CSCI-GA.3033-111: Protein Design HW2

Christine Wu (email:cw4459@nyu.edu)

Oct 27 2024

1 Explore the Alphafold2, Alphafold3, and Rosetta Fold colab notebooks

Solution: We have Alphafold2, Alphafold3, and Rosetta Fold colab notebooks present as followed: Alphafold2, Alphafold3, and Rosetta Fold.

2 Predict some protein structures. Choose at least 3 proteins from [Lysozyme (P00698), Cytochrome C (P99999), Myoglobin (P02144), GFP (P42212), Beta-Lactamase (P62593), Thioredoxin (P0AA25), Ribonuclease A (P61823)], and use all three methods to generate structures for them (Sequence can be obtained on uniprot).

Solution: We first picked 4 proteins from the list, which are Lysozyme (P00698), Cytochrome C (P99999), Myoglobin (P02144), GFP (P42212). Then use UniProt to find their sequences. Lastly, use the three notebooks to generate the structures. We have the results as the following:

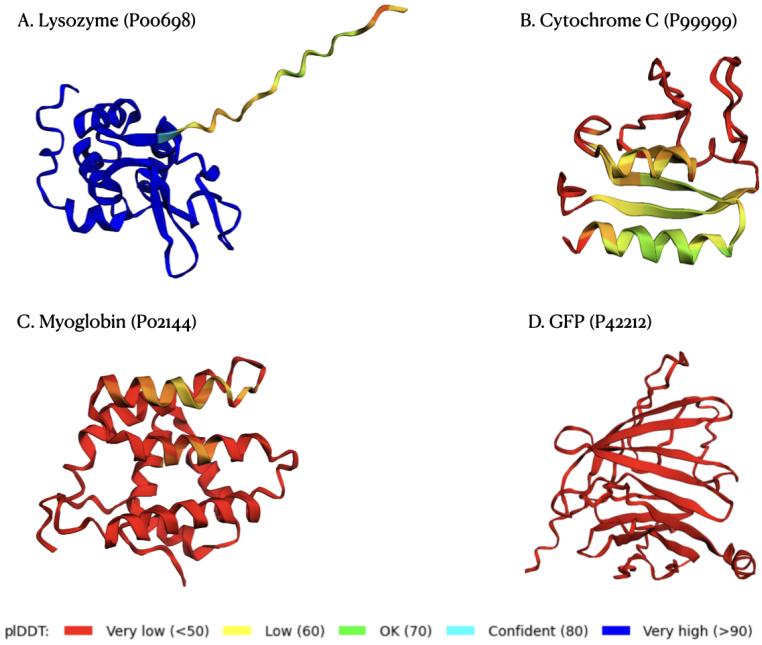

Figure 1: Alphafold2 Predictions

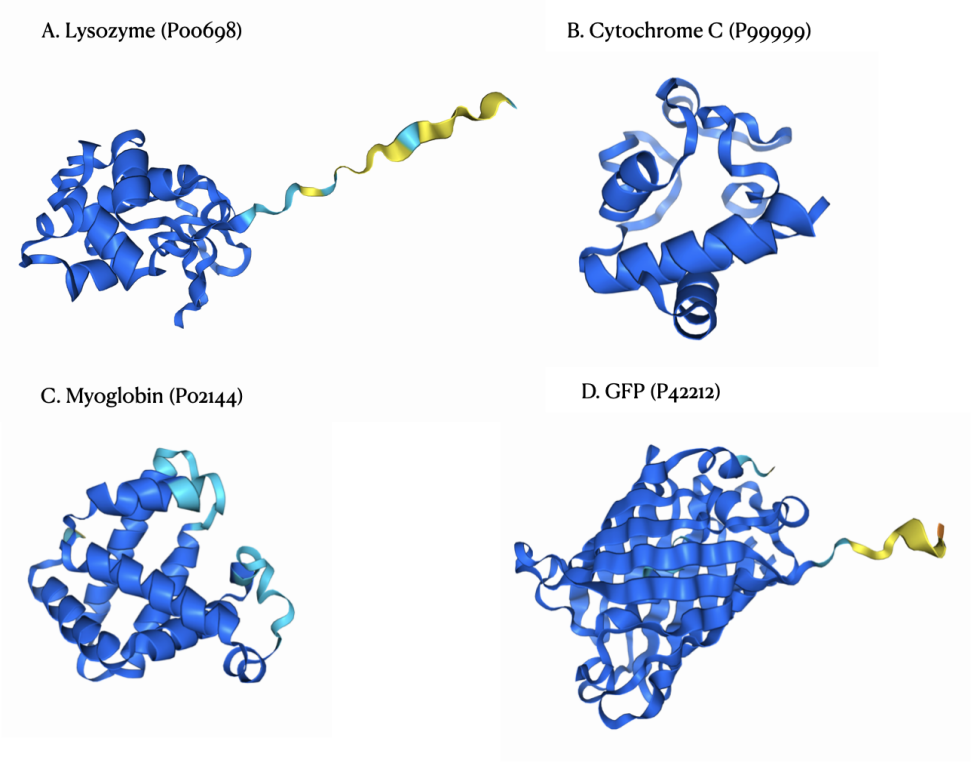

Figure 2: Alphafold3 Predictions

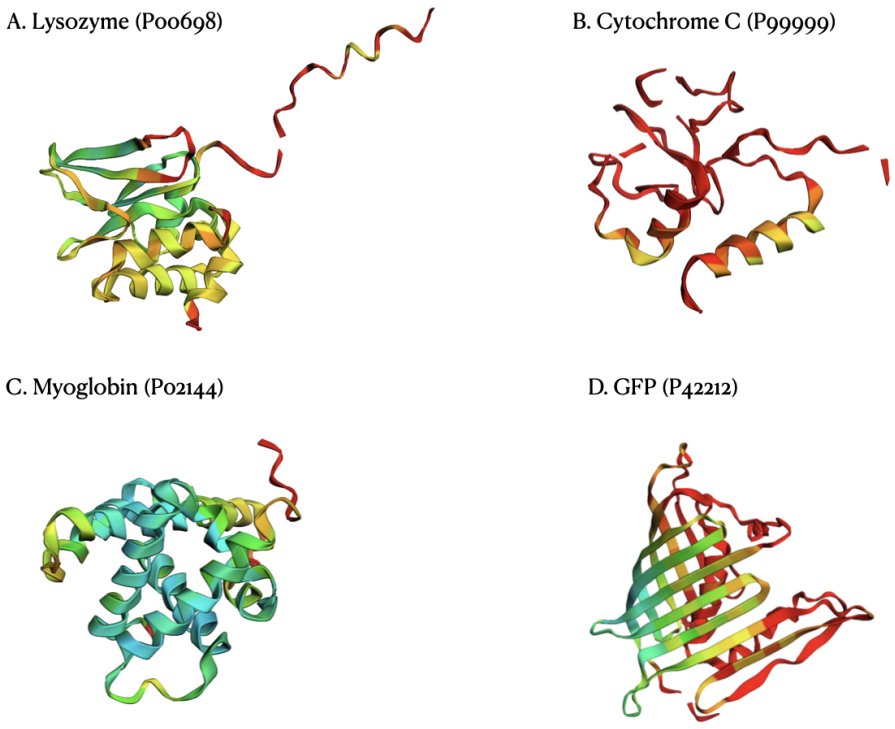

Figure 3: Rosetta Fold Predictions

The sequences we used are:

- **Lysozyme (P00698)** (Lysozyme (P00698)):
  `MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWW`
  `CNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL`
- **Cytochrome C (P99999)** (Cytochrome C (P99999)):
  `MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKM`
  `IFVGIKKKEERADLIAYLKKATNE`
- **Myoglobin (P02144)** (Myoglobin (P02144)):
  `MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKG`
  `HHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG`
- **GFP (P42212) Myoglobin (P02144)** (GFP (P42212)):
  `MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKQH`
  `DFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIK`
  `VNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK`

3 Take note of the pLDDT charts created. What type of region did each respective model predict well? Why might the models have their respective strengths and weaknesses?

Solution: We have the following pLDDT charts presented. For alphafold3, we have our own python code to generated the plots with the downloaded data by using the `fold_2024_10_27_16_37_full_data_0`, `fold_2024_10_27_16_37_full_data_1`, `fold_2024_10_27_16_37_full_data_2`,`fold_2024_10_27_16_37_full_data_3`, `fold_2024_10_27_16_37_full_data_4`. As in the webpage, there were no pLDDT plots presented. Moreover, we also have the Predicated aligned error (PAE) diagrams. The following are the results:

