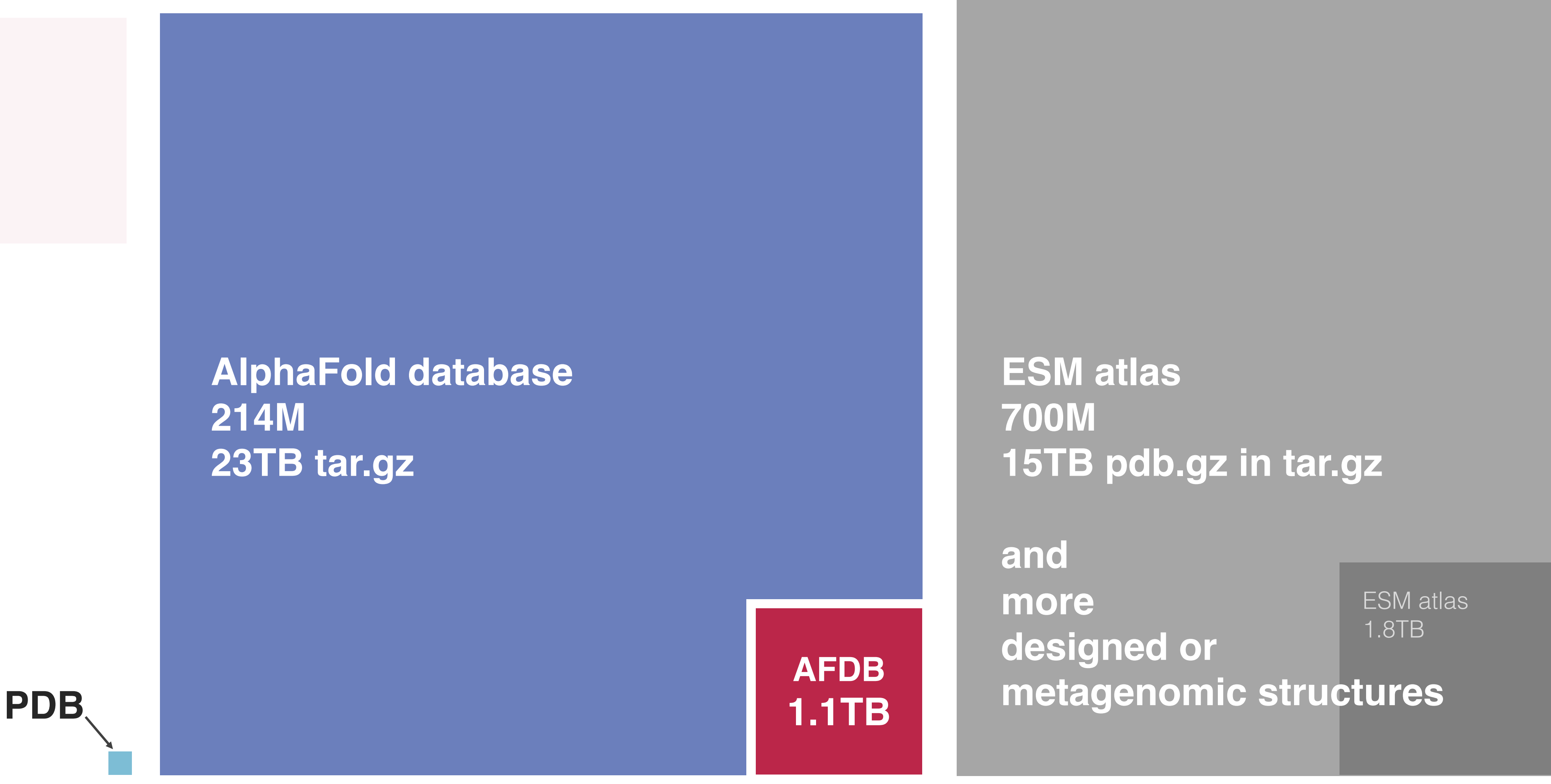
Foldcomp: scalable solution for compressing huge protein structure database

