Electronic Supplementary Material (ESI) for RSC Advances.

SUPPLEMENTARY INFORMATION

Combining Enhanced Sampling and Deep Learning Dimensionality Reduction for the Study of the Heat Shock Protein B8 and its Pathological Mutant K141E.

Daniele Montepietra[1,2], *Ciro Cecconi*[1,2], *Giorgia Brancolini*[2*]

[1] Department of Physics, Computer Science and Mathematics, University of Modena and Reggio Emilia, Via Campi 213/A 4100 Modena, Italy

[2] Istituto Nanoscienze – CNR-NANO, Center S3, via G. Campi 213/A, 41125 Modena, Italy

Corresponding author e-mail: giorgia.brancolini@nano.cnr.it

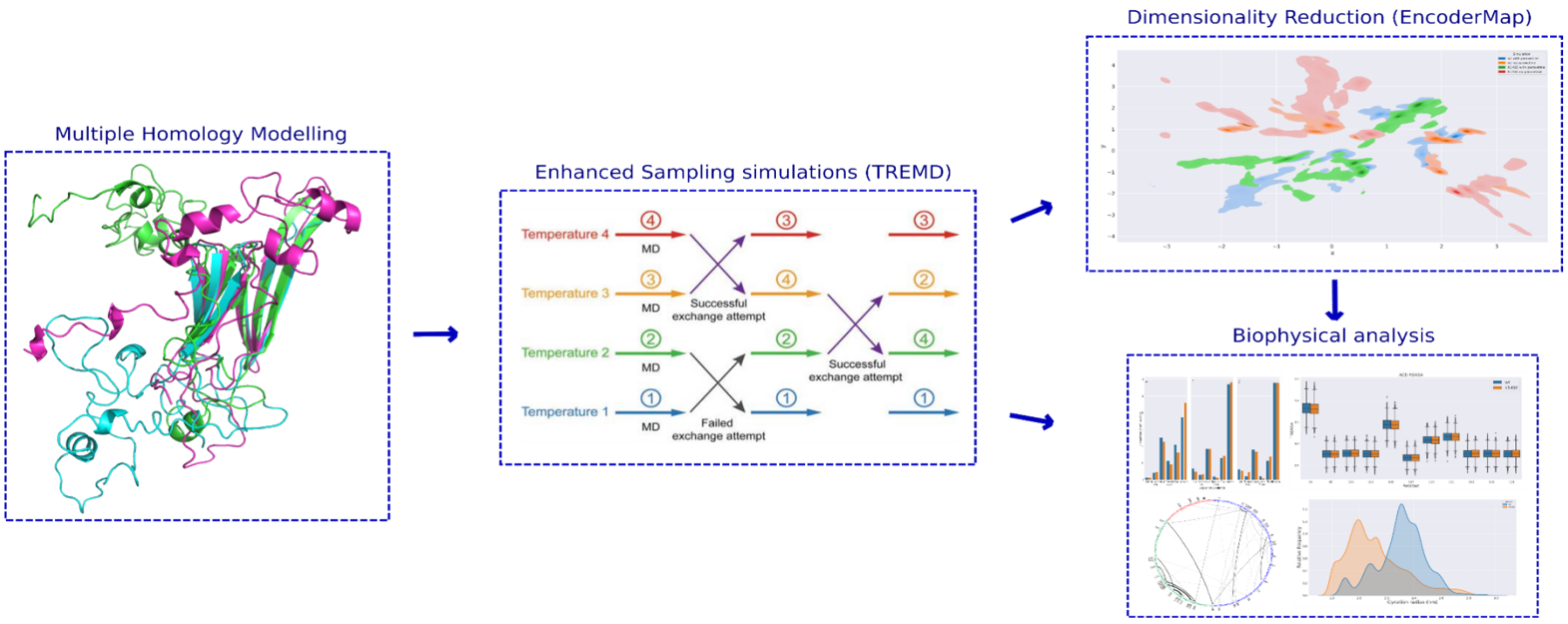


Figure S1: Schematic representation of the workflow used to study HSPB8. Starting from the Homology Models obtained from the homology model servers ROSETTA, I-TASSER, and MODELLER, we obtained different starting structures for HSPB8, which were then simulated with the Enhanced Sampling method Temperature Replica Exchange (TREMD). Classical analyses were first used to analyze the resulting trajectories, e.g., Radius of Gyration, Salt bridges network, RMSD, and RMSF. Then, using EncoderMap, the proteins' trajectories were projected to a 2D map to allow more effective visualization of the difference in conformations assumed by the two variants of HSPB8.

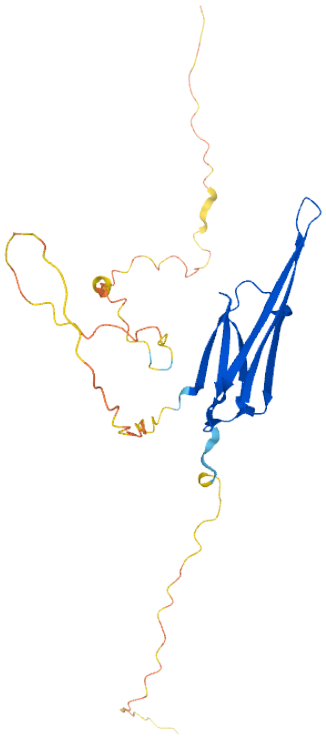


Figure S2: Homology Model of *wt* HSPB8 obtained with Alphafold2 on UNIPROT (entry Q9UJY1). Protein residues are colored according to their confidence. AlphaFold2 produces a per-residue confidence score (pLDDT) between 0 and 100. Blue refers to very high confidence (pLDDT > 90), cyan to high confidence (90 > pLDDT > 70), yellow to low confidence (70 > pLDDT > 50), and orange to very low confidence (pLDDT < 50). HSPB8 IDRs are systematically predicted with low or very low model confidence.

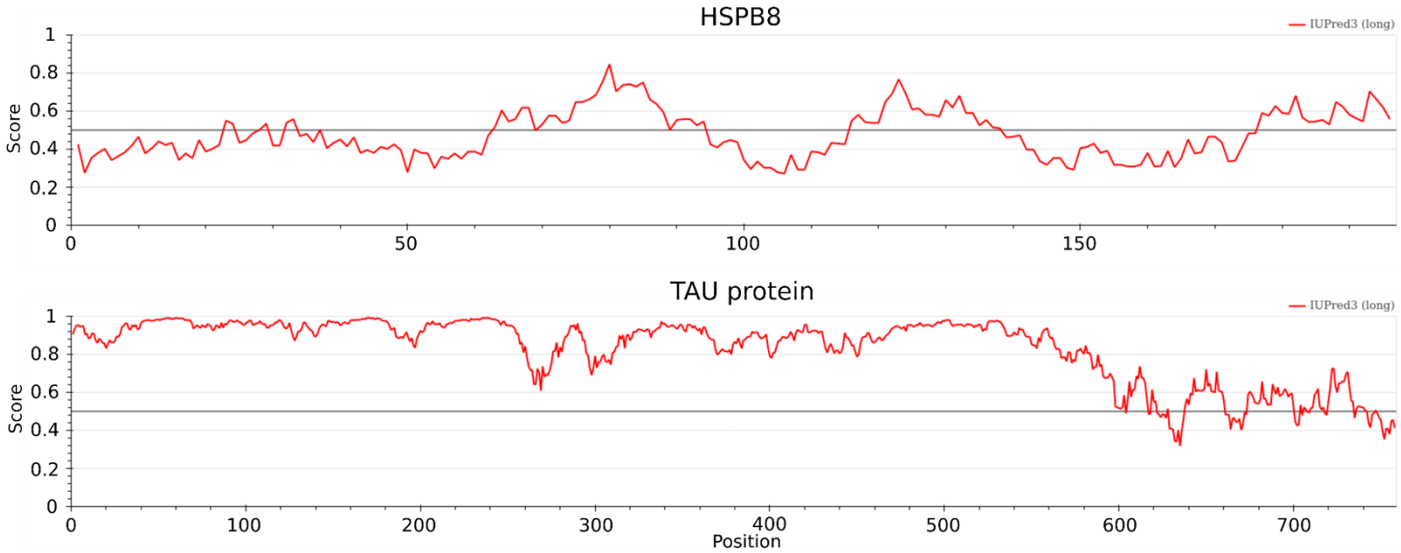


Figure S3: Disorder score for HSPB8 (top panel) and tau protein (bottom panel) from IUPRED3. IUPred3 is a combined web interface that allows users to identify disordered protein regions. Compared to a highly disordered protein such as the tau protein, HSPB8 IDRs are comparable more to a molten globule state. The webserver is available at https://iupred3.elte.hu/.

