Insights into the stability of a therapeutic antibody Fab fragment by molecular dynamics and its stabilization by computational design

Computational stabilization of an antibody Fab fragment

Nuria Codina[1], Cheng Zhang[1], Nesrine Chakroun[1] and Paul A. Dalby[1]

[1]Department of Biochemical Engineering, University College London, Gordon Street, London, WC1E 7JE, UK

Corresponding Author:

Prof. Paul A. Dalby, Department of Biochemical Engineering, University College London, Gordon Street, London, WC1E 7JE, UK. Tel: +44-20-7679-9566. Email: p.dalby@ucl.ac.uk.

Abstract

Successful development of protein therapeutics depends critically on achieving stability under a range of conditions, while retaining their specific mode of action. Gaining a deeper understanding of the drivers of instability across different stress conditions, will potentially enable the engineering of protein scaffolds that are inherently manufacturable and stable. Here, we compared the structural robustness of a humanized antibody fragment (Fab) A33 using atomistic molecular dynamics simulations under two different stresses of low pH and high temperature. RMSD calculations, structural alignments and contact analysis revealed that low pH unfolding was initiated through loss of contacts at the constant domain interface (C_L-C_H1), prior to C_L domain unfolding. By contrast, thermal unfolding began with loss of contacts in both the C_L-C_H1 and variable domain interface (V_L-V_H), followed by domain unfolding of C_L and also of V_H, thus revealing divergent unfolding pathways. FoldX and Rosetta both agreed that mutations at the C_L-C_H1 interface have the greatest potential for increasing the stability of Fab A33. Additionally, packing density calculations found these residues to be under-packed relative to other inter-domain residues. Two salt bridges were identified that possibly drive the conformational change at low pH, while at high temperature, salt bridges were lost and reformed quickly, and not always with the same partner, thus contributing to an overall destabilization. Sequence entropy analysis of existing Fab sequences revealed considerable scope for further engineering, where certain natural mutations agreed with FoldX and Rosetta predictions. Lastly, the unfolding events at the two stress conditions exposed different predicted aggregation-prone regions (APR), which would potentially lead to different aggregation mechanisms. Overall, our results identified the early stages of unfolding and stability-limiting regions of Fab A33, which provide interesting targets for future protein engineering work aimed at stabilizing to both thermal and pH-stresses simultaneously.

Author Summary

Currently, antibody-based products are the most rapidly growing class of pharmaceuticals due to their high specificity towards their targets, such as biomarkers on the surface of cancer cells. However, they tend to aggregate at all stages of product development, which leads to decreased efficiency and could elicit an immunological response. Improvements in the stability of therapeutic antibodies are generally made during the development phase, by trial and error of the composition of the formulated product, which is both costly and time consuming. There is great demand and potential for identifying the drivers of instability across different stress conditions, early in the discovery phase, which will enable the rational engineering of protein scaffolds. This work elucidated the stability-limiting regions of the antibody fragment Fab A33 using several computational tools: atomistic molecular dynamics simulations, *in-silico* mutational analysis by FoldX and Rosetta, packing density calculators, analysis of existing Fab sequences and predictors of aggregation-prone regions. Results identified particular regions in which mutagenesis has the potential to stabilize Fab against both thermal and pH-stresses simultaneously. Overall, the methodology used here could improve the developability screening of candidate antibody products for many diseases, such as cancer, chronic inflammatory diseases and infectious diseases.

Key Words

Antibody fragment; protein stability; protein aggregation; protein engineering; molecular dynamics simulations

Introduction

In the last 30 years, monoclonal antibody products have become the main drug class for new approvals in the pharmaceutical industry [1]. To date, over 60 antibody-based drugs are on the market, representing half of the total sales, with over 550 further antibodies in clinical development [2]. They are used as therapeutic drugs to treat human diseases, mainly in oncology, auto-immune diseases and cardiovascular diseases. The use of antibody fragments, such as the antigen-binding antibody fragment (Fab) studied here, brings additional advantages, including deeper tissue penetration due to their smaller size, which has proven beneficial to treat tumors [3]. In addition, Fab fragments lack the Fc domain, and thus are not glycosylated which allows simpler and less costly manufacture due to their expression in prokaryotic systems [4]. However, the lack of the Fc domain leads to their more rapid clearance in humans than for full antibodies.

The stabilization of therapeutic proteins against aggregation remains one of the biggest challenges facing approvals as biopharmaceutical products [5–7]. Thus, not only their mode of action, but protein stability is a crucial factor to their becoming successful products. Novel antibody products such as Fabs, single-chain variable fragments (ScFvs) and bi-specifics are currently being developed and their properties remain largely unknown. Knowledge about the stability of these pharmaceutical products, especially in the early development stages, would aid in their engineering and the design of antibody fragments that are more aggregation resistant.

Native protein conformations are only marginally stable, and are highly dynamic, hence they are more realistically described as a native ensemble. There is increasing evidence to suggest that under native conditions, aggregation takes place primarily from partially unfolded native-like states [8–12]. However, little is known about the structures of native conformers that initiate aggregation, or how these are affected by different stress conditions. Local unfolding of proteins can expose aggregation prone-regions (APR), that have the potential to trigger aggregation [13,14]. APRs are the regions in the protein most likely to form and stabilize the cross-

β structures that are characteristic of many aggregates, notably hydrophobic sequences with low net charge and a strong β-sheet forming propensity. Generally, APRs are located in the protein core, protected from the solvent, and thus blocked from forming cross-β structure. Under certain stresses, such as an increase in temperature, a change in pH, addition of denaturants, or elevated shear force, structural regions in the protein may destabilize and partially unfold to expose any underlying APRs [15]. Each structural region of the protein can respond differently to stress, and so the overall pattern of responses is likely to vary with each type of stress. Thus, determining the conformational changes that a protein experiences under different stress conditions is important for identifying common routes towards stabilization across all stress conditions via either mutagenesis or formulation.

Molecular dynamics (MD) simulations have been extensively used to study protein stability [16–21]. MD simulations offer atomic resolution insights into the early conformational events that can take place under different conditions. To date, not many all-atom MD studies on antibody fragments have been reported. MD simulations were used previously to study the aggregation potential of an antibody Fab fragment, from a human IgG1k antibody, via multiple elevated temperature MD simulations at 300 K, 450 K and 500 K [22]. This revealed that domain interfaces deformed prior to the unfolding of individual domains, and that two V_H domain sites were potentially labile to aggregation. Their structural deformation increased the solvent-accessible surface area of the APRs in those regions. The unfolding process of an antibody Fab fragment was also studied using an elastic network model, to reveal that the constant regions were more flexible, and unfolded earlier, than the variable regions [23]. MD simulations at 450 K and 500 K have also revealed the stability-limiting regions of an antibody single-chain variable fragment (scFv) [24]. Disruption of the V_L-V_H interface was found to precede the unfolding of the domain structures. In contrast to the study on the Fab above, the V_H domain of the scFv was found to be more thermally resistant than the V_L domain.

Each Fab is composed of one light and one heavy chain (Fig 1 and Fig S1), each

122 comprising a variable (V_L or V_H) and a constant (C_L or C_H1) domain. Each domain forms an
123 immunoglobulin fold, having two layers of β-sheets, an inner β-sheet and an outer β-sheet.
124 Constant domains are formed of seven β-strands in a 4+3 arrangement, while variable domains
125 have two additional β-strands in a 5+4 arrangement. The variable domains contain the antigen-
126 binding site at their complementary determining regions (CDRs), formed by three loops in V_L and
127 three loops in C_L. There are five disulfide bonds in Fab, four of them intra-domain and the last one
128 between the light and heavy chains at the hinge region. Individual domains interact to form the
129 variable region interface (V_L-V_H), and the constant region interface (C_L-C_H1). Interface contacts
130 are shown in Fig 1 and the residues involved in the contacts are listed in Table S1. The variable
131 region interface is mainly formed by aromatic side chains that are tightly packed and located at
132 the center of the interface (six Tyr, two Trp, and two Phe), forming hydrophobic interactions.
133 However, fewer aromatic side-chains are involved in the constant region domain interface (four
134 Phe). Furthermore, no contacts at <3.5 Å were found between part of one of the β-strands (at
135 residues 177-180) in the C_L domain, and the C_H domain in our Fab A33 homology model (Fig 1).
136
137 **Fig 1. Fab A33 structure with interface contacts highlighted.** Fab is composed of light
138 (magenta) and heavy (yellow) chains. Each chain contains variable (V_L and V_H) and constant (C_L
139 and C_H1) domains. The antigen-binding region at the complementary determining regions (CDRs;
140 blue), are located in the variable domains. There are five disulfide bonds (gray highlights).
141 Contacts between heavy and light chains within 3.5 Å are indicated with green dashed arrows. β
142 -strand 177-180 in C_L domain does not have contacts with C_H domain, zoom in right-inset.
143
144 Here, we compare the early unfolding events of Fab A33 at both high temperature and
145 low pH, using all-atom MD simulations. A common feature to both stress conditions was that
146 unfolding was initiated by the loss of interfacial contacts between neighboring domains, and that
147 domain unfolding occurred later. However, our results revealed different unfolding pathways for

the two stress conditions, leading to partial unfolding of only the C_L domain at low pH, compared to destabilization of both C_L and V_H domains in the high temperature condition. These conformational changes exposed different predicted aggregation-prone regions (APR), which would additionally support divergent aggregation mechanisms. Salt-bridge analysis provided insights into the location of those that were broken most rapidly due to protonation in the low pH simulation, and also showed that high temperature led to an increased fluctuation of salt bridge formation and breaking, more generally throughout the structure. An *in-silico* mutational analysis by both FoldX and Rosetta, predicted that the constant domain interface had the greatest potential for further stabilization, a finding that was also supported by lower packing-density calculations. Taken together, our results determined the stability-limiting regions at low pH and high temperature for Fab A33, and also identified those with the greatest potential for mutations that simultaneously improve stability under both low pH and high temperature conditions.

Results

Interface contacts, RMSD of individual domains and structural alignments revealed different unfolding pathways at low pH and high temperature

To determine which domains of Fab A33 are more susceptible to unfolding under low pH and high temperature, we first followed the RMSD of each individual domain (V_L, V_H, C_L and C_H1) along the simulations, as changes in RMSD are indicative of a conformational change. Simulations in the unfolding trajectories (pH 3.5 and pH 4.5 at 300 K, for low pH; pH 7.0 at 340 K and 380 K, for high temperature) were compared to the simulations in the native trajectory (pH 7.0 at 300 K). For every condition of pH and temperature, three independent simulations were performed, and their average RMSD are shown in Figures 2 and 3. Additionally, structures from the unfolding trajectories (pH 3.5, for low pH; 380 K, for high temperature) were aligned to structures from the native trajectory (pH 7.0 at 300 K), to visualize the structural changes that individual domains experienced. For each domain alignment, ten structures were taken every 3 ns

from each simulation repeat, from the 20-50 ns range at which the RMSD had stabilized. Thus, a total of thirty structures from each stress condition were compared to thirty from the native trajectory. We also monitored the number of interface contacts between domains (V_L-V_H and C_L-C_H1) during the simulations using a cutoff of 4 Å, to understand the temporal relationship between breakage of contacts in each interface, and the unfolding of each domain. The RMSD and radius of gyration (R_g) of the whole protein are also shown at every condition in Fig S2; where increased RMSD and R_g were observed at the conditions of low pH (3.5) and high temperature (380 K).

At low pH, almost no change was observed in the number of interfacial contacts within the variable region (V_L-V_H) of Fab A33, when compared to at pH 7.0. For pH 7.0, 333 ± 24 contacts were maintained (discarding the first frame), and 309 ± 24 contacts were maintained at pH 3.5 (Fig 2A). By contrast, a loss of interfacial contacts in the constant region (C_L-C_H1) was observed between pH 7.0 with 335 ± 17 contacts, and pH 3.5 with 265 ± 12 contacts (Fig 2B). Interestingly, this loss of constant region interfacial contacts at low pH took place very quickly, with pH 7.0 retaining 384 ± 14 contacts after 5 ns of the simulation, while simulations at pH 3.5 retained only 270 ± 11 contacts. This could be attributed to the lack of a well-defined hydrophobic core in the C_L-C_H1 interface, resulting in numerous early-disrupted contacts. Notably, C_L was the only domain to show a noticeable conformational change at low pH, revealed as an increase in RMSD from 1.8 ± 0.1 Å at pH 7.0 (calculated between 20-50 ns of the simulation), to 2.2 ± 0.1 Å at pH 3.5 (Fig 2E). This domain displacement occurred in the first 20 ns, after many interface contacts had already been lost with respect to pH 7.0, which suggests that destabilization of the C_L-C_H1 interface preceded and potentially accelerated the unfolding of the C_L domain. The other domains (V_L, V_H and C_H1) did not unfold significantly during the low pH simulations (Fig 2C, 2G and 2I). Structural alignments confirmed this result, showing remarkably good alignment between the pH 7.0 and pH 3.5 structures in each case for V_L, V_H and C_H1 (Fig 2D, 2H and 2J). Alignments of the C_L domain at pH 7.0 and pH 3.5 revealed a slight displacement at low pH, especially visible in the loop regions (Fig 2F). These findings agreed with previous experimental work, which combined

SAXS, atomistic modelling and smFRET to reveal the displacement of the C_L domain in Fab A33 at low pH [14].

