# Prediction of H-Bond Rotations from Protein H-Bond Topology

Andersen, Jørgen Ellegaard; Koyanagi, Yuki; Nielsen, Jakob Toudahl; Villemoes, Rasmus

Go to publication entry in University of Southern Denmark's Research Portal

# Prediction of H-Bond Rotations from Protein H-Bond Topology

Jørgen Ellegaard Andersen[1, 2], Yuki Koyanagi[2], Jakob Toudahl Nielsen[3, 4], and Rasmus Villemoes[5]

[1]*Danish Institute for Advanced Study, University of Southern Denmark*
[2]*Centre for Quantum Mathematics, Department of Mathematics and Computer Science, University of Southern Denmark*
[3]*Interdisciplinary Nanoscience Center (iNANO), Aarhus University*
[4]*Department of Chemistry, Aarhus University*
[5]*Prevas A/S*

## Abstract

H-bonds are known to play an important role in the folding of proteins into three-dimensional structures, which in turn determine their diverse functions. The conformations around H-bonds are important, in that they can be non-local along the backbone and are therefore not captured by the methods such as Ramachandran plots. We study the relationship between the geometry of H-bonds in proteins, expressed as a spatial rotation between the two bonded peptide units, and their topology, expressed as a subgraph of the protein fatgraph. We describe two experiments to predict H-bond rotations from their corresponding subgraphs. The first method is based on sequence alignment between sequences of the signed lengths of H-bonds measured along the backbone. The second method is based on finding an exact match between the descriptions of subgraphs around H-bonds. We find that 88.14% of the predictions lie inside the ball, centred around the true rotation, occupying just 1% of the volume of the rotation space SO(3).

## 1 Introduction

A protein is a biological molecule consisting of a linear polymer of amino acids. Different proteins fold into unique three-dimensional structures, also called native conformations, and it is widely recognised that the function of a protein is highly dependent on the three-dimensional structure of its native folded state. A classical way to describe the backbone conformation of proteins is the Ramachandran plots, which plots the dihedral angles $(\varphi, \psi)$ before and after each $C^\alpha$ atoms in two-dimensional distributions [26]. With the increase in the amount of the available protein structural data, the plots have been updated and extended to be used in structural validation [13, 27] and a number of other purposes (see, for example, [7] for a review). A natural extension of Ramachandran plots may be to amalgamate more than one pair of conformation angles to describe the backbone conformation at a larger scale. Levitt [12] proposed a descriptor that combines two consecutive pairs of conformation angles, while Carugo and Djinović-Carugo [8] used an average over the dihedral angles $\varphi$ and $\psi$ to characterise entire proteins. Other methods that do not use dihedral angles, include "curvature" and "torsion" of backbones analogous to the notion of curvature and torsion in differential geometry, computed from the coordinates of $C^\alpha$ atoms [24, 25], the writhe and the average crossing number in knot theory [29, 28], projection of nearby atoms to a small sphere centred at each $C^\alpha$ atoms [19], and the coordinates of atoms in two consecutive peptide units [23].

A common feature for the above methods is that they describe conformation of a backbone by moving along it from one end to the other. One of the main mechanisms determining and stabilising the folded structure is H-bonds [6, 30, 18], whose geometry has also been studied
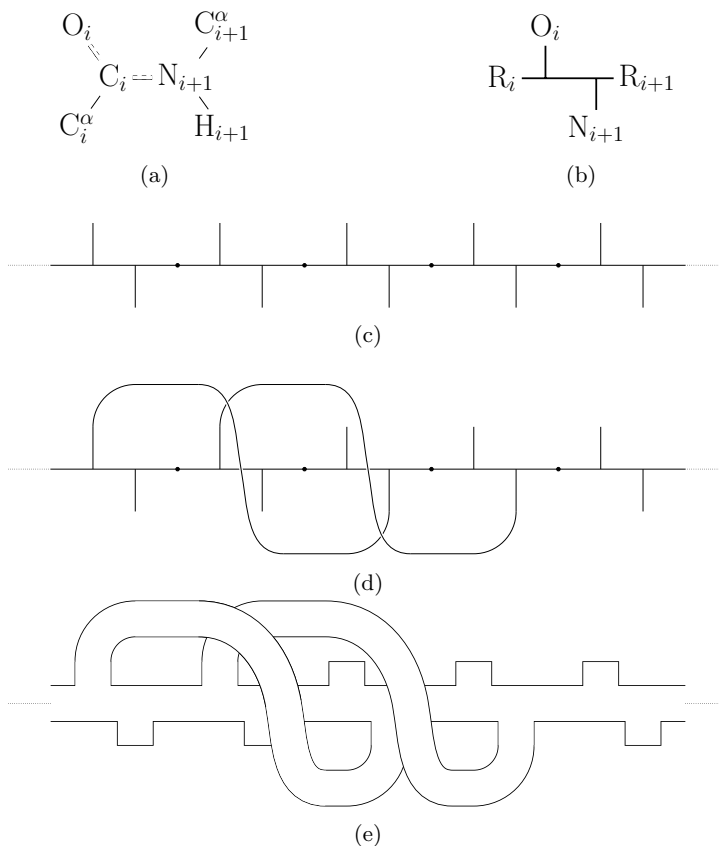
1

Figure 1: Construction of the protein model. Each peptide unit (a) is represented by a building block with the donor nitrogen and the acceptor oxygen half-edges drawn below and above the backbone, respectively (b). These building blocks are concatenated at so-called $\alpha$-carbon linkages to form a model of the backbone (c). The H-bonds are drawn as edges between corresponding donor and acceptor half-edges (d). One can associate a surface to a given fatgraph (e).

in relation to the structural analysis of proteins [9, 14]. Spatial rotations were introduced as a systematic three-dimensional descriptor of H-bond geometry in [20], and were found to correspond well to the concrete secondary structures and other local structural motifs [20]. If we ignore the geometry of H-bonds and instead concentrate on their topology, we obtain a graph, with backbone atoms as vertices and the covalent bonds and H-bonds as edges. In [22], an extension to this structure, called a *fatgraph*, was used to study protein structures. A fatgraph can be thought of as a surface obtained by "fattening" the underlying graph; i.e. by expanding the vertices to discs, and edges to "ribbons" connecting these discs (Figure 1e). In this paper, we utilise the construction of protein fatgraphs described in

[22], with some minor simplifications. The construction is summarised in Figure 1. The model is based on a representation of peptide units as fatgraph "building blocks" (Figure 1b), which are then glued together to represent a backbone (Figure 1c). Finally, the H-bonds are added as edges between the donor and the acceptor half-edges (Figure 1d).

The correspondence between fatgraphs and surfaces allows us to compute topological invariants, such as genus and the number of boundary components, for protein structures. In [21] the relation between these topological invariants and the protein domains was discussed. In addition, the planarity of peptide units allows us to assign an element of the rotation group SO(3) to each H-bond, as described in [20] (See Supplementary Material

for a detailed description). In [20], it was discovered that the H-bond rotations computed from PDB [5] are concentrated around well-defined clusters, and that these clusters correlated with some well-known local structures.

In this paper, we study the relation between topology and geometry of proteins around each H-bond. The study forms a part of an effort to solve the problem of predicting proteins' geometric structure from their primary structures, also called the protein folding problem. Our intended approach is in two steps, and relies on the fatgraph structure as an intermediate step. In the first stage a fatgraph structure is predicted from a primary sequence, and in the second stage a geometric structure is predicted from a protein fatgraph. Progress is being made on the different aspect of the programme, including the enumeration of possible protein fatgraph structures [1], and the study of relationship between protein fatgraphs and geometric structures [3, 2].

The current study is a contribution to the second stage, but instead of predicting the entire geometric structure from a protein fatgraph, we aim to predict local geometry of proteins, expressed as spatial rotations along the H-bonds, using fatgraph structures around the H-bonds. We call these local fatgraph structures *H-bond local patterns*, or simply *H-bond patterns*. For a given H-bond $a$, the *H-bond local pattern* or simply the *H-bond pattern* of *window size* $w$ around $a$ is the subfatgraph of the protein fatgraph consisting of the set of backbone atoms, whose distance along the backbone to one of the endpoints of $a$ is no more than $w$ atoms, together with all backbone and H-bond edges between them. We call $a$ the *central bond* of the pattern. The central bond determines the *signed length* of the bond, which is the distance between the central bond's endpoints, measured from the donor to the acceptor. Formally, if the donor of the bond is in the $i$'th peptide unit and the acceptor in the $j$'th peptide unit, the the signed distance $d$ is defined to be $j - i$ if $i < j$ and $j - i - 1$ otherwise (Figure 2).

