




1 proteusPy: A Python Package for Protein Structure 2 and Disulfide Bond Modeling and Analysis

3 Eric G Suchanek  ^{1*}

4 ¹ Flux-Frontiers * These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: Richard Gowers 

Reviewers:

- [@AnjaConev](#)

Submitted: 11 November 2023

Published: unpublished

License

Authors of papers retain copyright
and release the work under a
Creative Commons Attribution 4.0
International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

5 Summary

6 **proteusPy** is a Python package specializing in the modeling and analysis of proteins of known
7 structure with an initial focus on Disulfide bonds. This package significantly extends the
8 capabilities of the molecular modeling program **proteus**, (Pabo & Suchanek, 1986), and utilizes
9 a new implementation of the **Turtle3D** class for disulfide and protein modeling. This initial
10 implementation focuses on the **Disulfide** class, which implements methods to analyze the
11 protein structure stabilizing element known as a **Disulfide Bond**.

12 The work has resulted in a freely-accessible database of over 120,494 disulfide bonds contained
13 within 35,818 proteins in the **RCSB Protein Databank**. The routines within the library are
14 capable of extracting, comparing, and visualizing the disulfides contained within the database,
15 facilitating analysis and understanding. In addition, the package can readily model disulfide
16 bonds of arbitrary conformation, facilitating predictive analysis.

7 General Capabilities

- 18 ▪ Interactively display disulfides contained in the RCSB in a variety of display styles
- 19 ▪ Calculate geometric and energetic properties about these disulfides
- 20 ▪ Create binary and sextant structural classes by characterizing the disulfide torsional
21 angles into n classes
- 22 ▪ Build idealized disulfide bonds from disulfide dihedral angle input
- 23 ▪ Find disulfide neighbors based on dihedral angles
- 24 ▪ Overlap disulfides onto a common frame of reference for display
- 25 ▪ Build protein backbones from backbone phi, psi dihedral angle templates
- 26 ▪ More in development

27 See <https://suchanek.github.io/proteusPy/proteusPy.html> for the API documentation with
28 examples

29 Statement of Need

30 Disulfide bonds represent the sole naturally occurring covalent bond in proteins, playing a
31 pivotal role in structural stabilization within and between protein subunits. Moreover, they
32 participate in enzymatic catalysis, regulate protein activities, and offer protection against
33 oxidative stress. Establishing an accessible structural database of these disulfides would serve
34 as an invaluable resource for exploring these critical structural elements. While the capability
35 to visualize protein structures is well established with excellent protein visualization tools like
36 Pymol, Chimera and the RCSB itself, the tools for disulfide bond analysis are more limited.
37 (Wong & Hogg, 2010) describe a web-based disulfide visualization tool; this is currently
38 unavailable.

39 Accordingly, I have developed the **proteusPy** package to delve into the RCSB Protein Data
40 Bank, furnishing tools for visualizing and analyzing the disulfide bonds contained therein.
41 This endeavor necessitated the creation of a python-based package containing data structures
42 and algorithms capable loading, manipulating and analyzing these entities. Consequently,
43 an object-oriented database has been crafted, facilitating introspection, analysis, and display.
44 The package's API is accessible online at: [proteusPy API](#), offering comprehensive details and
45 numerous illustrative examples.

46 Requirements

- 47 1. PC running MacOS, Linux, Windows with git, git-lfs and make installed
- 48 2. 8 GB RAM
- 49 3. 1 GB disk space

50 Installation

51 It's simplest to clone the repo via GitHub since it contains all of the notebooks, data and
52 test programs. Installation includes installing my Biopython fork. This is required to rebuild
53 the database. I highly recommend using Miniforge since it includes mamba. The installation
54 instructions below assume a clean install with no package manager or compiler installed.

