

RESEARCH ARTICLE

SPECS: Integration of side-chain orientation and global distance-based measures for improved evaluation of protein structural models

Rahul Alapati¹, Md. Hossain Shuvo¹, Debswapna Bhattacharya^{1,2*}

1 Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama, United States of America, **2** Department of Biological Sciences, Auburn University, Auburn, Alabama, United States of America

* bhattacharyad@auburn.edu**OPEN ACCESS**

Citation: Alapati R, Shuvo M.H, Bhattacharya D (2020) SPECS: Integration of side-chain orientation and global distance-based measures for improved evaluation of protein structural models. PLoS ONE 15(2): e0228245. <https://doi.org/10.1371/journal.pone.0228245>

Editor: Claudio M. Soares, Universidade Nova de Lisboa Instituto de Tecnologia Quimica e Biologica, PORTUGAL

Received: July 22, 2019

Accepted: January 11, 2020

Published: February 13, 2020

Copyright: © 2020 Alapati et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All source code and underlying data for this study are freely available at <https://github.com/Bhattacharya-Lab/SPECS> and <http://watson.cse.eng.auburn.edu/SPECS/>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Significant advancements in the field of protein structure prediction have necessitated the need for objective and robust evaluation of protein structural models by comparing predicted models against the experimentally determined native structures to quantitate their structural similarities. Existing protein model versus native similarity metrics either consider the distances between alpha carbon (C α) or side-chain atoms for computing the similarity. However, side-chain orientation of a protein plays a critical role in defining its conformation at the atomic-level. Despite its importance, inclusion of side-chain orientation in structural similarity evaluation has not yet been addressed. Here, we present SPECS, a side-chain-orientation-included protein model-native similarity metric for improved evaluation of protein structural models. SPECS combines side-chain orientation and global distance based measures in an integrated framework using the united-residue model of polypeptide conformation for computing model-native similarity. Experimental results demonstrate that SPECS is a reliable measure for evaluating structural similarity at the global level including and beyond the accuracy of C α positioning. Moreover, SPECS delivers superior performance in capturing local quality aspect compared to popular global C α positioning-based metrics ranging from models at near-experimental accuracies to models with correct overall folds—making it a robust measure suitable for both high- and moderate-resolution models. Finally, SPECS is sensitive to minute variations in side-chain χ angles even for models with perfect C α trace, revealing the power of including side-chain orientation. Collectively, SPECS is a versatile evaluation metric covering a wide spectrum of protein modeling scenarios and simultaneously captures complementary aspects of structural similarities at multiple levels of granularities. SPECS is freely available at <http://watson.cse.eng.auburn.edu/SPECS/>.

Introduction

The biological function of a protein molecule is intimately linked to its three dimensional (3D) structure. The knowledge of the 3D structure of a protein, therefore, helps us in understanding its function [1] and enables improved drug design [2]. However, experimental determination of the 3D structure of a protein is expensive and time consuming. Furthermore, the rapid accumulation of protein sequence data without available structures make it practically impossible to solve the structures of all the proteins experimentally [3]. Protein structure prediction methods aim to address these challenges by computationally predicting the 3D structure of proteins in a time-efficient manner. Computational protein 3D structure prediction, therefore, has become an integral part of structural bioinformatics [4]. Contemporary protein structure prediction methods [5–10] typically generate a large number of protein models for a given target protein and select a finite subset (typically 5 to 10) of chosen models as candidates for the final prediction. The evaluation of the accuracy of these candidate predicted models via 3D structure comparison approaches, in which the predicted models are compared against the experimentally solved native conformation of the protein in order to quantitate their similarities or differences, is critically important [11] for assessing the success of the structure prediction pipelines.

A number of model vs. native comparison-based accuracy evaluation measures have been developed over the last decade [12]. Majority of the existing model-native evaluation measures rely on superposition-based or superposition free distance-based measures [13–17], in which degrees of similarities or differences are determined based on the corresponding distances between either the main chain atoms or the side-chain atoms of the model and native. $C\alpha$ Root Mean Square Deviation ($C\alpha$ RMSD) [18] is one of the most commonly used main chain superposition-based model-native dissimilarity scores. It is the measure of the overall disagreement between the $C\alpha$ atoms of the corresponding residues after optimal structural superposition. The lower the $C\alpha$ RMSD, the better the model is in agreement with respect to the native. RMSD can be extended to include all the backbone atoms or even all atoms. However, one major limitation of RMSD is its dependence on the length of the target protein in that it is easier to obtain lower RMSD values for smaller proteins compared to larger proteins. Furthermore, RMSD is overly sensitive to minute modeling errors such as in the flexible loop regions of the structure [12].

LG-score [19] is a popular superposition-based model-native similarity metric proposed by Levitt and Gerstein. It is measured as the sum of the reciprocated distances between the aligned $C\alpha$ atoms minus gap penalties. Siew, Elofsson, Rychlewski and Fischer proposed MaxSub score [20] by identifying the maximum substructure in which the distances between equivalent residues of two structures after superposition are below some threshold value, such as 3.5Å. Max-Sub score lies between 0 and 1 with higher scores indicating better agreement between the model and the native. Zhang and Skolnick developed TM-score [14] by exploiting a length-dependent normalizing distance scale to eliminate the inherent protein size dependence. TM-score lies between 0 and 1, with higher scores indicating better model-native similarity.

Global Distance Test (GDT) [13], a popular structural superposition-based global model-native similarity metric, on the other hand, uses a distance threshold based approach. It is defined by the average proportion of model residues having their $C\alpha$ atom distances from the corresponding residues in the native structure below a few predefined distance thresholds. Multiple superpositions of the pair of structures, each including the largest set of superimposable atoms are considered and the maximal residue set for each cutoff is selected, followed by averaging over several predetermined thresholds. For GDT-TS [13], predetermined thresholds of 1, 2, 4 and 8Å are considered for calculation of the maximal residue set. The high accuracy version of the GDT measure, GDT-HA [17], uses lower thresholds of 0.5, 1, 2 and 4Å for the calculation of the maximal residue set. The range of GDT-TS and GDT-HA measures are from

0 to 1 with higher scores indicating better agreement of the models compared to the native. GDT-TS and GDT-HA are widely used assessment metrics in the Critical Assessment of protein Structure Prediction (CASP) experiments [21,22].

LG-score, MaxSub score, TM-score, GDT-TS or GDT-HA consider only the main chain $C\alpha$ atoms for quantitating the structural similarity. However, protein side-chains play a major role in defining its conformation at the atomic detail. Therefore, quantifying the side-chain similarities or differences can improve the sensitivities of model-native similarity metrics [23]. Global Distance Calculation for Side-Chains (GDC-SC) [24] is a measure, which determines the correctness of the side-chain positioning. GDC-SC metric is similar to GDT-TS in that it uses a characteristic atom for each residue type instead of relying on the $C\alpha$ atom. Similar to GDT-TS, GDC-SC computes the optimal structural superposition based on the $C\alpha$ atoms of the model vs. native, and subsequently uses the residue-specific side-chain characteristic atom using a distance threshold based approach to quantitate model-native similarity scores. The range of GDC-SC is from 0 to 1 with higher scores indicating better model-native similarity. Although, GDC-SC quantifies the positioning of the side-chain, it only takes into consideration the distances between the side-chain atoms and not their orientation with respect to the backbone—crucial for highly sensitive structural and functional studies based on protein structures that mandates atomistic resolution [25–27]. While existing model-native similarity measures such as LG-score, MaxSub score, TM-score, GDT-TS or GDT-HA, and GDC-SC consider either $C\alpha$ atom distances or side-chain distances, an integrated structural similarity metric that can simultaneously capture the distances between the backbone and side-chain atoms as well as the orientation of side-chain atoms with respect to backbone may offer some advantages.