We will also record the H-bonds whose two endpoints are no more than $w$ atoms away from one of the endpoints of $a$ along the backbone. An H-bond pattern may be expressed as a sequence of letters as follows (Figure 3a);
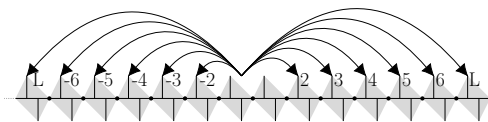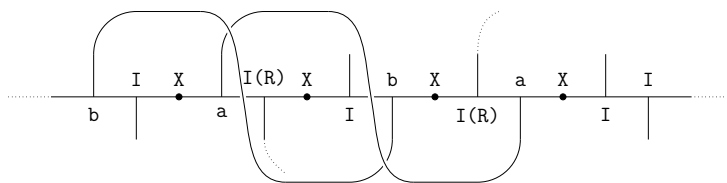
bIXaIXIbXIaXII



Figure 2: Signed length of H-bonds. Peptide units are shaded grey. All bonds longer than $\pm 6$ are denoted by L.

Each pair of lowercase letters indicate a pair of atoms with an H-bond between them. The letter "a" is given a special meaning as an indication of the central bonds. The remaining H-bonds are ordered by the position of the donor atoms (starting from the N-terminus) and given letters "b, c, ... ". The uppercase letter X indicates a $C^\alpha$ atom, and the uppercase letter I indicates an "isolated" N or O half-edge (with no H-bond attached). Hence we see that the above pattern corresponds to the structure shown in Figure 3a, with the second H-bond as the central bond. If the distance between the two endpoints of the central bond is greater than $2w$, the H-bond pattern is not connected, which may be indicated by ":". So we may obtain a pattern such as (Figure 3b);
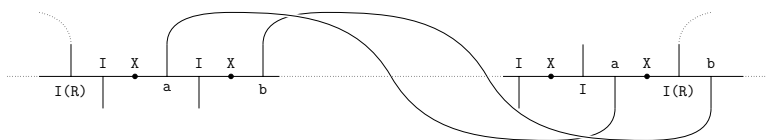
IIXaIXb:IXIaXIb

This local pattern description can be enhanced further by encoding various extra pieces of information:

- A backbone atom within the window may have an H-bond attached, whose other endpoint lies outside the window. In the basic pattern description, this would be indistinguishable from an unbonded backbone atom; using the letter R (for Remotely bonded) instead of I allows us to distinguish these cases. We may also restrict this information for atoms located at most $r$ from either endpoint of the central bond, for some $0 \le r \le w$. Atoms which are remotely bonded but lie further than $r$ from the central bond are then treated as if they were isolated atoms.

- We may record whether an H-bond is twisted or not. Formally, this is determined by whether the inner product between the normal vectors to the peptide planes is positive or negative. If the inner product is negative (that is, if the bond

(a) Connected pattern. The pattern can be expressed as `bIXaIXIbXIaXII_-3`.



(b) Disconnected pattern. The pattern can be expressed as `IIXaIXb:IXIaXIb_L`.

Figure 3: Connected (a) and disconnected (b) H-bond local patterns of window size 3. The parts of backbone and H-bonds, that do not participate in the pattern are shown by dotted lines. Note the remotely-bonded atoms may be replaced by the letter `R`, depending on the parameter specified (see text for details).

is twisted), we replace the lowercase letter with the corresponding letter from the other end of the alphabet (that is, `z` instead of `a`, `y` instead of `b` etc.). Similarly to the remote bonds, this information may be restricted to the atoms located at most $t$ atoms from either endpoint of the central bond, for some $0 \leq t \leq w$.

- Beside the local pattern itself, we may record separately the residues at the four $C^{\alpha}$'s closest to the central bond's endpoints. To reduce the number of possible bond descriptions and obtain reasonable clusters, we choose a grouping of the 20 residues into 1, 2, 3 or 4 groups according to their chemical properties (Supplementary Material), and simply record the 4-tuple of group identifiers. The four residues around an H-bond are chosen and recorded in the following order;

  1. The residue preceding the N-donor amino acid residue.
  2. The N-donor amino acid residue.
  3. The O-acceptor amino acid residue.
  4. The residue following the O-acceptor amino acid residue.

The resulting pattern description may look as follows:

$$\texttt{10XbRXcRXdRXzbXvcXIdXIvXRzXRRXRRXRRX}$$
$$\texttt{\_5\_4LLLL} \tag{1}$$

Here the number `10` indicates the window size, and the central bond is twisted; as shown by the use of the letter `z` instead of `a`. There are several remotely-bonded atoms, indicated by the letter `R`. The last segment, `4LLLL`, indicates the number of groups used in the grouping of residues and the four group identifiers.

The above description of H-bond patterns only relies on the fatgraph structure of a protein, with its backbone and H-bonds. It is also possible to generate a local pattern which includes information about a protein's tertiary contacts, by adding an edge to the fatgraph, where there is a tertiary contact. These edges may be labelled to indicate their status as tertiary contacts.

In the context of protein structure research, in particular the protein structure prediction, the effectiveness of deep learning approach has been demonstrated by AlphaFold and AlphaFold2 systems [31, 10] in CASP [15]. The current study differs from the protein structure prediction as studied in CASP in two points. Firstly, our input is the fatgraph structure of proteins, and not their primary sequences. Our method is not dependent on the primary sequence and the information on residues only plays an ancillary role. Secondly, our output is a spatial rotation along a given H-bond, a descriptor of the local structure around the H-bond. Our main focus is to study the relationship between the topology and the geometry of proteins, and to investi-

4

gate the use of spatial rotation as a descriptor of protein structures.

# 2 Methods and Results

## 2.1 Dataset

The dataset used was based on the HQ60 dataset in [20]. Here we give a brief definition of this dataset. PISCES [32] is a service that, among other things, creates subsets of sequences from PDB based on specified threshold for structure quality and sequence identity. For the HQ60 dataset, we use only X-ray structures, with a resolution threshold of 2.0Å, Rfac threshold of 0.2, and maximum sequence homology of 60%. The data was extracted from PDB [5] in March 2021, resulting in a collection of 16262 proteins. The H-bonds are taken from the DSSP program [11], with the additional conditions [4];

$$\text{HO-distance} < 2.7\text{Å}$$
$$\text{angle(NHO)}, \text{angle(COH)} > 90°.$$

The resulting dataset contained approximately 2.4 million H-bonds.

A subset of 200 proteins was randomly selected as the test data from the dataset, and the remaining 16,062 proteins were used as the training data.

## 2.2 H-bond pattern alignment

For each H-bond in the training data, H-bond local pattern was generated using the following parameter combination;

- Window size: 10

- Remote bonds: 10

- Twisted bonds: 10

- Number of residue classes: 1

Each H-bond pattern was translated to a sequence, which we call an H-bond length sequence, based on the lengths of the H-bonds. The translation was done as follows (see Figure 4).

1. Recall that each H-bond pattern corresponds to a repeated sequence of N-C$^\alpha$-O atoms, representing an amino acid, in our protein model. At either end of the pattern we may have only a part of this three-atoms sequence; for example it may start with a C$^\alpha$ or an O atom. Given an H-bond pattern, we split it into three-atom segments corresponding to amino acids (we may end up with segments at either end of the pattern which consist of fewer than three atoms). In the following procedure we only consider the atoms in the H-bond pattern.

2. For each of the resulting segments, check whether the N atom is a donor. If so, we assign a symbol encoding the signed length followed by the twistedness of the bond ("+" for twisted, and "-" for not twisted). E.g. "+4-" for a bond of length +4, which is not twisted ("-").

3. For each of the remaining segments, check whether the O atom is an acceptor. If so, we assign a symbol "A" to the segment.

4. The remaining segments are assigned a symbol "U" for unbonded.

We note the H-bond length sequences constructed as above completely determine the local H-bond fatgraph structure of proteins. Using the above translation method, we can compute the list of H-bond length sequence and associated rotation value from our training dataset. The same procedure is applied to the proteins whose H-bond rotations are to be predicted. For an H-bond local pattern of window size 10 around the central bond $a$ in the prediction dataset, let $s(a)$ be the corresponding H-bond length sequence (whose length is the number of whole or partial amino acids contained in the H-bond pattern). We then compute alignment score for $s(a)$ against each of the H-bond length sequences in the training dataset, using Needleman-Wunsch algorithm [17], with the substitution matrix constructed as follows. Let

$$\begin{aligned} S_1 &= \{-6-, -5-, \dots, +5-, +6-, \\ &\quad -6+, -5+, \dots, +5+, +6+\} \\ S_2 &= \{L-, L+\} \\ S &= S_1 \cup S_2 \\ K &= S \cup \{\text{U}, \text{A}\}. \end{aligned}$$

Define $l : S \to \mathbb{Z} \cup \{L\}$ to be the function that returns the length part of $x \in S$, i.e. it removes the last character in $x$. Define also
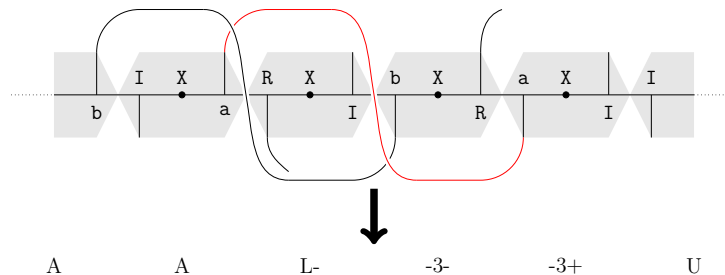
Figure 4: Translation of H-bond pattern to H-bond length sequence. Amino acids are shaded grey. A twisted bond is shown in red.

