

Antibody side chain conformations are position-dependent.

Short title

Antibody side chain conformations are position-dependent.

Keywords

Side chain prediction — Protein structure prediction — Antibodies — Antibody Design — Rotamer Library

Authors

Jinwoo Leem¹ , Guy Georges², Jiye Shi³, and Charlotte M. Deane¹ 

Affiliations

¹ Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB

² Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Nonnenwald 2, 82377, Penzberg, Germany

³ Chemistry Department, UCB, 208 Bath Road, Slough SL1 3WE, UK

Corresponding author

Professor Charlotte M. Deane

email: deane@stats.ox.ac.uk

address: Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/prot.25453

© 2018 Wiley Periodicals, Inc.

Received: Sep 26, 2017; Revised: Dec 15, 2017; Accepted: Jan 05, 2018

Abstract

Side chain prediction is an integral component of computational antibody design and structure prediction. Current antibody modelling tools use backbone-dependent rotamer libraries with conformations taken from general proteins. Here we present our antibody-specific rotamer library, where rotamers are binned according to their IMGT position, rather than their local backbone geometry. We find that for some amino acid types at certain positions, only a restricted number of side chain conformations are ever observed. Using this information, we are able to reduce the breadth of the rotamer sampling space. Based on our rotamer library, we built a side chain predictor, PEARS. On a blind test set of 95 antibody model structures, PEARS had the highest average χ_1 and χ_{1+2} accuracy (78.7% and 64.8%) compared to three leading backbone-dependent side chain predictors. Our use of IMGT position, rather than backbone ϕ/ψ , meant that PEARS was more robust to errors in the backbone of the model structure. PEARS also achieved the lowest number of side chain-side chain clashes. PEARS is freely available as a web application at <http://opig.stats.ox.ac.uk/webapps/pears>.

Introduction

Antibodies are proteins of the adaptive immune system. They form an important class of biotherapeutics – over 50 are currently in phase II or III studies [1]. Currently, experimental antibody design is a resource-intensive process, and there has been a growing interest in developing computational antibody design methods. Several studies [2–5] have demonstrated the potential of using computational techniques; however, *in silico* methods are not yet accurate enough for more extensive, large-scale therapeutic antibody design applications [6, 7]. A key component that can be improved within these pipelines is structure modelling – specifically, side chain prediction [8, 9]. Having the ability to correctly predict side chains, particularly at the antigen binding site, may improve the accuracy of scoring functions [10, 11] and facilitate computational mutation [12, 13]. In addition, side chains can affect the structure of the complementarity-determining region (CDR) loops [14, 15] and/or the elbow angle [16].

During side chain prediction, the torsion angles of the side chains, the χ angle(s), must be determined using only the existing backbone structure [12, 17–19]. This is often done by describing the series of χ angles as discrete forms, known as rotamers [20]. Each amino acid has a distinct set of χ angle preferences due to steric constraints between atoms [21]. For instance, the χ_1 angle distribution of serine shows peaks at 60°, 180°, and 300°; these correspond to the gauche + ($g+$), trans (t), and gauche – ($g-$) forms, respectively. Although side chain prediction can be simplified to predicting each residue's rotamer, the solution space still remains large [18].

Side chain predictors generally rely on three major components: an energy function, a search method, and a rotamer library [22]. The energy function is used as an objective function for search methods, *e.g.* dead-end elimination (DEE) or machine learning, to identify a solution [12, 22–24]. A rotamer library is a statistical model of the χ angle distributions of side

chains. There are three broad classes of rotamer libraries: backbone-independent [25], backbone-dependent [20, 26], and 'other', where the rotamer probability depends on features such as the secondary structure [27], or the amino acid's position in a protein fragment [28]. Backbone-dependent rotamer libraries are the most common and used in many leading side chain predictors, for example, SCWRL, RASP, and SIDEpro [12, 22, 24]. Backbone-dependent libraries calculate the probability of an amino acid's rotamer as a function of its ϕ/ψ angles, as the ϕ, ψ, χ angles are correlated [29].

Backbone-dependent libraries bin rotamers as a function of the ϕ/ψ angles, so the accuracy of the backbone has a strong influence on the accuracy of predicted χ angles. Only a handful of studies have used backbone-dependent libraries to predict side chains on models [9, 30], where the backbone may be incorrect. In addition, most side chain predictors use a rotamer library developed using all proteins [20, 26]. This assumes that all proteins have similar χ angle preferences for a given ϕ/ψ . A benchmark on membrane proteins has shown that their side chains are more difficult to predict than soluble proteins [31], suggesting that current rotamer libraries and thus, side chain predictors, may not be optimal for membrane proteins. Extending from these observations, and the fact that family-specific predictors are more accurate [32], it may be that family-specific predictors, where enough data is available, could improve side chain prediction.

In this paper, we describe a family-specific side chain predictor for antibodies. As of March 2017, there are over 2600 antibody structures in the PDB [33], making it feasible to develop an antibody-specific rotamer library and prediction method. A feature of antibody structures is their degree of structural conservation [34]. This conservation is captured and described using antibody numbering schemes [35–37]. In this paper, we use the IMGT numbering [37]. Each IMGT position describes the local environment of that position within the antibody and there is often a limited repertoire of amino acids at a given position [38, 39].

We first describe the rotameric preferences of the side chains of antibodies, and show that at specific IMGT positions, certain amino acids adopt specific χ_1 angle states. Using these patterns and our antibody-specific side chain data, we apply our side chain prediction algorithm, PEARS (Position-dEpendent Antibody Rotamer Swapper), to predict side chains on a non-redundant set of 639 antibody structures and a blind test set of 95 antibody structures. We also benchmark our method on model structures of both these sets. We find that PEARS is comparable to other methods in predicting the side chains of crystal structures, while on model structures, PEARS achieves the highest average accuracy. Furthermore, PEARS predicts most targets with almost zero side chain-side chain clashes.

Our position-dependent rotamer library offers a new way of analysing and harnessing rotamer data for side chain prediction. It is best suited to scenarios where there are uncertainties in the model backbone, such as computational antibody design applications. PEARS is freely available as a web application in <http://opig.stats.ox.ac.uk/webapps/pears>.

Methods

Datasets

Antibody variable fragment (Fv) structures with resolution $\leq 2.5\text{\AA}$ were downloaded from SAbDab on 27 Jan 2016 [33]. Structures were clustered using CD-HIT [40], with a 90% Fv sequence identity cutoff. This non-redundant set of 639 antibodies (617 V_H , 562 V_L chains) was used to build our position-dependent rotamer library. The same set of 639 antibodies was used to make two test sets. The ‘crystal set’ is identical to the training set. We also made a ‘model set’, where ABodyBuilder [9] was used to generate a model structure for each structure in the crystal set. Briefly, ABodyBuilder searches for a template structure from SAbDab [33] to model the framework region. It then uses FREAD to model the CDR loops [41]; if a suitable decoy cannot be found for a particular CDR loop, it is predicted using MODELLER [42]. For predictions, information from identical sequences was not used.