Here, we integrate side-chain (SC) orientation and global distance based metrics to propose a new superposition-based model-native similarity metric, Superposition-based Protein Embedded $C\alpha$ -SC (SPECS) score. SPECS integrates global $C\alpha$ positioning based distance and side-chain distance and orientation in a singular framework using the united-residue representation [28] for an integrated model-native similarity metric. To the best of our knowledge, this is the first study to propose a protein model evaluation metric that includes side-chain orientation. Furthermore, the seamless integration of $C\alpha$ and SC in the united-residue representation is novel. Experimental results demonstrate that SPECS is a reliable and sensitive model-native similarity measure across a wide range of protein modeling scenarios in that SPECS not only is a reliable measure for evaluating the accuracy of global $C\alpha$ positioning but also captures other aspects of model-native accuracy at the global level beyond just the realm of $C\alpha$ positioning. Moreover, SPECS captures local quality aspect better than some of the most popular global $C\alpha$ positioning-based metrics, for both high-resolution models at near-experimental accuracy and moderate-resolution models with correct backbone positioning. Finally, SPECS successfully captures minute variations of side-chain χ angles even for protein models having perfect $C\alpha$ trace—revealing the effectiveness of including side-chain orientation. Collectively, SPECS is a reliable and sensitive evaluation metric for improved assessment of protein models covering a wide range of modeling scenarios and is highly effective at simultaneously capturing structural aspects at both global and local levels, thereby being a valuable new measure for comprehensive evaluation of protein structural models.

Materials and methods

United-residue representation for structural alignment

We use the united-residue representation of polypeptide conformation [28] as shown in Fig 1. In united-residue representation, the polypeptide chain of the protein is represented by a sequence of $C\alpha$ atoms and characteristic side-chain (SC) atoms, which are attached to the $C\alpha$

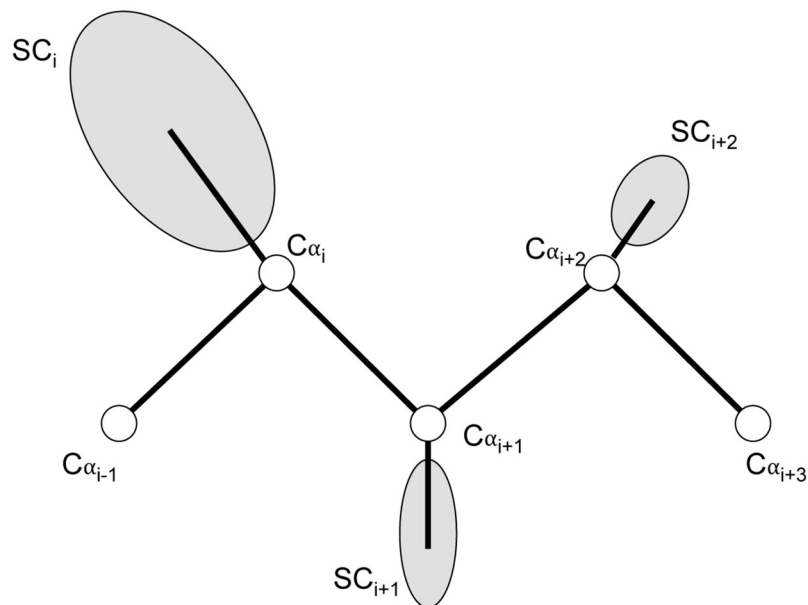


Fig 1. United-residue representation of polypeptide conformation. The polypeptide chain of a protein is represented as a sequence of $C\alpha$ atoms and SC atoms, which are attached to the $C\alpha$ atoms. All the atoms in the united-residue representation are connected using virtual bonds.

<https://doi.org/10.1371/journal.pone.0228245.g001>

atoms. The side-chain characteristic atom is obtained by computing the centroid of all the heavy atoms present in the side-chain of a given residue in its all-atom representation. All atoms of the polypeptide chain in united-residue representation are connected using virtual bonds. The location of the residue i in the polypeptide chain is completely defined by the positioning of $C\alpha_i$ and positioning of the corresponding side-chain characteristic atom SC_i attached to $C\alpha_i$.

In Fig 2, we show structurally aligned model and native structures in the united residue representation. We represent the $C\alpha$ position of the residue i in the model as $C\alpha_i$ and the corresponding aligned residue j in the native as $C\alpha_j$. Consequently, the corresponding side-chain characteristic atom i in the model is represented as SC_i and the characteristic atom j in the native is represented as SC_j . While the distance between $C\alpha_i$ and $C\alpha_j$ is denoted purely by their Euclidean distance d_{ij} , the relative positioning between the side-chain characteristic atoms is represented by the vector \vec{r}_{ij} , the magnitude of which is their Euclidean distance r_{ij} . $\hat{u}_{ij}^{(1)}$, $\hat{u}_{ij}^{(2)}$ are the unit vectors, which represent the direction of the $C\alpha$ and SC virtual bonds in the model and native. $\theta_{ij}^{(1)}$ is the virtual planar angle between $\hat{u}_{ij}^{(1)}$ and \vec{r}_{ij} in the model and $\theta_{ij}^{(2)}$ is the virtual planar angle between $\hat{u}_{ij}^{(2)}$ and \vec{r}_{ij} in the native and they are computed as follows [28]:

$$\theta_{ij}^{(1)} = \cos^{-1}(\hat{u}_{ij}^{(1)} \cdot \vec{r}_{ij}) \tag{1}$$

$$\theta_{ij}^{(2)} = \cos^{-1}(\hat{u}_{ij}^{(2)} \cdot \vec{r}_{ij}) \tag{2}$$

Φ_{ij} is the virtual dihedral angle of counterclockwise rotation between $\hat{u}_{ij}^{(2)}$ and \vec{r}_{ij} in the plane defined by $\hat{u}_{ij}^{(1)}$ and \vec{r}_{ij} and is computed as follows [28]:

$$\Phi_{ij} = \cos^{-1} \left(\frac{\hat{u}_{ij}^{(1)} \cdot \hat{u}_{ij}^{(2)} - \cos \theta_{ij}^{(1)} \cos \theta_{ij}^{(2)}}{\sin \theta_{ij}^{(1)} \sin \theta_{ij}^{(2)}} \right) \tag{3}$$