$t : S \rightarrow \{+, -\}$ to be the function that returns the twistedness of $x \in S$. Construct the substitution matrix $M^{(1)}$ with entries $M^{(1)}_{k_1,k_2}$, $k_1, k_2 \in K$ by the pseudocode shown in Algorithm 1. The gap score was set to -1.

---

**Algorithm 1** Pseudocode for the construction of substitution matrix $M^{(1)}$.

---

1: **for** $k_1 \in K$ **do**
2:     **for** $k_2 \in K$ **do**
3:         **if** $k_1 == k_2$ **then**
4:             $s = 1$
5:         **else if** $\{k_1, k_2\} \cap (K \setminus S) \neq \emptyset$ **then**
6:             $s = -1$
7:         **else if** $l(k_1) == l(k_2)$ **then**
8:             $s = 0$
9:         **else if** $\{k_1, k_2\} \cap S_2 \neq \emptyset$ **then**
10:             $s = -0.75$
11:         **else**
12:             $s = -\left|l(k_1) - l(k_2)\right|/20$
13:         **end if**
14:         **if** $\{k_1, k_2\} \subset S$ and $t(k_1) \neq t(k_2)$ **then**
15:             $s = s - 0.1$
16:         **end if**
17:         $M^{(1)}_{k_1,k_2} = s$
18:     **end for**
19: **end for**

---

Our prediction for the rotation along $a$ is the rotation value associated to the H-bond length sequence with the highest alignment score for $s(a)$.

Three further substitution matrices, $M^{(2)}, M^{(3)}$, and $M^{(4)}$ were constructed to investigate the effect of different penalties for mismatch. In $M^{(2)}$ the penalty for when $k_1, k_2 \in S_1$ ("short to short" substitution) was made exponential instead of linear, by replacing line 12 with

$d = -\left|l(k_1) - l(k_2)\right|$

$s = -0.6 \left((\exp d - 1)/(\exp 12 - 1)\right).$

In $M^{(3)}$ the penalty was made logarithmic by replacing line 12 with

$d = -\left|l(k_1) - l(k_2)\right|$
$s = -0.6 \left(\log(d + 1)/\log(12 + 1)\right).$

Finally, in $M^{(4)}$, the penalty for different twistedness values was increased by replacing line 15 by

$s = s - 0.8.$

The effect of different gap scores was investigated by using the simplest substitution matrix; 1 along the diagonal (match) and -1 elsewhere (mismatch), with different gap scores. We call these simple substitution matrices $M^{(p)}_{\text{simple}}$, where $p$ is the gap penalty score.

## 2.3  H-bond pattern matching

In the second method, we attempt to find an exact match for a given H-bond local pattern in the training data. To ensure we find a match, we generate a sequence of local patterns ranging in their complexity.

For each H-bond in the training data, H-bond local patterns were constructed with the following parameter combinations;

- Window size: 0 (only the central bond), $1, 2, \ldots, 10$

- Remote bonds: 0 (do not indicate remote bond), $1, 2, \ldots,$ window size

- Twisted bonds: none (-1), only the central bond (0), window size

- Number of residue classes: 1, 2, 3, 4

This resulted in 792 parameter combinations. For each of these parameter combinations, the H-bond patterns with less than 30

6

occurrences were discarded. For each of the remaining H-bond patterns with the associated SO(3) rotations, we performed a clustering analysis. A description of the clustering algorithm can be found in [20] (Method section). We note that rotations are represented as axis-angle pairs, and the algorithm uses a discretised rotational space; it divides the cube $(-\pi, \pi)^3$ into $81 \times 81 \times 81$ small boxes, and finds a mode box for each well-defined cluster. Each box can belong to at most one cluster, even when the algorithm finds several clusters.

We then compute the score $s$ for each clustering run by the following formula;

$$
s = \begin{cases} \pi - m & \text{if there is only one cluster} \\ -1 & \text{if there are } > 1 \text{ cluster,} \end{cases}
$$

$$(2)$$

where $m$ is the mean distance between all boxes in the cluster and the mode box, which of course is bounded by $\pi$. In this way we associate a score for each bond description. The result is a table where each row contains an H-bond pattern, a rotation value and a score. This data is then used to predict rotation of a given H-bond from a topological model of the protein, which we describe below.

Prediction was done on the test data of 200 proteins, using the same procedure as the training. For each H-bond in the test data, we obtain H-bond patterns using the same sets of parameter combinations as used in the training stage. Each resulting H-bond pattern is looked up in the table of patterns for which a clustering was performed. If a match is found, we obtain an estimate for the bond's rotation, which is the centre of the largest cluster, along with a score for that estimate, which is the score associated to the cluster. Our final prediction for the rotation associated to the H-bond is the estimate with the highest associated score. If two estimates have the same score, the result with more detailed H-bond pattern is used for the prediction.

In order to promote the selection of larger pattern for prediction, we modified the score function (2) to penalise the use of smaller patterns. The new score function is given by

$$
s = \begin{cases} \pi - & m - \exp(3 - w) \\ & \text{if there is only one cluster} \\ -1 & \text{otherwise,} \end{cases}
$$

$$(3)$$

where $m$ is the mean distance to the cluster mode, and $w = \max\{3, \text{window size}\}$.

The above analysis was repeated using the local H-bond patterns generated including the tertiary contact information, where the edges corresponding to tertiary contacts and backbone $\alpha$-carbon linkages contribute equally to the computation of window size, and we do not consider atoms more than one tertiary contact away from the central bond. In other words, an atom is only included in the H-bond pattern if the distance (along the backbone and the tertiary contact edges) from it to either end of the central bond is less than or equal to the window size, and that no more than one tertiary contact is traversed in computing the distance. This criteria was applied to limit the pattern to the atoms most likely to exert some influence on the central H-bond.

## 2.4 Results

The results of the prediction using H-bond length sequence alignment with the substitution matrices $M^{(1)}$ to $M^{(4)}$ are shown in Table 1. With the substitution matrix $M^{(1)}$, about half (49.56%) of the predictions were placed inside a ball centred around the true rotation, occupying just 0.1% of the volume of SO(3) (Figure 5a), and 72.91% inside a ball occupying 1% of the SO(3) volume (Figure 5b). We see that modifying the penalty for ("short to short") substitution appears to have positive effect on the prediction accuracy compared to the linear penalty ($M^{(1)}$), and the improvements appear to be similar for the different penalties ($M^{(2)}$ to $M^{(4)}$).

The effect of different gap scores using the simple substitution matrix $M_{\text{simple}}$ is shown in Table 2. Perhaps surprisingly, the prediction accuracy using $M_{\text{simple}}^{(-1)}$ was very similar to the result obtained by using $M^{(4)}$, the substitution matrix with a large twistedness penalty (Table 1). There was a reduction in accuracy when the gap score was set to 0, and a small improvement when it was set to -5.

The results from the pattern matching are shown in Table 3 under the column "Orig. score". We see in 61.10% of all cases, the predicted rotation lies within a ball comprising just 0.1% of the total volume of SO(3) centred at the true rotation, and in 86.90% of all cases, the prediction was within a ball corresponding to 1% of the volume of SO(3). To anal-
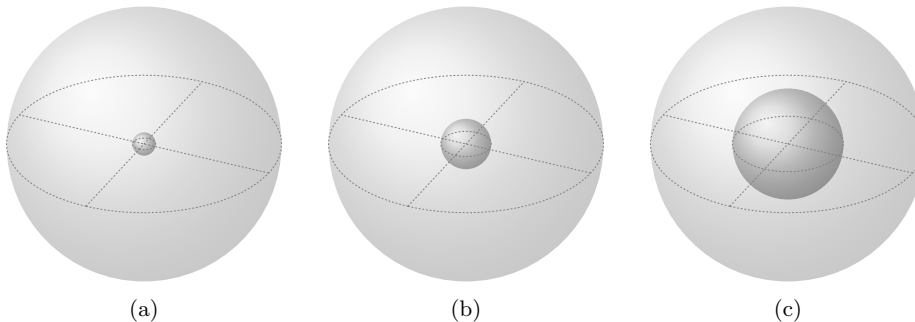
Figure 5: Illustrations of balls comprising 0.1% (a), 1% (b) and 10% (c) of the total volume of SO(3). The entire SO(3) is shown as a ball of radius $\pi$; a point $p$ in the ball defines a rotation via the angle-axis pair $(\theta, v)$, where $\theta = \|p\|$ and $v = p/\|p\|$.

