Deciphering structural artifacts: A machine learning framework to differentiate AlphaFold predictions from crystallographic PDB structures

Hadi Vareno[1], Tobias Harren[1], Malte Korn[1], Matthias Rarey[1], and the Student Team*

[1]*University of Hamburg, ZBH - Center for Bioinformatics, 22761, Hamburg, Germany.*

**This work resulted from a student project within the Computing in Science Bachelor Program. The student team includes: Ann Sophie Bäurle, Svea Dosdahl, Torben Kröpke, Lennard Liebsch, Lina Schumacher, Pina Spengler, Mehmet Tulgar and Luis Valentin*

Background: The paradigm "structure determines function" has made the high-throughput prediction of protein structures by AlphaFold [1] a cornerstone of modern structural biology. As AI-generated models increasingly populate scientific databases, understanding the subtle structural discrepancies between these models and gold-standard data from crystallographic experiments available in the Protein Data Bank (PDB) [2] becomes critical for downstream applications like molecular docking and drug design.

Methods: We developed a robust machine learning pipeline to classify protein structures based on their origin (experimental PDB vs. computational AlphaFold). To eliminate biological bias, a "UniProt Bridge" protocol was implemented, which pairs structures with maximal sequence identity via local alignment. A comprehensive library of 155 structural features was engineered, which encompasses Solvent Accessible Surface Area (SASA), backbone geometry (including ϕ and ψ dihedral angles), and side-chain torsion (χ) angles. To isolate algorithm-specific artifacts from general physical instabilities, all structures were subjected to energy minimization using the AMBER ff14SB force field [3].

Results: Initial classification models achieved near-perfect accuracy, primarily driven by systematic backbone biases: 96% of AlphaFold structures exhibited significantly increased *Cα–Cα* bond lengths and expanded *N–Cα–C* bond angles relative to PDB entries. Remarkably, even after energy minimization, which resolved these primary geometric artifacts, the models maintained over 95% classification accuracy. Feature importance analysis identified the variance of $\chi 3$ torsion angles and the packing density of long-chain residues (e.g., Lysine and Glutamine) as the most discriminative factors. Furthermore, we identified a systematic idealization in AlphaFold predictions, characterized by the frequent misclassification of rare π-helices as standard α-helices.

Conclusion: Our findings demonstrate that AlphaFold predictions possess distinct statistical signatures that differentiate them from experimental structures. While highly accurate, AlphaFold reflects algorithm-specific artifacts, particularly regarding side-chain rotamer distributions and rare secondary structure elements. This study highlights specific areas for the future refinement of AI-based structure prediction.

[1] Jumper, J. et al. (2021) Nature, 596, 583-589.

[2] Berman, H.M. et al. (2000) Nucleic Acids Res, 28, 235-242.

[3] Maier, J.A. et al. (2015) J. Chem. Theory Comput., 11, 3696-3713.