AI-Assisted Protein Analysis: From Sequence Representation to Drug-Binding Hypothesis Generation

Debabrata Pruseth
AI Architect & Applied AI Researcher
Singapore

Author Note

This article is a research-style companion version of the author's blog post "AI-Assisted Protein Analysis: From Sequence to Drug-Binding Hypothesis" and the associated GitHub project AI-Assisted-Structure-Based-Drug-Discovery-Pipeline.

In this study, the author developed an applied AI workflow that demonstrates how publicly available computational biology tools and large language models can be combined to move from protein sequence representation to an early-stage drug-binding hypothesis. The work uses **E. coli dihydrofolate reductase**, also known as **DHFR / folA**, as a practical case study.

The workflow combines protein sequence representation, AlphaFold/ColabFold-style structure prediction, confidence evaluation using pLDDT scores, receptor and ligand preparation, molecular docking using AutoDock Vina, residue-level interaction analysis, and LLM-assisted scientific interpretation.

This work is exploratory and educational. It does not claim biological validation, drug efficacy, experimentally confirmed binding, or therapeutic relevance. The objective is to demonstrate a transparent and reproducible approach for early-stage computational hypothesis generation.

Abstract

Artificial intelligence and computational biology tools are making early-stage structure-based drug discovery workflows more accessible to learners, researchers, and applied AI practitioners. This paper presents an AI-assisted protein analysis pipeline that begins with protein sequence representation and ends with a cautious drug-binding hypothesis. The case study uses **E. coli DHFR / folA**, a 159-residue enzyme involved in folate metabolism and DNA synthesis. The author developed a workflow that uses ColabFold / AlphaFold-style structure prediction to generate a three-dimensional protein model, evaluates local structural confidence through pLDDT scores, performs exploratory molecular docking using AutoDock Vina, identifies candidate ligand-proximal residues, and uses an LLM-assisted interpretation step to convert computational outputs into a structured scientific summary. The predicted DHFR structure showed high local confidence, with a mean pLDDT of 95.49 and 146 out of

159 residues above 90. Docking produced a best AutoDock Vina score of approximately −5.686 kcal/mol, suggesting a modest but computationally plausible ligand-protein interaction. The study concludes that AI-assisted protein analysis can support early hypothesis generation, but docking scores and predicted structures require further computational refinement and experimental validation before biological claims can be made.

Keywords: AlphaFold, ColabFold, Protein Structure Prediction, DHFR, folA, AutoDock Vina, Molecular Docking, Drug Discovery, Structure-Based Drug Discovery, pLDDT, Computational Biology, Large Language Models, AI-Assisted Interpretation, Binding Hypothesis, Bioinformatics

---

1. Introduction

Proteins are fundamental molecular machines in biological systems. They catalyze reactions, regulate pathways, transmit signals, support cellular structure, and enable many essential biological functions. Although a protein is encoded as a one-dimensional amino acid sequence, its biological behavior depends heavily on its three-dimensional structure. Functional regions such as binding pockets, catalytic sites, interaction surfaces, and conformational grooves emerge only after the protein folds into its spatial form.

In structure-based drug discovery, a key question is whether a small molecule can bind to a target protein in a way that may influence biological function. Traditionally, this type of analysis depended on experimentally resolved structures obtained through X-ray crystallography, nuclear magnetic resonance, or cryo-electron microscopy. More recently, AI-based protein structure prediction methods such as AlphaFold and ColabFold have made high-confidence predicted protein structures accessible to a much wider group of researchers and learners.

This study explores the following practical question:

Can an AI-predicted protein structure be used to support early-stage drug-binding hypothesis generation when combined with molecular docking and structured scientific interpretation?

To address this question, the author developed a compact pipeline that moves through five analytical stages:

1. Protein sequence representation
2. Structure prediction and confidence evaluation
3. Docking-based interaction analysis
4. Residue-level interpretation
5. LLM-assisted scientific synthesis

The pipeline is intentionally framed as a hypothesis-generation workflow. It is not presented as a validated drug discovery method or as evidence of experimentally confirmed binding.