Fig 2. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0, 4.5 and 3.5 (all 300 K). A, B) Contacts between light and heavy chains within 4.0 Å with simulation time, for variable (V_L-V_H) and constant (C_L-C_H) regions, respectively, pH values as labelled. C, E, G, I) RMSD of individual domains with simulation time for V_L, V_H, C_L and C_H1, respectively, pH values as labelled. In all cases, the average of three independent simulations is shown with the SEM as error. D, F, H, J) Alignments of structures from simulations at pH 7.0 and 3.5 for V_L, V_H, C_L and C_H1, respectively. Ten structures from the last 30 ns of each simulation were used, totaling thirty structures from pH 7.0 and thirty from pH 3.5.

For thermal denaturation, MD simulations are commonly run at temperatures as high as 500 K to attempt to fully denature the protein. Here, we aimed to capture the early thermal unfolding events of Fab A33, which involve only partial unfolding of the protein. For this reason, and to reflect experimental conditions more closely, lower temperatures of 340 K and 380 K were used in our simulations. Interfacial contacts were found to break across both the variable and the constant regions (Fig 3A and 3B). At 380 K, there was an average of only 220 ± 24 contacts in the variable interface and 204 ± 13 contacts in the constant interface. Those between the constant domains broke earlier than those of the variable domains, with only 218 ± 14 present after just 5 ns of the simulations at 380 K. This is consistent with previous reports, which also found that the constant region interface lost a larger fraction of its total interface contacts consistently faster than the variable region interface at high temperature, and also that domain unfolding occurred later than the loss of interfacial contacts [22]. Overall, more contacts were lost at both interfaces with high temperature than at low pH.

While V_L and C_H1 experienced only small domain displacements (Fig 3C and 3I), clear domain unfolding was observed for both C_L and V_H (Fig 3E and 3G). At 380 K from 20 ns to 50 ns, the V_H domain displayed an increase in RMSD from 2.4 Å at pH 7.0, to 3.2 Å at pH 3.5. That for the C_L domain increased from 1.8 Å at pH 7.0 to 2.4 Å at pH 3.5 (all averages were ± 0.1 Å). In these cases, many interface contacts were also lost with respect to pH 7.0 and 300 K, before the unfolding of individual structural domains, again consistent with destabilization of the interface contributing to the loss of stability for the individual domains. For both V_L and C_H1, structures from the simulations at 300 K and 380 K aligned well (Fig 3D and 3J), whereas for V_H and C_L (Fig 3H and 3F) the domains were structurally perturbed at the higher temperature. The V_H domain experienced a displacement of the loops on the N-terminal region, including the three CDR loops (Fig 3H). Differences in the C_L domain at high temperature were found in the loops and within an internal β-strand (Fig 3F). This was consistent with previous work which identified instability and structural changes in the V_H domain of another Fab at high temperature [22]. Taken together, these findings suggest a different unfolding pathway for Fab A33 at low pH and at high temperature.

Fig 3. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0 and temperatures 300 K, 340 K and 380 K. A, B) Contacts between light and heavy chains within 4.0 Å with simulation time, for variable (V_L-V_H) and constant (C_L-C_H) regions, respectively, temperature values as labelled. C, E, G, I) RMSD of individual domains with simulation time for V_L, V_H, C_L and C_H1, respectively, temperature values as labelled. In all cases, the average of three independent simulations is shown with the SEM as error. D, F, H, J) Alignments of structures from simulations at temperatures of 300 K and 380 K for V_L, V_H, C_L and C_H1, respectively. Ten structures from the last 30 ns of each simulation were used, totaling thirty structures from 300 K and thirty from 380 K.

Loss in β-strand secondary structure confirms regions of unfolding

The unfolding of individual domains was additionally followed by their loss in secondary structure (SS); specifically, we monitored the change in β-strand structure. Constant domains are composed of seven β-strands named A to G, while variable domains contain two more strands, a total of nine, with the two additional strands termed C' and C'' (Fig 4A and 4B). To calculate the loss in β-strand structure for each of the strands, we first tracked the secondary structure designation for each residue in Fab A33 throughout the simulations (Fig S3). The percentage of time occupied within β-strand was calculated for each residue, and then summed for each of the 32 β-strands in Fab A33. This value was averaged for each of the three repeats at each condition. The percentage change in β-strand occupancy was then calculated, to determine the loss relative to the reference simulations at pH 7, 300 K (Fig 4 and Fig S4).

Fig 4. Loss of secondary structure for each of the 32 β-strands of Fab A33. A, B) Strand order shown by lettering (A-G) for variable and constant domains, respectively. C, D) Percentage increase/decrease in β-strand secondary structure for each strand in Fab during the simulations, respect to pH 7.0 and 300 K, for: C) pH 3.5 and 300 K, D) pH 7.0 and 380 K. Error bars are the same and equal for positive and negative values.

At pH 3.5, the C_L domain had an overall loss in secondary structure content, confirming the results found in the previous section (Fig 4C). Strands F (-13 ± 3 %) and G (-12 ± 7 %) of the C_L domain showed the highest β-sheet structure loss, both located in the outer β-sheet. Strands B (-8 ± 6 %), C (-6 ± 4 %) and E (-8 ± 6 %) also experienced significant losses. Additionally, the β-strand C'' of the V_L domain also showed a large variability between repeat trajectories. C'' is the shortest strand, and is located at the extreme of the outer β-sheet connecting the CDR-2 loop, and so the large variability suggests that this is a flexible region prone to the loss of secondary structure in some but not all simulations.

At 380 K, the losses in β-strand content were located in both the C_L and V_H domains, consistent with the unfolding described in the previous section (Fig 4D). Many strands in the C_L domain showed significant losses, A (-7 ± 2 %), C (-8 ± 3 %), D (-16 ± 12 %) and G (-5 ± 1 %), located at the extremes of the inner and outer β-sheets. The V_H domain also showed high losses of β-strand. However, of these strands A (-18 ± 17 %) and G (-8 ± 9 %) also showed high variability between repeats. Interestingly, these same two strands in V_H were previously found to deform at high temperature in a different Fab [22]. Strand C" (-13 ± 4 %) of the V_H domain also showed a significant loss of β-strand content.

Salt bridge analysis identifies key stabilizing salt bridges

To identify the ionizable residues that potentially drive the conformational changes at low pH, a salt bridge analysis was performed. Salt bridges were identified over the simulation time for all the MD simulations carried out, using an O-N bond distance cutoff of 3.2 Å. From these, the occurrence (%) of each salt bridge during the simulation was calculated, and averaged for the three independent repeats at each condition. Lastly, the most persistent salt bridges at each condition were highlighted in the Fab A33 structure (Fig 5 and Fig S5).

Fig 5. Salt bridge analysis. A, B, C) Salt bridges formed during the simulation time for representative MD simulations at A) pH 7.0 and 300 K, B) pH 3.5 and 300 K and C) pH 7.0 and 380 K. Presence of a salt bridge is indicated in white and absence in black. D, E, F) List of salt bridges and its occurrence for simulations at D) pH 7.0 and 300 K, E) pH 3.5 and 300 K and F) pH 7.0 and 380 K. Values shown are the average of three independent simulations with their SEM as error. The more persistent salt bridges are highlighted for pH 7.0 (green), pH 3.5 (red) and pH 7.0 and 380 K (blue). G, H, I) The more persistent salt bridges are mapped into the Fab A33 structure. Two key stabilising salt bridges (Glu165-Lys103 and Glu195-Lys149) are highlighted in a dashed circle.

303

304 At pH 7.0 and 300 K, a total of 36 salt bridges were present. Interestingly, many of these
305 salt bridges were flexible and able to form with different partners during a single trajectory, such
306 as Asp122 which partnered with both Lys126 and Lys436, or Asp304 which paired with both
307 Arg252 and Arg281. This is consistent with previous work, which found that salt bridges break
308 and reform, and not always with the same partner [25]. The most persistent (as % of time present)
309 salt bridges at pH 7.0 were Glu165-Lys103 (67 ± 4 %), Glu195-Lys149 (65 ± 4 %), Asp82-Arg61
310 (61 ± 14 %), Asp151-His189 (58 ± 24 %), and Asp304-Arg252 (55 ± 27 %), as shown in Fig 5A,
311 5D and 5G.

312 At low pH, pH 3.5, 300 K, a total of 27 salt bridges were observed, but most of them were
313 very short lived. The more persistent salt bridges at pH 3.5 were Asp151-His189 (86 ± 6 %),
314 Asp82-Arg61 (78 ± 11 %), Asp122-Lys126 (73 ± 1 %), Asp362-Lys335 (56 ± 6 %) (Fig 5B, 5E
315 and 5H). The protonation state at the end of the pH 3.5 simulations was calculated again using
316 these Fab conformations, which revealed these salt bridges to be still present due to predicted
317 pK_a values of below 3.5 for these aspartates. Comparison of the salt bridges at pH 7.0 and 3.5,
318 indicated the presence of two salt bridges that potentially trigger the conformational change
319 observed at low pH, and thus the loss of Fab A33 stability. Glu165-Lys103 and Glu195-Lys149,
320 were the most persistent contacts at pH 7.0, but were not present at pH 3.5. Glu165-Lys103
321 bridges the C_L domain to the V_L domain, and Glu195-Lys149 bridges the outer β-strands C and
322 F of the C_L domain. Loss of these salt bridges at low pH, would therefore destabilize the C_L
323 domain, and promote the observed C_L domain displacement.

324 At high temperature, pH 7.0 and 380 K, a total of 45 salt bridges were observed. The
325 greater number than at 300 K, reflects an increased conformational flexibility of many salt bridges
326 at the higher temperature, in which they often broke, but then reformed with a different partner.
327 Indeed, at the high temperature, salt bridges broke and reformed much faster (Fig 5C). At 380 K,
328 the total time present for the most persistent salt bridges observed at 300 K, had decreased to 47

329 ± 9 % for Glu165–Lys103, 58 ± 3 % for Glu195-Lys149, 21 ± 7 % for Asp151-His189, and 43 ±
330 22 % for Asp304-Arg252. However, Asp82-Arg61 increased in occurrence to 70 ± 3 % (Fig 5C,
331 5F and 5I). These findings indicate that the increased dynamics at high temperature, results in
332 constant rupture and formation of salt bridges, and this transient disruption leaves Fab A33 more
333 susceptible to unfolding.
334
335 **FoldX, Rosetta and packing density calculations predict sub-optimal stability of C_L and the**
336 **C_L-C_H1 interface**
337 Computational tools such as FoldX and Rosetta-ddG [26,27] predict the relative changes
338 in folding free energy ($\Delta\Delta G$) between the Gibbs free energies (ΔG) of the protein carrying a
339 simulated point mutation and the wild-type protein, to find those mutations that will most
340 significantly reduce the free energy of the protein. These approaches are often also combined to
341 find consensus predictions [28,29]. To predict stabilizing mutations in Fab A33, we calculated the
342 $\Delta\Delta G$ with both FoldX and Rosetta-ddG, of all possible single-mutant variants when accessing all
343 19 other substitutions across the 442 residue positions in Fab A33, totaling 8398 mutations. FoldX
344 identified 1879 of these mutations as stabilizing (22.4 %), while Rosetta-ddG identified 2386 (28.4
345 %). Of these, 956 (11.4 %) were predicted by both algorithms. Fig S6 shows the correlation
346 between the mutations predicted by FoldX and Rosetta, and Table S2 lists the 25 most stabilizing
347 mutations predicted by both algorithms, with their respective $\Delta\Delta G$ values.
348 Fig 6A compares the greatest stabilization predicted by FoldX and Rosetta, at each of the
349 442 residues in Fab A33, regardless of the specific mutation selected by each algorithm. Six
350 residues were highlighted in magenta in Fig 6A and 6B, as those predicted by both algorithms to
351 have the greatest potential for stabilization. These residues corresponded to S176, N137, S397,
352 S159, S12 and T180, and all of their mutations predicted to be most stabilizing, were to more
353 hydrophobic amino acids, such as Trp, Leu, Ile, Phe and Tyr (Fig S6 and Table S2). Four of the
354 six residues (S176, N137, S397 and T180), are located in the constant domain interface, between

355 C_L and C_H1 domains (Fig 6B), suggesting significant potential for further stabilization within the
356 C_L-C_H1 interface. Furthermore, S159 is in the C_L domain, but interacts with an outer β-strands of
357 C_L, and S12 is in the V_L domain, but interacts with the C_L domain (Fig 6B). Thus overall, the C_L
358 domain has a relatively high potential for stabilization, through repacking of the C_L-C_H1 interface,
359 within the C_L domain itself, and also through improved interaction between C_L and V_L. This is
360 consistent with the MD simulations which found the displacement of C_L away from the interface
361 with C_H, and subsequent unfolding of the C_L domain, to be critical steps in early or partial
362 unfolding.