55 MacOS/Linux

- 56 ■ Install Miniforge: <https://github.com/conda-forge/miniforge> (existing Anaconda instal-
57 lations are fine but please install mamba)
- 58 ■ Install git-lfs:
 - 59 – <https://help.github.com/en/github/managing-large-files/installing-git-large-file-storage>
- 60 ■ Install make on your system.
- 61 ■ From a shell prompt while sitting in your repo dir:

```
$ git clone https://github.com/suchanek/proteusPy.git  
$ cd proteusPy  
$ make pkg  
$ mamba activate proteusPy  
$ mamba install vtk  
$ make install
```

62 Windows

- 63 ■ Install Miniforge: <https://github.com/conda-forge/miniforge> (existing Anaconda instal-
64 lations are fine but please install mamba)
- 65 ■ Install git for Windows and configure for Bash:
 - 66 – <https://git-scm.com/download/win>
- 67 ■ Install git-lfs:
 - 68 – <https://git-lfs.github.com/>
- 69 ■ Install GNU make:
 - 70 – <https://gnuwin32.sourceforge.net/packages/make.htm>
- 71 ■ Open a Miniforge prompt and cd into your repo dir:

```
(base) C:\Users\egs\repos> git clone https://github.com/suchanek/proteusPy.git
(base) C:\Users\egs\repos> cd proteusPy
(base) C:\Users\egs\repos\proteuspy> make pkg
(base) C:\Users\egs\repos\proteuspy> conda activate proteusPy
(proteusPy) C:\Users\egs\repos> make install
```

72 Testing

73 I currently have pytest and docstring testing for the modules in place. To run them cd into
74 the repository and run:

```
$ make tests
```

75 The modules will 1) run pytest for the main modules and 2) perform docstring tests. This will
76 result in a number of disulfide visualization windows to open. Simply close them. If all goes
77 normally there will be no errors. (you may need to install pytest via `pip install pytest`.)

78 Usage

79 Once the package is installed it's possible to load, visualize and analyze the Disulfide bonds in
80 the RCSB Disulfide database. The general approach is:

- 81 ▪ Load the database
- 82 ▪ Access disulfide(s)
- 83 ▪ Analyze
- 84 ▪ Visualize

85 A simple example is shown below:

```
import proteusPy
from proteusPy import Load_PDB_SS, Disulfide

PDB_SS = Load_PDB_SS(verbose=True)

best_ss = PDB_SS["2q7q_75D_140D"]
best_ss.display(style="sb", light=True)
```

86 The [notebooks](#) directory contains my Jupyter notebooks and is a good place to start:

- 87 ▪ [Analysis_2q7q.ipynb](#) provides an example of visualizing the lowest energy Disulfide con-
88 tained in the database and searching for nearest neighbors on the basis of conformational
89 similarity.
- 90 ▪ [Anearest_relatives.ipynb](#) gives an example of searching for disulfides based on sequence
91 similarity.

92 The [programs](#) subdirectory contains the primary programs for downloading the RCSB disulfide-
93 containing structure files, extracting the disulfides and creating the disulfide database:

- 94 ▪ [DisulfideDownloader.py](#): Downloads the raw RCSB structure files.
- 95 ▪ [DisulfideExtractor.py](#): Extracts the disulfides and creating the database loaders
- 96 ▪ [DisulfideClass_Analysis.py](#): Performs binary or sextant analysis on the disulfide database.

97 The first time one loads the database via `Load_PDB_SS()` the system will attempt to download
98 the full and subset database from Google Drive. If this fails it's possible to rebuild the database
99 from the repo's `data` subdirectory (not the package's) by: `pip install -e .` at the repository
100 top-level. If you've downloaded from github this will work correctly. If you've installed from
101 pyPi via `pip` it will fail.

102 Quickstart

103 After installation is complete, launch jupyter lab:

```
104 $ jupyter notebook
```

105 and open [Analysis_2q7q](#). This notebook analyzes the disulfide bond with the lowest energy in the entire database and performs some searches in dihedral angle space to find similar conformations. There are several other notebooks in this directory that illustrate using the program. Some of these reflect active development work so may not be 'fully baked'.

