Sequence-structure-function relationships in class I MHC: a local frustration perspective

Short Title: Biophysical basis of MHC I polymorphism

Onur Serçinoğlu[1], Pemra Ozbek[2]

[1]Recep Tayyip Erdogan University, Faculty of Engineering, Department of Bioengineering, 53100, Fener, Rize, Turkey, e-mail: onur.sercinoglu@erdogan.edu.tr

[2]Marmara University, Faculty of Engineering, Department of Bioengineering, 34722, Goztepe, Istanbul, Turkey, e-mail: pemra.ozbek@marmara.edu.tr

*Corresponding author: pemra.ozbek@marmara.edu.tr

Abstract

Class I Major Histocompatibility Complex (MHC) binds short antigenic peptides with the help of Peptide Loading Complex (PLC), and presents them to T-cell Receptors (TCRs) of cytotoxic T-cells and Killer-cell Immunglobulin-like Receptors (KIRs) of Natural Killer (NK) cells. With more than 10000 alleles, the Human Leukocyte Antigen (HLA) chain of MHC is the most polymorphic protein in humans. This allelic diversity provides a wide coverage of peptide sequence space, yet does not affect the three-dimensional structure of the complex. Moreover, TCRs mostly interact with pMHC in a common diagonal binding mode, and KIR-pMHC interaction is allele-dependent. With the aim of establishing a framework for understanding the relationships between polymorphism (sequence), structure (conserved fold) and function (protein interactions) of the MHC, we performed here a local frustration analysis on pMHC homology models covering 1436 HLA I alleles. An analysis of local frustration profiles indicated that (1) variations in MHC fold are unlikely due to minimally-frustrated and relatively conserved residues within the HLA peptide-binding groove, (2) high frustration patches on HLA helices are either involved in or near interaction sites of MHC with the TCR, KIR, or Tapasin of the PLC, and (3) peptide ligands mainly stabilize the F-pocket of HLA binding groove.

Author Summary

A protein complex called the Major Histocompatibility Complex (MHC) plays a critical role in our fight against pathogens via presentation of antigenic peptides to receptor molecules of our immune system cells. Our knowledge on genetics, structure and protein interactions of MHC revealed that the peptide-binding groove of Human

Leukocyte Chain (HLA I) of this complex is highly polymorphic and interacts with different proteins for peptide-binding and presentation over the course of its lifetime. Although the relationship between polymorphism and peptide-binding is well-known, we still lack a proper framework to understand how this polymorphism affects the overall MHC structure and protein interactions. Here, we used computational biophysics methods to generate structural models of 1436 HLA I alleles, and quantified local frustration within the HLA I, which indicates energetic optimization levels of contacts between amino acids. We identified a group of minimally frustrated and conserved positions which may be responsible for the conserved MHC structure, and detected high frustration patches on HLA surface positions taking part in interactions with other immune system proteins. Our results provide a biophysical basis for relationships between sequence, structure, and function of MHC I.

Introduction

The sequence-structure-function paradigm plays a central role in structural biology: the primary structure (i.e. amino acid sequence) of a protein dictates the three-dimensional structure (fold), which in turn influences the function [1–3]. The close relationship between the structural architecture and function of a protein implies that disruptions to the native folded state can destabilize the protein structure, and eventually cause loss of function. In line with this perspective, several studies integrating sequence evolution with protein biophysics reported strong correlations between positional conservation levels or rates of synonymous/non-synonymous mutation and stability-related residue metrics such as solvent accessibility or hydrogen bonding patterns [4–11]. On the other hand, stability is not the main determinant of protein function. Residues that are not involved in forming the protein scaffold, such as the catalytic sites of an enzyme or binding hot spots located on protein surface, are important for protein function as well.

Protein evolution is thus driven by an interplay between functional and structural constraints [2,10,12,13]. With these constraints at play, mutations may occur via two alternative mechanisms: positive and negative selection [14]. In the negative (or purifying) selection, mutations leading to detrimental effects on protein stability or function are eliminated during the selection process, leading to increased levels of conservation at sites crucial for function or stability. In the positive selection model, the protein is under constant pressure to acquire “new-functions”, thus change/variation in the amino acid sequence is favored, especially at sites with direct relevance to protein function [15].

From this perspective of molecular evolution and protein function, the human class I Major Histocompatibility Complex (MHC I) is an interesting system [16–19]. MHC I consists of heavy (H) and light (L) chains called Human Leukocyte Antigen (HLA I) and β-2 microglobulin (β-2-m), respectively (Fig 1). Following assembly within the Endoplasmic Reticulum (ER), the MHC is unstable in the absence of a peptide ligand [20,21]. Intracellularly derived antigenic peptides originating from self or foreign proteins are loaded into the binding groove of the HLA chain of MHC with the help of a multiprotein complex called the Peptide Loading Complex (PLC). Here, the peptide contacts six different binding pockets (A,B,C,D,E and F) within the peptide binding groove of HLA [22]. Upon the formation of a stable peptide-loaded MHC (pMHC), the complex is transported to cell surface, and presents peptide ligands to immune cell receptors such as the T-Cell Receptor (TCR) of cytotoxic CD8+ T-cells, and Killer-cell Immunoglobulin-like Receptors (KIRs) of Natural Killer (NK) cells [23–26].

Fig 1. The three-dimensional structure of the pMHC.

With more than 10000 identified protein alleles, HLA I is the most polymorphic protein in humans [27]. The HLA I protein is clearly undergoing a process of positive selection, where the acquisition of new variants modulate the function of antigen presentation via modifying the peptide ligand recognition specificities [16,28–31]. In other words, functional constraints require HLA to maintain a very high level of allelic diversity. This allelic diversity helps the immune system cover a large space of potential peptide binders, and thereby fight against pathogens effectively [16,19,24,32]. On the other hand, HLA genes are also strongly linked to infectious and autoimmune diseases due to high polymorphism levels [33].

Despite this high sequence variation in HLA and the enormous diversity of peptide ligands, experimentally elucidated pMHC structures of different alleles display an “ultra-conserved” fold with minimal variation [34]. Moreover, most of the TCR-pMHC crystal structures display a conventional docking mode of TCR over the HLA chain [35,36]. On the other hand, HLA allelic diversity may affect interactions of the pMHC with other proteins as well. Only certain allele groups containing specific epitopes interact with KIR [25,37]. Structural details of the PLC-pMHC interaction were also only recently revealed [34,38–41].

HLA polymorphism should be taken into account in order to truly comprehend the mechanisms involved in the formation of stable pMHC and interactions with TCR and KIR molecules. However, incorporating such extreme levels of polymorphism into wet-lab experiments is unfeasible. Hence, computational biophysics methods are preferred in large-scale modeling studies. In particular, homology modeling was used to classify a high number of HLA alleles into distinct groups based on peptide interaction patterns [42], binding pocket similarities [43] and surface electrostatics [44]. These studies mainly focused on classifying HLA alleles based on their peptide-binding or protein interaction behaviors, and therefore only the peptide-binding groove was modeled. However, characterization of the relationship between HLA polymorphism and pMHC stability as well as TCR/KIR/PLC interactions requires the modeling of the whole complex and the use of proper methods for identifying residue-level effects on stability and protein interactions. To this end, quantification of local frustration in the structure can be utilized [45–49]. Local frustration analysis has already been applied successfully to perform similar sequence-structure-function studies on calmodulin

116 [50,51], repeat proteins [5,52], and TEM beta-lactamases [6].

117 Here, we provide an analysis of local frustration patterns of 1436 HLA I alleles.
118 We explain how class I MHC retains a conserved fold while maintaining highly versatile
119 ligand specificities using homology-modeled pMHC structures. Using the frustration
120 data, we also show the existence of local frustration-based energetic footprints of
121 polymorphism, providing a biophysical basis for the previously observed differences
122 between molecular stabilities/cell surface expression levels of different allele groups
123 and TCR/KIR/PLC interactions. Finally, we also provide a local frustration based
124 explanation of how peptides stabilize peptide binding pockets.

125 ## Results and Discussion

126 ### Sequence variation in HLA binding groove

127 We began with an analysis of sequence variation in the HLA I peptide binding
128 groove. A total of 8696 HLA I binding groove (α-1 and α-2 domains) sequences
129 (including 2799, 3433 and 2464 alleles from HLA-A, -B, and -C gene loci, respectively)
130 were included in the analysis. We identified amino acid variation at each respective
131 position in the HLA peptide binding groove by constructing sequence logos (Fig 2A).
132 In line with findings of a recent analysis [53], we detected high levels of variation at
133 most binding groove positions. However, some positions were relatively conserved,
134 and dominated by a single amino acid. This dominance is due to the imbalance
135 between the number of occurrences of the respective amino acid and those of the
136 others. Glycine, phenylalanine, proline, tryptophan, leucine, aspartic acid and arginine
137 were found to be the most conserved amino acid types.

Fig 2. Sequence conservation/variation and evolutionary importance of HLA I peptide-binding groove positions

(A) Sequence logo of the HLA I peptide-binding groove (residues 1-180). Polar, neutral, basic, acidic and hydrophobic amino acids are colored green, purple, blue, red, and black, respectively. (B) real-value Evolutionary Trace (rvET) scores of binding groove positions. Low rvET scores indicate high evolutionary importance, and vice versa.

The sequence logos also display several “hyper-variable” positions as well (positions 9, 24, 45, 67, 97, 116, 138, 152, 156, and 163). This variation is expected, since all of these positions are located within the peptide binding pockets, and were previously shown to define allele-specific peptide ligand repertoires [42,43,54].

We also quantified the evolutionary importance of each position in the binding groove by computing the real-value Evolutionary Trace (rvET) ranks per position [55]. rvET is an absolute rank of a given position in terms of its evolutionary importance. Here, lower rvET ranks/scores indicate higher evolutionary importance and vice versa. The ET analysis here is based on both sequence variation/conservation and the closeness of sequence divergence to the root of the constructed phylogenetic three at a given sequence position. Thus, while conserved positions in a multiple sequence alignment tend to have low rvET scores, relatively less conserved positions may also obtain low rvET scores as well, provided that the sequence divergence occurs near the root of the tree and variation occurs within small rather than large branches of the tree [55–58]. In general, we observed higher rvET values at positions with the highest sequence variation and vice versa (Fig 2B).