We downloaded an additional 192 antibody structures with resolution $\leq 2.5\text{\AA}$, which were released between 27 January 2016 and 23 March 2017. Structures were clustered using CD-HIT [40] with a 90% Fv sequence identity cutoff, leading to an initial set of 126 antibodies. In order to avoid redundancy with the training set, we removed antibodies that shared 90% sequence identity to the training set; this led to a final set of 95 antibodies. We refer to the crystal and model structures of these 95 antibodies as the ‘blind crystal’ and ‘blind model’ sets. For all structures, we numbered the structures in the IMGT numbering scheme [37] using ANARCI [43].

Construction of the position-dependent rotamer library

For every observed side chain in our training set of 639 antibodies, we discretised the χ angles using definitions from the Dynaemomics library [20]. Although eight amino acids are non-rotameric [10], we treated them as rotameric, following previous studies [12].

For each IMGT position i with amino acid α (henceforth known as a ‘side chain type’), the probability of rotamer r was calculated as the relative frequency of r among all observed rotamers of α at i . We also calculated the position-independent probability of r . For each rotamer, we calculated the average and standard deviation for all χ angle(s). We only used side chain structures where all χ angle(s) were present. For example, H120 glutamine in 4eig:B is missing the χ_2 and χ_3 angles; thus, it was not used.

For position-independent rotamers with fewer than ten observations, we calculated the average and standard deviation based on all rotamers with the same χ angle bin. For example, there were only eight examples of the arginine $\{g-, g-, g+, t\}$ rotamer across all positions and structures in our non-redundant set. Thus, to calculate the average χ_1 angle of this rotamer, we used the average χ_1 angle of all observed arginine side chain structures with $\chi_1 = g-$. Similarly, to calculate the average χ_3 angle, we considered the χ_3 angles of all arginine side chain structures with $\chi_3 = g+$.

Density estimation of χ_1 angle modes

For each side chain type, we estimated the χ_1 angle density using a Gaussian mixture model (GMM). To estimate the density, we imposed a minimum number of 20 examples. For example, at IMGT position L109, 41 asparagine residues were observed in our non-redundant set; in contrast only eight lysine residues were observed at the same position. Thus the χ_1 angle mode was estimated for L109 asparagine but not for L109 lysine.

A GMM represents a weighted sum of K Gaussian distributions,

$$\rho(\chi_1) = \sum_{k=1}^K \lambda_k \mathcal{N}(\chi_1 | \mu_k, \sigma_k). \quad (1)$$

The k^{th} Gaussian distribution has a set of three parameters, $\{\lambda_k, \mu_k, \sigma_k\}$, corresponding to the weight, mean, and covariance, respectively. We represent the K -long vector of weights as Λ .

Depending on the values of λ_k , we determined the mode of the χ_1 angle of i, α , $\hat{\chi}_1(i, \alpha)$, as

$$\hat{\chi}_1(i, \alpha) = \begin{cases} \text{Unimodal} & \text{if } \max(\Lambda) \geq 0.8 \\ \text{Bimodal} & \text{if } \min(\Lambda) \leq 0.1 \text{ and } \max(\Lambda) < 0.8 \\ \text{Trimodal} & \text{otherwise.} \end{cases} \quad (2)$$

The number of components, K , was determined by calculating the Akaike Information Criterion for each GMM [44].

In total, there were 2383 observed side chain types. A χ_1 angle mode was calculated for 720 side chain types; for the remaining 1663 side chain types, no χ_1 angle mode was calculated as there were fewer than 20 observations.

In our training set, 113456 side chain structures were observed. The 720 side chain types with a χ_1 angle mode were based on 104999 observations, whereas the 1663 side chain types with no mode were based on 8457 observations. Thus, 92.5% of observed side chains were from a side chain type with a defined χ_1 angle mode.

Side chain prediction algorithm – PEARS

A flow chart of PEARS is given in Figure S1. We first describe the order in which we make our predictions, then the filters that are used for each step.

For all target structures, PEARS first detects and builds disulphide bridges. In antibodies, disulphide bridges are normally formed between H23–H104 and L23–L104. For each pair of cysteines, the disulphide score, BS , is calculated using an implementation of the SCWRL3 function (Equation 10).

Next, PEARS predicts IMGT positions with a unimodal side chain type (Figure 2). When predicting these positions, PEARS initially builds side chain structures with the same χ_1 angle bin. Positions with unimodal χ_1 side chain types are then sorted by the number of rotamers at each position [12], and the rotamer with the lowest self-energy (E_{self}) is fitted

John Wiley & Sons, Inc.

(Equation 3). If the fitted side chain structure has clashes, it is subjected to 200 rounds of Gaussian relaxation. If clashes still remain, the rotamer with the next-lowest E_{self} is used. If none of the predicted side chain structures can fit at a given position, PEARS repredicts these positions at a later stage, allowing variation in the χ_1 angle.

Once the unimodal positions have been predicted, the remaining positions are predicted in a similar manner to other side chain prediction methods [12, 22]. Rotamers are read from our position-dependent rotamer library in descending order of probability up to a cumulative probability of 99%. If there are fewer than 20 examples of a particular side chain type, the position-independent probability of a rotamer r is used.

A rotamer s at position i , s_i , is eliminated via DEE [45] if PEARS identifies another rotamer r at i , r_i , which satisfies Goldstein's criterion:

$$E_{self}(s_i) - E_{self}(r_i) + \sum_{j=1, j \neq i}^N \min_{r_j} \{E_{pair}(s_i, r_j) - E_{pair}(r_i, r_j)\} > 0.$$

If a position has only one rotamer after DEE, it is fixed. Otherwise a graph is constructed, where an edge is formed if the difference between the maximum and minimum pair energies (E_{pair} ; Equation 5) between rotamers at positions i and j is greater than 3 [22]. The graph is then solved by searching for biconnected components [46].

Calculation of self and pair energies

For each rotamer, PEARS calculates the self-energy (E_{self}) using our implementations of RASP's energy functions [22]. The self-energy of a rotamer r at i , $E_{self}(r_i)$, is calculated as

$$E_{self}(r_i) = E_{position}(r_i) + \sum_{j \in \text{fixed positions}} E_{pair}(r_i, r_j). \quad (3)$$

The term $E_{position}(r_i)$ is similar to the library term in SCWRL3; here, the probability term represents the probability that an amino acid α will have the rotamer r at IMGT position i .

$$E_{position}(r_i) = -Q \log \frac{Pr(r_i|i, \alpha)}{\sum Pr(r_i|i, \alpha)} + \sum_{|i-j| \geq 1} \sum_{a \in SC(i)} \sum_{b \in BB(j)} E(a, b) \quad (4)$$