| Range of $d$ | | $M^{(1)}$ | $M^{(2)}$ | $M^{(3)}$ | $M^{(4)}$ |
|---|---|---|---|---|---|
| $d < 0.2664$ | (0.1%) | 49.56 | 52.57 | 52.62 | 52.66 |
| $d < 0.4567$ | (0.5%) | 66.49 | 72.59 | 72.66 | 72.70 |
| $d < 0.5766$ | (1.0%) | 72.91 | 79.63 | 79.71 | 79.77 |
| $d < 0.7862$ | (2.5%) | 80.65 | 87.08 | 87.16 | 87.30 |
| $d < 0.9968$ | (5.0%) | 85.33 | 90.97 | 91.02 | 91.14 |
| $d < 1.2689$ | (10.0%) | 89.17 | 93.92 | 93.94 | 94.11 |
| $d < 1.7663$ | (25.0%) | 94.49 | 97.07 | 97.04 | 97.24 |
| $d < 2.3099$ | (50.0%) | 97.48 | 98.61 | 98.60 | 98.84 |
| $d < 2.7437$ | (75.0%) | 98.86 | 99.35 | 99.36 | 99.50 |
| $d < 3.1416$ | (100%) | 100.00 | 100.00 | 100.00 | 100.00 |

Table 1: Accumulative % of H-bonds whose predicted rotation values lies within the specified distance from the true rotation values, for different substitution matrices $M^{(1)}, M^{(2)}, M^{(3)}$ and $M^{(4)}$ (see section 2.2 for the details of how these are constructed). The numbers in parentheses show the volume of the ball whose radius is the upper limit of the range as a proportion of the volume of entire SO(3), w.r.t. the invariant metric.

yse the prediction results further, we looked at the DSSP classes [11] of four residues around each H-bond. DSSP assigns seven secondary structure classes (plus "unclassified") to each residue (Table 4).

We analysed the frequencies of four-tuples of DSSP classes and the associated prediction accuracies (Supplementary Material), and observed the frequencies concentrated on a few classes, which is also evident if we look at the ten most frequent DSSP class combinations (Table 5). We also observe that the residue class combination for the sheet structure ("EEEE") has the high frequency and relatively low accuracy.

We then analysed each prediction and looked at the parameters used to produce it. By looking at the distance between predicted and true rotations ($\Delta$) and the mean distance to cluster mode ($m$) for each prediction,

we found a group of predictions made using small window sizes, with many of them having large $\Delta$ values (Supplementary Material). This could happen, for example, if the H-bond pattern matched with a smaller window size has the associated cluster with a well-defined "peak" (thus having a low $m$ value and high score), while the cluster for a match with a larger window size has a lower "peak" (and a high $m$ value). The prediction was repeated with the modified score function (3), with the results shown in Table 3. We see the prediction accuracy was improved to 89.05% inside 1% SO(3) volume. Analysis of the other parameters (Remote bonds, Twisted bonds, and Residue groups) did not show any similar anomalies, and applying similar modifications to the score function to prioritise matches with more detailed patterns did not result in an improvement (results not shown).

| Range of $d$ | | $M_{\text{simple}}^{(-1)}$ | $M_{\text{simple}}^{(-5)}$ | $M_{\text{simple}}^{(0)}$ |
|---|---|---|---|---|
| $d < 0.2664$ | $(0.1\%)$ | 52.84 | 53.52 | 50.27 |
| $d < 0.4567$ | $(0.5\%)$ | 72.82 | 73.69 | 69.01 |
| $d < 0.5766$ | $(1.0\%)$ | 79.94 | 80.97 | 75.92 |
| $d < 0.7862$ | $(2.5\%)$ | 87.37 | 88.43 | 83.31 |
| $d < 0.9968$ | $(5.0\%)$ | 91.24 | 92.26 | 87.71 |
| $d < 1.2689$ | $(10.0\%)$ | 94.13 | 95.13 | 91.01 |
| $d < 1.7663$ | $(25.0\%)$ | 97.15 | 97.64 | 95.17 |
| $d < 2.3099$ | $(50.0\%)$ | 98.82 | 99.04 | 97.68 |
| $d < 2.7437$ | $(75.0\%)$ | 99.51 | 99.61 | 99.01 |
| $d < 3.1416$ | $(100\%)$ | 100.00 | 100.00 | 100.00 |

Table 2: Accumulative % of H-bonds whose predicted rotation values lie within the specified distance from the true rotation values, for the simple substitution matrices with different gap scores (see section 2.2 for how the definition of these matrices). The numbers in parentheses show the volume of the ball whose radius is the upper limit of the range as a proportion of the volume of entire SO(3), w.r.t. the invariant metric.

| Range of $d$ | | Orig. score | New score | Tertiary | 2020 data |
|---|---|---|---|---|---|
| $d < 0.2664$ | $(0.1\%)$ | 61.10 | 62.74 | 53.16 | 62.63 |
| $d < 0.4567$ | $(0.5\%)$ | 80.72 | 82.17 | 74.51 | 82.42 |
| $d < 0.5766$ | $(1.0\%)$ | 86.90 | 88.14 | 82.01 | 88.54 |
| $d < 0.7862$ | $(2.5\%)$ | 92.87 | 93.80 | 89.70 | 93.97 |
| $d < 0.9968$ | $(5.0\%)$ | 95.67 | 96.11 | 93.60 | 96.34 |
| $d < 1.2689$ | $(10.0\%)$ | 97.38 | 97.62 | 96.16 | 97.76 |
| $d < 1.7663$ | $(25.0\%)$ | 98.89 | 98.95 | 98.16 | 98.99 |
| $d < 2.3099$ | $(50.0\%)$ | 99.69 | 99.72 | 99.69 | 99.80 |
| $d < 2.7437$ | $(75.0\%)$ | 99.87 | 99.90 | 99.95 | 99.93 |
| $d < 3.1416$ | $(100\%)$ | 100.00 | 100.00 | 100.00 | 100.00 |

Table 3: Accumulative % of H-bonds whose predicted rotation values lie within the specified distance from the true rotation values, with the original score function (2), with the new score function (3), with tertiary bonds, and training only with data from 2020 to predict newly added proteins.

Table 3 also shows the results from the analysis, where the H-bond local patterns were generated including the tertiary contact information. We see that 82.01% of all predictions lay within 1% SO(3)-volume of true values.

Furthermore, we ran the pattern matching prediction algorithm with the modified score function (3) again, but only using PDB data from March 2020 for training, and and using 200 of the proteins added to PDB after March 2020 for validation. The results are shown in the column "2020 data" in Table 3, and demonstrates the stability of our method with new structural data.

## 3 Discussion

In the first method, where H-bond pattern alignment score was used to find the best match, we were able to achieve over 70% of the predictions within 1% volume of SO(3) space from the true rotations. Changing the substitution matrices to simulate non-linear penalties did not result in any significant change in prediction accuracy. A minor improvement was observed when the penalty for the twistedness was increased (Table 1). This may indicate the fact that the twistedness of an H-bond is directly related to its associated rotation value; if a bond is twisted and another is not twisted, it is unlikely that the two have similar rotation values. More surprisingly, an improvement similar to using large twistedness

| DSSP class | DSSP symbol | Local structure |
|---|---|---|
| $\alpha$-helix | H | |
| $3_{10}$-helix | G | Helix |
| $\pi$-helix | I | |
| Strand | E | Sheet |
| Isolated $\beta$-bridge residue | B | |
| Turn | T | Coil |
| Bend | S | |
| Unclassified | - | Unclassified |

Table 4: DSSP classes and corresponding local structure patterns.

| Residues | Frequency | Accuracy |
|---|---|---|
| HHHH | 9296 | 99.83% |
| EEEE | 5604 | 79.59% |
| HH-H | 780 | 99.36% |
| HTHH | 694 | 97.84% |
| T–T | 405 | 80.74% |
| TTHH | 390 | 96.15% |
| T-HH | 339 | 88.79% |
| GG-G | 290 | 93.45% |
| TS-T | 281 | 87.54% |
| H-HH | 263 | 91.25% |

Table 5: 10 most frequent combinations of DSSP classes around H-bonds, together with the proportions of predictions, which lie inside a ball centred at the true rotation having a volume corresponding to 1% of the total volume of SO(3).

penalty was observed when using the "simple" substitution matrix where a match is given the score of 1 and a mismatch -1. A further, even larger improvement was observed by using the gap score of -5. As the window size is constant for the training and test datasets, larger gap score has the effect of making the algorithm more like the one that simply counts the match/mismatch between the two H-bond sequences. It is however not immediately clear why this may increase the prediction accuracy compared to allowing gaps, or making penalties dependent on the change in the bond lengths. It should be noted, that even though the Needleman-Wunsch algorithm itself does not place any restriction on its use, in the study of protein structures it is typically used for aligning segments of primary sequences. Accordingly, there are substitution matrices which are generally accepted in the field and tested in various applications [16]. The sequences considered in this study have not been studied previously, and the substitution matrix was developed specifically for this study. It is therefore possible that a further development of the substitution matrix may improve the prediction accuracy.