363 Further highly stabilizing mutations at residues S395 and S267, as predicted only by
364 FoldX, are highlighted in green in Fig 6. S395 is also located in the C_L-C_H1 interface. However,
365 S267 is in CDR2 of the V_H domain, and so not a good candidate for general framework
366 stabilization due to its role in antigen binding. Highly stabilizing mutations at residues K103 and
367 T72, as predicted only by Rosetta-ddG are highlighted in yellow. These were both located in the
368 V_L domain, but K103 also interacts with the C_L domain, further suggesting that the interactions
369 within and around the C_L domain are the least optimized for stability.
370
371 **Fig 6. Predicted residues that can be stabilized further by FoldX and Rosetta-ddG.** A)
372 Correlation between FoldX and Rosetta predictions. Residues predicted by both software to be
373 most stabilizing are shown in magenta on the bottom left. Residues predicted only by FoldX to be
374 stabilizing are shown in green and mutations predicted only by Rosetta in yellow. B) Residues
375 predicted to be stabilized further the most are mapped in Fab A33 structure, following the same
376 colour scheme as in A).
377
378 The packing density of each Fab A33 residue was calculated using the package occluded
379 surface (OS) software, which calculates occluded surface and atomic packing [30,31]. The
380 occluded surface packing (OSP) value of each atom is calculated from normal vectors that extend

outward from the atom surface until they intersect a neighboring van der Waals surface (Fig S7). This value is 0.0 for completely exposed residues and 1.0 where 100% of molecular surface is in contact with other van der Waals surface. The average OSP for all 28 β-strand residues within domain interfaces (V_L-V_H and C_L-C_H1) was 0.49 ± 0.01 (OSP values shown in Table S3). By contrast, the average OSP of the five constant-domain interface residues (S176, N137, S397, T180, and S395), identified by FoldX and Rosetta as having high stabilization potential, was 0.41 ± 0.05 (OSP values shown in Table S3). This can be visualized in Fig 7, where OSP values were mapped in the structure of Fab A33, with red to indicate high packing density, and blue to indicate low packing density. β-strand residues within domain interfaces were highlighted as sticks (Fig 7A). Residues in the constant interface (C_L-C_H1) were lighter colored than residues in the variable interface (V_L-V_H), indicating less tight packing of the constant interface. An insight of the residues identified by FoldX and Rosetta is provided in Fig 7B, where a lighter color than the residues in the variable interface was also observed. This result shows that the predicted residues are under-packed, and therefore have the potential to be mutated to pack the C_L-C_H1 interface more tightly.

Fig 7. Packing density of every residue in Fab A33, computed using Occluded Surface. A) The occluded surface packing (OSP) values were added as B-factors to the PDB file for the Fab A33 homology model. High packing values are shown in red and low values in blue. Residues in β-strands within domain interfaces (V_L-V_H and C_L-C_H1) are highlighted in sticks and ball representation. B) Residues identified by FoldX and Rosetta that could be stabilised further (S176, N137, S397, T180, and S395) are highlighted in sticks and ball, and sphere representation.

Comparison to natural sequence variations in Fabs

The natural variability of Fab sequences was identified from within one hundred light chains and one hundred heavy chains curated from those available in the Protein Data Bank [32]. Sequence alignment and sequence entropy calculations for each residue were obtained using

Bioedit [33]. An entropy of zero indicates a fully conserved residue, whereas 3.04 is the maximum entropy, originating from 21 possibilities (all amino acids plus the stop codon). There is significant positional bias in the sequence variability of Fabs due to the hypervariability of the CDRs [34], the presence of kappa (κ) and lambda (λ) light chain isotypes, allotypic diversity across individuals, and idiotypic variability within the variable domains of individuals. The sequence entropy analysis (Fig S8) clearly shows this, with the highest sequence entropy (>2), for the six CDRs, and a slightly lower variability on average within the C_H domain. Similarly, the higher sequence entropy on average for the C_L domain, compared to the C_H domain results from the grouping of kappa (κ) and lambda (λ) light chain isotypes.

Even though residues in the constant regions have more restricted variability, many natural variations are still observed that may affect stability. Except for the fully conserved S176, the sites predicted as having the most potential for stabilizing mutations by both FoldX and Rosetta, had natural variations, with sequence entropy values of N137: 0.74, S397: 0.15, S159: 0.96, S12: 0.96, T180: 0.43. This was also true for mutations identified only by FoldX (S267: 1.69 and S395: 0.69), and only by Rosetta-ddG (K103: 0.16, T72: 0.91). Comparisons between the existing mutations and the stabilizing mutations suggested by FoldX and Rosetta are shown in Table 1. In general, the mutations found in existing Fabs were conservative changes to residues with properties similar to the original residue. For instance, S397 and S395 are only naturally mutated to Thr, whereas T180 can also be Ser, and K103 can be Arg. By contrast, FoldX and Rosetta predictions were typically from polar to more hydrophobic residues, typically to Ile, Leu, Trp, Val, and Tyr. A few suggested mutations were also found naturally, such as N137I, S12Y, and S267P. S159 also shows the potential to be mutated to more hydrophobic residues, Val and Met. Overall, this analysis shows that despite their low natural variability, many residues in the constant domains had significant scope for stabilization through non-natural mutations.

Table 1. Comparison between the mutations in existing human and mouse Fabs and the stabilizing mutations suggested by FoldX and Rosetta.

Original Residue	Light chain				Heavy Chain		Suggested mutation by FoldX and Rosetta
	Kappa (κ)		Lambda (λ)				
	Human	Mouse	Human	Mouse	Human	Mouse	
S176	-	-	-	-	N.A.	N.A.	W, M, R, L Y
N137	-	-	-	T, I	N.A.	N.A.	L, M, I
S397	N.A.	N.A.	N.A.	N.A.	-	T	I, L, W, V
S159	-	V	V	M	N.A.	N.A.	R, F
S12	-	Y, P, A, T	-	T	N.A.	N.A.	F, Q, Y
T180	-	-	S	-	N.A.	N.A.	Y, W
S395	N.A.	N.A.	N.A.	N.A.	-	T	L, M, R, I, V
S267	N.A.	N.A.	N.A.	N.A.	P, W, Y, T, G, D, N, A	P, I, G, N, D, Q, L	P
K103	R	-	-	-	N.A.	N.A.	Y, F, T
T72	S	S	S	A	N.A.	N.A.	Y

N.A. (Not Applicable), mutation does not apply to the chain (light or heavy); -, no existing mutations were found on that chain.

Solvent exposure of different aggregation-prone regions promotes different aggregation pathways for low pH and high temperature

The aggregation pathways of Fab A33 at low pH and high temperature at pH 7.0, are already known to result in different aggregate morphologies [12]. Here we explored whether the

two conditions also exposed different aggregation-prone regions (APRs). APRs can be predicted from sequence information, and either assume a fully unfolded protein, or otherwise refine the prediction by factoring solvent exposure of the APR based on structure and dynamics information. The sequence-based predictions are based on either the intrinsic properties of amino acids, or their compatibility with protein structural features in known amyloid fibril structures. Examples of sequence-based predictors include PASTA 2.0 [35], TANGO [36], AGGRESCAN [37], MetAmyl [38], FoldAmyloid [39] and Waltz [40]. As the ability of APRs to trigger aggregation depends upon their solvent accessibility, more recent structure-based predictors consider the three-dimensional structure of the protein and in some cases also their potential modes of partial unfolding. Examples include AGGRESCAN 3D [41], AggScore [42], SAP [43] and Solubis [44]. Here, we want to compare the solvent accessibility of APRs in Fab A33, between our MD simulations at the unfolding conditions and at the reference trajectory. Thus, we used sequence-based APR predictors to determine the APRs in Fab A33, and then determined their solvent accessible surface area (SASA) from the MD simulations, for relative comparisons.

We used four sequence-based predictors to determine the APRs in Fab A33, PASTA 2.0, TANGO, AGGRESCAN and MetAmyl. APRs were selected when three out of the four predictors identified an aggregation-prone sequence (Fig S9A). Seven APRs were found, namely residues 31-36, 47-51, 114-118 and 129-139 in the light chain and residues 261-165, 325-329 and 387-402 in the heavy chain. The presence of these APRs was confirmed using Amylpred2 [45], a consensus tool of eleven existing algorithms (Fig S9A). The location of the seven APRs revealed that they are all located in the interior of Fab A33, and thus protected from the solvent (Fig S9B). Exposure of one of these APRs as a result of a conformational change by an environmental stress, has the potential to trigger aggregation. Thus, the SASA of each APR during the simulations was calculated, as well as the difference in solvent accessibility, ΔSASA, between unfolding conditions and the reference simulation (Table 2).

At low pH, only one APR (residues 387-402), was found to increase its solvent accessibility

467 significantly at pH 3.5, with an increase of 57 ± 25 Å^2 (10 % increase) (Table 2), consistent with
468 previous experimental findings [14]. This APR is located in the C_H1 domain and its exposure can
469 be explained by the C_L domain displacement observed at low pH (Fig 8A). At high temperature,
470 two APRs were found to increase their solvent accessibility, APR 261-265 located in V_H and APR
471 129-139 located in C_L (Fig 8B). APR 261-265 increased its SASA 15 ± 6 Å^2 (107 % increase) and
472 APR 129-139 increased its SASA 20 ± 16 Å^2 (13 % increase) (Table 2). The location of these
473 APRs agrees with the domains found to unfold previously at high temperature. Notably, the APRs
474 exposed at low pH and high temperature are different, suggesting the potential to follow different
475 aggregation mechanisms depending on the stress applied.
476

477 **Table 2. SASA of the APRs in Fab A33 during simulations and SASA differences between**
478 **unfolding simulations and the reference simulation.** Solvent accessible surface area of the
479 seven aggregation-prone regions in Fab A33 during all simulations, and relative differences
480 (ΔSASA) between the unfolding trajectories (pH 3.5 and 4.5 at 300 K, for low pH; pH 7.0 at 340
481 K and 380 K, for high temperature) and the reference trajectory (pH 7.0 and 300 K).

APR	Fab domain	SASA (Å^2) pH 7.0 300 K	SASA (Å^2) pH 4.5 300 K	ΔSASA (Å^2) pH(4.5-7.0)	SASA (Å^2) pH 3.5 300 K	ΔSASA (Å^2) pH(3.5-7.0)	SASA (Å^2) pH 7.0 340 K	ΔSASA (Å^2) T(340-300 K)	SASA (Å^2) pH 7.0 380 K	ΔSASA (Å^2) T(380 - 300 K)
31-36	V_L	118 ± 4	112 ± 5	-5 ± 7	115 ± 16	-3 ± 17	120 ± 8	3 ± 9	106 ± 4	-12 ± 6
47-51	V_L	100 ± 1	104 ± 7	4 ± 13	115 ± 20	15 ± 23	101 ± 6	1 ± 13	96 ± 7	-4 ± 13
114-118	C_L	125 ± 4	121 ± 7	-4 ± 8	110 ± 7	-15 ± 8	112 ± 2	-13 ± 5	120 ± 13	-5 ± 13
129-139	C_L	152 ± 8	147 ± 11	-5 ± 13	151 ± 12	0 ± 14	165 ± 10	13 ± 13	172 ± 14	20 ± 16

261-265	V_H	14 ± 2	13 ± 0	-1 ± 2	13 ± 1	-1 ± 3	19 ± 1	5 ± 3	29 ± 5	15 ± 6
325-329	V_H	122 ± 16	115 ± 14	-7 ± 21	119 ± 7	-3 ± 17	88 ± 10	-34 ± 19	120 ± 22	-1 ± 27
387-402	C_H1	552 ± 21	547 ± 25	-5 ± 33	609 ± 12	57 ± 25	544 ± 7	-9 ± 22	489 ± 48	-15 ± 21

Fig 8. Fab A33 predicted APRs that increase its solvent accessibility at low pH and high temperature. A) APR 387-402 increases its SASA at pH 3.5 and B) APR 261-265 and APR 129-139 increase its SASA at 380 K (Table 1). All mapped in red in Fab A33 homology structure.