The value of Q is fixed to 3. Here, $SC(i)$ and $BB(i)$ refer to the set of side chain and backbone atoms of the i th position, respectively. Thus, the second term in $E_{position}(r_i)$ represents the sum of energies between the side chain atoms of r_i with the backbone atoms of the j th position. Finally, for all j fixed positions, we calculate the sum of energies between the side chain atoms of r_i with the side chain atoms of r_j . For each pair of rotamers r_i and r_j in 'effective contact' (Equation 7), we calculate the pair energy, E_{pair} , as the sum of energies between the side chain atoms of r_i and r_j . In other words,

$$E_{pair}(r_i, r_j) = I(r_i, r_j) \sum_{a \in SC(i)} \sum_{b \in SC(j)} E(a, b). \quad (5)$$

where $I(r_i, r_j)$ represents an indicator function that is 1 if r_i and r_j are in effective contact. For any side chain–side chain and backbone–side chain contacts, the energy between two atoms a and b is the sum of the van der Waals and hydrogen bond terms (Equations 8, 9),

$$E(a, b) = E_{vdw}(a, b) + E_{hb}(a, b). \quad (6)$$

Calculation of effective contacts

Following the method in RASP, two rotamers at positions i and j are in effective contact if their $C\beta$ atoms are within the hemispheres formed by each side chain [22]. Thus,

$$I(r_i, r_j) = \begin{cases} 1 & \text{if } d(C\beta_i, C\beta_j) < t_i + t_j + 5\text{\AA} \\ & \text{and } \angle C\alpha_i C\beta_i C\beta_j > 90^\circ \text{ or } \angle C\alpha_j C\beta_j C\beta_i > 90^\circ \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Here, $d(C\beta_i, C\beta_j)$ represents the distance between the $C\beta$ atoms of r_i and r_j ; t_i and t_j are the hemispheric radii of the side chains at the i^{th} and j^{th} positions (Figure 1A). The hemisphere radius represents the distance between the $C\beta$ and the furthest atom in the side chain.

Calculation of van der Waals energy

The van der Waals energy between any two atoms a and b , $E_{vdw}(a, b)$, is calculated using our implementation of the van der Waals term in RASP, where,

$$E_{vdw}(a, b) = \begin{cases} 50\epsilon_{ab} & \text{if } d' < 0.465 \\ \epsilon_{ab}(80 - 64.5d') & \text{if } 0.465 \leq d' < 0.75 \\ 1.63\epsilon_{ab}[(\frac{1}{d'}^{12}) - 2(\frac{1}{d'}^6)] & \text{if } 0.75 \leq d' < 0.8929 \\ 0.99\epsilon_{ab}[(\frac{1}{d'}^{12}) - 2(\frac{1}{d'}^6)] & \text{if } 0.8929 \leq d' < 2.3 \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

d' represents the ratio of the distance between a and b with respect to the sum of their van der Waals radii; ϵ_{ab} represents the square-root of the well-depths of a and b from CHARMM36 [47], *i.e.*, $\epsilon_{ab} = \sqrt{\epsilon_a \epsilon_b}$. In PEARS, we use the following van

der Waals radii: 1.7Å (C), 1.55Å (N), 1.52Å (O), and 1.8Å (S).

Calculation of hydrogen bonds

Since PEARS does not predict the coordinates of hydrogen atoms, we use our implementation of the hydrogen bond function in RASP. However, we also calculate hydrogen bonds formed from amine and amide nitrogens. For histidine residues, we assume that only the N δ atom is a hydrogen donor. For the set of atoms shown in Figure 1B, the energy of the hydrogen bond E_{hb} is calculated as

$$E_{hb}(a, b) = -1.8 \sqrt{\frac{(\cos(\nu - 111.5^\circ) - \cos 37^\circ)(\cos(\tau - 120^\circ) - \cos 47^\circ)}{(1 - \cos 37^\circ)(1 - \cos 47^\circ)}} \quad (9)$$

where ν is the angle $\angle bDA$ and τ is the angle $\angle cAD$ in degrees. If the multiplicand under the square root is negative, we set E_{hb} to 0.

Calculation of disulphide bridges

To predict disulphide bridges between two cysteine side chains, we use our implementation of the SCWRL3 disulphide bridge scoring function [46], where

$$\begin{aligned} BS = & \frac{|d - 2\text{\AA}|}{0.05\text{\AA}} + \frac{|A_1 - 104^\circ|}{5^\circ} + \frac{|A_2 - 104^\circ|}{5^\circ} \\ & + \frac{\min\{||\theta_1| - 80^\circ|, ||\theta_1| - 180^\circ|\}}{10^\circ} \\ & + \frac{\min\{||\theta_2| - 80^\circ|, ||\theta_2| - 180^\circ|\}}{10^\circ} \\ & + \frac{||\theta_3| - 90^\circ|}{20^\circ} + \frac{E_{\text{self}}(r_1) + E_{\text{self}}(r_1)}{2}. \end{aligned} \quad (10)$$

- r_i and r_j represent the cysteine rotamers for the putative disulphide bridge at the i^{th} and j^{th} positions.
- d represents the distance between the two S γ atoms, S γ_i and S γ_j .
- A_1 and A_2 correspond to the C β_i - S γ_i - S γ_j and S γ_i - S γ_j - C β_j bond angles.
- $\theta_1, \theta_2, \theta_3$ angles correspond to the dihedral angles for C α_i - C β_i - S γ_i - S γ_j , S γ_i - S γ_j - C β_j - C α_j , and C β_i - S γ_i - S γ_j - C β_j , respectively.

Building side chain coordinates and Gaussian relaxation

Side chain coordinates are predicted using the natural extension reference frame method [48]. Each side chain structure is initially built using the rotamer's mean χ angle(s). If there are fewer than 10 examples of a particular rotamer, we use the

position-independent average χ values.

For example, the $\{t, g+\}$ rotamer of H65 isoleucine was only seen in 4bz2:HL. The side chain of this structure was predicted using the average χ angles of the position-independent $\{t, g+\}$ isoleucine rotamer (76 cases). If there were less than 10 cases of a position-independent rotamer (*e.g.* the position-independent $\{t, g-, g+\}$ methionine rotamer is only observed once), then its average χ angles were calculated based on all methionine rotamers with $\chi_1 = t$, $\chi_2 = g-$, and $\chi_3 = g+$.

Subsequently, if the constructed side chain structure clashes with any part of the target structure, the side chain is re-constructed by sampling χ angles from a Gaussian distribution with parameters from our rotamer library. The side chain's coordinates are adjusted until there are no clashes, or we reach 200 iterations, whichever occurs sooner. For every 50 sampling iterations, we double the standard deviation; if there are clashes after 200 iterations, we do not predict the side chain. Across our four test sets, 653 out of 260675 side chains were not predicted; these were considered to be χ_1 and χ_{1+2} incorrect.

Benchmarking accuracy and clash detection

A side chain prediction was considered to be correct if it was within 40° of the native side chain structure (Table S1; [12]). The χ_1 accuracy represents the fraction of predictions with a correct χ_1 angle; the χ_{1+2} accuracy represents the fraction of predictions where both χ_1 and χ_2 angles were correct. Positions without a side chain prediction were considered to be incorrect.

PEARS was benchmarked against three other side chain predictors: SCWRL, SIDEpro, and RASP. SCWRL (version 4) [12] was used with default parameters. SCWRL samples from a backbone-dependent library [26], using an energy function that combines van der Waals and hydrogen bond terms. The solution is determined by DEE and graph decomposition. SIDEpro [24] uses a backbone-dependent rotamer library [49] and an artificial neural network to predict the side chains. RASP [22] uses a backbone-dependent rotamer library [49], and a combination of DEE and a Monte Carlo search algorithm to determine the solution. If a structure is missing a backbone atom, RASP does not make any predictions. In total, RASP did not generate any predictions for 35 structures in the crystal set, 30 structures in the model set, 10 structures in the blind crystal set, and four structures in the blind model set. For RASP, the reported average accuracies were calculated for targets with side chain predictions.

