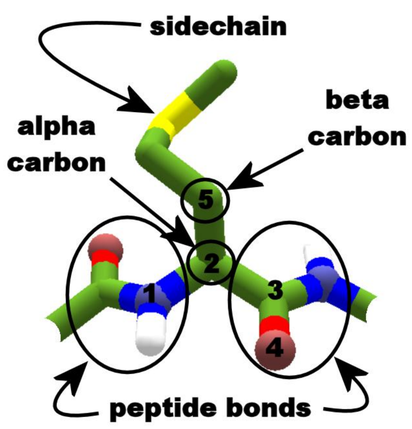
A Complete and Bi-Continuous Invariant of Protein Backbones

Ziqiu Jiang, Olga Anosova, and Vitaliy Kurlin
Materials Innovation Factory, University of Liverpool, Liverpool, UK

Challenges of Protein Comparison

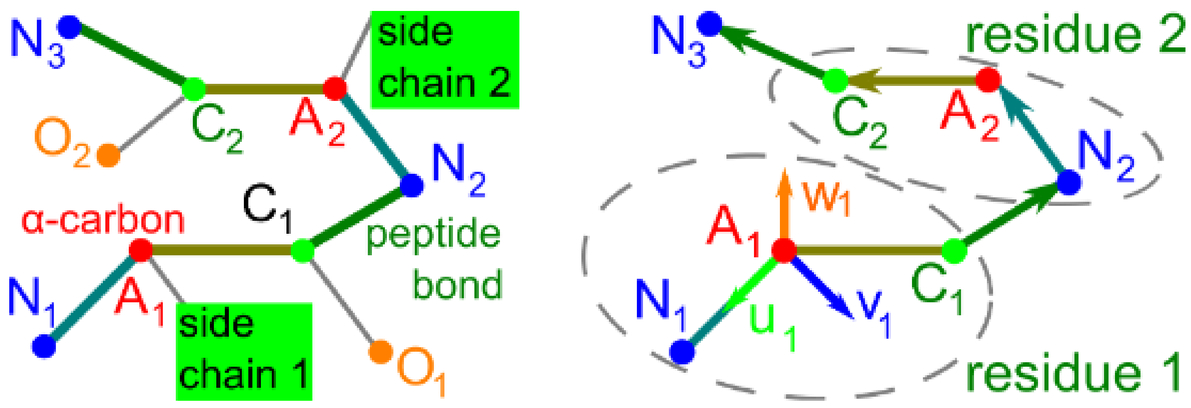
A protein's functional properties depend on its tertiary structure. Standard metrics for comparing protein shapes possess limitations:

TM-score and LDDT fail basic metric axioms, which could lead to pre-determined results for clustering algorithms [1, 2].
RMSD is too slow for all pairs in large databases such as the Protein Data Bank (PDB).
We designed a complete linear-time invariant that uniquely determines any backbone structure under rigid motion.



The Backbone Rigid Invariant (BRI)