Homology modeling of HLA alleles in the context of pMHC

Next, we investigated how binding groove sequence variation is reflected on the pMHC structure. As reported previously [34,43,44,59], the number of experimentally determined pMHC structures with different HLA alleles is significantly lower than the total allele number. Hence, a homology modeling approach is necessary. Moreover, homology models of the full pMHC, including the binding groove as well as distant β-2-m and α-3 domains, are needed, since these domains make extensive contacts with each other as well as the rest of the structure, and were previously shown to be essential for peptide binding [60,61]. However, the α-3 domain sequence is not available for many alleles. Moreover, it is also necessary model peptides within the binding groove as peptide ligand is an integral part of the pMHC structure. Similar to limitations in pMHC structures, the number of identified peptide ligands with binding affinity measurements for individual HLA alleles is limited as well, with some alleles having no peptide ligands identified so far [62]. Therefore, it is also necessary to predict peptide ligands using computational approaches. We thus selected 1436 HLA alleles (464, 689 and 283 from HLA-A, -B and -C loci, respectively) with complete HLA sequence (including all three domains α-1, α-2, and α-3) for homology modeling (the complete list is given in Supplementary Table 1). The list of homology modelled alleles included 41 out of 42 core alleles representing the functionally significant sequence variation [53]. We predicted up to 10 strong binder peptides for each of the 1436 alleles using netMHCpan 3.0 [63], and obtained homology models in the context of these peptides.

Integrating biophysics into HLA I evolution

After generating the homology models, we analyzed the local frustration within pMHC structures. Local frustration analysis is based on calculation of pairwise contacts between amino acids, and comparison of these contacts to possible alternative contacts made by other amino acid pairs at each site in the structure. A Single Residue Frustration Index (SRFI) is then obtained as a position-specific local frustration score: amino acids with optimized energetic contacts are minimally frustrated, where those that are the least preferred at their respective positions are highly frustrated. If neither, then the frustration is termed “neutral”. SRFI values thus indicate how ideal (minimally frustrated) the contacts of each position are or how much frustration is present (highly frustrated).

We used the *frustratometer2* tool [64] to quantify SRFI at each position in homology models. Since multiple peptides (and hence structures) were modeled for each allele, we calculated median SRFI per allele per position, and used these median SRFI values in further analyses.

In order to get an overview of local frustration against evolutionary importance of each position, we first mapped position-specific median SRFI and rvET values onto the HLA I binding groove (Fig 3). SRFI distributions of several selected positions are also given in Fig 4A. Minimal frustration was observed at conserved positions located in the β-sheet floor of the binding groove or in α-1 and α-2 domains contacting the β-sheet floor (5, 7, 8, 25, 27, 34, 36, 101, and 164) (Fig 4A). As expected, the top two minimally frustrated positions were cysteines responsible for forming the conserved disulfide bridge between positions 101 and 164. Mutations at these positions abolish

206 HLA expression [27,65]. Our observation of minimal frustration at conserved positions
207 are in line with recent findings of Dib et al. [66], where co-evolving residues were
208 shown to avoid residues important for protein stability. Haliloglu et al. previously
209 analyzed several HLA structures using the Gaussian Network Model (GNM), and
210 identified positions 6, 27, 101, 103, 113, 124 and 164 as “energetically active” and
211 possibly important for stability [67]. Here, either the same positions or their sequential
212 neighbors of were found to be minimally frustrated.

213 **Fig 3. Position-specific Single Residue Frustration Index (SRFI) and rvET scores**
214 **mapped onto the HLA I binding groove.**

215 (A, B, C) SRFI mapped onto binding groove positions. (D, E, F) log(rvET) mapped
216 onto binding groove positions. Selected positions are also shown on the structure.

217 **Fig 4. Box-plots of position-specific SRFI values.**

218 (A) SRFI box-plots of selected minimally and highly frustrated residues (B) SRFI box-
219 plots of positions with neutral frustration. (C) SRFI box-plots of peptide-binding pocket
220 positions. Coloring according to log(rvET) values.

221 On the other hand, we also observed several positions on α-1 and α-2 helices
222 with relatively high frustration levels (58, 61, 68, 80, 84, 141, 144, 145, 146, 155) (Fig
223 3). Most of these positions are located within interfaces of interaction with either TCR
224 or Tapasin of the PLC.

225 The interaction between the TCR and pMHC I is a central event in adaptive
226 immunity [26,68,69]. The TCR recognizes a peptide antigen only when presented by
227 an MHC molecule (MHC restriction) [69,70]. TCR-pMHC structures determined to date

228 indicate a conserved diagonal binding geometry, where the hypervariable CDR3 loop
229 of TCR contacts the peptide, and germline-encoded variable α and β domains (Vα
230 and Vβ) contact HLA I α-1 and α-2 helices, respectively [71]. Although deviations from
231 this conventional docking mode exist, including a reversed polarity docking mode
232 [72,73], TCR signaling in such unconventional modes is limited [74]. The importance
233 of the conventional docking mode for TCR signaling adds support to the germline-
234 encoded model of TCR-pMHC interaction, in which evolutionarily conserved TCR-
235 pMHC contacts were proposed to govern MHC restriction [75]. In this regard, previous
236 analyses of TCR-pMHC structures highlighted the importance of contacts made by
237 HLA α-1 and α-2 helix residues 69 and 158 [76] or 65, 69 and 155 as a "restriction
238 triad" [77]. We observed high SRFI levels at or near these positions.

239 Local frustration data can also provide a basis for PLC-MHC interactions.
240 Tapasin is the main component of the PLC which contacts the HLA binding groove.
241 Predicted Tapasin-MHC interactions [39,78] and a recently elucidated crystal structure
242 of pMHC with TAPBPR (a Tapasin homolog) [79–81] indicated that Tapasin cradles
243 the HLA molecule via contacts with α-3 domain and $\alpha_{2\text{-}1}$ helix segment. Our
244 observation of high SRFI levels at positions 141, 144, 145, and 146 located on the $\alpha_{2\text{-}1}$
245 helix segment is in line with these findings.

246 Median SRFI values of most binding groove positions were found to indicate
247 neutral frustration (Fig 4B). On the other hand, HLA polymorphism at peptide binding
248 pocket positions apparently caused significant drifts towards minimal or high
249 frustration as well, even though the median SRFI remained within the neutral
250 frustration range (-1 to 1).

Overall, these results suggest that the human MHC evolves to maintain its structural fold, as evidenced by the dominance of conserved core positions showing minimal frustration within the HLA peptide binding groove. Moreover, the presence of relatively higher frustration at or near TCR and Tapasin contact positions on α-1 and α-2 helices may also provide a biophysical basis for protein interactions of pMHC.

Clustering of HLA alleles into distinct groups based on frustration data

Next, we performed a hierarchical clustering analysis of allele-specific SRFI profiles. For simplification, we excluded positions with less than 0.5 SRFI variation from the analysis. Clustering results are shown in Fig 5A, and complete lists of alleles in each cluster are given in Supplementary Table 1. Strikingly, alleles from different gene loci were clustered into three separate groups based on frustration data of only 14 positions (45, 66, 67, 69, 74, 76, 79, 80, 82, 116, 131, 152, 156, 163). An exception was the HLA-B*46 group, which was clustered along with the HLA-C alleles. This is not surprising, as alleles in this group share the KYRV motif at 66, 67, 69, and 76 with HLA-C [82].

Fig 5. Clustering of allele-specific SRFI profiles based on 14 binding groove positions.

(A) SRFI cluster heatmap. Clustering was performed for all 1436 alleles included, yet not all of these alleles are indicated on axis label for clarity. (B) SRFI of three identified clusters corresponding to three HLA gene loci mapped onto the binding groove. Coloring as in Figure 3.

Clustering of HLA alleles from distinct loci into separate groups has been

273 previously demonstrated based on peptide binding pocket similarities [43], peptide-
274 HLA contacts [42], surface electrostatics [44] and peptide binding repertoires [83].
275 Here, HLA-A, -B and –C alleles were clustered into distinct groups from a different
276 perspective using local frustration data. Moreover, the structural energetics aspect
277 provides an additional level of detail.

278 Unlike data used in previous studies, the SRFI may explain previously reported
279 differences between HLA alleles in terms of pMHC stability and hence, the cell surface
280 expression levels. Compared to HLA-A and HLA-B, lower cell surface expression
281 levels were previously observed for HLA-C alleles [84–87]. Moreover, peptide
282 repertoires of HLA-C alleles are known to be more limited [84,88]. KYRV motif was
283 also shown to be responsible for intrinsic instability of HLA-C [86]. Relatively higher
284 frustration in HLA-C group, especially at binding pocket positions of 66, 74 and 80, is
285 in line with these findings: higher frustration may introduce a destabilizing effect into
286 the binding pockets, lead to restrictions in peptide binding, and thereby reduce cell
287 surface expression levels. Among these positions, 66 is a member of the B pocket and
288 increased frustration in this position may indicate a less stable B pocket in HLA-C.
289 Likewise, position 80 is a member of F pocket and a similar effect may be valid for the
290 F pocket as well. Around position 80, positions 79 and 82 also display high frustration
291 as well, further contributing to F pocket frustration.

292 HLA-C additionally differs from HLA-B and HLA-A in terms of interactions with
293 KIRs of NK cells [89,90]. The alleles in this group contain either the C1 or C2 epitopes
294 with an arginine or lysine at position 80, respectively [91]. Moreover, HLA-C (and HLA-
295 B46, which is clustered alongside HLA-C) is distinguished by a KYRV motif at

296 positions 66, 67, 69 and 76 [88]. Previous crystal structures of HLA-C and KIR clearly
297 show that the KIR contacts residues on α-1 helix of HLA near peptide C terminus
298 [92,93]. High levels of frustration in this contact area of HLA-C (Figure 5B) may
299 explain why HLA-C better interacts with the KIR than HLA-A and HLA-B alleles.