Clashes in the side chain predictions were detected using a KD-tree algorithm. Two atoms were considered to clash if they were separated by less than 63% of the sum of their van der Waals radii, similar to previously established cutoffs [22, 24].

Identification of regions

CDR and framework positions were identified using the definitions from [50]. A position was classified to be at the V_H – V_L interface if it had at least one heavy atom within 5 Å of the opposing chain, and the position was observed in over 90% of paired antibodies in the training set (485/539 antibodies). A position was classified as ‘buried’ if it had a relative solvent

accessibility of $\leq 10\%$. For positions that satisfied multiple criteria, *e.g.* H44 (framework and V_H – V_L interface position), we used its accuracy for both regions.

Results

An antibody-specific rotamer library

Our antibody-specific, position-dependent rotamer library was built using 639 non-redundant antibody structures downloaded from SAbDab [33]. This is a smaller dataset than other rotamer libraries. For instance, the backbone-dependent rotamer library by Shapovalov *et al.* is based on 3974 protein chains, whereas our library has 1178. However, our method bins rotamers by IMGT positions rather than the ϕ/ψ angles, meaning this smaller number of observations should still be an ample training set. In the backbone-dependent rotamer library of Shapovalov *et al.*, $10^\circ \times 10^\circ$ bins were used [26], leading to a total of 23328 amino acid/ ϕ/ψ combinations. Across all structures in our training set, we recorded 286 different IMGT positions including insertions [37], meaning that only 5184 different side chain types are possible. Only 2383 are observed, reflecting the sequence conservation of antibodies.

For 339 side chain types, the χ_1 angle is classified to have a unimodal distribution under our framework (Equation 2). These unimodal side chain types (coloured blue; Figure 2) are generally found in the framework region, and are often associated with known conserved amino acids [51]. For example, the highly-conserved glutamine residues at H44 and L44 show unimodal χ_1 angle profiles, with both positions adopting the $\chi_1 = t$ configuration. This phenomenon is also observed in some positions of the CDR loops, such as L116 tyrosine (Figure 2). Furthermore, all of the 18 amino acids had at least one IMGT position in which their χ_1 angle distribution was unimodal. Among the 113456 observed side chain structures, 55.1% (62528 observations; coloured blue in Figure 2) of the data has a unimodal χ_1 angle, while 22.1% of the data (25074 observations; coloured red in Figure 2) has a bimodal χ_1 , and 15.3% is trimodal (coloured green in Figure 2).

One thousand six hundred and sixty-three side chain types do not have a defined χ_1 mode, as there was less than 20 observed side chain structures in our training set (coloured grey in Figure 2). However, these side chain types only account for 7.5% (8457 side chain structures) of the data in our dataset. The remaining side chain types (coloured white in Figure 2) are not observed in our non-redundant set of 639 structures.

To compare our rotamer library with the backbone-dependent rotamer library from Shapovalov *et al.* [26], we first identified the ten most common ϕ/ψ bins for each amino acid type in our non-redundant set. For the rotamers in these dihedral bins, we found that their probabilities were fairly well correlated ($r = 0.78$). At the other end of the spectrum, rotamers that were rare ($< 1\%$ of the data) in the backbone-dependent library [26] also appear to be rare in our library. There are examples that were considered rare in the backbone-dependent rotamer library, while they were not rare in our position-dependent library, *e.g.* the glutamine $\{g+, g+, Ot\}$ rotamer (Figure S2). Likewise, there were also some rotamers that were considered rare in

our position-dependent rotamer library, but not rare in the backbone-dependent rotamer library [26], such as the tryptophan $\{t, g+\}$ rotamer.

Benchmarking PEARS

Previously, side chain predictors have been assessed on non-redundant sets of general proteins from the PDB [12, 24, 31]. Only two of the 379 structures in SCWRL's test set were antibodies [12], while the dataset in Peterson *et al.*'s study had none [31]. Here we carry out an antibody-specific evaluation of side chain predictors using four separate test sets (see Methods), comparing our method, PEARS, to three other side chain predictors: SCWRL, RASP, and SIDEpro [12, 22, 24]. All four methods required less than 10 seconds to predict the side chains for an input structure. On average, PEARS required 7.20 seconds per structure, while SCWRL, RASP, and SIDEpro required 5.41, 0.13, and 0.87 seconds per structure, respectively.

Predictions on crystal structures

We initially tested the four methods on two separate sets of crystal structures: the 'crystal set', which is the same set of 639 antibody structures that were used for building our rotamer library, and the 'blind crystal set', a set of 95 antibody structures that was released between 27 Jan 2016 and 23 Mar 2017 (see Methods). Since the rotamer library encodes rotamers by the average χ angle (or the position-independent average χ angle where there are fewer than ten observed side chain structures for a particular rotamer), there should be little bias, even on the 639 structures used to build our rotamer library.

All four methods showed comparable performance (Table S2). SIDEpro had the highest average χ_1 accuracy for the crystal set (83.2%), whereas PEARS had the lowest average χ_1 accuracy (81.1%). On the other hand, PEARS had the highest average χ_{1+2} accuracy (66.6%) while RASP had the lowest average χ_{1+2} accuracy (62.5%). Likewise, for the blind crystal set, SIDEpro had the best χ_1 accuracy (85.8%) while PEARS had the lowest average χ_1 accuracy (81.7%). Again, PEARS had the highest average χ_{1+2} accuracy for the blind crystal set (69.5%), with RASP having the lowest average χ_{1+2} accuracy (66.5%; Figures 3A, 3B).

Figure 4A shows that PEARS was more accurate on unimodal positions of the blind crystal set, but less on other types, particularly those without a defined χ_1 angle mode. However, PEARS had a higher overall χ_{1+2} accuracy as it produced the correct χ_2 angle prediction for positions with a unimodal χ_1 angle (Figure 4B). For instance, glutamine at L106 is a unimodal side chain type; for 5d96:CB, PEARS produced a χ_{1+2} correct rotamer, while SCWRL was only χ_1 correct (Figure 5A).

Predictions on model structures

We used ABodyBuilder to generate a model structure for each of the 639 antibodies in the crystal set and the 95 antibodies in the blind crystal set (see Methods). Here, unlike the crystal set, the backbone may be incorrect. Building side chains on models is a more realistic test of the methods, especially in the context of computational antibody design [7].

