Topology of protein metastructure and β-sheet topology

Andersen, Jørgen Ellegaard; Fuji, Hiroyuki; Koyanagi, Yuki

Publication date:
2021

Document version:
Submitted manuscript

Topology of protein metastructure and β-sheet topology

Jørgen Ellegaard Andersen[1,2], Hiroyuki Fuji[3,4], and Yuki Koyanagi[2]

[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] *Faculty of Information Science and Technology, Osaka Institute of Technology*
[4] *Yukawa Institute for Theoretical Physics, Kyoto University*

Abstract

We introduce a new, simplified model of proteins, which we call protein metastructure. The metastructure of a protein carries information about its secondary structure and β-strand conformations. Furthermore, protein metastructure allows us to associate an object called a fatgraph to a protein, and a fatgraph in turn gives rise to a topological surface. It becomes thus possible to study the topological invariants associated to a protein. We discuss the correspondence between protein metastructures and fatgraphs, and how one can compute topological invariants, such as genus and the number of boundary components, from fatgraphs. We then describe an algorithm for generating likely candidate metastructures using the information obtained from topology of protein fatgraphs. This algorithm is further developed to predict β-sheet topology of proteins, with a possibility to combine it with an existing algorithm. We demonstrate the algorithm on the data from PDB, and improve the performance of and existing algorithm by combining with it.

1 Introduction

The configurations of β-strands in a protein, often called β-sheet topologies, have been studied since the 1970's [28]. β-sheets, along with α-helices, are one of the fundamental structural components in the proteins. As opposed to helices, their structures involve interactions between residues which are far apart along the backbone. A better understanding of their structures and foldings is therefore crucial, if we are to understand the folding mechanism of entire proteins. The problem is further complicated by the intrinsic flexibility of β-sheet structures compared to α-helices [16]. Early studies [28, 27, 31] have identified some general rules (such as the preference for the right-handedness in parallel β-sheets) from investigation of individual proteins. As the amount of available data increased, studies have used computer programs to survey the database and found frequent patterns in the β-strand configurations [34, 29]. The information can be used to filter and rank a series of decoy structures by computing probabilities for different patterns [29]. Another approach is to assign pseudoenergy to each pair of β-strand residues and solve the β-sheet topology prediction problem as an optimisation problem [15]. At least one study [19] has compared the two methods, and found that the latter's performance to be better. One may also combine the two methods by, for example, forbidding certain β-strand configurations that are not found in the database [32], or by incorporating the two in Bayesian modelling [12]. Other studies used integer programming techniques to predict β-sheet topologies [30, 18].

Fatgraph is a mathematical object, that has been used successfully to study topological structures of another biological macromolecule, RNA [24, 26, 2]. A fatgraph can be thought of as a standard graph, where the edges and vertices have been "fattened" to ribbons and discs to form a surface (see Section 1.2 for details). It has been particularly useful in solving the problem of enumerating topologically distinct RNA [5, 3, 4] and protein structures [6]. Furthermore, the technique can be adopted to lower levels of abstraction by considering more parameters,

thus allowing for enumeration of more realistic structures [1, 9, 8, 7]. It has also been shown that the topology of proteing fatgraphs is strongly linked to their geometric structures [11, 10]. Inspired by their success, we introduce a new model for studying β-sheet topology of proteins, which we call protein metastructure. This model greatly simplifies the study of β-sheet topologies by amalgamating consecutive residues belonging to the same secondary structure, but still retains the information needed to understand the configuration of β-strands. We give a detailed definition in Section 1.1. Furthermore, each metastructure corresponds to a fatgraph, and this transition to fatgraphs allows us to compute topological invariants such as the number of boundary components and genus associated to each protein. The details of this correspondence are described in Section 1.2. Our use of fatgraphs in studying β-sheet topologies was inspired by [25], but our construction is much simpler, and can be constructed without the knowledge of proteins' geometric structures. We will use the topology of fatgraphs associated to proteins to predict β-strand conformations of proteins. More specifically, we will use the distribution of genus and number of boundary components to filter the candidate structures, whose topology does not agree with the distribution (see Section 1.3 and Section 3 for details).

1.1 Protein Metastructure

Given a protein, its primary structure is the sequence of amino acids in the polypeptide chain. There are 20 different amino acids in the standard gene code, so a primary structure can be expressed as a finite word in an alphabet with 20 letters;

$$\texttt{EKKSINECDLKGKKVLIRVDFNVP}... \qquad (1)$$

The secondary structure of a protein can be defined as a set of local substructures, most frequent of which are α-helices and β-sheets. The DSSP-algorithm [21] is an algorithm commonly used to classify residues into 3 or 7 secondary structure classes. When used (with 3-class output) on the above protein it produces a word in an alphabet with 3 letters;

$$\gamma\gamma\gamma\gamma\alpha\alpha\alpha\gamma\gamma\gamma\gamma\gamma\gamma\beta\beta\beta\beta\beta\beta\beta\gamma\gamma\gamma\gamma\gamma... \qquad (2)$$

Here we used the letter α for Helices, β for Sheets, and γ for Coils. When we apply this reduction to the data extracted from PDB [14] (see Section 2 for details of data selection), we begin to see some patterns in the proportion of these classes in proteins. There are, for example, few proteins which contain less than 25% γ residues, or more than 75% of any one class (Figure 1a). This can be explained by the rigidness of helix and sheet structures; a protein composed (almost) exclusively of α or β residues will not have the necessary flexibility to bend and fold into its native structure. For that to occur, a certain proportion of γ residues are required. On the other hand, too much γ residues would most likely result in lack of stability and will be energetically unfavourable. The largest concentration appears to be around 30$\sim$50% α, 10$\sim$30% β, and 30$\sim$50% γ residues (Figure 1a).

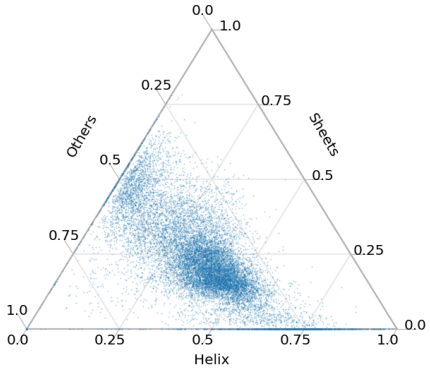
We now introduce the *reduced secondary structure sequence* by reducing each segment of identical letters in a secondary structure sequence (2) to a single letter;

$$\gamma\alpha\gamma\beta\gamma\beta\gamma\beta\gamma\alpha...$$

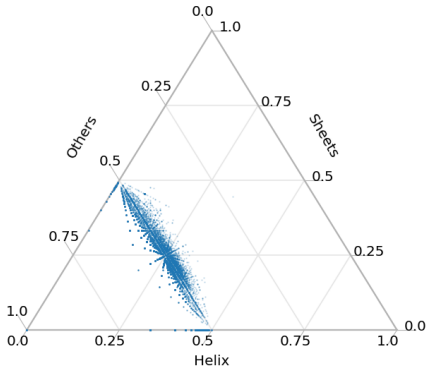
Not surprisingly, the distribution of proportions of the 3 classes in such reduced sequences are concentrated around $\gamma = 50\%$ (Figure 1b), since the reduced sequences are mostly sequences of $\gamma\alpha$s and $\gamma\beta$s by construction.

In a reduced sequence, each letter β corresponds to a β-strand. We may therefore add an additional data to a reduced sequence to specify β-sheet structure of the protein. To do this, we define the *protein metastructure* as the triple (r, P, A), where r is a finite word in an alphabet of three letters, α, β and γ, and P and A are sets of pairs of integers (i, j) for some $1 \leq i < j \leq s$, where s is the number of letter β in r. We also put a further condition, that $P \cap A = \emptyset$. Then for a given protein, we obtain its metastructure by setting r to be the reduced sequence, and populating P and A as follows;

1. Number the letters β in r along the backbone, starting from the N-terminus.
2. Identify all pairs (i, j), where there is at least one hydrogen bond between ith and jth strands.
3. Let I be the set of all pairs (i, j) identified in the previous step. Partition I into two sets P and A, where P consists of all parallel connections and A all anti-parallel connections.



(a) Proportion of 3 classes in secondary structure sequences



(b) Proportion of 3 classes in reduced sequences

Figure 1: Proportion of 3 classes in 16264 selected proteins

If there is only a single bond between two strands, thus making it impossible to determine the configuration between the two, we extend the strands by up to three residues. If it is still not possible to determine the configuration (because the extended strands has a single bond between them), then we assign the pair to P, as parallel configuration. This is because the standard anti-parallel configuration requires two hydrogen bonds between a pair of residues, thus making it less likely that there is only one hydrogen bond present. This forced assignment was necessary only in 181 out of 10141 proteins in the dataset, representing 1.8% of the data.