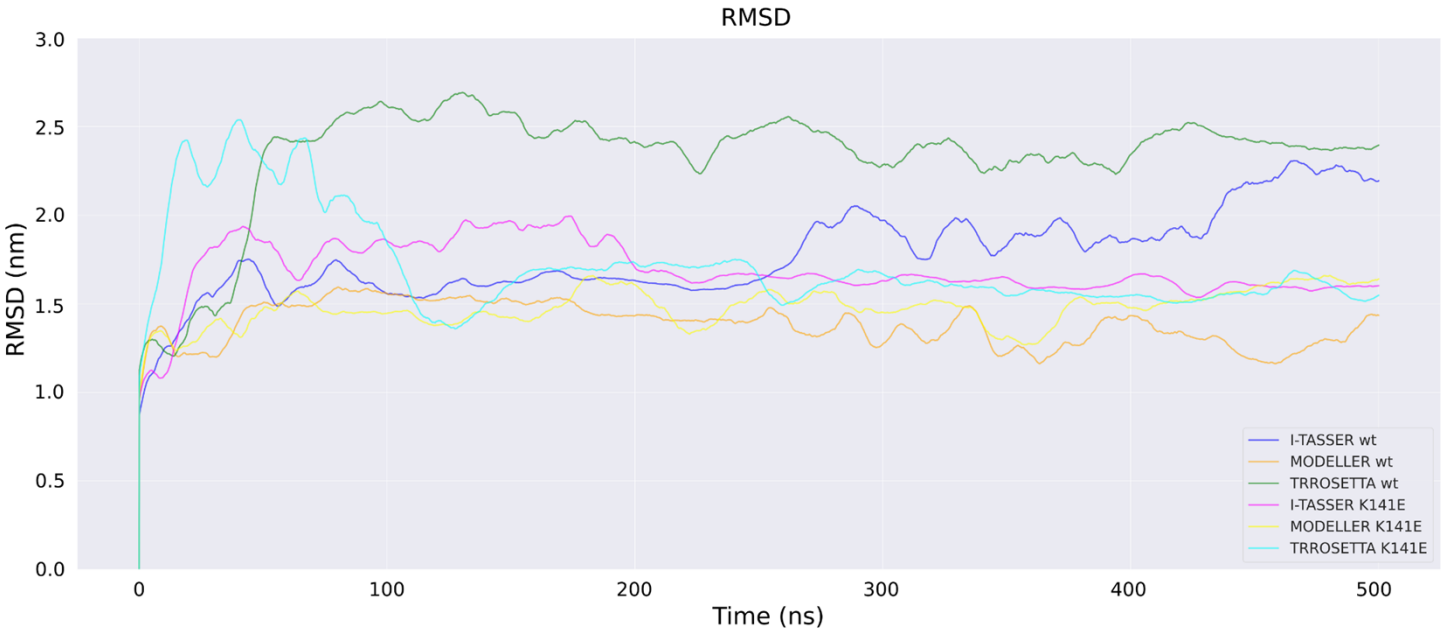


Figure S4: Root Mean Square Deviation (RMSD) computed on the MD trajectories for each of the three homology models obtained with I-TASSER, MODELLER, ROSETTA for both wild-type (*wt*) and mutated HSPB8 (K141E) proteins. We observe that after the Homology Model structure stabilization in the first 50 ns, the RMSD undergoes minor changes in every MD. This underlines how even 500 ns MD simulations are unsuited to simulate proteins containing IDRs.

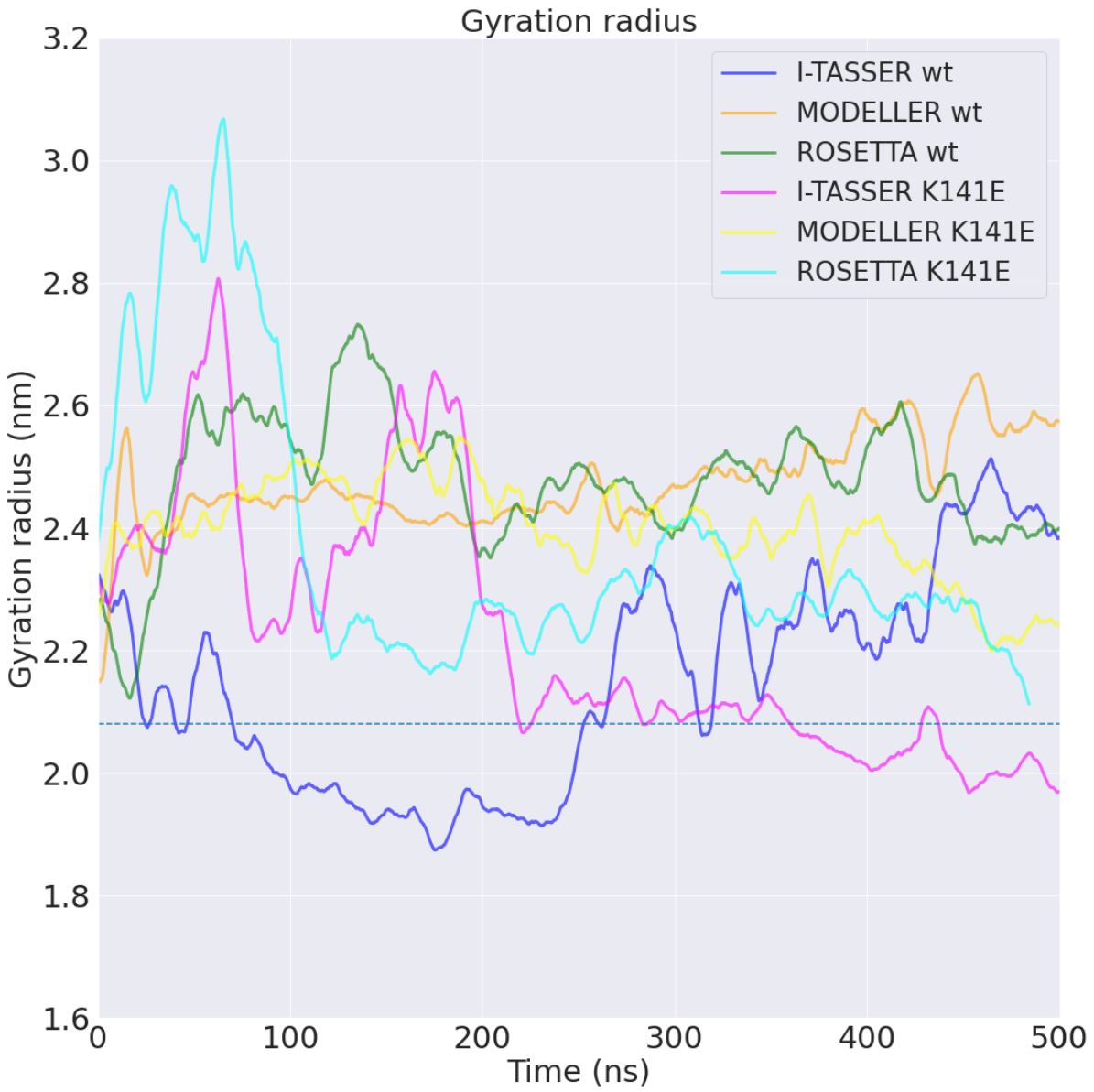


Figure S5: Radius of gyration (Rg) computed on the MD trajectories for each of the three homology models obtained with I-TASSER, MODELLER, ROSETTA for both wild-type (*wt*) and mutated HSPB8 (K141E) proteins.

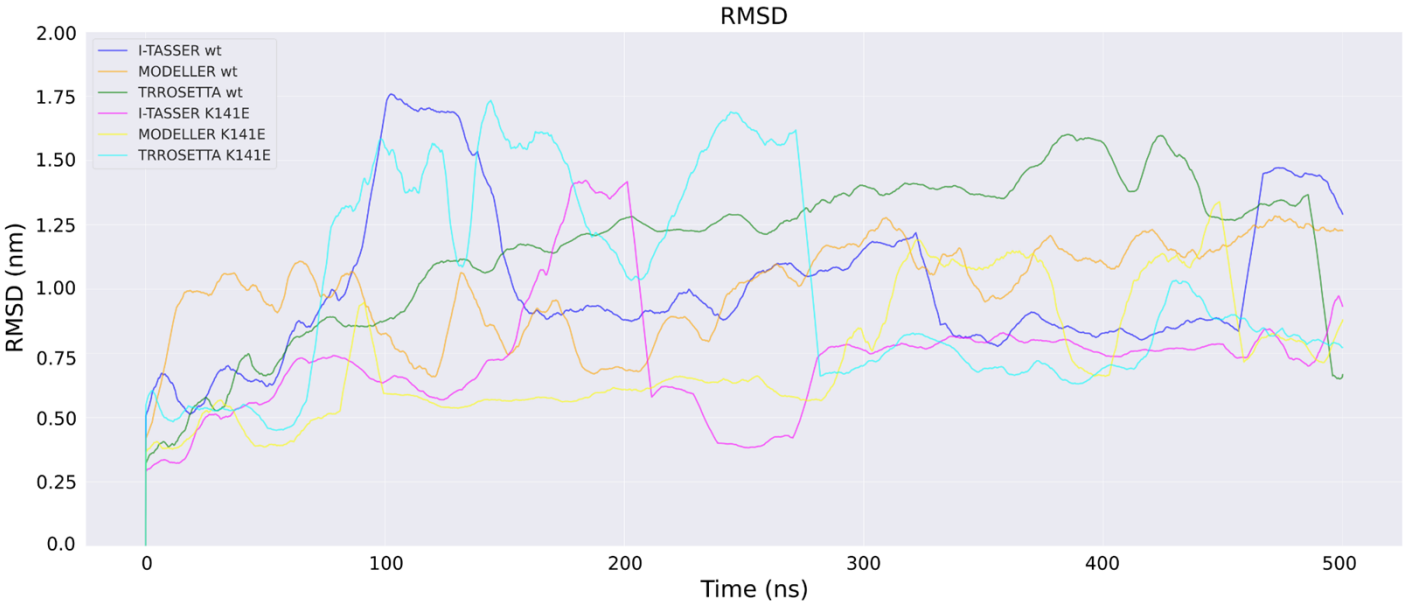


Figure S6: Root Mean Square Deviation (RMSD) computed on the TREMD trajectories for each of the three Homology Models obtained with I-TASSER, MODELLER, and ROSETTA for both *wt* and K141E HSPB8. As can be seen from the graph, all RMSD profiles show sharp increases and decreases. These structure changes occur at temperature swappings during TREMDs, which indicates to us how this Enhanced Sampling method allows the protein to explore the available conformational space more extensively.