Discussion

Antibody-based products are the main class of approved biopharmaceuticals, due to their high target specificity [1]. However, there are many barriers to their successful development into therapeutics, with protein aggregation being perhaps the most common and challenging to prevent [5]. There is a need to identify potential instabilities of therapeutic proteins during their early development, particularly against stresses that they will encounter during manufacture, storage and delivery. This would allow their early elimination from further development, or otherwise rational mutagenesis into more stable products. In this context, we have elucidated the first unfolding events that take place on a humanized Fab A33 using atomistic MD simulations, and compared these to predictions of potentially stabilising mutations using computational tools.

Our simulations showed that contacts at the interface between domains (V_L-V_H and C_L-C_H1) were lost before individual domains unfolded. Interfacial contacts in the constant domain were the least stable, and were lost very quickly during the simulations under both stresses of low pH and high temperature. In line with these results, FoldX and Rosetta agreed that the residues

503 that can be stabilized the most, are located in the constant domain interface. Further validation
504 was provided by packing density calculations, which revealed that the residues identified by the
505 stability predictors, were under packed relative to the other residues located in the interface
506 between domains. Based on these findings, we speculate that improvement of Fab A33 stability
507 should be targeted to the constant domain interface. Only one of the top mutations suggested by
508 FoldX and Rosetta, N137I was found to be present in our analysis of natural variation within
509 existing Fab sequences. However, there was significant scope for improvement through mutating
510 the interfacial residues S176, N137, S397, T180, and S395, to the suggested hydrophobic
511 residues.

512 The further goal could be to improve the stability of the individual domains. The C_L domain
513 was found to unfold at both, low pH and high temperature. Salt bridge analyses identified two key
514 salt bridges at the heart of this domain unfolding at low pH, Glu165-Lys103 and Glu195-Lys149.
515 Glu165-Lys103 bridges the C_L domain to the V_L domain, and Glu195-Lys149 is located in outer
516 β-strands of the C_L domain, bridging β-strands C and F. FoldX and Rosetta also identified
517 stabilizing mutations in the C_L domain. To stabilize the interaction between the C_L and V_L domain,
518 hydrophobic mutants of S12 and K103 were predicted. Interestingly, the mutation S12 to Tyr is
519 found naturally. In the C_L domain, S159 was identified, which interacts with an outer β-strand,
520 suggesting this interaction can also be improved. Lastly, the V_H domain was also found to unfold
521 at high temperature. The only mutation identified in this domain is S267, to a Pro, which notably
522 is found naturally. Overall, the results found with MD simulations and stabilizing software
523 predictors strongly agree in the domains of Fab A33 that can be stabilized further.

524 In order to gain insights into the mechanisms by which aggregation might occur, APRs in
525 Fab A33 were identified, and their solvent accessibilities were compared. All APRs in Fab A33
526 are located in the interior of the protein, however, at low pH and high temperature the SASA of
527 certain APRs increased. Notably, different APRs were exposed under both stresses, suggesting
528 that different aggregation mechanisms occur under each stress. This result stresses the

importance of identifying the stability of a protein under the different stresses it might encounter. Taken together, this work provides insights into the stability and robustness of the therapeutically relevant Fab A33, and offers a path to the engineering and design of a more aggregation resistant antibody fragment.

Materials and Methods

Fab A33 homology model

The homology model of wild-type Fab A33 was built from the crystal structure of human germline antibody 5-51/O12 (PDB ID: 4KMT) and the amino-acid sequence of Fab A33 (Fig. S1) [46,47]. The C226S heavy-chain variant was used to avoid the formation of linked Fab dimers. We used the Rosetta method “minirosetta”, as detailed in previous works [48].

Molecular dynamics simulations

Molecular dynamic (MD) simulations on the Fab A33 homology model were conducted in Gromacs v5.0 [49]. MD simulations were carried out at neutral pH and room temperature (pH 7.0 and 300 K) and under two stresses, low pH (pH 3.5 and pH 4.5 at 300 K) and high temperature (pH 7.0 at 340 K and 380 K). Many high temperature simulations are performed at relatively high temperatures (e.g. 500 K), to achieve complete denaturation of the protein, however, in this case, we aim to partially unfold Fab A33 and detect the regions prone to early unfolding. Simulations were carried out using the OPLS-AA/L all-atom force field [50]. The Fab PDB file was first converted to a topology file with its five (four intra-chain and one inter-chain) disulfide bonds retained. The protonation state of each residue was entered manually, and these were determined

553 at each pH using the PDB2PQR server, which performed the pKa calculations by PropKa [51].
554 This gave the following total charges: +9 (pH 7.0), +18 (pH 4.5) and +35 (pH 3.5). The Fab A33
555 structure was centered in a cubic box with a layer of water up to at least 10.0 Å from the protein
556 surface. The box was solvated with SPC/E water molecules, Cl- added to neutralize the net
557 charges, and NaCl added to an ionic strength of 50 mM for all simulations. The system was energy
558 minimized using the steepest descent algorithm (2000 steps) followed by the conjugate gradient
559 method (5000 steps). The solvent and ions around the protein were equilibrated in position-
560 restricted simulations for 100 ps under NVT ensemble to stabilize at the specified temperature,
561 and then at 100 ps under NPT ensemble to stabilize at atmospheric pressure. Lastly, MD
562 simulations were carried out for 50 ns in triplicates under the five conditions (pH 7.0 and 300 K;
563 pH 4.5 and 300 K; pH 3.5 and 300 K; pH 7.0 and 340 K; pH 7.0 and 380 K). Jobs were submitted
564 to the UCL Legion High Performance Computing Facility. The time step of the simulations was
565 set to 2 fs and trajectories were saved every 10 ps.
566
567 **Analysis of MD trajectories**
568 MD trajectories were saved reduced, every 0.2 ns (total of 250 frames). Interface contacts
569 over simulation time were calculated using the native contacts extension of the visual molecular
570 dynamics (VMD) program [52]. A cutoff distance of 4 Å was used in the calculations. Variable
571 domain contacts (V_L-V_H) were calculated between residues 1-108 (V_L) and 215-334 (V_H). Constant
572 domain contacts (C_L-C_H1) were calculated between residues 109-214 (C_L) and 335-442 (C_H1).
573 RMSD of individual domains during the simulations were calculated using the RMSD trajectory
574 tool in VMD. All the structures of the trajectory were first aligned and the RMSD was calculated
575 (no hydrogens included). Domains were V_L (1-108), V_H (215-334), C_L (109 to 214) and C_H1 (335
576 to 429). Averages and SEM of three independent repeats are shown. Structural alignments of the
577 last 30 ns of the trajectories were also performed using VMD. Secondary structure (SS)
578 assignments of each residue along the trajectory were done using the DSSP module [53,54]. To

analyze the loss in β-strand structure, we monitored the percentage of β-sheet SS per residue. These values were summed for each of the 32 β-strands in Fab A33 and differences were calculated between the unfolding simulations and the reference simulations (pH 7 and 300 K). Lastly, salt bridges were calculated along the trajectories using VMD and a cutoff distance between O and N groups of 3.2 Å. From these, the occurrence (%) of each salt bridge during the simulation was calculated, and averaged for the three independent repeats at each condition.

Mutational study and ΔΔG calculations by FoldX and Rosetta

The effect of mutations on the stability of Fab A33 was studied using FoldX (foldx.crg.es) [26] and the Rosetta method "ddg_monomer" (www. rosettacommons.org) [27]. Both tools predicted the difference in folding free energy, ΔΔG, between the protein carrying a point mutation and the wildtype. Each of the 442 residues in the Fab A33 were mutated to the other 19 possibilities, totaling 8398 single mutants. FoldX was used as a plugin in the graphical interface YASARA [55]. The "Repair" command was used first to energy minimize the homology model of Fab A33, by rearranging the amino acid side chains. Next, the "BuildModel" command was used to introduce the point mutations, optimize the structure of the new protein variant, and calculate the stability change upon mutation. The Rosetta "ddg_monomer" method was used, where an example of mutation and option files, listing the parameters of the executable, can be seen in previous work [48]. Jobs were submitted to the UCL Legion High-Performance Computing Facility.

Packing Density

Occluded surface (OS) program was used to calculate the atomic packing of Fab A33 [30,31]. The occluded surface packing (OSP) values are useful for identifying regions of loose packing in a protein. OSP value for each residue are calculated from the collection of extended normals (ray-lengths) that extend outward from the molecular surface until they intersect

neighboring van der Waals surface. Analysis of these normals, their respective lengths and the surface area involved in the interaction, defines the packing of each atom in the protein.

Sequence Entropy of Fab sequences

Fab sequences were retrieved from the Protein Data Bank (PDB) [32], totaling one hundred light chains and one hundred heavy chains. For light chains, kappa (κ) and lambda (λ) chains were included. For κ light chains, λ light chains and heavy chains, sequences from the species human and mouse were used. Sequence alignment and calculation of the sequence entropy for each residue were calculated using Bioedit [33]. The maximum entropy for 21 possible amino acids (including stop codon) is 3.04 and zero represents a fully conserved residue.

Aggregation-prone regions (APR) predictions

Aggregation prone regions (APR) of Fab A33 were predicted using PASTA 2.0 [35], TANGO [36], AGGRESCAN [37] and MetAmyl [38], using the protein sequence as input. The regions in which three out of the four software identified an APR were selected, resulting in seven APRs. Amylpred2 consensus tool was used to confirm the presence of these APRs [45]. To calculate the solvent accessible surface area (SASA) of each APR during the trajectories, the average area per residue over the trajectory was calculated first, using gromacs analysis tool "sasa", then summed for each APR.

ACKNOWLEDGMENTS

We thank the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Emergent Macromolecular Therapies (EP/L015218/1) (N.C.C.), the EPSRC Future Targeted Healthcare Manufacturing Hub (EP/P006485/1, EP/I033270/1) (N.C.), and EPSRC EP/N025105/1 (C.Z.).

655

656

657 **REFERENCES**

658 1. Ecker DM, Jones SD, Levine HL. The therapeutic monoclonal antibody market. MAbs.
659 2015;7(1):9–14.
660 2. Carter PJ, Lazar GA. Next generation antibody drugs: Pursuit of the “high-hanging fruit.”
661 Nat Rev Drug Discov. 2018;17:197–223.
662 3. Nelson AL. Antibody fragments: hope and hype. MAbs. 2010;2(1):77–83.
663 4. Enever C, Batuwangala T, Plummer C, Sepp A. Next generation immunotherapeutics-
664 honing the magic bullet. Curr Opin Biotechnol. 2009;20(4):405–11.
665 5. Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein
666 pharmaceuticals: An update. Pharm Res. 2010;27(4):544–75.
667 6. Wang W, Nema S, Teagarden D. Protein aggregation-pathways and influencing factors.
668 Int J Pharm. 2010;390(2):89–99.
669 7. Wang W. Protein aggregation and its inhibition in biopharmaceutics. Int J Pharm.
670 2005;289(1–2):1–30.
671 8. Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. Nat
672 Chem Biol. 2009;5:15–22.
673 9. Neudecker P, Robustelli P, Cavalli A, Walsh P, Lundström P, Zarrine-Afsar A, et al.
674 Structure of an intermediate state in protein folding and aggregation. Science.
675 2012;336(6079):362–6.
676 10. Canet D, Last AM, Tito P, Sunde M, Spencer A, Archer DB, et al. Local cooperativity in
677 the unfolding of an amyloidogenic variant of human lysozyme. Nat Struct Biol.
678 2002;9:308–15.
679 11. Kendrick BS, Carpenter JF, Cleland JL, Randolph TW. A transient expansion of the
680 native state precedes aggregation of recombinant human interferon-γ. Proc Natl Acad

Sci. 1998;95(24):14142–6.

