BMED 3600 Data Skills Exercise: Protein Folding Prediction

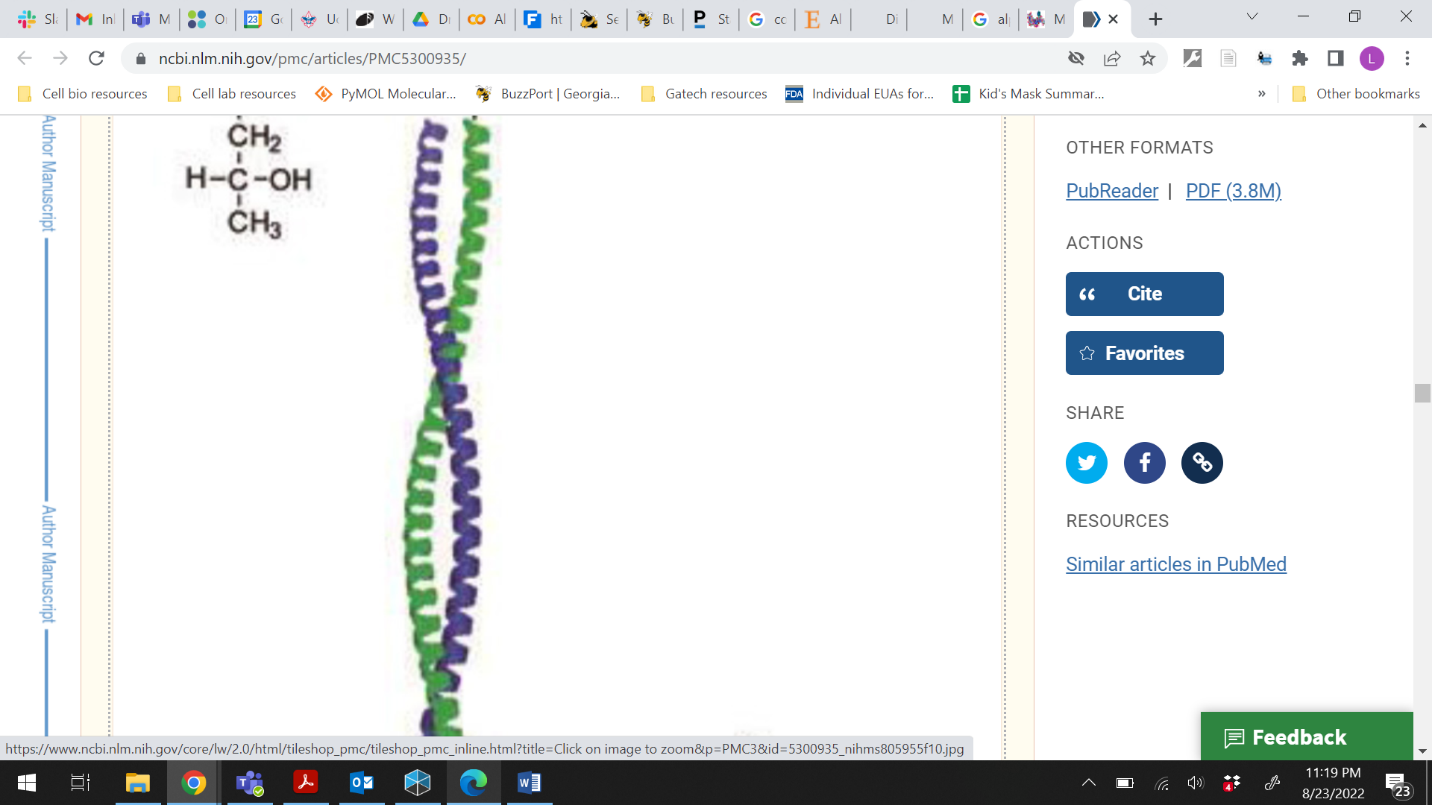
Spring 2025

After completing this exercise, you should be able to:

1. Describe different levels of protein structure
2. Describe experimental and computational ways of predicting protein structure, and their limitations
3. Use AlphaFold to make protein structure predictions using a primary protein sequence
4. Interpret the results of an AlphaFold prediction
5. Describe potential limitations that AlphaFold may have when making structure predictions

You will be asked a series of questions, some of which have more “obvious” answers and some of which you will need to do some critical thinking for. Do NOT divide up the questions between group members, they are meant to be done sequentially. You may use class resources but do not ask the internet for hints on how to answer questions – this totally defeats the purpose of the exercise!

1. You work for a biotech company whose goal is to design peptides or peptide polymers as drug carriers. Many drugs have properties that make a carrier peptide necessary for drug delivery. What does “polymer” mean in this context? Why might encapsulating a drug in one of these polymers help with drug delivery?



