Aug 27, 2025

Workflow for Protein Structure Prediction, Refinement, and Validation

DOI
dx.doi.org/10.17504/protocols.io.dm6gpmz2dgzp/v1

Marco Palma[1,2]

[1]IGDORE; [2]Biointelix LLC

BioLab

vriskone

DOI: https://dx.doi.org/10.17504/protocols.io.dm6gpmz2dgzp/v1

Protocol Citation: Marco Palma 2025. Workflow for Protein Structure Prediction, Refinement, and Validation. **protocols.io https://dx.doi.org/10.17504/protocols.io.dm6gpmz2dgzp/v1**

Protocol status: Working
We use this protocol and it's working

Created: August 26, 2025

Last Modified: August 27, 2025

Protocol Integer ID: 225561

Keywords: protein structure prediction services, protein structure prediction, protein structure, reproducible computational workflow for protein structure prediction, specialized protein structure prediction service, protein structure prediction, modeling protein structure, protein structure, bioinformatics, reproducible computational workflow, structural biology, researchers in structural biology, structure generation, drug discovery, sequence preparation, epitope mapping, workflow, docking

Abstract

This protocol provides a reproducible computational workflow for protein structure prediction using a combination of freely available software, open-source pipelines, and specialized **protein structure prediction services**. The protocol covers sequence preparation, structure generation, refinement, validation, and downstream applications such as docking and epitope mapping. This workflow is intended for researchers in structural biology, immunology, bioinformatics, and drug discovery who require reliable approaches for modeling protein structures when crystallographic or cryo-EM data are not available.

Materials

Input Data

- Protein sequence in FASTA format (single-letter amino acid code).
- Metadata: UniProt accession number, domain annotations, organism of origin.

Software and Tools

- **Primary prediction**: AlphaFold2 (local installation) or ColabFold (cloud-based).
- **Alternative approaches**: I-TASSER, Phyre2, SwissSidechain.
- **Refinement**: GalaxyRefine, ModRefiner.
- **Validation**: ProSA-web, PDBsum, MolProbity.
- **Docking/interaction analysis**: PyDockWEB, HADDOCK.
- **Optional commercial support**: Specialized **protein structure prediction services** providing expert-guided modeling and validation.

Troubleshooting

Step-by-Step Procedure

1 **Sequence Preparation (30–60 min)**

1.1 Retrieve protein sequence from UniProt or NCBI in FASTA format.

1.2 Inspect the sequence for non-standard residues, signal peptides, or tags.

- *Tip*: Remove signal peptides using SignalP if the modeling target is the mature protein.

1.3 For multi-domain proteins, split the sequence into predicted domains using Pfam or InterPro.

2 **Primary Protein Structure Prediction (6–24 h depending on size)**

2.1 AlphaFold2/ColabFold

- Input the full sequence.
- Set number of recycles = 3–6 for improved accuracy.
- Enable template search (MMseqs2 or PDB70 database).
- Output: 5 ranked models with confidence scores (pLDDT).

2.2 Alternative approaches

- Use I-TASSER or Phyre2 when AlphaFold2 fails for low-complexity regions.
- Generate comparative models if homologous structures are available.

2.3 Protein Structure Prediction Services

- Submit the sequence to a professional modeling provider if large-scale computing resources are unavailable.
- Services often deliver optimized models, quality metrics, and optional docking simulations.

3 **Model Refinement (2–6 h)**

3.1 Select top models based on:

- pLDDT score > 70 (high confidence).
- Predicted TM-score > 0.5 (correct topology).

3.2 Refine structure using GalaxyRefine:

- Input predicted PDB file.
- Generate up to 5 refined models.
- Compare RMSD and MolProbity scores.

3.3 Optional: Run ModRefiner for atomic-level energy minimization.

4 Model Validation (2–4 h)

4.1 Stereochemistry:

- PDBsum for Ramachandran plot and hydrogen bond networks.
- MolProbity for clash score and rotamer outliers.

4.2 Global quality assessment:

- ProSA-web: Z-score comparison with experimentally determined structures.
- ERRAT for non-bonded atomic interactions.

4.3 Consensus evaluation:

Compare refined models and select the one with the highest structural consistency.

5 Functional Annotation and Downstream Applications (variable)

5.1 Ligand docking:

- Use PyDockWEB or AutoDock Vina to simulate protein–ligand interactions.
- Analyze binding energy and interface residues.

5.2 Protein–protein docking:

- HADDOCK for macromolecular interactions.
- Identify key interface regions.

5.3 Epitope mapping (for vaccine research):

- Predict linear and conformational epitopes using ElliPro or Discotope.
- Validate predicted epitopes on refined protein models.

5.4 Comparative analysis:

Align predicted model to homologous crystal structures for benchmarking.

Note

Notes and Recommendations

- **Computational resources**: Full-length proteins >1000 aa may require GPUs and high-memory servers. Domain-based modeling can reduce complexity.
- **Disordered regions**: Low-confidence predictions (pLDDT < 50) often correspond to intrinsically disordered segments; interpret with caution.
- **Multi-conformation proteins**: Ensemble modeling provides insights into functional flexibility.
- **Professional services**: Researchers without bioinformatics expertise may rely on **protein structure prediction services**, which offer curated pipelines, expert validation, and consulting for specific projects.

Expected result

Expected Outcomes

- Generation of reliable 3D protein models.
- Validated structures with stereochemical quality and global metrics.
- Application-ready models for docking, epitope mapping, or rational drug design.

Protocol references

Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

Yang, J. et al. (2015). The I-TASSER Suite: protein structure and function prediction. Nat Methods, 12, 7–8.

Laskowski, R. A. et al. (2011). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res, 33.

Acknowledgements

We acknowledge the contributions of the open-source computational biology community and the availability of expert-led protein structure prediction services, which enhance accessibility to structural modeling for the broader scientific community.