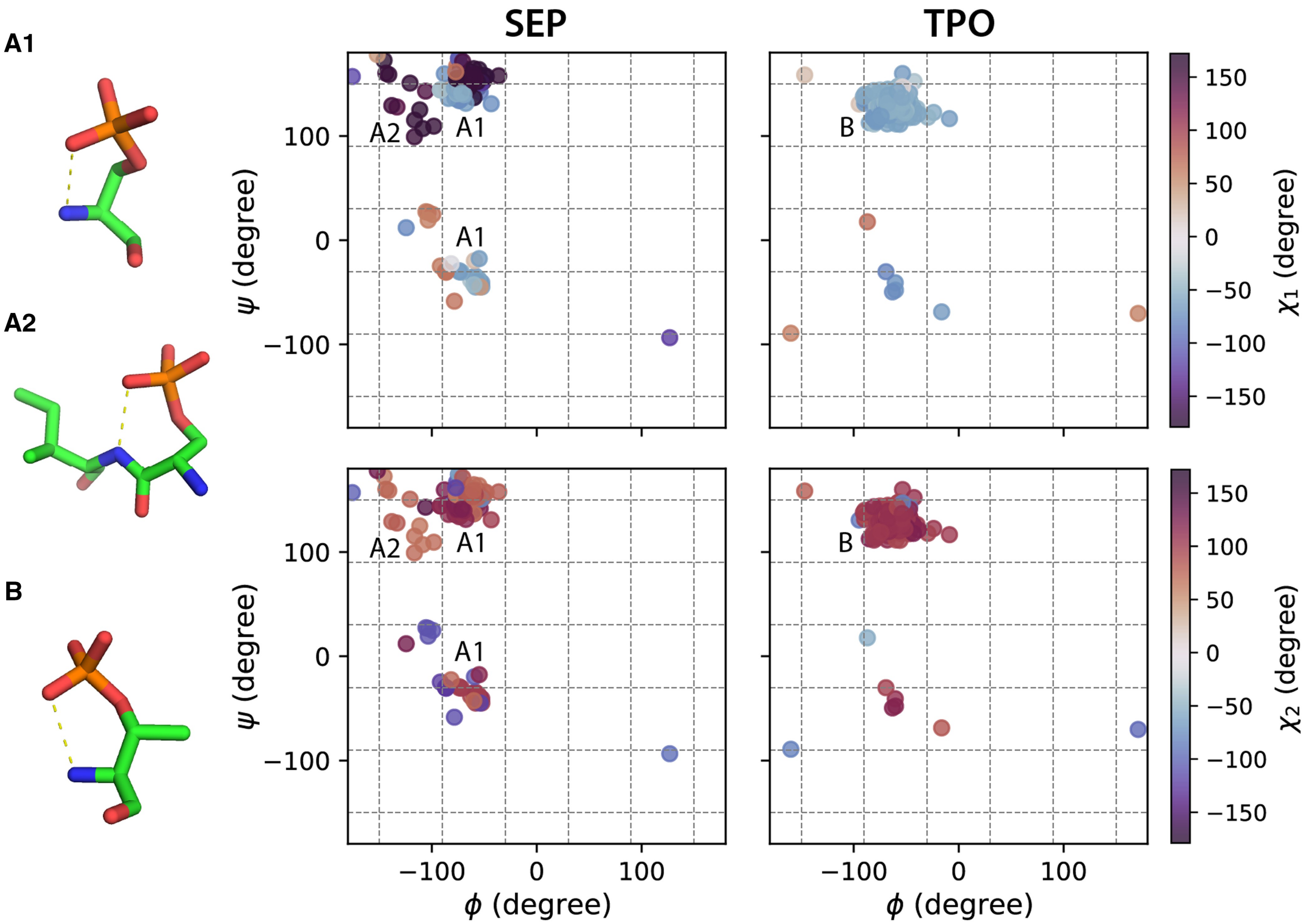


Figure 2. Ramachandran plots color coded by χ_1 and χ_2 angle ranges for SEP and TPO with hydrogen bond formation using the backbone-dependent library. Backbone bins in 60° separation are shown in gray dash lines. Hydrogen bonds are defined by within a donor–acceptor distance cutoff of 3.5 Å. (A1, A2, and B) Representative configurations for SEP (A1, A2) and TPO (B) associated with the annotated backbone regions.

predicting side chain conformations produces much better results for protein structure refinement using NMR and X-ray data as opposed to a BI-rotamer library. The BI-rotamer library focuses solely on the distribution of sidechain dihedral angles, and for PTMs may be a necessity if the data is too sparse to differentiate it from the BD-rotamer case.