In the second method, where the rotation prediction was done by finding an exact match for a given H-bond pattern, we were able to achieve close to 90% of our predictions lying inside 1% SO(3) volume of the true rotations. We believe there are potentials for further improvement, since an analysis of the clustering results show that if we could choose the "best" clustering result, i.e. the clustering result that lies closest to the true value, it will result in over 95% and 80% of "predictions" inside 1% and 0.1% SO(3)-volume, respectively. So it is possible that a better score function than (3) may give the necessary improvement. An improvement may also be possible by a method based on a machine-learning approach, which has proved so successful in protein structure prediction. It is also clear from the results of the DSSP class analysis, that we need to improve our predictions in the sheet structure, if we are to reach our goal. Our hypothesis is that due to the structural flexibility of sheets compared to helices, the bonds in sheets are affected by the nearby atoms (which are not necessarily near the bond in the secondary structure) to a greater extent than the bonds in helices. We have attempted, unsuccessfully so far, to take this into account by using tertiary bond information in our H-bond pattern generation. A better understanding of $\beta$-sheet structures and their topology may be needed to improve prediction of H-bond rotations inside $\beta$-sheets.

# References

[1] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, and Koyanagi, Yuki. "Enumeration of protein structures by matrix model techniques". In preparation.

[2] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, and Koyanagi, Yuki. "Topology of protein metastructure and $\beta$-sheet topology". Preprint.

[3] Andersen, Jørgen Ellegaard, Jensen, Jens Ledet, Koyanagi, Yuki, and Nielsen, Jakob Toudahl. "Using topology to estimate structural similarities of proteins". Preprint.

[4] Baker, E.N. and Hubbard, R.E. "Hydrogen bonding in globular proteins". *Progress in Biophysics and Molecular Biology* 44.2 (1984), pp. 97–179. ISSN: 0079-6107.

[5] Berman, Helen M., Westbrook, John, Feng, Zukang, Gilliland, Gary, Bhat, T. N., Weissig, Helge, Shindyalov, Ilya N., and Bourne, Philip E. "The Protein Data Bank". *Nucleic Acids Research* 28.1 (2000), pp. 235–242.

[6] Bordo, Domenico and Argos, Patrick. "The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins". *Journal of molecular biology* 243.3 (1994), pp. 504–519.

[7] Carugo, Oliviero and Djinović-Carugo, K. "Half a century of Ramachandran plots". *Acta Crystallographica Section D: Biological Crystallography* 69.8 (2013), pp. 1333–1341.

[8] Carugo, Oliviero and Djinović-Carugo, Kristina. "A proteomic Ramachandran plot (PRplot)". *Amino acids* 44.2 (2013), pp. 781–790.

[9] Grishaev, Alexander and Bax, Ad. "An empirical backbone- backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation". *Journal of the American Chemical Society* 126.23 (2004), pp. 7281–7292.

[10] Jumper, John, Evans, Richard, Pritzel, Alexander, Green, Tim, Figurnov, Michael, Tunyasuvunakool, Kathryn, Ronneberger, Olaf, Bates, Russ, Žídek, Augustin, Bridgland, Alex, Meyer, Clemens, Kohl, Simon A A, Potapenko, Anna, Ballard, Andrew J, Cowie, Andrew, Romera-Paredes, Bernardino, Nikolov, Stanislav, Jain, Rishub, Adler, Jonas, Back, Trevor, Petersen, Stig, Reiman, David, Steinegger, Martin, Pacholska, Michalina, Silver, David, Vinyals, Oriol, Senior, Andrew W, Kavukcuoglu, Koray, Kohli, Pushmeet, and Hassabis, Demis. "High Accuracy Protein Structure Prediction Using Deep Learning". In: *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*. 2020.

[11] Kabsch, Wolfgang and Sander, Christian. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22.12 (1983), pp. 2577–2637.

[12] Levitt, Michael. "A simplified representation of protein conformations for rapid simulation of protein folding". *Journal of molecular biology* 104.1 (1976), pp. 59–107.

[13] Lovell, Simon C, Davis, Ian W, Arendall III, W Bryan, De Bakker, Paul IW, Word, J Michael, Prisant, Michael G, Richardson, Jane S, and Richardson, David C. "Structure validation by C$\alpha$ geometry: $\phi$, $\psi$ and C$\beta$ deviation". *Proteins: Structure, Function, and Bioinformatics* 50.3 (2003), pp. 437–450.

[14] Morozov, Alexandre V, Kortemme, Tanja, Tsemekhman, Kiril, and Baker, David. "Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations".

*Proceedings of the National Academy of Sciences* 101.18 (2004), pp. 6946–6951.

[15] Moult, John, Pedersen, Jan T, Judson, Richard, and Fidelis, Krzysztof. "A large-scale experiment to assess protein structure prediction methods". *Proteins: Structure, Function, and Bioinformatics* 23.3 (1995), pp. ii–iv.

[16] Mount, David W. "Comparison of the PAM and BLOSUM amino acid substitution matrices". *Cold Spring Harbor Protocols* 2008.6 (2008), pdb–ip59.

[17] Needleman, Saul B. and Wunsch, Christian D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 0022-2836.

[18] Pace, C Nick, Scholtz, J Martin, and Grimsley, Gerald R. "Forces stabilizing proteins". *FEBS letters* 588.14 (2014), pp. 2177–2184.

[19] Peng, Xubiao, Chenani, Alireza, Hu, Shuangwei, Zhou, Yifan, and Niemi, Antti J. "A three dimensional visualisation approach to protein heavy-atom structure reconstruction". *BMC structural biology* 14.1 (2014), pp. 1–16.

[20] Penner, Robert, Andersen, Ebbe Sloth, Jensen, Jens Ledet, Kantcheva, Adriana Krassimirova, Bublitz, Maike, Nissen, Poul, Rasmussen, Anton Michael Havelund, Svane, Katrine Louise, Hammer, Bjørk, Rezazadegan, Reza, Nielsen, Niels Christian, Nielsen, Jakob Toudahl, and Andersen, Jørgen Ellegaard. "Hydrogen bond rotations as a uniform structural tool for analyzing protein architecture". *Nature Communications* 5 (2014).

[21] Penner, Robert, Knudsen, Michael, Wiuf, Carsten Henrik, and Andersen, Jørgen Ellegaard. "An Algebro-Topological Description of Protein Domain Structure". *P L o S One* 6.5 (2011).

[22] Penner, Robert, C., Knudsen, Micheal, Wiuf, Carsten, and Andersen, Jørgen Ellegaard. "Fatgraph models of proteins". *Communications on Pure and Applied Mathematics* 63.10 (2010), pp. 1249–1297.

[23] Pereira, Joana. "Structural Bioinformatics and Crystallography Tools for Automated Protein Model Building and Validation". PhD thesis. Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2016.

[24] Rackovsky, S and Scheraga, HA. "Differential geometry and polymer conformation. 1. Comparison of protein conformations1a, b". *Macromolecules* 11.6 (1978), pp. 1168–1174.

[25] Rackovsky, S and Scheraga, HA. "Differential geometry and protein folding". *Accounts of Chemical Research* 17.6 (1984), pp. 209–214.

[26] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. "Stereochemistry of polypeptide chain configurations". *J. Mol. Biol.* 7 (1963), pp. 95–99.

[27] Read, Randy J, Adams, Paul D, Arendall III, W Bryan, Brunger, Axel T, Emsley, Paul, Joosten, Robbie P, Kleywegt, Gerard J, Krissinel, Eugene B, Lütteke, Thomas, Otwinowski, Zbyszek, et al. "A new generation of crystallographic validation tools for the protein data bank". *Structure* 19.10 (2011), pp. 1395–1412.

[28] Røgen, Peter and Bohr, Henrik. "A new family of global protein shape descriptors". *Mathematical Biosciences* 182.2 (2003), pp. 167–181.

[29] Røgen, Peter and Fain, Boris. "Automatic classification of protein structure by using Gauss integrals". *Proceedings of the National Academy of Sciences* 100.1 (2003), pp. 119–124.

[30] Rose, George D and Wolfenden, Richard. "Hydrogen bonding, hydrophobicity, packing, and protein folding". *Annual review of biophysics and biomolecular structure* 22.1 (1993), pp. 381–415.