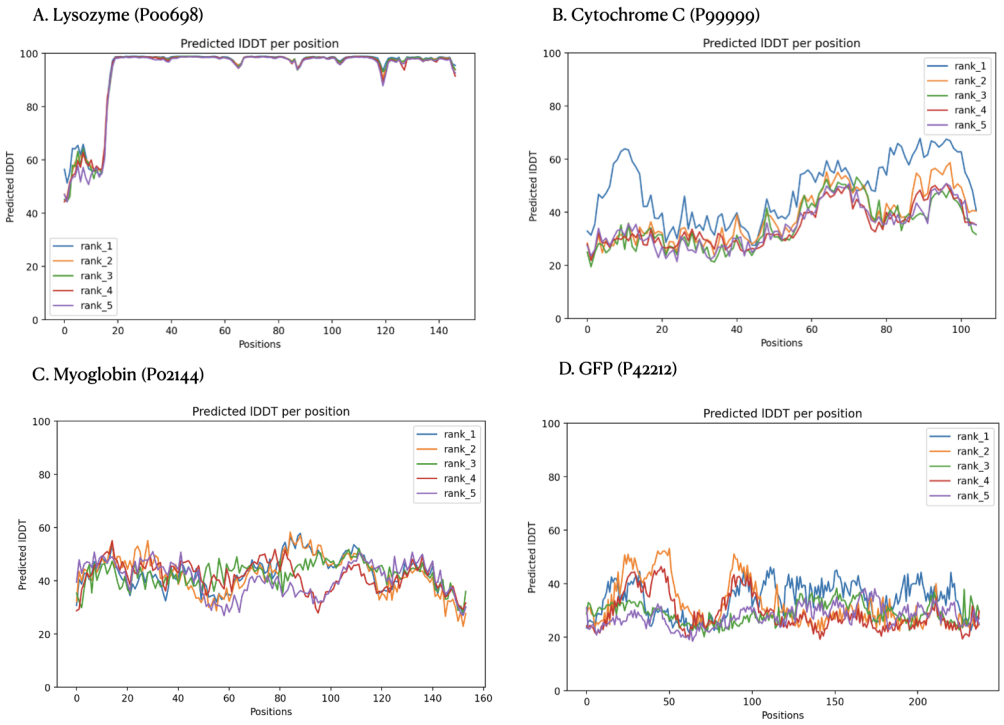

Figure 4: Alphafold2 pLDDT charts

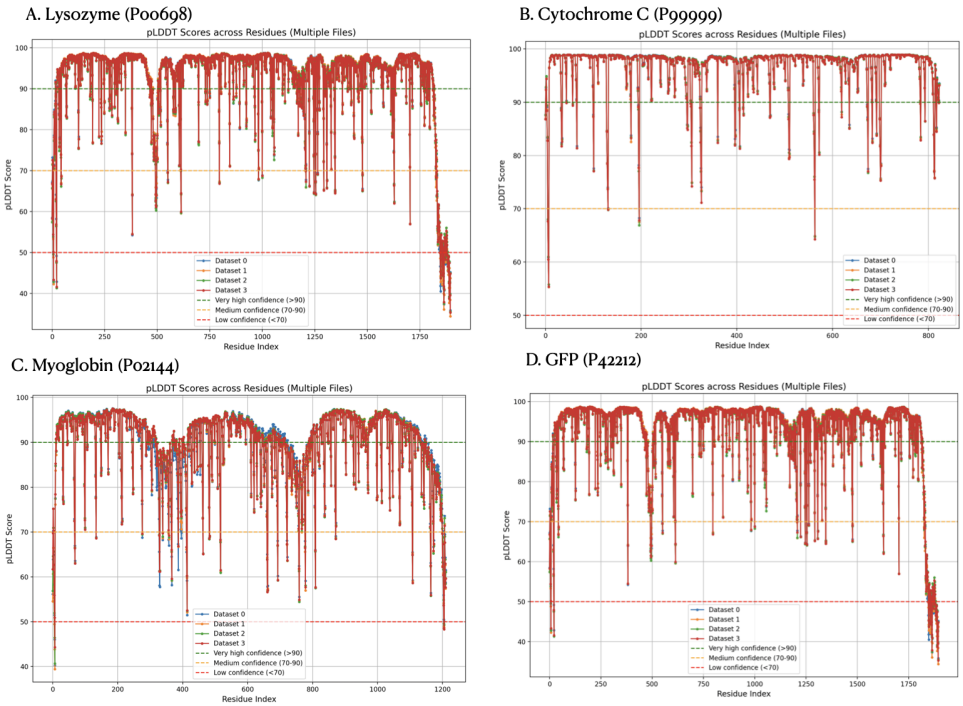

Figure 5: Alphafold3 pLDDT charts

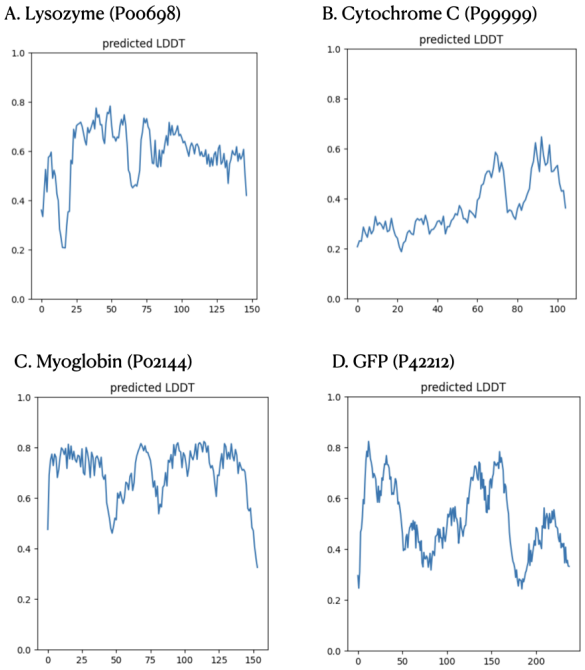

Figure 6: Rosetta Fold pLDDT charts

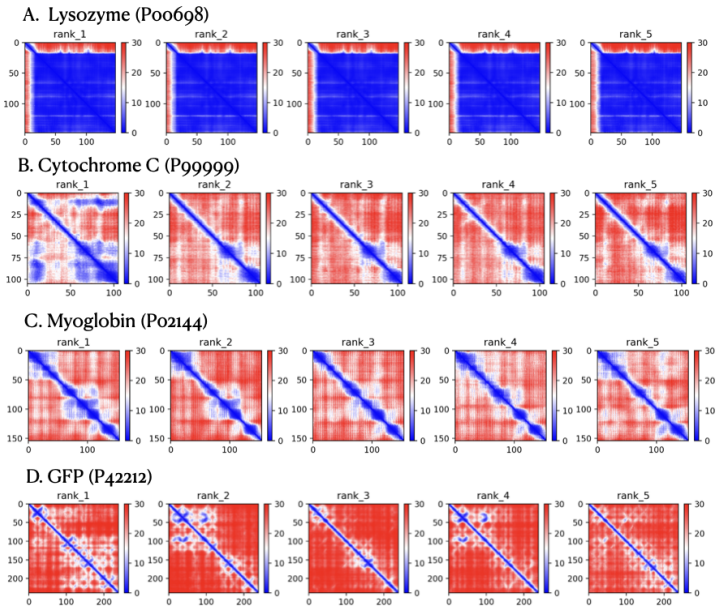

Figure 7: Alphafold2 PAE charts

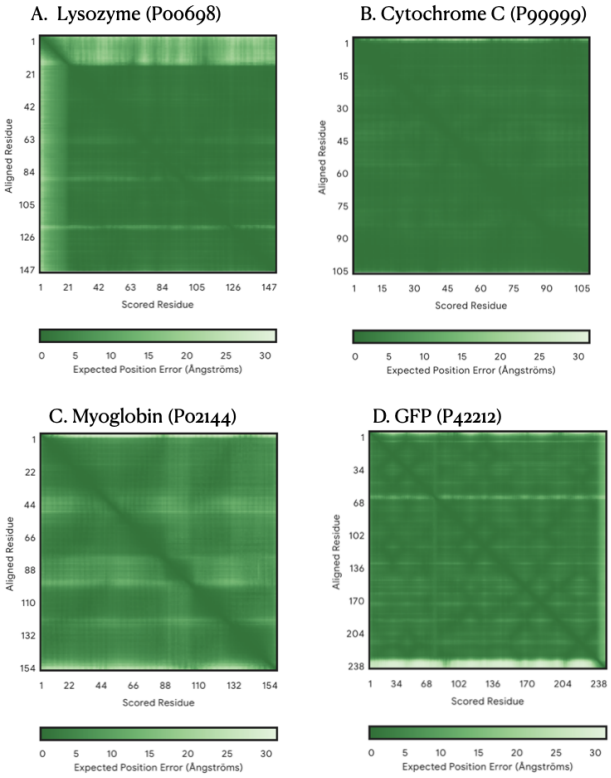

Figure 8: Alphafold3 PAE charts