---

2. Research Objective

The objective of this study is to demonstrate how accessible AI and computational tools can be combined into a beginner-friendly but scientifically cautious protein analysis workflow.

The specific objectives are:

1. To use a protein amino acid sequence as the starting representation for structure-based analysis.
2. To generate a predicted three-dimensional structure using a ColabFold / AlphaFold-style workflow.
3. To evaluate the predicted structure using local confidence scores.
4. To conduct exploratory docking between the predicted protein structure and a demonstration ligand.
5. To identify candidate residues located near the docked ligand pose.
6. To use an LLM-assisted interpretation step to convert technical outputs into a structured scientific summary.
7. To define the limitations of the workflow and identify future steps required for more rigorous validation.

---

3. Target Protein and Case Study Context

The target protein selected for this study is **E. coli dihydrofolate reductase**, commonly referred to as **DHFR** and associated with the **folA** gene. The selected target is a 159-residue enzyme involved in folate metabolism and nucleotide biosynthesis.

DHFR is a useful educational target for computational drug-discovery demonstrations because it is biologically meaningful, well studied, and historically relevant in enzyme inhibition research. In biological systems, DHFR contributes to pathways required for DNA synthesis. For this reason, DHFR has been widely studied in drug-discovery contexts.

In this study, DHFR is used as a representative protein target for demonstrating an AI-assisted computational workflow. The study does not claim to identify a new inhibitor or validate any therapeutic compound.

---

4. Conceptual Approach

The author designed the workflow around a simple research logic:

A protein sequence can be transformed into a predicted structure; the structure can be evaluated for confidence; the structure can be used for exploratory docking; docking can identify plausible interaction regions; and AI-assisted interpretation can help convert technical outputs into a cautious scientific hypothesis.

This approach reflects a broader pattern in applied AI research: AI tools are not used as isolated black boxes, but as components in a structured analytical pipeline. Each stage produces an output that informs the next stage.

The conceptual workflow is:

Stage	Purpose	Research Output
Sequence representation	Define the biological target	Protein amino acid sequence
Structure prediction	Generate 3D structural hypothesis	Predicted protein structure
Confidence evaluation	Assess local reliability	pLDDT confidence profile
Docking analysis	Explore ligand-protein compatibility	Docking poses and affinity estimates
Residue interaction analysis	Identify candidate binding region	Ligand-proximal residues
Scientific interpretation	Convert outputs into cautious insight	Binding hypothesis and next steps

The workflow is designed to be transparent and reproducible. However, each stage is also subject to uncertainty. The study therefore emphasizes cautious interpretation throughout.

---

5. Methodology

5.1 Protein Sequence Representation

The starting point of the study was the amino acid sequence of **E. coli DHFR / folA**. In this pipeline, the sequence acts as the biological representation from which the downstream structure prediction is initiated.

The use of a sequence-first approach is important because it reflects how modern AI-assisted protein analysis can begin without an experimentally resolved structure. This makes the workflow accessible for educational and exploratory use cases where only the protein sequence is available.

5.2 Structure Prediction

The author used a ColabFold / AlphaFold-style structure prediction workflow to generate a three-dimensional model of the DHFR target. The purpose of this step was to transform the one-dimensional amino acid sequence into a spatial model suitable for structural analysis.

The resulting predicted structure served as the receptor model for downstream confidence evaluation and docking. Since the structure was computationally predicted rather than experimentally resolved, it was not assumed to be automatically valid for all downstream uses. Instead, it was evaluated using local model confidence scores.

5.3 Confidence Evaluation

The predicted DHFR structure was evaluated using **pLDDT** scores, which provide residue-level confidence estimates in AlphaFold-style outputs. Higher pLDDT values indicate greater local confidence in the predicted structural position of residues.

The confidence analysis produced the following summary:

Metric	Value
Number of residues	159
Mean pLDDT	95.49
Minimum pLDDT	77.25
Maximum pLDDT	98.75
Residues above 90	146
Residues between 70 and 90	13
Residues below 70	0

The predicted model showed high confidence across most of the structure. A mean pLDDT of 95.49 and 146 residues above 90 suggest that the model is locally reliable for exploratory structural analysis. No residues were below 70, indicating the absence of low-confidence regions under the applied threshold.

