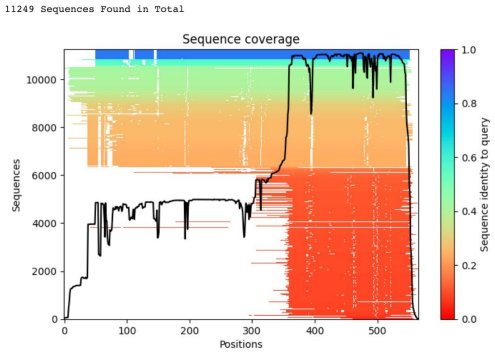
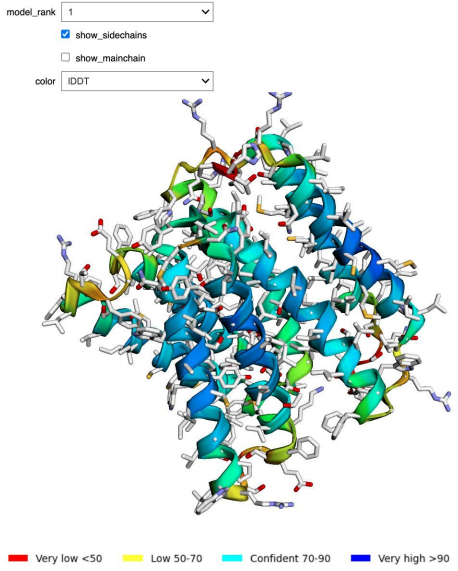
Protein Ai "A Platform for Predicting Protein Structures based on the Molecule's Sequence with ChatBot Interface for Analysis"

Hasan Aldhahi
Supervisor: Narges Lux, Julian Kunkel
Website: protein-ai.academiccloud.de





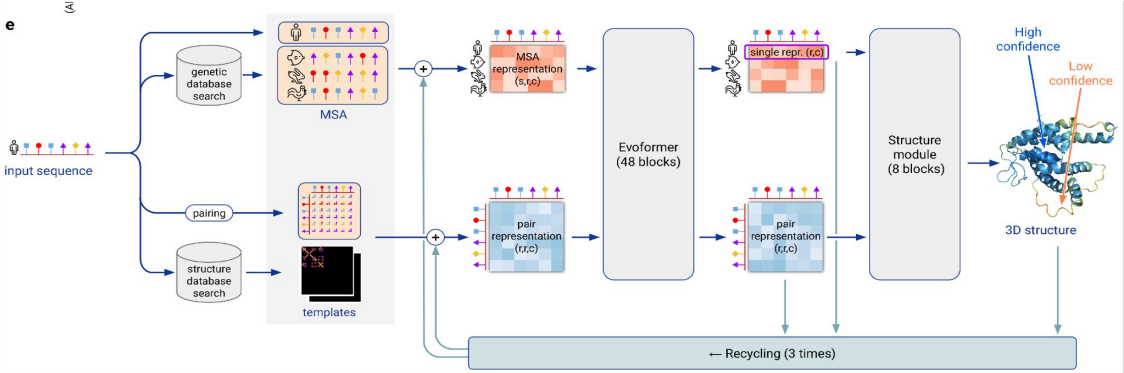
Introduction

nature

https://doi.org/10.1038/s41586-021-03819-2

Accelerated Article Preview

Highly accurate protein structure prediction with AlphaFold



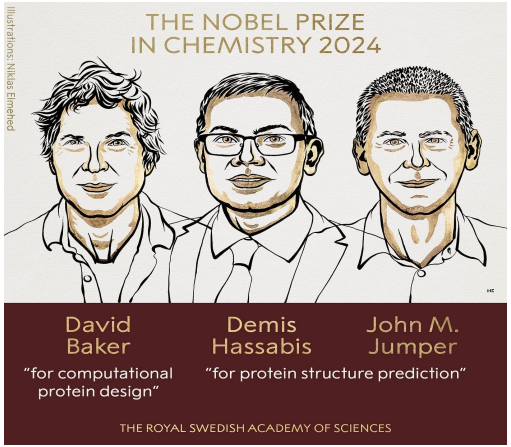
Received: 11 May 2021

Accepted: 12 July 2021

Accelerated Article Preview Published online 15 July 2021

Cite this article as: Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli & Demis Hassabis