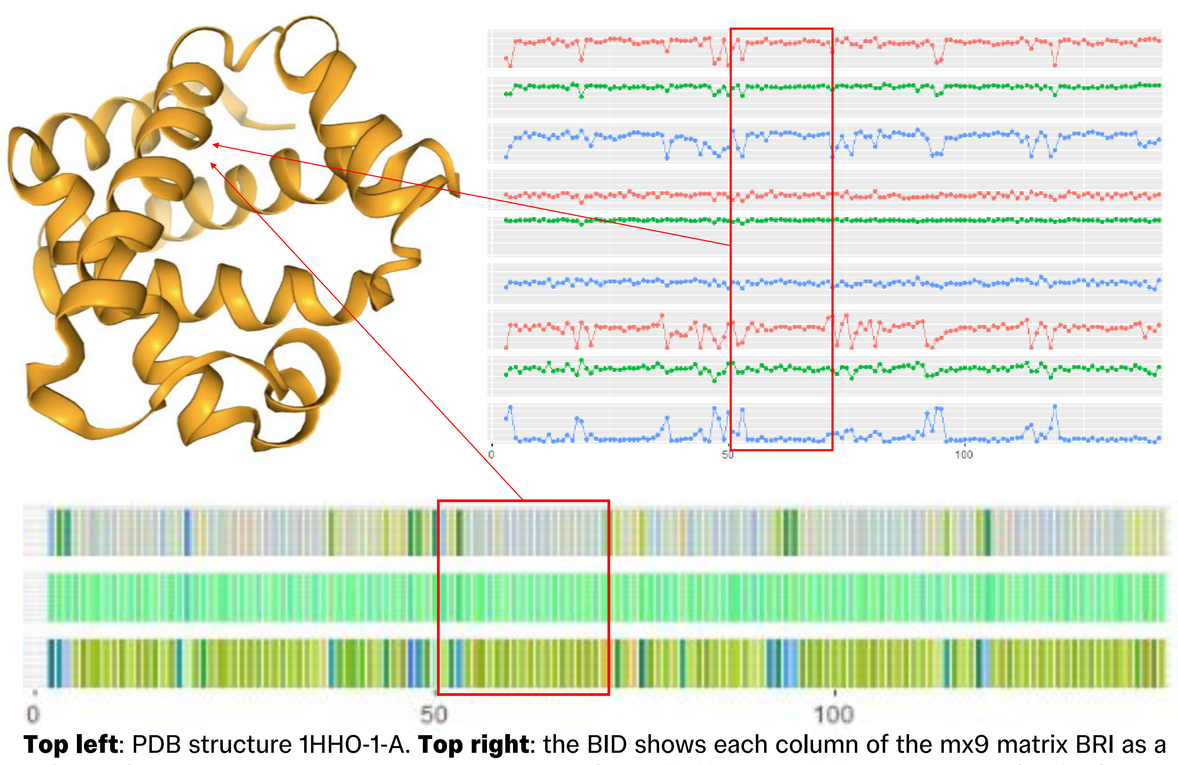
For any protein backbone S of m residues, the Backbone Rigid Invariant BRI(S) is an $\mathbf{m \times 9}$ matrix describing three main atoms (nitrogen, alpha-carbon, and carbonyl carbon) in a basis associated with the previous residue [3].



The computational time $O(m)$ of BRI is linear (proportional to the number m of residues), outperforms quadratic-size distance matrices and avoids unstable alignment-based RMSD.

BRI Visualization

The Backbone Invariant Diagram (BID) visualises the BRI matrix via 9 invariant curves whose values can be converted into the Backbone Invariant Barcode (BIB) of colour bars along the residue indices, enabling a fast detection of alpha-helices.