300 All in all, these results support the view that HLA-C is intrinsically less stable on
301 the cell-surface via a post-translational mechanism [84,86]. This mechanism may
302 simply involve a less than optimal packing in the binding groove of HLA-C. Our results
303 also may also explain the limited diversity of HLA-C peptide ligands: higher frustration
304 in ligand binding sites may indicate a lower peptide binding capability.

305 **Effect of peptide binding on local frustration profiles**

306 Peptide-free (empty) MHC does not have a well-defined 3D structure (no crystal
307 structure of a peptide-free MHC could be obtained so far). Instead, empty MHC
308 continuously switches between alternative conformations until a sufficiently stable
309 peptide-HLA interaction is achieved [21]. Peptide binding to HLA occurs via six
310 pockets in the binding groove named A, B, C, D, E and F [22]. For many alleles, A, B
311 and F pockets are decisive for peptide binding (A/B and F pockets binding N- and C-
312 terminus of peptide ligands, respectively). Structural integrity of the F pocket of some
313 alleles was previously shown to be more sensitive to peptide truncation in MD
314 simulations than those of A and B pockets [41,94–96]. This implies that contacts
315 between peptide C-terminus and HLA F pocket are highly important for molecular
316 stability. We reasoned that, by comparing peptide-free and -loaded MHC frustration
317 profiles, the positions that depend on peptide contacts least/most for stability can be

identified. Thus, we additionally quantified local frustration in peptide-free homology models, and calculated average changes in SRFI upon binding of the peptide ligands to each allele. In line with previous observations, SRFI increase (hence reduction in frustration) upon peptide binding was highest near the F-pocket (positions 81, 84, 95, 142, and 147) (Fig 6).

Fig 6. Box-plots of SRFI change upon peptide binding.

Positions are ordered according to decreasing SRFI change.

Material and Methods

Sequence variation analysis

Aligned amino acid sequences of the α-1 and α-1 domains (i.e. the binding groove) of the HLA were retrieved from the IMGT/HLA database [27]. The sequences included in this file were converted to FASTA format for further analysis using Biopython [97]. Seq2Logo 2.0 web server was used to generate sequence logos to indicate conservation/variation at specific sites [98] using Shannon's Information Content (IC) [99] as follows:

$$I = \sum_{a} p_a \cdot log_2 p_a / q_a$$

Here, I denotes information content, p_a is the probability of observing amino acid a at the respective sequence position (calculated from input multiple sequence alignment) and q_a is the pre-defined (background) probability of observing amino acid at the given position. An equal probability was used for each amino acid type (1/20).

The Evolutionary Trace (ET) method is an improvement over the classical IC

approach described above [55,56]. In this method, the IE is calculated after constructing a phylogenetic tree of sequences and for each branch (node) of the tree. Then, the conservation/variation level at a given position is calculated using the following equation:

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left\{ - \sum_{a} p_a \cdot log_2 p_a / q_a \right\}$$

Here, N is the number of sequences in the alignment and $N - 1$ is the number of possible nodes in the phylogenetic tree. The final score ρ_i is also termed "real-value Evolutionary Trace" score (rvET), and denotes the rank of a position (i) with respect to all other positions. Hence, lower rvET values indicate a higher evolutionary importance and vice-versa.

Prediction of peptide binders to class I HLA alleles and homology modeling of pMHC structures

Due to the limited number of HLA alleles with structure information at the Protein Data Bank (PDB), homology modeling was used to generate pMHC structures for 1436 alleles selected as follows. For selecting peptide ligands for homology modeling, a data-driven peptide-binding prediction tool (netMHCpan 3.0) was used [63] (stand-alone version of netMHCpan 3.0 was downloaded from http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?netMHCpan). First, 25000 random nonamer peptide sequences were generated using equal probability for each of the twenty naturally occurring amino acids at each peptide position. Then, binding affinities of all generated peptide sequences were predicted for each HLA allele recognized by netMHCpan 3.0 and with identified α-3 domain sequence. The results were then filtered to extract up to 10

peptides with the highest binding affinities (i.e. lowest binding free energies) and classified as strong binders by netMHCpan 3.0 for each allele. These peptides were then homology modelled in the context of HLA alleles as well as β-2-m using Modeller version 9.19 [100]. An X-ray structure of the HLA-B*53:01 allele was used as template (PDB ID: 1A1M).

Local frustration analysis

Local frustration analysis was performed on all produced homology models using *frustratometer2* tool [64] (Stand-alone version available from https://github.com/gonzaparra/frustratometer2 was used). This tool uses the AWSEM (Associative Memory, Water Mediated, Structure and Energy Model) coarse-grained potential [101] to calculate residue-residue interaction energies. A sequence separation of 3 was used to calculate local amino acid densities, as defined by AWSEM. In addition to the interactions predicted by the AWSEM, long-range electrostatic interactions were also included - as offered by *frustratometer2* - using a Debye-Hückel potential:

$$V_{DH} = K_{elect} \sum_{i<j} \frac{q_i q_j}{\epsilon_r} e^{-r_{ij}/l_D}$$

where q_i and q_j are charges of residues *i* and *j*, r_{ij} is the distance between residues *i* and *j*, ϵ_r is the dielectric constant of the medium (in vacuum this value is 1) and l_D is the Debye-Hückel screening length, which is calculated using physiological values of temperature (25 °C), dielectric constant of water (80) and ionic strength of water (0.1 M), yielding a l_D of 10 Å. Here, it is possible to define an electronic strength of the protein system using a parameter called electrostatic constant (k):

383 $$k = {K_{elect}}/{\epsilon_r}$$

384 Here, a k value of 4.15, which corresponds to an aqueous solution, was used.

385 Once the energies are computed using the AWSEM potential, *frustratometer2*
386 assesses the local frustration present in each residue-residue contact. Here, the
387 contacts between two residues are compared to contacts in decoy structures or
388 molten globule configurations. The tool can produce decoys by simultaneously
389 mutating both residues to all other amino acids. A normalization is applied in order to
390 compare the energies of the native structure to decoys. A "mutational frustration
391 index" (F_{ij}^{m}) then captures how favorable the contacts of the residues present in the
392 native structure to decoys as follows:

393 $$F_{ij}^{m} = \frac{E_{i,j}^{T,N} - \langle E_{i,j}^{T,U} \rangle}{\sqrt{1/N \sum_{k=1}^{n} \left(E_{i,j}^{T,U} - \langle e_{i,j}^{T,U} \rangle \right)^2}}$$

394 where $E_{i,j}^{T,N}$ is the total energy of native protein and $E_{i',j'}^{T,U}$ is the average energy of decoy
395 structures. By changing the amino acid identity at both positions simultaneously, not
396 only the contact between the respective two residues are changed, but those contacts
397 made by the two residues with other residues are changed as well.

398 This index can also be calculated by changing the amino acid type of *only one of*
399 *the residues* instead of applying simultaneous mutations at both positions. The index
400 then represents how favorable the contacts made by a given amino acid are at a given
401 position in the structure. When applied this way, the index is termed "Single Residue
402 Frustration Index" (SRFI).

Frustration-based clustering of HLA class I alleles

Upon application of local frustration analysis on pMHC structures, SRFI profiles averaged over peptides per HLA allele are obtained. This is represented in the form of an "SRFI matrix" with rows corresponding to positions (residues) in pMHC structure and columns corresponding to HLA alleles. A column-wise agglomerative clustering was applied on this matrix to identify similarities and differences between different HLA alleles in terms of their local structural energetics. Distances between clusters were defined by the "single linkage" criteria, in which inter-cluster distances were taken as the shortest distance between any two points of a pair of clusters:

$$L(r,s) = \min\left(D(x_{ri},x_{sj})\right)$$

where $L(r,s)$ is the inter-cluster distance between clusters r and s and $D(x_{ri},x_{sj})$ is the distance between points x_{ri} and x_{sj} of the two clusters. Here, each data point represents the SRFI profile of each HLA allele. The distance between data points were computed using Manhattan distance:

$$D(x_{ri},x_{sj}) = \sum_{k}^{n} |x_{ri,k} - x_{sj,k}|$$

Here, the summation is performed over rows of data (k → n), which are actually positions in the pMHC structure.

Acknowledgments

Marmara University BAP FEN-A-101018-0526 is acknowledged. The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources). OS would like to

acknowledge access to computing resources at RTEU Samsung Center of Excellence (SMM) for data analysis.

Author Contributions

Conceived and designed the experiments: OS. Performed the experiments: OS. Analyzed the data: OS. Contributed reagents/materials/analysis tools: OS PO. Wrote the paper: OS PO

References

1. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Science. Wiley-Blackwell; 2012. pp. 769–785. doi:10.1002/pro.2071

2. Bastolla U, Dehouck Y, Echave J. What evolution tells us about protein physics, and protein physics tells us about evolution. Current Opinion in Structural Biology. Elsevier Current Trends; 2017. pp. 59–66. doi:10.1016/j.sbi.2016.10.020

3. Bahar I, Jernigan RL, Dill KA. Protein actions: principles and modeling. 2017. Available: https://www.crcpress.com/Protein-Actions-Principles-and-Modeling/Bahar-Jernigan-Dill/p/book/9780815341772

4. Kimura M, Ohta T. On some principles governing molecular evolution. Proc Natl Acad Sci U S A. 1974;71: 2848–52. Available: http://www.ncbi.nlm.nih.gov/pubmed/4527913

5. Parra RG, Espada R, Verstraete N, Ferreiro DU. Structural and Energetic Characterization of the Ankyrin Repeat Protein Family. Orengo CA, editor. PLoS

Comput Biol. 2015;11: e1004659. doi:10.1371/journal.pcbi.1004659

6. Abriata LA, Merijn ML, Tomatis PE. Sequence-function-stability relationships in proteins from datasets of functionally annotated variants: The case of TEM β-lactamases. FEBS Lett. 2012;586: 3330–3335. doi:10.1016/j.febslet.2012.07.010