Motivation

Importance of Protein Prediction:

- **Drug Discovery & Development:** Helps identify potential drug targets by predicting protein structures and interactions with drug molecules
- **Disease Understanding:** Assists in uncovering the molecular basis of diseases by predicting how mutations in protein sequences may alter structure and function.
- **Synthetic Biology & Bioengineering:** Facilitates the design of novel proteins for industrial, environmental, and therapeutic applications, such as enzymes for biofuels or bioplastics.

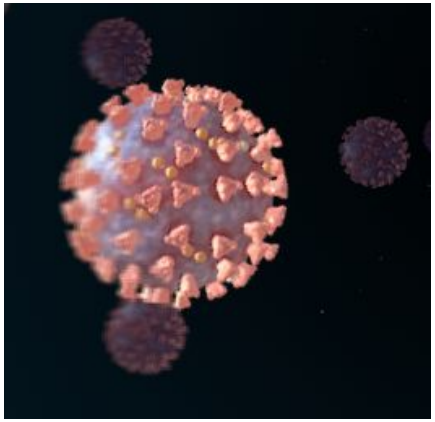
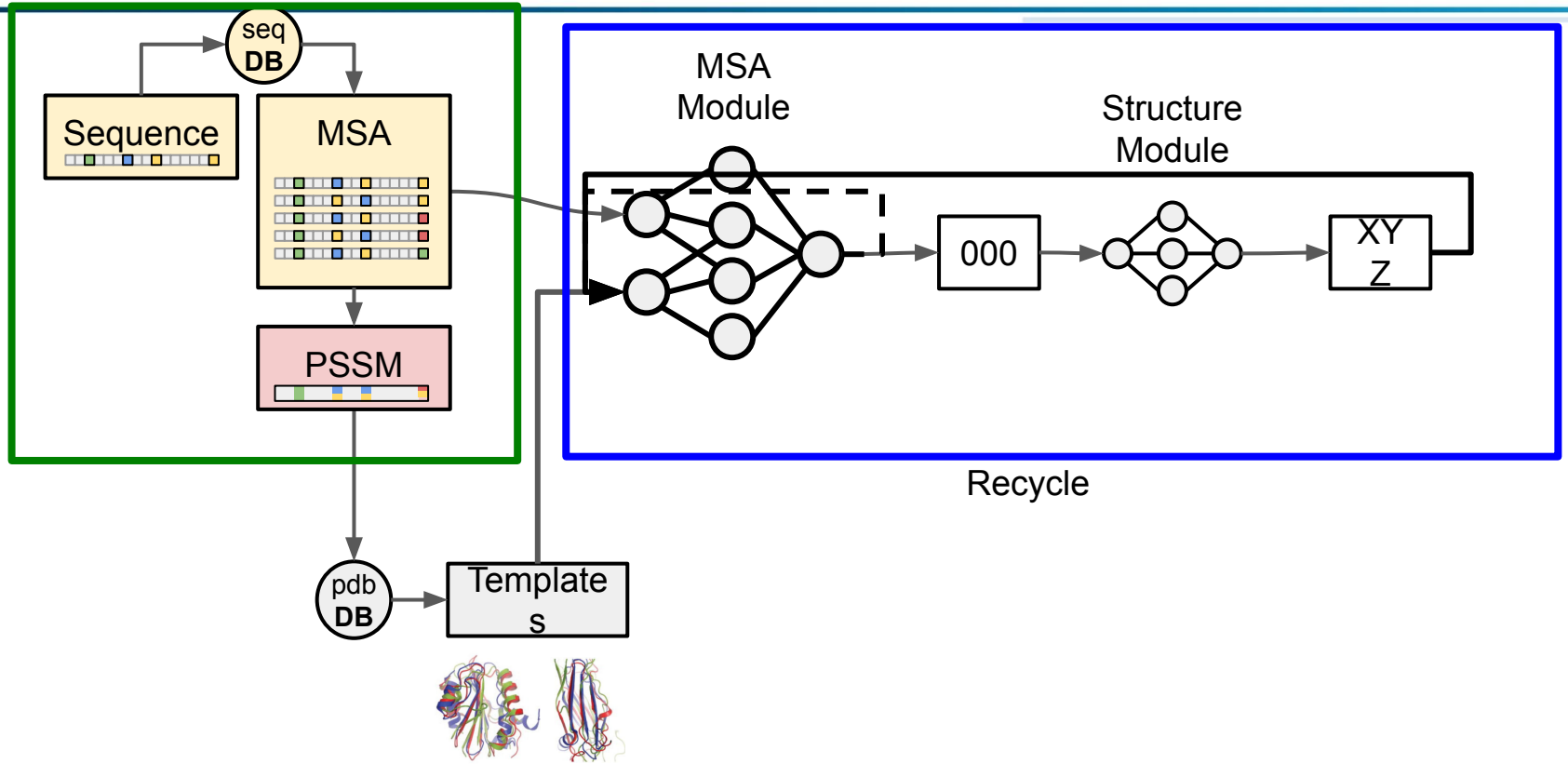