The influence of backbone conformations on PTM side-chain rotamer states is illustrated in Figs. 1 and 2 (and Supplementary Figs. S1–S3). In Fig. 1, the χ_1 of PTM-modified amino acids show distinct rotamer populations in different ϕ/ψ regions (and in relation to secondary structure), and more importantly the χ_1 distributions for the PTM-modified amino acids also shift considerably from the unmodified canonical residues in these regions. For example, while the χ_1 of SER, THR and LYS show similar populations regardless of ϕ/ψ ranges, the χ_1 of SEP, TPO, and M3L have visibly different preferences for certain backbone regions (Fig. 1). This illustrates that the SIDEPro rotamer library that utilizes an unmodified χ_1 will be unable to fully capture these structural changes exhibited by the PTMs.

The χ_2 of the phosphorylated amino acids SEP and TPO adopt rotamers that cannot be easily explained using bond hybridization models, but instead arise from favorable hydrogen bonding interactions (Wong *et al.* 2005) (Fig. 2). In particular, TPO has a dominating χ_1/χ_2 population centered at $-60°/120°$, a configuration that encourages formation of an internal P–O/N–H hydrogen bond (Fig. 2B). χ_2 of SEP exhibits a mixed population that spans the angular range from 60° to −120°. In addition to the χ_1/χ_2 $-60°/\pm120°$ configuration associated with an internal hydrogen bond (Fig. 2A1), we observe the formation of a hydrogen bond between P–O of the SEP residue i and N–H of its adjacent residue $i+1$ with a χ_1/χ_2 configuration around $-180°/60°$ (Fig. 2-A2). These varieties of sidechain rotamer states are consistent with the cooperative transition between a state in which the phosphate group is well-solvated and a state that forms intra- and inter-residue hydrogen bonds, as noted in reference (Wong *et al.* 2005). As shown in Fig. 2 (and Supplementary Figs. S1–S3), the observed higher percentage of hydrogen bond formations in the polyproline helical backbone region, and the resulting selectivity in the χ_1/χ_2 configurations, help rationalize the distinct rotamer populations for phosphorylated amino acids such as SEP and TPO that are dependent on backbone configurations. All these differences highlight the need to better characterize the rotamer states of PTM-modified amino acids apart from the canonical amino acids.

3.2 Side chain ensembles with PTMs for folded proteins

We verify the fidelity of our constructed BD-rotamer and BI-rotamer libraries for PTMs by sampling their sidechain torsion distributions to generate new side chain packings and compare the resulting accuracies of the repacked structures to the experimental PDB data. In Supplementary Fig. S4, we show that the sidechain rotamer distribution of the constructed libraries is in good agreement with the PDB data when sampled based on the same backbone conformation. When repacking PTMs sidechains, we removed the original PTM sidechain structures and then resampled the rotamers for PTMs with all other residues intact. We compare the RMSD values of the full protein structures with PTMs, regenerated with the BD-rotamer and BI-rotamer libraries from this work against the SIDEpro and Rosetta libraries, and compare the distributions as boxplots in Fig. 3A.

It is evident that the BD-rotamer library has a smaller interquartile range of somewhere between 0.25 and 1.0 Å and is skewed toward lower RMSD values across all of the PTM types compared to the other rotamer libraries (Fig. 3A and

