Fast lossy protein structure compression algorithm

https://github.com/steineggerlab/foldcomp

Hyunbin Kim[1], Johannes Söding[2], Martin Steinegger[1]

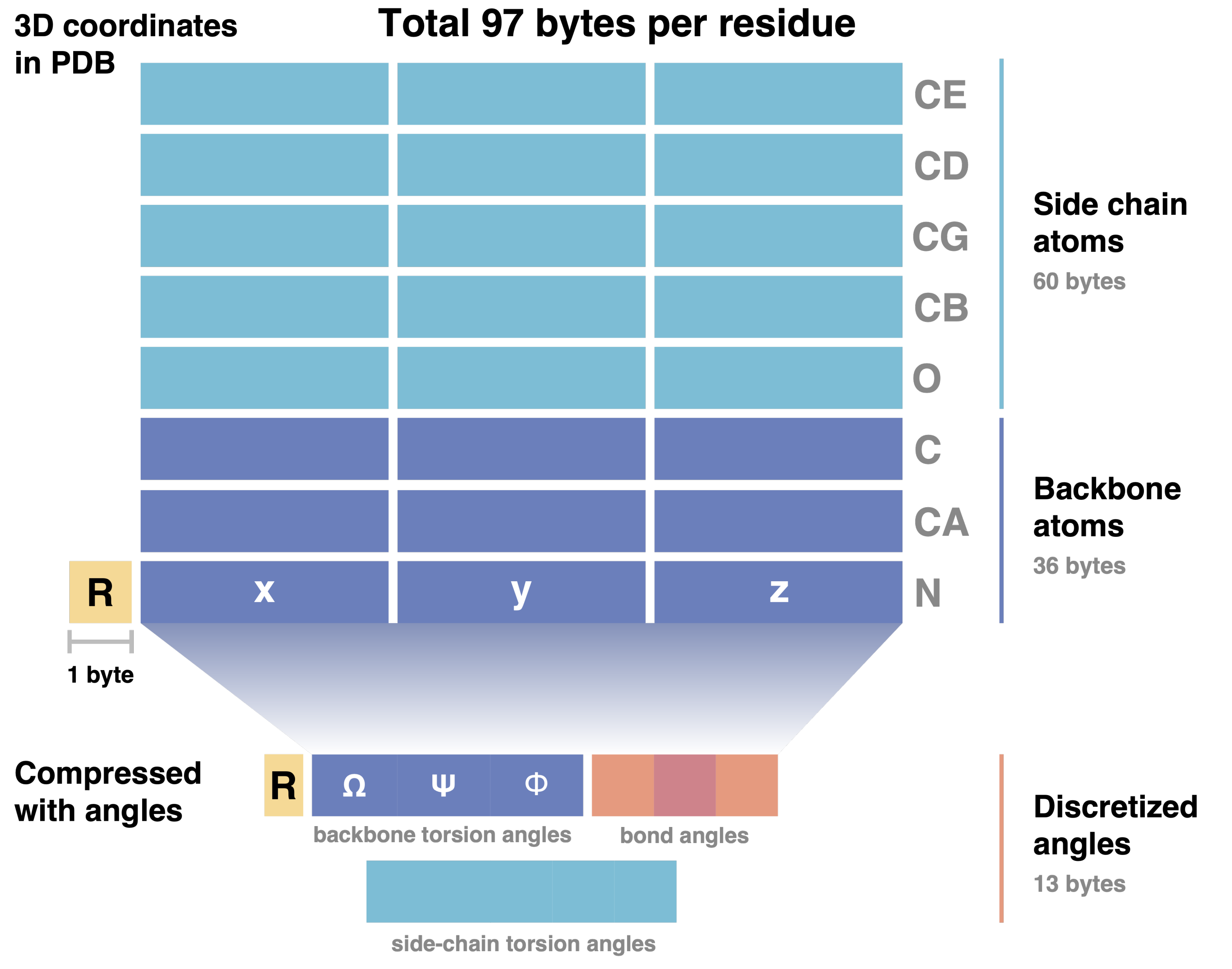
[1]Seoul National University, South Korea; [2]Max Planck Institute, Germany; martin.steinegger@snu.ac.kr

ABSTRACT

AlphaFold2 produces structure predictions at high quality and speed. EMBL and DeepMind have announced to soon release a database containing over 100 million predicted structures covering the UniRef90. Thus, a future with billions of predicted structures is soon imaginable. Additionally, the prediction speed is constantly improved. e.g., ColabFold is ~100x faster compared to baseline AF2.

