Structural Analysis of Disulfide Bonds in the RCSB Protein Data Bank Using proteusPy

Eric G. Suchanek, Ph.D.

Flux-Frontiers
Cincinnati, OH

suchanek@flux-frontiers.com

This is a preprint of the following chapter: Eric G. Suchanek,
Structural Analysis of Disulfide Bonds in the RCSB Protein Data Bank Using proteusPy,
published in *Functional Disulphide Bonds: Methods and Protocols, Second Edition*,
edited by Philip Hogg, 2026, Springer Nature, reproduced with permission of Springer Nature.
The final authenticated version is available online at:
https://doi.org/10.1007/978-1-0716-5158-2

Abstract

A structural analysis of disulfide bonds extracted from the RCSB Protein Data Bank (PDB) is presented. This analysis utilized the **proteusPy** Python package and provides a statistical basis for yielding a dataset of geometrically reasonable disulfide bond structures. Additionally, a method for quantization of dihedral angles to reduce structural complexity and create disulfide structural classes is described. This method allows for the segmentation of the disulfide bonds into structural families. The results of this analysis provide a framework for understanding the geometric and structural families of disulfide bonds in the RCSB databank.

1 Introduction

Disulfide bonds are formed when two Cysteine residues are oxidized resulting in a sulfur-sulfur covalent bond. A schematic of a modelled disulfide bond is shown in Figure 1. The proximal and distal residues are '1', the chain is 'A' and the pdbid is '1egs'. This figure represents an illustration of disulfide bond connectivity and torsions.

These structural elements play pivotal roles in structural stabilization within and between protein subunits. Moreover, they participate in enzymatic catalysis, regulate protein activities, and offer protection against oxidative stress. Disulfide bonds also play an important role in the development of neurodegenerative diseases like Alzheimer's since they are intimately involved in tau protein aggregation[1, 2, 3].

The Python package **proteusPy** has been developed to facilitate analysis and characterization of disulfide bonds. This package is available on the GitHub repository[4] and contains the disulfide bond structural database extracted from the RCSB Protein Data Bank: `https://www.rcsb.org/`. **proteusPy** has a number of features including:

- Interactively display disulfides contained in the RCSB in a variety of display styles
- Calculate geometric and energetic properties about these disulfides

- Create binary and octant structural classes by quantizing the disulfide torsional angles
- Build idealized disulfide bonds from dihedral angle input
- Find disulfide neighbors based on dihedral angle input
- Overlap disulfides onto a common frame of reference for display
- Build protein backbones from backbone phi, psi dihedral angle templates

Refer to the **proteusPy** documentation for more information on the package features and usage. This is available at `https://suchanek.github.io/proteusPy`. The package is available for installation via GitHub at `https://github.com/suchanek/proteusPy`.

2 Methods

The general process employed to create the database involves the following steps:

1. Extraction of disulfide bonds from the RCSB Protein Data Bank. This required the development of a parser to extract the disulfide bond information. This seemingly simple task was complicated by the fact that there are many errors in the PDB files, and the parser had to be able to manage them. This led to the development of a robust logging system to track the parsing and ensure that the parser could handle the anomalies without crashing.
2. Construction of a master list of disulfide bonds using the **DisulfideClassManager** class. This list is stored on Google Drive and is downloaded automatically by the proteusPy package upon installation.
3. Filtering based on geometric criteria to ensure structural validity for analysis purposes.
4. Quantization of dihedral angles for structural classification.
5. Visualization of disulfide bonds via the `rcsbviewer.py` program.

2.1 Disulfide Database Creation

A query to identify proteins containing one or more disulfide bonds using the query builder on `https://www.rcsb.org` was used. This yielded 38,214 entries. These were downloaded via a Python script **DisulfideDownloader.py**. Once downloaded, the structures were extracted with the **proteusPy.DisulfideExtractor** entry point. This multi-processing program can extract over 36,000 PDB structure files in just under 3 minutes using 14 cores on a M3 Max Macbook Pro. Structures that were non-parsable were discarded. The remaining 36,968 were used as the basis for the geometric analysis.

2.2 Disulfide Bond Distance Analysis

The structures used are the product of X-Ray crystallography, and as such all models have errors associated with their atomic positions. Structure quality is generally tied to the **resolution** structural parameter. Resolutions < 2.0 Å are considered high quality with sufficient detail to easily distinguish individual atoms.

To ensure that the structures used in the analysis are geometrically reasonable, structures were filtered based on the criteria indicated above:

- The distance between the two sulfur atoms in the disulfide bond. This should be close to the covalent bond length of 2.04 Å.

- The distance between the two C_α atoms. This distance should not exceed 8 Å.
- Bond length deviations from ideality.
- Bond angle deviations from ideality.

Since the S_γ-S_γ bond is covalent, the interatomic distance should be close to the covalent bond length of 2.04 Å. The overall S_γ-S_γ and C_α-C_α distance distributions are shown in Table 1. The mean distance is 2.12 Å, with a standard deviation of 1.95 for the entire database. The distance between the proximal and distal residue C_α atoms is another useful screening property. This distance cannot exceed 8 Å, since this would yield a structure with physically impossible geometry. The C_α-C_α distance distribution before any filtering is shown in Figure 2. The mean distance is 5.67 Å, with a standard deviation of 1.87. This is a wide distribution, indicating that the quality of the structures varies significantly. There are also a number of structures with C_α-C_α distances well above 8 Å, indicating physically impossible geometry.

The longest disulfide bond in the database is not shown, since the C_α-C_α bond distance is over 100 Å. For interested readers, the ID is: 6zz8-178C-178D. This structure is clearly an outlier and would be discarded in the filtering process. Conversely, the shortest disulfide has a C_α-C_α distance of 2.71 Å, which is extremely short and well below the average. This is illustrated in Figure 3. The structure exhibits a highly strained configuration with a high torsional energy and practically *cis* S_γ-S_γ bond (Figure 4).

Table 1: Disulfide Bond Distances, Full Database

Distance (Å)	**Mean (Å)**	**Std**
S_γ-S_γ	2.12	1.95
C_α-C_α	5.67	1.87

Statistics (mean, standard deviation, confidence intervals) were calculated for the bond lengths and angles in order to filter them for reasonable geometry.

The cutoffs for the C_α-C_α and S_γ-S_γ bond distances are shown in Table 2.

Table 2: Disulfide Bond Distance Cutoffs for 95th Percentile

Bond	**Distance (Å)**
S_γ-S_γ	2.52
C_α-C_α	6.71

The application of the two filters resulted in the loss of 16,312 structures, leaving a total of 35,367 structures containing 158,965 disulfide bonds. The overall C_α-C_α distance distribution after filtering is shown in Figure 5. The S_γ-S_γ distance distribution after filtering is shown in Figure 6.