Hyunbin Kim
Milot Mirdita
Martin Steinegger

Seoul National University
South Korea

We compressed AlphaFold Database (23TB) into 1.1TB



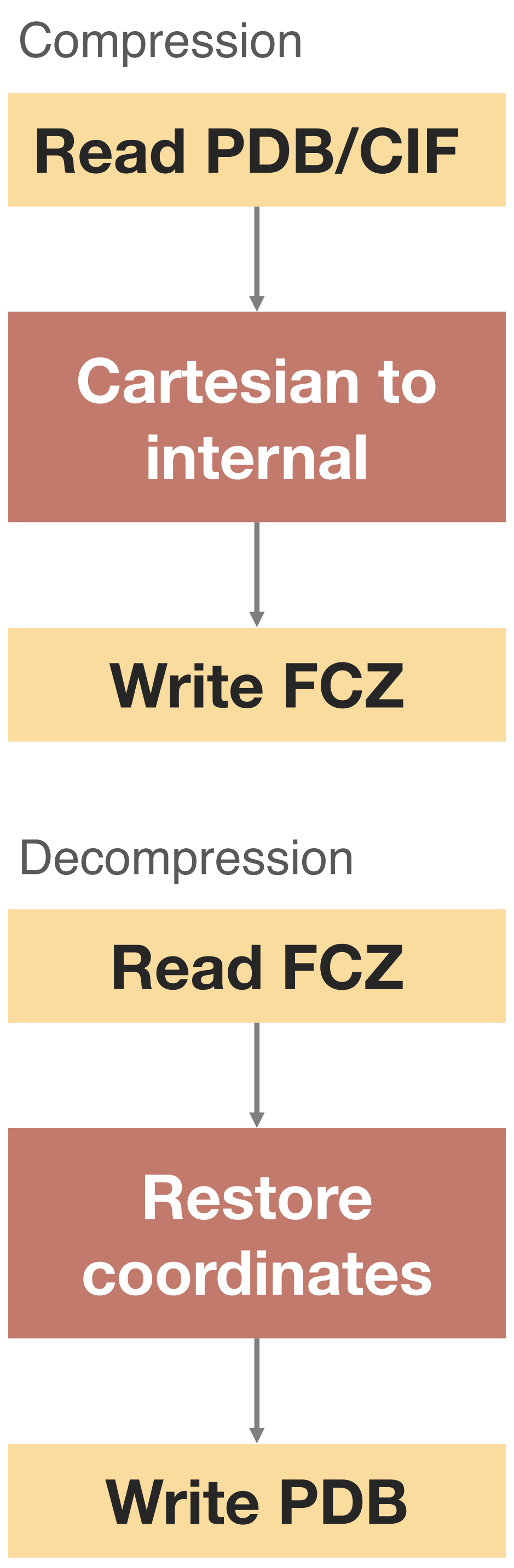
Advent of unprecedentedly large protein structure databases

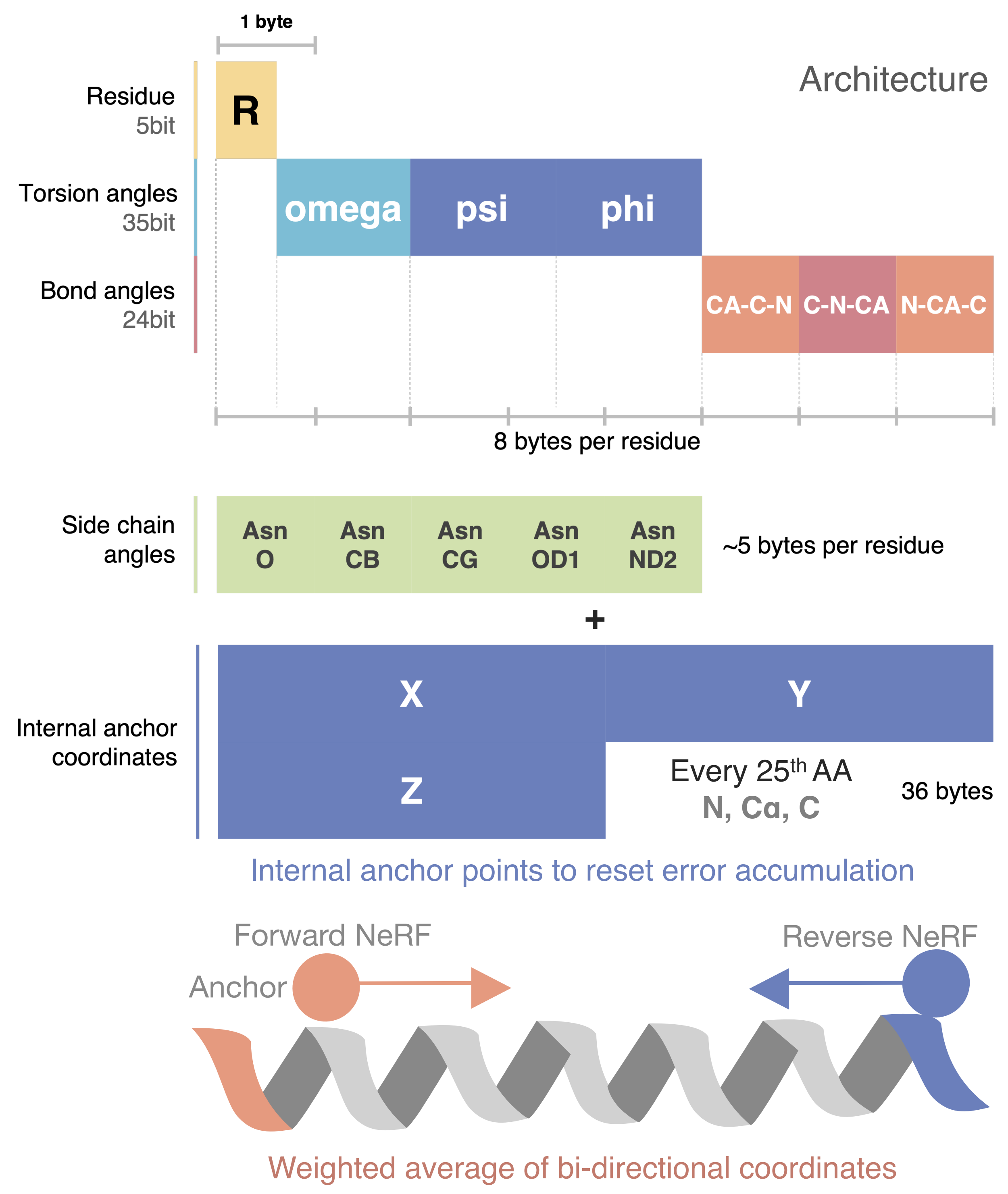
The AlphaFold databases of 214M UniProt proteins and the ESMatlas catalog of nearly 700M metagenomic protein structures provide valuable resources to the community. However, their extensive sizes of 23TB and 15TB, respectively, exceed the capacity of standard workstations and pose a challenge even to well-equipped cluster environments.