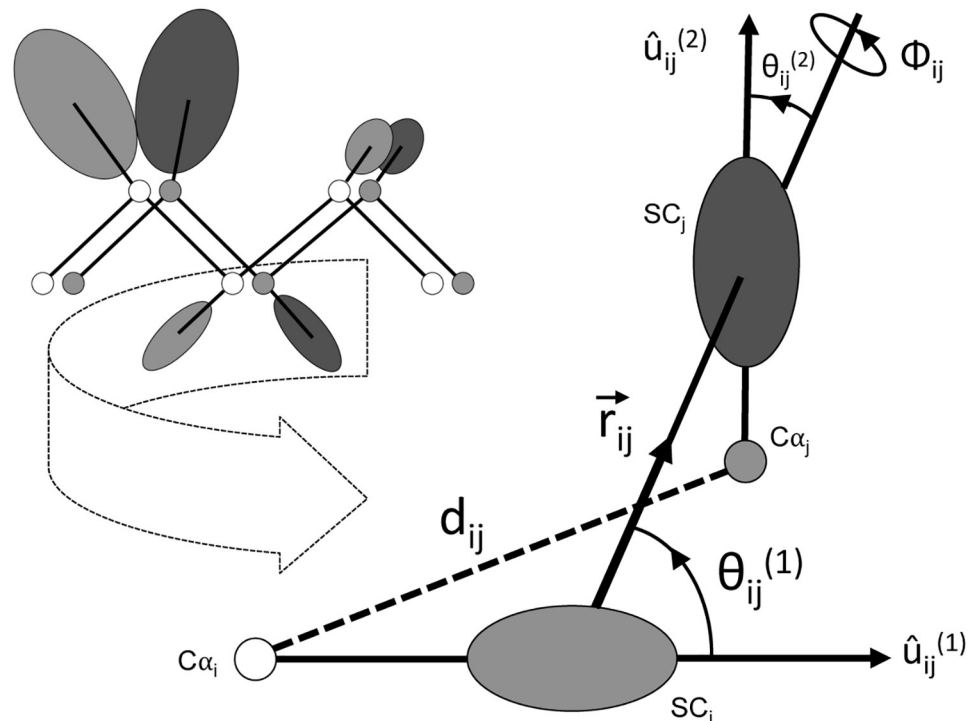


Fig 2. Parameterization of structurally aligned model and native structures in the united-residue representation. The structural alignment between the residue *i* in the model and the corresponding aligned residue *j* in the native is fully captured by two distances d_{ij} and r_{ij} , two planar angles $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(2)}$, and one dihedral angle Φ_{ij} .

<https://doi.org/10.1371/journal.pone.0228245.g002>

Structural alignment between model and native in the united residue representation, therefore, is fully captured by two distances d_{ij} and r_{ij} , two planar angles $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(2)}$, and one dihedral angle Φ_{ij} .

Formulating the side-chain-orientation-included structural similarity metric: SPECS

We utilize the aforementioned united residue representation to formulate side-chain-orientation-included structural similarity metric called SPECS, which stands for Superposition-based Protein Embedded CA SC score. SPECS is a weighted combination of five different components consisting of two distance components based on d_{ij} and r_{ij} , two planar angle components based on $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(2)}$, and one dihedral angle component based on Φ_{ij} .

For computing the first component of SPECS, the optimal structural superposition between model and native is determined based on the $C\alpha$ atom positioning, in order to calculate their Euclidean distances, d_{ij} . Average proportion of model residues having $C\alpha$ atom distances from the corresponding residues in the native structure below four different distance thresholds of 0.5, 1, 2 and 4Å are then calculated, followed by averaging the proportion of residues in four different distance thresholds as:

$$SPECS_{dCA} = \frac{P_{dCA_{0.5}} + P_{dCA_{1}} + P_{dCA_{2}} + P_{dCA_{4}}}{4.0} \quad (4)$$

where $P_{dCA_{0.5}}$, $P_{dCA_{1}}$, $P_{dCA_{2}}$ and $P_{dCA_{4}}$ are the proportions of the set of residues for which d_{ij} values are below distance thresholds of 0.5, 1, 2 and 4Å, respectively. Consequently

$SPECS_{dCA}$, ranges from [0, 1] with higher values indicating better model-native similarity in terms of $C\alpha$ atom distances.

For computing the remaining four components of SPECS, we rely on the optimal structural superposition previously determined based on the positioning of the $C\alpha$ atoms of the model vs. native to rotate and translate the side-chain atoms using the $C\alpha$ positioning-based rotation and translation matrices. For the side-chain distance based component of SPECS, Euclidean distances between the aligned SC atoms in model and native, r_{ij} , are calculated. Subsequently each r_{ij} value is assigned to a distance bin i , with $i = 1$ corresponding to values $\leq 0.5\text{\AA}$ and $i = 10$ corresponding to values $\leq 5.0\text{\AA}$, followed by averaging the proportion of residues in ten different distance bins as:

$$SPECS_{rSC} = \frac{2 \sum_{i=1}^{10} (k - i + 1) p_{rSCi}}{k(k + 1)} \tag{5}$$

where $k = 10$ is the number of bins and p_{rSCi} is the proportion of reference atoms assigned to distance bin i . It should be noted here that a reference atom assigned to a lower distance bin based on its r_{ij} value is, by definition, also assigned to higher distance bins. For example, if the r_{ij} value of a reference atom is less than 0.5\AA , it would be assigned to all the ten bins. $SPECS_{rSC}$ also ranges from [0, 1] with higher values indicating better model-native similarity in terms of SC atom distances.

Next, for computing the side-chain planar angle based components, we divide the $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(2)}$ planar angles into four planar angle bins of $\leq 30^\circ$, $\leq 60^\circ$, $\leq 90^\circ$ and $\leq 120^\circ$, followed by averaging the proportion of residues in four different planar angle bins as:

$$SPECS_{\theta^{(1)}} = \frac{2 \sum_{i=1}^4 (k - i + 1) p_{\theta^{(1)}i}}{k(k + 1)} \tag{6}$$

where $k = 4$ is the number of bins and $p_{\theta^{(1)}i}$ is the proportion of residues assigned to planar angle bin i .

$$SPECS_{\theta^{(2)}} = \frac{2 \sum_{i=1}^4 (k - i + 1) p_{\theta^{(2)}i}}{k(k + 1)} \tag{7}$$

where $k = 4$ is the number of bins and $p_{\theta^{(2)}i}$ is the proportion of residues assigned to planar angle bin i . Analogous to the distance bins, a residue belonging to a lower planar angle bin automatically falls in all higher planar angle bins. For example, if a residue's $\theta_{ij}^{(1)}$ value is less than 30° , the residue would be assigned to all the four bins in Eq 6. Also, if a residue's $\theta_{ij}^{(2)}$ value is less than 30° , the residue would be assigned to all the four bins in Eq 7.

Once again, $SPECS_{\theta^{(1)}}$ and $SPECS_{\theta^{(2)}}$ also range from [0, 1] with higher values indicating better model-native similarity in terms of the planar angle components of the side-chain orientations.

Next, for computing the side-chain dihedral angle based component, we divide the Φ_{ij} dihedral angle into ten bins of $\leq 30^\circ$, $\leq 60^\circ$, $\leq 90^\circ$, $\leq 120^\circ$, $\leq 150^\circ$, $\leq 180^\circ$, $\leq 201^\circ$, $\leq 240^\circ$, $\leq 270^\circ$ and $\leq 300^\circ$, followed by averaging the proportion of residues in ten different dihedral angle bins as shown below:

$$SPECS_{\Phi} = \frac{2 \sum_{i=1}^{10} (k - i + 1) p_{\Phi i}}{k(k + 1)} \tag{8}$$

where $k = 10$ is the number of bins and $p_{\Phi i}$ is the proportion of residues assigned to dihedral angle bin i . Of note, the assignment of a residue to a lower dihedral angle bin automatically

qualifies the residue to be assigned to all higher dihedral angle bins. For instance, if a residues' Φ_{ij} value is less than 30° , the residue would be assigned to all the ten bins. Once again, $SPECS_\Phi$ also ranges from $[0, 1]$ with higher values indicating better model-native similarity in terms of the dihedral angle component of the side-chain orientations.