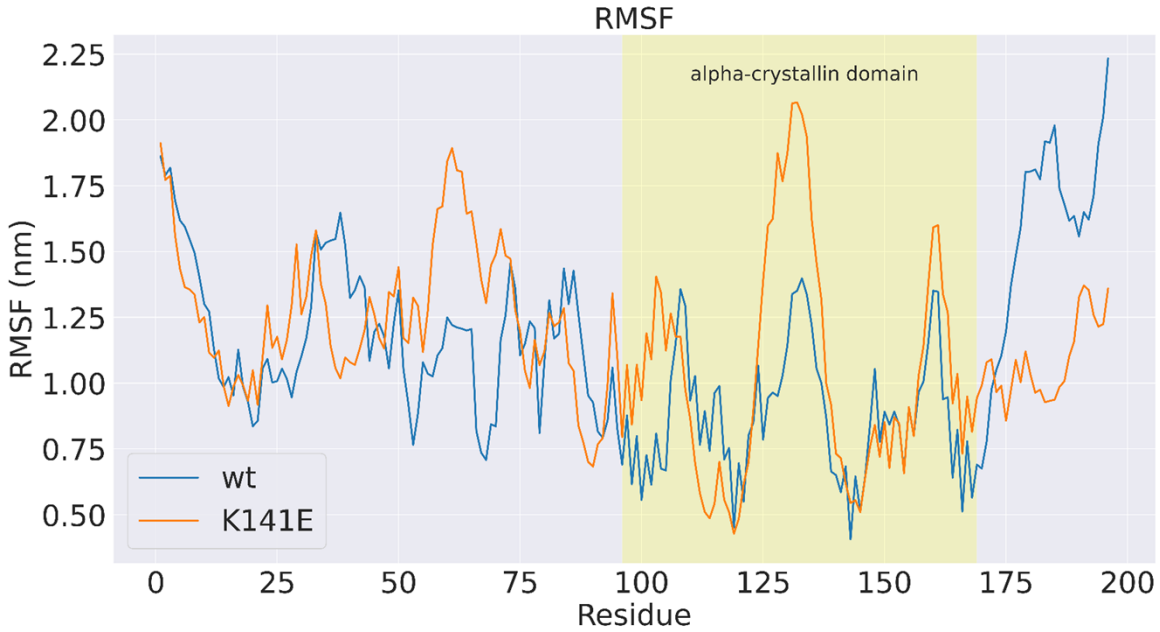


Figure S7: Root Mean Square Fluctuations (RMSF) per residue for the TREMD simulations are plotted to compare the flexibility of *wt* (blue) and K141E (orange). For the *wt* and K141E TREMD, all frames have been concatenated.

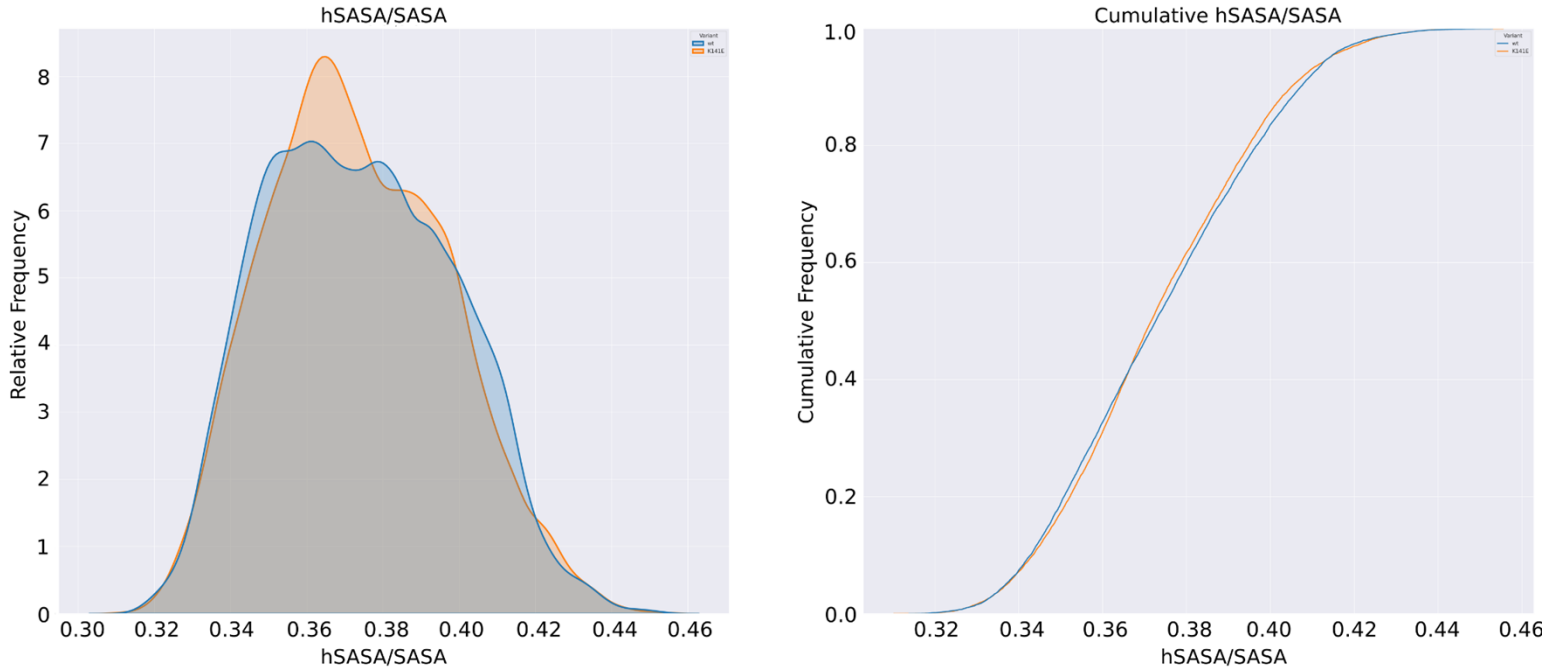


Figure S8: The left panel shows the hSASA/SASA distributions of the *wt* (blue) and K141E (orange) HSPB8, computed concatenating the TREMD trajectories collected for each HM. The distributions are normalized so that their underlying area is equal to one. The right panel displays the cumulative frequency distributions of the *wt* (blue line) and K141E (orange line) HSPB8, computed on the combined TREMD trajectories.

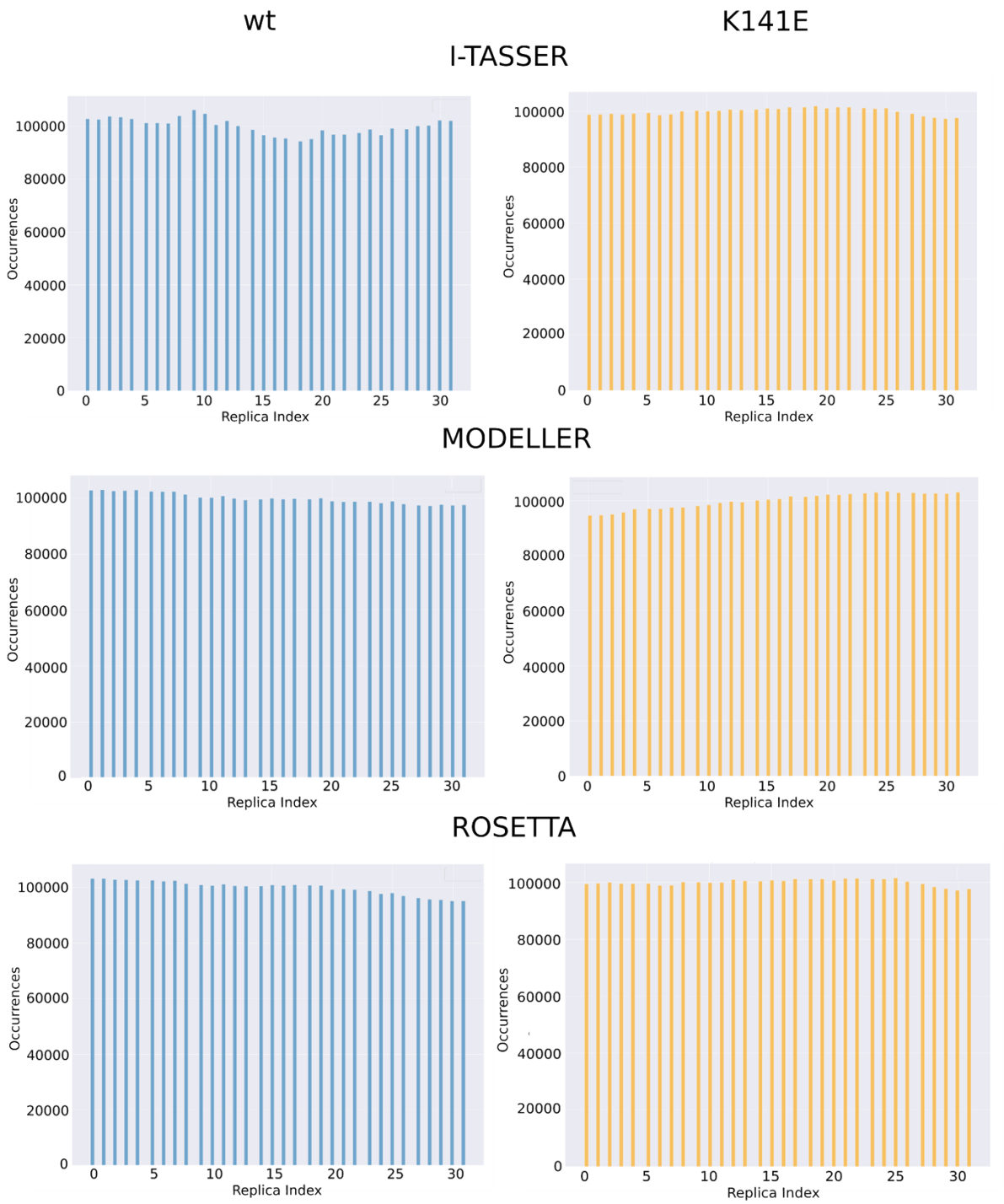
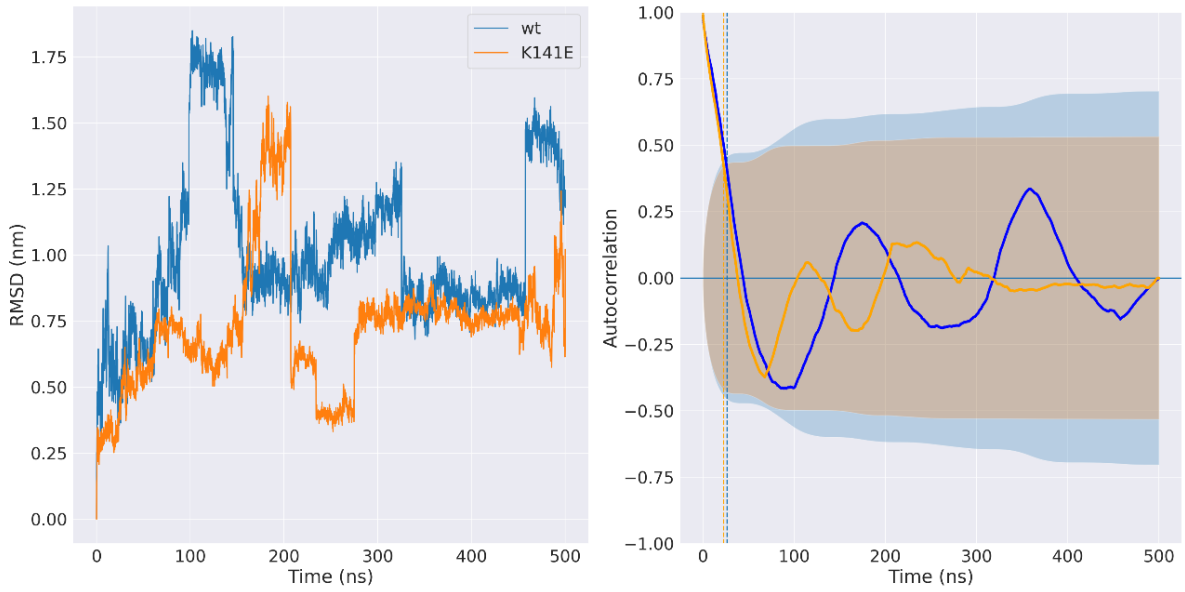
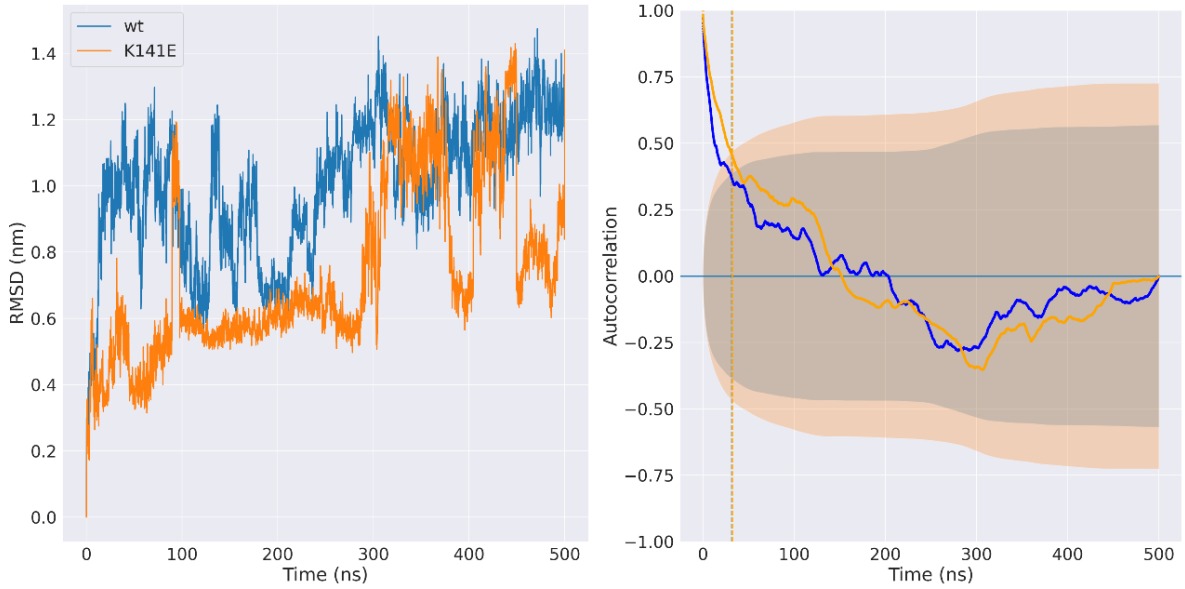