Let $\mathcal{S}$ be the set of all possible metastructures, and let $\mathcal{S}_{\text{bif}} \subset \mathcal{S}$ be the subset consisting of all metastructures, where at least one β-strand is connected to more than 2 other strands (bifurcations). Similarly, let $\mathcal{S}_{\text{bar}}$ be the subset of metastructures with β-barrels (see Table 1 for the size of these subsets of metastructures). Consider $\tilde{\mathcal{S}} = \mathcal{S} \setminus (\mathcal{S}_{\text{bif}} \cup \mathcal{S}_{\text{bar}})$. For each $s \in \tilde{\mathcal{S}}$, we can associate a metastructure motif diagram (Figure 2) as follows;

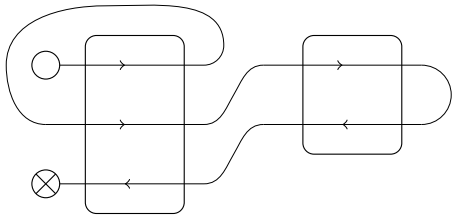


Figure 2: An example metastructure motif diagram. The associated metastructure may be $(\gamma\beta\gamma\alpha\gamma\beta\gamma\beta\gamma\beta\gamma\beta\gamma, \{(1,2)\}, \{(3,4),(2,5)\})$

1. Each β-strand is denoted by a straight line segment with an arrowhead in the middle.
2. If $(i,j) \in P$, draw the i'th and j'th strands next to each other, with arrowheads on both segments pointing the same direction.
3. If $(i',j') \in A$, draw the i'th and j'th strands next to each other, with arrowheads pointing the opposite direction.
4. Draw a "sheet" around each stack of strands.
5. Connect the strands, from the 1st to last, following the directions of arrowheads, and avoiding the interior of the sheets.
6. N-terminus is denoted by $\circ$, and C-terminus is denoted by $\otimes$.
7. Note for each sheet, we have a choice of 2 strands to draw on the top, and a choice of which way the first strand points to (see Figure 3; orientation of all other strands are then decided by parallel/anti-parallel configurations). We can make this canonical by requiring that;
   (a) the top strand comes before the bottom strand in the backbone-ordering
   (b) the first strand (in the backbone-ordering) in a sheet points from left to right

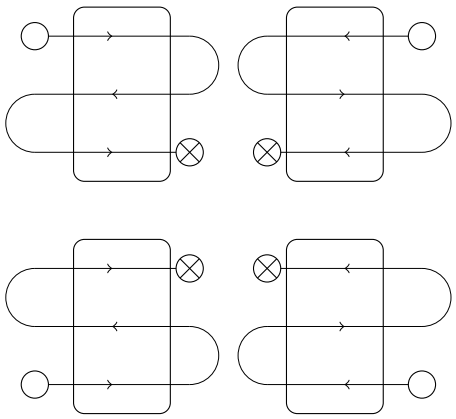
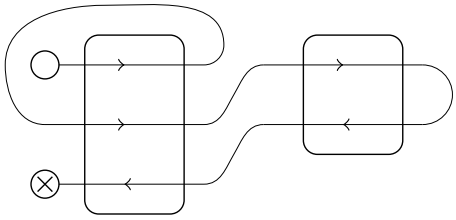


Figure 3: Four equivalent metastructure diagrams. The two requirements force us to choose the top-left diagram.



	1	2	3	4	5
1	0	1	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	1	0	0
5	0	1	0	0	0

Figure 4: A protein metastructure diagram and the corresponding pairing matrix. The 1 in (1,2)-th entry corresponds to the parallel configuration between the first and the second strand in the backbone, and the two 1's in the lower-triangular part correspond to the anti-parallel configurations between the third and the fourth strands, and the second and the fifth strands.

We note that the metastructure diagram, hence the β-sheet topology of a protein with n β-strands can be recorded in an $n \times n$ matrix, whose entries are either 0 or 1. This can be done by setting (i, j)'th entry to 1 if the i'th and j'th strands are paired in the parallel configuration, and setting (j, i)'th entry to 1 if the pairing is anti-parallel. All other entries (where there is no pairing observed) are set to 0. We call this matrix $\mathbb{P}$ the protein's pairing matrix (Figure 4). The 1's in the upper-triangular part show parallel pairings, and the 1's in the lower-triangular part show anti-parallel pairings. The number of paired strands the ith strand has can be computed as the total number of 1 cells in the ith row and column. In a pairing matrix $\mathbb{P}$, an isolated strands can be seen as zero row and column; the i'th strand is isolated (has no paired strand), if and only if the i'th row and the i'th column do not have a 1. Similarly, the i'th strand has bifurcation, if and only if the total number of 1 cells in the i'th row and column is strictly greater than 2. A β-sheet manifests itself as a "chain" of strands, with the edge strand having only one non-zero entry in the corresponding row or column. A β-barrel is a circular chain without edge strands (Figure 5).

Note, if $\mathcal{T}$ is the set of metastructure motifs, the map $\varphi : \tilde{\mathcal{S}} \to \mathcal{T}$ described above corresponds to "forgetting" r in $(r, P, A) \in \tilde{\mathcal{S}}$.

1.2 Fatgraph

In order to understand topological characteristics of protein metastructures, we need to pass from metastructure diagrams to topological surfaces. The main idea is to "thicken" the non-β segments in a given metastructure diagram to (untwisted) bands or ribbons, as in Figure 6, to produce a fatgraph $\mathbb{D}$. Formally, a fatgraph $\mathbb{D}$ is a graph D together with a cyclic ordering of the incident half-edges at each vertex. It can be obtained from a metastructure diagram by contracting each sheet to a point, and ordering the resulting half-edges at each vertex anti-clockwise from the N-terminus, or the starting end of the first strand in the sheet (Figure 7). A fatgraph $\mathbb{D}$ gives rise to a unique (orientable) surface $X_{\mathbb{D}}$ by thickening each edge to a band and each vertex to a disc. As an orientable surface, it obeys Euler's formula

$$\chi(X_{\mathbb{D}}) = v - e + n = 2 - 2g,$$

where v is the number of vertices (which correspond to the β-sheets in the metastructure diagram), e the number of edges or bands (corresponding to the non-β segments, excluding

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	1	0	0

Figure 5: An example pairing matrix with three sheets. The first sheet consists of strands 1, 2, and 3 in anti-parallel configuration. The edge strands are 1 and 3, which can be seen by the fact that the total number of 1-cells in the first (or third) row and column is 1. On the other hand, the total number of 1-cells in the second row and column is 2, indicating the strand 2 is paired to two other strands. The second sheet, consisting of strands 4, 5, and 6, has no edge strand, and forms a barrel. The third sheet consists of strands 7, 8, 9, and 10. We see the strand 8 has three paired strands, thus indicating a bifurcation.

the N- and C-terminal segments), n the number of boundary components, and g the genus of $X_{\mathbb{D}}$.

Note this map ψ from $\mathcal{T}$ to the set Σ of fatgraphs with two marked half-edges is not injective (Figure 7). Nonetheless the composition $\psi \circ \varphi$ allows us to compute topological invariants, such as genus and number of boundary components for protein metastructures.

1.3 Topological characteristics of proteins

We compute genera and numbers of boundary components for 10141 selected proteins from PDB ([14]; see Section 2 for details of the selection process), which do not contain β-barrels or bifurcations in β-sheets. Figure 8 shows frequency distribution of actual proteins by their genera and numbers of boundary components.

The same distribution was also computed

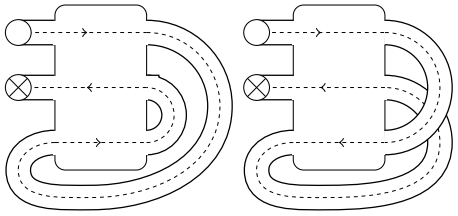


Figure 6: Thickening edges of metastructure diagrams to obtain fatgraphs (or more precisely, surfaces associated to fatgraphs). The surface on the left has genus 0, whereas the one on the right has genus 1.

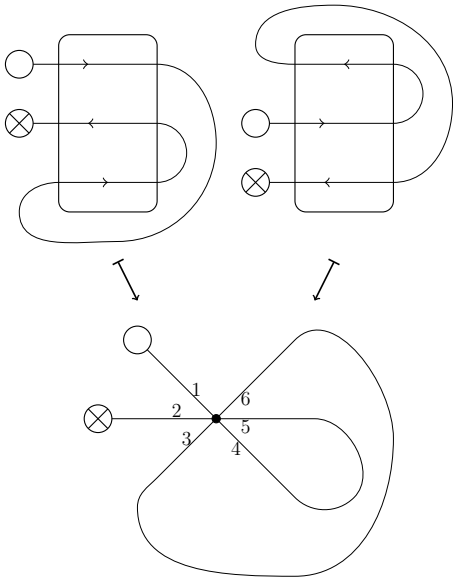


Figure 7: Construction of fatgraph from metastructure diagrams. Note the two different motifs result in an identical fatgraph.

from 10141 simulated metastructures, produced as follows;

1. Reduced sequences were generated in the following manner.
   (a) The length was chosen such that the distribution of lengths for the simulated data is the same as the distribution for the PDB data.
   (b) Each pair of letters (1st and 2nd, 3rd and 4th, and so on) was given 50% chance of being "$\gamma\alpha$" and 50% chance of being "$\gamma\beta$". If the sequence has odd number of letters, the letter "γ" was attached at the end.