108 Class Details

109 The primary classes developed for **proteusPy** are described briefly below. Please see the [API](#) for details.

111 Disulfide

112 This class provides a Python object and methods representing a physical disulfide bond either extracted from the RCSB protein databank or a virtual one built using the [Turtle3D](#) class. The disulfide bond is an important intramolecular stabilizing structural element and is characterized by:

- 116 ■ Atomic coordinates for the atoms $N, C_\alpha, C_\beta, C', S_\gamma$ for both amino acid residues. These are stored as both raw atomic coordinates as read from the RCSB file and internal local coordinates.
- 117
- 118
- 119 ■ The dihedral angles $\chi_1 - \chi_5$ for the disulfide bond
- 120 ■ A name, by default: $\{\text{pdb_id}\}\{\text{prox_resnum}\}\{\text{prox_chain}\}_{\{\text{distal_resnum}\}}\{\text{distal_chain}\}$
- 121
- 122 ■ Proximal residue number
- 123 ■ Distal residue number
- 124 ■ Approximate bond torsional energy (kcal/mol):

$$125 E_{kcal/mol} \approx 2.0 * \cos(3.0 * \chi_1) + \cos(3.0 * \chi_5) + \cos(3.0 * \chi_2) + \cos(3.0 * \chi_4) + 3.5 * \cos(2.0 * \chi_3) + 0.6 * \cos(3.0 * \chi_3) + 10.1$$

- 126 ■ Euclidean length of the dihedral angles (degrees) defined as:

$$\sqrt{\chi_1^2 + \chi_2^2 + \chi_3^2 + \chi_4^2 + \chi_5^2}$$

- 127 ■ $C_\alpha - C_\alpha$ distance (Å)
- 128 ■ $C_\beta - C_\beta$ distance (Å)
- 129 ■ The previous C' and next N coordinates for both the proximal and distal residues. These are needed to calculate the backbone dihedral angles ϕ, ψ .
- 130
- 131 ■ Backbone dihedral angles ϕ and ψ , when possible. Not all structures are complete and in those cases the atoms needed may be undefined. In this case the ϕ and ψ angles are set to -180° .
- 132
- 133

134 The class also provides 3D rendering capabilities using the excellent [PyVista](#) library, and can display disulfides interactively in a variety of display styles:

- 136 ■ 'sb' - Split Bonds style - bonds colored by their atom type
- 137
- 137 ■ 'bs' - Ball and Stick style - split bond coloring with small atoms
- 138
- 138 ■ 'pd' - Proximal/Distal style - bonds colored *Red* for proximal residue and *Green* for the distal residue.
- 139

140 ▪ 'cpk' - CPK style rendering, colored by atom type:

- 141 – Carbon - Grey
- 142 – Nitrogen - Blue
- 143 – Sulfur - Yellow
- 144 – Oxygen - Red
- 145 – Hydrogen - White

146 Individual renderings can be saved to a file and animations can be created. The *cpk* and *bs*
147 styles are illustrated below:

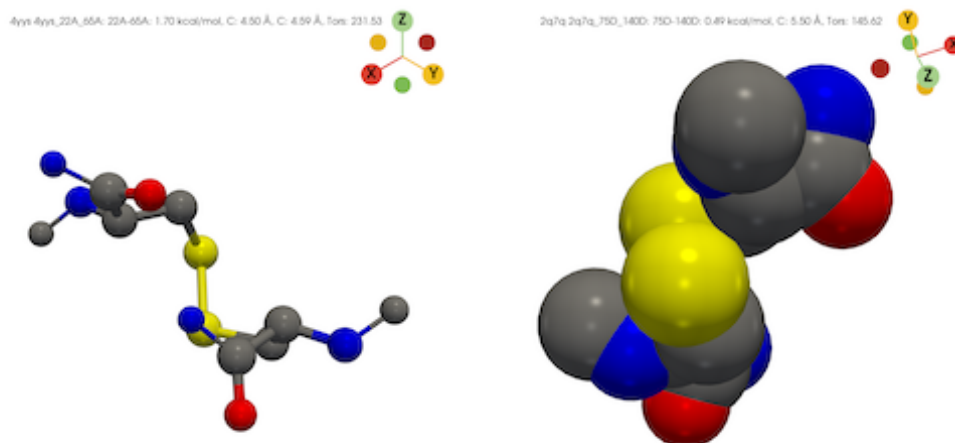


Figure 1: CPK & BS Disulfide Rendering

148 **DisulfideLoader**

149 This class encapsulates the disulfide database itself and is its primary means of accession.
150 Instantiation takes 2 parameters: **subset** and **verbose**. Given the size of the database, one can
151 use the **subset** parameter to load the first 1000 disulfides into memory. This facilitates quicker
152 development and testing new functions. I recommend using a machine with 16GB or more to
153 work with the full dataset.