2.3 Disulfide Bond Strain Energy Analysis

The approximate bond torsional energy (kcal/mol) of a given disulfide bond is calculated using the following equation[5]:

$$\begin{aligned} E_{kcal\,mol^{-1}} \approx{} & 2.0 \cdot \cos(3.0 \cdot \chi_1) + \cos(3.0 \cdot \chi_{1'}) + \cos(3.0 \cdot \chi_2) \\ & + \cos(3.0 \cdot \chi_{2'}) + 3.5 \cdot \cos(2.0 \cdot \chi_3) \\ & + 0.6 \cdot \cos(3.0 \cdot \chi_3) + 10.1 \end{aligned} \tag{1}$$

The equation embodies the typical 3-fold rotation barriers associated with single bonds ($\chi_1, \chi_1', \chi_2, \chi_2'$), and a high 2-fold barrier for χ_3, resulting from the partial double bond character of the S_γ-S_γ bond. This property leads to two major disulfide families, characterized by the sign of χ_3. *Left-handed* disulfides have $\chi_3 < 0°$ and *right-handed* disulfides have $\chi_3 > 0°$. The overall distribution of disulfide bond torsional energies after filtering is shown in Figure 7[1].

2.4 Dihedral Angle Quantization

The overall conformation of a particular disulfide bond is characterized by five rotatable bonds. These are denoted χ_1, χ_2, χ_3, χ_2', χ_1', and reflect the overall 'twist' of the two residues between the first amino acid (proximal residue) and second amino acid (distal residue). The 'twist' between each of these sidechain bonds is known as the *torsion* or *dihedral* angles. Schmidt et al. [6] proposed a method for quantizing these angles into binary classes by calling a dihedral '+' if it is greater than 0 and '-' if it is less than 0. This method yielded $2^5 = 32$ possible classes. The proteusPy package uses:

- '0' represents a negative angle
- '2' represents a positive angle

For example, the class "00000" represents a disulfide bond where all five dihedral angles are negative, known as a "-LHSpiral" (Left-Handed Spiral).

2.4.1 Octant Quantization

The octant quantization system divides the 360° range of each dihedral angle into 8 equal segments of 45° each, numbered from 1 to 8 clockwise starting from 0° as shown in Table 3. This is illustrated in Figure 8. This creates $8^5 = 32,768$ possible classes, providing much finer granularity in classifying disulfide conformations.

Table 3: Dihedral Angle Octant Quantization

Octant	**Positive Range**	**Negative Equivalent**
1	315° to 360°	−45° to 0°
2	270° to 315°	−90° to −45°
3	225° to 270°	−135° to −90°
4	180° to 225°	−180° to −135°
5	135° to 180°	–
6	90° to 135°	–
7	45° to 90°	–
8	0° to 45°	–

2.4.2 Mapping Between Systems

The binary and octant systems are related as follows:

- Binary '0' (negative angle) maps to octants 1, 2, 3, and 4

---

[1] The distribution prior to filtering was similar

- Binary ‘2’ (positive angle) maps to octants 5, 6, 7, and 8

Since there are 32,768 possible octant classes, and only 32 binary classes, each binary class maps to 1024 octant classes. **proteusPy** provides a method (proteusPy.DisulfideLoader.toclass()) to map from binary to octant classes.

2.4.3 Example: Mapping Dihedrals to a Class String

Let’s consider a specific disulfide bond with the following dihedral angles:

$$\begin{aligned} \chi_1 &= -60^\circ \\ \chi_2 &= -60^\circ \\ \chi_3 &= -90^\circ \\ \chi_{2'} &= -60^\circ \\ \chi_{1'} &= -60^\circ \end{aligned}$$

Binary Classification: Since all angles are negative, they all map to ‘0’ in the binary system:

- Binary class string: “00000” (the “-LHSpiral” class)

Octant Classification:

- $\chi_1 = -60^\circ$ falls in the range -90° to $-45^\circ \rightarrow$ Octant 2
- $\chi_2 = -60^\circ$ falls in the range -90° to $-45^\circ \rightarrow$ Octant 2
- $\chi_3 = -90^\circ$ falls in the range -90° to $-45^\circ \rightarrow$ Octant 2
- $\chi_{2'} = -60^\circ$ falls in the range -90° to $-45^\circ \rightarrow$ Octant 2
- $\chi_{1'} = -60^\circ$ falls in the range -90° to $-45^\circ \rightarrow$ Octant 2

Thus, the octant class string is “22222”. This example demonstrates how the same disulfide bond is classified as “00000” in the binary system and “22222” in the octant system. The octant system then provides a structural filter that reduces the possible number of structures within a structural family.

2.5 Disulfide Bond Classification

2.5.1 Binary Classes

The binary class results are shown in Table 4. The class names are per Schmidt et al. [6]. The most common class is 00000b or ’—–’, representing a “-LHSpiral” conformation. This class represents 23.36% of the total disulfide bonds in the database. The least common class is 20202, representing a “+RHStaple” conformation, which represents 0.31% of the total disulfide bonds in the database.

Table 4: Disulfide Binary Classes in the Filtered (95%) RCSB Database

$\chi_1, \chi_2, \chi_3, \chi_2', \chi_1'$	**Count**	**Incidence (%)**	**Energy (kcal mol^{-1})**	C_α-C_α(Å)	**Classname**
-,-,-,-,-	39140	24.63	2.68 (0.84-4.53)	5.74 (5.37-6.11)	-LHSpiral
-,-,-,-,+	7489	4.71	3.25 (1.39-5.11)	6.04 (5.63-6.45)	-,-,-,-,+
-,-,-,+,-	4397	2.77	4.79 (1.84-7.73)	5.30 (4.55-6.05)	-LHHook
-,-,-,+,+	2189	1.38	3.90 (1.42-6.38)	5.37 (4.76-5.98)	-,-,-,+,+
-,-,+,-,-	15390	9.68	4.47 (2.98-5.96)	4.20 (3.75-4.66)	-RHStaple
-,-,+,-,+	1231	0.78	4.67 (1.86-7.48)	5.08 (4.48-5.69)	-,-,+,-,+
-,-,+,+,-	6638	4.18	4.19 (1.67-6.70)	5.33 (4.76-5.91)	-,-,+,+,-
-,-,+,+,+	6313	3.97	3.45 (1.08-5.82)	5.21 (4.74-5.68)	-,-,+,+,+
-,+,-,-,-	6512	4.10	3.93 (1.01-6.85)	5.71 (4.94-6.48)	-,+,-,-,-
-,+,-,-,+	7154	4.50	3.22 (0.73-5.70)	6.18 (5.63-6.74)	-,+,-,-,+
-,+,-,+,-	2910	1.83	4.64 (0.98-8.30)	5.42 (4.40-6.44)	-LHStaple
-,+,-,+,+	1046	0.66	4.22 (1.18-7.25)	5.52 (4.66-6.38)	-,+,-,+,+
-,+,+,-,-	6470	4.07	3.76 (1.03-6.49)	5.38 (4.88-5.88)	-RHHook
-,+,+,-,+	813	0.51	5.89 (3.10-8.68)	5.86 (5.38-6.33)	-,+,+,-,+
-,+,+,+,-	7910	4.98	3.20 (0.92-5.47)	5.81 (5.40-6.21)	-RHSpiral
-,+,+,+,+	6647	4.18	4.00 (2.20-5.80)	6.25 (5.87-6.63)	-,+,+,+,+
+,-,-,-,-	4211	2.65	3.32 (1.33-5.32)	6.03 (5.66-6.41)	±LHSpiral
+,-,-,-,+	1256	0.79	3.86 (1.08-6.64)	6.26 (5.92-6.61)	+LHSpiral
+,-,-,+,-	1043	0.66	5.65 (2.52-8.78)	5.75 (5.15-6.36)	±LHHook
+,-,-,+,+	703	0.44	4.45 (1.92-6.97)	5.67 (5.08-6.26)	+LHHook
+,-,+,-,-	629	0.34	5.86 (2.93-8.79)	5.29 (4.58-6.00)	±RHStaple
+,-,+,-,+	317	0.20	7.72 (4.51-10.9)	5.57 (4.76-6.38)	+RHStaple
+,-,+,+,-	1593	1.00	4.53 (1.82-7.23)	5.59 (5.10-6.08)	±RHHook
+,-,+,+,+	639	0.40	6.52 (3.44-9.60)	6.09 (5.53-6.65)	+,-,+,+,+
+,+,-,-,-	3401	2.14	3.27 (0.72-5.83)	5.52 (4.79-6.24)	-/+LHHook
+,+,-,-,+	1866	1.17	3.69 (0.78-6.60)	6.03 (5.55-6.52)	+,+,-,-,+
+,+,-,+,-	1632	1.03	4.58 (1.40-7.76)	5.12 (4.05-6.19)	±LHStaple
+,+,-,+,+	500	0.32	4.87 (1.25-8.49)	5.65 (4.92-6.37)	+LHStaple
+,+,+,-,-	7765	4.89	3.04 (0.89-5.20)	5.13 (4.72-5.54)	-/+RHHook
+,+,+,-,+	654	0.41	5.03 (2.13-7.93)	5.70 (5.17-6.23)	+RHHook
+,+,+,+,-	3260	2.05	3.43 (1.13-5.72)	5.94 (5.58-6.29)	±RHSpiral
+,+,+,+,+	7220	4.54	3.70 (2.20-5.21)	6.36 (6.06-6.67)	+RHSpiral