[31] Senior, Andrew W, Evans, Richard, Jumper, John, Kirkpatrick, James, Sifre, Laurent, Green, Tim, Qin, Chongli, Žídek, Augustin, Nelson, Alexander WR, Bridgland, Alex, et al. "Improved protein structure prediction using potentials from deep learning". *Nature* (2020), pp. 1–5.

[32]  Wang, G. and Dunbrack, R. L.
      "PISCES: a protein sequence culling
      server". *Bioinformatics* 19 (2003),
      pp. 1589–1591.

# S1 Supplementary Material

## S1.1 Spatial rotation along a H-bond

We give a short summary of the assignment of SO(3) element to a H-bond, as illustrated in Figure S1. First we associate a triple of orthogonal unit vectors $\mathcal{F}_i = (u_i, v_i, w_i)$, $u_i, v_i, w_i \in \mathbb{R}^3$ to the $i$'th peptide unit $P_i$, $i = 1, 2, \ldots$ as follows. $u_i$ is the unit vector parallel to the displacement vector $\overrightarrow{C_i N_{i+1}}$ along the peptide bond. Let $u_i^\perp$ be the vector in the plane of peptide unit which is perpendicular to $u_i$. Let $\bar{v}_i$ be the projection of the vector $\overrightarrow{C_i^\alpha C_i}$ onto the subspace $\mathbb{R} u_i^\perp$ and set $v_i = \bar{v}_i / \|\bar{v}_i\|$. Finally, we set $w_i = u_i \times v_i$, where $\times$ denotes the cross product. For each $i$, there is a unique element of SO(3) $A_{P_i}$, which takes the standard three frame $(x, y, z)$ to $\mathcal{F}_i$. Let $\mathcal{F}_j$ be the three frame associated to another peptide unit $P_j$. Then there is a unique element of the group SO(3), $A_{P_j}(A_{P_i})^{-1}$, which takes $\mathcal{F}_i$ to $\mathcal{F}_j$. However, this element of SO(3) is not invariant under rotation of the entire protein. The solution to this problem is to apply the rotation $A_{P_i}^{-1}$ to $\mathcal{F}_i$ and $\mathcal{F}_j$, so that $\mathcal{F}_i$ becomes the standard three-frame in $\mathbb{R}^3$. The element of SO(3) that takes the standard three-frame to $A_{P_i}^{-1}\mathcal{F}_j$ is $A_{P_i}^{-1}A_{P_j}$. If there is a H-bond from $P_i$ to $P_j$ in our model, we assign $A_{P_i}^{-1}A_{P_j} \in$ SO(3) to this edge.
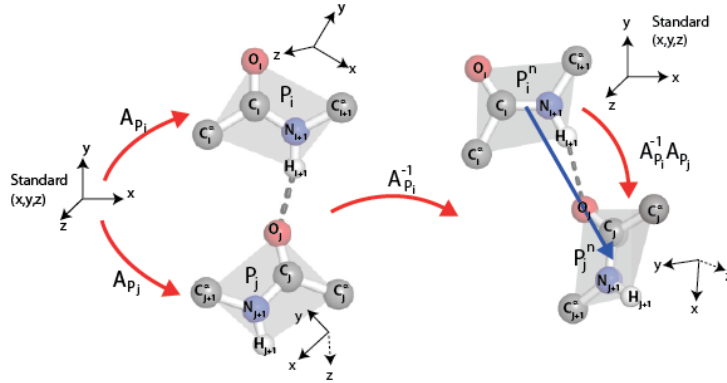


Figure S1: Assignment of SO(3) element to a H-bond. See text for details.

## S1.2 Grouping of amino-acid residues for construction of H-bond local patterns

When constructing H-bond local patterns, we divide the 20 "standard gene-encoded" amino acids into up to four groups according to the scheme shown in Table S1.

| Amino acids | (L) Leucine | (V) Valine | (I) Isoleucine | (F) Phenylalanine | (M) Methionine | (A) Alanine | (G) Glycine | (S) Serine | (C) Cysteine | (E) Glutamic acid | (K) Lysine | (R) Arginine | (D) Aspartic acid | (T) Threonine | (Y) Tyrosine | (N) Asparagine | (Q) Glutamine | (H) Histidine | (W) Tryptophan | (P) Proline | (X) N/A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of groups** 1 | X | | | | | | | | | | | | | | | | | | | | |
| 2 | L | | | | | | | | | E | | | | | | | | | | | X |
| 3 | L | | | | | A | | | | E | | | | | | | | | | | X |
| 4 | L | | | | | A | | | | E | | | | | | | | | | P | X |

Table S1: Grouping scheme for amino acids and group labels.

## S1.3 Analysis of prediction accuracy by the four residues around the central bond

We analysed accuracy (Figure S2a) and frequency (Figure S2b) of our predictions by the DSSP classes of four residues around the H-bonds. The residues are ordered along the backbone, so the two-letter combinations along the horizontal (donor-side) and the vertical (acceptor-side) axes show the DSSP class of the residue immediately preceding, followed by the residue immediately after, the H-bond donor (acceptor).
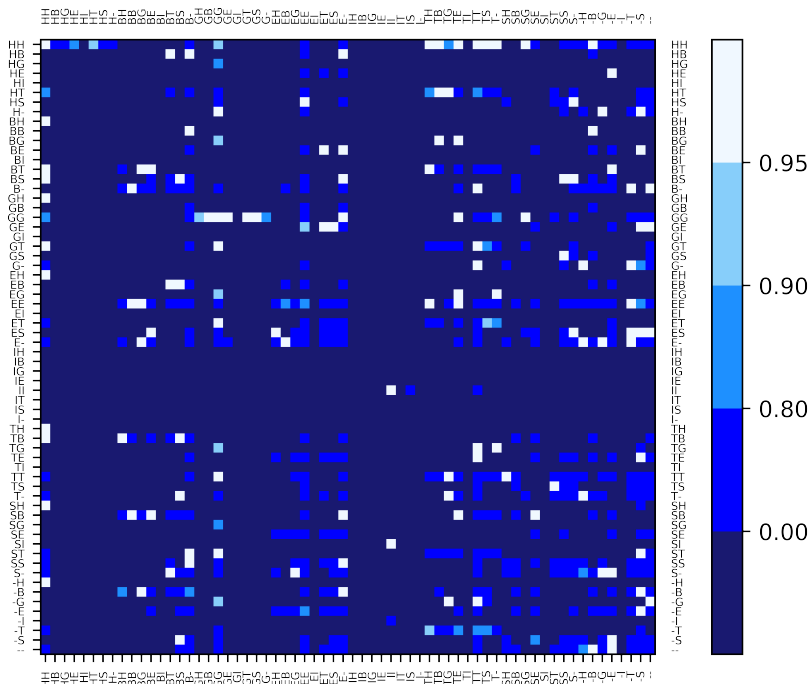
(a) Accuracy of predictions by the DSSP classes of four residues around the H-bond; % of predictions lying within 1% SO(3) volume.
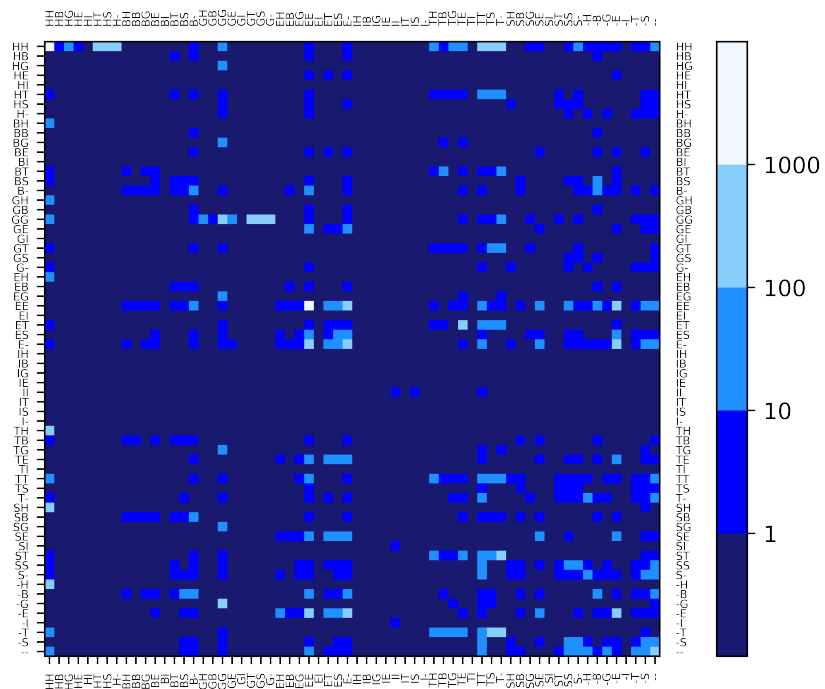


(b) Frequency by the DSSP classes of four residues around the H-bond.