However, pLDDT confidence should not be interpreted as proof of biological completeness. A high-confidence predicted structure may still fail to capture ligand-induced conformational changes, protein flexibility, alternate states, solvent effects, or dynamic behavior.

5.4 Docking-Based Binding Hypothesis Generation

After confidence evaluation, the predicted protein structure was used as the receptor for exploratory molecular docking. The author prepared the receptor and a demonstration ligand for docking-compatible representation and performed docking using **AutoDock Vina**.

The docking stage was designed to answer a limited question:

Does the demonstration ligand produce a computationally plausible docking pose against the predicted DHFR structure?

It was not designed to prove binding, biological inhibition, or pharmacological activity.

For the initial docking run, the search region was defined as a broad protein-centered box. This was suitable for a beginner-friendly demonstration, but it is not the most rigorous strategy for production-level docking. A more mature workflow should define the docking grid using known active-site residues, experimentally resolved ligand-bound structures, pocket-detection algorithms, or literature-supported binding regions.

5.5 Residue-Level Interaction Analysis

Following docking, the author identified protein residues located near the docked ligand pose. A proximity threshold of **4.5 Å** was used to identify candidate ligand-proximal residues.

This residue-level analysis helps move the interpretation beyond a single docking score. It provides a structural basis for discussing where the ligand may be positioned relative to the protein surface or pocket.

5.6 LLM-Assisted Scientific Interpretation

The final stage used a large language model as a scientific interpretation assistant. The LLM was not used to perform structure prediction or molecular docking. Instead, it was used to organize the computational outputs into a structured, cautious interpretation.

The interpretation step was constrained to avoid overclaiming. The language was deliberately framed around terms such as:

- computationally plausible
- exploratory
- hypothesis-generating
- requires validation

This ensured that the LLM-assisted summary remained aligned with the limitations of the computational workflow.

---

6. Results

6.1 Structure Prediction Outcome

The predicted DHFR structure showed strong local confidence across most of the 159-residue protein. The mean pLDDT score was **95.49**, and **146 residues** had pLDDT scores above 90.

This result indicates that the predicted structure is suitable for exploratory structural analysis and educational docking experiments. The absence of residues below 70 supports the overall local reliability of the model.

However, the predicted structure remains a static computational model. It should not be interpreted as a complete substitute for experimentally resolved structures or molecular dynamics-based conformational sampling.

6.2 Docking Outcome

The docking analysis produced five candidate docking modes. The best AutoDock Vina affinity score was approximately **−5.686 kcal/mol**.

Mode	Affinity kcal/mol	RMSD Lower Bound	RMSD Upper Bound
1	−5.686	0.000	0.000
2	−5.676	1.961	5.167
3	−5.446	2.529	3.526
4	−5.297	2.298	5.204
5	−5.215	4.070	5.080

The docking score suggests a modest computationally plausible interaction. The top two docking modes had very similar scores, indicating that more than one pose may be energetically comparable under the simplified docking setup.

The score range does not indicate strong binding evidence. Instead, it supports the weaker and more appropriate conclusion that the ligand-protein interaction is worth further computational exploration.

6.3 Candidate Interaction Region

Residue-level proximity analysis identified a cluster of residues near the docked ligand pose. The candidate interaction region included residues such as:

ILE5, ILE14, ASN18, ASP27, PHE31, HIS45, THR46, ILE94, GLY95–97, ARG98, TYR100, and THR123.

These residues define a putative ligand-proximal region in the predicted DHFR structure. The presence of a localized interaction region supports the generation of a binding hypothesis. However, because the docking setup used a broad protein-centered search region, the residue cluster should not be interpreted as a confirmed binding site without additional validation.

6.4 Interpretation Outcome

The integrated interpretation of the structure prediction, confidence analysis, docking score, and residue proximity analysis suggests that the demonstration ligand may form a computationally plausible interaction with the predicted DHFR structure.