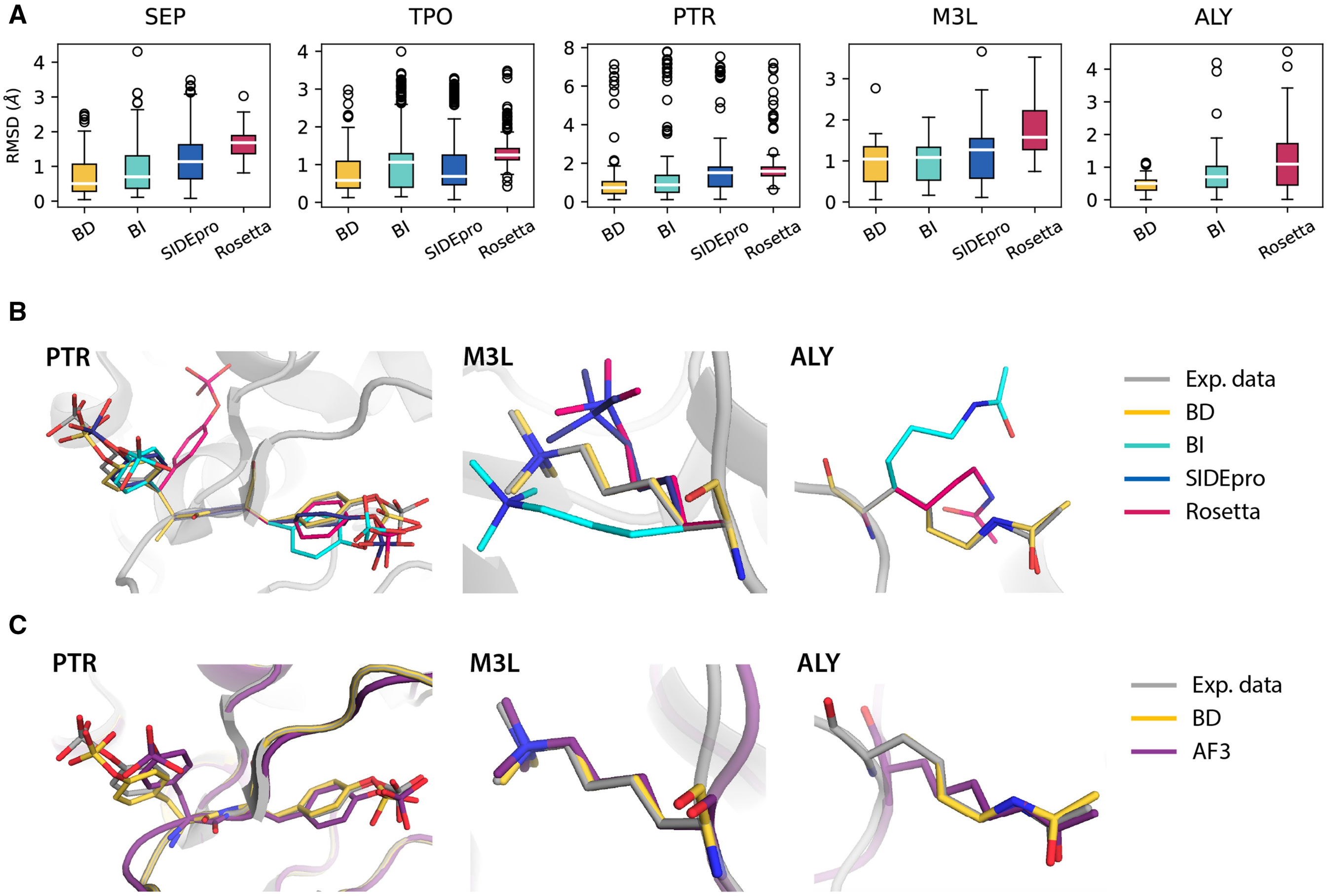


Figure 3. RMSD distributions of repacked PTM-modified residues using different rotamer libraries and compared to experimental structures and AlphaFold3. SIDEpro does not support ALY packing. (A) Boxplots for the RMSD distributions. Medians are highlighted in white and each box extends from the first quartile (*Q*1) to the third quartile (*Q*3). Outliers in circle are defined as points outside of 1.5 times the interquartile range below *Q*1 or above *Q*3. (B) Repacked PTM-containing structures using different rotamer libraries compared to the experimental PDB structures. (C) Repacked PTM-containing structures using BD-rotamer library compared to AlphaFold3 and the experimental PDB structures. Examples are taken from PDB ID 3CLY (PTR), 4EZH (M3L), and 4QUT (ALY).