Figure S9: Distribution of temperatures' indices assumed by replica 1 during all 500 ns TREMD simulations. The y-axis reports the number of occurrences of each index, counted as the sum of the simulation frames in which replica 1 assumed that temperature. Each row represents the starting homology model (I-TASSER, MODELLER, ROSETTA from top to bottom), with the left panels (in blue) reporting the *wt* variant distributions and the right panels (in orange) the K141E variants.

I-TASSER



MODELLER



ROSETTA

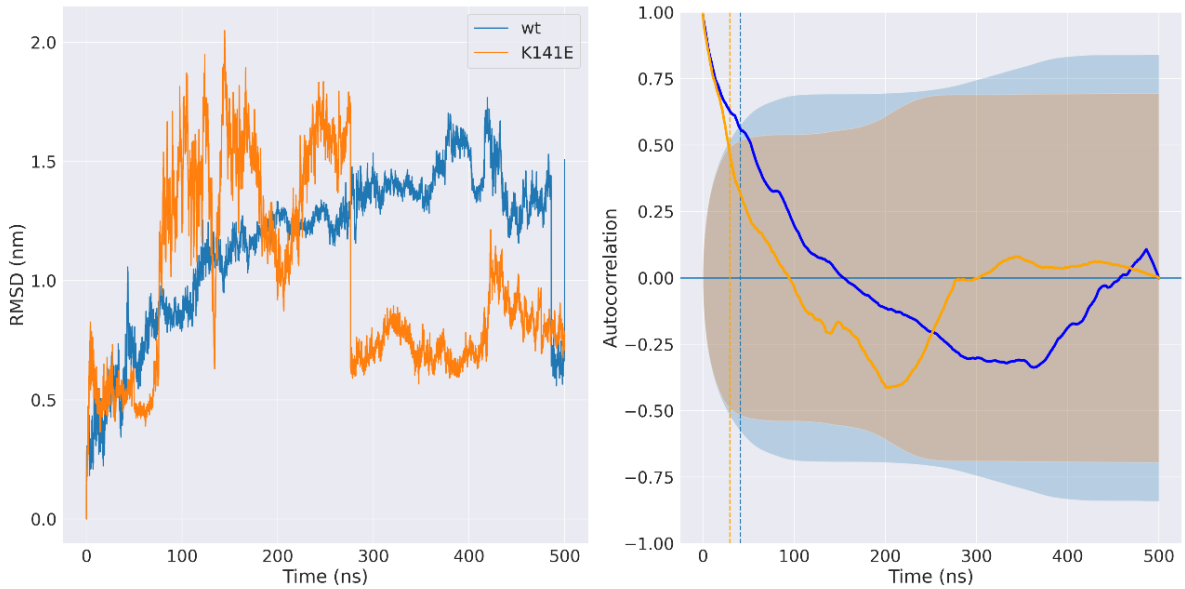
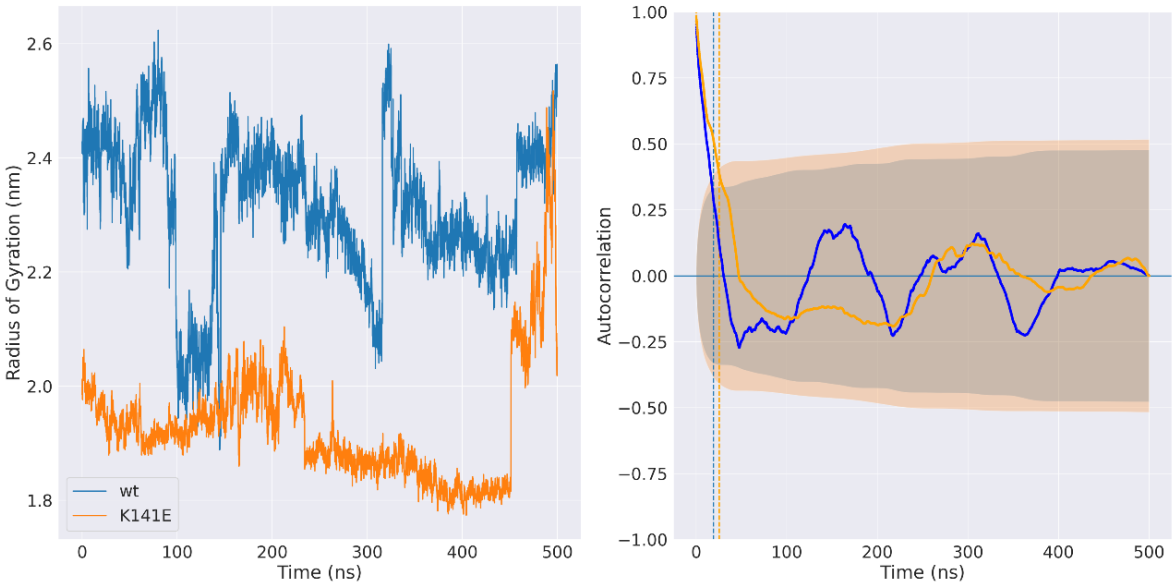
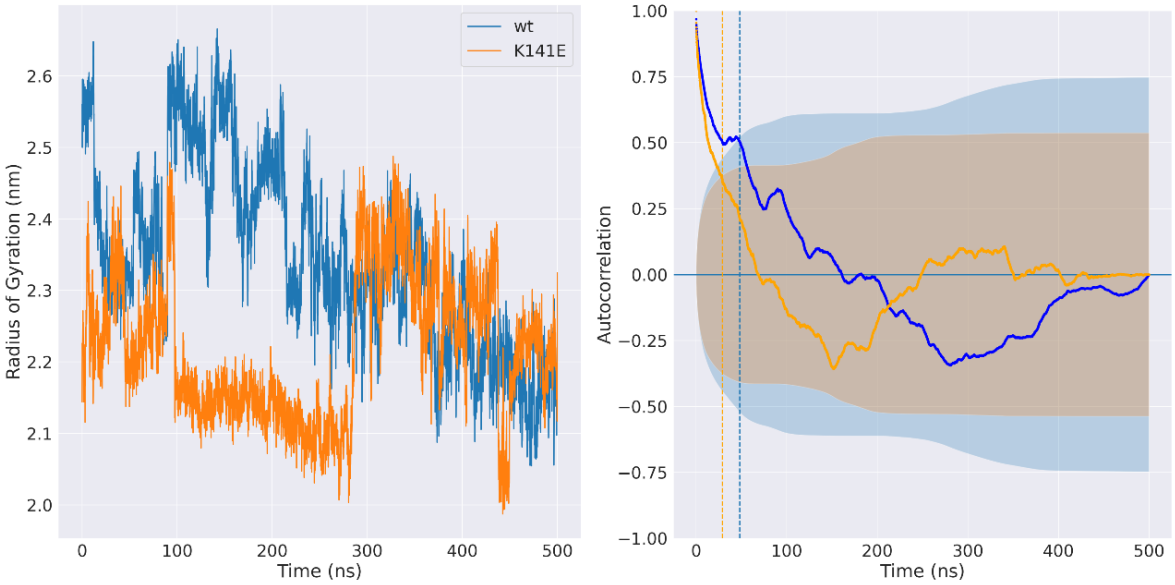


Figure S10: Plot of RMSD as a function of simulation time (left panels) and corresponding autocorrelation plots (right panels). Each row represents the starting homology model (I-TASSER, MODELLER, ROSETTA from top to bottom), with the *wt* variant in blue and the mutated K141E in orange. In the right panels, the x-axis shows the time lags values used for the autocorrelation calculation (reported on the y-axis). The shaded areas represent the 95% confidence interval for which two RMSD values taken at a time distance equal to the time lag can be considered uncorrelated (the areas' colors reflect the variant). The vertical dashed lines are drawn at the intersection of the autocorrelation curves and confidence intervals and represent the autocorrelation time, which is the minimum time distance between two RMSD observations to be considered independent. The autocorrelation times found are 26 ns and 22 ns for I-TASSER (*wt* and K141E, respectively); 31 ns and 32 ns for MODELLER (*wt* and K141E, respectively); 41 ns and 29 ns for ROSETTA (*wt* and K141E, respectively).