On model structures, PEARS outperformed the other methods (Table S3). PEARS had the highest average χ_1 and χ_{1+2} accuracy (78.1% and 62.5%) while RASP had the lowest average χ_1 and χ_{1+2} accuracies of 72.7% and 52.4%, respectively. For the blind model set, PEARS' average χ_1 and χ_{1+2} accuracies were 78.7% and 64.8% (Figures 3C, 3D); RASP had the lowest χ_1 and χ_{1+2} accuracies (73.9% and 53.7%). Similar to the blind crystal set, PEARS produced more accurate side chain structures for positions with a defined χ_1 mode (Figures 4C, 4D). We also compared χ_1 accuracy for each antibody region, *e.g.* buried residues and those in the CDRs (see Methods). PEARS had $\sim 5\%$ higher χ_1 accuracy across all regions of the antibody model in comparison to other methods (Figure S3).

These results demonstrate the benefits of a position-dependent approach: rotamer sampling in PEARS is more robust to minor inaccuracies in the ϕ/ψ angles. In fact, in model structures, PEARS outperforms the other methods in χ_1 and χ_{1+2} prediction for unimodal, bimodal and trimodal positions, which accounts for 91.1% of the data (Figures 4C, D). In particular, unimodal side chain types were predicted well. Figure 5B shows an example where a side chain of a model structure (H20 lysine), which is a unimodal side chain type, is only correctly predicted by PEARS for both the χ_1 and χ_2 angles. The accuracy of side chain prediction was weakly correlated with the quality of the models. On a global level, χ_1 accuracy and side chain RMSD were not related to the backbone RMSD of the model structure (Figures S5, S6). However, if local backbone RMSD is considered, the χ_1 accuracy for SCWRL, RASP, and SIDEpro decreased as soon as any deviation occurred ($>0.5\text{\AA}$ RMSD); beyond 1.5\AA RMSD, all methods showed a decrease in χ_1 accuracy (Figure S7).

The restriction of the χ_1 sampling appears to have a minor negative effect in predicting the side chains of crystal structures. However, when considering bimodal and trimodal side chain types, they lead to improved predictions for model structures as it is agnostic to deviations in the ϕ/ψ . To demonstrate the influence of our position-dependent rotamer library, we tested PEARS with the backbone-dependent rotamer library from Shapovalov *et al.* [26]. Using the backbone-dependent rotamer library, PEARS had a weaker performance on model structures (Figure S8).

PEARS generates the fewest side chain–side chain clashes.

A key feature of PEARS is its clash relaxation strategy. If a predicted side chain structure clashes with an existing side chain in the structure, PEARS introduces Gaussian noise into each χ angle of the rotamer using parameters from our rotamer library. If a suitable prediction cannot be made, PEARS only places a $C\beta$ atom, rather than forcing a prediction. Two atoms were considered to clash if they were separated by less than 63% of the sum of their van der Waals radii, similar to previously tested cutoffs [22, 24]. To test the four methods for clashes, we only compared positions that were predicted by all four methods in the blind model set. Thus, we omitted 112 side chain structures (0.7% of the data) from the clash calculations. PEARS consistently had the lowest number of side chain–side chain and side chain–backbone clashes (Figure 6). PEARS had an average of 0.5 clashes while the next-best performer, SIDEpro, had an average of 3.5 clashes (Table S4).

Discussion

In this manuscript, we present our position-dependent, antibody-specific rotamer library. The library was built using a non-redundant set of 639 antibodies from SAbDab [33]. We show that a side chain's χ_1 angle is dependent on its IMGT position. This pattern was not restricted to a particular amino acid type, nor position. Of the 5184 possible side chain types, 2801 are never observed in our dataset. For another 1663 side chain types, there were fewer than 20 observations, meaning that no attempt could be made to determine if the χ_1 distribution was uni or multimodal. The gaps in the data could be due to the inherent biases in antibody sequences (*e.g.* cysteine residues almost always at H23/H104/L23/L104), or more simply, a lack of side chain structural data. However, as more data becomes available, the gaps in the data (*e.g.* some of the positions in the CDR loops) will disappear.

Next, using this IMGT position-based rotamer library, we tested our prediction algorithm, PEARS. On two sets of crystal structures, all of the methods performed similarly, with PEARS showing the lowest χ_1 accuracy, and SIDEpro the highest. PEARS had the highest average χ_{1+2} accuracy, while RASP had the weakest performance. Over half the positions in both sets of crystal structures are unimodal χ_1 side chain types, and here PEARS has the highest accuracy for both χ_1 and χ_2 angles. However, for the remaining positions, particularly those with little data (*i.e.* side chain types considered as having 'no mode' in our framework), PEARS produced the least accurate χ_1 predictions. A possible route to improve the accuracy for PEARS on crystal structures would be the development of a hybrid approach, where PEARS is used to predict positions with unimodal χ_1 angle distributions, and backbone-dependent methods elsewhere, especially in positions where there are no known χ_1 angle modes.

On two sets of model structures, PEARS had the highest average χ_1 and χ_{1+2} accuracies. PEARS gave the most accurate predictions for side chain types with unimodal or multimodal χ_1 angles. PEARS also generated the fewest clashes in the final model structure, making it possible to use the PEARS predictions for subsequent computational analyses, such as antibody-antigen docking. As observed in the crystal set, backbone-dependent methods outperformed PEARS on side chain types with no χ_1 mode, which further reinforces the potential benefits of a hybrid side chain prediction method. Our results on models suggest that position-dependent information is useful for cases where there are uncertainties in the backbone geometry.

Our position-dependent rotamer library and side chain prediction algorithm, PEARS, provide a new approach to the current paradigm of side chain prediction. PEARS builds accurate side chain structures in model structures with almost zero clashes, making it particularly applicable for antibody design or blind test scenarios. The concept of PEARS is potentially generalisable, provided that there is a sufficient number of structures to define the relative topology; possible examples include T-cell receptors and G-protein coupled receptors. Recovering the correct side chains will lead on to more accurate results in subsequent applications, such as humanisation and antibody-antigen docking. PEARS is freely available as a web application at <http://opig.stats.ox.ac.uk/webapps/pears>.

Acknowledgments

J.L. and C.M.D. have been funded by the Engineering Physical Sciences Research Council and the Medical Research Council (EPSRC and MRC; grant numbers EP/G037280/1 and EP/L016044/1). This work has also received additional funding from UCB Pharma and Roche Diagnostics GmbH.

Conflict of interest

None declared.