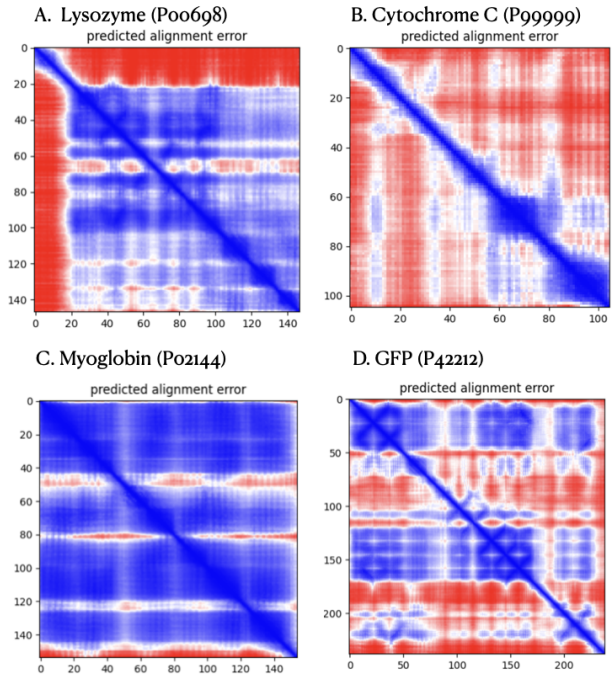

Figure 9: Rosetta Fold PAE charts

Summary: In order to tell which region each respective model predicted well, we can combine the answers from question 1 and the pLDDT and PAE plots from question 2. From question 1, we see the score visually by the bar indicating Regions with pLDDT > 90, Regions with pLDDT 70–90, and Regions with pLDDT < 70:

- **Alphafold2 predictions:** Proteins like Lysozyme (P00698) and Myoglobin (P02144) have high pLDDT confidence across their structured regions, indicated by blue in the models. These proteins have well-defined secondary structures, like alpha-helices and beta-sheets, which are predicted with high confidence. Cytochrome C (P99999) and GFP (P42212), however, display a combination of high- and low-confidence regions, with flexible loops and unstructured areas marked in yellow and red, indicating uncertainty in these regions. The pLDDT charts (Figure 4) confirm high confidence for structured regions in Lysozyme and Myoglobin, while Cytochrome C and GFP exhibit greater variability, particularly in loops or disordered regions. In the PAE diagrams (Figure 7), Alphafold2 shows low predicted alignment error (in blue) for secondary structures in Lysozyme and Myoglobin, while Cytochrome C and GFP display increased alignment errors (in red) in flexible loop regions, reflecting the model's difficulty in predicting these disordered areas accurately.

- **Alphafold3 predictions:** The Alphafold3 models demonstrate improved confidence across all proteins, with most regions, including well-structured secondary elements like alpha-helices and beta-sheets, colored blue, indicating pLDDT scores above 90. The pLDDT charts (Figure 5) also show consistently high confidence, with very few regions below pLDDT 70, suggesting greater predictive accuracy compared to Alphafold2. In the PAE diagrams (Figure 8), Alphafold3 maintains low alignment error across structured regions for Lysozyme and Myoglobin and shows improved alignment for Cytochrome C and GFP as well, although some loop regions still exhibit moderate alignment error. This suggests that Alphafold3 is better equipped to handle both structured secondary elements and flexible regions in complex proteins.

- **RosettaFold predictions:** RosettaFold predictions display more variation in confidence, particularly for Cytochrome C and GFP, where lower-confidence regions are indicated by red and yellow, highlighting difficulties in modeling complex and unstructured regions. While RosettaFold can reliably predict

well-defined secondary structures, like alpha-helices in Lysozyme and Myoglobin, it struggles more with regions that lack stable secondary structure, as seen in Cytochrome C and GFP. The pLDDT charts (Figure 6) show greater variability in confidence levels compared to the Alphafold models, especially in flexible or disordered regions, where pLDDT values frequently fall below 0.5. The PAE diagrams (Figure 9) reveal the most alignment error in RosettaFold predictions, especially for proteins with flexible regions, such as Cytochrome C and GFP, indicating that RosettaFold struggles more with reliably predicting disordered regions compared to Alphafold2 and Alphafold3.

Overall, AlphaFold2 and AlphaFold3 excel in predicting well-structured proteins due to their transformer-based architectures and the use of multiple sequence alignments. However, they both struggle with disordered regions and protein complexes (the latter issue being mitigated in AlphaFold3). RosettaFold, while flexible and good for novel folds, may not have the same level of accuracy for complex structures and shows more variability in predictions, particularly for dynamic or disordered regions.

The differences in performance between AlphaFold2, AlphaFold3, and RosettaFold arise from the core methodologies behind each model. AlphaFold2 uses deep learning and evolutionary information from multiple sequence alignments (MSAs) to predict protein structures, making it particularly strong at modeling well-structured proteins. AlphaFold3 builds on AlphaFold2's architecture but adds improved all-atom modeling, which allows it to handle more complex tasks like protein-ligand and protein-DNA interactions, making it more versatile. On the other hand, RosettaFold relies heavily on co-evolutionary signals and flexible sampling techniques, which makes it effective for predicting novel folds, especially when evolutionary data is available, but it struggles more with dynamic and disordered regions, showing more variability in confidence. These differences reflect the models' unique balance between machine learning, evolutionary data, and structural sampling.

```
## The code for alphafold3's pLDDT:
import json
import matplotlib.pyplot as plt
import numpy as np

def extract_plddt_from_json(json_file):
    # Load the JSON file
    with open(json_file, 'r') as f:
        data = json.load(f)

    # Extract pLDDT scores from the 'atom_plddts' field
    plddt_scores = data.get('atom_plddts', [])

    if not plddt_scores:
        print(f"No pLDDT scores found in {json_file}.")
        return None

    return plddt_scores

def plot_plddt(all_scores, labels):
    # Plot the pLDDT scores for multiple data sets
    plt.figure(figsize=(12, 8))

    for scores, label in zip(all_scores, labels):
        plt.plot(np.arange(len(scores)), scores, marker='o', linestyle='-', markersize=3, label=label)

    plt.title('pLDDT Scores across Residues (Multiple Files)', fontsize=14)
    plt.xlabel('Residue Index', fontsize=12)
    plt.ylabel('pLDDT Score', fontsize=12)

    # Add confidence threshold lines
    plt.axhline(y=90, color='green', linestyle='--', label='Very high confidence (>90)')
    plt.axhline(y=70, color='orange', linestyle='--', label='Medium confidence (70-90)')
    plt.axhline(y=50, color='red', linestyle='--', label='Low confidence (<70)')

    plt.legend()
    plt.grid(True)
    plt.show()