Novel compression format of torsion angles

We introduce **Foldcomp**, a novel compression algorithm that encodes the torsion and bond angles in a compact binary format, named FCZ. Foldcomp achieves up to 90% compression compared to float-encoded 3D coordinates, requiring only 13 bytes per residue.

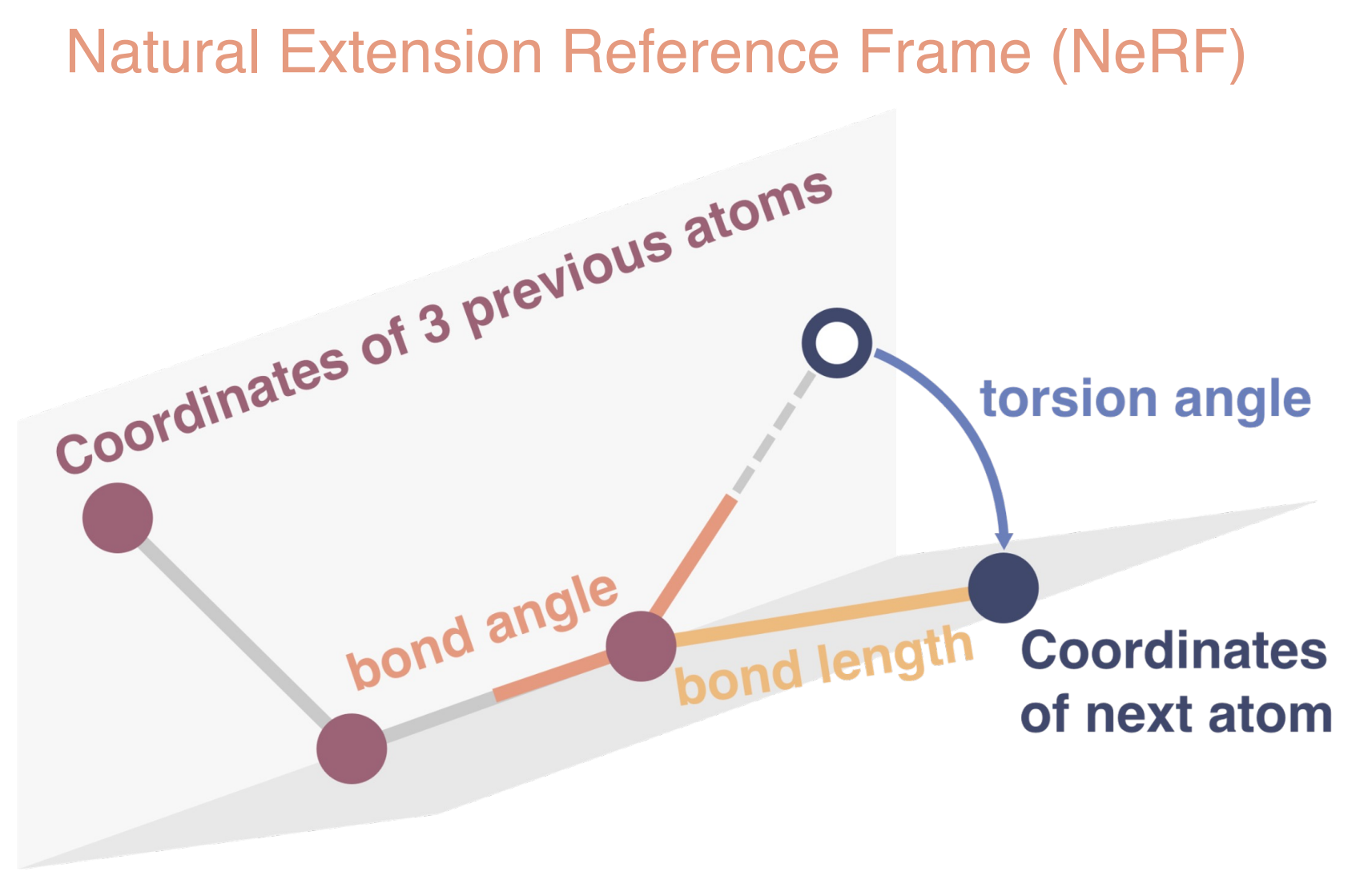