Supplementary Table S5). The BI-rotamer library for PTMs also demonstrates trends in improvement to the other standard methods, with the exception of TPO, and the statistical significance of improvement is not as strong compared to the BD-rotamer library. We also performed a Wilcoxon Signed Rank test to evaluate the RMSD differences between each distribution pair with a 95% confidence level, in which the *P* values from this test shown in Supplementary Fig. S5 show that the BD-rotamer library results in a statistically meaningful decrease in RMSDs compared to existing methods such as SIDEpro and Rosetta. For all PTM types considered in this work, the PTM BD-rotamer provides excellent performance in recovering the experimental structures; Fig. 3B provides examples of repacked PTM-modified structures using all four methods and compared to experimental PDB structures. Figure 3C also compares the BD-rotamer library and AlphaFold3 (Abramson *et al.* 2024) against experiment. Overall, the BD-rotamer library is in better agreement with experiment for both the PTR and ALY, but due to the limitations of AlphaFold3 software at present, we cannot exhaustively test this assertion across the whole data set that we did for SIDEPro and Rosetta.

3.3 Side chain ensembles with PTMs for intrinsically disordered proteins

To demonstrate how our PTM libraries can support ensemble generation for disordered proteins, we considered two cases for which the proteins contain IDRs. The Histone H3 N-terminal IDR within the nucleosome structure (chains C and G, PDB ID 8SIY) have four of the five types of PTMs investigated here, with methylation at K9, phosphorylation at S10 and T11, and acetylation at K14 (Fig. 4A). The ubiquitin

