Please choose one of the following two for a final project:

1. **Research project:** Write a brief paper (two to four pages in Physical Review format) about a topic of your choice at the interface of Statistical Physics and Biology. It should be formatted as a regular article with title, abstract, and bibliography. The main text should contain introductory and concluding paragraphs (whether or not they appear as subsections is not important). The ideal project will involve a combination of literature review, discussion of an analytical or computational model, and application/analysis of biological data.

   Students can collaborate in groups provided that the respective contributions of the author of the joint paper is clearly specified in a footnote. (The length of the paper may be proportionately longer in such collaborations.) Clearly the initial hurdle is coming up with an interesting project that is doable in a short time. We would thus like you to think about potential projects, and consult with the course staff about the choice of topic (preferably as soon as possible, but no later than beginning of April).

   You can see examples of such Final projects from previous years on the web at: https://web.mit.edu/8.592/www/grades/projects/ . Some potential projects, clearly in need of refinement and elaboration, are suggested in the following pages.

2. **Teaching site:** Design a web-site that can be used to teach a topic at the interface of Statistical Physics and Biology to non-specialists. For example, imagine that a high school teacher would use a one hour class to teach the material to an honor science class the material using your web-page. As such, you should include introductory materials, references that interested students can pursue on their own. The presentation must also be colorful and dynamic (e.g. by including figures, animations, applets, etc.) to engage and maintain the interests of a diverse non-specialist audience.

   The rules for collaborations, as well as timeline, are the same as for a Research Projects. You can see examples of such Final projects prepared before, on the same website: https://web.mit.edu/8.592/www/grades/projects/ .
1. **Epistatic interactions between mutations:** In class we described a model for the dynamics of a single mutation (or independently evolving mutations). Generalize the description to two mutations in which the fitness of the double mutant is not simply the sum of the fitnesses for single mutants.

2. **‘Finite temperature’ alignments:** The standard tests of sequence similarity (such as BLAST) result in a score which can be interpreted as the lowest energy of a directed polymer. One can use the transfer matrix to evaluate the partition function of the corresponding directed polymer at any temperature. The free energy would then converge to the standard similarity score in the limit of zero temperature. Is it possible to get a better sense of similarity of sequences by using the finite temperature free energy? One could test this by evaluating the free energies for alignment of proteins that are known to have structural similarity, but no obvious sequence similarity.

3. **Statistics of random alignments:** Consider proteins of various lengths. For each protein generate very many randomized sequences by permuting the proteins, also including some fraction of insertions and deletions. Implement and perform global gapped sequence alignments between the randomized sequences. For each length plot the distribution of scores of alignments, and compare to the Tracy–Widom distribution.

4. **Role of specific interactions in protein folding:** It is widely believed that all native interactions are important to fold a protein, but it may turn out that only few are sufficient to fold it. The aim is to test plausibility of folding a protein with minimal number of interactions, and to estimate the minimal number of interactions needed.

   Simulate folding of a 27-mer or 36-mer on a cubic lattice. First, use a Go-model (i.e. all native interactions are attractive and non-natives are zeros) and measure MFPT (mean folding time). Next, make all non-native interactions the same and weakly attractive (hydrophobic attraction) $B_{nn} = -0.1$. Make most of native interactions the same $B_n = -0.1$, while keeping very few (3–10) native interactions strongly attractive $B_n = -1$. Can you fold this protein using only few native interactions? What’s the minimal number of interactions required to fold a 27-mer? 36-mer? Are there any specific preferred location for these key interactions in the “protein” structure? Can you find a rational for the number and location of key interactions based on the native structure and other compact structures of the 27-mer?

5. **Topological state of a polymer:** The goal is to examine connection between topology of a polymer chain and its spatial organization. To this end, you can generate thousands of conformations of a Hamiltonian ring (closed space-filling polymer ring), characterize a topological state of each (i.e. knotted or unknot ted) by the value of Alexander polynomial, and compare statistical properties of knotted and unknot ted conformations, e.g. sizes of subchains. Our hypothesis is that unknot ted configurations are more likely
to have more compact subchains. This can be connected to the spatial organization of DNA in human cells. A potentially useful algorithm for generating configurations of Hamiltonian walks can be found in the following references:
http://prl.aps.org/pdf/PRL/v100/i11/e118102

6. **The fractal (crumpled) globule; 1:** The goal is to explore properties of the fractal globule, a non-equilibrium polymer state emerging as a result of polymer collapse. The fractal globule is an unknotted compact polymer state that is consistent with 3D organization of DNA in a human cell. While computer modeling provided structures of the fractal globules, little is understood about intrinsic organization of this unknotted polymer conformation. Our hypothesis is that the fractal globule can be approximated by a polymer folded without pseudoknots (like RNA secondary structure), which is then “crumpled” to form a 3D structure. This hypothesis can be tested by trying to approximate given fractal globule structures by non-pseudoknotted conformations, i.e. developing an algorithm to find a minimal set of interactions to be removed such that remaining globule becomes a planer (“RNA-like”) conformation. Statistical properties of obtained conformations can then be studied.