2.5.2 Octant Classes

The octant classification strategy was used since the 45 degree octants span the normal tri-fold dihedral minima of ±60°, and the number of possible classes (8^5 or 32,768) is manageable with current computational resources. These classes are constructed at runtime when the **DisulfideDownloader** class is instantiated.

In order to map from a binary class string to all possible octant class strings the **itertools** python library is used. This allows for the generation of all permutations of the class strings efficiently. The octant classification for the binary class 00000b is shown in Figure 9.

The **DisulfideClassAnalysis.py** program was used to generate the binary and octant class distributions and statistics. This program enumerates class membership for these classes and calculates and displays dihedral angle and distance statistics for each class. Given the large number of octant classes present (over 9000), a cutoff factor was introduced to reduce the number of classes displayed to a manageable number. The typical cutoff factor used was 0.04% which yields 325 described classes. This means that only classes with an incidence of 0.04% or greater are described.

Consider binary class 00000b (the "-LHSpiral" structural family, which has the highest representation in the entire database at 24.63%) once more (Figure 9); the two most prominent peaks correspond to octant classes 22222o and 22322o, respectively. These differ only by the movement of χ_3 into octant class 3 ($-135°$ to $-90°$). Only 617/1024 possible classes are present in this class. The dihedral angle statistics for the two classes are shown in Figure 10 and Figure 11. The error bars display relatively small values, suggesting that the class structures are structurally similar. Boxplots for the binary and octant classes are presented in Figure 12 and Figures 13 through 16. The boxplot for binary classes reveals a broad range of energy variation across classes, whereas the octant classes exhibit a narrower distribution. The tighter clustering of octant classes reflects the 1024x better resolution than a binary classification. The figures illustrate a wide range of torsional energies and there are several classes that occupy low energy regions of the space. The lowest energy disulfide in the entire database is from pdbid $2q7q75D - 140D$, with an approximate torsional energy of 0.49 kcal mol^{-1}. This is virtually identical to the theoretical minimum energy calculated through the **scikit-learn**[2]. The modeled and actual structures are shown in Figure 17.

The **proteusPy.DisulfideClassGenerator** class is used to visualize an ensemble of the average structure for each of these peaks. This class takes the mean and standard deviations for the dihedral angles for each octant class and generates an idealized disulfide bond structure ensemble. These represent the average structure for each class with their respective dihedral angle standard deviations. Each dihedral angle therefore has three possible states, (mean − std, mean, mean + std). This generates 3^5 or 243 structures. An example is shown in Figure 18 and Figure 19. The figures are unable to capture the real differences between the ensemble structures, but the areas of space covered and accessible to the subclasses are readily apparent when viewed in 3D.

3 Discussion

The analysis has yielded a high-quality dataset of disulfide bonds segmented into structural families based on eight-fold (octant) dihedral angle quantization. Notable observations include:

- Disulfide bond model quality varies significantly in the RCSB database. Filtering at 95% confidence removes over 16,000 bonds based on C_α-C_α and S_γ-S_γ distances alone.

- The disulfide structural classes present represent ≈ 28% of the total possible classes (≈ 8,500/32,768), which is a significant reduction in structural complexity. This indicates that the structural space of

---

[2] Data not shown but available in the proteusPy repository.

disulfide bonds is constrained heavily; local non-bonded interactions and steric hindrance are likely to be the primary factors.

- The octant approach allows for the further elucidation of structural families within the binary classification
- There are disulfides whose conformation is close to the global energy minimum predicted through scikit-learn.
- Ensemble structures for disulfide consensus classes can be used to predict sites in proteins of known structure capable of accommodating those bonds.

Acknowledgments

The author would like to acknowledge the wonderful resource provided by `http://www.rcsb.org` for providing the structural data used in this study. In addition, I'd like to thank Philip Hogg for his insight and counsel.