However, with advances in speed, storing all the structures is becoming a major issue. Storing the structure of a protein with 250 residues in PDB format takes approx. 200 kilobytes (only 3D coordinates 25 kilobytes), thus one billion structures would require hundreds of terabytes.

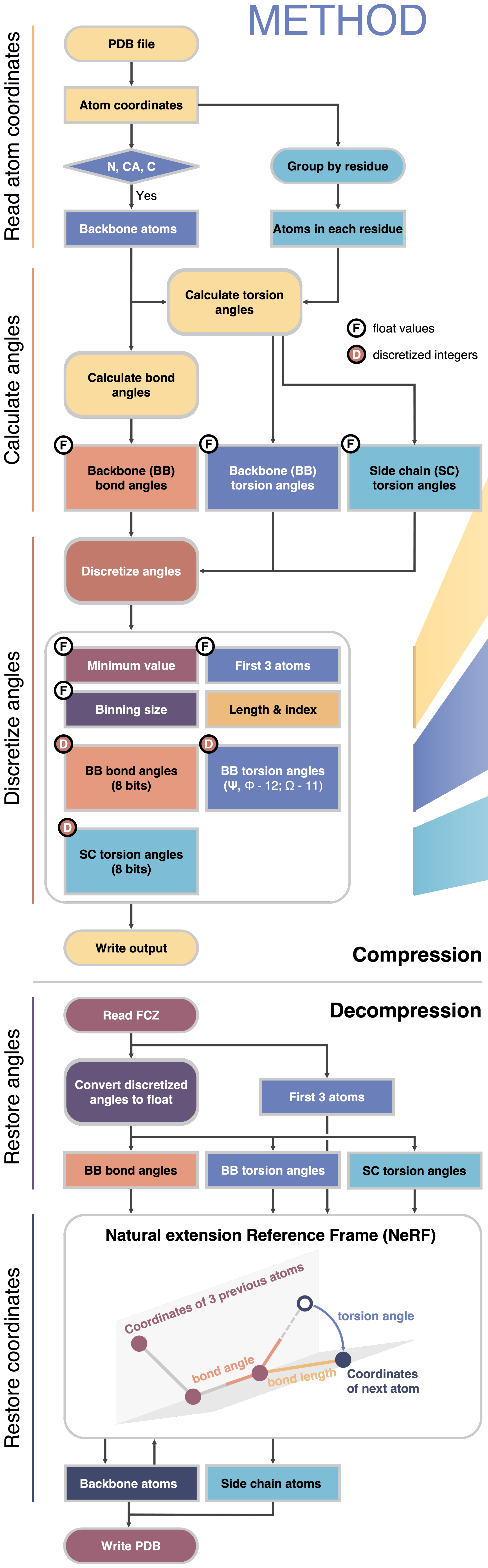
Here, we propose a novel format and method to compress protein structures requiring only 10 kilobytes for a protein structure of average size (4.8 kb for coordinates), reducing the required storage space by an order of magnitude.