Figure 1: Covid-19 3D structure

Alpha Fold 2 Architecture

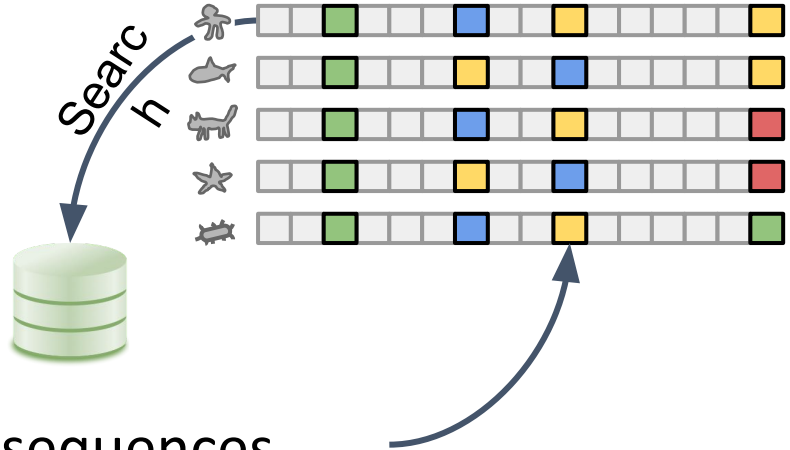


Sequence alignment

Generate a multiple sequence alignment

Sequence Example .pdb

- MKTLLAILAVLATAVATTFSRGSH…



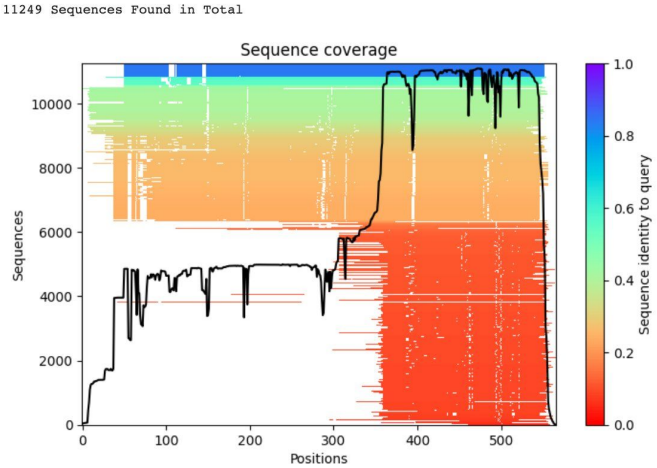


Figure 2: Output of the Alignment Sequence

Sequence alignment

- Sample pLDDT (predicted Local Distance Difference Test) for a predicted protein structure
- AlphaFold2 assign a low confidence to a region of a protein if:
  - highly flexible or intrinsically disordered
  - does not have any well-defined structure