7. Dean AM, Neuhauser C, Grenier E, Golding GB. The Pattern of Amino Acid Replacements in α/β-Barrels. Mol Biol Evol. 2002;19: 1846–1864. doi:10.1093/oxfordjournals.molbev.a004009

8. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, et al. Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. J Mol Evol. 2014;79: 130–142. doi:10.1007/s00239-014-9644-x

9. Shahmoradi A, Wilke CO. Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. Proteins Struct Funct Bioinforma. 2016;84: 841–854. doi:10.1002/prot.25034

10. Echave J, Wilke CO. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. Annu Rev Biophys. 2017;46: 85–103. doi:10.1146/annurev-biophys-070816-033819

11. Echave J, Jackson EL, Wilke CO. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys Biol. 2015;12: 025002. doi:10.1088/1478-3975/12/2/025002

12. Wilke CO. Bringing Molecules Back into Molecular Evolution. Kosakovsky Pond SL, editor. PLoS Comput Biol. 2012;8: e1002572.

doi:10.1371/journal.pcbi.1002572

13. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface. 2014;11: 20140419–20140419. doi:10.1098/rsif.2014.0419

14. Fay JC, Wyckoff GJ, Wu C-I. Positive and Negative Selection on the Human Genome. Genetics. 2001;158.

15. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nature Reviews Genetics Nature Publishing Group; Feb 19, 2016 pp. 109–121. doi:10.1038/nrg.2015.18

16. Meyer D, Vitor VR, Bitarello BD, Débora DY, Nunes K. A genomic perspective on HLA evolution. Immunogenetics. Springer; 2018. pp. 5–27. doi:10.1007/s00251-017-1017-3

17. Crux NB, Elahi S. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? Front Immunol. 2017;8: 832. doi:10.3389/fimmu.2017.00832

18. Vandiedonck C, Knight JC. The human Major Histocompatibility Complex as a paradigm in genomics research. Brief Funct Genomic Proteomic. 2009;8: 379–94. doi:10.1093/bfgp/elp010

19. Martínez-Borra J, López-Larrea C. The emergence of the Major Histocompatilibility Complex. Adv Exp Med Biol. 2012;738: 277–89. doi:10.1007/978-1-4614-1680-7_16

20. Kurimoto E, Kuroki K, Yamaguchi Y, Yagi-Utsumi M, Igaki T, Iguchi T, et al. Structural and functional mosaic nature of MHC class I molecules in their peptide-free form. Mol Immunol. 2013;55: 393–399. doi:10.1016/j.molimm.2013.03.014

21. Theodossis A. On the trail of empty MHC class-I. Mol Immunol. 2013;55: 131–4. doi:10.1016/j.molimm.2012.10.012

22. Saper MA, Bjorkman PJ, Wiley DC. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. J Mol Biol. 1991;219: 277–319. doi:10.1016/0022-2836(91)90567-P

23. Liu J, Gao GF. Major Histocompatibility Complex: Interaction with Peptides. eLS. Chichester, UK: John Wiley & Sons, Ltd; 2011. doi:10.1002/9780470015902.a0000922.pub2

24. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The major histocompatibility complex and its functions. Garland Science; 2001. Available: http://www.ncbi.nlm.nih.gov/books/NBK27156/

25. Rangarajan S, Mariuzza RA. Natural Killer Cell Receptors. Struct Biol Immunol. 2018; 101–125. doi:10.1016/B978-0-12-803369-2.00004-8

26. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. Annu Rev Immunol. 2015;33: 169–200. doi:10.1146/annurev-immunol-032414-112334

27. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: Allele variant databases. Nucleic Acids Res. 2015;43:

511 D423–D431. doi:10.1093/nar/gku1161

512 28. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC,
513 mechanisms and misunderstandings. Proc R Soc B Biol Sci. 2010;277: 979–988.
514 doi:10.1098/rspb.2009.2084

515 29. Sommer S. The importance of immune gene variability (MHC) in evolutionary
516 ecology and conservation. Frontiers in Zoology. BioMed Central; 2005. p. 16.
517 doi:10.1186/1742-9994-2-16

518 30. Parham P. Function and polymorphism of human leukocyte antigen-A,B,C
519 molecules. Am J Med. 1988;85: 2–5. doi:10.1016/0002-9343(88)90369-5

520 31. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility
521 complex class I loci reveals overdominant selection. Nature. 1988;335: 167–170.
522 doi:10.1038/335167a0

523 32. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast
524 preponderance of HLA-A and -B polymorphism. Immunogenetics. 1999;50: 201–
525 212. doi:10.1007/s002510050594

526 33. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease.
527 Nature Reviews Immunology. Nature Publishing Group; 2018. pp. 325–339.
528 doi:10.1038/nri.2017.143

529 34. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et
530 al. Major histocompatibility complex (MHC) class I and MHC class II proteins:
531 Conformational plasticity in antigen presentation. Frontiers in Immunology. 2017.
532 p. 292. doi:10.3389/fimmu.2017.00292

533 35. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell
534 Antigen Receptor Recognition of Antigen-Presenting Molecules. Annu Rev
535 Immunol. 2015;33: 169–200. doi:10.1146/annurev-immunol-032414-112334

536 36. Antunes DA, Rigo MM, Freitas M V., Mendes MFA, Sinigaglia M, Lizée G, et al.
537 Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-
538 Based Cancer Immunotherapy. Front Immunol. 2017;8: 1–16.
539 doi:10.3389/fimmu.2017.01210

540 37. Li Y, Mariuzza RA. Structural basis for recognition of cellular and viral ligands by
541 NK cell receptors. Front Immunol. 2014;5: 123. doi:10.3389/fimmu.2014.00123

542 38. Blees A, Januliene D, Hofmann T, Koller N, Schmidt C, Trowitzsch S, et al.
543 Structure of the human MHC-I peptide-loading complex. Nature. 2017;551: 525–
544 528. doi:10.1038/nature24627

545 39. Fisette O, Wingbermühle S, Tampé R, Schäfer L V. Molecular mechanism of
546 peptide editing in the tapasin-MHC I complex. Sci Rep. 2016;6: 19085.
547 doi:10.1038/srep19085

548 40. Hafstrand I, Sayitoglu EC, Apavaloaei A, Josey BJ, Sun R, Han X, et al.
549 Successive crystal structure snapshots suggest the basis for MHC class I
550 peptide loading and editing by tapasin. Proc Natl Acad Sci U S A. 2019;116:
551 5055–5060. doi:10.1073/pnas.1807656116

552 41. Fisette O, Wingbermühle S, Schäfer L V. Partial Dissociation of Truncated
553 Peptides Influences the Structural Dynamics of the MHCI Binding Groove. Front
554 Immunol. 2017;8: 408. doi:10.3389/fimmu.2017.00408

555 42. Harjanto S, Ng LFP, Tong JC. Clustering HLA class I superfamilies using
556 structural interaction patterns. Kobe B, editor. PLoS One. 2014;9: e86655.
557 doi:10.1371/journal.pone.0086655

558 43. Mukherjee S, Warwicker J, Chandra N. Deciphering complex patterns of class-I
559 HLA-peptide cross-reactivity via hierarchical grouping. Immunol Cell Biol.
560 2015;93: 522–32. doi:10.1038/icb.2015.3

561 44. Mumtaz S, Nabney IT, Flower DR. Scrutinizing human MHC polymorphism:
562 Supertype analysis using Poisson-Boltzmann electrostatics and clustering. J Mol
563 Graph Model. 2017;77: 130–136. doi:10.1016/J.JMGM.2017.07.033

564 45. Onuchic JN, Luthey-Schulten Z, Wolynes PG. THEORY OF PROTEIN
565 FOLDING: The Energy Landscape Perspective. Annu Rev Phys Chem. 1997;48:
566 545–600. doi:10.1146/annurev.physchem.48.1.545

567 46. Ferreiro DU, Komives EA, Wolynes PG. Frustration in biomolecules. Q Rev
568 Biophys. 2014;47: 285–363. doi:10.1017/S0033583514000092

569 47. Ferreiro DU, Komives EA, Wolynes PG. Frustration, function and folding. Curr
570 Opin Struct Biol. 2018;48: 68–73. doi:10.1016/J.SBI.2017.09.006

571 48. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. Localizing frustration in
572 native proteins and protein assemblies. Proc Natl Acad Sci U S A. 2007;104:
573 19819–24. doi:10.1073/pnas.0709915104

574 49. Freiberger MI, Guzovsky AB, Wolynes PG, Parra RG, Ferreiro DU. Local
575 frustration around enzyme active sites. Proc Natl Acad Sci U S A. 2019;116:
576 4037–4043. doi:10.1073/pnas.1819859116

577 50. Tripathi S, Waxham MN, Cheung MS, Liu Y. Lessons in Protein Design from
578 Combined Evolution and Conformational Dynamics. Sci Rep. 2015;5: 14259.
579 doi:10.1038/srep14259

580 51. Tripathi S, Wang Q, Zhang P, Hoffman L, Waxham MN, Cheung MS.
581 Conformational frustration in calmodulin-target recognition. J Mol Recognit.
582 2015;28: 74–86. doi:10.1002/jmr.2413

583 52. Espada R, Ferreiro DU, Parra R G. The Design of Repeat Proteins: Stability
584 Conflicts with Functionality. Biochem Mol Biol J. 2017;03. doi:10.21767/2471-
585 8084.100031

586 53. Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, et al.
587 Distinguishing functional polymorphism from random variation in the sequences
588 of >10,000 HLA-A, -B and -C alleles. Keating BJ, editor. PLoS Genet. 2017;13:
589 e1006862. doi:10.1371/journal.pgen.1006862

590 54. van Deutekom HWM, Keşmir C. Zooming into the binding groove of HLA
591 molecules: which positions and which substitutions change peptide binding
592 most? Immunogenetics. 2015;67: 425–36. doi:10.1007/s00251-015-0849-y

593 55. Mihalek I, Reš I, Lichtarge O. A Family of Evolution-Entropy Hybrid Methods for
594 Ranking Protein Residues by Importance. J Mol Biol. 2004;336: 1265–1282.
595 doi:10.1016/j.jmb.2003.12.078