References

- [1] REICHERT, J. M. 2017. Antibodies to watch in 2017. *mAbs* 9:167–181.
- [2] CHOI, Y., HUA, C., SENTMAN, C. L., ACKERMAN, M. E., AND BAILEY-KELLOGG, C. 2015. Antibody humanization by structure-based computational protein design. *mAbs* 7:1045–1057.
- [3] LEWIS, S. M., WU, X., PUSTILNIK, A., SERENO, A., HUANG, F., RICK, H. L., GUNTAS, G., LEAVER-FAY, A., SMITH, E. M., HO, C., HANSEN-ESTRUCH, C., CHAMBERLAIN, A. K., TRUHLAR, S. M., CONNER, E. M., ATWELL, S., KUHLMAN, B., AND DEMAREST, S. J. 2014. Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nat Biotechnol* 32:191–198.
- [4] LIPPOW, S. M., WITTRUP, K. D., AND TIDOR, B. 2007. Computational design of antibody–affinity improvement beyond in vivo maturation. *Nat Biotechnol* 25:1171–1176.
- [5] PANTAZES, R. J. AND MARANAS, C. D. 2010. OptCDR: A general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng Des Sel* 23:849–858.
- [6] KHOURY, G. A., SMADBECK, J., KIESLICH, C. A., AND FLOUDAS, C. A. 2014. Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol* 32:99–109.
- [7] KURODA, D., SHIRAI, H., JACOBSON, M. P., AND NAKAMURA, H. 2012. Computer-aided antibody design. *Protein Eng Des Sel* 25:507–521.
- [8] ALMAGRO, J. C., TEPLYAKOV, A., LUO, J., SWEET, R. W., KODANGATTIL, S., HERNANDEZ-GUZMAN, F., AND GILLILAND, G. L. 2014. Second Antibody Modeling Assessment (AMA-II). *Proteins* 82:1553–1562.
- [9] LEEM, J., DUNBAR, J., GEORGES, G., SHI, J., AND DEANE, C. M. 2016. ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation. *mAbs* 8:1259–1268.
- [10] ALFORD, R., LEAVER-FAY, A., JELIAZKOV, J., O'MEARA, M., DiMALO, F., PARK, H., SHAPOVALOV, M., RENFREW, P., MULLIGAN, V., KAPPEL, K., LABONTE, J., PACELLA, M., BONNEAU, R., BRADLEY, P., DUNBRACK, R., DAS, R., BAKER, D., KUHLMAN, B., KORTTEMME, T., AND GRAY, J. 2017. The rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* 13:3031–3048.
- [11] VREVEN, T., HWANG, H., PIERCE, B., AND WENG, Z. 2012. Prediction of protein–protein binding free energies. *Prot Sci* 21:396–404.
- [12] KRIVOV, G. G., SHAPOVALOV, M. V., AND DUNBRACK, R. L. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77:778–795.
- [13] WILLIS, J. R., BRINEY, B. S., DELUCA, S. L., CROWE, J. E., AND MEILER, J. 2013. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. *PLOS Comput Biol* 9:e1003045.

- [14] FERA, D., SCHMIDT, A. G., HAYNES, B. F., GAO, F., LIAO, H.-X., KEPLER, T. B., AND HARRISON, S. C. 2014. Affinity maturation in an HIV broadly neutralizing B-cell lineage through reorientation of variable domains. *Proc Natl Acad Sci USA* 111:10275–10280.
- [15] KURODA, D., SHIRAI, H., KOBORI, M., AND NAKAMURA, H. 2008. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* 73:608–620.
- [16] KOENIG, P., LEE, C. V., WALTERS, B. T., JANAKIRAMAN, V., STINSON, J., PATAPOFF, T. W., AND FUH, G. 2017. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc Natl Acad Sci USA* 114:E486–E495.
- [17] COLBES, J., CORONA, R. I., LEZCANO, C., RODRÍGUEZ, D., AND BRIZUELA, C. A. 2016. Protein side-chain packing problem: is there still room for improvement? *Brief Bioinform*.
- [18] GAILLARD, T., PANEL, N., AND SIMONSON, T. 2016. Protein side chain conformation predictions with an MMGBSA energy function. *Proteins* 84:803–819.
- [19] HARDER, T., BOOMSMA, W., PALUSZEWSKI, M., FRELLSEN, J., JOHANSSON, K. E., AND HAMELRYCK, T. 2010. Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics* 11:306.
- [20] TOWSE, C.-L., RYSAVY, S., VULOVIC, I., AND DAGGETT, V. 2016. New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities. *Structure* 24:187–199.
- [21] DUNBRACK, R. L. AND KARPLUS, M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Mol Biol* 1:334–340.
- [22] MIAO, Z., CAO, Y., AND JIANG, T. 2011. RASP: rapid modeling of protein side chain conformations. *Bioinformatics* 27:3117–3122.
- [23] LIANG, S., ZHOU, Y., GRISHIN, N., AND STANDLEY, D. M. 2011. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. *J Comput Chem* 32:1680–1686.
- [24] NAGATA, K., RANDALL, A., AND BALDI, P. 2012. SIDEpro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins* 80:142–153.
- [25] TUFFERY, P., ETCHEBEST, C., HAZOUT, S., AND LAVERY, R. 1991. A New Approach to the Rapid Determination of Protein Side Chain Conformations. *J Biomol Struct Dyn* 8:1267–1289.
- [26] SHAPOVALOV, M. V. AND DUNBRACK, R. L. 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19:844–858.
- [27] LOVELL, S. C., WORD, J. M., RICHARDSON, J. S., AND RICHARDSON, D. C. 2000. The penultimate rotamer library. *Proteins* 40:389–408.
- [28] CHINEA, G., PADRON, G., HOOFT, R. W. W., SANDER, C., AND VRIEND, G. 1995. The use of position-specific rotamers in model building by homology. *Proteins* 23:415–421.
- [29] DUNBRACK, R. L. AND KARPLUS, M. 1993. Backbone-dependent Rotamer Library for Proteins Application to Side-chain Prediction. *J Mol Biol* 230:543–574.
- [30] LU, M., DOUSIS, A. D., AND MA, J. 2008. OPUS-Rota: A fast and accurate method for side-chain modeling. *Prot Sci* 17:1576–1585.
- [31] PETERSON, L. X., KANG, X., AND KIHARA, D. 2014. Assessment of protein side-chain conformation prediction methods in different residue environments. *Proteins* 82:1971–1984.
- [32] ROSS, G. A., MORRIS, G. M., AND BIGGIN, P. C. 2013. One size does not fit all: The limits of structure-based models in drug discovery. *J Chem Theory Comput* 9:4266–4274.
- [33] DUNBAR, J., KRAWCZYK, K., LEEM, J., BAKER, T., FUCHS, A., GEORGES, G., SHI, J., AND DEANE, C. M. 2014. SAbDab: the structural antibody database. *Nucleic Acids Res* 42:D1140–1146.