Finally, SPECS is calculated as a weighted average of $SPECS_{dCA}$, $SPECS_{rSC}$, $SPECS_\theta^{(1)}$, $SPECS_\theta^{(2)}$, and $SPECS_\Phi$ as:

$$SPECS = \frac{4 * SPECS_{dCA} + SPECS_{rSC} + SPECS_{\theta^{(1)}} + SPECS_{\theta^{(2)}} + SPECS_\Phi}{8.0} \quad (9)$$

In this scoring scheme, equal weights are assigned to both the main chain and the side-chain based components, to equally emphasize the importance of $C\alpha$ and SC positioning. The $C\alpha$ distance based component, $SPECS_{dCA}$ is given a weight of 4, which makes half of the overall score and the four side-chain based components make the other half.

Datasets and similarity metrics used for benchmarking

We benchmark SPECS against four datasets. The first dataset is the CASP12 [29] and CASP13 regular target sets consisting of 55 and 32 regular domains for CASP12 and CASP13, respectively, with publicly available experimental structures. We use this dataset to compare SPECS against three popular model-native similarity metrics: GDT-TS [13], TM-score [14] and SphereGrinder [30]. GDT-TS and TM-score are both superposition-based global similarity scores, which determine the model-native similarity based on the distances between the $C\alpha$ atoms. SphereGrinder is based on an all-atom RMSD fit between the model and native structures, using a sphere constructed by considering the set of atoms within 6\AA of the $C\alpha$ atoms for each residue in the native structure.

The second dataset is the CASP12 [29] and CASP13 refinement target sets consisting of a total of 37 refinement target domains with publicly available experimental structures. We use this set to compare SPECS against four high-resolution model-native similarity metrics: GDT-HA [17], CAD-AA (all atoms) [31], GDC-SC [24] and IDDT [15]. GDT-HA is a superposition-based score, which determines the model-native similarity based on the distances between the $C\alpha$ atoms. GDC-SC is a superposition-based score, which determines the model-native similarity based on the distances between the side-chain characteristic atoms. CAD-AA and IDDT are all-atom based superposition-free scores.

The third dataset is the 3DRobot [32] decoy set, which consists of 200 non-homologous protein targets each having 300 decoys. 3DRobot generates a well-packed decoy pool with an even distribution of decoy accuracy over the Root Mean Square Deviation (RMSD) space with respect to the native. We use this set to evaluate the agreement between SPECS and MolProbity [33] as a local structure quality estimator and compare with two $C\alpha$ atom based model-native similarity metrics GDT-HA score and TM-score. MolProbity is a log-weighted combination of the clash score, percentage of Ramachandran not favored and the percentage of bad side-chain rotamers, giving one number that reflects the crystallographic resolution at which those values would be expected. Thus, lower MolProbity scores indicate enhanced stereochemistry and better physical realism. It should be noted here, that unlike the other scores, MolProbity does not determine the local quality of a model by comparing it with the native. MolProbity score is not native-dependent and hence significantly distinct from the other scoring functions used in this work.

The fourth dataset is a monomeric proteins dataset [37], which consists of 229 protein models and 33,461 residues. These models have perfect $C\alpha$ positioning with respect to the native, but possess varying side-chain conformations. We use this set to evaluate the ability of SPECS

to capture the correctness of side-chain χ angles. Three widely-used side-chain prediction methods RASP [34], Rosetta-fixbb [35] and SCWRL4 [36] are used to rebuild the side-chain given the $C\alpha$ trace [37]. RASP [34] is designed for rapid prediction of side-chain conformations by efficient elimination of atomic clashes and relaxation. Rosetta-fixbb [35] employs a Monte Carlo optimization approach to optimize the side-chain placement on a fixed backbone. SCWRL4 [36] utilizes a backbone-dependent rotamer library in conjunction with graph decomposition algorithms to solve the combinatorial side-chain packing problem. The prediction accuracies of these three methods are evaluated in terms of the Angular RMSDs of the χ_1 side-chain torsion angles. The χ_1 angle is the dihedral angle between the planes defined by the atoms N, $C\alpha$, $C\beta$, and $C\gamma$. We first calculate χ_1 angle for every residue using the PDB module [38] of the Biopython package [39] to compute the Angular RMSD at the target level, from the corresponding χ_1 angles [40] as:

$$\text{Angular RMSD} = \sqrt{\frac{1}{n} \sum_i (\min(|x_{2i} - x_{1i}|, 2\pi - |x_{2i} - x_{1i}|)^2)} \quad (10)$$

where x_1 is the vector of χ_1 angles for n residues in the target and x_2 is the vector of corresponding χ_1 angles for n residues in the native. Consequently, a lower Angular RMSD indicates better average accuracy in terms of side-chain dihedral angles. To facilitate a head-to-head comparison between SPECS and the average accuracy of side-chain dihedral angles, we subsequently normalize the Angular RMSD as:

$$\text{Normalised Angular RMSD} = \frac{1}{1 + \left(\frac{\text{Angular RMSD}}{\pi/4}\right)^2} \quad (11)$$

Results and discussion

SPECS is a reliable measure for evaluating the accuracy of global $C\alpha$ positioning

To investigate the ability of SPECS to quantitate model-native accuracy at the global level based on $C\alpha$ positioning, we benchmark SPECS on the regular target domain from CASP12 [29] and CASP13, and compare it with the existing $C\alpha$ based model-native similarity metrics. The CASP12 set consists of 55 target domains and the CASP13 decoy set consists of 32 target domains. The targets were divided into template-based (TBM), free modeling (FM) and overlapped (TBM/FM) categories as defined by the assessors. GDT-TS [13], TM-score [14] and SphereGrinder [30] are directly taken from the data archive of the Prediction Center (<http://www.predictioncenter.org/>), whereas SPECS is calculated by comparing the model with the native. Fig 3 shows the relationships between SPECS and GDT-TS, TM-score, SphereGrinder. The average Pearson and Spearman correlation coefficients, as shown in Fig 3, indicate that SPECS is highly correlated to other scores in that the average Pearson and Spearman correlations with respect to GDT-TS, TM-score and SphereGrinder are always greater than 0.8 in both CASP12 and CASP13 datasets, where SPECS attains the highest correlation with GDT-TS score. In CASP12 dataset, the average Pearson and Spearman correlation between SPECS and GDT-TS are 0.95 and 0.94 respectively followed by 0.89 and 0.83 respectively between SPECS and TM-score, followed by 0.87 and 0.82 respectively between SPECS and SphereGrinder. We find a similar trend in CASP13 dataset, where the average Pearson and Spearman correlations between SPECS and GDT-TS are both 0.94 respectively, followed by 0.94 and 0.93 respectively between SPECS and TM-score, followed by 0.89 and 0.85 respectively between SPECS and SphereGrinder. The persistency of strong correlations, therefore,