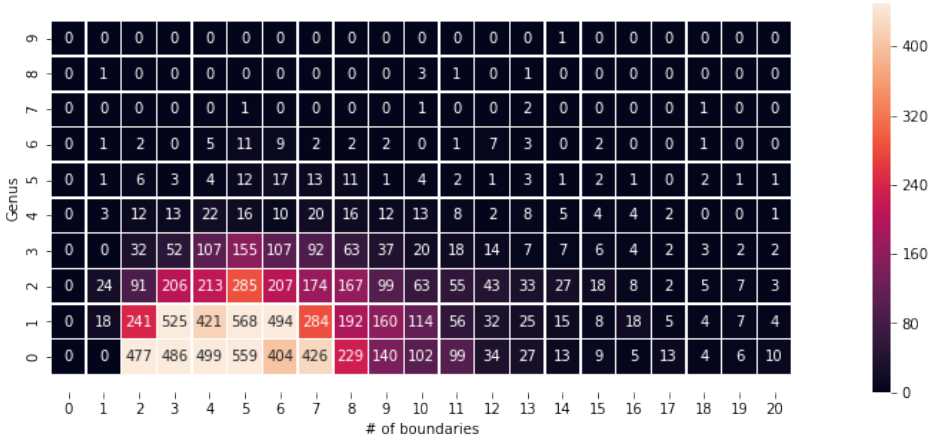


Figure 8: Frequency distribution (extract) of protein metastructures by genus and number of boundary components

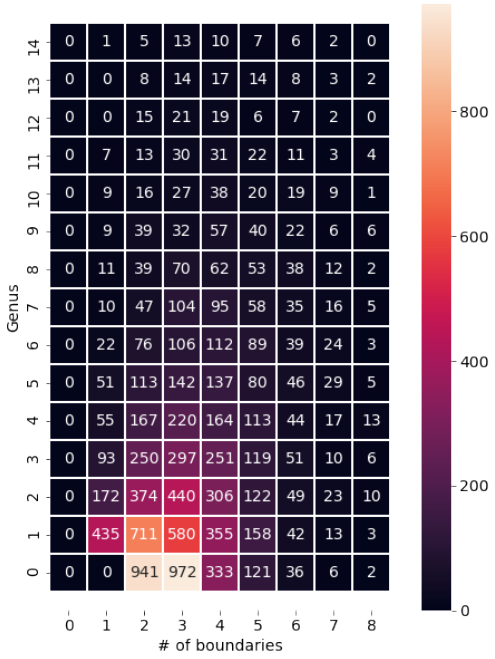


Figure 9: Frequency distribution (extract) of simulated protein metastructures by genus and number of boundary components

2. To each reduced sequence generated as above, a fatgraph structure was assigned as follows.

   (a) Let U be the set of letter βs in a given sequence, indexed with their positions in the sequence; $\beta_1, \beta_2, \ldots$. Then we partition U into a random number of subsets, each containing at least 2 elements.

   (b) For each subset U_i, choose a random ordering of β_is in the subset. This defines the ordering of strands in a beta-sheet.

   (c) For each ordered subset U_i with n_i elements, choose a random sequence of 1 and -1, of length n_i, but starting with 1. This sequence defines parallel/anti-parallel orientation of each strand with respect to the previous strand in a sheet.

We observe that the actual data tends to favour lower genera (and higher number of boundary components) compared to the simulated data (Figure 9). This implies that metastructures whose associated surfaces have lower genera are favoured over those that result in high genera in the nature. Inspired by this observation, we will develop a method for prediction of β-sheet topology using the characteristics of the protein's associated surface in Section 3.

For later use, we compute the distribution of the actual protein data by genus, num-

ber of boundary components and number of β-strands in the largest β-sheet. We call this the 3-dimensional genus-boundary distribution (see Supplementary Material).

2 Dataset

The dataset used was prepared similarly to the HQ60 dataset in [23]. Here we give a brief summary of this dataset. PISCES [33] 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 in May 2021, resulting in a set of 16262 proteins. The hydrogen bonds are taken from the DSSP program [21], with the additional conditions [13];

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

The secondary structures are also determined by DSSP, and they are recorded with three main secondary classes; [H]elix for H, G or I 8-state classes, [S]heet for E, and [C]oil for others. Thus we obtain, for each protein (of length n) in the dataset, a primary sequence $a_1a_2\cdots a_n$, where a_i is one of the 20 standard gene code amino acids, and a secondary structure sequence $b_1b_2\cdots b_n$, where $b_i = \alpha, \beta,$ or γ. We superimpose these two sequences to obtain a *hybrid sequence* $c_1c_2\cdots c_n$, where $c_i = b_i$ if $b_i = \alpha$ or β, and $c_i = a_i$ otherwise. Together with the information about hydrogen bonds, we are able to identify β-strands, their pairings and whether the pairing is parallel or anti-parallel (see Section 1.1 for details). For the purpose of the current analysis, we are only interested in proteins containing β-sheets. Furthermore, proteins containing β-barrels are excluded from the analysis. For bifurcated β-sheets, we performed the following pre-processing;

1. For each β-strand s, let p be the number of β-strands that are paired to it.
2. If $p = 3$, do the following.
3. Let s_1, s_2, s_3 be the β-strands paired to s, ordered by the number of H-bonds to/from s. Let n_1, n_2, n_3 be the number of H-bonds between s and s_1, s_2, s_3, respectively. We have $n_1 \geq n_2 \geq n_3$.
4. If $n_3 \leq n_2/2$, ignore the pairing between s and s_3.

If, after the above pre-processing, a protein still contains a bifurcated β-sheet, it is excluded from the analysis. By performing the pre-processing, the number of proteins excluded because of bifurcation was reduced from 3859 to 1946. The procedure resulted in 11853 proteins for the analysis. See Table 1 for the number of proteins in each category.

α only	1551	(9.5%)
Bifurcation	1946	(12.0%)
β-barrel	912	(5.6%)
Accepted for analysis	10141	(72.9%)
Total (HQ60)	16262	

Table 1: Number of proteins filtered from the dataset.

3 Methods

We will now describe a series of experiments to attempt to utilise the topological characteristics of protein metastructures described in Section 1.3.

3.1 Binary classification of candidate structures by their topology

200 proteins are randomly chosen for validation from the dataset, and the remaining 9941 proteins are used as the learning data. The idea is to use the learning data to decide the local configuration of β-strands, i.e. those strands, that are close to each other along the backbone. We then use the global topological data to decide the global configuration of the local blocks. We will now describe the first part of the method below. The aim is to first populate the pairing matrix $\mathbb{P}$ along the super- and sub-diagonals (i.e. the entries directly above and below the diagonal). We then repeat the procedure to populate the second entries above and below the diagonal, then the third entries, and so on. Pseudocode for populating the super- and sub-diagonals in the pairing matrix is given in Algorithm 1.

1. From each protein in the validation data, we consider its hybrid sequence and extract segments between two consecutive β-strands.

2. For each extracted segment s, compute alignment score for all segments from the learning data using the Needleman-Wunsch algorithm[1] [22].

3. Let t be the segment from the learning data with the highest alignment score. The configuration of two strands at either end of segment t (whether they are paired by hydrogen bonds, and if so whether parallel or anti-parallel) determines the configuration of two strands at either end of s.

4. Normalise the alignment score by dividing it by the score for perfect match, and record it in the appropriate entry in the pairing matrix $\mathbb{P}$. Specifically, if $p(s,t)$ is the alignment score for segments s and t, the normalised score $\tilde{p}(s,t)$ is given by $p(s,t)/p(s,s)$. Suppose s is the segment between i'th and $i+1$'th β-strands, and that they should be paired in the parallel configuration. Then set $\mathbb{P}_{(i,i+1)} = \tilde{p}(s,t)$. If, on the other hand, they should be paired in the anti-parallel configuration, set $\mathbb{P}_{(i+1,i)} = \tilde{p}(s,t)$

5. If there is a tie for the highest alignment score such that $p(s,t_1) = p(s,t_2) = \cdots = p(s,t_k)$, set $x = \#\{t_i|$the two strands at either ends of t_i are paired$\}$, where $\#S$ denotes the number of elements in a set S. Set $y = k - x$. The two strands at either end of s are paired, if and only if $x \geq y$. The parallel/anti-parallel configuration of the two strands is determined similarly.

The above procedure allows us to populate $\mathbb{P}$ along the super- and sub-diagonals. We now repeat the procedure with s being a segment containing k β-strands, $k = 1, 2, 3, \ldots$, such that s is the segment between i'th and $i+k+1$'th β-strands. We do this to populate $\mathbb{P}$ up to d entries above and below the diagonal, where

[1]For the substitution matrix we use blosum62 [20], extended by setting a match score with α or β to 4 and mismatch involving α or β to -4. See Supplementary Material for more details.