596 56. Wilkins A, Erdin S, Lua R, Lichtarge O. Evolutionary trace for prediction and
597 redesign of protein functional sites. Methods Mol Biol. 2012;819: 29–42.
598 doi:10.1007/978-1-61779-465-0_3

599 57. Wilkins AD, Lua R, Erdin S, Ward RM, Lichtarge O. Sequence and structure
600 continuity of evolutionary importance improves protein functional site discovery
601 and annotation. Protein Sci. 2010;19: 1296–1311. doi:10.1002/pro.406

602 58. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, et al.
603 Structural clusters of evolutionary trace residues are statistically significant and
604 common in proteins. J Mol Biol. 2002;316: 139–154. doi:10.1006/jmbi.2001.5327

605 59. Mukherjee S, Warshel A. Electrostatic origin of the mechanochemical rotary
606 mechanism and the catalytic dwell of F1-ATPase. Proc Natl Acad Sci U S A.
607 2011;108: 20550–5. doi:10.1073/pnas.1117024108

608 60. Omasits U, Knapp B, Neumann M, Steinhauser O, Kobler R, Schreiner W, et al.
609 Analysis of key parameters for molecular dynamics of pMHC molecules. Mol
610 Simul. 2008;34: 781–793. doi:10.1080/08927020802256298

611 61. Serçinoğlu O, Ozbek P. Computational characterization of residue couplings and
612 micropolymorphism-induced changes in the dynamics of two differentially
613 disease-associated human MHC class-I alleles. J Biomol Struct Dyn. 2018;36:
614 724–740. doi:10.1080/07391102.2017.1295884

615 62. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al.
616 The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2014;43: D405–
617 D412. doi:10.1093/nar/gku938

618 63. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC
619 class I molecules integrating information from multiple receptor and peptide
620 length datasets. Genome Med. 2016;8: 33. doi:10.1186/s13073-016-0288-x

64. Parra RG, Schafer NP, Radusky LG, Tsai MY, Guzovsky AB, Wolynes PG, et al. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. Nucleic Acids Res. 2016;44: W356–W360. doi:10.1093/nar/gkw304

65. Warburton RJ, Matsui M, Rowland-Jones SL, Gammon MC, Katzenstein GE, Wei T, et al. Mutation of the α2 domain disulfide bridge of the class I molecule HLA-A∗0201 Effect on maturation and peptide presentation. Hum Immunol. 1994;39: 261–271. doi:10.1016/0198-8859(94)90269-0

66. Dib L, Salamin N, Gfeller D. Polymorphic sites preferentially avoid co-evolving residues in MHC class I proteins. PLoS Comput Biol. 2018;14: e1006188. doi:10.1371/journal.pcbi.1006188

67. Haliloglu T, Gul A, Erman B. Predicting important residues and interaction pathways in proteins using gaussian network model: Binding and stability of HLA proteins. Nussinov R, editor. PLoS Comput Biol. 2010;6: 20. doi:10.1371/journal.pcbi.1000845

68. Courtney AH, Lo WL, Weiss A. TCR Signaling: Mechanisms of Initiation and Propagation. Trends in Biochemical Sciences. 1 Feb 2017: 108–123. doi:10.1016/j.tibs.2017.11.008

69. La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. Understanding the drivers of MHC restriction of T cell receptors. Nature Reviews Immunology. 10 Apr 2018: 1–12. doi:10.1038/s41577-018-0007-5

70. Zinkernagel RM, Doherty PC. Immunological surveillance against altered self

components by sensitised T lymphocytes in lymphocytes choriomeningitis. Nature. 1974;251: 547–548. doi:10.1038/251547a0

71. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. Annu Rev Immunol. 2015;33: 169–200. doi:10.1146/annurev-immunol-032414-112334

72. Beringer DX, Kleijwegt FS, Wiede F, van der Slik AR, Loh KL, Petersen J, et al. T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. Nat Immunol. 2015;16: 1153–1161. doi:10.1038/ni.3271

73. Gras S, Chadderton J, Del Campo CM, Farenc C, Wiede F, Josephs TM, et al. Reversed T Cell Receptor Docking on a Major Histocompatibility Class I Complex Limits Involvement in the Immune Response. Immunity. 2016;45: 749–760. doi:10.1016/j.immuni.2016.09.007

74. Adams JJ, Narayanan S, Liu B, Birnbaum ME, Kruse AC, Bowerman NA, et al. T Cell Receptor Signaling Is Limited by Docking Geometry to Peptide-Major Histocompatibility Complex. Immunity. 2011;35: 681–693. doi:10.1016/J.IMMUNI.2011.09.013

75. Garcia KC, Adams JJ, Feng D, Ely LK. The molecular basis of TCR germline bias for MHC is surprisingly simple. Nat Immunol. 2009;10: 143–147. doi:10.1038/ni.f.219

76. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW. Evolutionarily Conserved Amino Acids That Control TCR-MHC Interaction. Annu Rev Immunol.

2008;26: 171–203. doi:10.1146/annurev.immunol.26.021607.090421

77. Tynan FE, Burrows SR, Buckle AM, Clements CS, Borg NA, Miles JJ, et al. T cell receptor recognition of a “super-bulged” major histocompatibility complex class I-bound peptide. Nat Immunol. 2005;6: 1114–1122. doi:10.1038/ni1257

78. Fleischmann G, Fisette O, Thomas C, Wieneke R, Tumulka F, Schneeweiss C, et al. Mechanistic Basis for Epitope Proofreading in the Peptide-Loading Complex. J Immunol. 2015;195: 4503–13. doi:10.4049/jimmunol.1501515

79. Thomas C, Tampé R. Structure of the TAPBPR–MHC I complex defines the mechanism of peptide loading and editing. Science (80-). 2017;358: 1060–1064. doi:10.1126/science.aao6001

80. Jiang J, Natarajan K, Boyd LF, Morozov GI, Mage MG, Margulies DH. Crystal structure of a TAPBPR–MHC I complex reveals the mechanism of peptide editing in antigen presentation. Science (80-). 2017;358: 1064–1068. doi:10.1126/science.aao5154

81. Natarajan K, Jiang J, Margulies DH. Structural aspects of chaperone-mediated peptide loading in the MHC-I antigen presentation pathway. Crit Rev Biochem Mol Biol. 2019;54: 164–173. doi:10.1080/10409238.2019.1610352

82. Hilton HG, McMurtrey CP, Han AS, Djaoud Z, Guethlein LA, Blokhuis JH, et al. The Intergenic Recombinant HLA-B∗46:01 Has a Distinctive Peptidome that Includes KIR2DL3 Ligands. Cell Rep. 2017;19: 1394–1405. doi:10.1016/j.celrep.2017.04.059

83. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a

revised and updated classification. BMC Immunol. 2008;9: 1. doi:10.1186/1471-2172-9-1

84. Neisig A, Melief CJ, Neefjes J. Reduced cell surface expression of HLA-C molecules correlates with restricted peptide binding and stable TAP interaction. J Immunol. 1998;160: 171–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/9551969

85. Neefjes JJ, Ploegh HL. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. Eur J Immunol. 1988;18: 801–810. doi:10.1002/eji.1830180522

86. Sibilio L, Martayan A, Setini A, Monaco E Lo, Tremante E, Butler RH, et al. A single bottleneck in HLA-C assembly. J Biol Chem. 2008;283: 1267–1274. doi:10.1074/jbc.M708068200

87. Apps R, Meng Z, Del Prete GQ, Lifson JD, Zhou M, Carrington M. Relative Expression Levels of the HLA Class-I Proteins in Normal and HIV-Infected Cells. J Immunol. 2015;194: 3594–3600. doi:10.4049/jimmunol.1403234

88. Zemmour J, Parham P. Distinctive polymorphism at the HLA-C locus: implications for the expression of HLA-C. J Exp Med. 1992;176: 937–950. doi:10.1084/jem.176.4.937

89. Parham P, Moffett A. Variable NK cell receptors and their MHC class i ligands in immunity, reproduction and human evolution. Nature Reviews Immunology. Nature Publishing Group; 2013. pp. 133–144. doi:10.1038/nri3370

709 90. Winter CC, Long E. A single amino acid in the p58 killer cell inhibitory receptor
710 controls the ability of natural killer cells to discriminate between the two groups of
711 HLA-C allotypes. J Immunol. 1997;142: 142–144.
712 doi:10.4049/jimmunol.166.12.7260

713 91. Winter CC, Long EO. A Single Amino Acid in the p58 Killer Cell Inhibitory
714 Receptor Controls the Ability of Natural Killer Cells to Discriminate between the
715 Two Groups of HLA-C Allotypes. J Immunol. 1997;158: 4026–4028. Available:
716 http://www.ncbi.nlm.nih.gov/pubmed/9126959

717 92. Boyington JC, Sun PD. A structural perspective on MHC class I recognition by
718 killer cell immunoglobulin-like receptors. Molecular Immunology. Pergamon;
719 2002. pp. 1007–1021. doi:10.1016/S0161-5890(02)00030-5

720 93. Fan QR, Long EO, Wiley DC. Crystal structure of the human natural killer cell
721 inhibitory receptor KIR2DL1–HLA-Cw4 complex. Nat Immunol. 2001;2: 452–460.
722 doi:10.1038/87766

723 94. Abualrous ET, Saini SK, Ramnarayan VR, Ilca FT, Zacharias M, Springer S. The
724 Carboxy Terminus of the Ligand Peptide Determines the Stability of the MHC
725 Class I Molecule H-2Kb: A Combined Molecular Dynamics and Experimental
726 Study. PLoS One. 2015;10: e0135421. doi:10.1371/journal.pone.0135421

727 95. Abualrous ET, Fritzsche S, Hein Z, Al-Balushi MS, Reinink P, Boyle LH, et al. F
728 pocket flexibility influences the tapasin dependence of two differentially disease-
729 associated MHC Class I proteins. Eur J Immunol. 2015;45: 1248–57.
730 doi:10.1002/eji.201445307

731 96. Hein Z, Uchtenhagen H, Abualrous ET, Saini SK, Janßen L, Van Hateren A, et
732 al. Peptide-independent stabilization of MHC class I molecules breaches cellular
733 quality control. J Cell Sci. 2014;127: 2885–97. doi:10.1242/jcs.145334

734 97. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython:
735 freely available Python tools for computational molecular biology and
736 bioinformatics. Bioinformatics. 2009;25: 1422–3.
737 doi:10.1093/bioinformatics/btp163

738 98. Thomsen MCF, Nielsen M. Seq2Logo: A method for construction and
739 visualization of amino acid binding motifs and sequence profiles including
740 sequence weighting, pseudo counts and two-sided representation of amino acid
741 enrichment and depletion. Nucleic Acids Res. 2012;40: W281–W287.
742 doi:10.1093/nar/gks469

743 99. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J.
744 1948;27: 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

745 100. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER.
746 Curr Protoc Bioinformatics. 2014;47: 5.6.1-5.6.32.
747 doi:10.1002/0471250953.bi0506s47

748 101. Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA.
749 AWSEM-MD: Protein structure prediction using coarse-grained physical
750 potentials and bioinformatically based local structure biasing. J Phys Chem B.
751 2012;116: 8494–8503. doi:10.1021/jp212541y

752 ## Supporting Information

753 **Supplementary Table 1. HLA I alleles for which homology models were**
754 **generated**.