A Figures

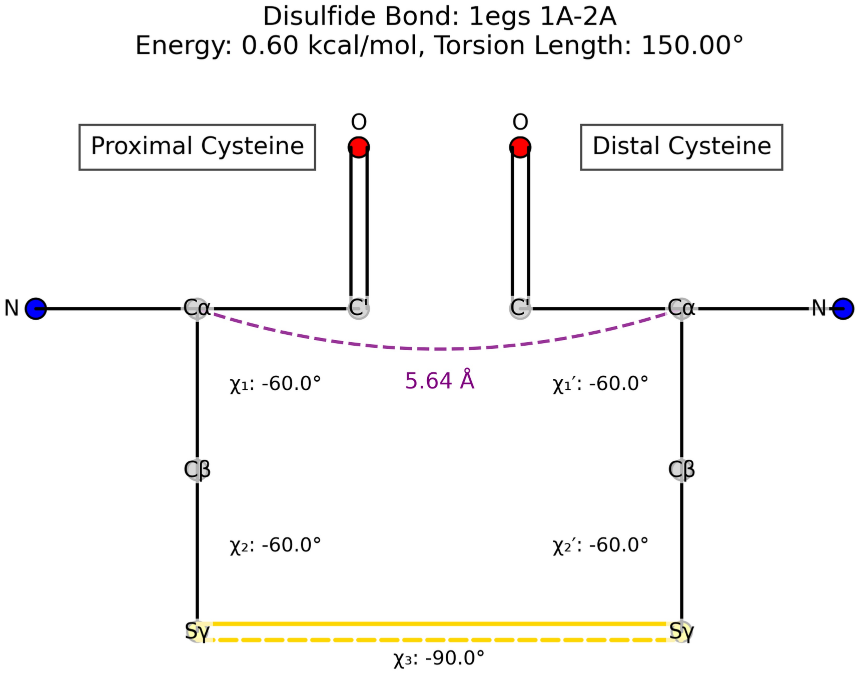


Figure 1: Schematic of a Modelled Disulfide Bond

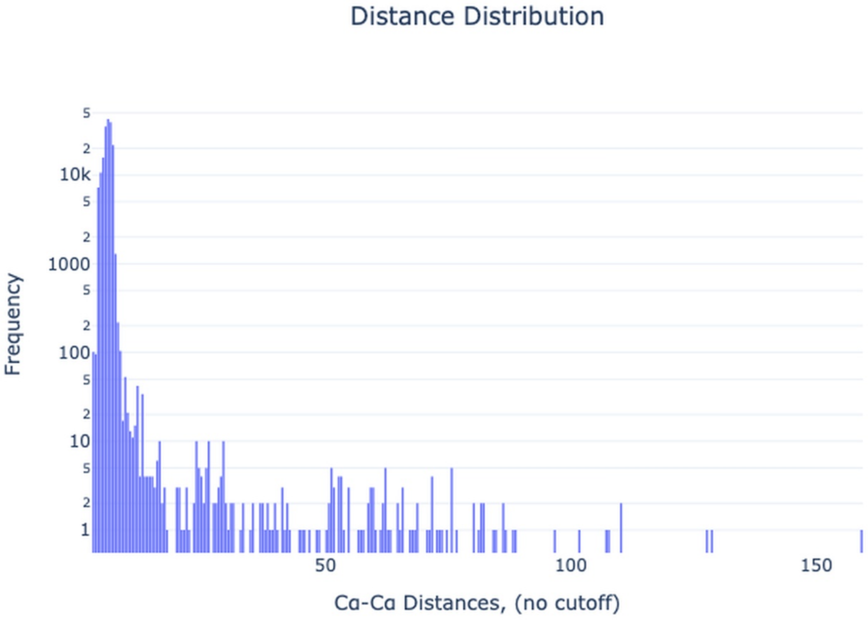


Figure 2: Histogram of Unfiltered Disulfide Bond C_α-C_α Distances

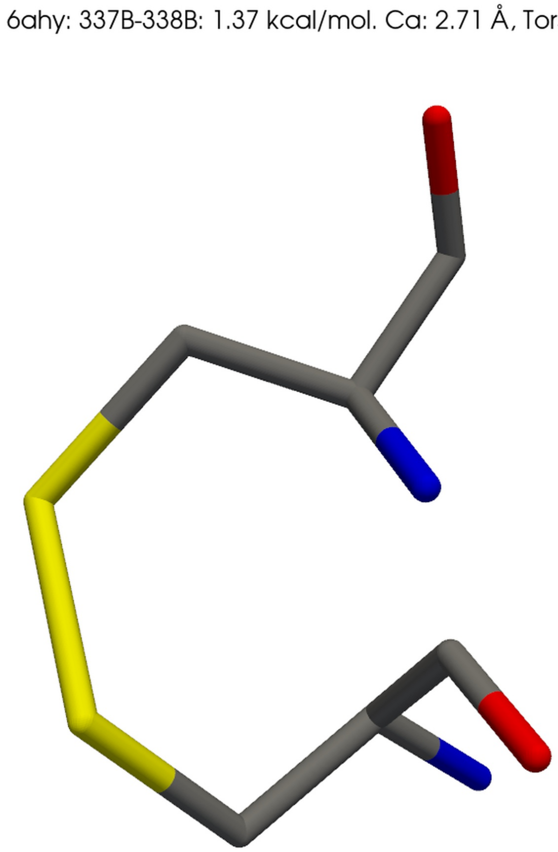


Figure 3: The Shortest Disulfide in RCSB

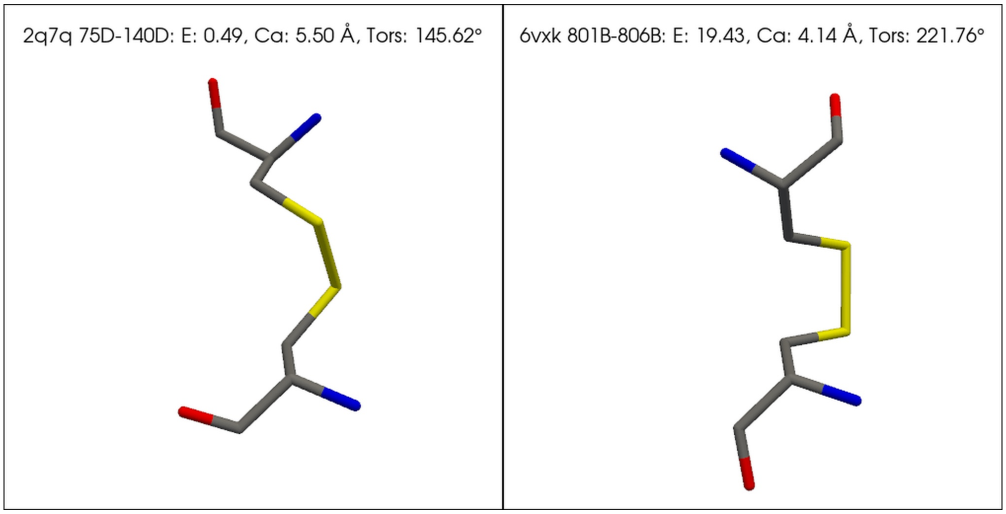


Figure 4: Minimum and Maximum Torsional Energy Structures

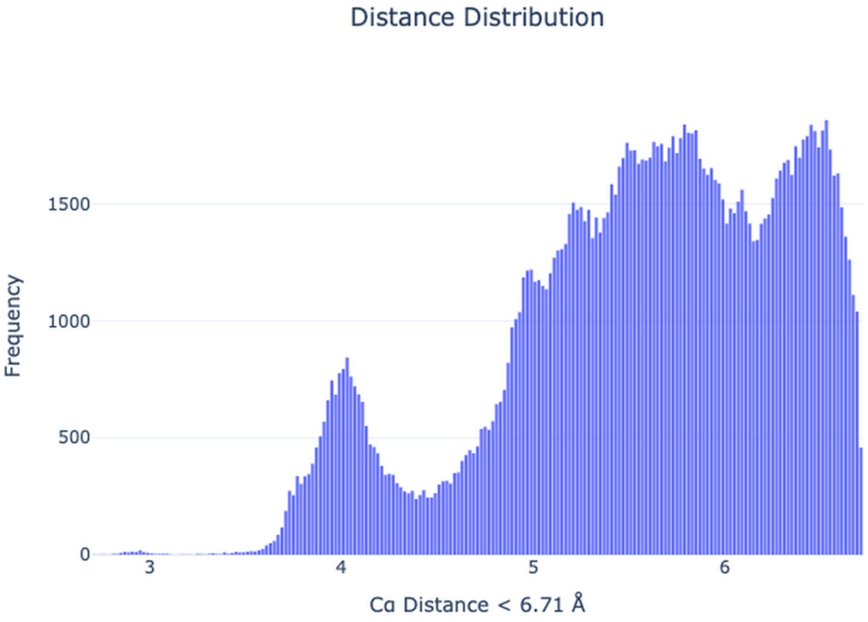