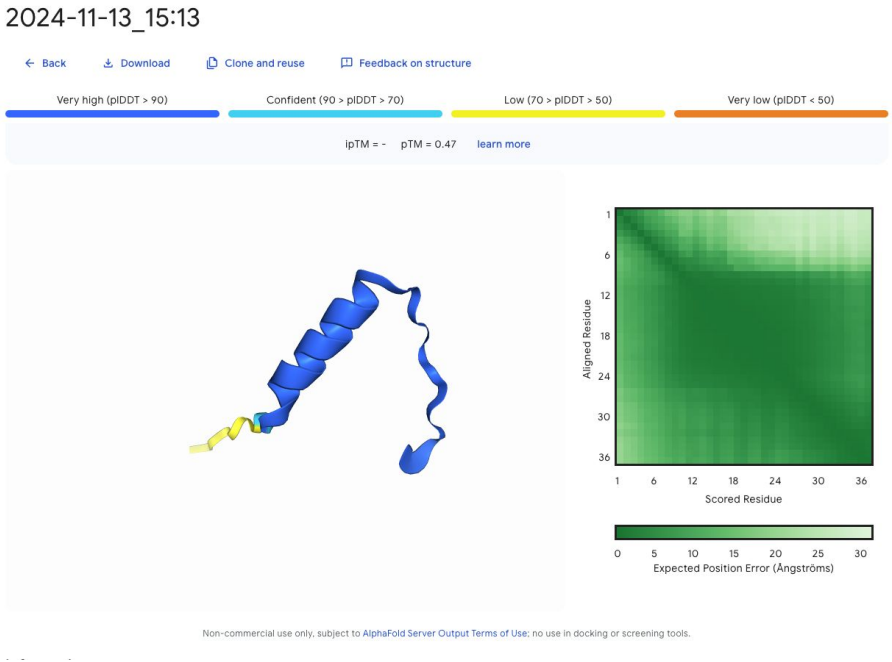


Figure 3: Final Output from AlphaFold

Making Advanced Prediction Accessible

1. **The Challenge:**
   - Predicting 3D molecular structures is crucial (Function, Therapeutics, Materials).
   - Breakthroughs (Alpha Fold, ColabFold) offer high accuracy.
   - BUT: Using these tools is complex:
     a. Requires High-Performance Computing (HPC), GPUs.
     b. Demands bioinformatics expertise.
     c. Often involves command-line interfaces, dependency management.
2. **The Gap:** Need for an easy-to-use, accessible interface for these powerful tools.
3. **Central Research Question:**
   How can GPU-intensive models (ColabFold, Boltz) be optimally integrated into a secure, user-friendly web app on HPC?
4. **Key Objectives:** Optimization, User Experience, Scientific Accuracy, Security & Privacy.

A Full-Stack Approach

- **Front-End:**
  - Complete overhaul from legacy JS to Next.js
  - Focus on performance, scalability, maintainability.
- **Back-End:**
  - Next.js acting as the intermediary API.
  - Designed for seamless, asynchronous data flow.
- **HPC Integration:**
  - Connecting web requests to GPU jobs via robust scripting.
  - Handling long-running computations efficiently.

Figure 4 : Full Architecture

Creating an Intuitive Interface

- Choosing the Model for prediction
  - Boltz
  - Alpha Fold 2
- Client-Side Validation:
  - Leveraged WebAssembly to run RDKit
  - Checks if entry is Protein, DNA, RNA
- Frontend-Fetching Jobs:
  - Downloadable Output
  - Write Description for Jobs
- Accept up to 2500 Molecules