755 Clusters are also indicated with numbers (1, 2, or 3). HLA core alleles (taken from
756 Robinson et al. [53]) are shown in bold.

757

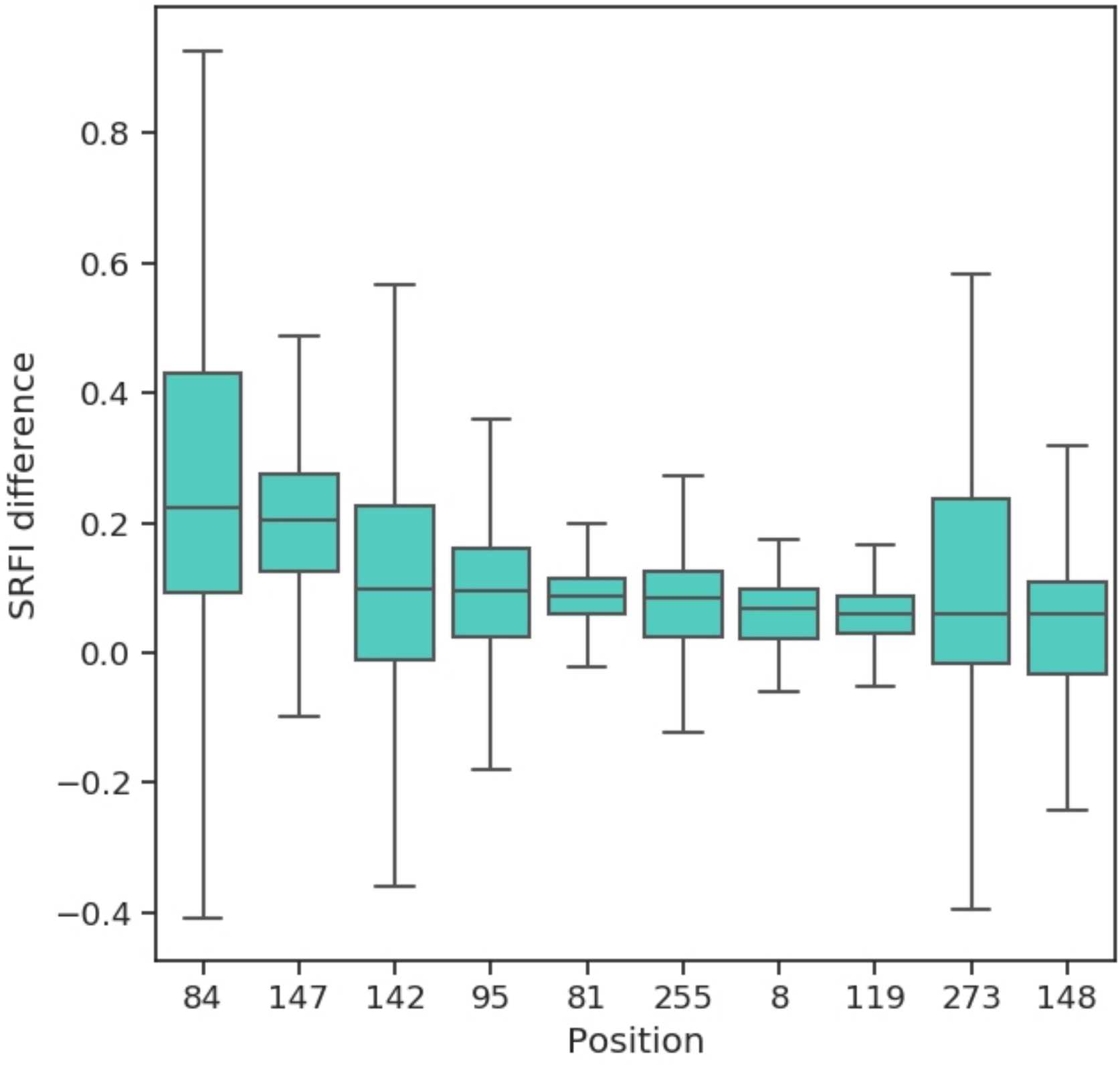

SRFI difference
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
84
147
142
95
81
255
8
119
273
148
Position

Figure 6

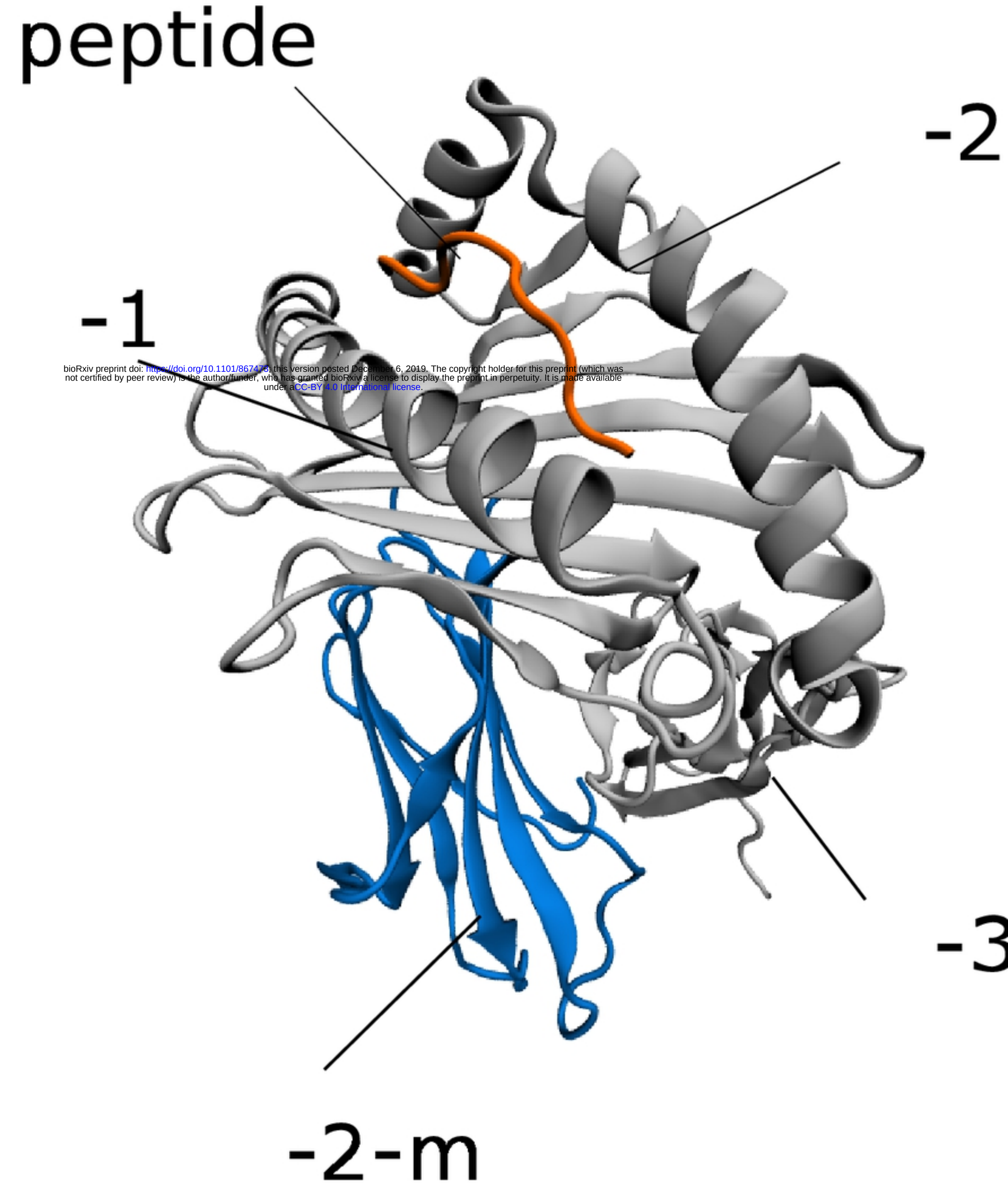
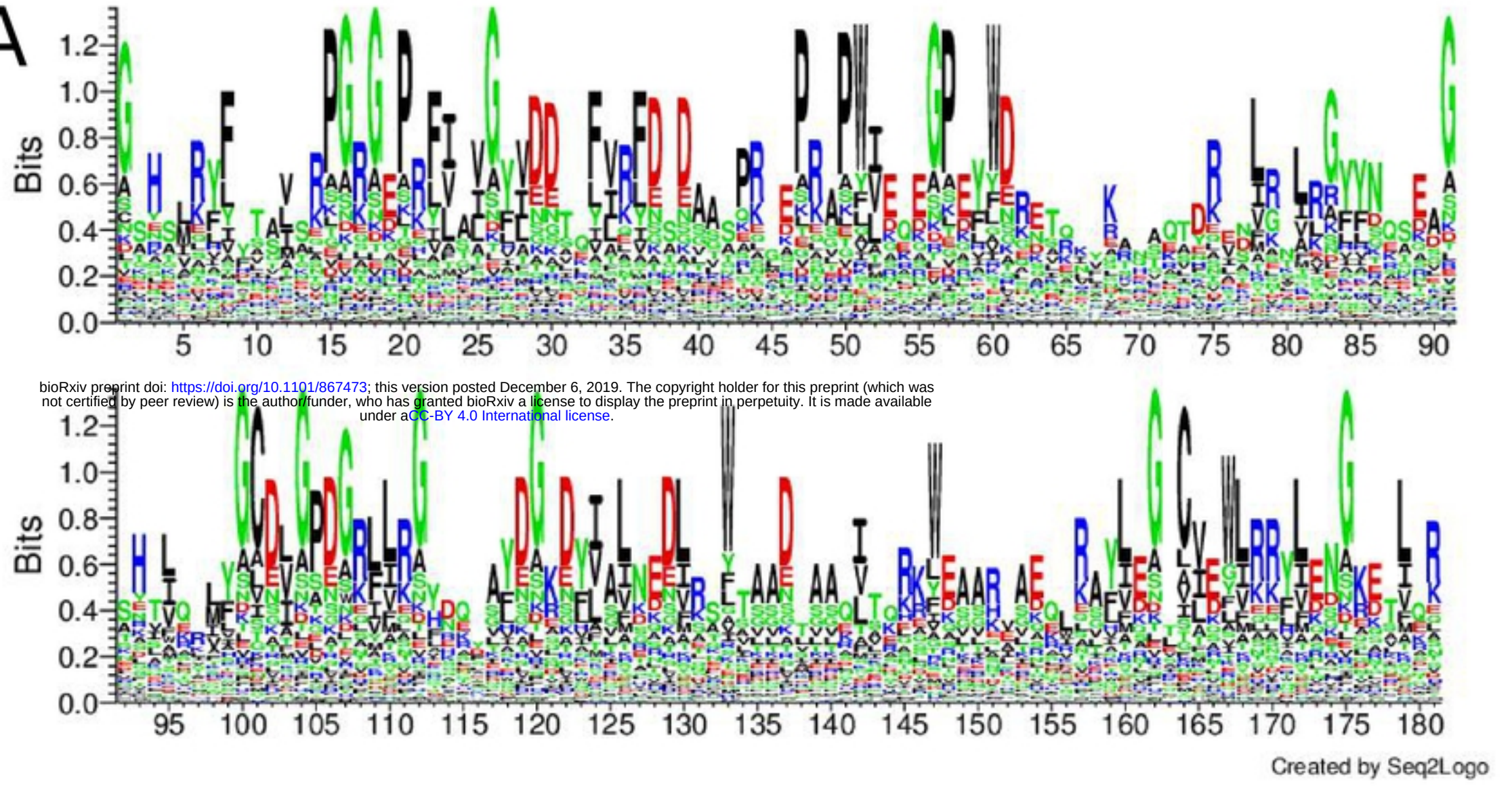