Algorithm 1 Pseudocode for populating the first diagonal in the pairing matrix $\mathbb{P}$

Let $t_1, t_2, \ldots, t_m$ be the hybrid segments between two consecutive β-strands, extracted from all proteins in the learning dataset.
Let $s_1, s_2, \ldots, s_n$ be the hybrid segments between two consecutive β-strands in a given protein in the validation data.
Let $\mathbb{P}$ by an empty $n \times n$ matrix.
for s_i in $s_1, s_2, \ldots, s_n$ **do**
  for t_j in $t_1, t_2, \ldots, t_m$ **do**
    Compute alignment score $p(s_i, t_j)$.
  end for
  Let $\tilde{j} \in \{1, 2, \ldots, m\}$ such that $p(s_i, t_{\tilde{j}}) = \max_j\{p(s_i, t_j)\}$.
  if $\tilde{j}$ is uniquely determined **then**
    if The two segments at either ends of $t_{\tilde{j}}$ are paired **then**
      Set $\tilde{p_i} = p(s_i, t_{\tilde{j}})/p(s_i, s_i)$.
      if The two segments are paired in parallel configuration **then**
        Set $\mathbb{P}_{(i,i+1)} = \tilde{p_i}, \mathbb{P}_{(i+1,i)} = 0$.
      else
        Set $\mathbb{P}_{(i+1,i)} = \tilde{p_i}, \mathbb{P}_{(i,i+1)} = 0$.
      end if
    else
      Set $\mathbb{P}_{(i,i+1)} = \mathbb{P}_{(i+1,i)} = 0$.
    end if
  else
    Let $\tilde{j_1}, \tilde{j_2}, \ldots, \tilde{j_k}$ be such that $p(s_i, t_{\tilde{j_h}}) = \max_j\{p(s_i, t_j)\}$ for all $h \in \{1, 2, \ldots, k\}$.
    Set $X = \{h|$two strands at either ends of $t_{\tilde{j_h}}$ are paired$\}$
    Set $Y = \{1, 2, \ldots, k\} \setminus X$
    if $\#X \geq \#Y$ **then**
      Set $\tilde{p_i} = p(s_i, t_{\tilde{j}})/p(s_i, s_i)$.
      Let $P \subset X$ be the subset such that the two segments at either ends of $t_{\tilde{j_h}}, h \in P$ are paired in parallel configuration.
      Let $A \subset X$ be the corresponding subset for anti-parallel configuration.
      if $\#P \geq \#A$ **then**
        Set $\mathbb{P}_{(i,i+1)} = \tilde{p_i}, \mathbb{P}_{(i+1,i)} = 0$.
      else
        Set $\mathbb{P}_{(i+1,i)} = \tilde{p_i}, \mathbb{P}_{(i,i+1)} = 0$.
      end if
    else
      Set $\mathbb{P}_{(i,i+1)} = \mathbb{P}_{(i+1,i)} = 0$.
    end if
  end if
end for

d is given by;

$$d = \begin{cases} 1 & \text{if } n < 7 \\ n - 5 & \text{if } 7 \leq n < 11 \\ 5 & \text{if } 11 \leq n. \end{cases}$$

Here the limit of 5 for d is forced by the fact that as the segments get longer, it becomes increasingly harder to obtain high alignment scores. This results in the chance of having $\mathbb{P}_{(i,j)} = 1$, in the discretisation process desribed below, being extremely small, when $|i - j| > 4$ (We were not able to get 1 in these cells in our tests). This is possibly related to the fact that the above method is essentially a method based on local data, and thus is not suited for predicting non-local configuration of β-strands. For that, another approach is needed which takes into account the global characteristics, which we will describe in the second part of the method. Before that, we need to translate the entries of the partial pairing matrix computed above, which are real numbers between 0 and 1, to either 0 or 1. We do this by changing the non-zero entries to 1, starting from the largest to the smallest. If, at any point, setting an entry to 1 results in a bifurcation or a β-barrel, the entry is set to 0 and we move onto the next largest entry (Figure 10). For later use, we name this procedure `MakeBinary()`, which takes a (partial) matrix of pairing scores as an input and returns a (partial) pairing matrix.

We now have a partial pairing matrix, populated up to d entries above and below the diagonal, without bifurcations or barrels. We populate the remaining entries by going through all possibilities, while avoiding bifurcations and β-barrels. We also require that the resulting matrix does not contain any isolated strand. The result is a number of candidate matrices, whose number depends on the partial pairing matrix computed in the first part of the method. We now construct a fatgraph from each candidate matrix, and compute its genus and number of boundary components, together with the number of strands in the largest sheet. We compare this data with the 3-dimensional genus-boundary distribution computed in Section 1.3. By a layer in the 3-dimensional genus-boundary distribution, we mean the 2-dimensional distribution of genus and number of boundary components for a specific value of number of strands in the largest sheet. Let g, n, l denote the genus, the number of boundary components and the number of strands in the largest sheet. Let $f(g, n, l)$ be the frequency of the cell (g, n, l) in the 3-dimensional genus-boundary distribution. We define the topology score $s_{\text{topo}}(\tau)$ for a metastructure τ with genus g, n boundary components and l strands in the largest sheet, by

$$s_{\text{topo}}(\tau) = \frac{f(g, n, l)}{T_l},$$

where T_l is the sum of frequencies for the lth layer. For a cutoff value $v \in (0, 1)$, a candidate metastructure τ is accepted, if $s_{\text{topo}}(\tau) \geq v$, and rejected if $s_{\text{topo}}(\tau) < v$. We also compute accuracy of each candidate structure, and look at the relationship between accuracy and acceptance of candidate structures.

3.2 Metastructure prediction by sequence alignment and topology

The method described in Section 3.1 was modified to provide a single, "best candidate" metastructure. The modification was made such that instead of classifying candidate metastructures as either accepted or rejected, a weighted sum of all candidate pairing matrices was produced, with weight given by the 3-dimensional genus-boundary distribution. More precisely, suppose a candidate pairing matrix $\mathbb{P}$ results in a structure with genus g, n boundary components and l strands in the largest sheet. Let $f(g, n, l)$ be the frequency of the cell (g, n, l) in the 3-dimensional distribution, and T_l be the sum of frequencies for the lth layer, as before. Then our final score matrix $\hat{\mathbb{P}}_{\text{score}}$ is given by

$$\hat{\mathbb{P}}_{\text{score}} = \sum_{\mathbb{P}} \frac{f(g, n, l)}{T_l} \mathbb{P},$$

where the sum is over all candidate pairing matrices for a protein. The final pairing matrix $\hat{\mathbb{P}}$ is computed from $\hat{\mathbb{P}}_{\text{score}}$ as before. A pseudocode for this procedure is shown in Algorithm 2.

3.3 Metastructure prediction by Betapro and topology

Betapro is a computer program for predicting β-sheet topology using recurrent neural net-

	1	2	3	4	5	6	7	8
1	0	0.95	0	0				
2	0	0	0	0.87	0			
3	0	0.98	0	0	0	0		
4	0	0	0	0	0	0	0	
5		0	0	0.72	0	0	0	0
6			0	0.23	0.46	0	0	0
7				0	0	0	0	0
8					0	0	0.60	0

	1	2	3	4	5	6	7	8
1	0	1	0	0				
2	0	0	0	0	0			
3	0	1	0	0	0	0		
4	0	0	0	0	0	0	0	
5		0	0	1	0	0	0	0
6			0	0	1	0	0	0
7				0	0	0	0	0
8					0	0	1	0

Figure 10: Illustration of the procedure `MakeBinary()`, which takes a partial score matrix (left) as an input and produces a pairing matrix (right). We start with the highest alignment score and set the first two, 0.98 and 0.95 to 1. The third highest, 0.87, would result in bifurcation, so it is set to 0. The next three scores are set to 1, but the last non-zero entry, 0.23 will result in a barrel involving strands 4, 5, and 6, so it is set to 0. The resulting partial pairing matrix has three blocks, listed as a set of strands, (1,2,3), (4,5,6) and (7,8). Filling this matrix by either 0 or 1 would result in $2^{20} = 1048576$ different matrices, but the restrictions placed on pairing matrices means there are only 97 valid completions.

Algorithm 2 Pseudocode for computation of prediction pairing matrix $\hat{\mathbb{P}}$.

Let $\mathbb{P}_{\text{partial}}$ be a given partial pairing matrix.
Let $\hat{\mathbb{P}}_{\text{score}}$ be a zero matrix of the same size as $\mathbb{P}_{\text{partial}}$.
for all Completion $\mathbb{P}$ of $\mathbb{P}_{\text{partial}}$ **do**
  if $\mathbb{P}$ contains a barrel, a bifurcation or an isolated strand **then**
    Move to next completion
  end if
  Compute genus g, number of boundary components n, and size of the largest sheet l for the metastructure corresponding to $\mathbb{P}$.
  Find the frequency of the cell (g, n, l) and the sum of frequencies for the lth layer T_l.
  $\hat{\mathbb{P}}_{\text{score}} = \hat{\mathbb{P}}_{\text{score}} + \frac{f(g,n,l)}{T_l}\mathbb{P}$
end for
Set $\hat{\mathbb{P}} = \textsc{MakeBinary}(\hat{\mathbb{P}}_{\text{score}})$

work (RNN) [15]. It takes a primary structure sequence as input, or a primary and secondary structure sequences, if the secondary structure is available from other sources. The output is a score matrix, where the entries are not restricted to $(0, 1)$, but positive real numbers computed as a sum of pseudoenergy for each residue pair in a β-strand pairing. The reported performances of Betapro are 0.54 for Recall and 0.59 for Precision [15].