Figure S2: Accuracy and frequency of predictions by DSSP classes. The horizontal and vertical axes show the donor-side and the acceptor-side residues, respectively.

## S1.4  Relation between the prediction accuracy and the size of clusters

### S1.4.1  Analysis by window size

We analysed the relation between the cluster scores, distance between the predicted and true rotations, and the window size used for prediction, using the results obtained by the original score function (2). Figure S3 shows the distribution of the scores (here using the mean distance to the mode box, or $m$ in (2)) against the distance between the predicted and true rotations ($\Delta$), filtered by the window size used for prediction. A cluster of predictions with low $m$ values and high $\Delta$ values can be seen in window size 2, and to a smaller extent in window size 3. We modified the score function to (3), to promote the selection of larger pattern for prediction. This had the desired effect on the prediction results, with few predictions made by small patterns (Figure S4).

### S1.4.2  Analysis by remote bond

We analysed the relation between the cluster scores, distance between the predicted and the true rotations, and the remote bond parameter used for predictions. Figure S5 shows the distribution of $m$ against $\Delta$, filtered by the difference between window size and remote bond parameters used for prediction, obtained by the score function (2). Similar clusters with low $m$ and high $\Delta$ values can be observed for the difference of 2. The cluster is linked to the one observed in **??**, where the remote bond parameter is 0. This can be verified by running the same analysis, but this time with the results obtained by the modified score function (3). Indeed, such cluster can not be observed with the data from the score function (3) (Figure S6.

### S1.4.3  Analysis by twistedness

We analysed the relation between the cluster scores, distance between the predicted and the true rotations, and the twistedness parameter used for predictions. We can again observe a cluster with low $m$ and high $\Delta$ values in the data produced with the original score function ((2); Figure S7), which is no longer present when the modified score function (3) is used (Figure S8.

### S1.4.4  Analysis by residue groups

We analysed the relation between the cluster scores, distance between the predicted and the true rotations, and the residue group parameter used for predictions. We can again observe a cluster with low $m$ and high $\Delta$ values in the data produced with the original score function ((2); Figure S9), which is no longer present when the modified score function (3) is used (Figure S10.

Figure S3: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the widnow size used for prediction. No prediction was made using window size 1, so the plot was omitted. The red line in dicates the mean for each plot.
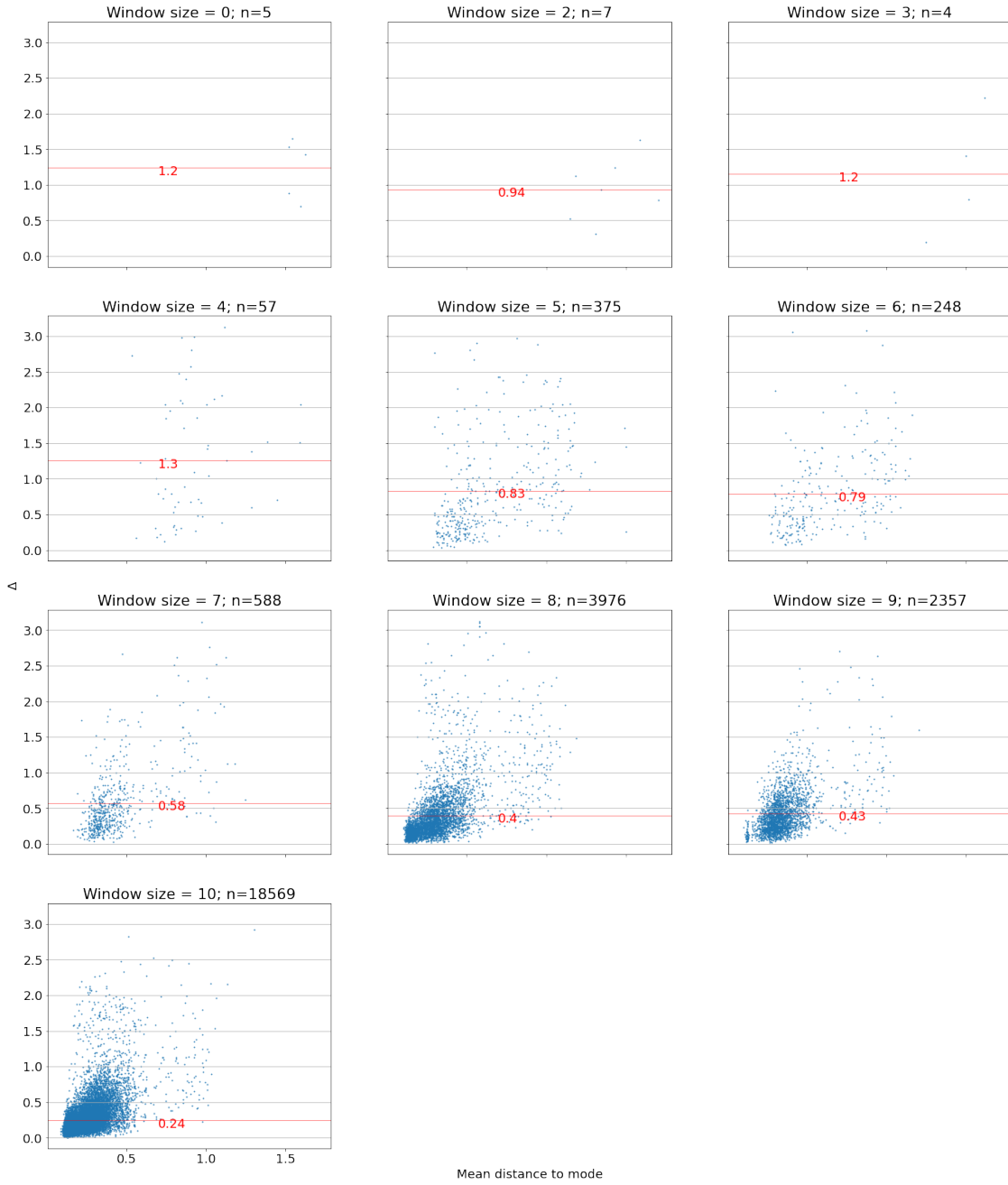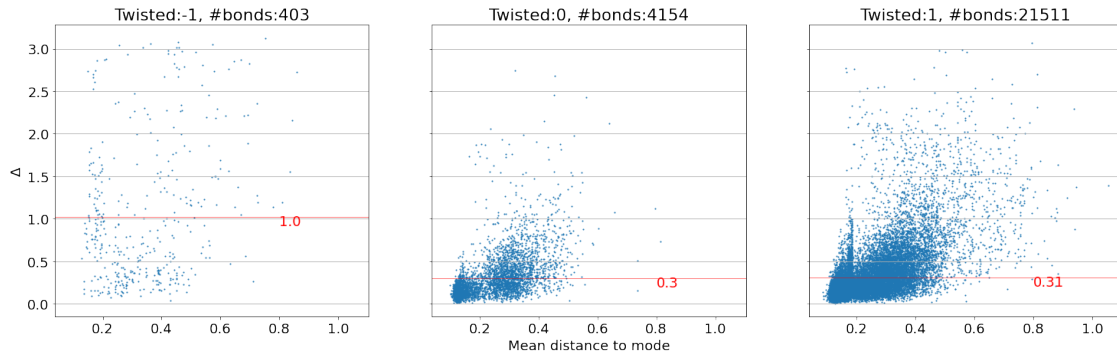
Figure S4: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the widnow size used for prediction. No prediction was made using window size 1, so the plot was omitted. The data is obtained by the modified score function (3). The red line in dicates the mean for each plot.

Figure S5: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the difference between the widnow size and remote bond parameter used for prediction. The red line in dicates the mean for each plot.

Figure S6: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the difference between the widnow size and remote bond parameter used for prediction. The data is obtained by the modified score function (3). The red line in dicates the mean for each plot.