The strongest evidence supporting this interpretation is:

1. The predicted receptor structure had high local confidence.
2. Docking produced a negative AutoDock Vina affinity score.
3. A localized set of ligand-proximal residues was identified.
4. The outputs could be converted into a coherent hypothesis for further study.

The result is best stated as:

The pipeline generated an early-stage computational drug-binding hypothesis for E. coli DHFR, but the hypothesis remains unvalidated and requires further computational and experimental evaluation.

---

7. Discussion

This study demonstrates how accessible AI and computational biology tools can be combined into a structured sequence-to-hypothesis workflow. The author developed a pipeline that begins with a protein sequence, generates a predicted structure, evaluates confidence, performs docking, identifies a candidate interaction region, and produces a cautious scientific interpretation.

The main value of the workflow lies in its integration. Each stage contributes a different type of evidence. Structure prediction provides a receptor model. pLDDT analysis provides confidence context. Docking provides a ligand-pose hypothesis. Residue proximity analysis provides structural interpretation. LLM-assisted interpretation helps convert these outputs into a readable scientific summary.

The results are useful for education, early exploration, and hypothesis generation. They are not sufficient for biological or pharmacological claims. This distinction is central to responsible AI-assisted scientific analysis.

The high pLDDT score supports exploratory use of the predicted protein structure, but it does not remove the need for further validation. The docking score provides a possible interaction signal, but docking scores are approximate and sensitive to receptor preparation, ligand preparation, docking grid definition, scoring functions, and protein flexibility.

The candidate interaction residues provide a structural region for future investigation. However, a more rigorous study should compare the predicted region against known DHFR active-site information, experimental structures, known inhibitors, or pocket-detection results.

The LLM-assisted interpretation layer adds value by improving communication and synthesis, but it must be governed carefully. LLMs should summarize computational outputs, not invent scientific conclusions. In this study, the interpretation step was constrained to remain cautious and hypothesis-oriented.

---

8. Research Contribution

This study contributes an educational applied AI framework for sequence-to-hypothesis protein analysis. The contribution is not a new drug discovery result, but a reproducible workflow demonstrating how several tools can be combined responsibly.

The main contributions are:

1. **A sequence-to-hypothesis workflow**
   The study demonstrates how a protein sequence can be transformed into a cautious drug-binding hypothesis using structure prediction, docking, and interpretation.
2. **Confidence-aware structural analysis**
   The workflow evaluates pLDDT confidence before using the predicted structure for docking.
3. **Docking-based exploratory hypothesis generation**
   The study shows how AutoDock Vina can be used to generate an initial ligand-protein interaction hypothesis.
4. **Residue-level interpretability**
   The pipeline identifies candidate residues near the docked ligand pose, making the docking result more interpretable than a score alone.
5. **LLM-assisted scientific communication**
   The study demonstrates how an LLM can convert computational outputs into a structured scientific summary while maintaining cautious language.
6. **Responsible framing**
   The study explicitly avoids claims of validated binding, biological activity, or drug efficacy.

---

9. Limitations

This study has several important limitations.

First, the protein structure used in the pipeline is a predicted model. Although the local pLDDT confidence is high, the structure may not capture ligand-bound conformations, flexible loops, alternate states, or dynamic molecular behavior.

Second, the docking grid was defined using a broad protein-centered search region. This is suitable for an educational demonstration but less rigorous than a binding-site-guided docking protocol.

Third, the ligand was used as a demonstration compound. The study does not evaluate ligand chemistry, known biological activity, toxicity, selectivity, solubility, off-target effects, or pharmacokinetic properties.

Fourth, AutoDock Vina scores are approximate. They can support exploratory ranking but cannot establish experimentally measured binding affinity.

Fifth, receptor and ligand preparation were simplified. Protonation states, charge assignment, tautomer selection, solvent effects, cofactors, and water molecules can significantly affect docking outcomes.

Sixth, no molecular dynamics simulation was performed. Therefore, the stability of the docked pose was not assessed over time.