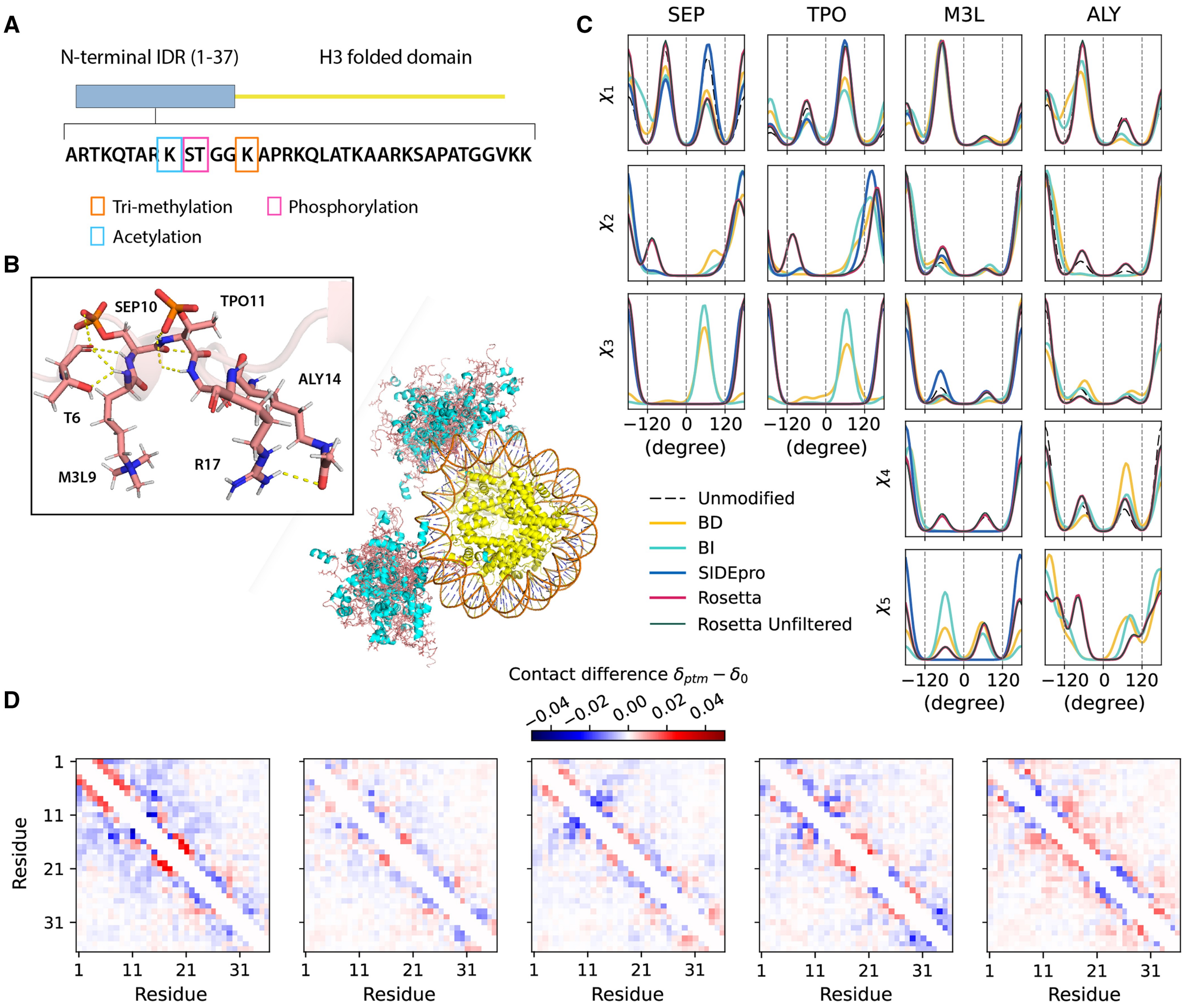


Figure 4. Histone H3 conformers generated with different PTMs libraries. (A) Modifications on the histone H3 N-terminus on chains C/G of the nucleosome. (B) Ensembles of 30 all-atom H3-modified nucleosome conformers (folded domains and DNA are taken from PDB ID 8SIY). PTM-modified residues are highlighted with stick representations. (C) Comparison of torsion angle probability distributions for PTM-modified residues in the H3 conformers with different libraries. (D) Fractional inter-residue contacts (C_α-C_α distances within 8 Å) of PTMs containing ensembles (δ_{ptm}) subtracted by the ensemble without modifications (δ_0) for the IDR regions. The maps were calculated with 500 randomly sampled conformers [based on convergence of Rgand averaged over 10 trials. From left to right: BD-rotamer, BI-rotamer, SIDEpro, Rosetta, and Rosetta without clash filtering, with contacts averaged from chain C and G.