In order to predict protein metastructure, we run Betapro using the primary and secondary structure sequences as input. From the output score matrix, we choose m entries with the highest scores, where m equals 4% of the number of entries in the score matrix, excluding the main diagonal. The entries that result in a bifurcation or a barrel, are ignored. The chosen entries are considered as β-strand pairings, and they are set to 1 in the partial pairing matrix. Next, all valid (i.e. avoiding isolated strands, bifurcations and barrels) completions of the partial pairing matrix are generated. Each completion is given two scores, one based on Betapro score matrix, and the other based on the genus-boundary distribution. The first, s_{bp}, is the sum of all scores in Betapro score matrix, where there is 1 in the pairing matrix. The second, s_{topo}, is given by $f(g, n, l)/T_l$, where g, n, l is the genus, the

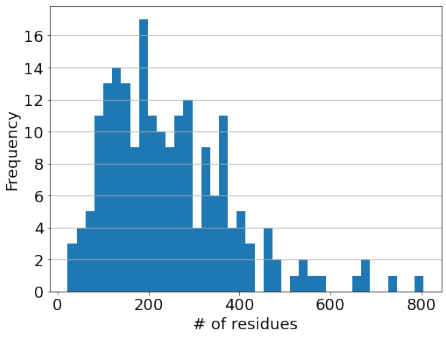
number of boundary components, and size of the largest sheet, as before. Our prediction is the structure with the highest combined score,

$$\hat{s} = as_{\text{bp}} + bs_{\text{topo}}, \tag{3}$$

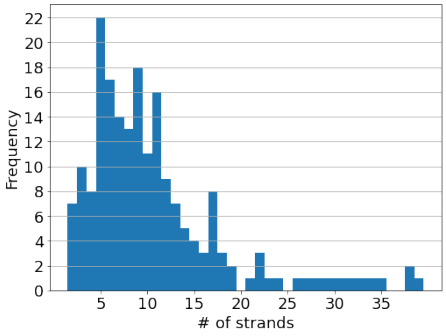
where $a, b \in [0, 1]$ with $a + b = 1$. The corresponding pseudocode is shown in Algorithm 3.

4 Results

Some of the larger proteins in the 200 test proteins could not be analysed using the method described, as there were too many possible ways to complete the pairing matrix. We therefore limit the analysis to the 181 proteins containing up to 20 β-strands. Their frequency distribution by the number of residues and β-strands is shown in Figure 11.



(a) By the number of residues



(b) By the number of strands

Figure 11: Frequency distribution of 200 proteins by the number residues (a) and by the number of strands (b).

The algorithm from Section 3.1 produced 91,431,292 candidate structures in total, but there are significant variations in the number of candidate structures per protein (Figure 12), as the possible number of candidates

Algorithm 3 Pseudocode for computation of prediction pairing matrix $\hat{\mathbb{P}}$ from Betapro score matrix $\mathbb{P}_{\text{bp}}$.

Let $\mathbb{P}_{\text{bp}}$ be the pairing score matrix produced by Betapro.
Let $\mathbb{P}_{\text{partial}}$ be an empty matrix of the same size as $\mathbb{P}_{\text{bp}}$.
Order the entries in $\mathbb{P}_{\text{bp}}$ from largest to smallest.
Set $c = 0$.
while $c \leq m$ **do**
    Let (i, j) be the index for the first element in the ordered list of entries in $\mathbb{P}_{\text{bp}}$.
    Set $\mathbb{P}_{\text{partial}(i,j)} = 1$.
    if $\mathbb{P}_{\text{partial}}$ results in a barrel or a bifurcation **then**
        Set $\mathbb{P}_{\text{partial}(i,j)} = 0$.
        $c = c - 1$.
    end if
    Remove the first element from the ordered list of entries in $\mathbb{P}_{\text{bp}}$.
    $c = c + 1$
end while
for all Completion $\mathbb{P}$ of $\mathbb{P}_{\text{partial}}$ **do**
    if $\mathbb{P}$ contains a barrel, a bifurcation or an isolated strand **then**
        Move to next completion
    end if
    Compute genus g, number of boundary components n, and size of the largest sheet l for the metastructure corresponding to $\mathbb{P}$.
    Find the frequency of the cell (g, n, l) and the sum of frequencies for the lth layer T_l.
    Set $s_{\text{topo}}(\mathbb{P}) = \frac{f(g,n,l)}{T_l}$.
    Set $\mathbb{P}_{\text{score}} = \mathbb{P} \dot{\times} \mathbb{P}_{\text{bp}}$, where $\dot{\times}$ denotes the entry-wise multiplication.
    Set $s_{\text{bp}}(\mathbb{P}) = \sum_{i,j} \mathbb{P}_{\text{score}(i,j)}$.
    Set $\hat{s}(\mathbb{P}) = as_{\text{bp}}(\mathbb{P}) + bs_{\text{topo}}(\mathbb{P})$.
end for
Set $\hat{\mathbb{P}}$ to be the completion $\mathbb{P}'$, such that $\hat{s}(\mathbb{P}') = \max\{\hat{s}(\mathbb{P}) | \mathbb{P} \text{ is a completion of } \mathbb{P}_{\text{partial}}\}$.

also depends on the partial structure determined using alignment of the α/γ segments between β-strands. In the current analysis, one protein (4UPIA) accounted for 63,907,920 candidate structures, representing 70% of the total number. Note, although some of these numbers are large, they still represent a significant reduction from the theoretically possible number of candidate structures, which is given by $n! \cdot 2^{n-2}$ for a protein with n strands, when considering only those structures with a single sheet. Naturally the numbers are even larger when considering multiple-sheet structures. We list the first few terms in Table 2.

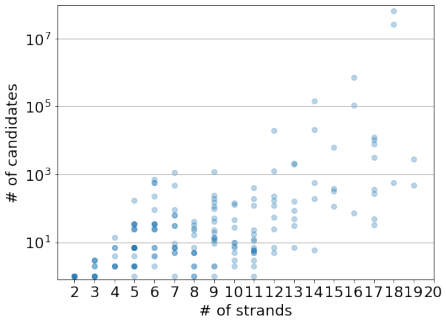


Figure 12: Number of candidate structures per protein, filtered by the number of strands. Note the log scale. There are large variations in the number of candidates among the proteins with the same number of strands.

Strands	Number of structures	
	Single sheet	Multiple sheets
2	2	2
3	12	12
4	96	108
5	960	1200
6	11520	15960
7	161280	246960

Table 2: The number of theoretically possible structures for a protein with n strands.

The topology filter, depending on the cutoff value and the number of strands, further reduces the number of candidate structures (Figure 13). Upon considering the balance between the ability to reduce the number of candidate structures and still retain high quality candidate structures, we decided to use the cutoff s_{topo} value of 0.02 for the subsequent analysis. The actual number of accepted candidate structures are shown in Figure 14. As we also can see from Figure 13, the topology filter is very effective at reducing the number of structures for proteins with larger number of strands (i.e. large number of candidates). In the current analysis, 3 proteins accounted for 99.6% of all candidate structures. For these the topology filter reduced the number of candidates by 92-97% (Table 3). When using any positive cutoff value for such a filter, there is a chance that no candidate structure for a protein is accepted. If it happens, we reduce the cutoff value only for the proteins with no accepted candidate structure, until one or more candidate structures are accepted. In the current analysis, the cutoff values were reduced by 0.005 down to 0.005. If, at the end of this iteration, we have proteins with no accepted structure, we randomly select one candidate structure for acceptance. This procedure, however, was not necessary for the current analysis, and all proteins had at least one candidate structure accepted at the cutoff value of 0.02.

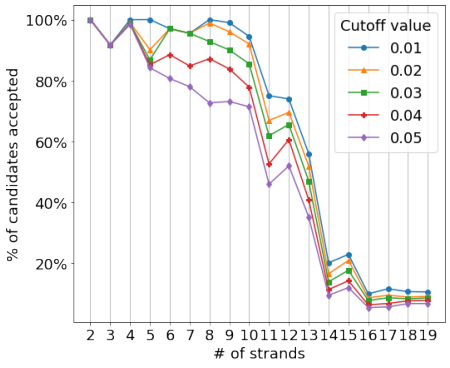


Figure 13: Percentages of accepted structures by cutoff values and the number of strands.

# candidates	# accepted	% accepted
63907920	4889934	7.7%
26437952	631882	2.4%
731584	20521	2.8%

Table 3: The numbers and percentages of accepted structures for the three proteins with most candidate structures, accounting for more than 99% of all candidates. The topology filter rejects more than 90% of candidates.