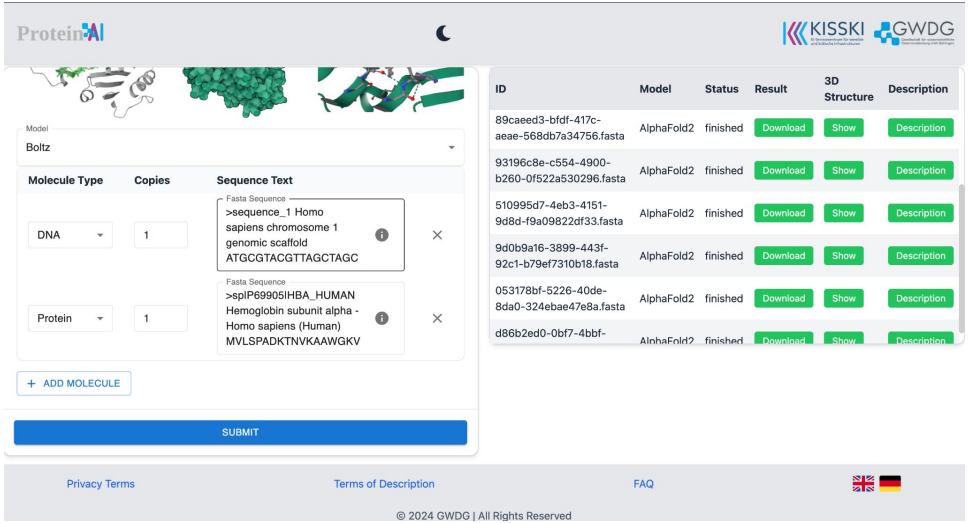


Figure 5: Home Page of Protein Ai

The LLM Analytical Assistant - Innovation & Security

- Context-Awareness:
  - Backend process parses output PDB file.
  - Extracts key structural features (helices, sheets, confidence scores).
- End-to-End Encryption (GDPR Compliance):
  - encrypted User conversations before storing using AES-256 with a unique key.
  - Adversarial Defense: Sets behavioral guardrails against: Code Injection and Payload Splitting techniques.