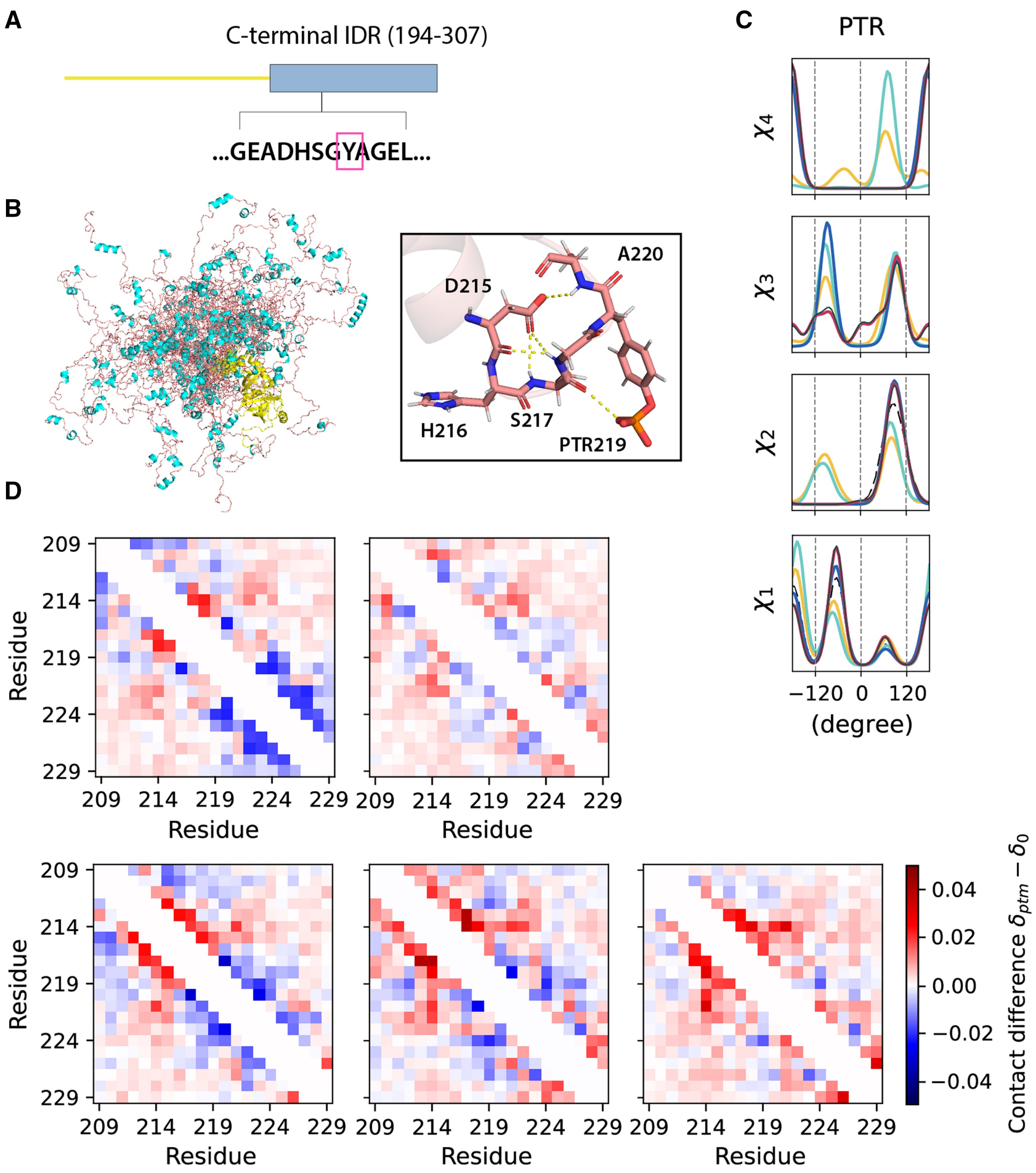


Figure 5. UDF1 conformers generated with different PTM libraries. (A) Modifications on the C-terminal IDR of UDF1 at Y219. (B) Cartoon representations of 30 all-atom UDF1 conformers (structure of the folded domain taken from PDB ID 2YUJ model). (C) Comparison of torsion angle probability distributions for PTM in the UDF1 conformers with different libraries. (D) Fractional inter-residue contacts (C_α–C_α distances within 8 Å) of PTMs containing ensembles (δ_{ptm}) subtracted by the ensemble without modifications (δ_0) for the IDR region around the PTM site. The maps were calculated with ensembles of 500 randomly sampled conformers and averaged over 10 trials. Top: BD-rotamer, BI-rotamer; bottom (L to R): SIDEpro, Rosetta, and Rosetta without clash filtering.

recognition factor in ER-associated degradation protein 1 (UFD1) with phosphorylation at Y219 (PDB ID 2YUJ) is shown in Fig. 5A. For both cases, we compare ensembles generated with and without PTMs using our BD-rotamer and BI-rotamer libraries as well as libraries from SIDEpro and Rosetta, although SIDEpro only contains M3L, SEP, and TPO since ALY is not supported by that method (Nagata *et al.* 2012). We also compared against results using the original Rosetta scoring function that ignores steric clashes for the Histone H3 nucleosome structure. Supplementary Figure S6 shows that the Ramachandran plots of the sidechain rotamer states with PTMs do not change for IDRs, nor among libraries, and only small changes are observed in secondary structure between rotamer libraries as seen in Supplementary Fig. S7. This indicates that the structural changes are concentrated in any differences in side chain packing.

Figures 4C and 5C show that the torsional properties of the PTM IDR ensembles are different to IDR ensembles without sidechain modifications at the modification sites. Furthermore, the sidechain rotamer state changes are reflected differently for the BD-, BI-, SIDEpro, and two Rosetta-rotamer ensembles. To better analyse what are the structural consequences of the IDR ensembles generated with the different PTM rotamer libraries, we constructed 2D contact maps subtracting the values from the ensemble without PTMs (Figs. 4D and 5D) to look for increases (red) or decreases (blue) of residue–residue contacts being made. For Histone H3, residues 5–14 show a higher population of contacts with the BD-rotamer library, which we see corresponds to a much denser hydrogen-bonded network on average in this area (Fig. 4B). A similar observation is found for UDF1 using the BD-rotamer library, and although it is only a single modification spot at residue 219, the PTM introduces a network of hydrogen bonds (example shown in Figure 5B). Especially for Histone H3, the other rotamer libraries show a smaller set of contacts or even net loss of contacts in this same region. In turn the BD-rotamer generated structures show a diminishment of contacts made by these same PTM residues with other regions of the protein, an effect which is muted in the other libraries. For the BI-rotamer library this is due to over-averaging, whereas for SIDEPro the differences with the unmodified ensembles is because it uses the same χ_1 as the unmodified residues. It is interesting that the Rosetta-rotamer library combined with no clash criteria gives a nearly opposite trend in sidechain packing for residues with PTMs. Given that repacking structures on the folded protein backbone showed greater reliability with the BD-rotamer set, we extrapolate that there is better justification for supporting the structural consequences observed for the Histone H3 and UDF1 IDRs.