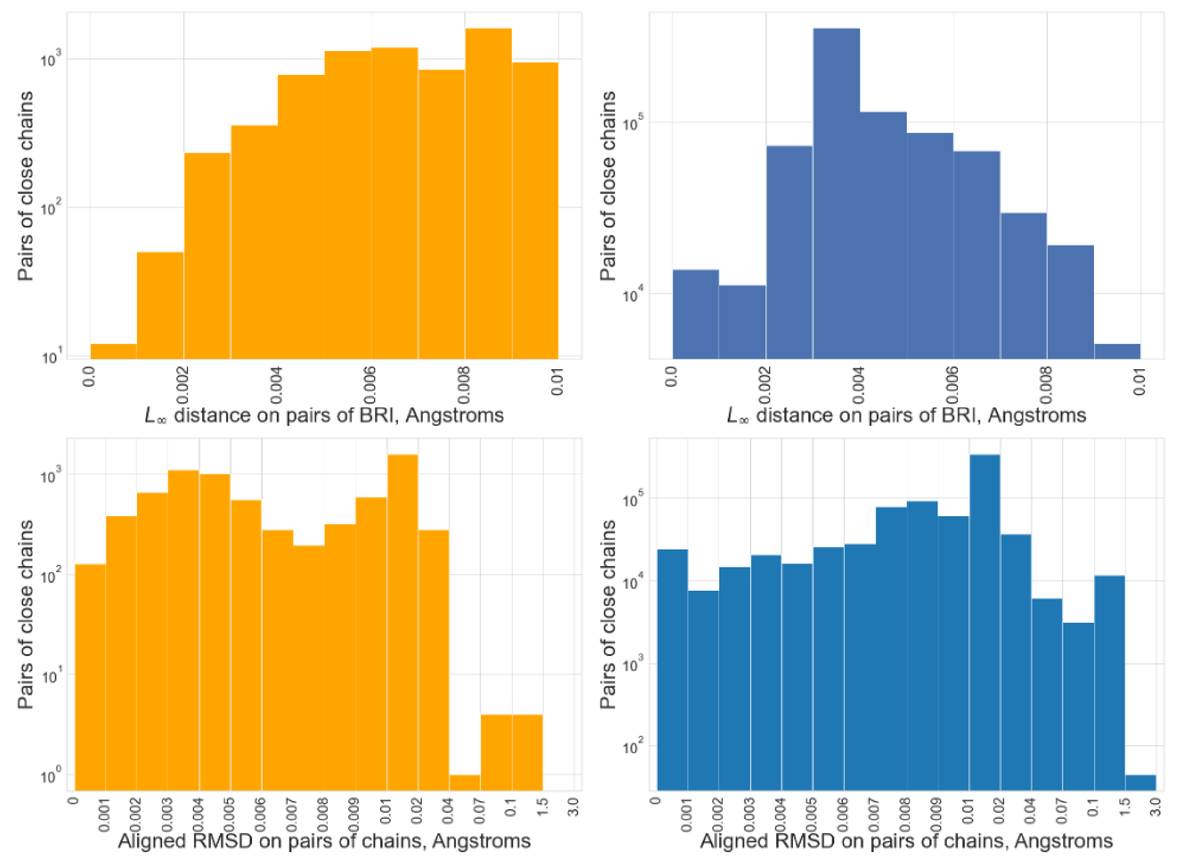
Top left: PDB structure 1HHO-1-A. **Top right**: the BID shows each column of the mx9 matrix BRI as a polygonal curve vs residue indices on the horizontal axis. **Bottom**: the BIB converts a triple of columns from BRI into one RGB coloured bar. Stable intervals in BIB and BID represent alpha-helices.

Results & PDB Analysis

- **Detecting Duplicates**

Within 6 hours on a modest desktop computer, we cleaned the PDB to 707K+ chains with full geometric data and no disorder, computed all BRI matrices, compared all-vs-all chains and found 13,907 pairs of rigid identical backbones, including 9,366 pairs with all identical x, y, z coordinates and 763 pairs from different PDB entries, often with different sequences, highlighting previously undetected redundancy and errors in the PDB [4]