12. Chakroun N, Hilton D, Ahmad SS, Platt GW, Dalby PA. Mapping the Aggregation Kinetics of a Therapeutic Antibody Fragment. Mol Pharm. 2016;13(2):307–19.

13. De Baets G, Schymkowitz J, Rousseau F. Predicting aggregation-prone sequences in proteins. Essays Biochem. 2014;56:41–52.

14. Codina N, Hilton D, Zhang C, Chakroun N, Ahmad SS, Perkins SJ, et al. An Expanded Conformation of an Antibody Fab Region by X-Ray Scattering, Molecular Dynamics and smFRET Identifies an Aggregation Mechanism. J Mol Biol. 2019;431(7):1409–25.

15. Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. Pharm Res. 2003;20(9):1325–36.

16. Lindorff-Larsen K, Trbovic N, Maragakis P, Piana S, Shaw DE. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. J Am Chem Soc. 2012;134(8):3787–91.

17. Rocco AG, Mollica L, Ricchiuto P, Baptista AM, Gianazza E, Eberini I. Characterization of the protein unfolding processes induced by urea and temperature. Biophys J. 2008;94(6):2241–51.

18. Salimi NL, Ho B, Agard DA. Unfolding simulations reveal the mechanism of extreme unfolding cooperativity in the kinetically stable α-lytic protease. PLoS Comput Biol. 2010;6(2):e1000689.

19. Settanni G, Fersht AR. High temperature unfolding simulations of the TRPZ1 peptide. Biophys J. 2008;94(11):4444–53.

20. Collu F, Spiga E, Chakroun N, Rezaei H, Fraternali F. Probing the early stages of prion protein (PrP) aggregation with atomistic molecular dynamics simulations. Chem Commun. 2018;54(57):8007–10.

21. Patel D, Kuyucak S. Computational study of aggregation mechanism in human

lysozyme[D67H]. PLoS One. 2017;12(5):1–17.

22. Buck PM, Kumar S, Singh SK. Insights into the potential aggregation liabilities of the b12 Fab fragment via elevated temperature molecular dynamics. Protein Eng Des Sel. 2013;26(3):195–206.

23. Su JG, Zhang X, Han XM, Zhao SX, Li CH. The intrinsic dynamics and unfolding process of an antibody fab fragment revealed by elastic network model. Int J Mol Sci. 2015;16(12):29720–31.

24. Wang T, Duan Y. Probing the stability-limiting regions of an antibody single-chain variable fragment: a molecular dynamics simulation study. Protein Eng Des Sel. 2011;24(9):649–57.

25. Kortkhonjia E, Brandman R, Zhou JZ, Voelz VA, Chorny I, Kabakoff B, et al. Probing antibody internal dynamics with fluorescence anisotropy and molecular dynamics simulations. MAbs. 2013;5(2):306–22.

26. Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E. Predicting folding free energy changes upon single point mutations. Bioinformatics. 2012;28(5):664–71.

27. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79(3):830–8.

28. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. Computationally designed libraries for rapid enzyme stabilization. Protein Eng Des Sel. 2014;27(2):49–58.

29. Buß O, Rudat J, Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J. 2018;16:25–33.

30. Pattabiraman N, Ward KB, Fleming PJ. Occluded molecular surface: analysis of protein packing. J Mol Recognit. 1995;8(6):334–44.

31. Fleming PJ, Richards FM. Protein packing: dependence on protein size, secondary structure and amino acid composition. J Mol Biol. 2000;299(2):487–98.

32. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, et al. The RCSB

Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013;41(D1):D475–82.

33. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999;41:95–8.

34. Elgert KD. Antibody structure and function. In: Immunology: Understanding the Immune System. John Wiley & Sons; 1998. p. 58–78.

35. Walsh I, Seno F, Tosatto SCE, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res. 2014;42(W1):301–7.

36. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol. 2004;22(10):1302–6.

37. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. BMC Bioinformatics. 2007;8(1):65.

38. Emily M, Talvas A, Delamarche C. MetAmyl: A METa-predictor for AMYLoid proteins. PLoS One. 2013;8(11).

39. Garbuzynskiy SO, Lobanov MY, Galzitskaya O V. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics. 2010;26(3):326–32.

40. Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods. 2010;7(3):237–42.

41. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Res. 2015;43(W1):W306–13.

42. Sankar K, Krystek SR, Carl SM, Day T, Maier JKX. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins.

2018;86(11):1147–56.

43. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci. 2009;106(29):11937–42.

44. Van Durme J, De Baets G, Van Der Kant R, Ramakers M, Ganesan A, Wilkinson H, et al. Solubis: A webserver to reduce protein aggregation through mutation. Protein Eng Des Sel. 2016;29(8):285–9.

45. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ. A Consensus Method for the Prediction of “Aggregation-Prone” Peptides in Globular Proteins. PLoS One. 2013;8(1):1–6.

46. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 2006;34(17):1–18.

47. Raman S, Vernon R, Thompson J, Tyka M, Pei J, Kim D, et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins. 2009;77(Suppl 9):89–99.

48. Zhang C, Samad M, Yu H, Chakroun N, Hilton D, Dalby PA. Computational Design to Reduce Conformational Flexibility and Aggregation Rates of an Antibody Fab Fragment. Mol Pharm. 2018;15(8):3079–92.

49. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2:19–25.

50. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B. 2001;105(28):6474–87.

51. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pK a values. Proteins Struct Funct Genet. 2005;61(4):704–21.

52. Humphrey W, Dalke A, Klaus S. VMD: Visual Molecular Dynamics. J Mol Graph.

1996;14(1):33–8.

53. Touw WG, Baakman C, Black J, Te Beek TAH, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015;43(D1):D364–8.

54. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.

55. van Durme J, Delgado J, Stricher F, Serrano L, Schymkowitz J, Rousseau F. A graphical interface for the FoldX forcefield. Bioinformatics. 2011;27(12):1711–2.

811

812

813 **Supporting information captions**

814

815 **Fig S1. Fab A33 sequence.** Fab A33 amino acid sequence separated by domains (V_L, C_L, V_H,
816 C_H1 and hinge region). The six CDRs in the V_L and V_H domains are highlighted in red.

817

818 **Table S1. Residues located in the interface between light and heavy chains in Fab A33.**

819

820 **Fig S2. All protein RMSD and radius of gyration (R_g) with simulation time.** A, B) RMSD of
821 the whole protein with simulation time for different A) pHs and B) temperatures, values as labelled.
822 In all cases, the average of three independent simulations is shown with the SEM as error. C, D)
823 Radius of gyration (R_g) of Fab A33 with simulation time for different A) pHs and B) temperatures,
824 values as labelled. In all cases, the average of three independent simulations is shown with the
825 SEM as error.

826

827 **Fig S3. Secondary structure (SS) of each residue in Fab A33 with simulation time,**
828 **calculated using DSSP**. Representative SS evolution are shown for A) pH 7.0 and 300K, B) pH
829 3.5 and 300K and C) pH 7.0 and 380 K, secondary structure type as indicated in the legend.

830

831 **Fig S4. Loss of secondary structure for each of the 32 β-strands of Fab A33.** A, B) Strand
832 order shown by lettering (A-G) for variable and constant domains, respectively. C, D) Percentage
833 increase/decrease in β-strand secondary structure for each strand in Fab during the simulations,
834 respect to pH 7.0 and 300 K, for: C) pH 4.5 and 300 K, D) pH 7.0 and 340 K. Error bars are the
835 same and equal for positive and negative values.

836

837 **Fig S5. Salt Bridge analysis for simulations at pH 4.5 at 300 K, and pH 7.0 at 340 K.** A, B)
838 Salt bridges formed during the simulation time for representative MD simulations at A) pH 4.5 and
839 300 K, B) pH 7.0 and 340 K. Presence of a salt bridge is indicated in white and absence in black.
840 C, D) List of salt bridges and its occurrence for simulations at A) pH 4.5 and 300 K, B) pH 7.0 and
841 340 K. Values shown are the average of three independent simulations with their SEM as error.
842 The more persistent salt bridges are highlighted for pH 4.5 (red) and for high temperature 340 K
843 (blue). E, F) The more persistent salt bridges are mapped into the Fab A33 structure.
844
845 **Fig S6. Stabilizing mutations predicted by FoldX and Rosetta. Correlation between FoldX**
846 **and Rosetta predictions.** Mutations predicted by both software to be most stabilizing are shown
847 in magenta and highlighted in a gray square on the bottom left. Mutations predicted only by FoldX
848 to be stabilizing are shown in green and mutations predicted only by Rosetta in yellow.
849
850 **Table S2. List of the most stabilizing mutations identified by FoldX and Rosetta-ddG.**
851 Mutation and ΔΔG of the 25 most stabilizing mutations predicted by FoldX and Rosetta.
852
853 **Fig S7. Normals used to calculate the packing of each atom in Fab A33 using occluded**
854 **surface software.** To calculate the occluded surface packing (OSP) value for each residue,
855 normals that extend from the surface outward until they intersect a neighboring van der Waals
856 surface are used. The normals used to calculate the OSP value of the inter-domain residues
857 identified by FoldX and Rosetta (S176, N137, S397, T180, and S395), are shown.
858
859 **Table S3. Packing indicated by the occluded surface packing (OSP) value of the residues**
860 **located in β-strands within domain interfaces (V_L-V_H and C_L-C_H1) of Fab A33 homology**
861 **model**. OSP values were calculated using the occluded surface software.
862 *, residues identified by FoldX and Rosetta in the constant domain interface that can be stabilized

863 further.

864

865 **Fig S8. Sequence entropy of Fab sequences.** (A) Entropy (Hx) of one hundred Fab light chains,
866 including κ and λ light chains. (B) Entropy (Hx) of one hundred Fab heavy chains. Light and heavy
867 chains are from human and mouse species. Alignment and entropy calculation were done with
868 Bioedit. Variable domains are indicated with a (V) and constant domains with a (C), separated by
869 a vertical line. CDRs are indicated in red.

870

871 **Fig S9. Aggregation prone regions (APR) in Fab A33.** A) The normalised aggregation
872 propensity for each residue in Fab A33 was predicted using PASTA 2.0, TANGO, AGGRESCAN
873 and MetAmyl software, each being colour-coded as shown in the legend. Aggregation-prone
874 sequences where three of the four software agreed were selected, and highlighted with an
875 asterisk. Amylpred2 consensus tool was also used to confirm the identification of those APRs,
876 indicated in red on the top. B) Using the native Fab A33 homology model, regions with greater
877 aggregation propensities are shown in red and reduced propensities in blue.