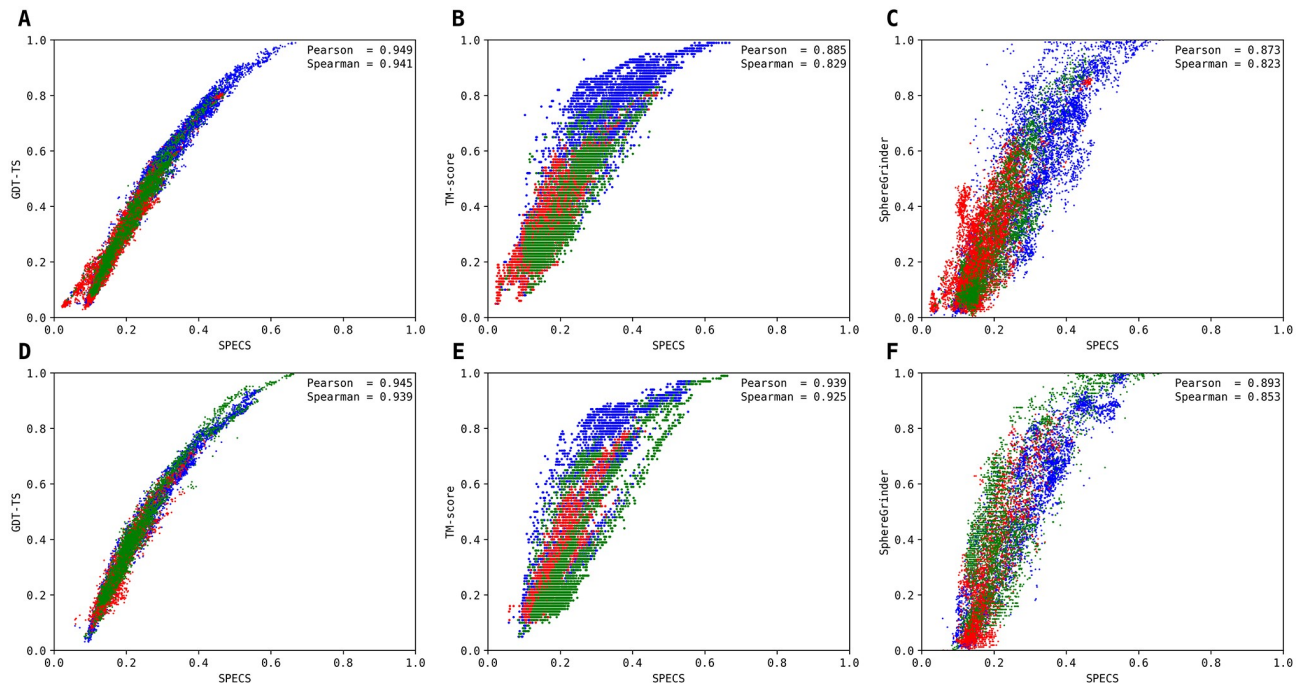


Fig 3. Comparisons between SPECS (horizontal axis) and GDT-TS, TM-score and SphereGrinder (vertical axis) using models in CASP12 (A-C) and CASP13 (D-F) regular single domain targets. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot. Blue, red, and green colors represent models assessed in template-based (TBM), free modeling (FM) and overlapped (TBM/FM) categories respectively.

<https://doi.org/10.1371/journal.pone.0228245.g003>

demonstrates that SPECS is a reliable measure for evaluating model-native similarity at the global level, determined purely based on the accuracy of $C\alpha$ positioning.

Beyond $C\alpha$ positioning: SPECS captures other aspects of accuracy at the global level

To examine the ability of SPECS to capture other accuracy aspects at the global level beyond just the realm of $C\alpha$ positioning, we next compare SPECS with one high-resolution $C\alpha$ positioning metric GDT-HA [17], and three other metrics capturing other aspects of accuracy at the global level: (i) CAD-AA [31], based on contact area difference; (ii) GDC-SC [24] based on side-chain placement; and (iii) IDDT [15] based on local distance difference; using refinement targets from CASP12 [29] and CASP13 refinement experiments. Overall there are 37 targets (34 from CASP12 and 3 from CASP13) for which the native structures are available. Once again, GDT-HA, CAD-AA, GDC-SC and IDDT scores are taken directly from the data archive of the Prediction Center (<http://www.predictioncenter.org/>), whereas SPECS is calculated by comparing the model against the native. Fig 4 shows the relationships between SPECS and superposition-based scores such as GDT-HA and GDC-SC as well as superposition-free scores such as CAD-AA and IDDT. The average Pearson and Spearman correlation coefficients, as shown in Fig 4, indicate that SPECS is well-correlated to other scores in that the average Pearson and Spearman correlations always remain greater than 0.8 with the only exception between the SPECS and the IDDT having a Spearman correlation of 0.76. Similar to GDT-TS, SPECS is highly correlated with GDT-HA where the Spearman and Pearson correlations are 0.99 and 0.96 respectively. Thereafter, SPECS achieves the Pearson correlation of 0.91 with GDC-SC followed by 0.88 with CAD-AA followed by 0.87 with IDDT. Similarly, the Spearman

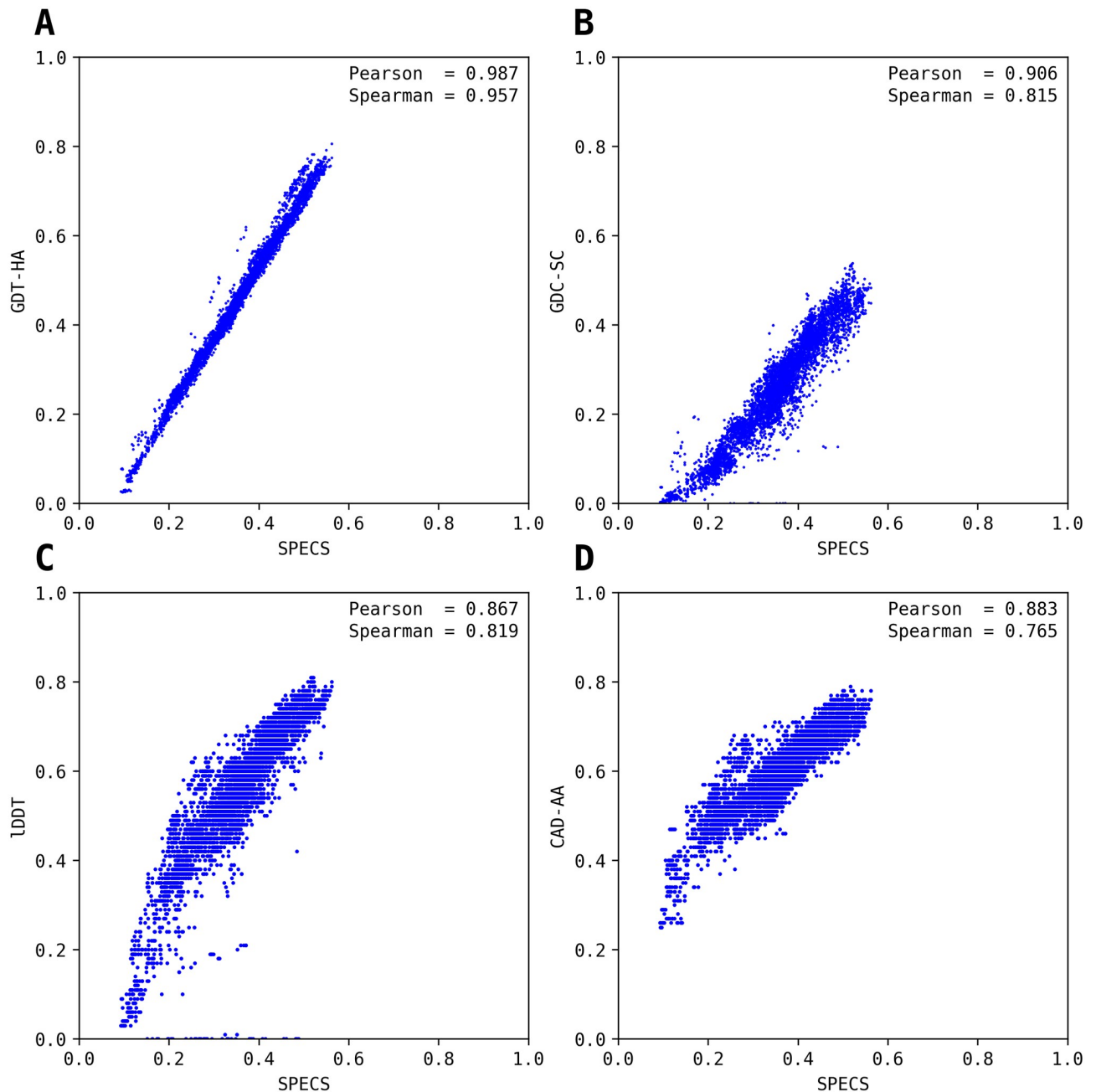


Fig 4. Comparisons between SPECS (horizontal axis) and existing model-native similarity metrics namely GDT-HA (A), GDC-SC (B), IDDT (C) and CAD-AA (D) (vertical axis) using models in CASP12 and CASP13 refinement targets. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot.