154 The entirety of the RCSB disulfide database is stored within the class via a **DisulfideList**, a
155 **Pandas** .csv file, and a **dict** of indices mapping the RCSB IDs into their respective list of
156 disulfide bond objects. The datastructures allow simple, direct and flexible access to the
157 disulfide structures contained within. This makes it possible to access the disulfides by array
158 index, RCSB structure ID or disulfide name.

159 Example:

```
160 import proteusPy
161 from proteusPy import Disulfide, DisulfideLoader, DisulfideList
162
163 SS1 = DisulfideList([], 'tmp1')
164 SS2 = DisulfideList([], 'tmp2')
165
166 PDB_SS = DisulfideLoader(verbose=False, subset=True)
167
168 # Accessing by index value:
169 SS1 = PDB_SS[0]
```

```
170 SS1
171 <Disulfide 4yys_22A_65A, Source: 4yys, Resolution: 1.35 Å>
172
173 # Accessing by PDB_ID returns a list of Disulfides:
174 SS2 = PDB_SS['4yys']
175 SS2
176 [<Disulfide 4yys_22A_65A, Source: 4yys, Resolution: 1.35 Å>,
177 <Disulfide 4yys_56A_98A, Source: 4yys, Resolution: 1.35 Å>,
178 <Disulfide 4yys_156A_207A, Source: 4yys, Resolution: 1.35 Å>]
179
180 # Accessing individual disulfides by their name:
181 SS3 = PDB_SS['4yys_56A_98A']
182 SS3
183 <Disulfide 4yys_56A_98A, Source: 4yys, Resolution: 1.35 Å>
184
185 # Finally, we can access disulfides by regular slicing:
186 SSlist = SS2[:2]
187 [<Disulfide 4yys_56A_98A, Source: 4yys, Resolution: 1.35 Å>,
188 <Disulfide 4yys_156A_207A, Source: 4yys, Resolution: 1.35 Å>]
```

189 The class can also render Disulfides overlaid on a common coordinate system to a pyVista window using the [DisulfideLoader.display_overlay\(\)](#) method.

191 **NB:** For typical usage one accesses the database via the `Load_PDB_SS()` function. This function loads the compressed database from its single source. Initializing a `DisulfideLoader` object will load the individual torsions and disulfide `.pkl` files, builds the classlist structures, and writes the completely built object to a single `.pkl` file. This requires the raw `.pkl` files created by the download process. These files are contained in the `repository data` directory, not in the `pyPi` distribution.

197 turtle3D

198 The `turtle3D` class represents an object that maintains a *local coordinate system* in three dimensional space. This coordinate system consists of:

- 200 ■ A Position in 3D space
- 201 ■ A Heading Vector
- 202 ■ A Left Vector
- 203 ■ An Up Vector

204 The *Heading*, *Left* and *Up* vectors are unit vectors that define the object's orientation in a *local* coordinate frame. The turtle developed in `proteusPy` is based on the excellent book by Abelson: ([Abelson & DiSessa, 1986](#)). The `to_local` and `to_global` methods convert between these two coordinate systems. These methods make it possible to readily compare different disulfides by:

- 209 1. Orienting the turtle at the disulfide's proximal residue in a standard orientation.
- 210 2. Converting the global coordinates of the disulfide as read from the RCSB into local coordinates.
- 211 3. Saving all of the local coordinates with the raw coordinates
- 212 4. Performing distance and angle calculations

214 By implementing the functions **Move**, **Roll**, **Yaw**, **Pitch** and **Turn** the turtle is capable of movement in a three-dimensional space. See ([Pabo & Suchanek, 1986](#)) for more details.

216 The turtle has several molecule-specific functions including `orient_at_residue` and `orient_from_backbone`. These routines make it possible to build protein backbones of arbitrary

218 conformation and to readily add sidechains to modeled structures. These functions are
219 currently used to build model disulfides from dihedral angle input.

220 Examples

221 I illustrate a few use cases for the package below. Use the **jupyter notebook** command from
222 your shell to launch jupyter. The examples illustrate the ease with which one can analyze and
223 visualize some disulfides.