Figure S7: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the twistedness parameter used for prediction (-1: no twisted bond recorded, 0: twistedness of central bond only, 1: twistedness of all bonds). The red line in dicates the mean for each plot.
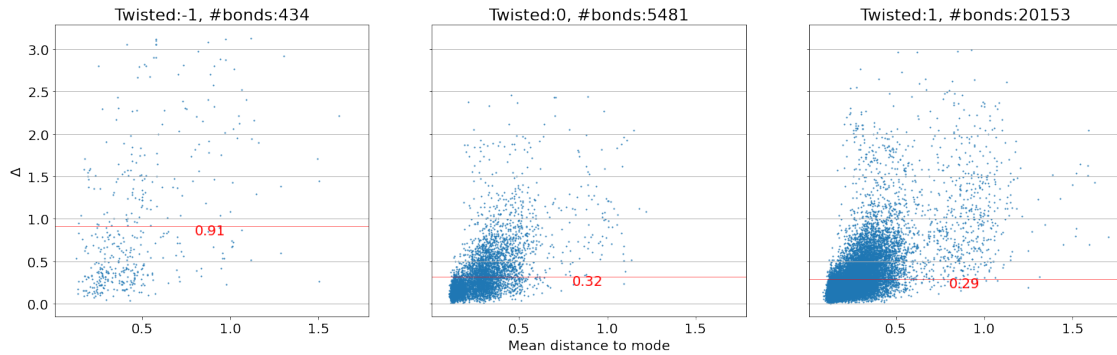


Figure S8: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the twistedness parameter used for prediction (-1: no twisted bond recorded, 0: twistedness of central bond only, 1: twistedness of all bonds). The data is obtained by the modified score function (3). The red line in dicates the mean for each plot.
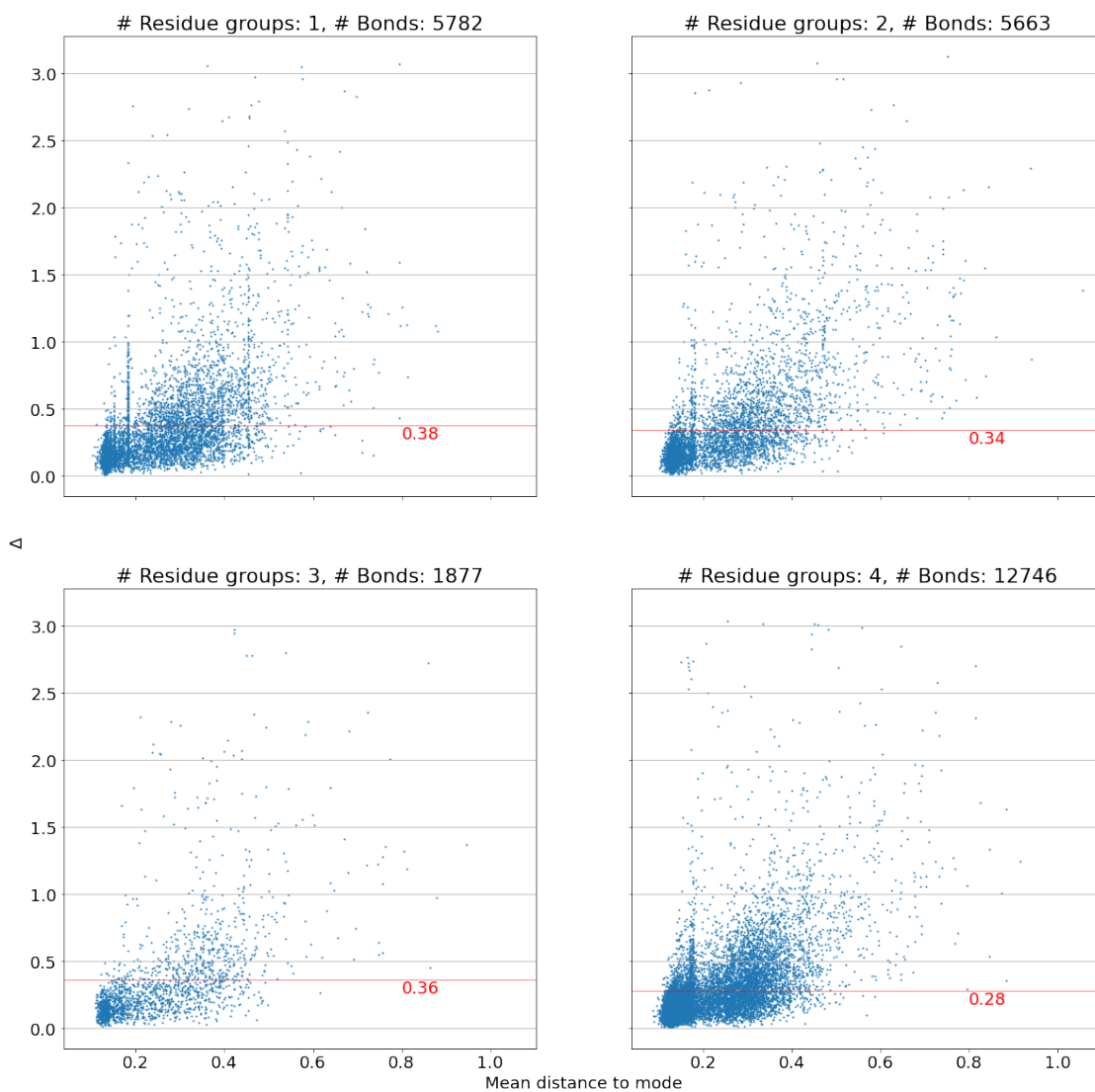
9

Figure S9: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the residue group parameter used for prediction (the parameter stores the number of groups used to classify residues). The red line in dicates the mean for each plot.
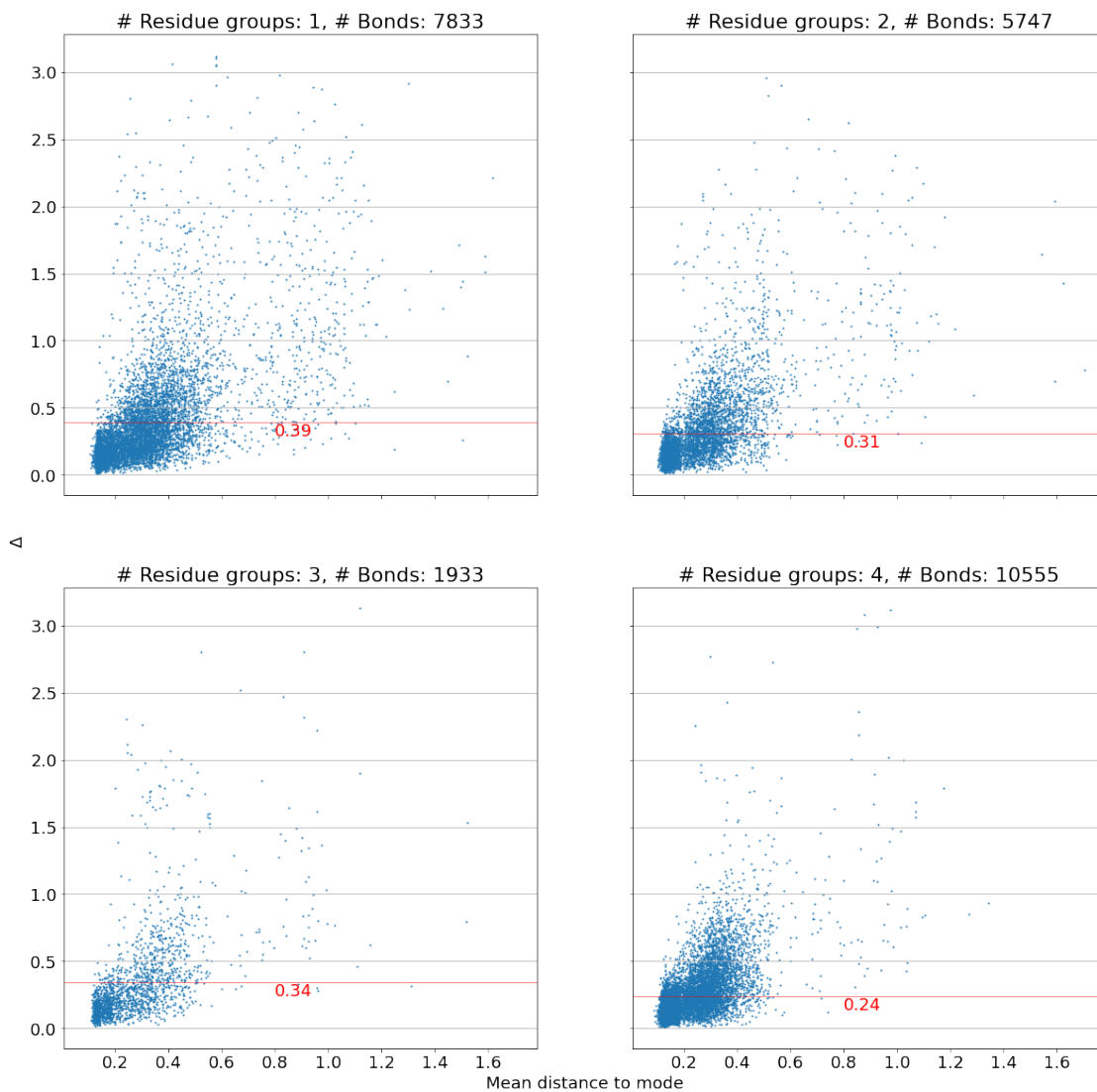
Figure S10: Distance to the true rotation ($\Delta$) and mean distance to cluster mode ($m$), filtered by the residue group parameter used for prediction (the parameter stores the number of groups used to classify residues). The data is obtained by the modified score function (3). The red line in dicates the mean for each plot.