1. To start on your project, you decide to make a peptide polymer that will use a “coiled-coil” of two alpha helices to attach one peptide to another (see image at left). Look at a table of the different amino acids and consider how the side chains are arranged in an alpha helix. Which amino acids from bonds/interactions would be useful to hold the two helices together? Which would not? Write some of each down.

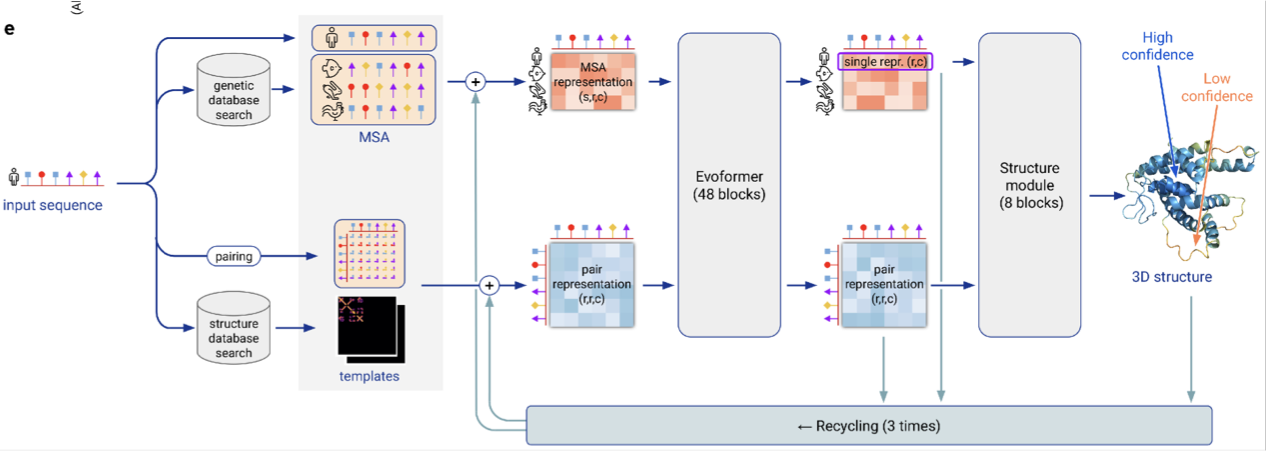
From Wu Y, Collier JH. α-Helical coiled-coil peptide materials for biomedical applications. Wiley Interdiscip Rev Nanomed Nanobiotechnol. 2017 Mar;9(2):10.1002/wnan.1424. doi: 10.1002/wnan.1424

1. You decide that your alpha helices should be 14 amino acids long (schematic below). If this alpha helix peptide is going to be part of a 2-alpha-helix coiled coil dimer, which amino acid side chain positions should have the specific characteristics you identified above?

1---2---3---4---5---6---7---8---9---10---11---12---13---14

Before ~2021, peptide and protein structure determination could be accurately accomplished by calculating energy minimization for a peptide chain (~days, not very reliable), or it could be determined experimentally by techniques such as X-ray crystallography and others (~years, requires protein to crystallize – we only got about 7% coverage of all known proteins in the 70 years since X-ray crystallography was first used with proteins!)

1. In 2021, a structure prediction tool that was very accurate when tested against a known set of structures (AlphaFold) was published. To predict how your peptide (from #3) will look, go to: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb> (or search for “colab fold” and click the first link) and copy-paste in the single letter amino acid sequence of your 14 amino acid-long peptide (from #3) in the “**query\_sequence”** section. Name the protein you are running in the “jobname” field.
   1. Click on Runtime tab at the top and select Run All. You can click “Run Anyway” when a warning pops up, this is Google’s warning that they did not generate the code you are executing. The step that is currently running has a circle with a stop sign/square in it. A green checkmark appears when a certain step is done. You DO NOT need to download the file that pops up.
   2. **While you are waiting** for the run to finish, please access the RCSB Protein Data Bank at the following url: <https://www.rcsb.org/>. Using the search box, search for “1IP1” (one, letter I, letter P, one), the G37A Human Lysozyme. Which of the structure prediction methods mentioned above was used to generate the structure for this file?
   3. Click the “Structure” tab above the picture of the lysozyme. Move the structure around with your mouse to investigate the protein from different angles at different levels of magnification. What is the number of alpha helices in the protein? Beta sheets?
   4. When your prediction from “a” is finished, scroll down to the Display 3D Structure box to check your structure. Did you make an alpha helix with the appropriate chemistry? If not, change something and re-run the structure prediction. When you have a structure that you like, copy and paste your single-letter amino acid sequence into the class spreadsheet (<https://gtvault.sharepoint.com/:x:/s/DataskillsinUndergradCurriculum/Ef2Ge1ibnGFLrfk1-QWEoX8B5qc60Kr2MyIo7m75sxkdBw?e=vLVokB>)
2. After running some lab experiments, you decide that this drug delivery approach isn’t feasible. Instead, you want to do some work with Intrinsically Disordered Protein (IDP)-based peptides. Your mentor gives you a specific protein that contains one of these regions and asks you to research it. Please copy and paste the sequence assigned to your team in the class spreadsheet (the sequences are at the bottom of this document) into the AlphaFold query (make sure to erase your other sequence first!), and click “run all” as before. This will take about 15 minutes, so while you wait, go to the next question.
3. When AlphaFold is told to predict a structure, it goes through several steps as in the image below.



Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

* 1. One of the first steps is a genetic database search and multiple sequence alignment (MSA). Another first step is to search through the Protein Data Bank (where you looked at the lysozyme structure) to compare the protein to existing/known 3D protein structures. What do you think an MSA is and why would it be useful in structure prediction?
  2. Next, the outputs are used as inputs for “Evoformer”, a neural network that begins to make structure predictions. The outputs are then used as inputs for the “Structure module”, which optimizes atom position based on energy minimization. The results are fed back into the Evoformer several times before the final structure prediction is reported. What energy would be “minimized” during protein folding/why do proteins fold? Why is it useful to perform this “recycling” step?
  3. The version of AlphaFold you are running doesn’t use all of these processes by default. Although you ignored them for now, several options are available for selection. The num\_relax option is currently set to zero. Setting this to 1 or 5 will have the effect of running an “Amber” based energy minimization on the models output from AlphaFold (see “Structure Module” in the picture above). What would be the benefit of using this? What would be a disadvantage?
  4. The template\_mode setting is “none” by default. Setting it to pdb100 allows AlphaFold to search the PDB100 database for similar structures (see “Structure Prediction Search” in the picture above. What would be the benefit of using this? What would be a disadvantage?

1. After the structure predictions are finished, answer the following questions: (saving the output files is not required)
2. How many models for the protein’s structure were predicted by AlphaFold?
3. Scroll down to the Display 3D Structure box to check the final predicted structure. What types of protein structure do you see? (primary/secondary/tertiary/quaternary)
4. The default coloring for the 3D structure is by “lDDT” (lowercase L, not capital i): predicted local distance difference test. plDDT is computed for each amino acid to estimate how well the predicted structure agrees with MSA data and PDB structure information. Values above 70 are considered high confidence whereas those scored below 50 are low confidence and hence unreliable. Does the AI tool have high or low confidence in the protein structure it produced? Are there any variations along the sequence in confidence? (It’s also shown graphically in the section below.) Enter your results into the class spreadsheet.
5. There are many proteins that have “intrinsically disordered regions”.
6. Why might a protein have disordered regions, in terms of bonds and energetics?
7. Do a Google search for your unknown protein. What does it do? Why might it be functionally useful for it to have intrinsically disordered regions? Enter this in the class spreadsheet.
8. Are there any patterns to the results in the class spreadsheet?
9. AlphaFold AI was trained on data from the Protein Data Bank (for example, like the lysozyme structure you saw earlier). Based on this fact, what do you think are limitations of AlphaFold correctly predicting certain protein structures?

Bonus, if you have time:

1. One of the most promising new types of biomaterials is polymers of intrinsically disordered proteins. Dr. Felipe Quiroz’s lab in our department studies these polymers, so if you’re interested in this, check out his work.
   1. Even though AlphaFold wasn’t very accurate at predicting the structure of protein regions containing intrinsic disorder, this doesn’t mean that IDP polymers have no structure. In fact, many of them self-assemble into orderly arrays when mixed together. What would be holding them together? Why would they do this instead of remaining single proteins?
   2. Would AlphaFold be a useful tool for predicting the assembled structure of one of these IDP polymers?
2. In your AlphaFold results, below the 3D structure, are informational plots for MSA sequence coverage (how many sequences are similar to different spans of amino acids along the sequence you submitted), and the lDDT for each amino acid. Later, if you are interested in exploring your results further, you could download the files as prompted and use the instructions found here to compare each predicted structure returned by AlphaFold: <https://bioboot.github.io/bimm143_F23/class-material/class11_alphafold.html#visualization-of-the-models-and-their-estimated-reliability>]
3. If you have extra time, take a screenshot of your first results, then choose a setting to change (from 6c) and re-run one or both structures. Are there noticeable differences in the predicted structure?