224 Find the lowest and highest energy disulfides present in the database

```
from proteusPy import Load_PDB_SS, DisulfideList, Disulfide

# load the database
PDB_SS = Load_PDB_SS(verbose=True, subset=False)

# retrieve the minimum and maximum energy structures
ssMin, ssMax = PDB_SS.SSList.minmax_energy

# make a list to hold them
minmaxlist = DisulfideList([ssMin, ssMax], "minmax")

# display them as ball and stick style
minmaxlist.display(style="bs", light=True)
```

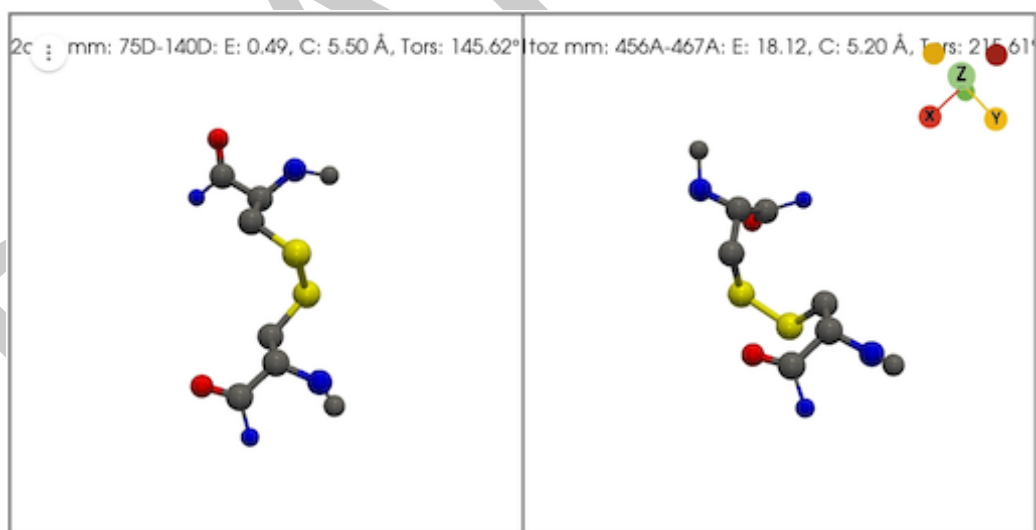


Figure 2: minmax

225 Find disulfides within 10 Å RMS in torsional space of the lowest energy 226 structure

227 In this example we load the disulfide database, find the disulfides with the lowest and highest
228 energies, and then find the nearest conformational neighbors. Finally, we display the neighbors
229 overlaid against a common reference frame. Note that the window title gives statistics about
230 the list of disulfides being displayed, including list name, resolution, number, average energy,
231 and average atom positional error.

```

import proteusPy
from proteusPy Load_PDB_SS, DisulfideList, Disulfide

PDB_SS = None
PDB_SS = Load_PDB_SS(verbose=False, subset=False)
ss_list = DisulfideList([], "tmp")

# Return the minimum and maximum energy structures. We ignore the maximum in this case.
ssmin_enrg, _ = PDB_SS.SSList.minmax_energy

# Make an empty list and find the nearest neighbors within 10 degrees avg RMS in
# sidechain dihedral angle space.

low_energy_neighbors = DisulfideList([], "Neighbors")
low_energy_neighbors = ssmin_enrg.Torsion_neighbors(sslist, 10)

# Display the number found, and then display them overlaid onto their common reference f

tot = low_energy_neighbors.length
low_energy_neighbors.display_overlay()
232 18