In order to examine how well our topological

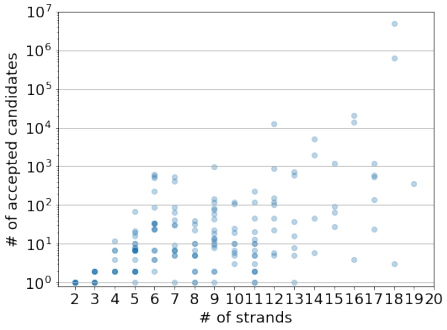


Figure 14: Number of accepted candidate structures per protein, filtered by the number of strands. Note the log scale. Compared to Figure 12, the numbers are significantly lower where there are large number ($> 10^4$) of candidate structures.

filter distinguishes between "good" and "bad" candidates, we investigate how the rate of acceptance changes for "good" and "bad" candidate structures. Precision and Recall are two measures often used for judging quality of predicted protein structures. They are given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP and FN stand for the number of true positive, false positive and false negative strand pairings.

For each target and for a given quality measure Q (=Precision or Recall), we divide the candidate structures into three classes; low quality (structures with $Q < 0.6$), medium quality ($0.6 \leq Q < 0.9$), and high quality ($0.9 \leq Q$). We then compute the acceptance rate for each class. The results are shown in Table 4. The acceptance rates increase with an increase in the quality levels.

Quality	Precision	Recall
Low	66.33%	65.45%
Medium	70.71%	72.04%
High	90.80%	89.62%

Table 4: Average acceptance rate by quality classes.

Metastructure prediction by sequence alignment and topology (Section 3.2) and by Betapro and topology (Section 3.3) were performed on the same set of proteins. The average Precision and Recall for the predictions are shown in Table 5. Different values of a in the combined score function (3) only had a very small effect (< 0.005) on Precision or Recall values (Table 5). The strand pairing scores from Betapro are strictly positive, potentially promoting the formation of large sheets which are topologically complex. To mitigate this, we applied logarithm to the strand pairing scores from Betapro and used them in the algorithm. This resulted in an increase in Precision but a (smaller) decline in Recall (Table 5). This change was seen across different number of strands (Figure 15). To investigate the effect of the number of selected pairings before computing completions, we ran the algorithm using 4, 5, and 6% for pre-selection, together with the "fewest possible" pre-selections, which is the number where a computation is possible within a reasonable amount of time (24 hours on a modern cpu). The number p of pre-selected pairs for a protein with n strands was;

$$p = \begin{cases} 0 & \text{if } n \leq 8 \\ n - 8 & \text{if } 9 \leq n \leq 11 \\ n - 7 & \text{if } 12 \leq n \leq 20 \end{cases}$$

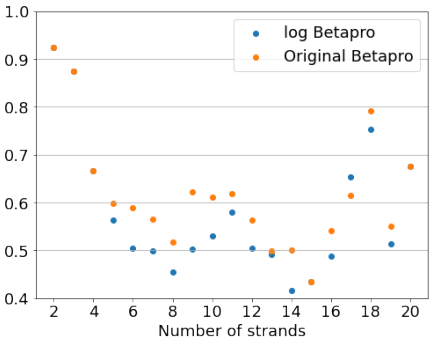
The results are shown in Table 6. Preselecting more pairings should have the effect of increasing false positive (FP) and decreasing false negative (FN), thereby reducing Precision and increasing Recall, which we observe here.

	Precision	Recall
Alignment	0.42	0.47
Betapro, a=0.1	0.56	0.62
Betapro, a = 1	0.56	0.62
logBetapro	0.67	0.57

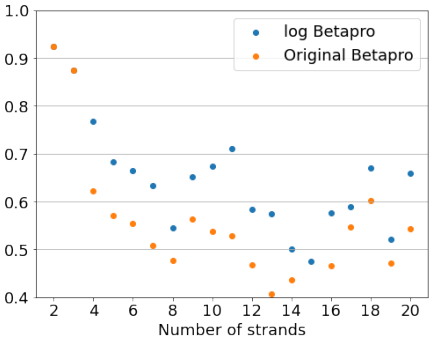
Table 5: Average Precision and Recall for different metastructure prediction methods.

	Fewest	4%	5%	6%
Precision	0.667	0.667	0.659	0.650
Recall	0.567	0.567	0.572	0.577

Table 6: Average Precision and Recall for different levels of pre-selection.



(a) Average Recall



(b) Average Precision

Figure 15: Average Recall (a) and Precision (b) by number of strands, Betapro and logBetapro scores.

5 Discussion

The difference in the distributions of the genera and the numbers of boundary components from the actual (Figure 8) and simulated data (Figure 9) indicate that the folding of β-sheets is not a completely random process. Indeed, it does appear that an increase in genus is costly and a structure that has lower genus is favoured over one with higher genus. This observation agrees with previous studies, which do not look at genus of β-sheets, but finds that certain β-sheet structures, many of which correspond to an increase in genus, are absent or very rare in proteins [29, 34]. The result of our binary classification analysis (Section 3.1) agrees with this observation. Even though the result is skewed by a highly uneven distribution of the number of candidate structures per protein, and the response of acceptance rate for an increase in quality is not linear, it does appear that the topology of protein metastructure captures some information about the native structure. Extending this result to prediction of metastructures proved more challenging. We did achieve a result comparable to that reported for Betapro when using strand-pairing scores as is, which was improved to be better than Betapro with an application of logarithm to the pairing scores. This is likely to be because the unprocessed Betapro scores are strictly greater than zero, thus encouraging formation of larger sheets in order to maximise the final score $\hat{s}$, even though the contribution from the topology score s_{topo} should, to some extent, prevent the formation of sheets that are too large and topologically complex. By applying logarithm to the Betapro scores, we encourage fewer pairings (and thus discourage large sheets), which resulted in improved Precision. We were, however, not able to outperform the figures reported by other, more recent studies (Table 7). The structure of the BCov and BetaProbe programs meant that it was not possible to combine them with our method in a similar manner to Section 3.3. It would be interesting to see if one can improve the results of Top-DBS program by combining with our method. Unfortunately the program code for Top-DBS was not available for inspection.

Program	Precision	Recall
Betapro [15]	0.59	0.54
BCov [30]	0.60	0.62
BetaProbe [18]	0.67	0.70
Top-DBS [17]	0.75	0.78
Current Study	0.67	0.57

Table 7: Comparison of Precision and Recall values for prediction of β-sheet topology.

One of the reasons why the results from our study could not match those from more recent studies may be that the topology filter, in its current form, is too coarse. Suppose we have a protein with three β-strands. There are 12 different protein metastructure configurations possible, but 8 of them have genus 0 and 3 boundary components, with the rest having genus 1 and 1 boundary component. This suggests a "finer" filter, which can distinguish between the structures having the same genus and number of boundary components (and maximum sheet size), may be able to produce a better result. However, with the size of the currently available dataset, making the filter finer would result in the frequency

in each cell being too small for sampling the distribution of genera or numbers of boundary components (or some other topological data).

The term β-sheet topology is commonly used to describe the configuration of β-strands in a β-sheet. However, to our knowledge, it has not been studied in relation to topological invariants. We have shown in this paper that the topological invariants such as genus and the number of boundary components can describe certain aspects of β-sheet topology of proteins, and how they might be used in prediction of β-sheet topologies. We believe the protein metastructure and topology of the associated fatgraph have a potential to provide a simpler, more mathematically natural way to analyse β-sheet topology.

Acknowledgement

This paper is partly a result of the ERC-SyG project, Recursive and Exact New Quantum Theory (ReNewQuantum) which received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810573. This work was supported by JSPS KAKENHI Grant Number JP20K03931, JP20K03601, JP18K03281.

References

[1] Alexeev, Nikita, Andersen, Jørgen Ellegaard, Penner, Robert C., and Zograf, Peter. "Enumeration of chord diagrams on many intervals and their non-orientable analogs". *Advances in Mathematics* 289 (Feb. 2016), pp. 1056–1081. ISSN: 0001-8708.

[2] Andersen, J. E., Huang, F. W. D., Penner, R. C., and Reidys, C. M. "Topology of RNA-RNA interaction structures". *Journal of Computational Biology* 19.7 (2012), pp. 928–943.

[3] Andersen, J. E., Penner, R. C., Reidys, C. M., and Waterman, M. S. "Topological classification and enumeration of RNA structures by genus". *Mathematical Biology* 67 (2013), pp. 1261–1278.

[4] Andersen, Jørgen E, Chekhov, Leonid O., Penner, Robert C, Reidys, Christian, and Sułkowski, Piotr. "Enumeration of RNA complexes via random matrix theory". *Biochemical Society. Transactions* 41.2 (2013), pp. 652–655. ISSN: 0300-5127.

[5] Andersen, Jørgen Ellegaard, Chekhov, Leonid O., Penner, Robert, Reidys, Christian M., and Sułkowski, Piotr. "Topological recursion for chord diagrams, RNA complexes, and cells in moduli spaces". *Nuclear Physics, Section B* 866.3 (2013), pp. 414–443. ISSN: 0550-3213.