I-TASSER



MODELLER



ROSETTA

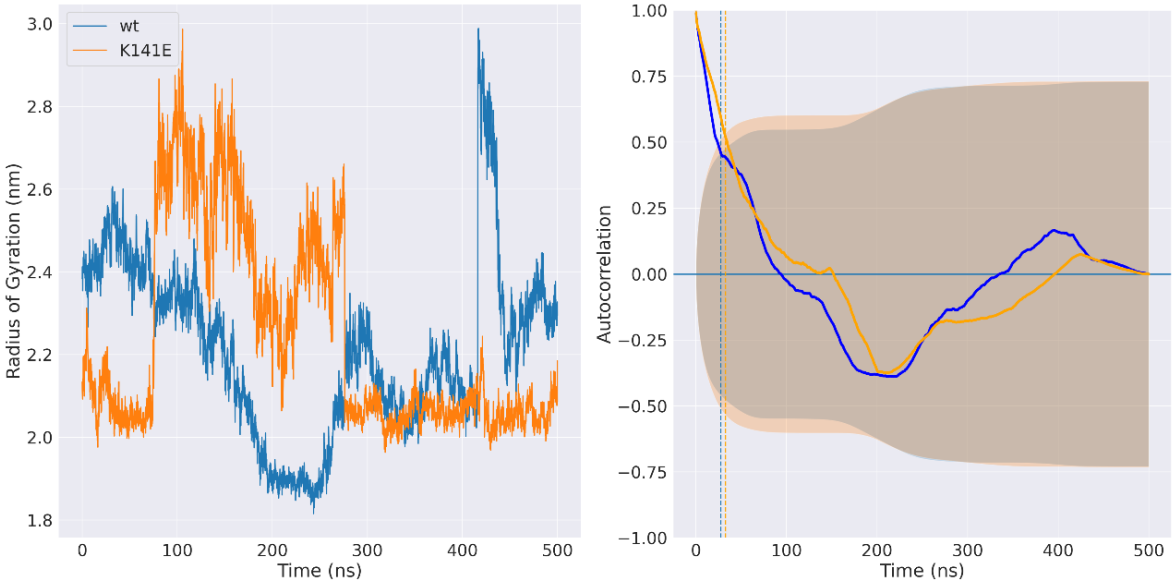


Figure S11: Plot of the Radius of gyration as a function of simulation time (left panels) and corresponding autocorrelation plots (right panels). Each row represents the starting homology model (I-TASSER, MODELLER, ROSETTA from top to bottom), with the *wt* variant in blue and the mutated K141E in orange. In the right panels, the x-axis shows the time lags values used for the autocorrelation calculation (reported on the y-axis). The shaded areas represent the 95% confidence interval for which two Rg values taken at a time distance equal to the time lag can be considered uncorrelated (the areas' colors reflect the variant). The vertical dashed lines are drawn at the intersection of the autocorrelation curves and confidence intervals and represent the autocorrelation time, which is the minimum time distance between two Rg observations to be considered independent. The autocorrelation times found are 19 ns and 25 ns for I-TASSER (*wt* and K141E, respectively); 48 ns and 29 ns for MODELLER (*wt* and K141E, respectively); 27 ns and 32 ns for ROSETTA (*wt* and K141E, respectively).

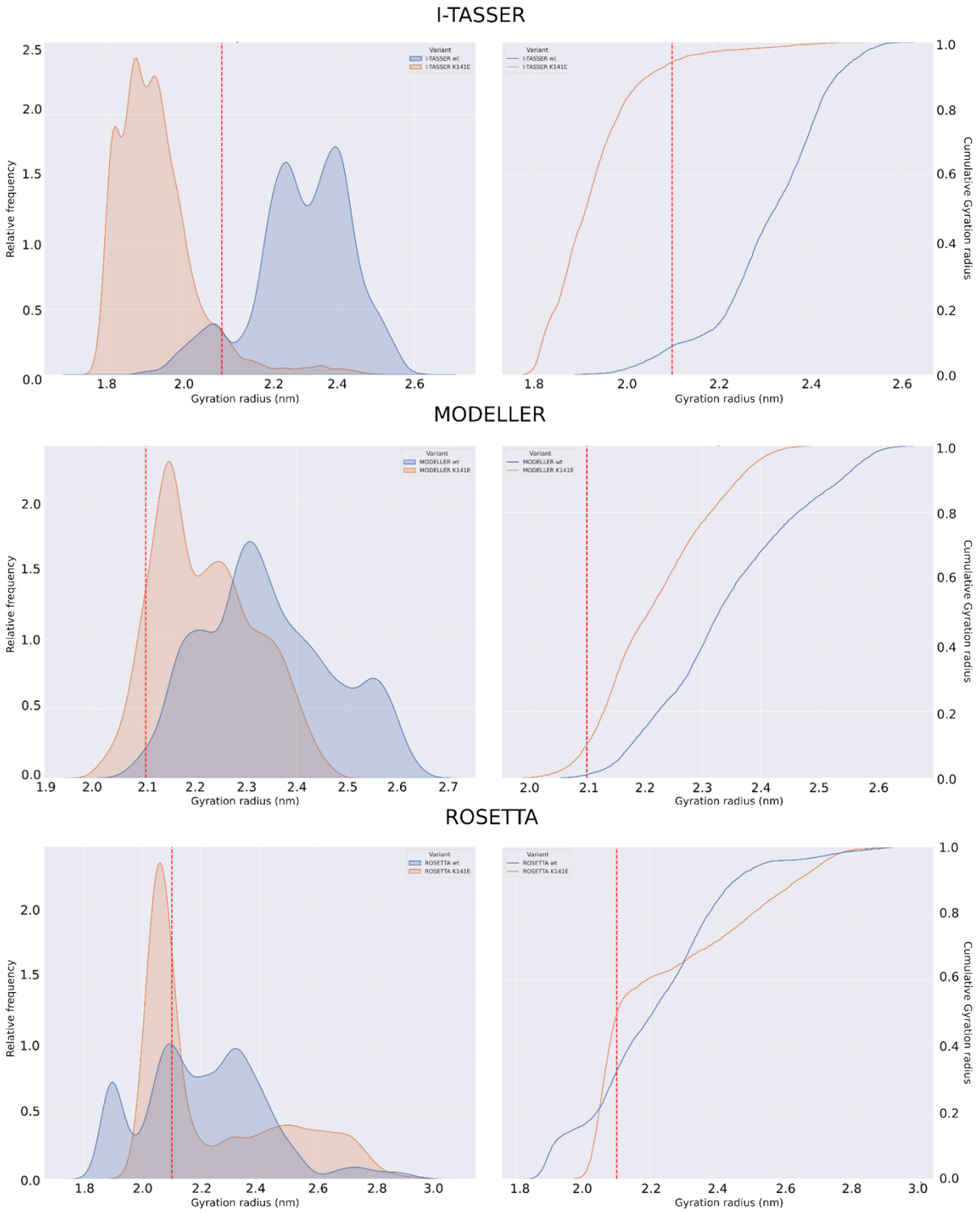


Figure S12: Radius of gyration distribution and cumulative plots for the I-TASSER (top panel), MODELLER (middle panel), and ROSETTA (bottom panel) TREMD. For each Homology Model, the left panel shows the Rg distributions of the *wt* (blue) and K141E (orange) HSPB8, computed concatenating the TREMD trajectories collected for each HM. The distributions are normalized so that their underlying area is equal to one. The right panels display the cumulative frequency distributions of the *wt* (blue line) and K141E (orange line) HSPB8, computed on the combined TREMD trajectories.