```

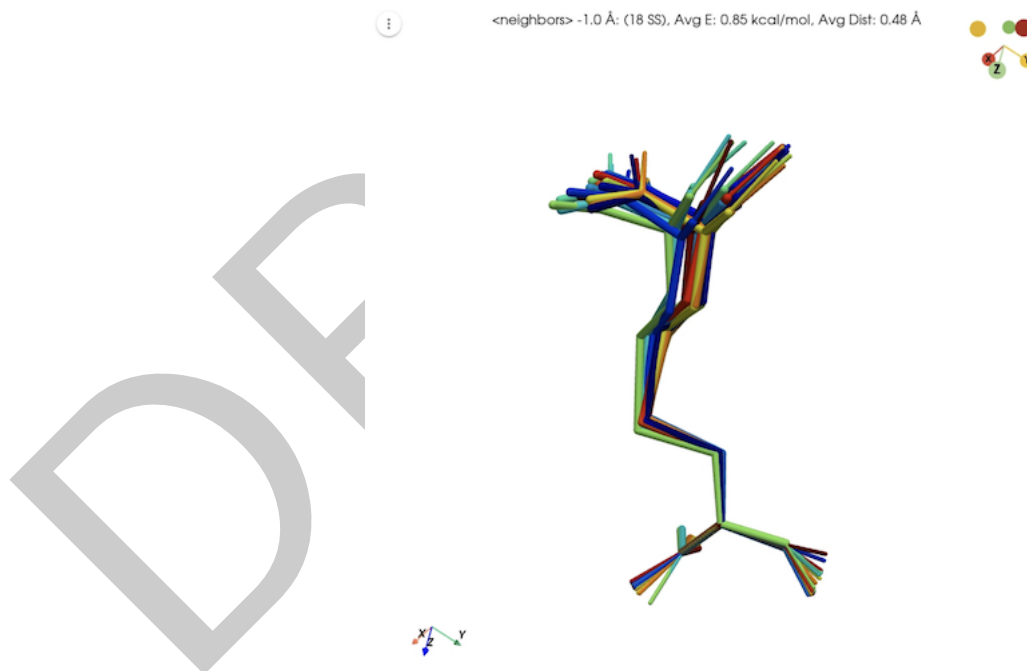


Figure 3: Low energy neighbors

233 Analyzing Disulfide Structural Class Distributions

234 The package includes the [DisulfideClassConstructor](#) class, which is used to create and manage
 235 Disulfide binary and sextant classes. A note about these structural classes is in order. ([Schmidt,
 236 2006](#)) described a method of characterizing disulfide structures by describing each individual
 237 dihedral angle as either + or - based on its sign. This yields 2^5 or 32 possible classes. The
 238 author was then able to classify protein functional families into one of 20 remaining structural

239 classes. Since the binary approach is very coarse and computers are much more capable than
240 in 2006 I extended this formalism to a *Sextant* approach. In other words, I created *six* possible
241 classes for each dihedral angle by dividing it into 60 degree segments. This yields a possible
242 6^5 or 7,776 possible classes. The notebook [DisulfideClassesPlayground.ipynb](#) contains some
243 initial results. This work is ongoing.

244 Summary

245 **proteusPy** is a python-based package capable of visualization and analysis of over 120,000
246 Disulfide bonds contained in the RCSB structural database. This work provides a strong
247 foundation to not only analyze these important structural elements but also provides flexible
248 tools for modeling proteins from dihedral angle input.

249 Appendix

250 Database Creation Workflow

251 The following steps were performed to create the RCSB disulfide database:

- 252 1. Identify disulfide containing proteins in the [RCSB](#): I generated a query using the web-
253 based query tool for all proteins containing one or more disulfide bond. The resulting
254 file consisted of 35,819 IDs. The file containing these is: [ss_ids.txt](#).
- 255 2. Download the structure files to disk. This resulted in the program [DisulfideDownloader.py](#).
256 The download took approximately twelve hours.
- 257 3. Extract the disulfides from the downloaded structures and build the **DisulfideLoader**
258 object. The program [DisulfideExtractor.py](#) was created and used to do this against the
259 individual structure files. This seemingly simple task was complicated by several factors
260 including:
 - 261 1. The PDB file parser contained in Bio.PDB described in ([Hamelryck & Manderick,](#)
262 [2003](#)) lacked the ability to parse the **SSBOND** records in PDB files. As a result I
263 forked the Biopython repository and updated the `parse_pdb_header.py` file. My
264 fork is available at: <https://github.com/suchanek/biopython>
 - 265 2. Duplicate disulfides contained within a multi-chain protein file.
 - 266 3. Physically impossible disulfides, where the $C_\alpha - C_\alpha$ distance is $> 8 \text{ \AA}$.
 - 267 4. Structures with disordered CYS atoms.

268 The disulfide extraction process is time consuming, and is only needed if the underlying
269 **Disulfide** class is changed.