878

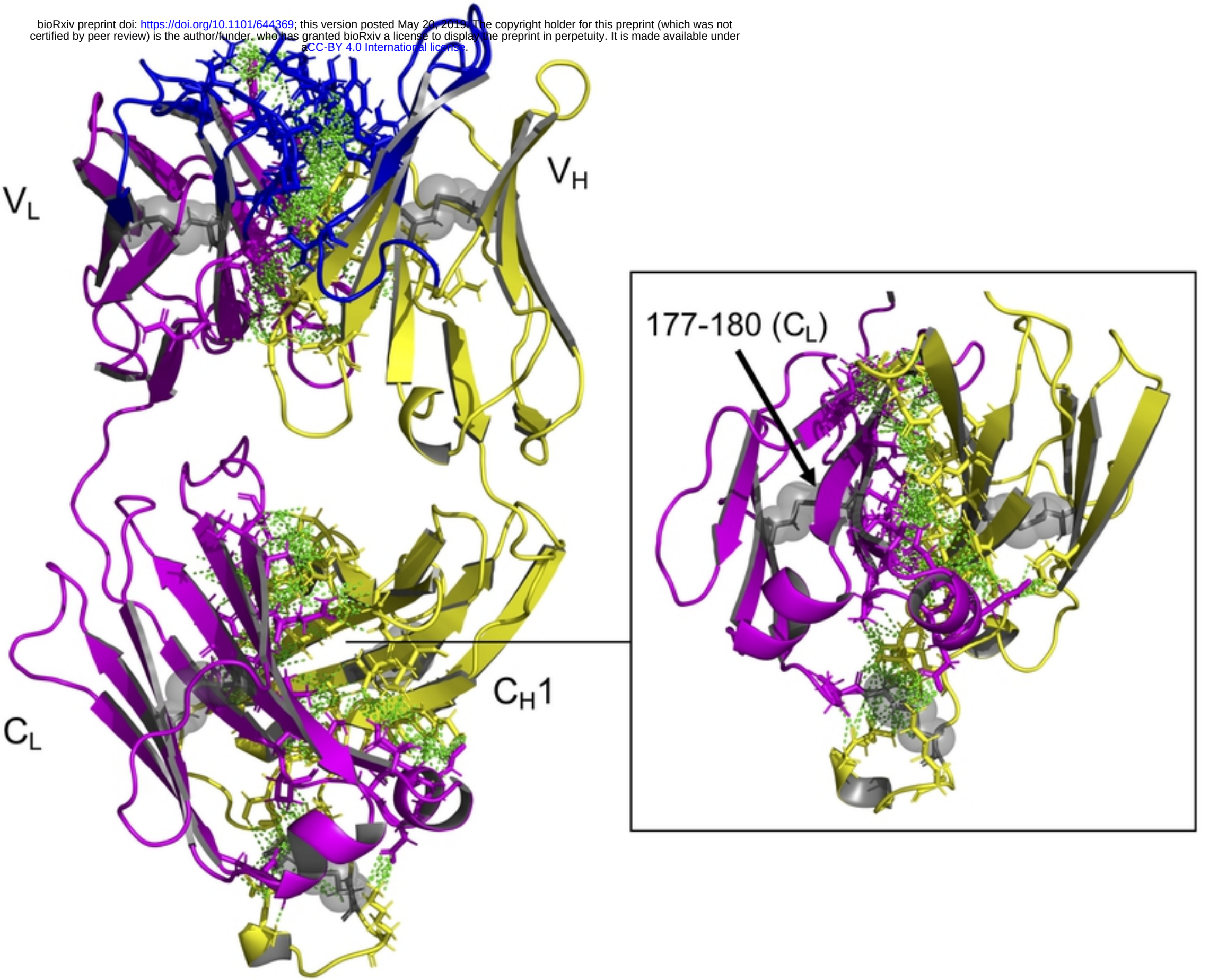


Figure 1

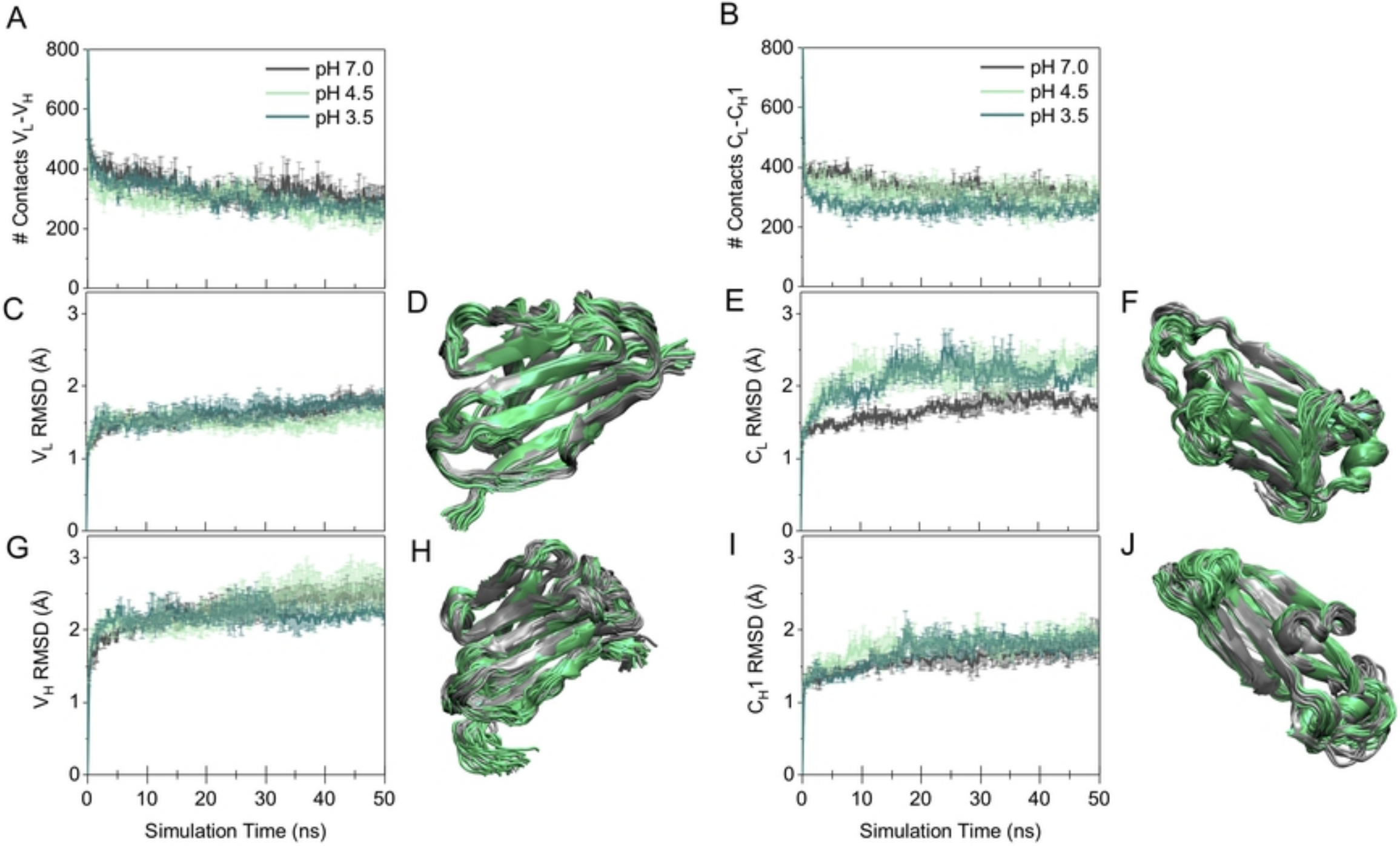

A
B
C
D
E
F
G
H
I
J
pH 7.0
pH 4.5
pH 3.5
Contacts V_L-V_H
Contacts C_L-C_H1
V_L RMSD (Å)
C_L RMSD (Å)
V_H RMSD (Å)
C_H1 RMSD (Å)
Simulation Time (ns)
Simulation Time (ns)

Figure 2

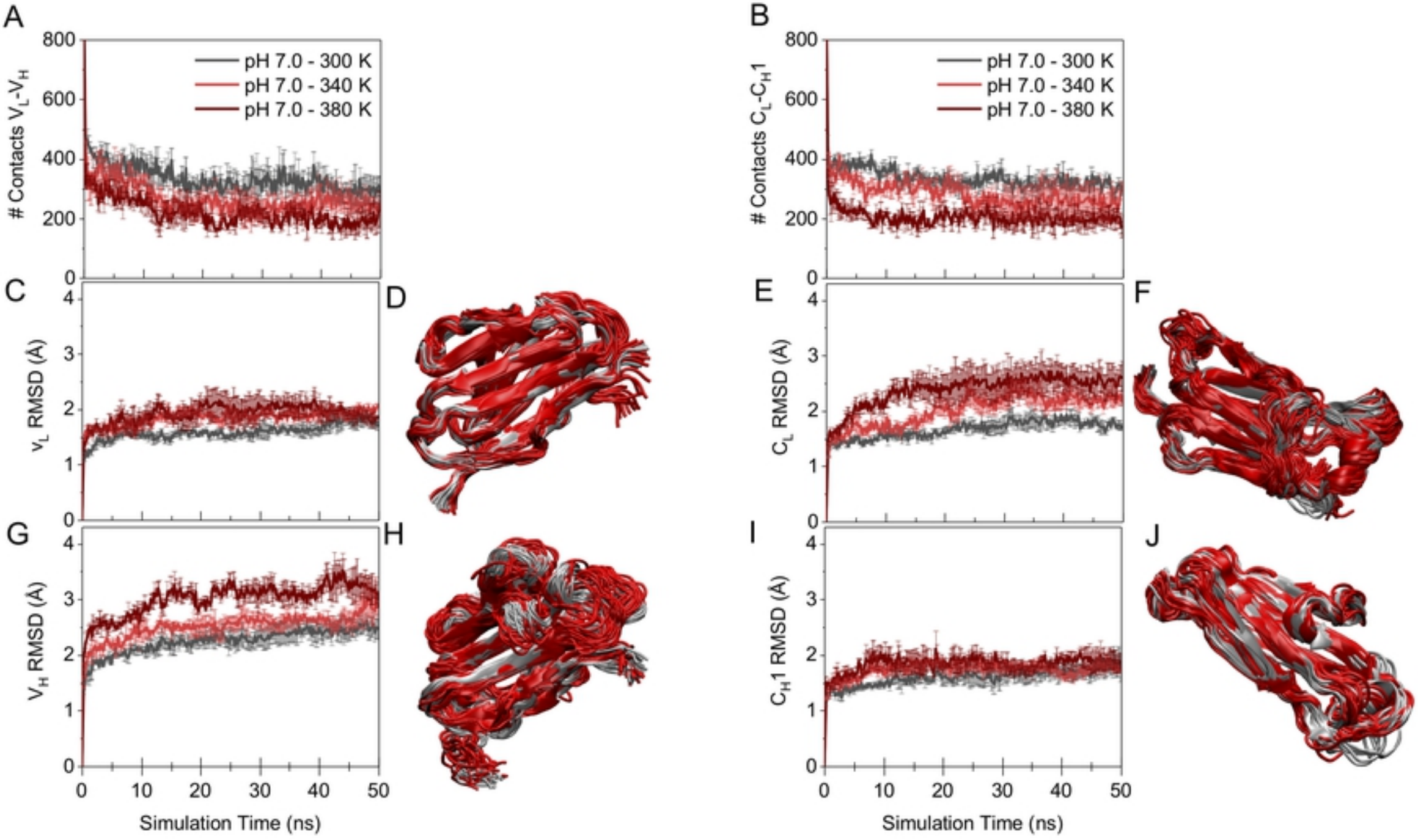
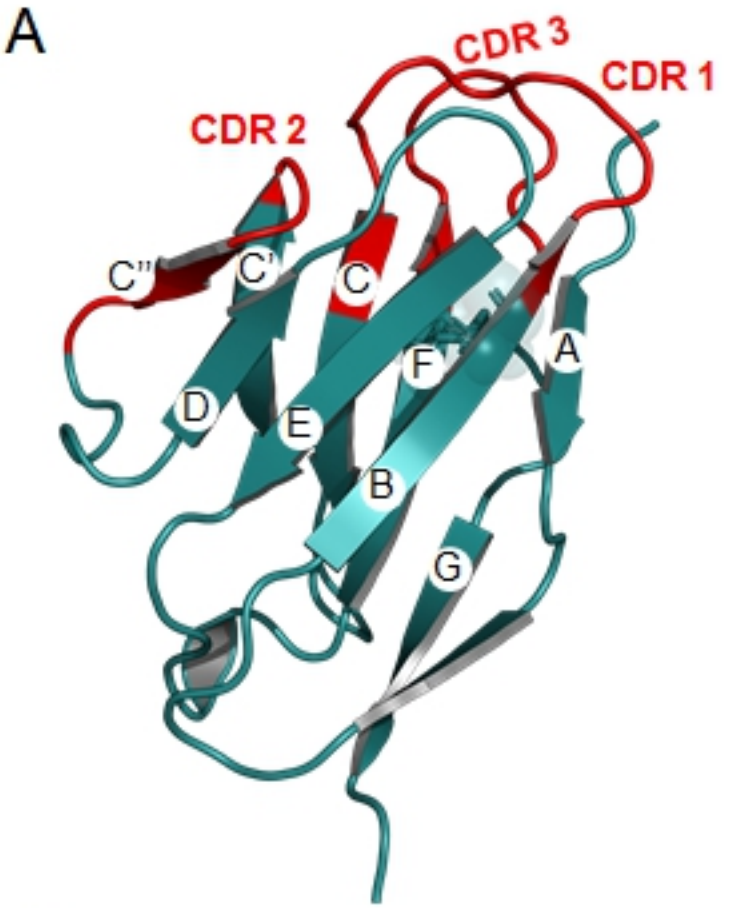
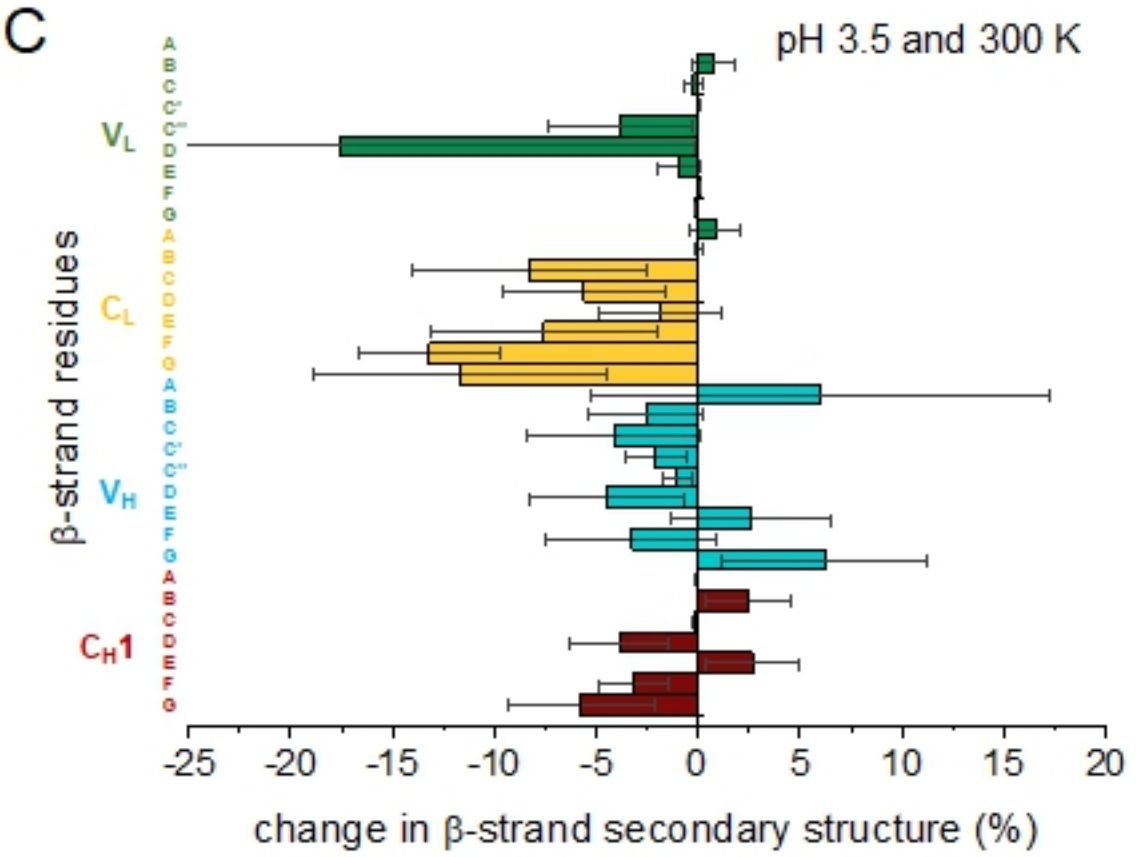
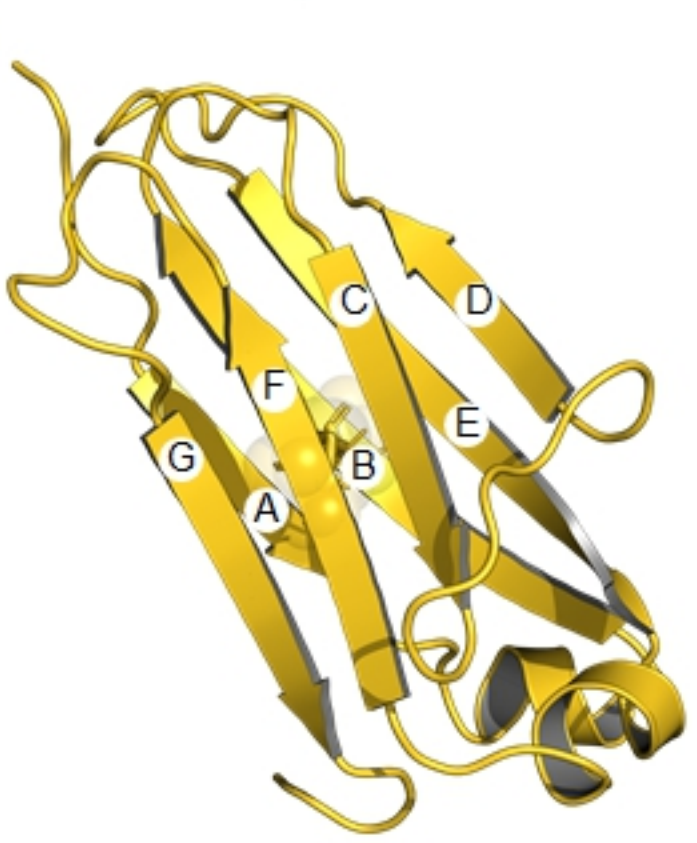