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

[7] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, Manabe, Masahide, Penner, Robert C., and Sułkowski, Piotr. "Enumeration of chord diagrams via topological recursion and quantum curve techniques". *Travaux Mathematiques, University of Luxembourg* 25 (Mar. 2017), pp. 285–323.

[8] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, Manabe, Masahide, Penner, Robert C., and Sułkowski, Piotr. "Partial Chord diagrams and Matrix models". English. *Travaux Mathematiques, University of Luxembourg* 25 (Mar. 2017), pp. 233–283.

[9] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, Penner, Robert C., and Reidys, Christian. *The boundary length and point spectrum enumeration of partial chord diagrams using cut and join recursion*. WorkingPaper. arXiv.org, Dec. 2016.

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

[11] Andersen, Jørgen Ellegaard, Koyanagi, Yuki, Nielsen, Jakob Toudahl, and Villemoes, Rasmus. "Prediction of H-bond rotations from protein H-bond topology". Preprint.

[12] Aydin, Zafer, Altunbasak, Yucel, and Erdogan, Hakan. "Bayesian Models and Algorithms for Protein β-Sheet Prediction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 8.2 (2011), pp. 395–409.

[13] 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.

[14] 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.

[15] Cheng, Jianlin and Baldi, Pierre. "Three-stage prediction of protein ß-sheets by neural networks, alignments and graph algorithms". *Bioinformatics* 21.suppl-1 (June 2005), pp. i75–i84. ISSN: 1367-4803.

[16] Chothia, Cyrus, Hubbard, Tim, Brenner, Steven, Barns, Hugh, and Murzin, Alexey. "Protein folds in the all-β and all-α classes". *Annual review of biophysics and biomolecular structure* 26.1 (1997), pp. 597–627.

[17] Dehghani, Toktam, Naghibzadeh, Mahmoud, and Sadri, Javad. "Enhancement of Protein β-sheet Topology Prediction using Maximum Weight Disjoint Path Cover". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.6 (2018), pp. 1936–1947.

[18] Eghdami, Mahdie, Dehghani, Toktam, and Naghibzadeh, Mahmoud. "BetaProbe: A probability based method for predicting beta sheet topology using integer programming". In: *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE. 2015, pp. 152–157.

[19] Fonseca, Rasmus, Helles, Glennie, and Winter, Pawel. "Ranking beta sheet topologies with applications to protein structure prediction". *Journal of Mathematical Modelling and Algorithms* 10.4 (2011), pp. 357–369.

[20] Henikoff, Steven and Henikoff, Jorja G. "Amino acid substitution matrices from protein blocks". *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919.

[21] 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.

[22] 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.

[23] 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).

[24] 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).

[25] 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.

[26] Reidys, C. M., Huang, F. W. D., Andersen, J. E., Penner, R. C., Stadler, P. F., and Nebel, M. E. "Topology and prediction of RNA pseudoknots". *Bioinformatics* 27.8 (2011), pp. 1076–1085.

[27] Richardson, Jane S. "β-Sheet topology and the relatedness of proteins". *Nature* 268.5620 (1977), pp. 495–500.

[28] Richardson, Jane S. "Handedness of crossover connections in beta sheets". *Proceedings of the National Academy of Sciences* 73.8 (1976), pp. 2619–2623.

[29] Ruczinski, Ingo, Kooperberg, Charles, Bonneau, Richard, and Baker, David. "Distributions of beta sheets in proteins with application to structure prediction". *Proteins: Structure, Function, and Bioinformatics* 48.1 (2002), pp. 85–97.

[30] Savojardo, Castrense, Fariselli, Piero, Martelli, Pier Luigi, and Casadio, Rita. "BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming". *Bioinformatics* 29.24 (2013), pp. 3151–3157.

[31] Sternberg, MJE and Thornton, JM. "On the conformation of proteins: The handedness of the connection between parallel β-strands". *Journal of molecular biology* 110.2 (1977), pp. 269–283.

[32] Subramani, Ashwin and Floudas, Christodoulos A. "β-sheet topology prediction with high precision and recall for β and mixed α/β proteins". *PloS one* 7.3 (2012).

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

[34] Zhang, Chao and Kim, Sung-Hou. "The anatomy of protein β-sheet topology". *Journal of molecular biology* 299.4 (2000), pp. 1075–1089.

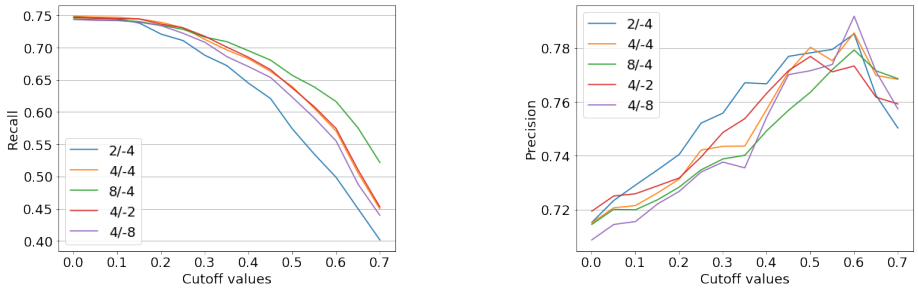
Supplementary Material

S1 Extension of Blosum62

The sequences we align consists of 20 letters representing standard gene code amino acids and two extra letters, α and β. We therefore need to extend the blosum62 substitution matrix to include scores for these two extra letters. We use 4 and -4 respectively for match and mismatch involving α and β. We investigated the effect of these scores by computing average Recall and Precision for different scores as follows;

1. For each match/mismatch score combination, compute alignment scores for the first and second diagonal (i.e. for the sequences involving zero or one β-strand).
2. Let v be a number between 0 and 1. For each cell $\mathbb{P}_{(i,j)}$ in the pairing matrix $\mathbb{P}$, where the alignment scores have been computed, set the value to 1 if $\mathbb{P}_{(i,j)} > v$, 0 otherwise.
3. Compute Recall and Precision for each protein.

The average Recall and Precision for various cutoff values and score combinations are shown in Figure S1. We see the variation in Precision is very small across different score combinations. The same holds for Recall, for small cutoff values. We also note that, compared to the average Recall, the average Precision does not vary much for different cutoff values. We therefore choose the cutoff value of 0 and match/mismatch score combination of 4 and -4 for the current analysis.



(a) Average Recall

(b) Average Precision

Figure S1: Average Recall (Figure S1a) and Precision (Figure S1b) for various cutoff values. The different lines represent different match/mismatch score combinations.

S2 The topology filter

The topology filter, as used in the paper, is the distribution of genera and numbers of boundary components for the protein metastructures filtered by the number of strands in the largest sheet. The filtering was done to be able to reflect the difference in the distributions between proteins with a small number of strands and those containing more strands, as one would expect those proteins with a large number of strands to have more (topologically) complex structures. The number of strands in the largest sheet was chosen as the filtering variable, because we expect the largest contribution to the genus (and the number of boundary components) to come from the largest sheet. We show the distributions up to the maximum sheet size of 10 in Figure S2. We also show the distributions of the same data, filtered by the number of strands Figure S3 and by the number of sheets Figure S4. It was thought that filtering by the number of sheets results in too few layers and will make the resulting topology filter less powerful, as it will not be able to distinguish subtler differences. To investigate whether filtering by the number of strands produces better results, we ran the binary classification (see Section 3.1) using the topology filter with the maximum sheet size as the third axis, and one with the number of strands as the third axis. The classification results were analysed by computing the proportion of the candidate structures above certain quality thresholds, that were accepted (Figure S5). For a good filter, we expect the percentages of acceptance to increase, as we restrict to candidate structures to only look at the high-quality structures. So we expect the lines to lie diagonally from the bottom-left to top-right. It was found the filter using the maximum sheet size performed better, particularly with Recall. We did not investigate why it is the case, but it may be that folding several large sheets into energetically favourable structure is complex, and in nature a combination of one large sheet and several smaller ones is more common.