270 I ultimately elected to only use a single example of a given disulfide from a multi-chain entry,
271 and removed any disulfides with a $C_\alpha - C_\alpha$ distance $> 8 \text{ \AA}$. This resulted in the current
272 database consisting of 35,808 structures and 120,494 disulfide bonds. While there are many
273 structure visualization and analysis packages available (PyMol, Chimera, RCSB) this is the
274 only centralized, locally available Disulfide database available.

275 The Future

- 276 ■ I am writing up the first analysis paper which will provide an overall survey of the RCSB
277 disulfide database in terms of structural statistics. This represents the outcome from my
278 initial desire for building the system in the first place.
- 279 ■ I am exploring disulfide structural classes using the sextant class approach as time permits.
280 This offers much higher class resolution than the binary approach and reveals subgroups

281 within the binary structural classes. I'd also like to explore the catalytic and allosteric
282 classes within the subgroups to look for common structural features at a higher level.

- 283 ■ I am working to deploy a Disulfide Database browser for further exploration and analysis.
284 There are several iterations of the viewer in the **programs** directory. The issue is I am
285 unable to refresh a **panel pyvista.plotter** object correctly into a single pane.

286 Miscellaneous

287 Performance

- 288 ■ Manipulating and searching through long lists of disulfides can take time. I've added
289 progress bars for many of these operations.
290 ■ Rendering many disulfides in **pyvista** can also take time to load and may be slow to
291 display in real time, depending on your hardware. I added optimization to reduce cylinder
292 complexity as a function of total cylinders rendered, but it can still be less than perfect.
293 The faster your GPU the better!

294 Visualizing Disulfides with pyVista

295 PyVista is an excellent 3D visualization framework and I've used it for the Disulfide visualization
296 engine. It uses the VTK library on the back end and provides high-level access to 3d rendering.
297 The menu strip provided in the Disulfide visualization windows allows the user to turn borders,
298 rulers, bounding boxes on and off and reset the orientations. Please try them out! There is
299 also a button for *local vs server* rendering. *Local* rendering is usually much smoother. To
300 manipulate: - Click and drag your mouse to rotate - Use the mouse wheel to zoom (3 finger
301 zoom on trackpad)

302 Developer's Notes:

303 The .pkl files needed to instantiate this class and save it into its final .pkl file are defined in
304 the [proteusPy.data](#) class and should not be changed. Upon initialization the class will load
305 them and initialize itself.

306 *NB:* disulfide database creation relies on my [fork](#) of the [Biopython](#) Python package to download
307 and build the database, (<https://github.com/suchanek/biopython>). This fork is installed
308 automatically.

309 Contributing/Reporting

310 I welcome anyone interested in collaborating on proteusPy! Feel free to contact me
311 at suchanek@mac.com, fork the [repo](#) and get coding. Issues can be reported to
312 <https://github.com/suchanek/proteusPy/issues>.

313 Bibliography

- 314 Abelson, H., & DiSessa, A. A. (1986). *Turtle geometry: The computer as a medium for*
315 *exploring mathematics*. MIT Press. <https://doi.org/10.7551/mitpress/6933.001.0001>
- 316 Hamelyrck, T., & Manderick, B. (2003). PDB file parser and structure class implemented
317 in python. *Bioinformatics*, 19(17), 2308–2310. [https://doi.org/10.1093/bioinformatics/](https://doi.org/10.1093/bioinformatics/btg299)
318 [btg299](https://doi.org/10.1093/bioinformatics/btg299)
- 319 Pabo, C. O., & Suchanek, E. G. (1986). Computer-aided model-building strategies for protein
320 design. *Biochemistry*, 25(20), 5987–5991. <https://doi.org/10.1021/bi00368a023>

- 321 Schmidt, B. (2006). Multiple disulfide-bonded states of native proteins: Estimate of number
322 using probabilities of disulfide bond formation. *Biochemistry*, 45(24), 7429–74334. <https://doi.org/10.1021/bi0603064>
323
- 324 Wong, J. H., & Hogg, P. J. (2010). Analysis of disulfide bonds in protein structures. *Journal*
325 *of Thrombosis and Haemostasis*, 8(10), 2345. [https://doi.org/10.1111/j.1538-7836.2010.](https://doi.org/10.1111/j.1538-7836.2010.03894.x)
326 [03894.x](https://doi.org/10.1111/j.1538-7836.2010.03894.x)

DRAFT