```

```
# List of file paths
json_files = [
    'fold_2024_10_27_16_37/fold_2024_10_27_16_37_full_data_0.json',
    'fold_2024_10_27_16_37/fold_2024_10_27_16_37_full_data_1.json',
    'fold_2024_10_27_16_37/fold_2024_10_27_16_37_full_data_2.json',
    'fold_2024_10_27_16_37/fold_2024_10_27_16_37_full_data_3.json'
]

json_files = [
    'fold_2024_10_27_16_37-2/fold_2024_10_27_16_37_full_data_0.json',
    'fold_2024_10_27_16_37-2/fold_2024_10_27_16_37_full_data_1.json',
    'fold_2024_10_27_16_37-2/fold_2024_10_27_16_37_full_data_2.json',
    'fold_2024_10_27_16_37-2/fold_2024_10_27_16_37_full_data_3.json'
]

json_files = [
    'fold_2024_10_27_16_37-3/fold_2024_10_27_16_37_full_data_0.json',
    'fold_2024_10_27_16_37-3/fold_2024_10_27_16_37_full_data_1.json',
    'fold_2024_10_27_16_37-3/fold_2024_10_27_16_37_full_data_2.json',
    'fold_2024_10_27_16_37-3/fold_2024_10_27_16_37_full_data_3.json'
]

json_files = [
    'fold_2024_10_27_16_37-4/fold_2024_10_27_16_37_full_data_0.json',
    'fold_2024_10_27_16_37-4/fold_2024_10_27_16_37_full_data_1.json',
    'fold_2024_10_27_16_37-4/fold_2024_10_27_16_37_full_data_2.json',
    'fold_2024_10_27_16_37-4/fold_2024_10_27_16_37_full_data_3.json'
]

all_plddt_scores = []
labels = ['Dataset 0', 'Dataset 1', 'Dataset 2', 'Dataset 3']

# Extract pLDDT scores for all files
for json_file in json_files:
    plddt_scores = extract_plddt_from_json(json_file)
    if plddt_scores:
        all_plddt_scores.append(plddt_scores)

# Plot the results
if all_plddt_scores:
    plot_plddt(all_plddt_scores, labels)
```

4 Find each protein's experimentally derived structure. Load the experimental and predicted structures into PyMol. Align them, and note any differences in structures.

Solution: We used the experimental structures, listed as: 2CDS for Lysozyme (P00698), 2N9I for Cytochrome C (P99999), 3RGK for Myoglobin (P02144), 2G16 for GFP (P42212). The results and the summary are the following:

