For your semester project, choose a structural modeling or design problem of your interest. You may pull ideas from your own research, the current literature, or other sources. Your project should allow you to apply your modeling abilities to realistic systems and/or to push your modeling skills beyond the standard applications we covered in class by combining protocols or creating new protocols. You are encouraged to visit and email Prof. Gray and Shourya as you formulate your ideas to identify the appropriate level of complexity. You may work in teams of two people.

The **project proposal** should contain (limit one page):

1. A brief problem statement describing your goal.
2. A statement of why your method or model would be important.
3. A brief outline of your computational strategy.
4. A list of any PyRosetta capabilities that you would need to complete your project that we have not covered in class (e.g., the use of nucleic acids, different solvent conditions, etc.)

The **final report** should be written like a research paper with the sections Background, Methods, Results, Discussion, and Conclusion. Furthermore, it should contain:

1. A clear statement of the problem you are solving.
2. A detailed description of the approach you used to solve the problem. A flowchart might be helpful. Share your full PyRosetta program via GitHub; include the URL in your report.
3. A description of your results, including quantitative data and figures showing the important portions of the model you created.
4. Limitations and caveats of your model, i.e., likely sources of error.
5. A discussion of how you could test your model/design experimentally.
6. Appropriate citations including full author lists and article titles.

The final report should be limited to 3-5 single-spaced pages for those enrolled in 540.414 and 5-7 for those enrolled in 540.614. The page limit includes figures and tables, but not the references.

You will also present your project to the class in a 15-minute presentation with appropriate media. Your **presentation** should include the same sections as listed above for the report.

**Timeline:**

- Nov 3 before class: Project topic due by email
- Nov 10: Project proposal due
- Week of Nov 17: Meet with Dr. Gray and/or Shourya
- Dec. 3: Student Presentations and Final Report due
**Project Grading Scheme**

Technical (60%)
- Problem statement is clear
- Background and impact explained clearly
- Algorithm is appropriate for the task and correct
- Code is easy to follow, commented, and follows guidelines
- Model is of sufficient quality
- Model is validated / limitations are identified
- Data are analyzed appropriately
- Weaknesses in data are identified
- Conclusions are appropriate for the data
- Modeling is tied to existing experimental data and new experiments

Report (20%)
- Appropriate figures and tables
- Figures and tables present data clearly
- Appropriate length
- Logical organization
- Prior work is cited appropriately

Presentation (20%)
- Clear presentation
- Appropriate use of visuals
- Stay within time limit

Other
- Team members share work appropriately

Subcategories are weighted evenly in each category. Teams will be given peer-evaluation forms to help split credit between team members.
You are welcome to choose your own project idea. New ideas, especially those drawn from research labs, are highly encouraged. The suggestions below would be suitable choices, and they demonstrate the scope appropriate for this assignment.

1. Create a **homology model** of a protein of interest. Use this model in a **design or docking application** or to analyze properties of the protein such as stability.

2. **Antibody design for Ebola.** Determine which proteins of the Ebola virus would constitute good targets for steric blocking of their function. Design an antibody to bind an epitope on Ebola as a potential therapeutic. Alternately, you could use a different scaffold protein, or focus your design toward diagnostics instead of therapeutics. (A therapeutic is meant to be safely injected into the patient to stop the disease, whereas a therapeutic would be part of a device and so would need greater emphasis on stability and reliability and less on biocompatibility.)

3. **Calcification disorder complexes.**
   
   **Background:** Calcification disorders are debilitating diseases that have a major impact on the affected individuals leading to chronic pain and blindness. They are also potentially fatal due to heart attack and stroke, even in infants under 6 months of age. Mutations in four genes (ABCC6, ENPP1, TNAP, N5TE) lead to the calcification disorders PseudoXanthoma Elasticum (PXE), Generalized Arterial Calcification of Infancy (GACI), Hypophosphatasia and Arterial Calcification with Deficiency of CD73 (ACDC).
   
   **Goals and methods:** Use known structures of homologous proteins to create a homology model in the membrane bilayer of one of the proteins ABCC6, ENPP1, TNAP, or T5NE. Display the structure with respect to the membrane and model the trans-membrane domain in these structures. Use secondary structure information and multiple-sequence alignments where necessary.