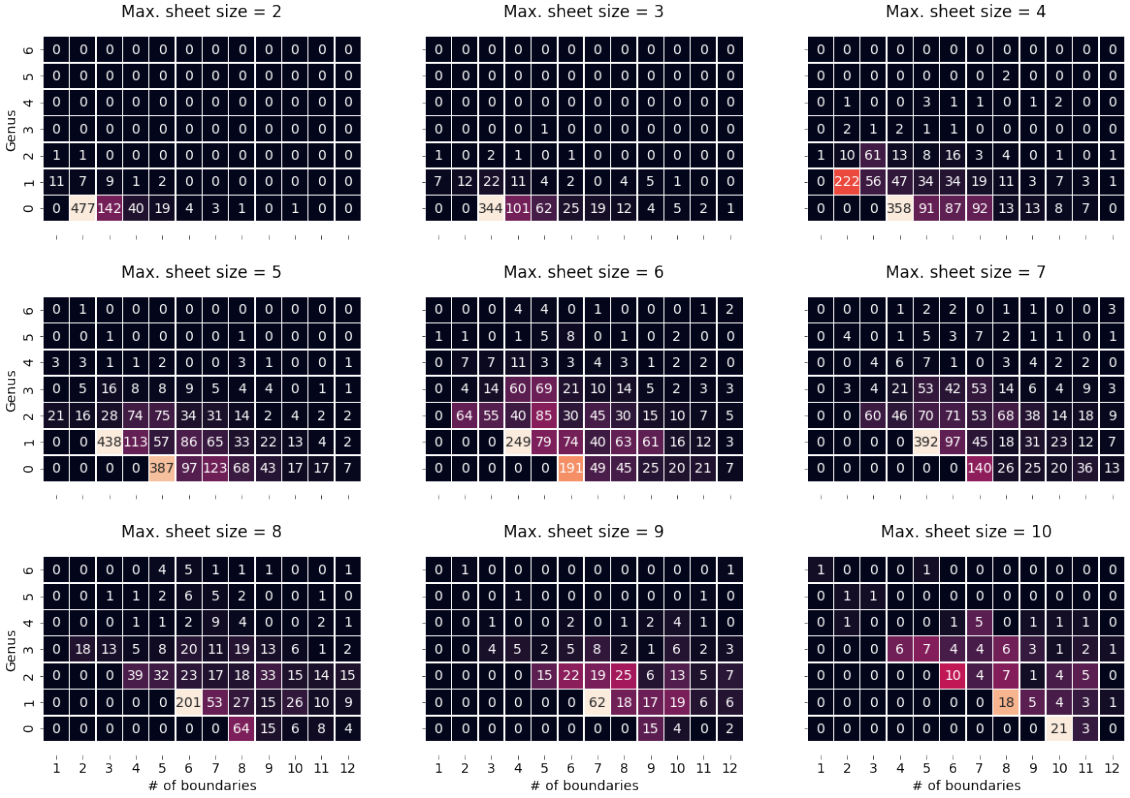


Figure S2: The distribution of genus and number of boundary components, filtered by the size of the largest sheet.

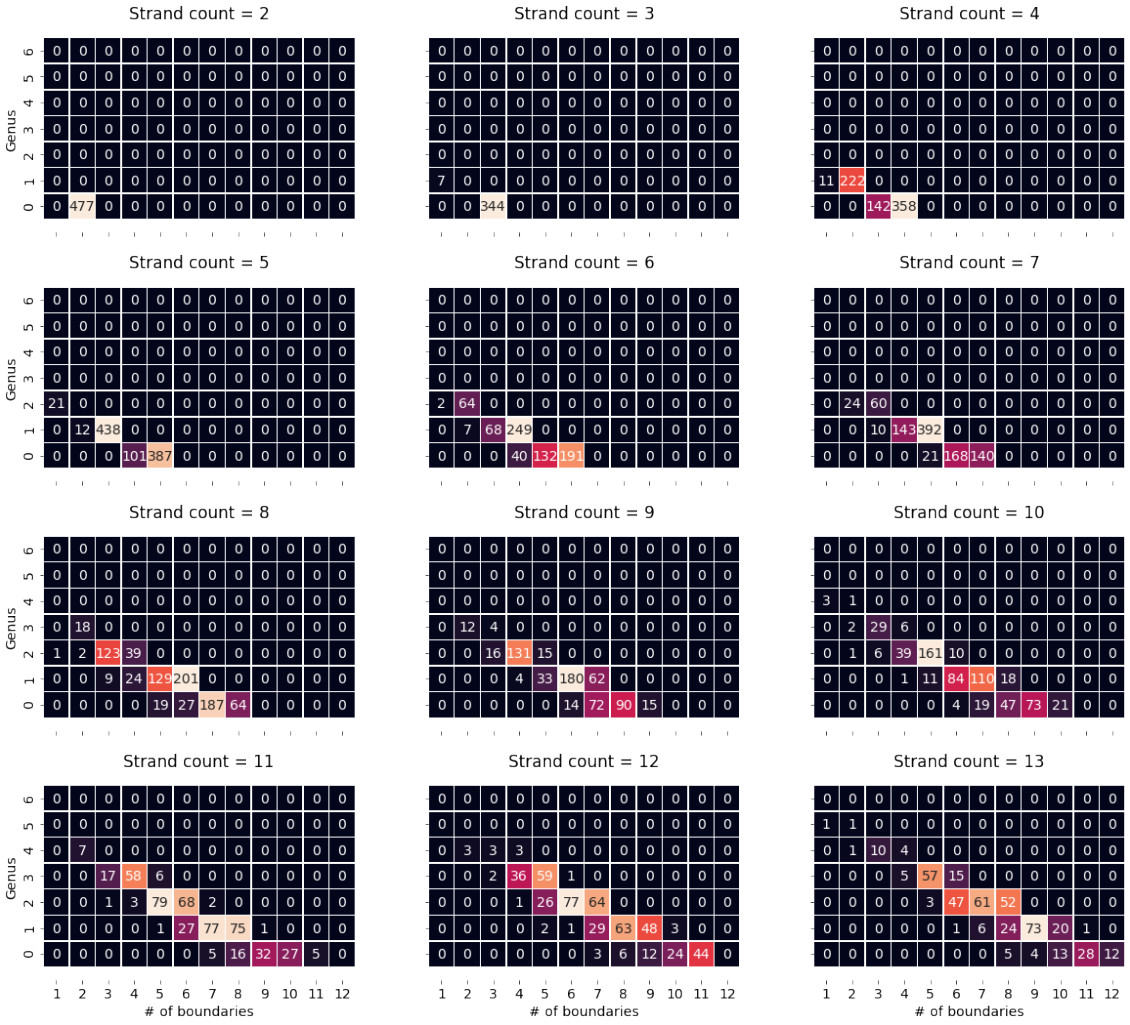


Figure S3: The distribution of genus and number of boundary components, filtered by the number of strands.

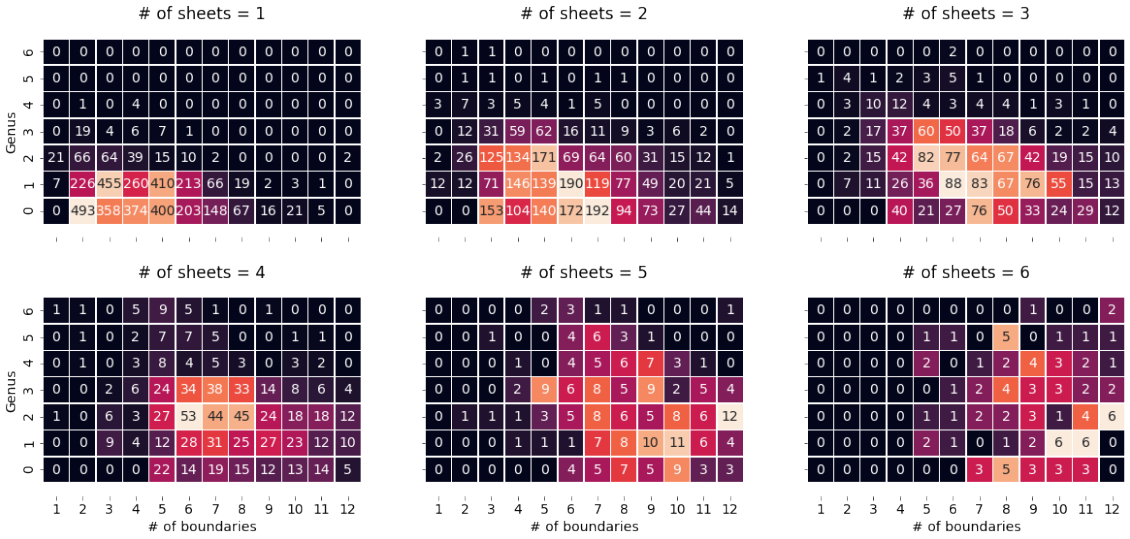
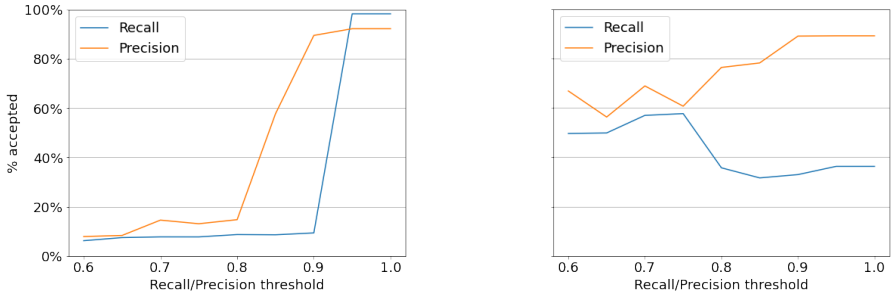


Figure S4: The distribution of genus and number of boundary components, filtered by the number of sheets.



(a) Maximum sheet size

(b) Number of strands

Figure S5: Acceptance rates for candidates above quality thresholds (measured in Recall and Precision). The topology filter with maximum sheet size (Figure S5a) shows increasing acceptance rates when the candidates are restricted to high-quality structures. On the other hand, the filter with number of strands (Figure S5b) shows relatively high acceptance rates for lower-quality candidates, and the rate drops for Recall, when the candidates are restricted to high-quality structures.