<https://doi.org/10.1371/journal.pone.0228245.g004>

correlation between SPECS and IDDT is 0.8 followed by 0.82 between SPECS and GDC-SC, followed by 0.76 between SPECS and CAD-AA. This strong correlation, therefore, substantiates that SPECS is not only strongly correlated with $C\alpha$ positioning based accuracy metrics like GDT-HA, but also side-chain based similarity metrics like GDC-SC, and all-atom based similarity metrics like CAD-AA and IDDT. Overall, the results demonstrate the ability of the SPECS to capture other aspects of model-native accuracy at the global level including and beyond the realm of $C\alpha$ positioning.

SPECS captures local quality aspect better than global C α based metrics

To assess the effectiveness of SPECS in capturing the local qualities of the models including stereochemistry and physical reasonableness, we evaluate it on 3DRobot set [32]. 3DRobot set consists of 200 non-homologous protein targets each with 300 decoys. From the entire pool consisting of 60,000 protein models, we consider models belonging to three RMSD bins namely $< 2\text{\AA}$, $< 4\text{\AA}$ and $< 6\text{\AA}$ based on their C α RMSD scores with respect to the natives and one TM-score bin consisting of decoys with TM-score > 0.5 . The three C α RMSD bins represent near-native accuracy, high accuracy, and medium accuracy protein models, respectively and TM-score > 0.5 represents protein models with correct overall fold. Models not belonging to any of these four bins are incorrectly folded and therefore not suitable for local quality analyses are excluded. To understand the relationship between the SPECS score assigned to a model and its physical realism, we analyze pairs of models for which SPECS vs. TM-score [14] and SPECS vs. GDT-HA [17] are in conflict. Between these conflicting pairs of models, we compare the agreement of the SPECS vs. GDT-HA and TM-score with MolProbity, which is a local quality estimator [33]. Fig 5A and 5B shows that the percentage of agreement in the ranking between SPECS and MolProbity score is consistently better compared to that between GDT-HA and MolProbity score across the $< 2\text{\AA}$ and $< 4\text{\AA}$ C α RMSD bins, indicating that SPECS is a robust measure for capturing local quality compared to GDT-HA for high-resolution protein models. Fig 5C and 5D shows that the percentage of agreement in the ranking between SPECS and MolProbity score is better compared to that between TM-score and MolProbity in $< 6\text{\AA}$ RMSD bin and when TM-score > 0.5 , indicating that SPECS is a robust measure for capturing local quality compared to TM-score for moderate-resolution protein models and for those with correct overall folds. Consistently better agreement between SPECS and MolProbity in all the four bins indicates that SPECS captures local quality aspect better than global C α positioning-based metrics, both for high- and moderate-resolution models.

SPECS is sensitive to minute variations in side-chain χ angles

To analyze the ability of SPECS to capture variations in side-chain χ angles in models having perfect C α trace, we analyze the side-chain χ angles of the monomeric proteins predicted by three widely used side-chain prediction methods RASP [34], Rosetta-fixbb [35], and SCWRL4 [36]. The C α atoms of the predicted models in the dataset are perfectly aligned with respect to the native resulting in 0\AA C α -RMSDs, enabling the assessment of the structural similarity purely based on the side-chain variations. The average Angular RMSD values of the side-chain conformation predicted by the three methods are shown in Fig 6, showing the relative accuracies of the three methods based on their Angular RMSD values, with RASP ranked as the best, followed by SCWRL4, followed by Rosetta-fixbb. In Table 1, we report the correlations between SPECS and the normalized Angular RMSD values of the side-chain conformation predictions for three methods. The results demonstrate that there is a weak but positive correlation between SPECS and normalized Angular RMSD with the most accurate side-chain predictor RASP attaining the highest correlation, followed by SCWRL4, followed by Rosetta-fixbb. It should be noted here that because of the perfect C α traces, accuracies of these models appear to be perfect (i.e., having scores of 1.0) when measured with some of the widely used structural similarity metrics such as GDT-TS, GDT-HA, and TM-score. In contrast, SPECS offers an added ability to rank these models, albeit based on the minute variations in the side-chain χ angles, thus making it more sensitive for evaluation of protein structural models.