4 Conclusions

Modeling sidechain conformations for PTMs have important applications in understanding protein conformational switches, including changes in dynamics and disordered protein interactions, leading to functional consequences modulated by PTMs. Although publicly available high-resolution structures including residues modified by many of the known PTMs are still limited, there is enough structural data for residues modified by phosphorylation, methylation, and acetylation. Thus, we have developed BD-rotamer and BI-rotamer libraries for PTM-modified residues, using structural data curated and cleaned from the recent PTM-SD update. While all PTMs are sourced from the current PDB, our protocol and available software allow for future updates of the BB- and BI-rotamer libraries when more structural data becomes available.

We evaluate the constructed libraries in comparison to SIDEpro and Rosetta in the context of generating conformers for folded domains and for proteins with disordered regions. The BD-rotamer libraries outperform the BI-rotamer libraries as well as SIDEpro and Rosetta in retrieving the experimental structures for PTMs on the folded proteins, as evidence of its ability to capture the correlation more accurately between backbone and sidechain, and within sidechain dihedral conformations. For phosphorylated residues specifically, the ability to predict the sidechain rotamer cooperatively is crucial to modeling local hydrogen bonding interactions, thus allowing one to better delineate the structural features of single conformer and ensembles upon chemical modifications. We also show that the constructed PTM-modified residue rotamer libraries can be used along with MCSCE and IDPConformerGenerator to produce all-atom conformers for disordered proteins with PTMs. If experimental data such as nuclear magnetic resonance, small-angle X-ray scattering, and fluorescence resonance energy transfer are available, a reweighting protocol using X-EISD46 or evolving underlying ensembles to agree with experimental data using DynamICE (Zhang *et al.* 2023) can be applied to these all-atom sidechain-modified conformers to generate more realistic ensemble representations to investigate PTM-regulated activities and interactions.

Acknowledgements

J.D.F.-K. and T.H.-G. acknowledge funding from the National Institutes of Health (2R01GM127627-05). J.D.F.-K. also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC, 2016–06718) and from the Canada Research Chairs Program.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

J.D.F.-K. and T.H.-G. acknowledge funding from the National Institutes of Health (2R01GM127627-05). J.D.F.-K. also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC, 2016-06718) and from the Canada Research Chairs Program.

Data availability

The data and code for this work are available at https://github.com/THGLab/ptm_sc.git. We also note that the PTM library is part of the MCSCE program seamlessly integrated into the IDPConformerGenerator platform.

References

Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* 2024; **630**:493–500.

Alford RF, Leaver-Fay A, Jeliazkov JR *et al.* The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* 2017;**13**:3031–48.

Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Bhowmick A, Brookes DH, Yost SR *et al.* Finding our way in the dark proteome. *J Am Chem Soc* 2016;**138**:9730–42.

Bhowmick A, Head-Gordon T. A Monte Carlo method for generating side chain structural ensembles. *Structure* 2015;**23**:44–55.

Craveur P, Rebehmed J, de Brevern AG. PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database* 2014;**2014**:bau041.

Dicks L, Wales DJ. Exploiting sequence-dependent rotamer information in global optimization of proteins. *J Phys Chem B* 2022; **126**:8381–90.

Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Sci* 1997;**6**:1661–81.

Ester M, Krieger H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, 1996, 26–231.

Fenwick RB, van den Bedem H, Fraser JS *et al.* Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc Natl Acad Sci U S A* 2014;**111**:E445–54.

Fraser JS, van den Bedem H, Samelson AJ *et al.* Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc Natl Acad Sci U S A* 2011;**108**:16247–52.

Ghafouri H, Lazar T, Del Conte A *et al.* PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res* 2024;**52**:D536–44.

Huang K-Y, Lee T-Y, Kao H-J *et al.* dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res* 2019;**47**:D298–308.