**p53**

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA

PRVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT

CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN

TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVHVCACPGR

DRRTEEENLRKKGEPHHELPPGSTKRALSNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL

KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

**Fos**

MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANFIPTVTAISTS

PDLQWLVQPALVSSVAPSQTRAPHPFGVPAPSAGAYSRAGVVKTMTGGRAQSIGRRGKVEQLSPEEEEKR

RIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDL

GFPEEMSVASLDLTGGLPEVATPESEEAFTLPLLNDPEPKPSVEPVKSISSMELKTEPFDDFLFPASSRP

SGSETARSVPDMDLSGSFYAADWEPLHSGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADS

FPSCAAAHRKGSSSNEPSSDSLSSPTLLAL

**Cadherin 1**

MGPWSRSLSALLLLLQVSSWLCQEPEPCHPGFDAESYTFTVPRRHLERGRVLGRVNFEDCTGRQRTAYFS

LDTRFKVGTDGVITVKRPLRFHNPQIHFLVYAWDSTYRKFSTKVTLNTVGHHHRPPPHQASVSGIQAELL

TFPNSSPGLRRQKRDWVIPPISCPENEKGPFPKNLVQLFSHAVSSNGNAVEDPMEILITVTDQNDNKPEF

TQEVFKGSVMEGALPGTSVMEVTATDADDDVNTYNAAIAYTILSQDPELPDKNMFTINRNTGVISVVTTG

LDRESFPTYTLVVQAADLQGEGLSTTATAVITVTDTNDNPPIFNPTTYKGQVPENEANVVITTLKVTDAD

APNTPAWEAVYTILNDDGGQFVVTTNPVNNDGILKTAKVCMVPGKMQKLASSQLFIPLSPGLGF

**BRCA1 (isoform 45)**

MGYRNRAKRLLQSEPENPSLQETSLSVQLSNLGTVRTLRTKQRIQPQKTSVYIELAACEFSETDVTNTEH

HQPSNNDLNTTEKRAAERHPEKYQGEAASGCESETSVSEDCSGLSSQSDILTTQQRDTMQHNLIKLQQEM

AELEAVLEQHGSQPSNSYPSIISDSSALEDLRNPEQSTSEKAVLTSQKSSEYPISQNPEGLSADKFEVSA

DSSTSKNKEPGVERSSPSKCPSLDDRWYMHSCSGSLQNRNYPSQEELIKVVDVEEQQLEESGPHDLTETS

YLPRQDLEGTPYLESGISLFSDDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAESAQSPAAAHTTD

TAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLITEETTHVVMKT

DAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIF

RGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNGFHAIGQMCEAP

VVTREWVLDSVALYQCQELDTYLIPQIPHSHY

**Oct4**

MAGHLASDFAFSPPPGGGGDGPGGPEPGWVDPRTWLSFQGPPGGPGIGPGVGPGSEVWGIPPCPPPYEFC

GGMAYCGPQVGVGLVPQGGLETSQPEGEAGVGVESNSDGASPEPCTVTPGAVKLEKEKLEQNPEESQDIK

ALQKELEQFAKLLKQKRITLGYTQADVGLTLGVLFGKVFSQTTICRFEALQLSFKNMCKLRPLLQKWVEE

ADNNENLQEICKAETLVQARKRKRTSIENRVRGNLENLFLQCPKPTLQQISHIAQQLGLEKDVVRVWFCN

RRQKGKRSSSDYAQREDFEAAGSPFSGGPVSFPLAPGPHFGTPGYGSPHFTALYSSVPFPEGEAFPPVSV

TTLGSPMHSN

**Fibrillin-1 (partial)**

ALKGEGWGDPCELCPTEPDEAFRQICPYGSGIIVGPDDSAVDMDECKEPDVCKHGQCINTDGSYRCECPF

GYILAGNECVDTDECSVGNPCGNGTCKNVIGGFECTCEE

**Resilin (Bombyx mori, precursor)**

MILSTFFISILAWSAIRCEPPVNSYLPPSNSANGRPSSQYGPPGSGGAPGHNQQGSGAQFSPSNQPQSSY

LPPSRSGQSSVDTQYGPPSSGGFGSGSSSQGQQLRQGSGQAKPDSQYGTPNQNGRFSSGLTGGHSQPGAS

SNSHPTSNSQGRQSFGNGGSSNGGSFGPDNQSFDQKPGSSYNPSTAGFRSSNSDSNSNGNFGSHAGGARL

TDGSQSSFSGTPSSSYGTPGFGGNGASHGNGGGSGSGFGGGNGGFGGDDNNEPAKYQFSYDVDDEQTGTK

FGHSEQRDGDLATGEYNVVLPDGRKQVVEYEAGLEGYKPQIRYEGFGESQRLSQGYTENGPNEGGFQSNN

VDSFGGSGEQTFGQFRGSNAQAGNGYPSGGQNGQNQGYPNGGSGQNGNQGFSSGNNGANRLNGGSQGYPS

GGPRGNGRGSAADPGNFSNASGRGSSRNSGGTSGSDGYPSGGPNGLRGSGY