Figure 5: Distribution of Disulfide Bond C_α-C_α Distances After Filtering

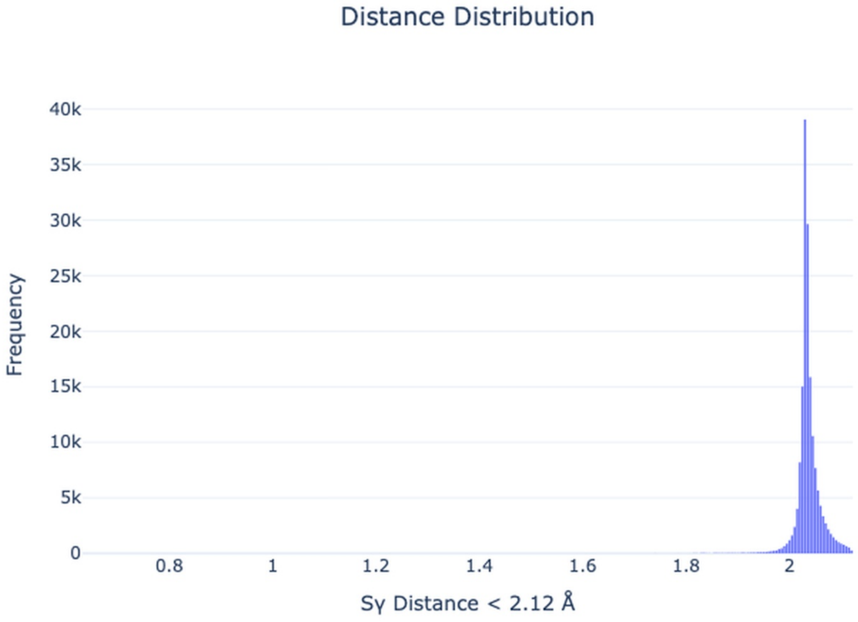


Figure 6: Distribution of Disulfide Bond S_γ-S_γ Distances After Filtering

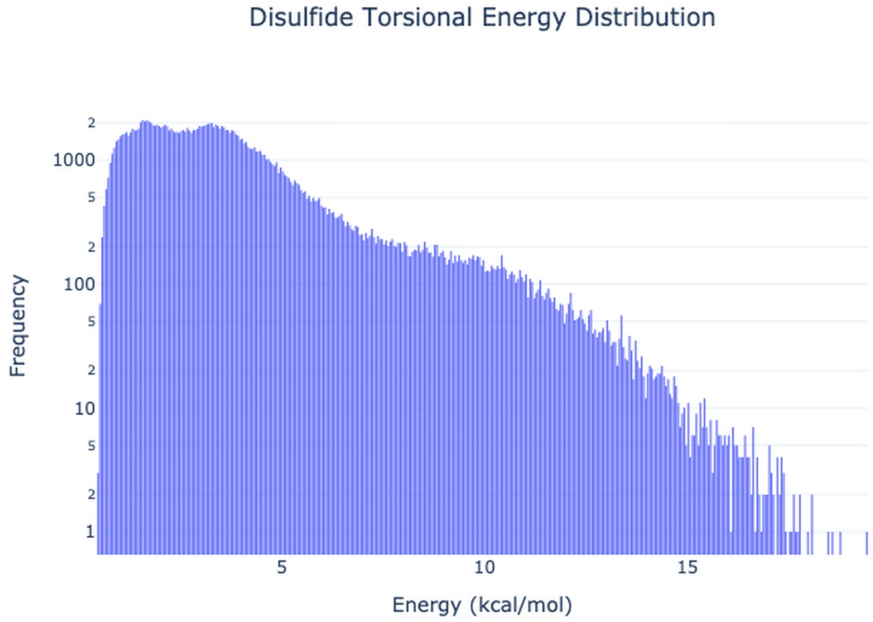


Figure 7: Distribution of Disulfide Bond Torsional Energies

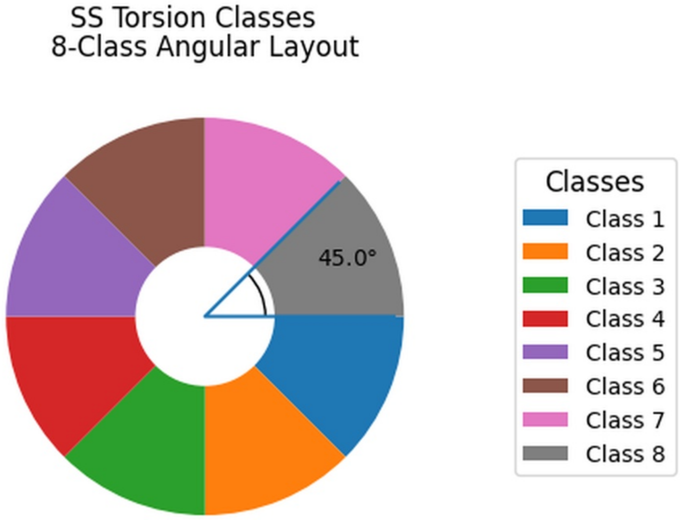


Figure 8: Disulfide Bond Octant Classification

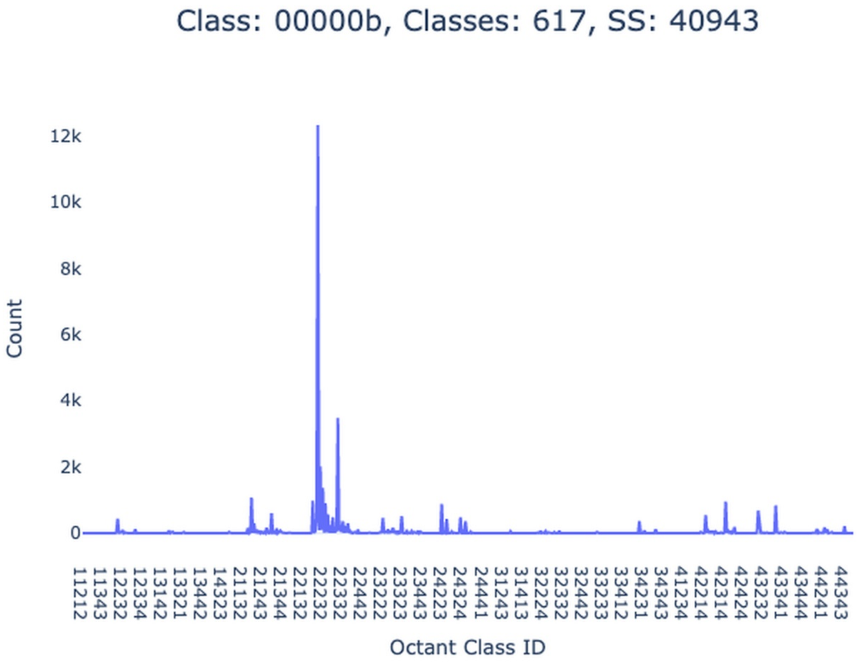