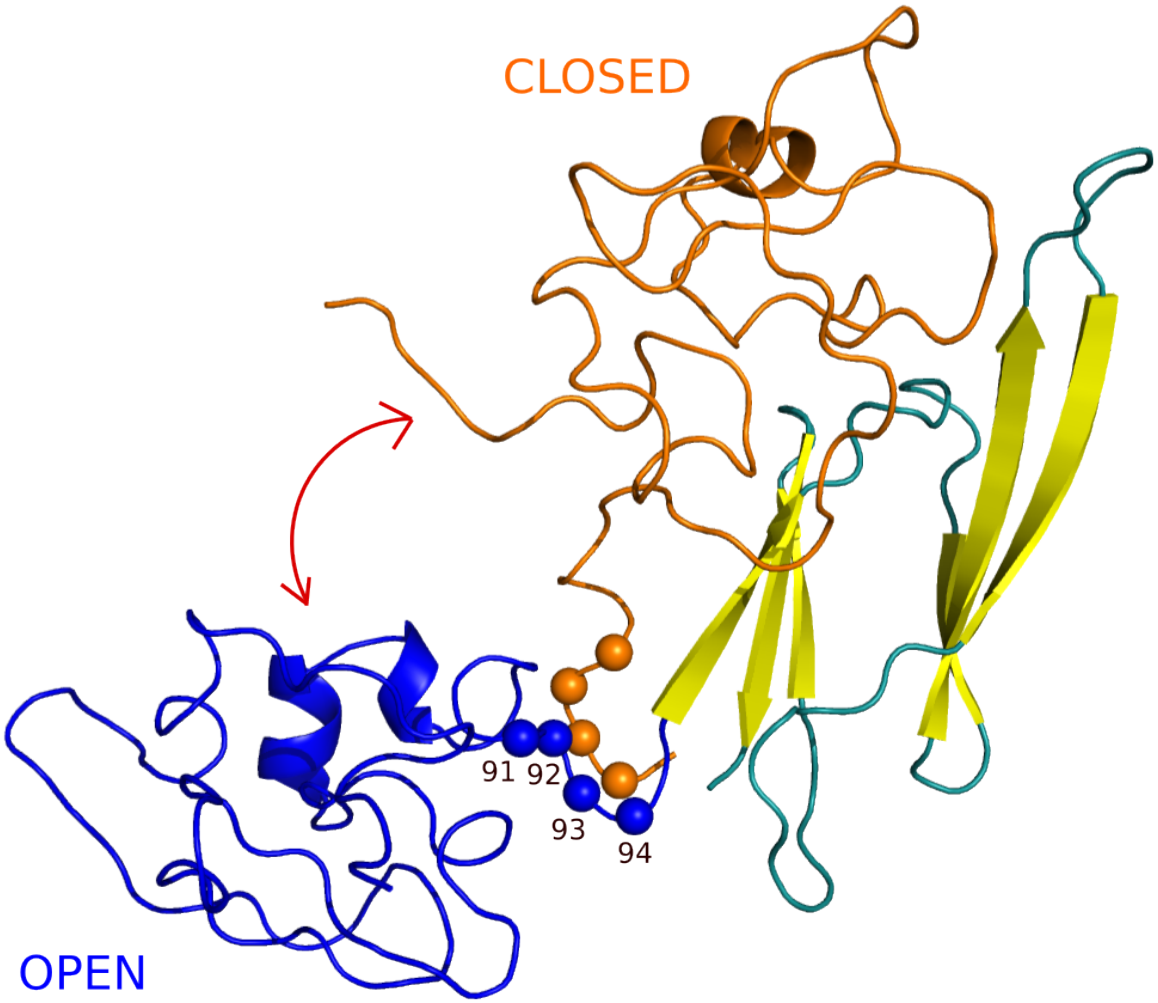


Figure S13: Representative IDR structures for the category of open (blue) and closed (orange) conformations. For both structures, the C_α of residues 91 to 94 are highlighted as spheres. To obtain this image, we superimposed two HSPB8 structures, and then one of the two ACDs, along with both CTDs, was removed. The value of the dihedral angle formed by these atoms, in combination with the radius of splicing of the protein, allows a more quantitative distinction of the continuum of conformations of HSPB8 *wt* and K141E into closed and open structures.

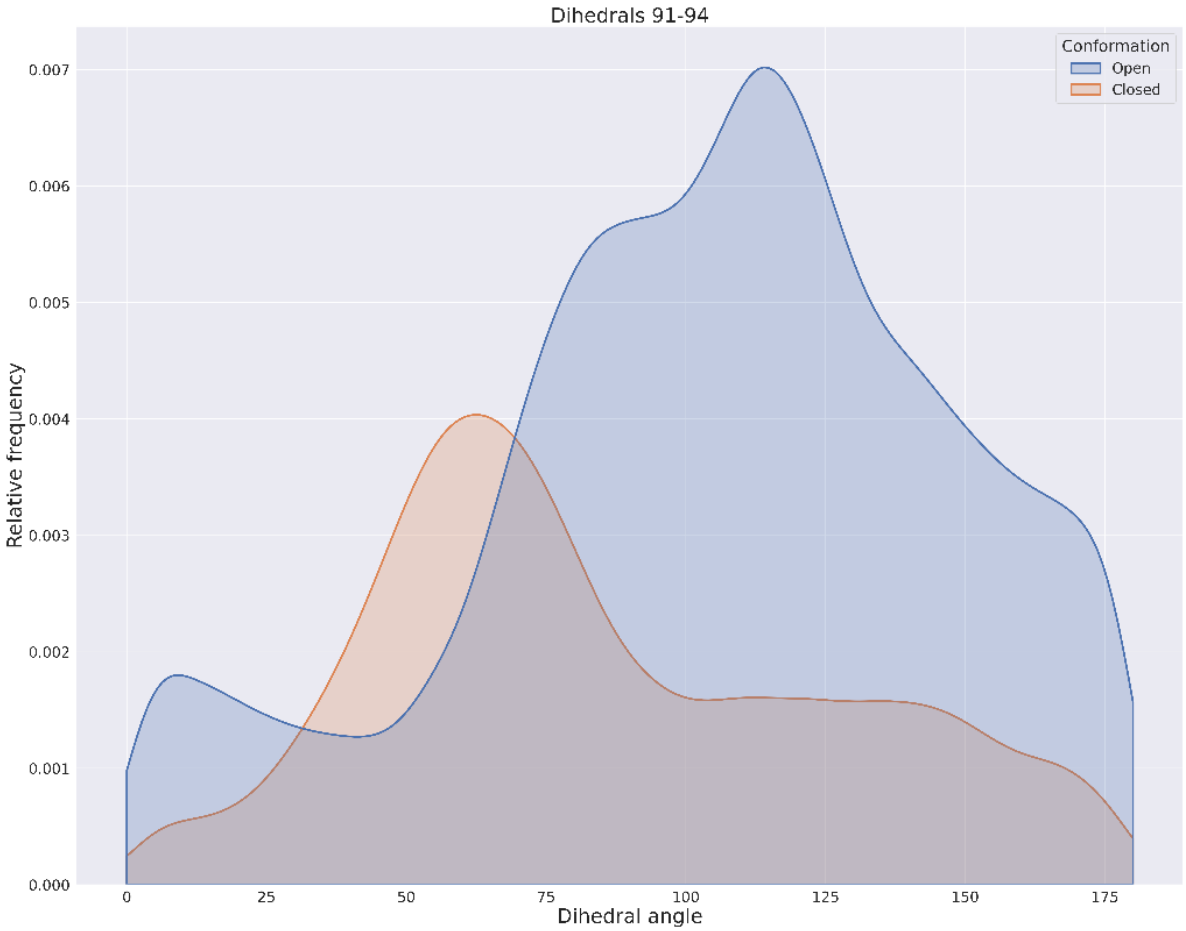


Figure S14: Distribution of the values assumed by the dihedral angle (from 0 to 180°) calculated from the C_α of residues 91 to 94. The distribution of dihedral angles associated with open conformations (Rg > 2.1 nm) are colored in blue, while the distribution of dihedral angles values for closed conformations in orange. The values were obtained on the concatenated TREMD trajectories collected for each HM. The distributions are normalized so that their underlying area is equal to one. The most frequent value assumed by the dihedral angle for closed conformations is 64°, while the most frequent value for open conformations is 114°. Together with the radius of gyration the dihedral angle corresponding to the hinge where the NTD fits into the ACD allows us to distinguish more closed conformations from more open ones.

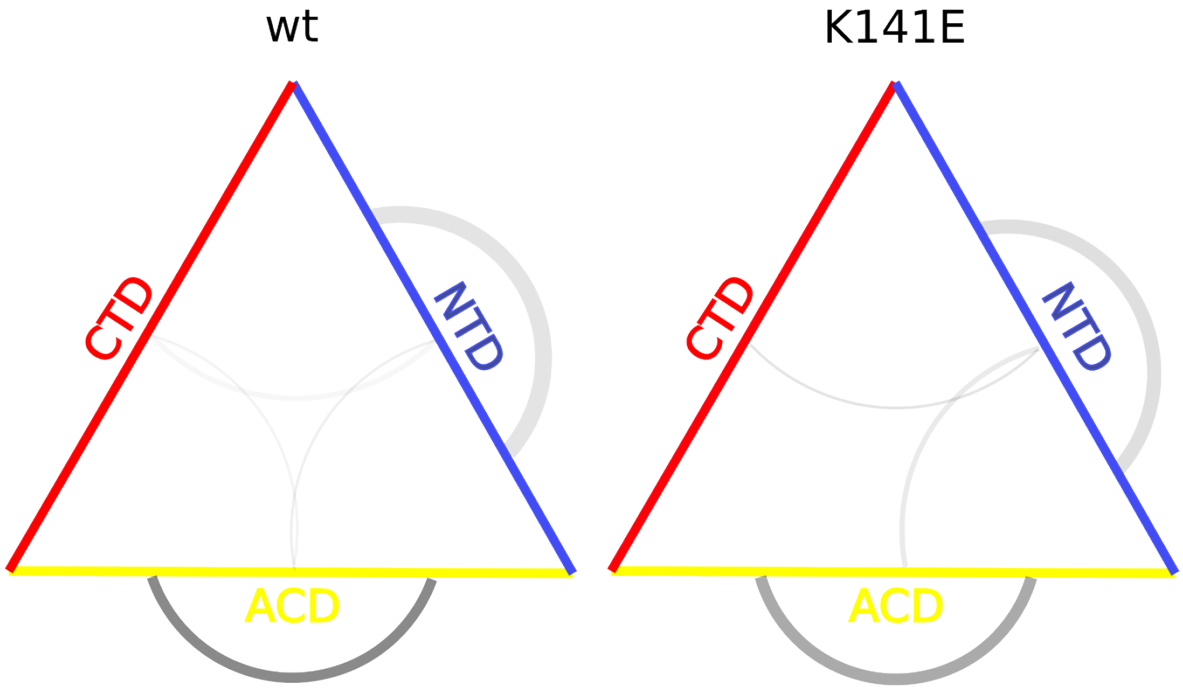


Figure S15: Graphical representation of the salt bridges between the three different HSPB8 domains: the structured ACD and the disordered NTD and CTD, for the *wt* and K141E concatenated TREMD trajectories. Both intra- and inter-domain salt bridges are represented as lines that connect a domain to itself or to another domain, respectively. Each line represents a mean salt bridge of the same kind: the opacity is proportional to the average number of frames in which the salt bridges were found, while the width is proportional to the number of different salt bridges.