- [34] SCHROEDER, H. W. AND CAVACINI, L. 2010. Structure and function of immunoglobulins. *J Allergy Clin Immunol* 125:41–52.
- [35] CHOTHIA, C. AND LESK, A. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196:901–917.
- [36] KABAT, E. A., WU, T. T., BILOFSKY, H., REID-MILLER, M., AND PERRY, H. M. 1983. Sequences of proteins of immunological interest. National Institutes of Health, 3rd edition.
- [37] LEFRANC, M.-P., POMMIÉ, C., RUIZ, M., GIUDICELLI, V., FOULQUIER, E., TRUONG, L., THOUVENIN-CONTET, V., AND LEFRANC, G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 27:55–77.
- [38] KABAT, E., WU, T., PERRY, H., GOTTESMAN, K., AND KOELER, C. 1991. Sequences of proteins of immunological interest. U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, 5th edition.
- [39] SAERENS, D., CONRATH, K., GOVAERT, J., AND MUYLDERMANS, S. 2008. Disulfide Bond Introduction for General Stabilization of Immunoglobulin Heavy-Chain Variable Domains. *J Mol Biol* 377:478–488.
- [40] LI, W. AND GODZIK, A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
- [41] CHOI, Y. AND DEANE, C. M. 2010. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins* 78:1431–1440.
- [42] ŠALI, A. AND BLUNDELL, T. L. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol* 234:779–815.
- [43] DUNBAR, J. AND DEANE, C. M. 2016. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* 32:298–300.
- [44] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer New York Inc, New York, NY, USA, 2nd edition.
- [45] DESMET, J., MAEYER, M. D., HAZES, B., AND LASTERS, I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539–542.
- [46] CANUTESCU, A. A., SHELENKOV, A. A., AND DUNBRACK, R. L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Prot Sci* 12:2001–2014.
- [47] HUANG, J. AND MACKERELL, A. D. 2013. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 34:2135–2145.
- [48] PARSONS, J., HOLMES, J. B., ROJAS, J. M., TSAI, J., AND STRAUSS, C. E. M. 2005. Practical conversion from torsion space to Cartesian space for in silico protein synthesis. *J Comput Chem* 26:1063–1068.
- [49] DUNBRACK, R. L. AND COHEN, F. E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Prot Sci* 6:1661–1681.
- [50] NORTH, B., LEHMANN, A., AND DUNBRACK, R. L. 2011. A new clustering of antibody CDR loop conformations. *J Mol Biol* 406:228–256.
- [51] IGAWA, T., TSUNODA, H., KIKUCHI, Y., YOSHIDA, M., TANAKA, M., KOGA, A., SEKIMORI, Y., ORITA, T., ASO, Y., HATTORI, K., AND TSUCHIYA, M. 2010. VH/VL interface engineering to promote selective expression and inhibit conformational isomerization of thrombopoietin receptor agonist single-chain diabody. *Protein Eng Des Sel* 23:667–677.

Figure legends

Figure 1. Calculation of effective contacts and hydrogen bonds.

A. An effective contact is determined using the same criteria as RASP [22]. B. Hydrogen bonds are calculated between donors (D; amide, amine nitrogens and hydroxyl oxygens) and acceptors (A; carboxyl oxygens and carbonyl oxygens). Adapted from [22].

Figure 2. Rotamer preferences of amino acids are IMGT position–dependent.

For each residue type at a given position, we determined the modality of the χ_1 angle distribution using a GMM. Amino acids are coloured blue if they have a unimodal χ_1 angle distribution (*e.g.* H40 asparagine), red if they have a bimodal χ_1 angle distribution (*e.g.* L1 aspartate), or green if they have a trimodal χ_1 angle distribution (*e.g.* H107 tyrosine). Amino acids are coloured grey if they have <20 data points ('no mode'; *e.g.* L55 arginine), and white if they are never observed (*e.g.* L69 valine). For simplicity, positions with insertions (*e.g.* H111A) are not visualised; in terms of prediction, these positions are treated in the same manner as any other IMGT position. A. Heavy chain positions; B. Light chain positions.

Figure 3. Boxplots of χ_1 and χ_{1+2} accuracy on the blind crystal and blind model sets.

A side chain prediction was evaluated using a 40° cutoff (Table S1). PEARS did not predict one side chain structure in the blind crystal set, and 112 side chain structures were not predicted in the blind model set. For these cases, PEARS was considered to have an incorrect prediction. RASP did not produce a prediction for 10 structures in the blind crystal set and four structures in the blind model set.

Figure 4. χ_1 and χ_{1+2} accuracy on the blind crystal and blind model sets, split by the χ_1 mode.

Unimodal side chain types account for 54.1% of the target positions in the blind crystal set, whereas bimodal side chain types account for 21.9%, trimodal side chain types 15.1%, and the remaining 8.9% have no mode.

Figure 5. Unimodal positions are predicted with greater accuracy by PEARS.

A. Glutamine residues at L106 have a unimodal χ_1 angle profile, with a peak for the $\chi_1 = g+$ rotamer. PEARS is good at predicting the χ_1 angle of these side chain types. In 5d96:CB, a structure of our blind crystal set, both PEARS and SCWRL produced accurate χ_1 predictions, but only PEARS was χ_{1+2} correct. B. Lysine residues at H20 have a unimodal χ_1 angle profile at $\chi_1 = t$. For 5epm:AB, a structure in the blind crystal set, we generated a model structure using ABodyBuilder (backbone RMSD 1.3Å), and predicted its side chains. Only PEARS generated χ_{1+2} correct predictions.

Figure 6. Clashes in the blind model set (95 structures).

Two atoms were considered to clash if they were separated by less than 63% of the sum of their van der Waals radii. The clash calculations consider side chain–side chain and side chain–backbone clashes. Clashes were only calculated across positions that were predicted by all four methods.