Figure 9: Octant classes for binary class 00000b

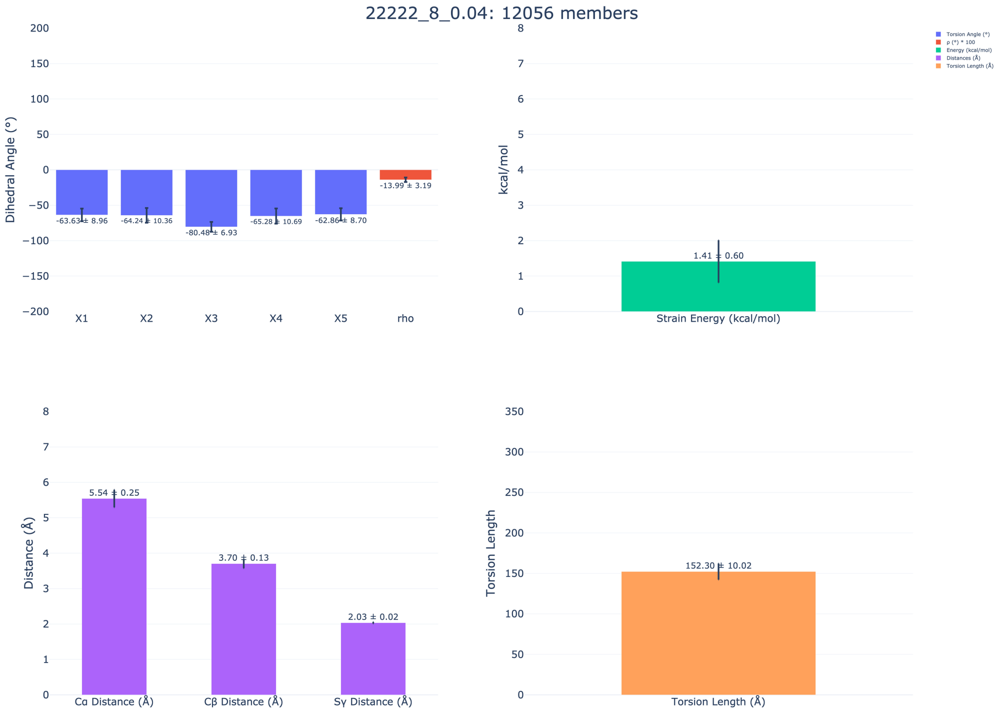


Figure 10: Class 22222o Dihedral Angle Details

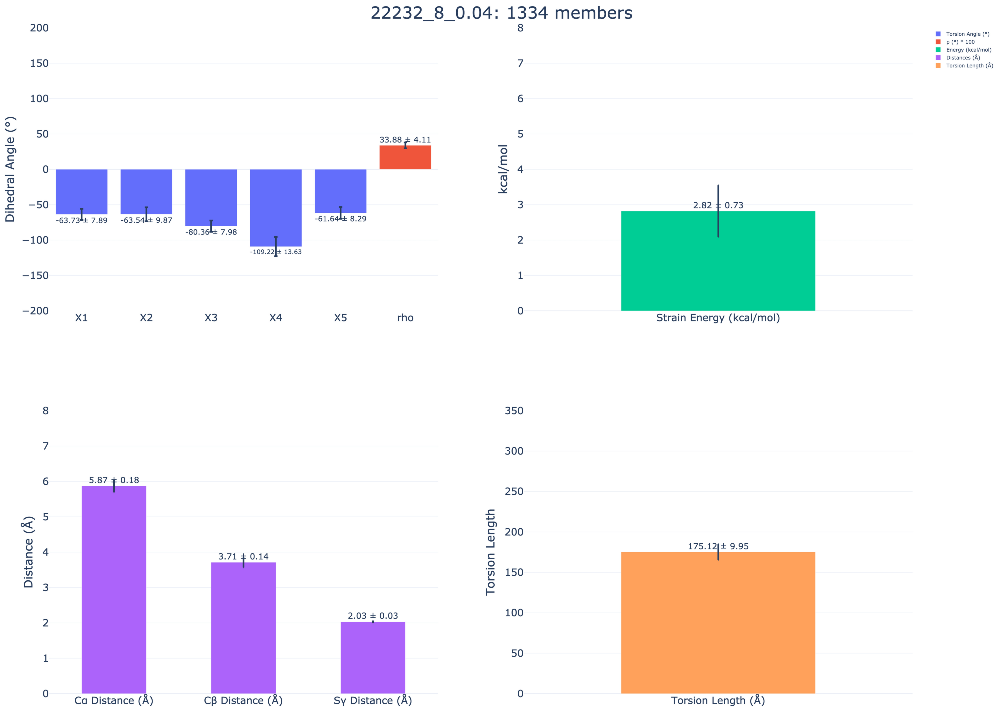


Figure 11: Class 22322o Dihedral Angle Details

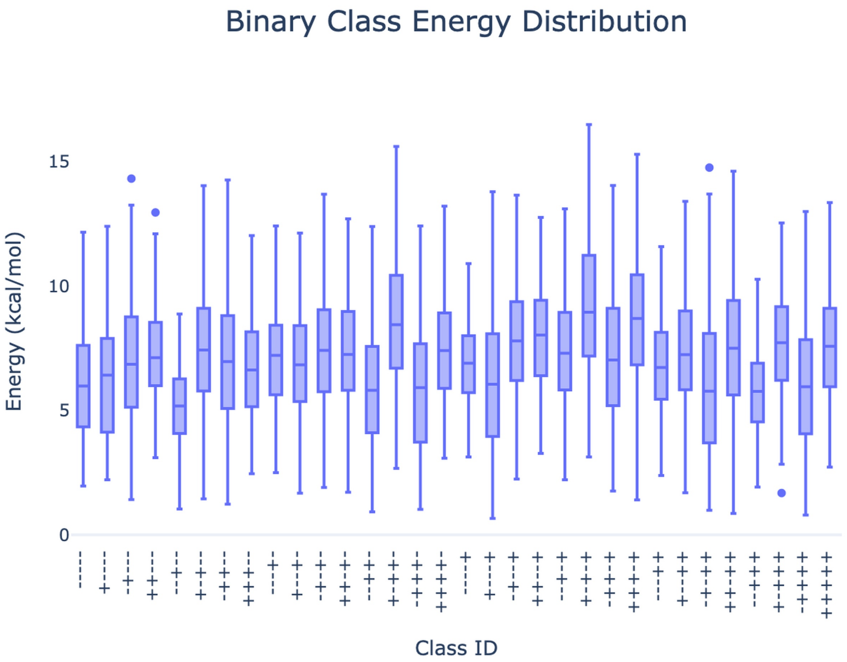


Figure 12: Disulfide Bond Torsional Energies by Binary Class

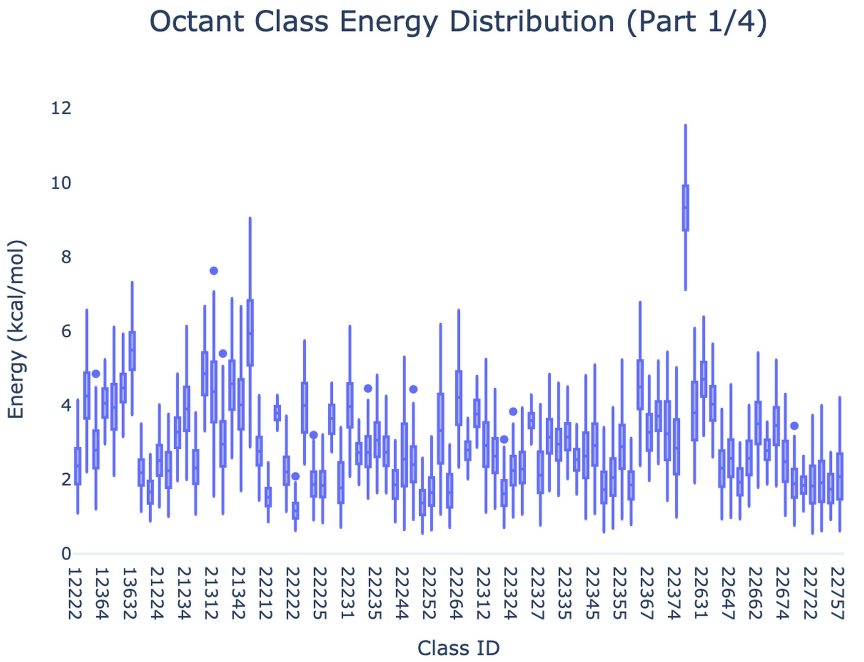


Figure 13: Disulfide Bond Torsional Energies by Octant Class (Part 1)