Huang X, Pearce R, Zhang Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* 2020;**36**:3758–65.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Jumper JM, Faruk NF, Freed KF *et al.* Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS Comput Biol* 2018;**14**:e1006342.

Lane TJ. Protein structure prediction has reached the single-structure frontier. *Nat Methods* 2023;**20**:170–3.

Lazar T, Martínez-Pérez E, Quaglia F *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res* 2021;**49**:D404–11.

Leaver-Fay A, Tyka M, Lewis SM *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 2011;**487**:545–74.

Liang S, Zheng D, Zhang C *et al.* Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* 2011;**27**:2913–4.

Liu ZH, Teixeira JMC, Zhang O *et al.* Local disordered region sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments. *Bioinformatics* 2023;**39**:btad739.

Mardia KV, Zemroch PJ. The Von Mises distribution function. *J Roy Stat Soc Ser C: Appl Stat* 1975;**24**:268–72.

McPartlon M, Xu J. An end-to-end deep learning method for protein side-chain packing and inverse folding. *Proc Natl Acad Sci U S A* 2023;**120**:e2216438120.

Misiura M, Shroff R, Thyer R *et al.* DLPacker: deep learning for prediction of amino acid side chain conformations in proteins. *Proteins: Struct Funct Bioinform* 2022;**90**:1278–90.

Moorman VR, Valentine KG, Wand A The dynamical response of hen egg white lysozyme to the binding of a carbohydrate ligand. *Protein Sci* 2012;**21**:1066–73.

Nadaraya EA. On estimating regression. *Theory Probab Appl* 1964; **9**:141–2.

Nagata K, Randall A, Baldi P. SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins* 2012;**80**:142–53.

Ollikainen N, de Jong RM, Kortemme T. Coupling protein side-chain and backbone flexibility improves the re-design of protein–ligand specificity. *PLoS Comput Biol* 2015;**11**:e1004335.

Petrovskiy DV, Nikolsky KS, Rudnev VR *et al.* Modeling side chains in the three-dimensional structure of proteins for post-translational modifications. *Int J Mol Sci* 2023;**24**:13431.

Ramazi S, Zahiri J. Post-translational modifications in proteins: resources, tools and prediction methods. *Database* 2021; **2021**:baab012.

Renfrew PD, Craven TW, Butterfoss GL *et al.* A rotamer library to enable modeling and design of peptoid foldamers. *J Am Chem Soc* 2014;**136**:8772–82.

Richard JP. Protein flexibility and stiffness enable efficient enzymatic catalysis. *J Am Chem Soc* 2019;**141**:3320–31.

Schymkowitz J, Borg J, Stricher F *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**:W382–8.

Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 2011;**19**:844–58.

Teixeira JMC, Liu ZH, Namini A *et al.* IDPConformerGenerator: a flexible software suite for sampling the conformational space of disordered protein states. *J Phys Chem A* 2022;**126**:5985–6003.

Tsai Y-X, Chang N-E, Reuter K *et al.* Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. *Cell* 2024;**187**:1296–311.e26.

Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;**37**:215–46.

Watson GS. Smooth regression analysis. *Sankhyā: Indian J Stat, Ser A (1961–2002)* 1964;**26**:359–72.

Wayment-Steele HK, Ojoawo A, Otten R *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 2024; **625**:832–9.

Welborn VV, Head-Gordon T. Fluctuations of electric fields in the active site of the enzyme ketosteroid isomerase. *J Am Chem Soc* 2019; **141**:12487–92.

Wolff AM, Nango E, Young ID *et al.* Mapping protein dynamics at high spatial resolution with temperature-jump X-ray crystallography. *Nat Chem* 2023;**15**:1549–58.

Wong SE, Bernacki K, Jacobson M. Competition between intramolecular hydrogen bonds and solvation in phosphorylated peptides: simulations with explicit and implicit solvent. *J Phys Chem B* 2005; **109**:5249–58.

Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 2015;**16**:18–29.

Zhang O, Haghighatlari M, Li J *et al.* Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *J Chem Phys* 2023;**158**:174113.

Published by Oxford University Press 2024.

Bioinformatics, 2024, 40, 1–9
https://doi.org/10.1093/bioinformatics/btae444
Original Paper