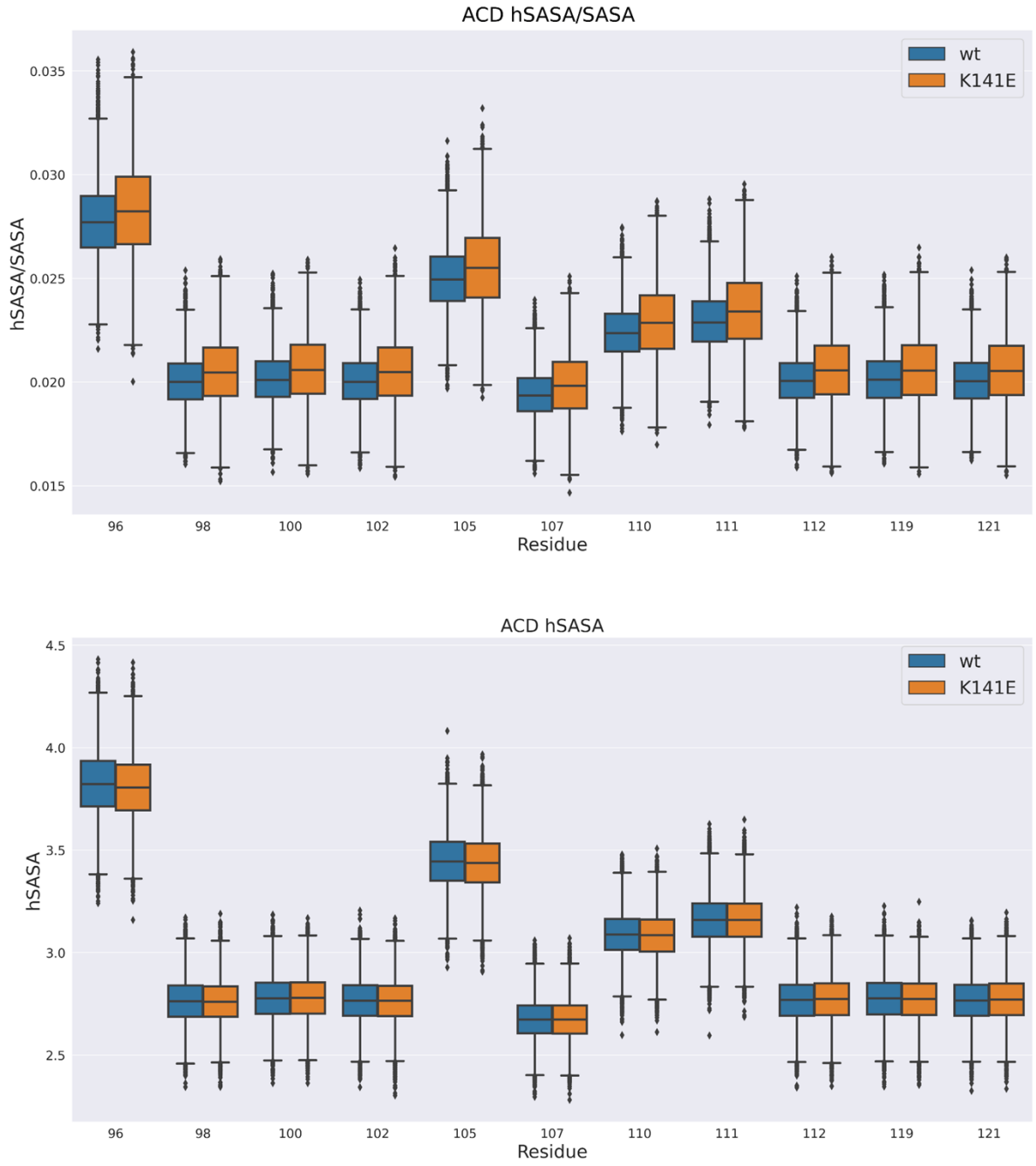


Figure S16: hSASA/SASA (top panel) and hSASA (bottom panel) boxplots of the ACD hydrophobic residues, for the *wt* (blue) and K141E (orange) HSPB8, calculated on the concatenated TREMD trajectories. The horizontal line in each boxplot represents the median value.

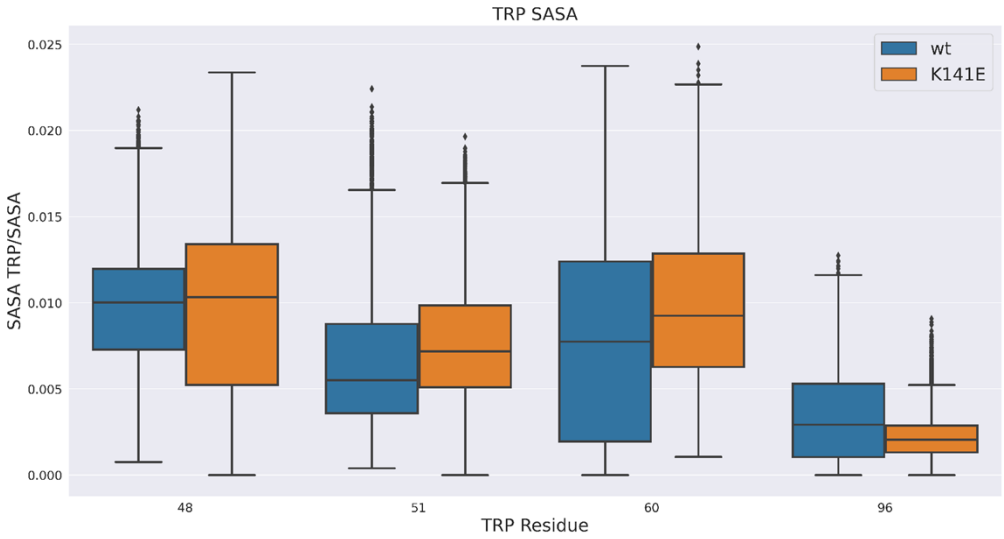


Figure S17: Boxplots of the distributions of single Trp residue SASA/protein SASA, for the *wt* (blue) and K141E (orange) HSPB8. The boxplots have been calculated on the concatenated TREMD trajectories. Four Trp residues are naturally found in HSPB8. The horizontal line in each boxplot represents the median value.

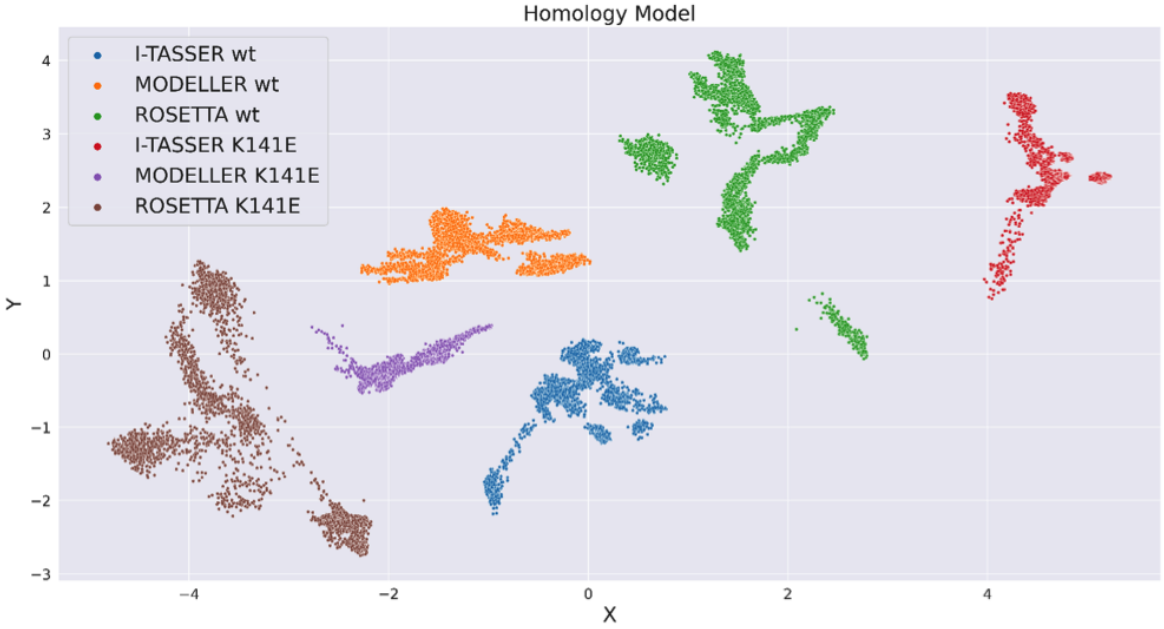


Figure S18: 2D KDE plot of each HMs using TREMD trajectories of *wt* and K141E HSPB8 encoded with the EncoderMap algorithm. The x-axis and y-axis correspond to the coordinates assigned by EncoderMap to the structures present in the trajectories. Different colors correspond to different HM TREMD trajectories.