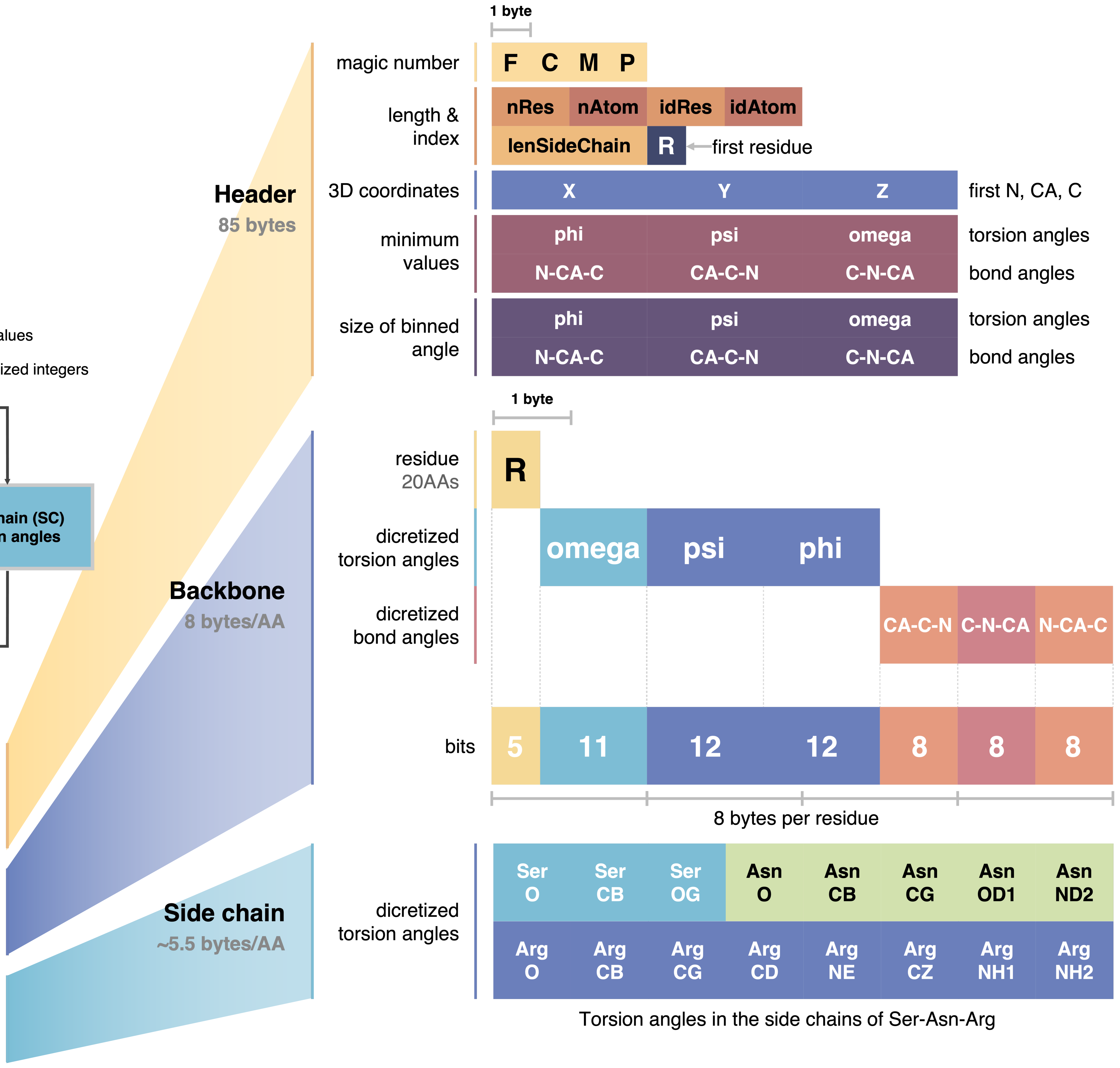
13 bytes per residue

We achieve this reduction by efficiently encoding the torsion angles of the backbone as well as the side-chain angles in a compact format. We show that using our lossy compression has no impact on structural downstream analysis. By storing angles with an optimized bit-format, we can reduce the storage required by 90% compared to float-encoded 3D coordinates, while maintaining a high compression and decompression speed.