Figure 1

A



B

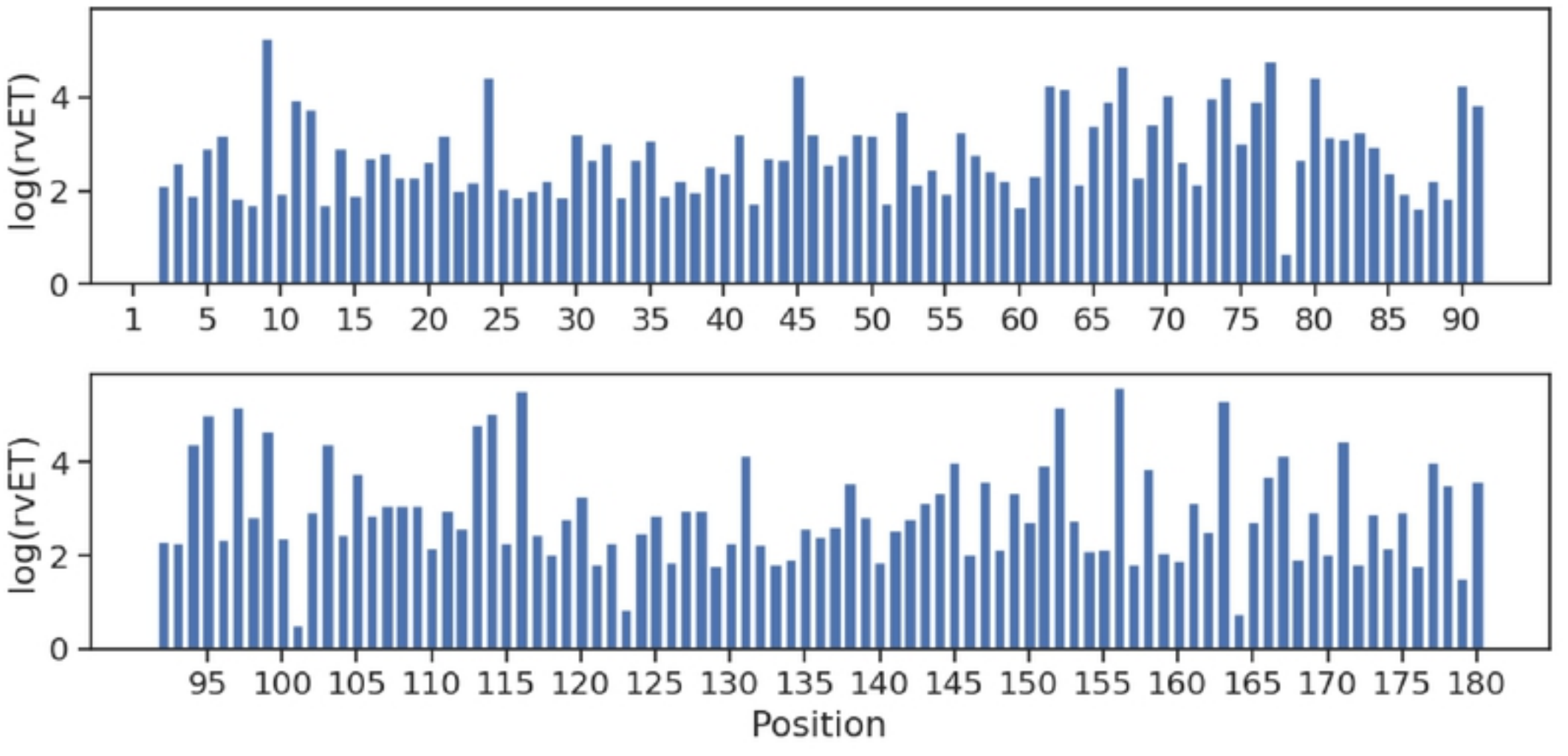


Figure 2

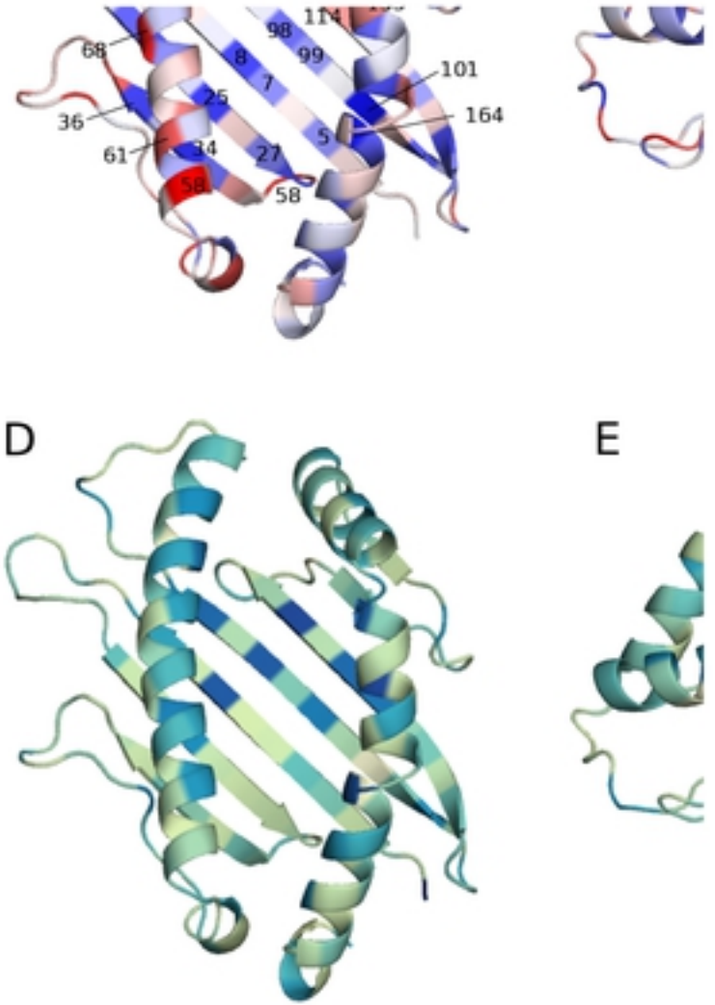


Figure 3

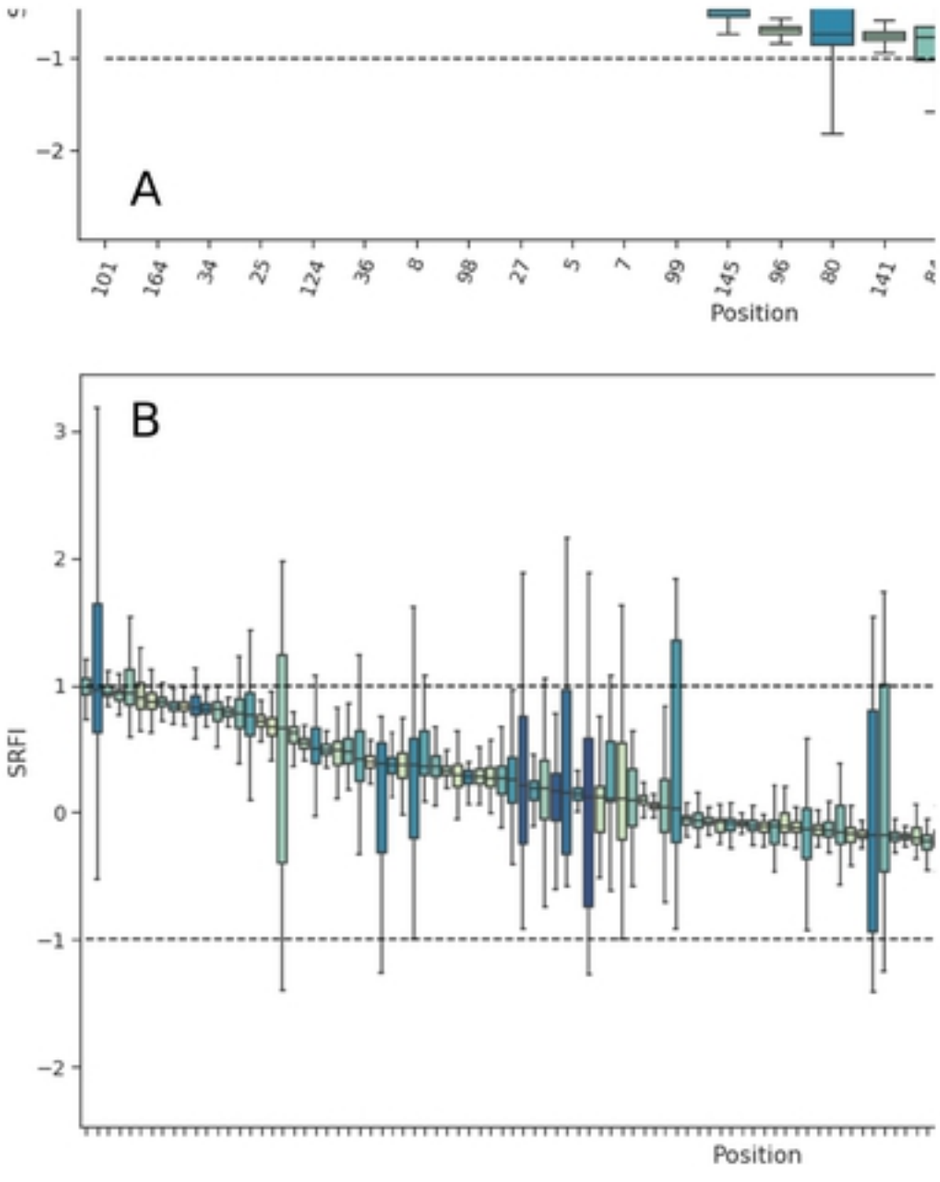