4. Examine the effect of a **post-translational modification** of an amino acid (phosphoserine, phosphotyrosine, methylhistidine, glycosylated residues, etc.) on a particular protein.

5. Examine algorithmic variations of the **cyclic coordinate descent** loop building algorithm. Does it matter if the loop cutpoint is at the N or C end of the loop? What if it is in the middle? Does the order matter when iterating over the backbone torsional degrees of freedom? What if the torsion angle in each cycle is chosen randomly instead of sequentially? The goal of this project is to examine several variations, benchmark them rigorously, and find a more efficient/accurate algorithm.

6. Use RosettaDesign in conjunction with ligand or protein docking to examine **sequence recovery at protein interfaces.** How well are sequences recovered at interfaces? If you generate an ensemble of sequences, are similar sequence conservation patterns observed as in nature? Can we use sequence conservation patterns of putative protein interfaces to distinguish between near-native and non-native interfaces?
7. **Structure of poly-ubiquitin.** Ubiquitin is a protein which is covalently attached to proteins in the cell, typically to target them for destruction by the proteome, but also for various other regulatory processes. Ubiquitin is attached by connecting its C-terminal residue to the side chain of a lysine, and ubiquitins can be attached to other ubiquitins forming a poly-ubiquitin chain. The project would be to try different linkages (via combinations or repetitions of the seven lysines on the surface of a single Ub) to see what kind of structures are possible. Are any combinations forbidden due to steric clashes? Are certain conformations more flexible due to more space around the C-terminal linking residues? How do the poly-Ub structures possible compare with those which are known to be used in biology?

8. **Model of intrinsically disordered protein in cell signaling.**

*Background:* Intrinsically disordered segments in proteins are thought to be present in 33-50% of eukaryotic proteins. Due to their inherent flexibility, they play an important role in binding other proteins and surfaces. These segments are often high in charge and low on hydrophobic residues, which allow them to bind strongly and specifically to their target. Xue *et al.* (Biochemistry, 2014, doi: 10.1021/bi500904f) have identified a model of a small, artificial disordered segment binding to a larger protein domain based on NMR constraints.

*Goals:* Dock an intrinsically disordered segment to a larger protein domain and then analyze:

- a) which terms are the largest contributors to the interface score of the proteins and
- b) whether the protein loses its inherent flexibility on binding, as measured by the variability in the low-energy structures.

9. **Generate structural models of bone-binding peptides to explain experimental binding**

*Background:* Hard tissues such as bone are formed by inorganic crystals such as calcium phosphate salts. Tissue formation is guided by various mineralization proteins, but little is understood about how these proteins are able to nucleate crystals, alter crystal morphology, or inhibit growth. Structural models will provide insight into these processes. Gungormus *et al.* (Biomaterials 2008) have identified two hydroxyapatite binding peptides of different levels of affinity; these can serve as a test system.

*Goals:* Design an algorithm to dock the two Gungormus peptides to the hydroxyapatite mineral surface. Identify structural features and explain differing binding affinity. For a control calculation, dock the peptides to a calcite surface.

*Methods:* Create a flexible-backbone docking algorithm for binding a mineral surface. Test alternate scoring functions and flexible docking algorithms.

10. **Identify recognition motifs in linear peptides in enzyme-substrate interactions**

*Background:* Substrate specificity is a key feature of enzymes that determines both their activity and function within the cell. A large number of enzymes interact with substrate proteins through a linear peptide of a specific sequence (e.g., protein kinase A). These recognition motifs are typically identified through sequence-based methods, and identifying the structural mechanisms that underlie this recognition has implications for drug design.

*Goals:* Design an algorithm that accurately distinguishes binders from non-binders and identify which peptide sequence features in the substrate are most responsible for binding.

*Methods:* For a specific protein-substrate complex and sets of known binding and non-binding peptide sequences, use sequence threading and/or docking and scoring to distinguish binders from non-binders. Identify sequence motifs that are responsible for binding activity through statistics or observation.