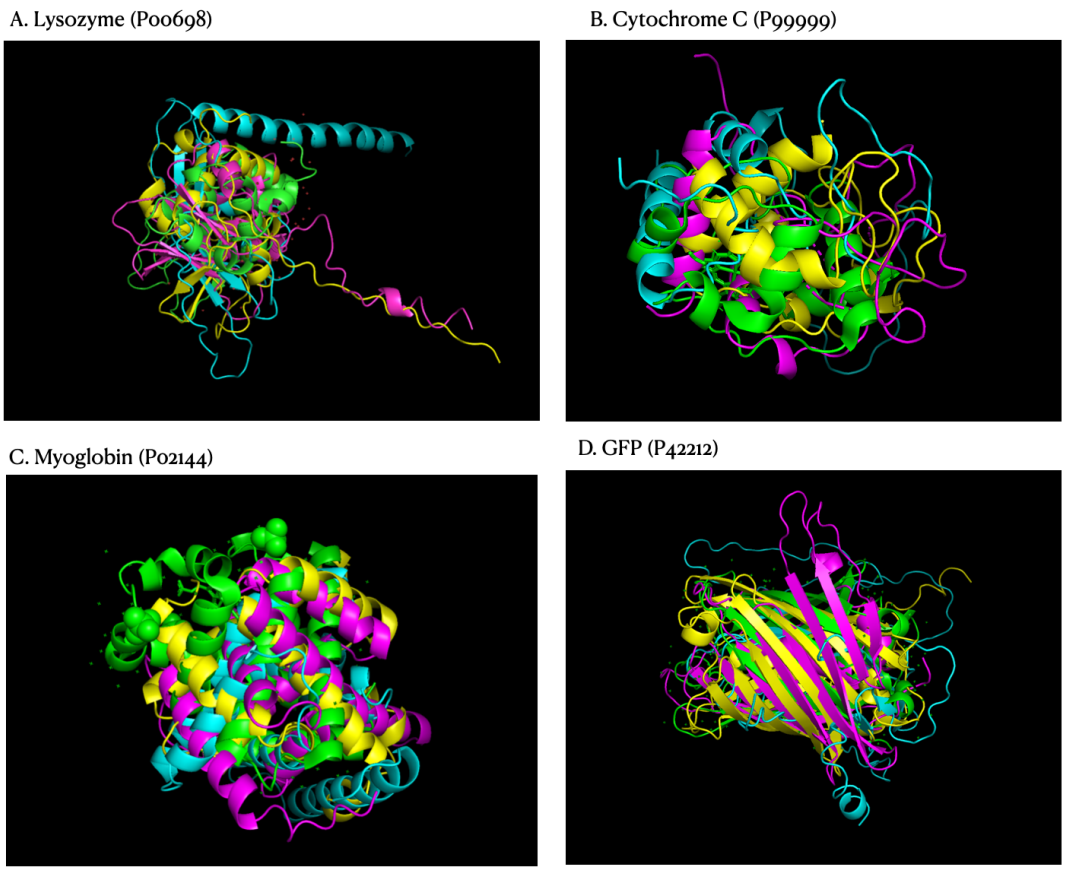

Figure 10: Experimental: green, Alpha2: blue, rosetta: pink, Alpha3: yellow.

Summary: For structural differences, Cytochrome C shows strong similarity across both the experimental and predicted structures, standing out compared to the other three proteins. In Lysozyme, the predicted structures differ slightly, especially in the extended tail region present in the predictions but absent in the experimental structure. For Myoglobin, the experimental structure includes an ion and a bound molecule, missing in all the predictions, though the core structural elements align well. Lastly, for GFP, while all models accurately predict the beta-sheet formation, the experimental structure's beta-sheets are more compact and tightly packed compared to the more extended forms seen in the predictions.

5 Restriction Enzyme ECORI binds to the DNA recognition site GAATTC. Generate structures for ECORI using Alphafold2, and Alphafold3. This is an opportunity to explore Alphafold3's capabilities as an all-atom capable model.

Solution: We picked ECORI from here: ECORI, and we obatined the alphafold2 and alphafold3 results as the following:

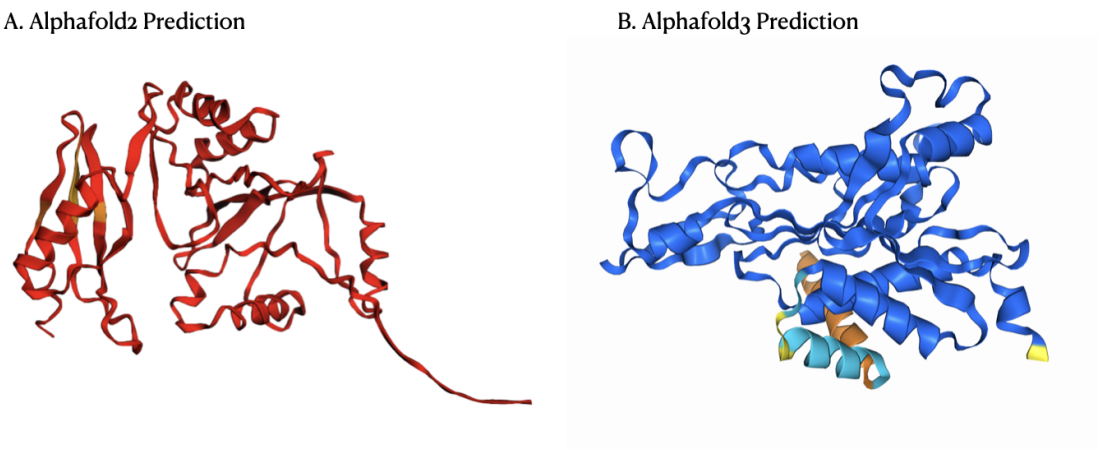

Figure 11: ECORI Structures

Summary: The comparison between Alphafold2 and Alphafold3 highlights the advantages of Alphafold3's all-atom capabilities. While Alphafold2 provides a reliable prediction of the protein's backbone and general fold, Alphafold3 offers a more refined, detailed atomic-level model, including accurate side-chain positioning. This added detail is particularly important for understanding specific molecular interactions, such as how ECORI binds to and cleaves DNA. The all-atom features of Alphafold3 allow for more precise modeling of protein-DNA complexes and active site interactions, making it a superior tool for studying enzymes and their mechanisms in finer detail.

6 Generate multiple different structures with Alphafold3 by prepending and appending nucleotides, both including and not including the complementary DNA recognition site. Is there a difference in Alphafold3's prediction when DNA is present?

Solutions: We will try three things here:

- Alphafold3 with just the ECORI sequence (no DNA).
- Alphafold3 with ECORI + GAATTC prepended or appended, since ECORI binds to the sequence GAATTC.
- Alphafold3 with ECORI + a random DNA sequence prepended or appended, and we picked TGCATG, and we do it with prepended.