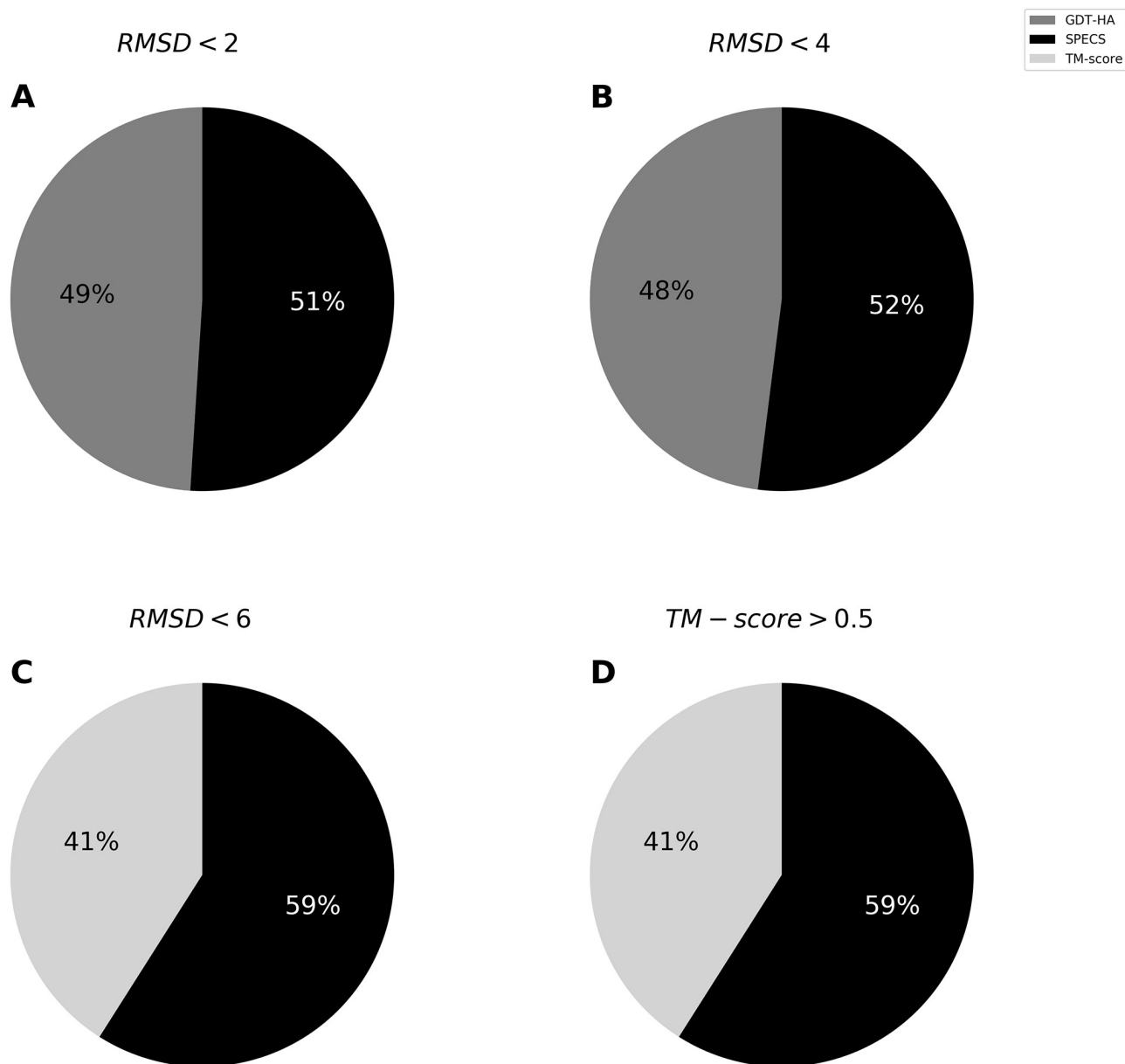


Fig 5. Pairs of 3DRobot models with conflicting ranking by SPECS vs. GDT-HA and TM-score with MolProbity. The 3DRobot models are divided into three bins $< 2\text{\AA}$, $< 4\text{\AA}$ and $< 6\text{\AA}$ based on their $C\alpha$ RMSD scores with respect to the natives and an additional bin consisting of decoys with TM-score > 0.5 . Pie charts represent the percentages of MolProbity score agreement with rankings by SPECS vs. GDT-HA (A-B), SPECS vs. TM-score (C-D).

<https://doi.org/10.1371/journal.pone.0228245.g005>

Conclusion

We present a side-chain-orientation included model-native similarity score, SPECS, which seamlessly combines side-chain orientation and the global distance based measures at the united-residue representation for improved assessment of protein structural models. SPECS is a weighted combination of five different components comprising of two distance based components quantifying the positioning of the $C\alpha$ and SC atoms and three angle based components capturing side-chain orientation. Experimental results demonstrate that SPECS is a

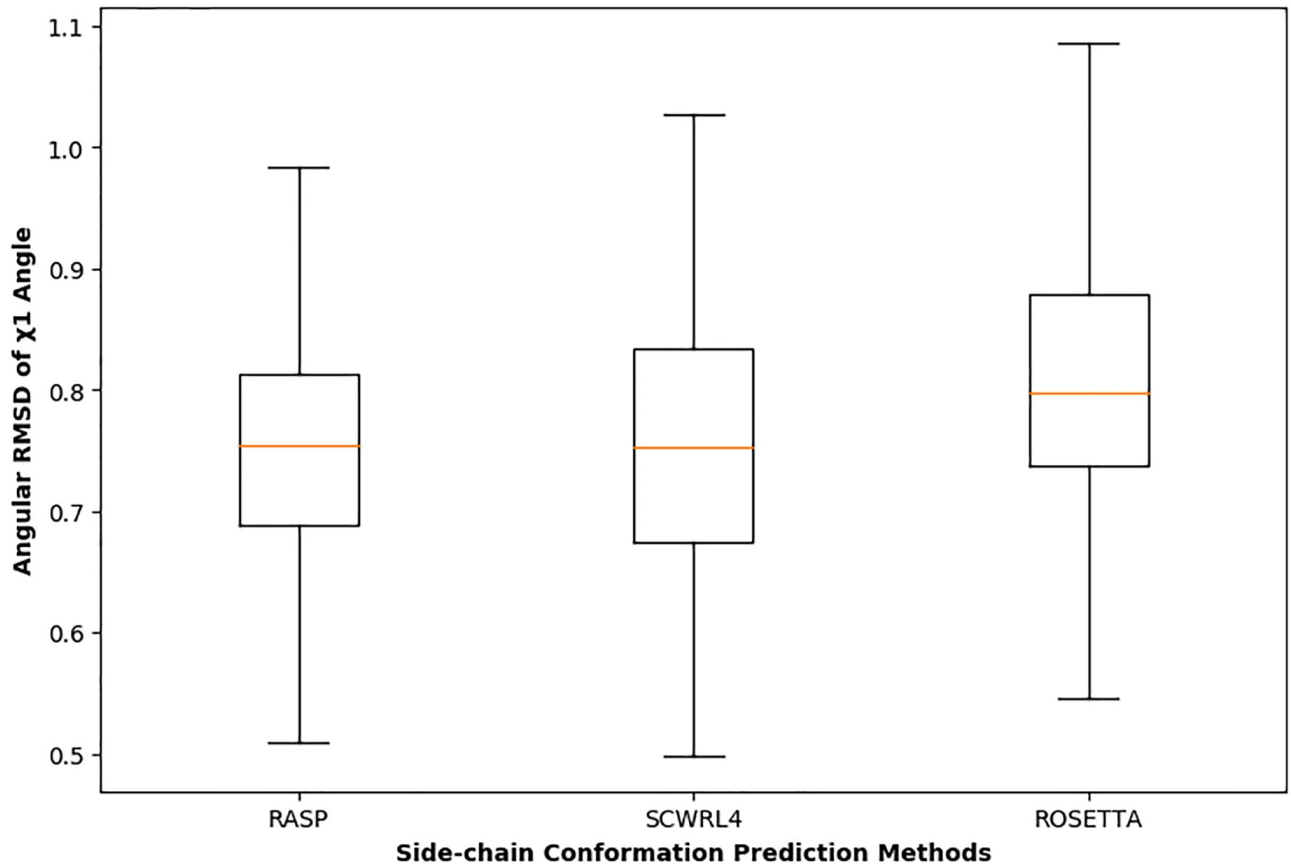


Fig 6. Distributions of angular RMSDs of side-chain χ angles. Lower and upper hinges: 1st and 3rd quartile. Whisker length: 1.5 times the interquartile range.

<https://doi.org/10.1371/journal.pone.0228245.g006>

reliable and robust evaluation measure for protein models covering various structural aspects at both the global and local levels by being highly correlated with several global model-native similarity metrics including superposition-based scores such as GDT-TS, GDT-HA, GDC-SC, SphereGrinder, TM-score, and superposition-free scores such as CAD-AA and IDDT as well as local quality measures such as MolProbity. Moreover, SPECS offers an added ability to rank models having only minute variations in the side-chain χ angles but with perfect $C\alpha$ traces, which are indistinguishable by various popular global structural similarity metrics. Collectively, these results demonstrate that SPECS is a reliable, robust, and sensitive model-native similarity metric for improved assessment of protein models that covers a wide range of protein modeling scenarios and encapsulates various aspects of structural similarity.

Table 1. Spearman correlations between SPECS and normalized Angular RMSDs of side-chain conformation prediction methods.