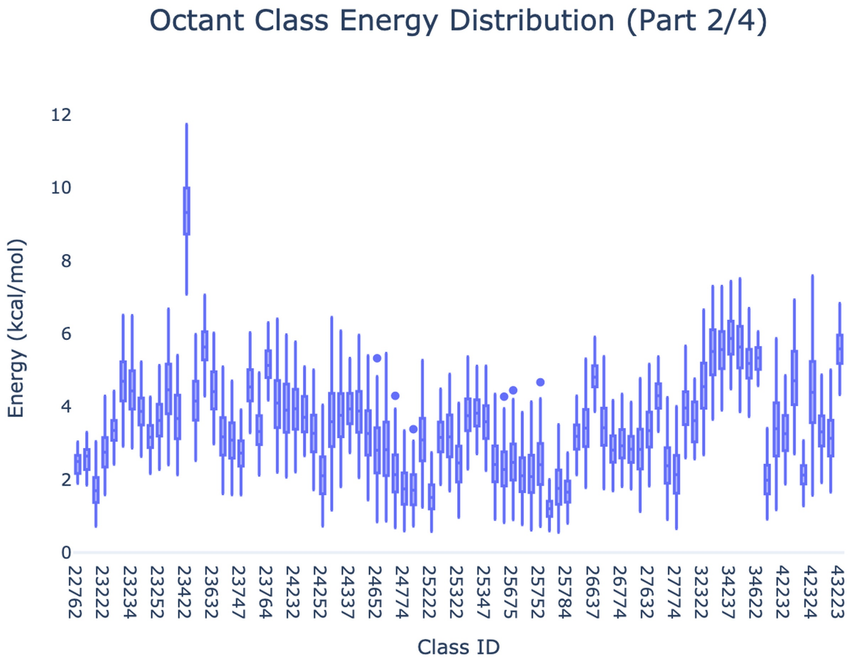


Figure 14: Disulfide Bond Torsional Energies by Octant Class (Part 2)

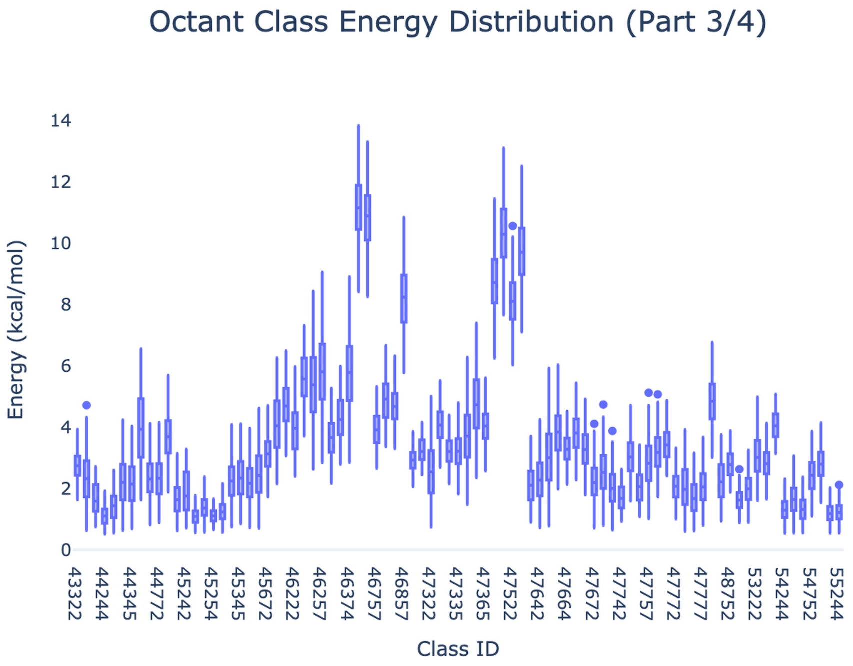


Figure 15: Disulfide Bond Torsional Energies by Octant Class (Part 3)

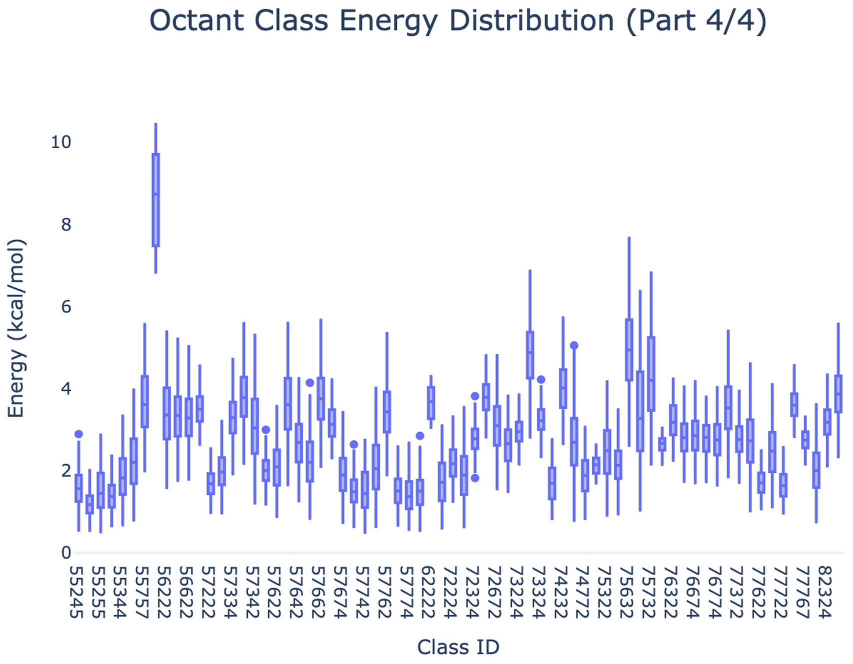


Figure 16: Disulfide Bond Torsional Energies by Octant Class (Part 4)