The following is the result:

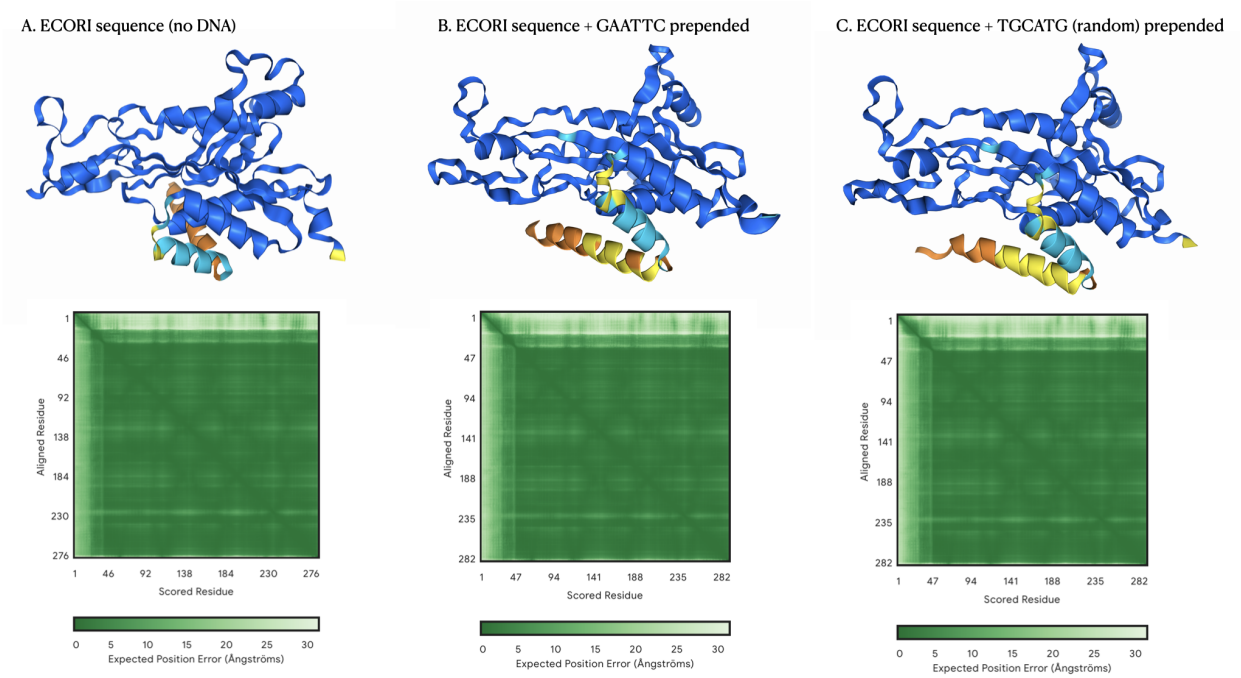

Figure 12: Alphafold3 Results

Summary: The predicted structures of ECORI, with and without DNA, show that the overall fold of the enzyme remains stable across all conditions. In all three cases—no DNA, GAATTC prepended, and TGCATG (random DNA) prepended—the core alpha-helical structure of ECORI is maintained. The pLDDT charts demonstrate high confidence in Alphafold3's predictions, with minimal changes in confidence across the different inputs. This indicates that the protein's structure is consistently well predicted, regardless of the presence or type of DNA sequence.

However, there are subtle differences between the predictions when the recognition site GAATTC is present compared to the random sequence. These variations, though slight, may suggest that Alphafold3 predicts some interaction between ECORI and its specific DNA substrate. The random DNA sequence (TGCATG) does not appear to induce any significant structural changes, indicating that non-specific DNA does not strongly affect Alphafold3's predictions in this context.

Overall, while the predicted structures are similar, the inclusion of the recognition sequence hints at potential protein-DNA interactions, even if they are not captured dramatically in these predictions. The high confidence in all cases supports the robustness of Alphafold3 in predicting ECORI's general structure, though capturing fine-grained binding interactions might require further refinement.

7 Look into Restriction Enzyme Biology a little deeper. What else would you propose adding to the Alphafold3 input to sharpen the prediction?

Solution: In order to further sharpen Alphafold3's predictions for ECORI, several biological elements could be added to the input:

- One important factor is the inclusion of cofactors like Mg^{2+} ions, which are essential for ECORI's enzymatic activity. Incorporating these metal ions in the prediction process could help Alphafold3 capture the enzyme's active conformation more accurately, as Mg^{2+} is crucial for DNA cleavage and could induce specific structural rearrangements.

- Another potential improvement would be to better model DNA-protein interactions by allowing Alphafold3 to account for the binding between ECORI and its recognition site, GAATTC. While Alphafold3 is capable of predicting protein-DNA complexes, refining its ability to capture the dynamics of specific binding interactions could lead to more accurate predictions. This is especially relevant for restriction enzymes, where binding to DNA often induces conformational changes necessary for catalysis.
- Lastly, incorporating environmental factors such as pH and temperature could also enhance the predictive accuracy. These factors can affect protein folding and stability, particularly for enzymes that function under specific conditions. Although Alphafold3 does not currently model such factors, future versions could benefit from simulating proteins in environments that mimic their physiological conditions.