7. **The fractal (crumpled) globule; mechanical properties:** Unknotted topology of a polymer in the fractal globule can affect its mechanical properties. Here you can use computer simulations (existing code and set-up) to study mechanical properties of the fractal globule. E.g. subject it to compression between two planes and obtain force-compression curve. Equilibrium globules are highly knotted and can be used for comparison.

8. **Role of ‘topology’ of the native structure in protein folding:** Simulate folding of a 27-mer on a cubic lattice. Use Go-model (i.e. all native interactions are attractive and non-natives are zeros) and measure MFPT (mean first passage time, i.e. mean folding time). Try folding 27-mer “proteins” with different native structure. Try some of possible native structures and find structures that fold fast and those that fold slowly. How significant is the difference in MFPT? Can you explain why some structures fold fast and others fold slowly? Study correlation between MFPT and the number of local (e.g. i, i + 3 or i, i + 5) interactions. Study correlation between MFPT and the number of native-like folds in the space of structures. Simulate folding of a 27-mer on a cubic lattice (using “designed sequences” or “Go-model”) and study how folding time depends on the statistical properties of the native structure. Our hypothesis is that more “knotted” structures make folding slower, while folding is faster into structures that resemble the fractal globule.

9. **Correlated mutations:** While evolving, proteins sequences accumulate mutations. Effect of one mutations can be compensated by another mutation. This leads to correlations between mutations. Revealing and understanding patterns of correlation can help to predict protein structure, as amino acids that interact are more likely to exhibit
correlated mutations. The system is analogous to a spin system at low temperature, where spins become correlated. Inferring the coupling constants of the spin system is then analogous to the inference of correlated mutations. Consider this “inverse statistical mechanics” problem, where the goal is to reconstruct the couplings of the spin system, given a series of observations of the states of spins. For some recent articles relevant to this problem, see

10. Protein folding and evolution: “Foldable” proteins constitute a tiny fraction of all possible protein sequences. Little is know about their organization in the sequence space. Understanding their distribution in the sequence space can help to answer several important questions. For example, what is the minimal number of mutations in a random (non-foldable) protein that can make it foldable? What is the minimal number of mutations that can make a foldable protein, fold into a different structure? Using available code for folding lattice proteins one can try to answer these questions.

11. Symmetry of protein structures: Protein structures are known to possess many interesting symmetries (two/four/eight-fold). Most of these were observed by a naked eye. We suggest systematically characterizing symmetry in natural protein structures. Recently developed algorithms in computer graphics allow to detect approximate symmetry in a cloud of points and can be applied to protein structures. This approach can also help to understand how such symmetric structures emerged in evolution. E.g. comparison of sequences of symmetric units within a single protein can help to test whether such units emerged by a series of sequence duplications.

12. Active matter: The interior of cells is crowded with many different biological molecules. Some of these molecules are consuming energy to actively self-propel (for example molecular motors moving on cytoskeletal filaments) while others are simply diffusing due to thermal motion. However, it has been shown recently, both analytically and in numerical simulations, that mixtures of active and passive particles tend to segregate. Could this effect be relevant inside cells or other biological systems? There are a variety of projects pertaining ‘Active matter,’ which can be discussed with Dr. Alex Solon.

13. Information theory of immune system The goal of the immune system is to recognize foreign proteins from the host ones. Such recognition is accomplished by receptor (MHC) molecules that recognize short peptides (of 8-10 amino acids) by binding them. Receptors that bind host peptides are eliminated during the training process. Each receptor can be described by the consensus sequence that it binds the strongest and by its specificity, i.e. the number of changes in the consensus sequence that it tolerates. Consider a repertoire of $N$ receptors: what is the optimal length of the peptide and the optimal specificity, which provide the best coverage of all possible peptide sequences, while minimizing the probability of an erroneous recognizing a host protein?
14. **Self-assembly of the spindle** In cell biology, the spindle is the structure that separates the chromosomes into the daughter cells during cell division. The spindle is made of chromosomes and microtubules. Recent data suggest that the spindle consists of thousands of short microtubules. Neighboring microtubules tend to align with each other, while the density of microtubules decays with the distance from the chromosome. Such system can be modeled by a modified XY model developed in statistical mechanics for spin systems. Each spin corresponds to a microtubule, the coupling favors alignment of microtubules and the external field models the effect of chromosomes. Can such modified XY model explain the structure and the self-organization of the spindle?

15. **‘Hubs’ and diameter of a network:** The ‘diameter’ of a network is a global measure of the effectiveness of a network, and changes if nodes are removed from the network. It was proposed that the change in the network diameter depends on the type of the network (i.e. degree distribution). Consider two types of networks: a random one and a network with a power-law degree distribution, that resembles some biological networks and Internet. How sensitive is the diameter to removal of nodes in these two classes of networks? You can try different strategies of removal: picking a node for removal randomly (uniformly), or picking more connected node (hubs) with a higher probability.

*****