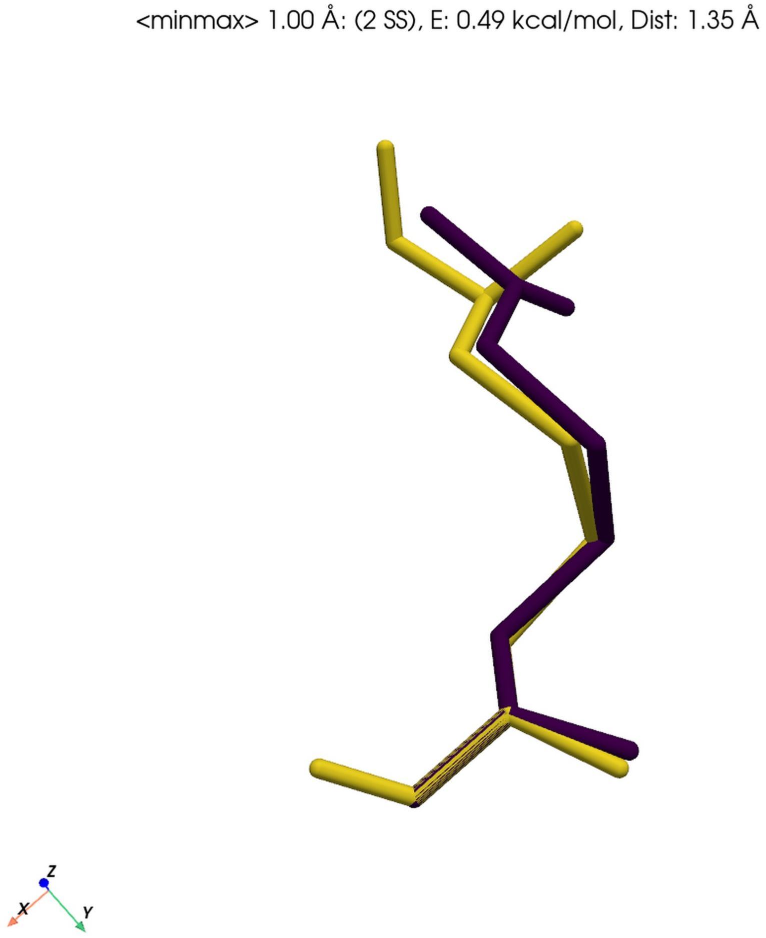


Figure 17: Overlay of Modeled and Actual 2q7q Disulfide Structure

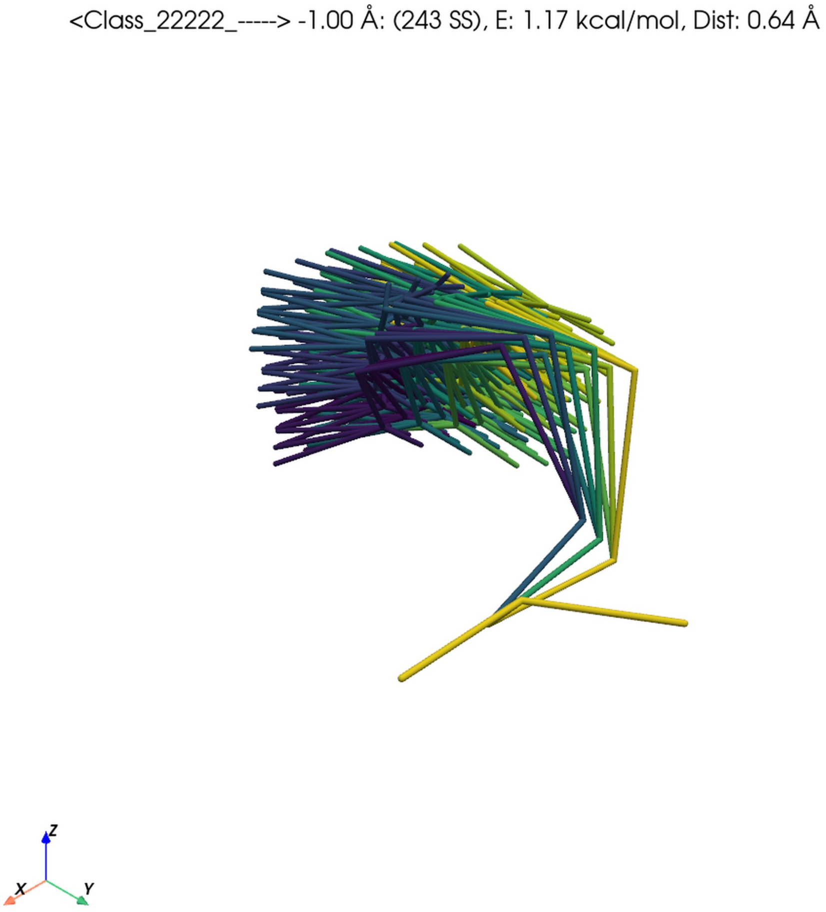
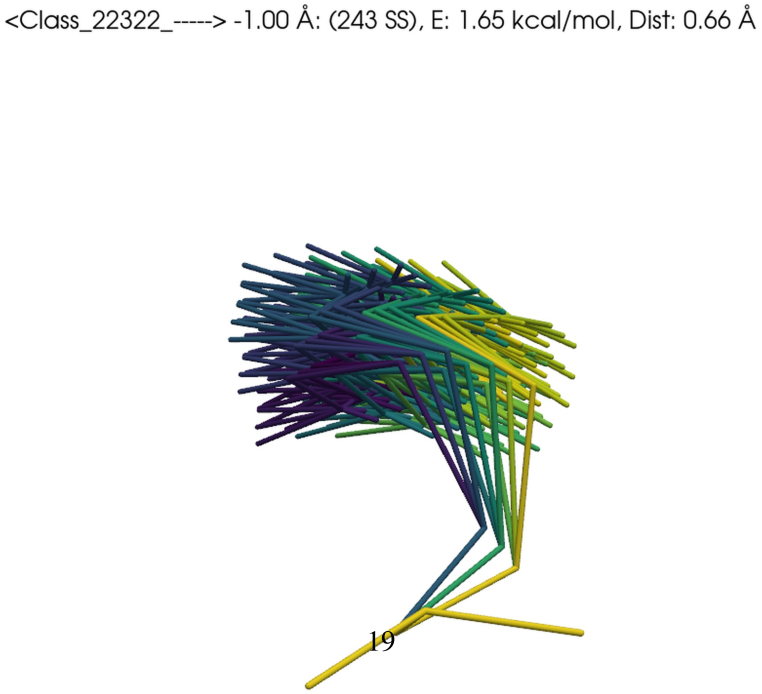


Figure 18: Class 22222o ensemble



References

[1] M. A. Meraz-Ríos, K. I. Lira-De León, V. Campos-Peña, M. A. De Anda-Hernández, and R. Mena-López. Tau oligomers and aggregation in alzheimer's disease. *Journal of Neurochemistry*, 112(6):1353–1367, mar 2010.

[2] T. Saito, T. Chiku, M. Oka, S. Wada-Kakuda, M. Nobuhara, T. Oba, K. Shinno, S. Abe, A. Asada, A. Sumioka, A. Takashima, T. Miyasaka, and K. Ando. Disulfide bond formation in microtubule-associated tau protein promotes tau accumulation and toxicity in vivo. *Human Molecular Genetics*, 30(21):1955–1967, nov 2021.

[3] M. F. Mossuto. Disulfide bonding in neurodegenerative misfolding diseases. *International Journal of Cell Biology*, 2013:318319, aug 2013.

[4] Eric G. Suchanek. proteuspy: A package for modeling and analyzing proteins of known structure, 2024. GitHub repository.

[5] C. O. Pabo and E. G. Suchanek. Computer-aided model-building strategies for protein design. *Biochemistry*, 25(20):5987–5991, oct 1986.

[6] Brian Schmidt, Lorraine Ho, and Philip J. Hogg. Allosteric disulfide bonds. *Biochemistry*, 45(24):7429–7433, jun 2006.