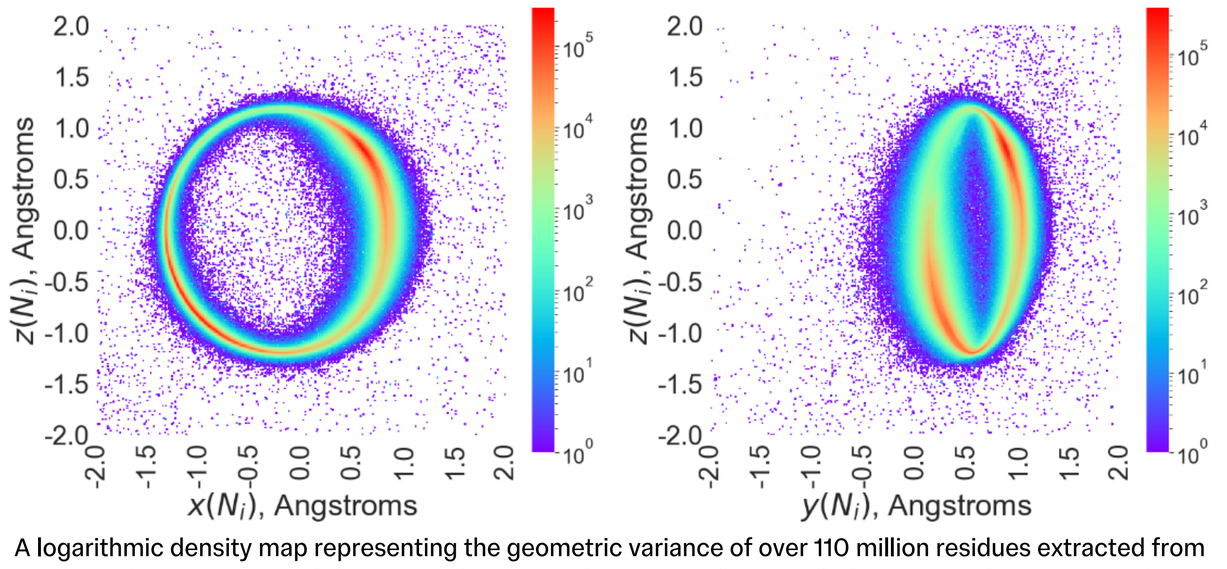
PDB id1 & chain	method and resolutions, Å	PDB id2 & chain	all atoms have identical x, y, z	different residues
1a0t-B	X-ray, 2.4, 2.4	1oh2-B	all 3×413	9
1ce7-A	X-ray, 2.7, 2.7	2mll-A	all 3×241	1, GLY≠HIS
1ruj-A	X-ray, 3, 3	4rhv-A	all 3×237	1, GLY≠SER
1gli-B/D	X-ray, 2.5, 1.7	3hhb-B/D	all 3×146	1, MET≠VAL
2hqe-A	X-ray, 2, 2	2o4x-A	all 3×217	1, GLN≠GLU
5adx-T	EM, 4, 8.2	5afu-Z	all 3×165	1, ILE≠VAL
5lj3-O	EM, 3.8, 10	5lj5-P	all 3×252	1, ALA≠VAL
8fdz-A	X-ray, 2.5, 2.2	8fe0-A	all 3×200	1, THR≠SER



Logarithmic histograms of near-duplicate chains of the same length. **Top row**: 783075 pairs with $L_\infty \leq 0.01$Å on BRIs including 13907 pairs of exact duplicates with $L_\infty = 0$. **Bottom row**: the same pairs with traditional RMSD. **Left column**: 7151 pairs with different sequences of amino acids. **Right column**: 775852 pairs with identical sequences.

- **Structural Variability**

The invariants revealed substantial variations in the geometry of residues across the PDB and challenged AlphaFold2 [5] assumptions that all residue triangles of backbone atoms have identical rigid shapes.



A logarithmic density map representing the geometric variance of over 110 million residues extracted from the PDB. Contrary to the rigid geometric assumptions coded into predictive models like AlphaFold2, the spatial dispersion demonstrates substantial natural variability in local residue triangle configurations.

Reference

[1] Y. Zhang, J. Skolnick, "Scoring function for automated assessment of protein structure template quality," Proteins: Structure, Function, and Bioinform. 57 (2004) 702–710.
[2] V. Mariani, et al., "lddt: a local superposition-free score for comparing protein structures and models using distance difference tests," Bioinformatics 29 (2013) 2722–2728.
[3] O. Anosova et al., "A Complete and Bi-Continuous Invariant of Protein Backbones under Rigid Motion," Match Communications in Mathematical and in Computer Chemistry, 94 (2025) 97–134.
[4] A. Wlodawer et al., "Duplicate entries in the Protein Data Bank: How to detect and handle them," en, Acta Crystallographica Section D: Structural Biology, 81 (2025) 170–180.
[5] J. Jumper et al., "Highly accurate protein structure prediction with alphafold," Nature, 596 (2021) 583–589.