Prediction Method	Spearman Correlation
RASP	0.41
SCWRL4	0.32
ROSETTA	0.30

<https://doi.org/10.1371/journal.pone.0228245.t001>

Supporting information

S1 Table. Target by target Pearson and Spearman correlations of SPECS with GDT-TS, TM-score and SphereGrinder scores on CASP12 regular single domain Targets.
(DOCX)

S2 Table. Target by target Pearson and Spearman correlations of SPECS with GDT-TS, TM-score and SphereGrinder scores on CASP13 regular single domain targets.
(DOCX)

S3 Table. Target by target Pearson and Spearman correlations of SPECS with GDT-HA, GDC-SC, IDDT and CAD-AA scores on CASP12 and CASP13 refinement targets.
(DOCX)

S4 Table. Target by target Angular RMSD of χ_1 angle and SPECS on side chain conformations predicted by RASP.
(DOCX)

S5 Table. Target by target Angular RMSD of χ_1 angle and SPECS on side chain conformations predicted by Rosetta-fixbb.
(DOCX)

S6 Table. Target by target Angular RMSD of χ_1 angle and SPECS on side chain conformations predicted by SCWRL4.
(DOCX)

Acknowledgments

The authors would like to thank Rahmatullah Roche for helpful discussions.

Author Contributions

Conceptualization: Debswapna Bhattacharya.

Data curation: Rahul Alapati, Md. Hossain Shuvo, Debswapna Bhattacharya.

Funding acquisition: Debswapna Bhattacharya.

Investigation: Rahul Alapati, Md. Hossain Shuvo, Debswapna Bhattacharya.

Methodology: Debswapna Bhattacharya.

Project administration: Debswapna Bhattacharya.

Resources: Debswapna Bhattacharya.

Software: Debswapna Bhattacharya.

Supervision: Debswapna Bhattacharya.

Validation: Debswapna Bhattacharya.

Visualization: Debswapna Bhattacharya.

Writing – original draft: Rahul Alapati, Md. Hossain Shuvo, Debswapna Bhattacharya.

Writing – review & editing: Rahul Alapati, Md. Hossain Shuvo, Debswapna Bhattacharya.

References

1. Arakaki AK, Zhang Y, Skolnick J. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*. 2004; 20: 1087–1096. <https://doi.org/10.1093/bioinformatics/bth044> PMID: 14764543
2. Wieman H. Homology-Based Modelling of Targets for Rational Drug Design. *Mini Reviews in Medicinal Chemistry*. 2012; 4.
3. Zhang Y. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*. 2009; 19: 145–155. <https://doi.org/10.1016/j.sbi.2009.02.005> PMID: 19327982
4. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*. 2009; 14: 676–683. <https://doi.org/10.1016/j.drudis.2009.04.006> PMID: 19422931
5. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*. 2012; 7: 1511–1522. <https://doi.org/10.1038/nprot.2012.085> PMID: 22814390
6. Li J, Bhattacharya D, Cao R, Adhikari B, Deng X, Eickholt J, et al. The MULTICOM Protein Tertiary Structure Prediction System. In: Kihara D, editor. *Protein Structure Prediction*. New York, NY: Springer New York; 2014. pp. 29–41.
7. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology*. Elsevier; 2004. pp. 66–93. [https://doi.org/10.1016/S0076-6879\(04\)83004-0](https://doi.org/10.1016/S0076-6879(04)83004-0) PMID: 15063647
8. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. 2010; 5: 725–738. <https://doi.org/10.1038/nprot.2010.5> PMID: 20360767
9. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*. 2012.
10. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008; 9.
11. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*. 1993; 233: 123–138. <https://doi.org/10.1006/jmbi.1993.1489> PMID: 8377180
12. Kufareva I, Abagyan R. Methods of Protein Structure Comparison. In: Orry AJW, Abagyan R, editors. *Homology Modeling*. Totowa, NJ: Humana Press; 2011. pp. 231–257.
13. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. 2003; 31: 3370–3374. <https://doi.org/10.1093/nar/gkg571> PMID: 12824330
14. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*. 2004; 57: 702–710.
15. Mariani V, Biasini M, Barbato A, Schwede T. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013; 29: 2722–2728. <https://doi.org/10.1093/bioinformatics/btt473> PMID: 23986568
16. Koehl P. Protein structure similarities. *Current Opinion in Structural Biology*. 2001; 11: 348–353. [https://doi.org/10.1016/S0959-440x\(00\)00214-1](https://doi.org/10.1016/S0959-440x(00)00214-1) PMID: 11406386
17. Kopp J, Bordoli L, Battey JND, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins: Structure, Function, and Bioinformatics*. 2007; 69: 38–56.
18. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallograp Sec A*. 1976; 32: 922–923.
19. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences*. 1998; 95: 5913–5920.
20. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*. 2000; 16: 776–785. <https://doi.org/10.1093/bioinformatics/16.9.776> PMID: 11108700
21. Moulton J, Fidelis K, Kryshchuk A, Schwede T, Tramontano A. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Progress and New directions in Round XI. *Proteins*. 2016; 84: 4–14. <https://doi.org/10.1002/prot.25064> PMID: 27171127
22. Moulton J, Fidelis K, Kryshchuk A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins*. 2018; 86 Suppl 1: 7–15.
23. Miao Z, Cao Y. Quantifying side-chain conformational variations in protein structure. *Scientific Reports*. 2016; 6.
24. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Assessment of the protein-structure refinement category in CASP8. *Proteins: Structure, Function, and Bioinformatics*. 2009; 77: 66–80.

25. Buczek A, Siodlak D, Bujak M, Broda MA. Effects of Side-Chain Orientation on the Backbone Conformation of the Dehydrophenylalanine Residue. Theoretical and X-ray Study. *The Journal of Physical Chemistry B*. 2011; 115: 4295–4306. <https://doi.org/10.1021/jp200949t> PMID: 21443240
26. Chien Y-T, Huang S-W. Accurate Prediction of Protein Catalytic Residues by Side Chain Orientation and Residue Contact Density. Srinivasan N, editor. *PLoS ONE*. 2012; 7: e47951. <https://doi.org/10.1371/journal.pone.0047951> PMID: 23110141
27. Fraser R.D.B. Side-chain orientation in fibrous proteins. *Nature*. 1955; 176: 358. <https://doi.org/10.1038/176358a0> PMID: 13253577
28. Liwo A, Odziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry*. 1997; 18: 849–873.
29. Moulton J, Fidelis K, Kryshchuk A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins: Structure, Function, and Bioinformatics*. 2018; 86: 7–15.
30. Kryshchuk A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL: CASP Prediction Center. *Proteins: Structure, Function, and Bioinformatics*. 2014; 82: 7–13.
31. Olechnovič K, Kulberkytė E, Venclovas Č. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins: Structure, Function, and Bioinformatics*. 2013; 81: 149–162.
32. Deng H, Jia Y, Zhang Y. 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*. 2016; 32: 378–387. <https://doi.org/10.1093/bioinformatics/btv601> PMID: 26471454
33. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*. 2010; 66: 12–21.
34. Miao Z, Cao Y, Jiang T. RASP: rapid modeling of protein side chain conformations. *Bioinformatics*. 2011; 27: 3117–3122. <https://doi.org/10.1093/bioinformatics/btr538> PMID: 21949272
35. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*. 2000; 97: 10383–10388.
36. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*. 2009; 77: 778–795.
37. Peterson LX, Kang X, Kihara D. Assessment of protein side-chain conformation prediction methods in different residue environments: Side-Chain Conformation Prediction Accuracy. *Proteins: Structure, Function, and Bioinformatics*. 2014; 82: 1971–1984.
38. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. *Bioinformatics*. 2003; 19: 2308–10. <https://doi.org/10.1093/bioinformatics/btg299> PMID: 14630660
39. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009; 25: 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163> PMID: 19304878
40. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*. 2008; 105: 8932–8937.