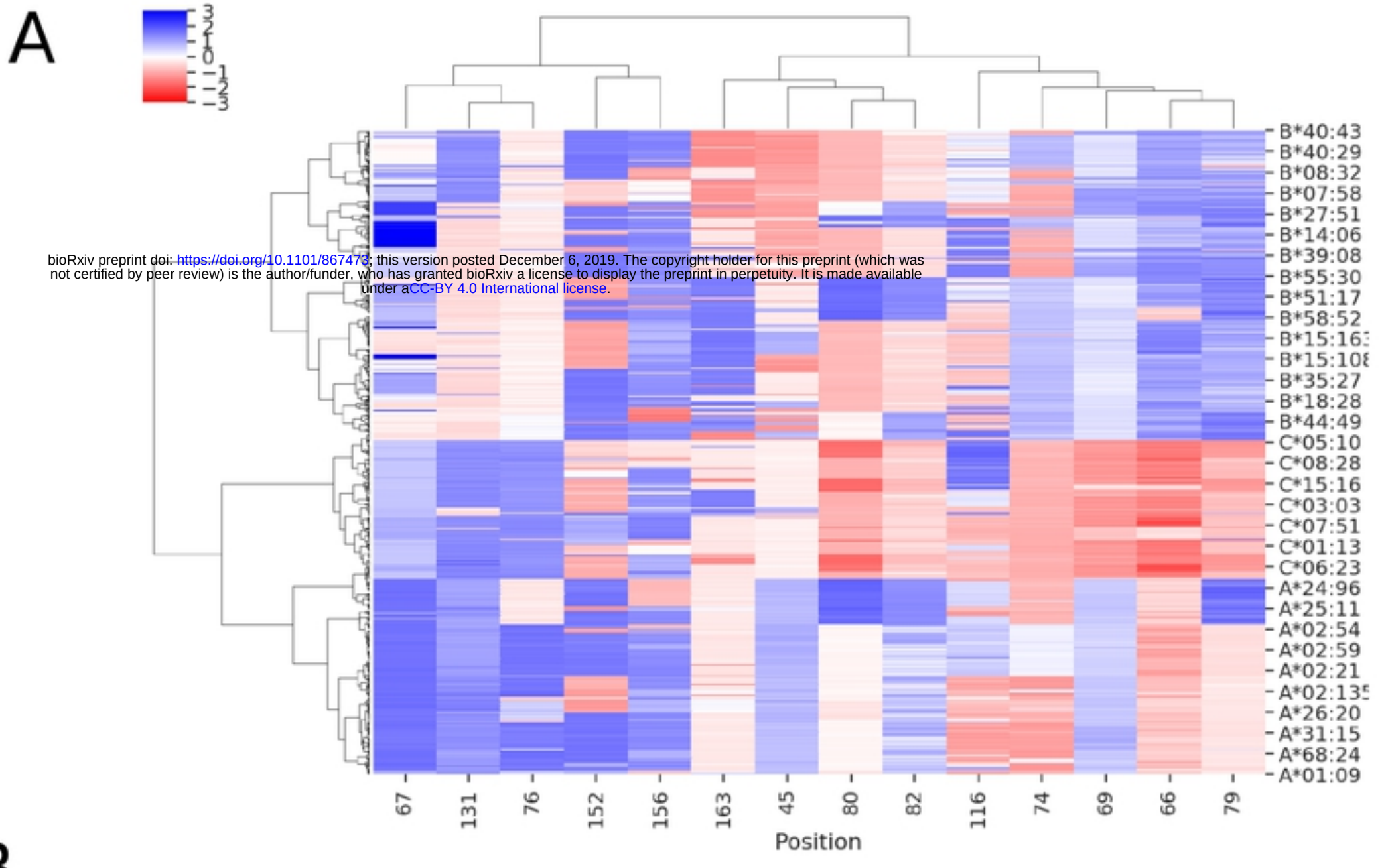

A
B
101
164
34
25
124
36
8
98
27
5
7
99
145
96
80
141
Position
SRFI
Position

Figure 4

A



B

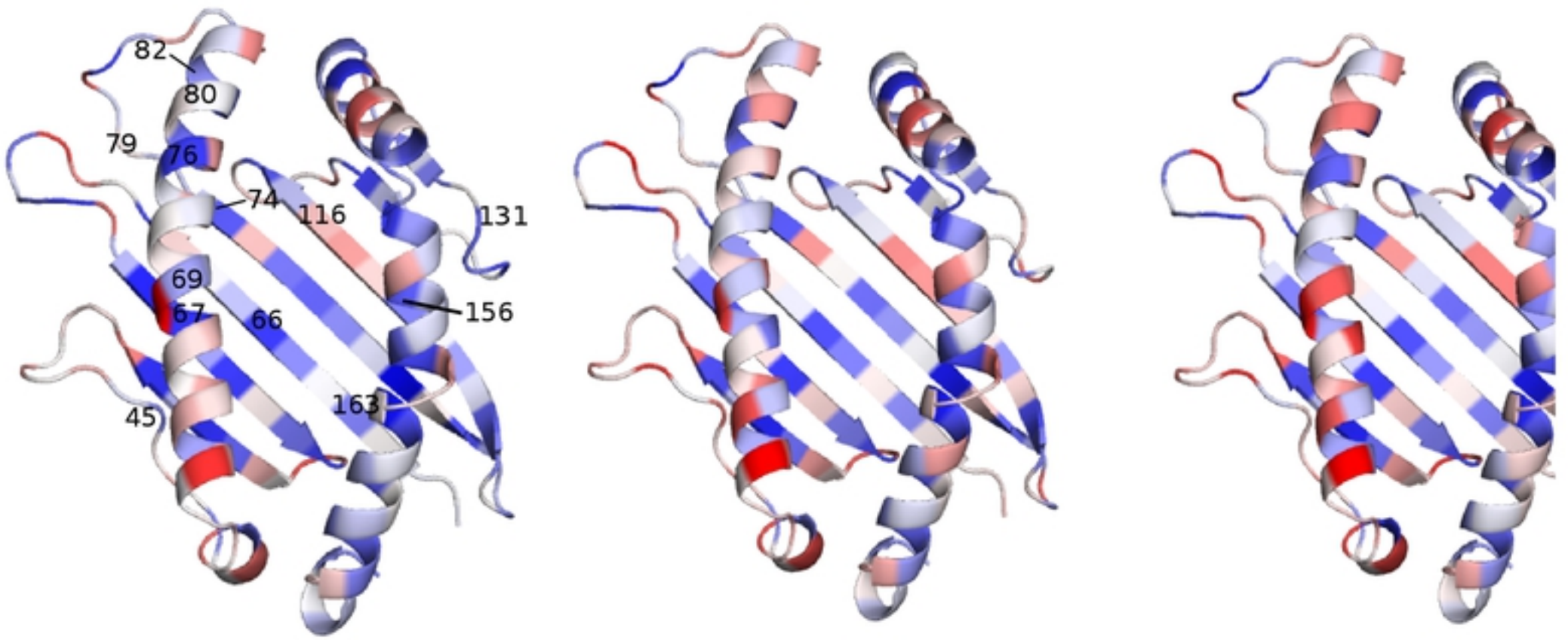


Figure 5