90% 13byte/AA 7.7kb/structure





Reducing loss to less than experimental error

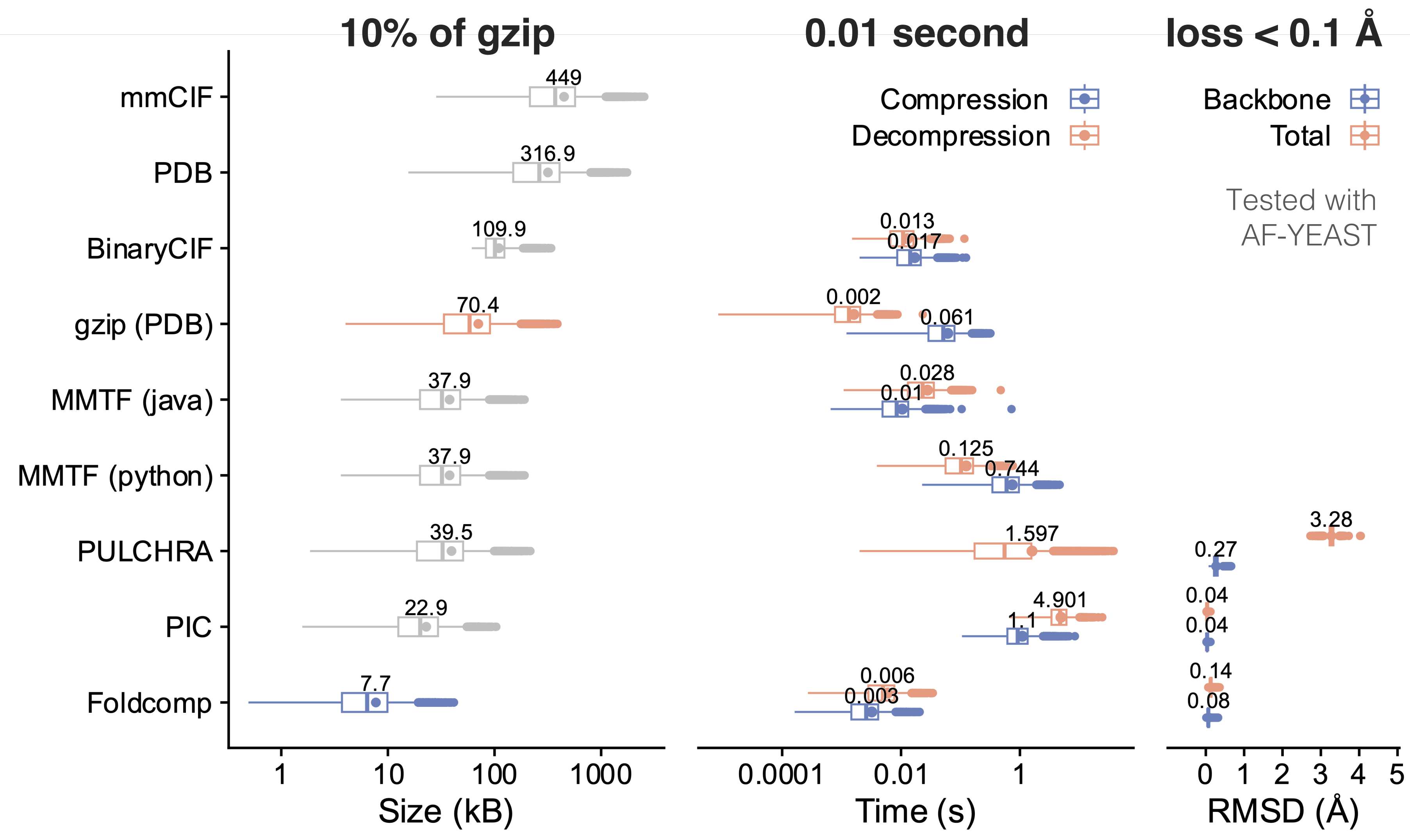
Reconstruction of original coordinates is accomplished by utilizing the NeRF algorithm with internal anchor points. By averaging bi-directional reconstructed coordinates, we were able to reduce reconstruction loss to ~0.08Å range.