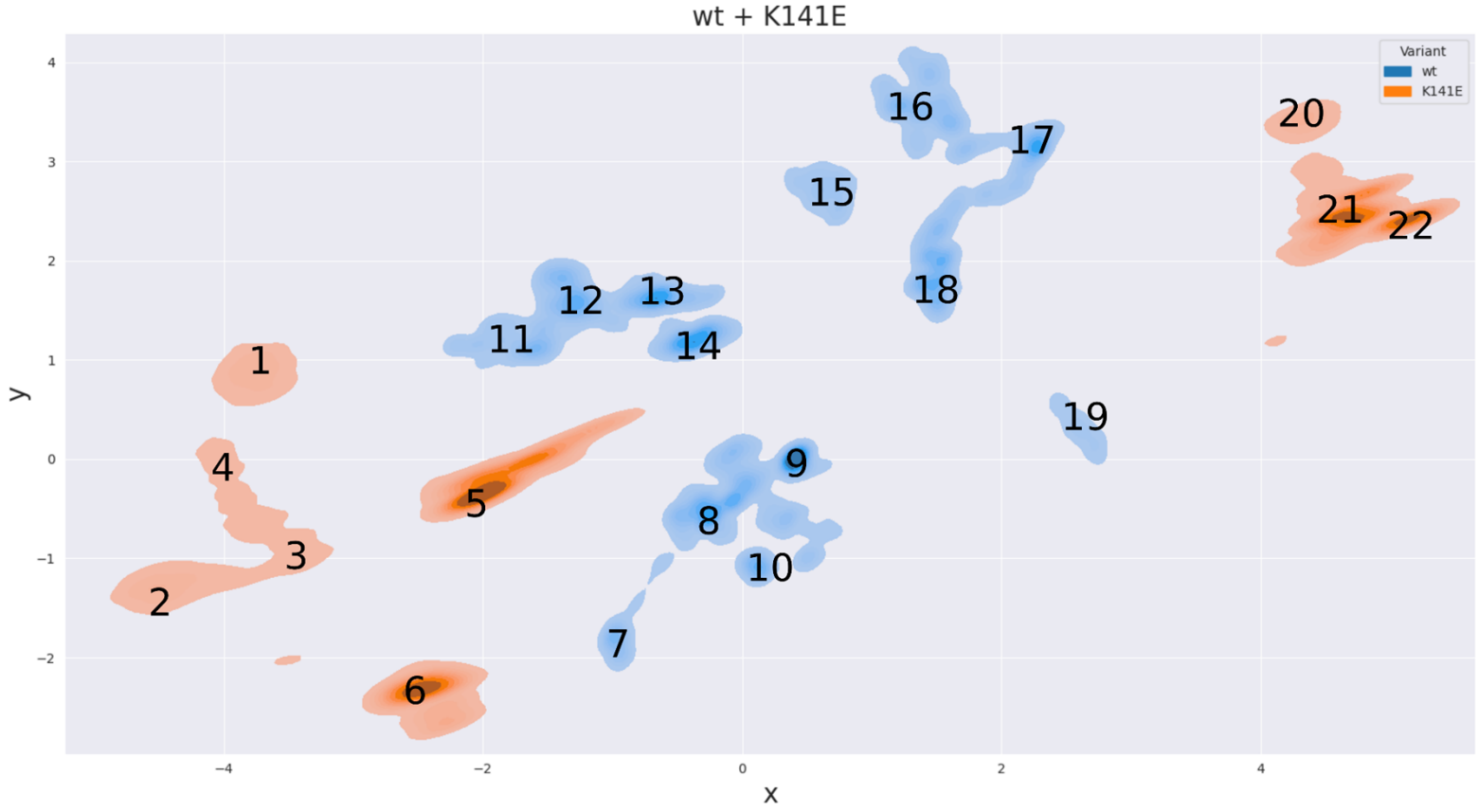


Figure S19: 2D KDE plot of the combined TREMD trajectories of *wt* (blue) and K141E (orange) HSPB8 encoded with the EncoderMap algorithm, as depicted in Fig. 5. At the points of minimum (the red crosses in Fig. 5), the numbers assigned to the structures of *wt* and K141E have been reported, used in Fig. S14.

	Rg (nm)	SASA (nm²)	hSASA (nm²)	hSASA/SASA
1 K141E	1.94	134.474	49.770	0.370
2 K141E	2.00	148.626	56.987	0.383
3 K141E	1.83	125.241	48.352	0.386
4 K141E	2.14	131.261	40.707	0.310
5 K141E	2.67	151.100	52.402	0.347
6 K141E	2.71	157.908	55.890	0.354
7 wt	2.04	135.919	45.090	0.332
8 wt	2.41	140.797	46.908	0.333
9 wt	2.30	129.081	42.198	0.327
10 *wt*	2.36	137.316	47.134	0.343
11 *wt*	2.31	137.656	51.415	0.373
12 *wt*	2.46	134.111	46.392	0.346
13 *wt*	2.21	143.266	55.325	0.386
14 *wt*	2.25	145.076	55.694	0.384
15 *wt*	1.90	130.774	41.493	0.317
16 *wt*	2.09	137.539	43.395	0.316
17 *wt*	2.71	149.247	50.879	0.341

18 *wt*	2.27	145.500	46.008	0.316
19 *wt*	2.32	138.924	46.357	0.334
20 K141E	2.33	145.973	50.394	0.345
21 K141E	2.46	141.130	47.671	0.338
22 K141E	2.04	133.771	43.605	0.326

Figure S20: Summary of the values of Radius of Gyration, SASA, hSASA, hSASA/SASA for the structures obtained from the EncoderMap plot, corresponding to the minimum points. The numbering is shown in Fig. S13. *Wt* structures are colored in blue, while K141E structures are in orange.

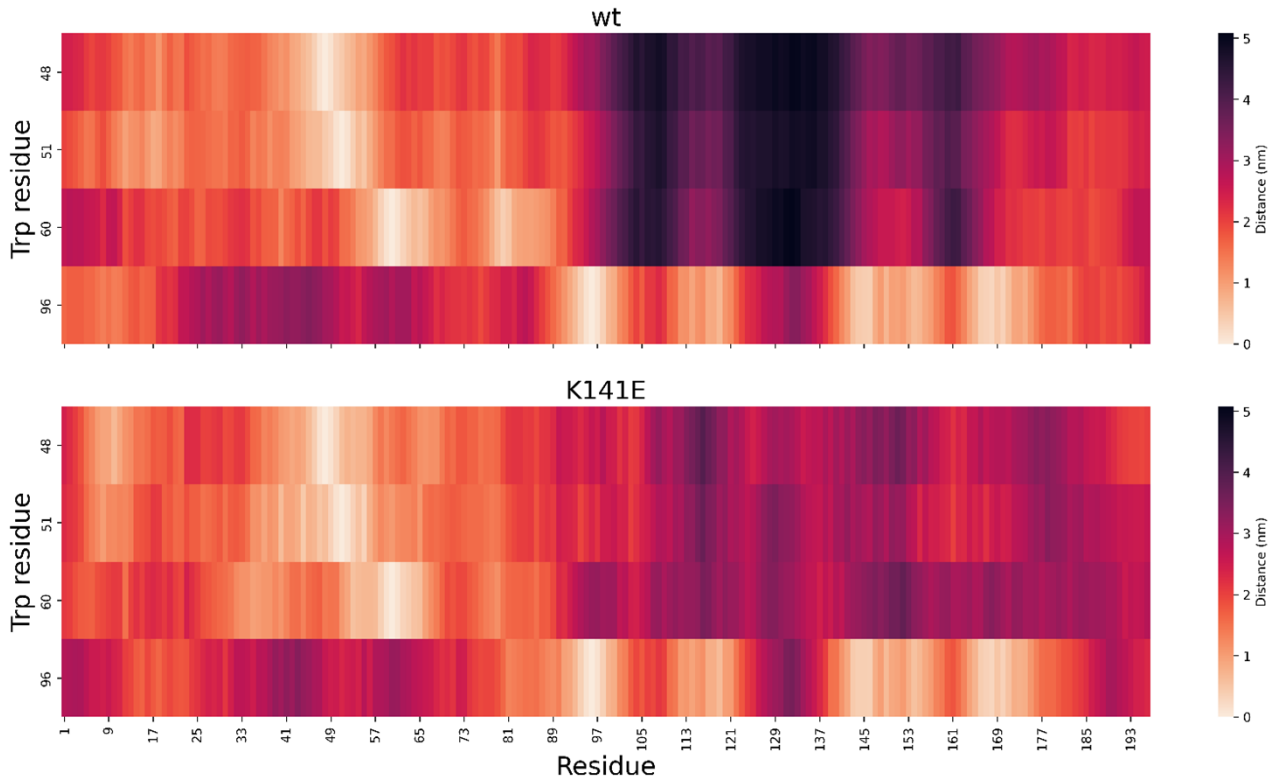


Figure S21: Heatmap of the distance between single Trp residues and other HSPB8 residues in *wt* (top panel) and K141E (bottom panel). For each Tpr-residue combination, displayed value correspond to the median distance in all the concatenated TREMD trajectory. The color scale associates lighter colors with smaller distances. We therefore observe that the environments of the Trp residues change from the *wt* to the K141E HSPB8, as for the mutated protein the Trp residues are more closely surrounded by the other residues.

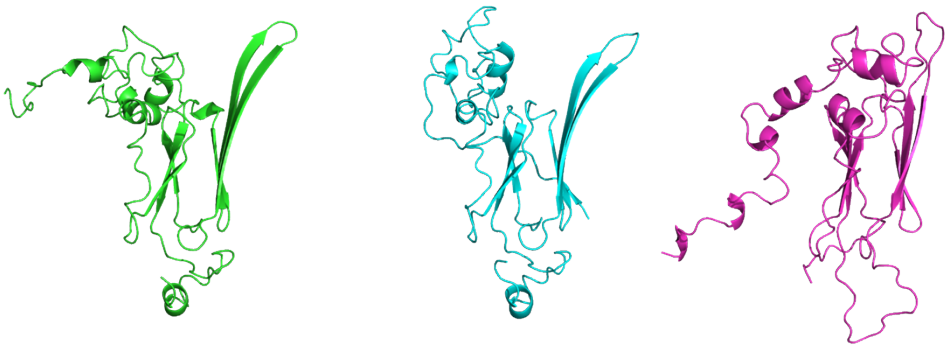


Figure S22: Pymol cartoon visualization of the homology models obtained for *wt* HSPB8 sequence with I-TASSER (green), MODELLER (cyan), and ROSETTA (magenta) algorithms.

SASA (p-value)			
	I-TASSER wt	MODELLER wt	ROSETTA wt
I-TASSER wt	-	-	-
MODELLER wt	6.69E-245	-	-
ROSETTA wt	0	9.10E-96	-
	I-TASSER K141E	MODELLER K141E	ROSETTA K141E
I-TASSER K141E	-	-	-
MODELLER K141E	0	-	-
ROSETTA K141E	0	1.60E-255	-
hSASA (p-value)			
	I-TASSER wt	MODELLER wt	ROSETTA wt
I-TASSER wt	-	-	-
MODELLER wt	0	-	-
ROSETTA wt	6.49E-209	0	-
	I-TASSER K141E	MODELLER K141E	ROSETTA K141E
I-TASSER K141E	-	-	-
MODELLER K141E	3.95E-291	-	-
ROSETTA K141E	2.59E-35	5.30E-140	-
Rg (p-value)			
	I-TASSER wt	MODELLER wt	ROSETTA wt
I-TASSER wt	-	-	-
MODELLER wt	1.68E-29	-	-
ROSETTA wt	3.25E-255	3.44E-309	-
	I-TASSER K141E	MODELLER K141E	ROSETTA K141E
I-TASSER K141E	-	-	-
MODELLER K141E	0	-	-
ROSETTA K141E	0	0	-

Figure S23: KS-test performed on different distributions of Rg, SASA, hSASA of the single homology models are compared pairwise. Results show that the distributions of SASA, hSASA, and Rg, taken in pairs, are statistically different.