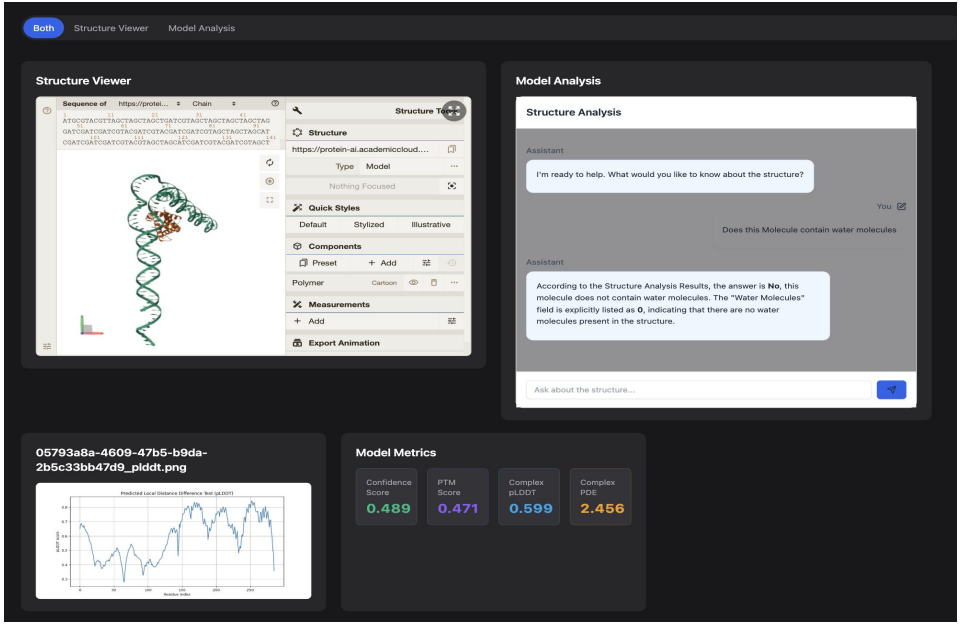


Figure 6: Analysis Dashboard

Test Cases

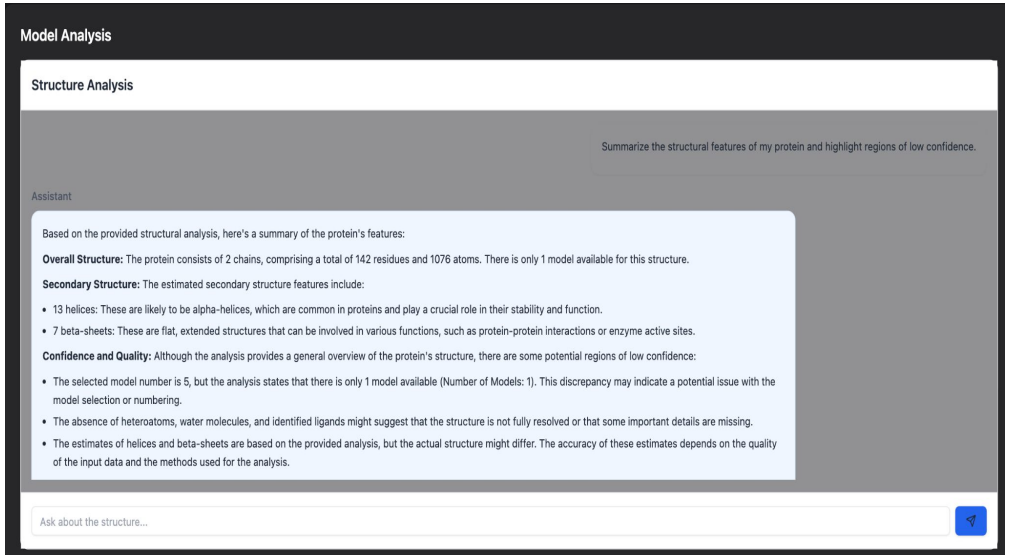


Figure 7: A screenshot showing the LLM chat interface. The user's prompt, "Summarize the structural features of my protein and highlight regions of low confidence."

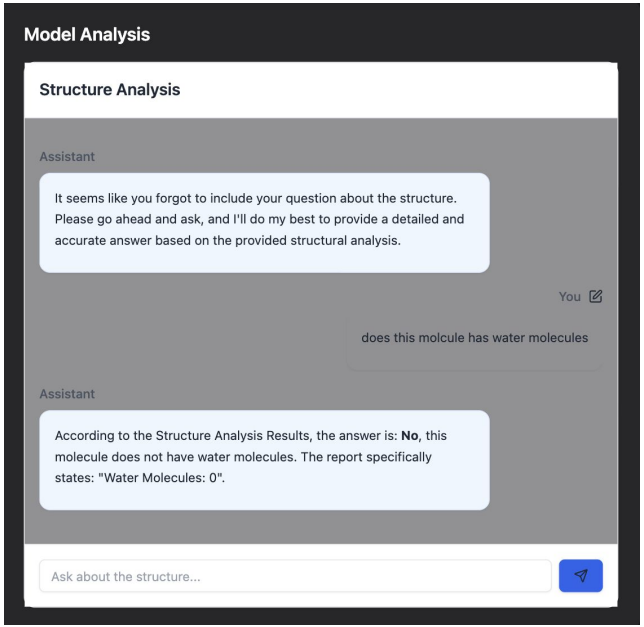


Figure 8: A screenshot of the LLM's response, answering the question "Does this Molecule contain water molecules"