Seventh, no experimental validation was performed. The study does not provide evidence of enzymatic inhibition, binding affinity, biological effect, antimicrobial effect, or therapeutic value.

Finally, the LLM-generated interpretation should be treated as a communication aid. It is not a substitute for expert biochemical review or experimental validation.

---

10. Governance and Responsible Use

AI-assisted scientific workflows require careful governance because computational outputs can easily be overinterpreted. In drug discovery, this risk is especially important because docking results and AI-generated summaries may appear more conclusive than they actually are.

A responsible workflow should follow these principles:

1. Avoid claiming validated binding without experimental evidence.
2. Avoid calling a ligand a drug based only on computational docking.
3. Separate computational outputs from biological conclusions.
4. Report assumptions, parameters, and limitations transparently.
5. Use expert review before making scientific or medical claims.
6. Validate computational hypotheses through additional methods.
7. Treat LLM outputs as structured summaries, not as scientific proof.

This study follows these principles by framing the result as exploratory and hypothesis-generating.

---

11. Future Work

The pipeline can be strengthened in several ways.

First, the docking region should be refined using known DHFR active-site residues, experimental structures, pocket-detection tools, or literature-supported binding-site information.

Second, known DHFR inhibitors should be docked as positive controls. This would provide a benchmark for comparing the demonstration ligand against compounds with known relevance.

Third, multiple ligands should be screened instead of a single demonstration molecule. A small ligand library would allow ranking, clustering, and prioritization.

Fourth, receptor and ligand preparation should be improved using more specialized methods for protonation, charge assignment, tautomer selection, and water or cofactor treatment.

Fifth, molecular dynamics simulation should be added to assess pose stability and protein-ligand behavior over time.

Sixth, docking outputs should be compared across multiple tools or scoring functions to assess robustness.

Seventh, LLM-assisted interpretation should be combined with retrieval-augmented references from UniProt, PDB, PubChem, and scientific literature to reduce hallucination risk.

Finally, experimental validation would be required before any biological or pharmacological claim could be made.

---

12. Reproducibility Note

The implementation code and workflow artifacts are available in the author's GitHub repository:

AI-Assisted-Structure-Based-Drug-Discovery-Pipeline

The repository provides the notebook-based implementation and supports reproducibility of the workflow. Code-level details, file paths, intermediate artifacts, and execution-specific outputs are intentionally kept outside the main body of this paper to preserve the research focus on approach, outcome, interpretation, and limitations.

---

13. Conclusion

This paper presented an AI-assisted protein analysis workflow that moves from protein sequence representation to cautious drug-binding hypothesis generation. The author developed and implemented a compact pipeline using **E. coli DHFR / folA** as the case study target.

The predicted DHFR structure showed high local confidence, with a mean pLDDT of **95.49** and **146 out of 159 residues** above 90. Docking produced a best AutoDock Vina score of approximately **−5.686 kcal/mol**, indicating a modest computationally plausible interaction. Residue-level analysis identified a candidate ligand-proximal region, supporting the generation of an early binding hypothesis.

The central conclusion is that AI-assisted protein analysis can support early-stage hypothesis generation, education, and structured scientific interpretation. However, predicted structures, docking scores, and LLM-generated summaries cannot replace rigorous computational refinement or experimental validation.

The value of the pipeline lies in its transparency, reproducibility, and responsible framing. It demonstrates how AI and computational biology tools can be combined to generate cautious, explainable hypotheses while avoiding unsupported biological or therapeutic claims.

References

Pruseth, D. (2026). *AI-Assisted Protein Analysis: From Sequence to Drug-Binding Hypothesis*. Debabrata Pruseth AI Blog.

Pruseth, D. (2026). *AI-Assisted-Structure-Based-Drug-Discovery-Pipeline*. GitHub.

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*.

Mirdita, M., Schütze, K., Moriwaki, Y., et al. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking. *Journal of Computational Chemistry*.

Suggested Citation

Pruseth, D. (2026). *AI-Assisted Protein Analysis: From Sequence Representation to Drug-Binding Hypothesis Generation*. Debabrata Pruseth AI Blog.