Figures

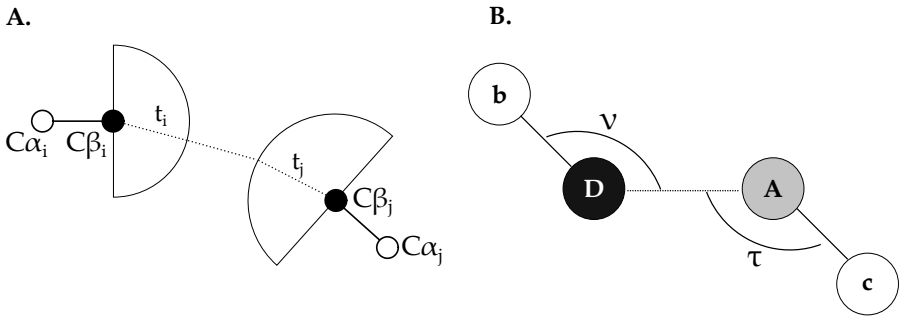


Figure 1

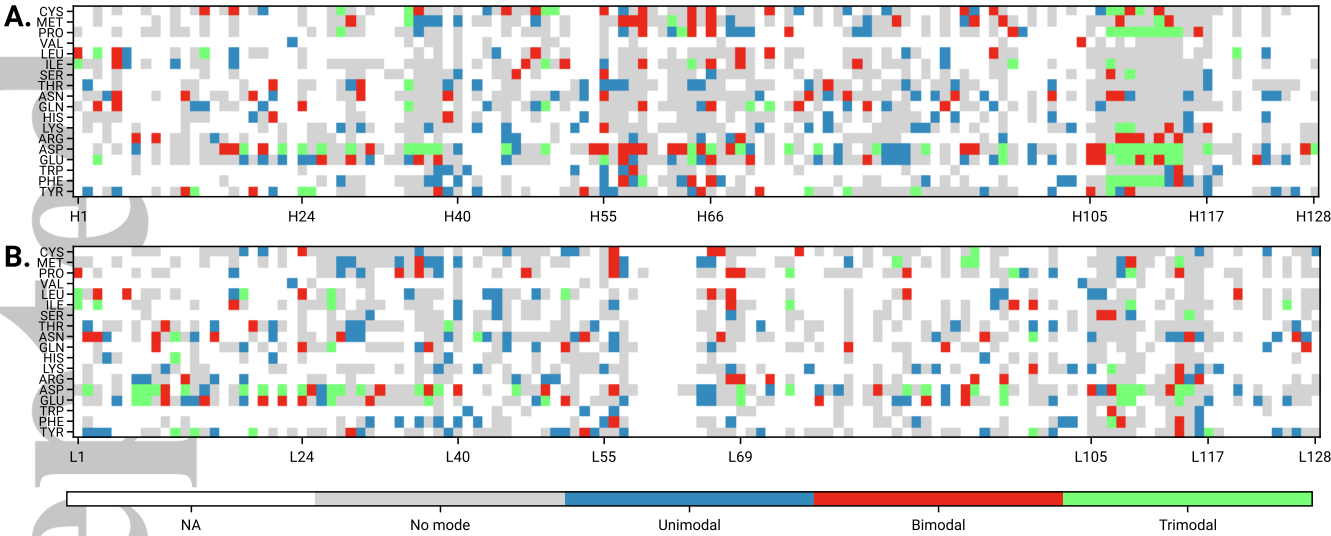


Figure 2

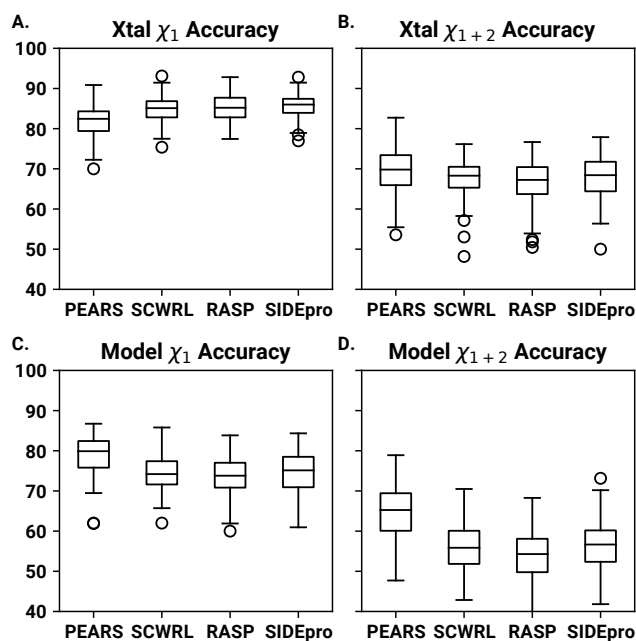


Figure 3

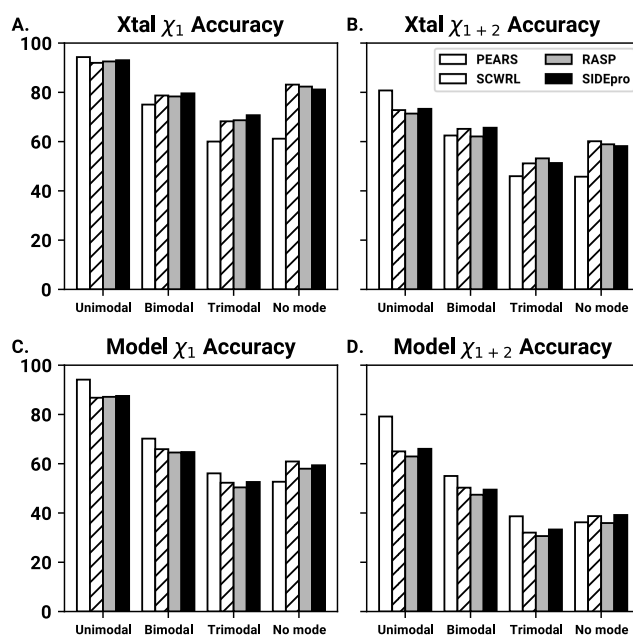


Figure 4

Accepted Article

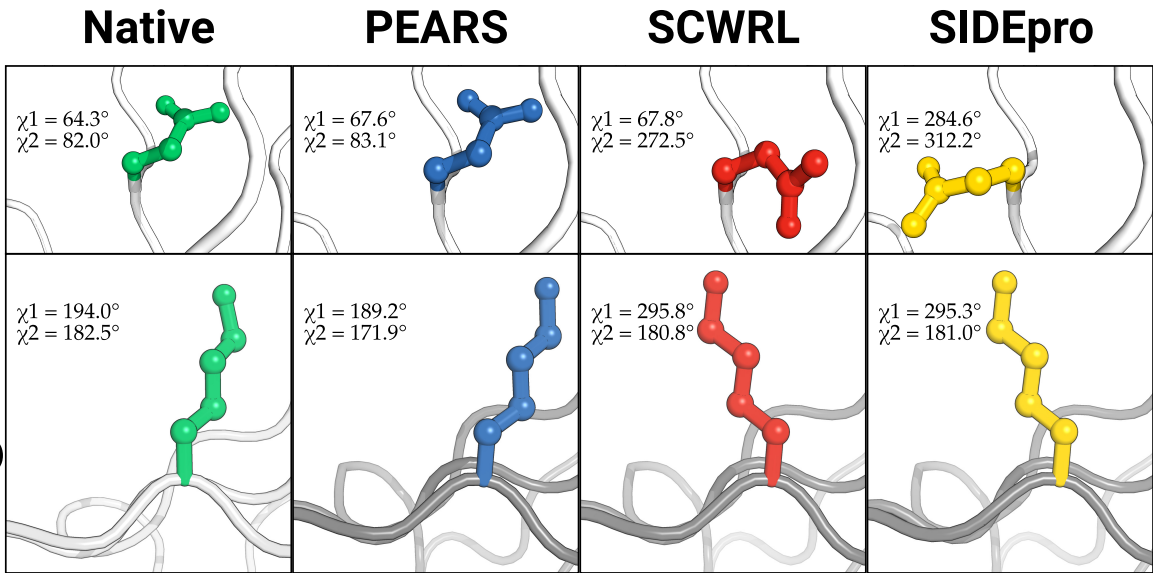


Figure 5

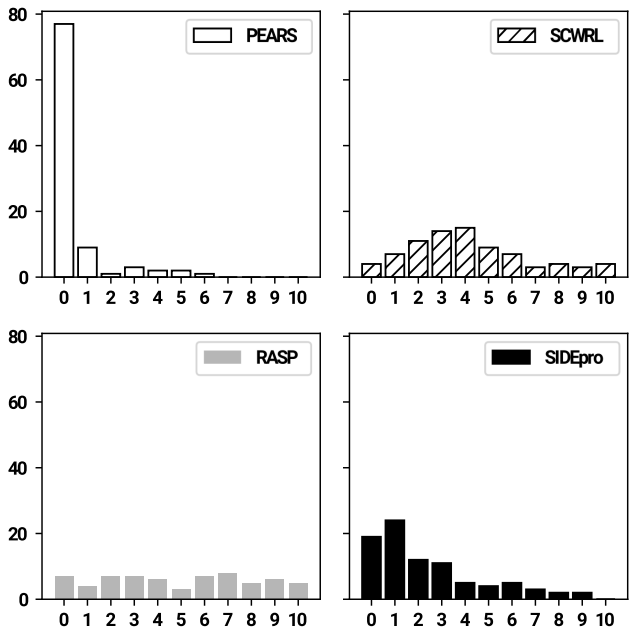


Figure 6