3D Protein Structure

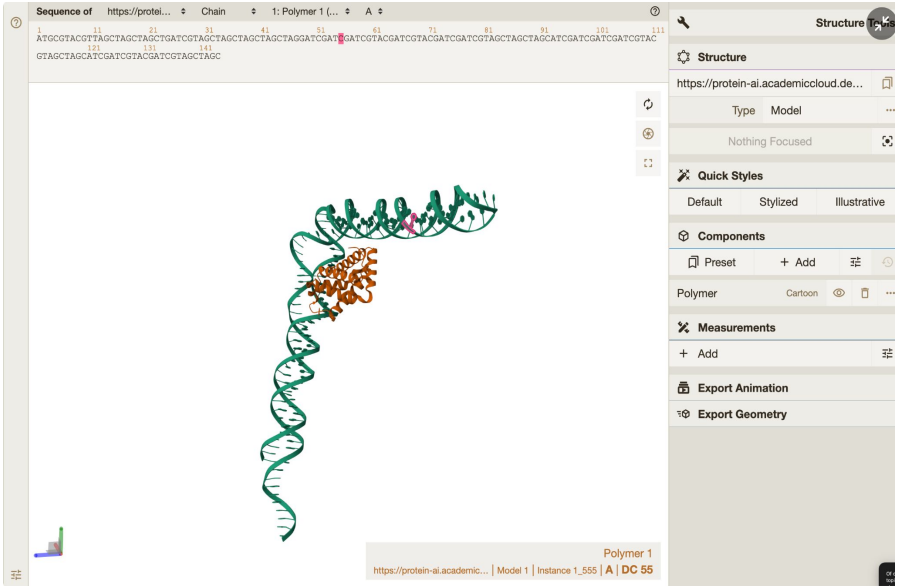


Figure 9: Protein Structure from Boltz

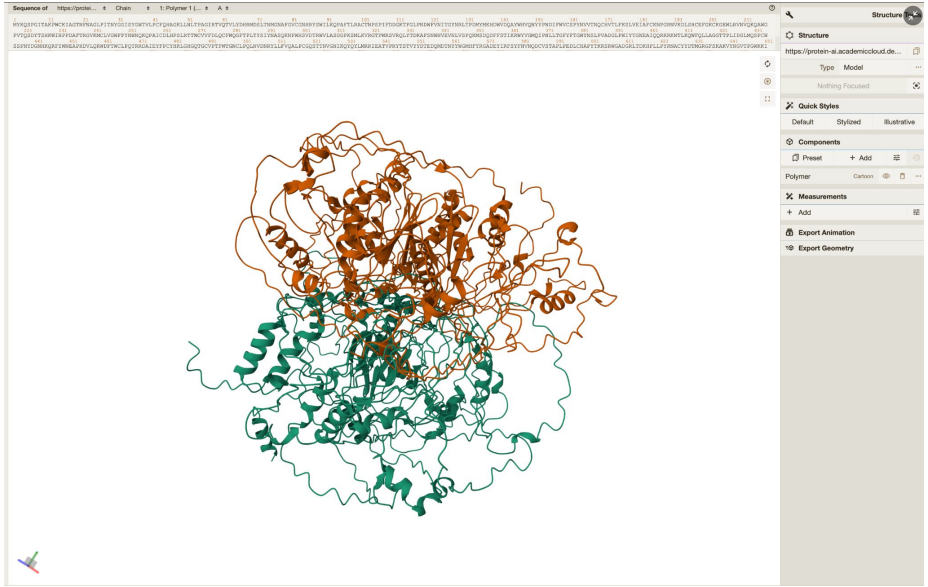


Figure 10: Protein Structure from AlphaFold

Results

- Performance: Reduced job failures, improved web app uptime, efficient job queuing.
  User Experience: Responsive interface, accessible features (dark mode), helpful notes.
- Security & Privacy: Full GDPR compliance, end-to-end encryption, defense against LLM attacks.
- Scientific Accuracy: Real-time input validation (RDKit), comprehensive analysis tools, context-aware LLM.

Future Work: BioSmol

- Moving beyond single models to a Multi-Agent Orchestration Framework.
- Leveraging the Smol-agent library.
  Potential Application: End-to-End Drug Design Pipelines.
- Includes Human-in-the-Loop (HITL) for complex decisions.
- MCP tools for Alpha Fold and Boltz

References

- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). *ColabFold: making protein folding accessible to all*. *Nature Methods*, 19(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … & Hassabis, D. (2021). *Highly accurate protein structure prediction with AlphaFold*. *Nature*, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
- RDKit: Open-source cheminformatics. http://www.rdkit.org
- *Boltz Model Github Repository* . https://github.com/jwohlwend/boltz
- *Mol JavaScript Library**. https://github.com/molstar/molstar
- Next.js Documentation. https://nextjs.org/docs

Acknowledgements and Questions

- Deepest gratitude to supervisors Dr. Narges Lux, Prof. Dr. Julian Kunkel, and the GWDG HPC team.

Website: protein-ai.academiccloud.de

Questions ?