A
B
C
D
E
F
G
H
I
J
pH 7.0 - 300 K
pH 7.0 - 340 K
pH 7.0 - 380 K
Contacts V_L-V_H
Contacts C_L-C_H1
v_L RMSD (Å)
C_L RMSD (Å)
V_H RMSD (Å)
C_H1 RMSD (Å)
Simulation Time (ns)
0
200
400
600
800
1
2
3
4
10
20
30
40
50

Figure 3


A
CDR 3
CDR 1
CDR 2
C''
C'
C
F
A
D
E
B
G


C
pH 3.5 and 300 K
β-strand residues
V_L
C_L
V_H
C_H1
change in β-strand secondary structure (%)


B
C
D
F
E
G
B
A

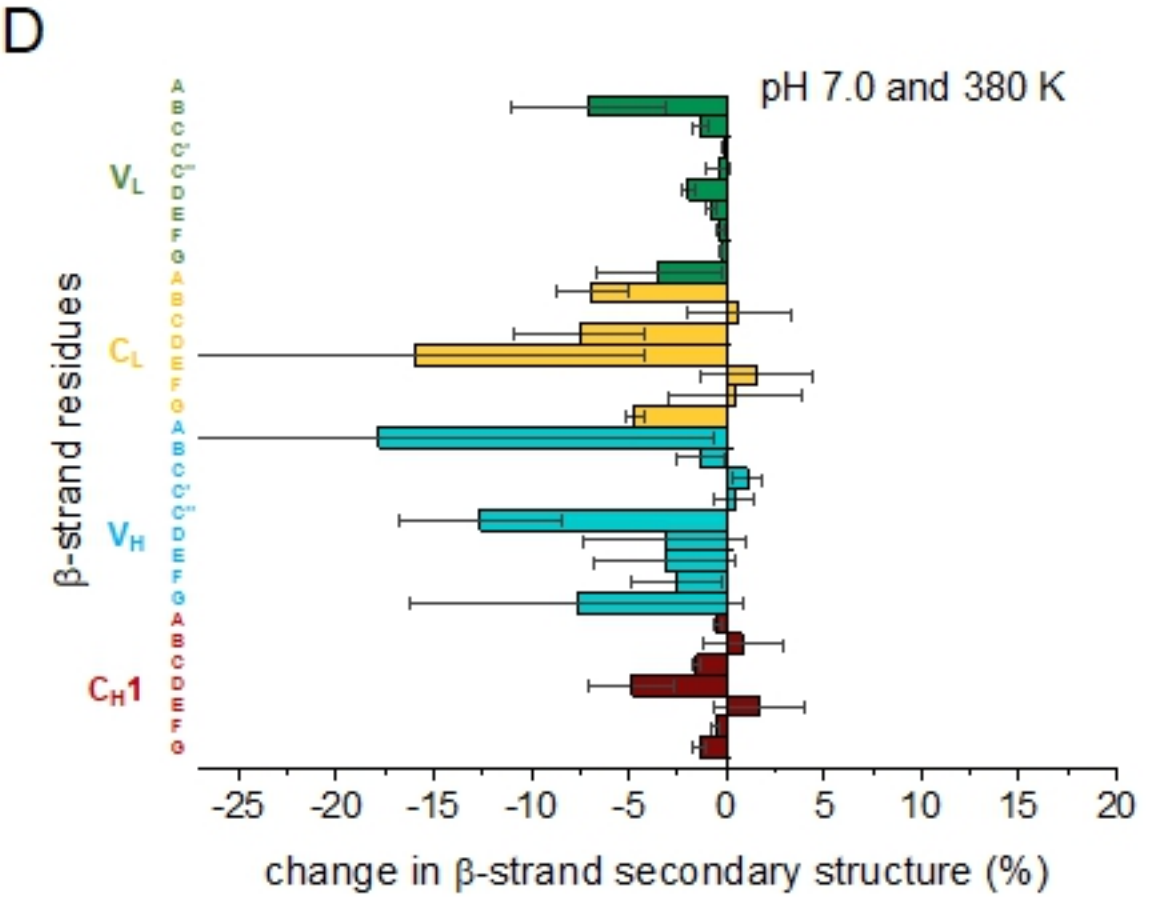

D
pH 7.0 and 380 K
β-strand residues
V_L
C_L
V_H
C_H1
change in β-strand secondary structure (%)

Figure 4

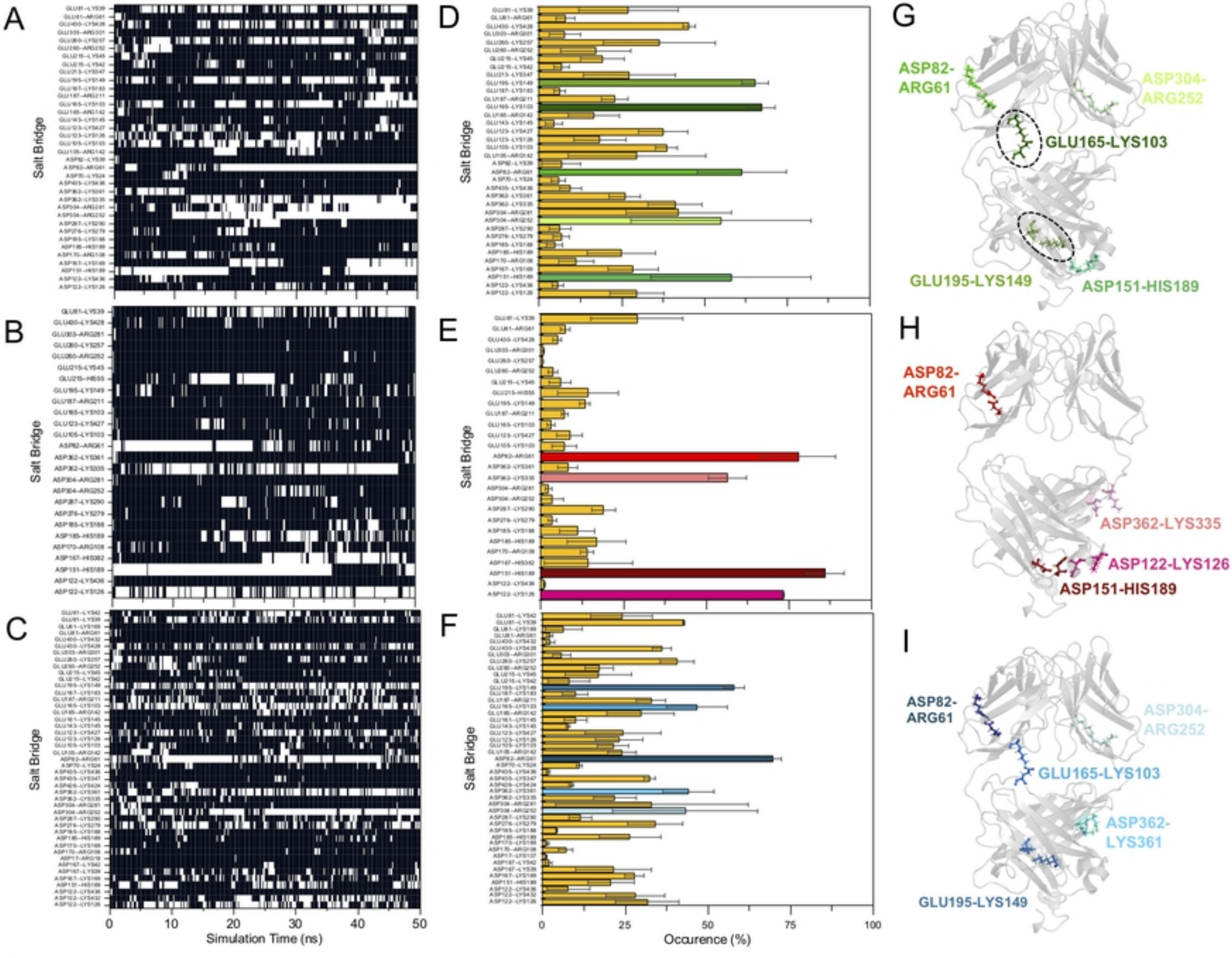

A
B
C
D
E
F
G
H
I
Salt Bridge
Simulation Time (ns)
Occurence (%)
ASP82-ARG61
ASP304-ARG252
GLU165-LYS103
GLU195-LYS149
ASP151-HIS189
ASP362-LYS335
ASP122-LYS126
ASP362-LYS361