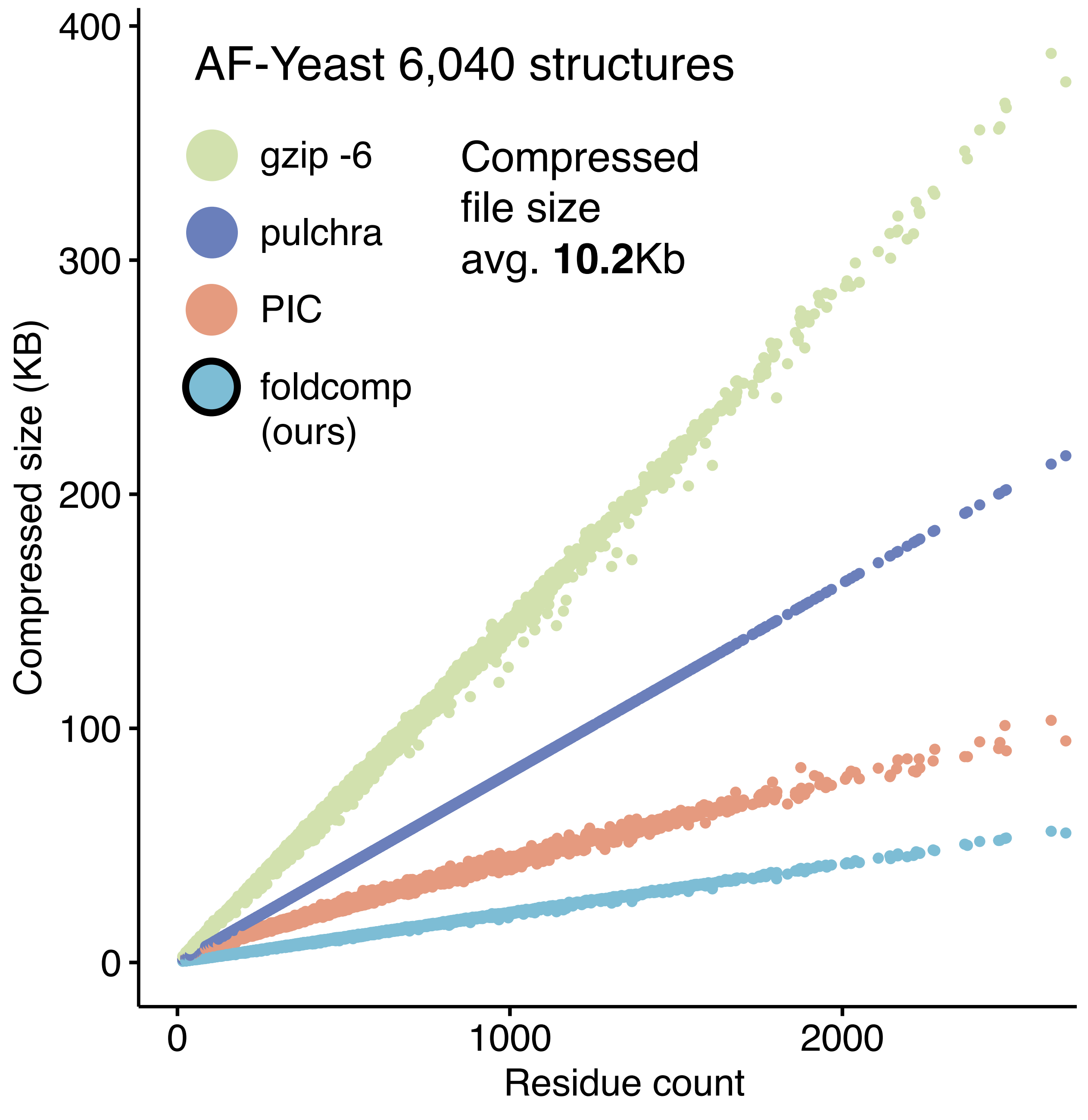
METHOD

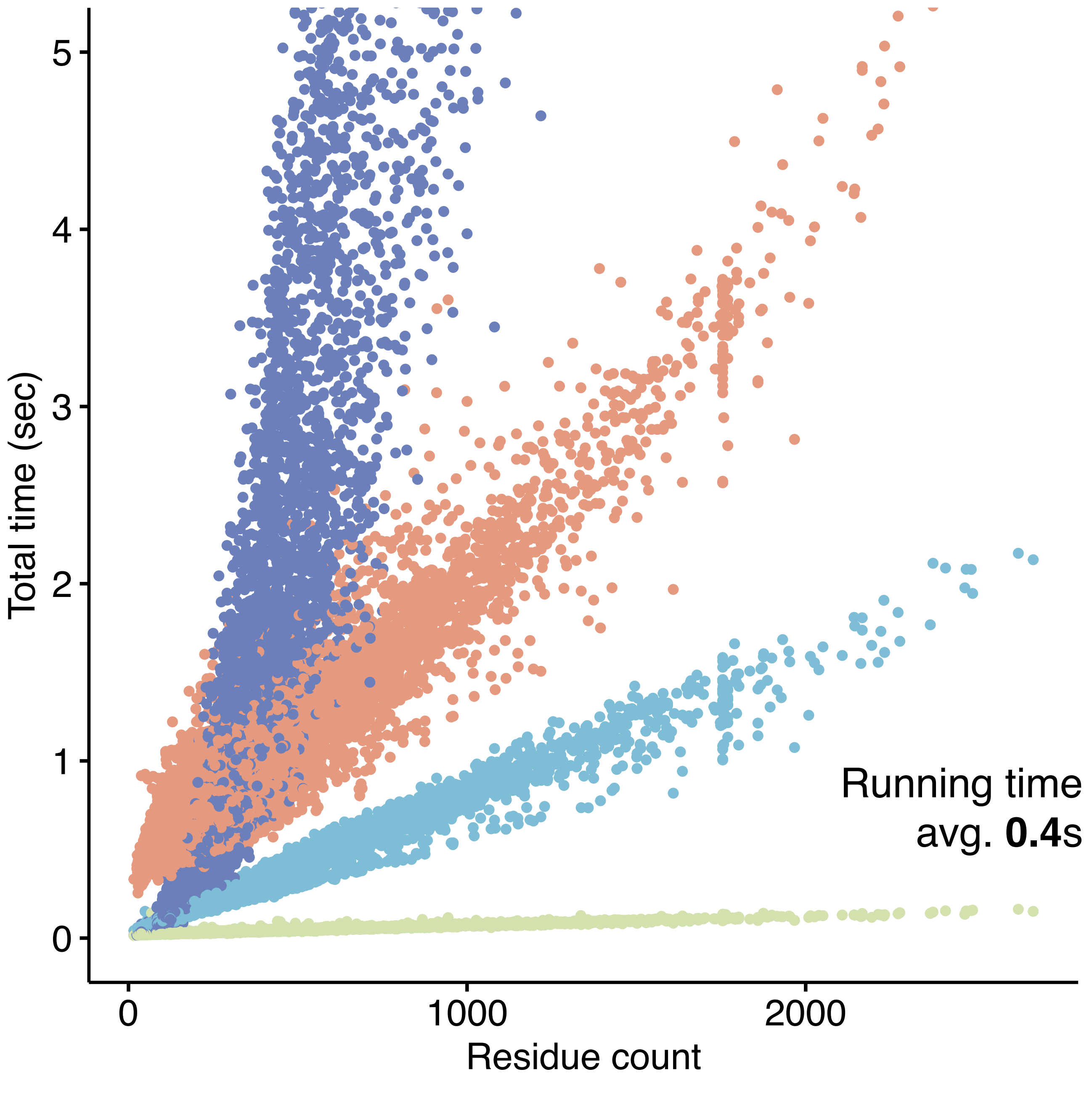


FILE STRUCTURE



BENCHMARK RESULT





Loss in compression

PDB ID	Tool	RMSD
1a0fA	foldcomp	**0.227**
	pic	18.976
	pulchra	3.208
1a0aA	foldcomp	**0.154**
	pic	20.688
	pulchra	3.343
1a0p_	foldcomp	6.744
	pic	19.091
	pulchra	**3.476**
1a0i_	foldcomp	4.241
	pic	21.420
	pulchra	**3.370**
1a0tP	foldcomp	**0.443**
	pic	20.958
	pulchra	3.120

5 randomly selected PDB files