Figure 5

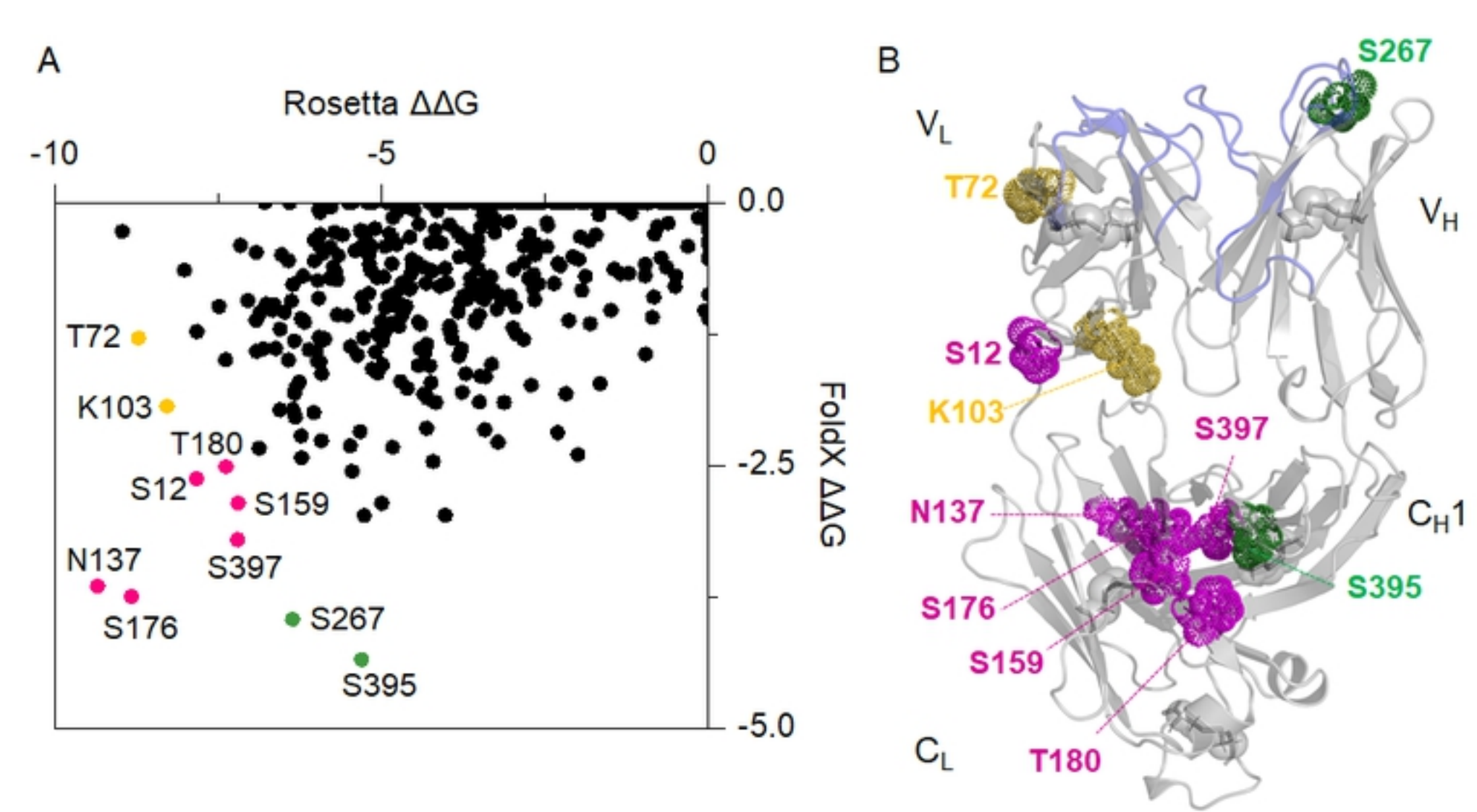

A
Rosetta ΔΔG
-10
-5
0
0.0
-2.5
-5.0
FoldX ΔΔG
T72
K103
T180
S12
S159
N137
S397
S176
S267
S395
B
S267
V_L
T72
V_H
S12
K103
S397
N137
C_H1
S395
S176
S159
C_L
T180

Figure 6

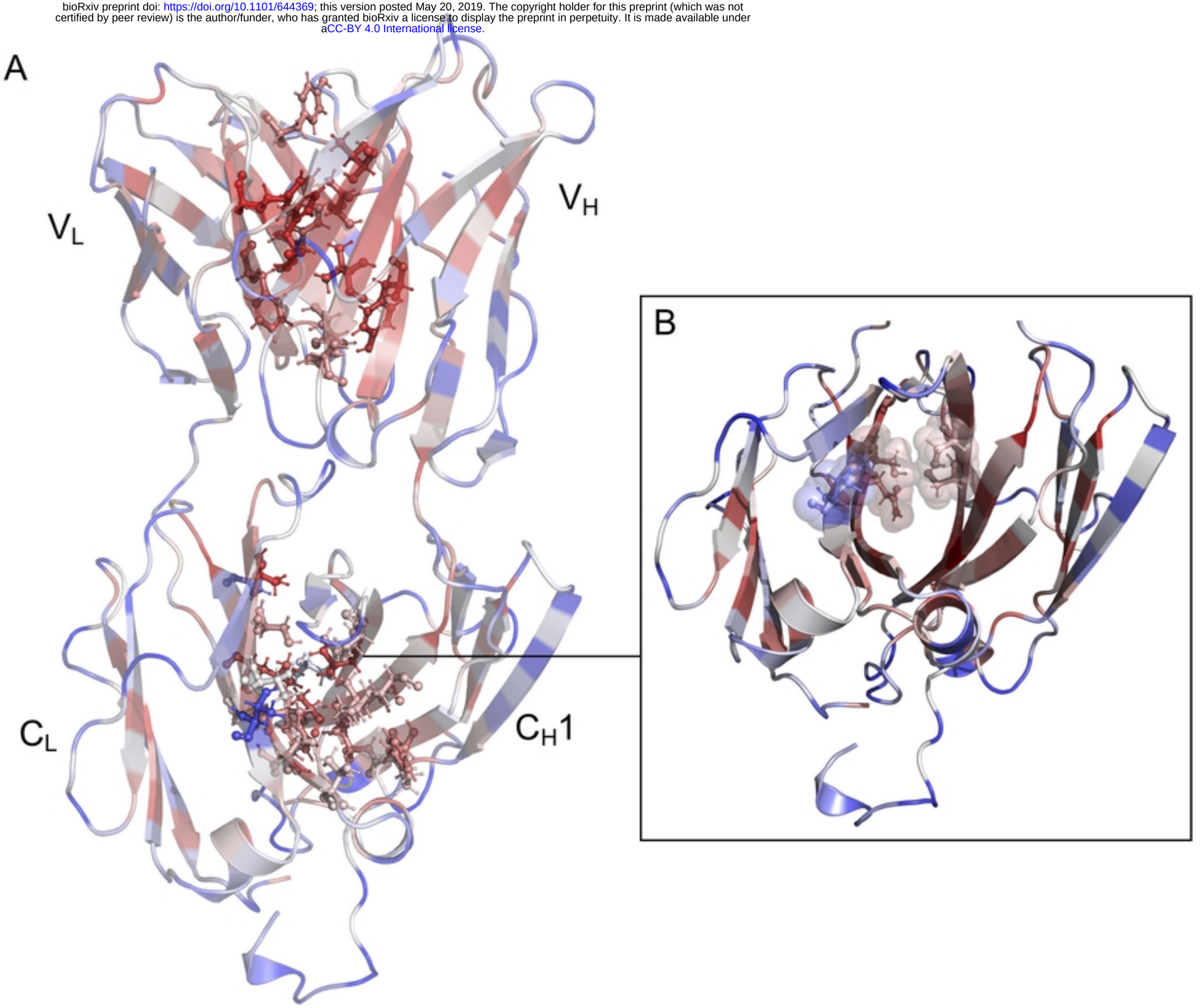


Figure 7

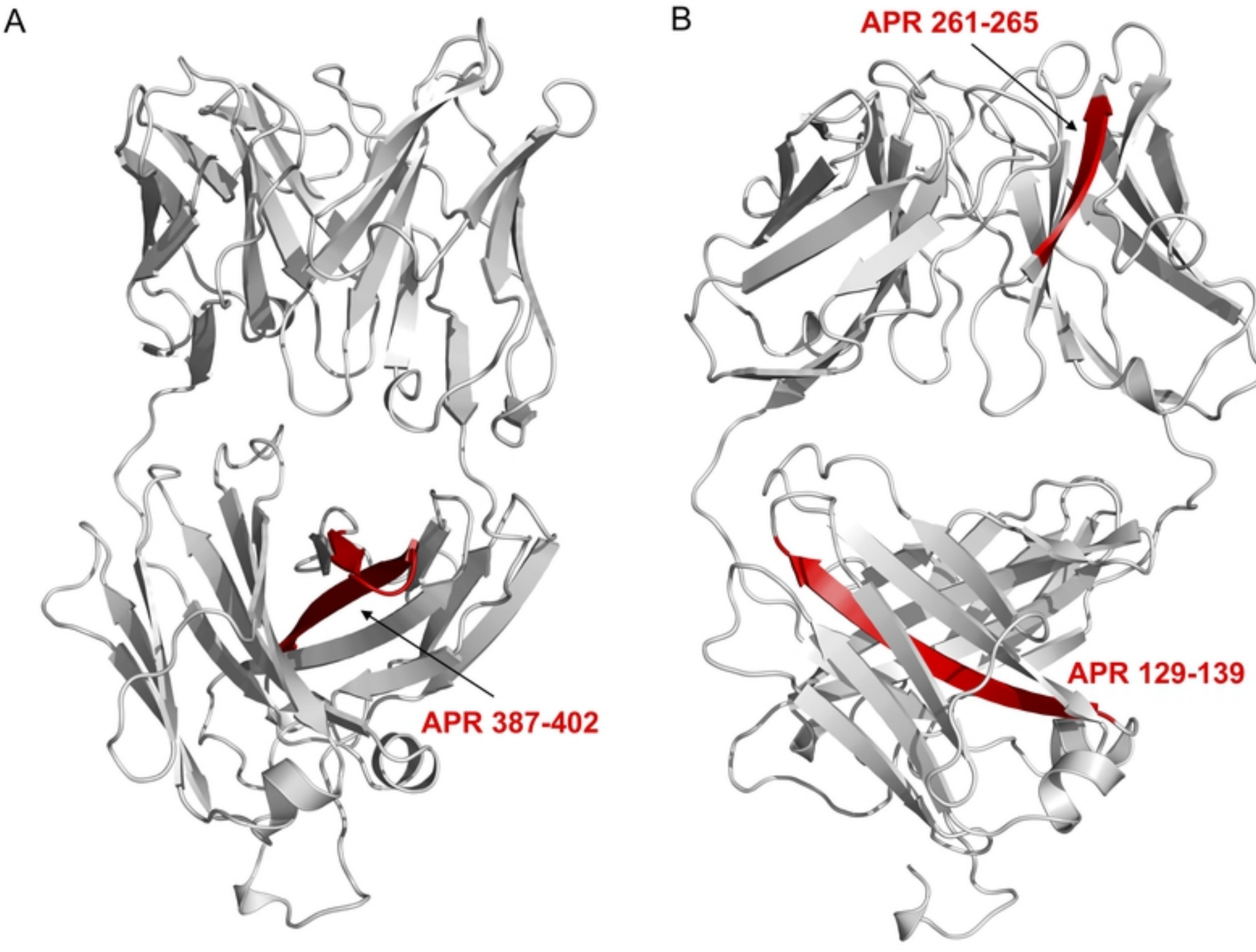


Figure 8