Natural Extension Reference Frame (NeRF)
Coordinates of 3 previous atoms
torsion angle
bond angle
bond length
Coordinates of next atom

As fast as gzip

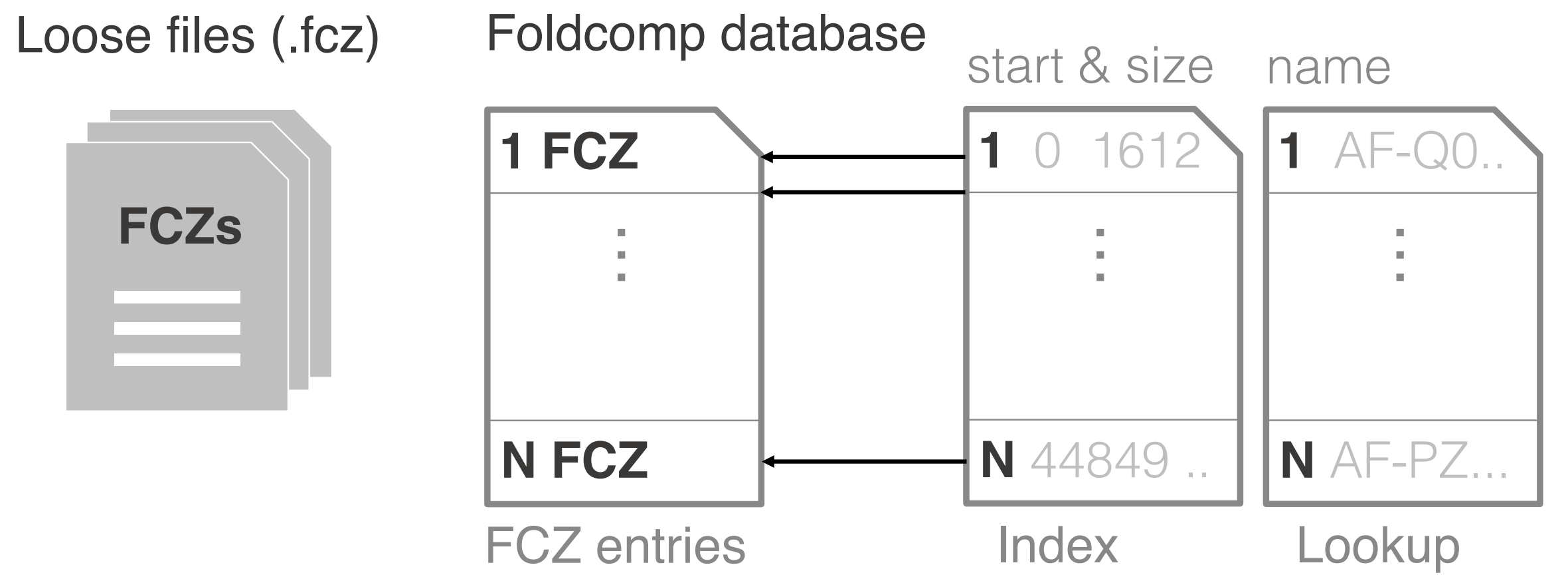
Our method is as fast as gzip, with 3ms and 6ms for compression and decompression, respectively.

10ms 0.08Å of RMSD



DB format for less overhead

MMseqs2 database format reduced unnecessary padding bytes and overhead from file numbers



Foldcomp is publicly available

as a command line interface and a Python API at https://foldcomp.foldseek.com.

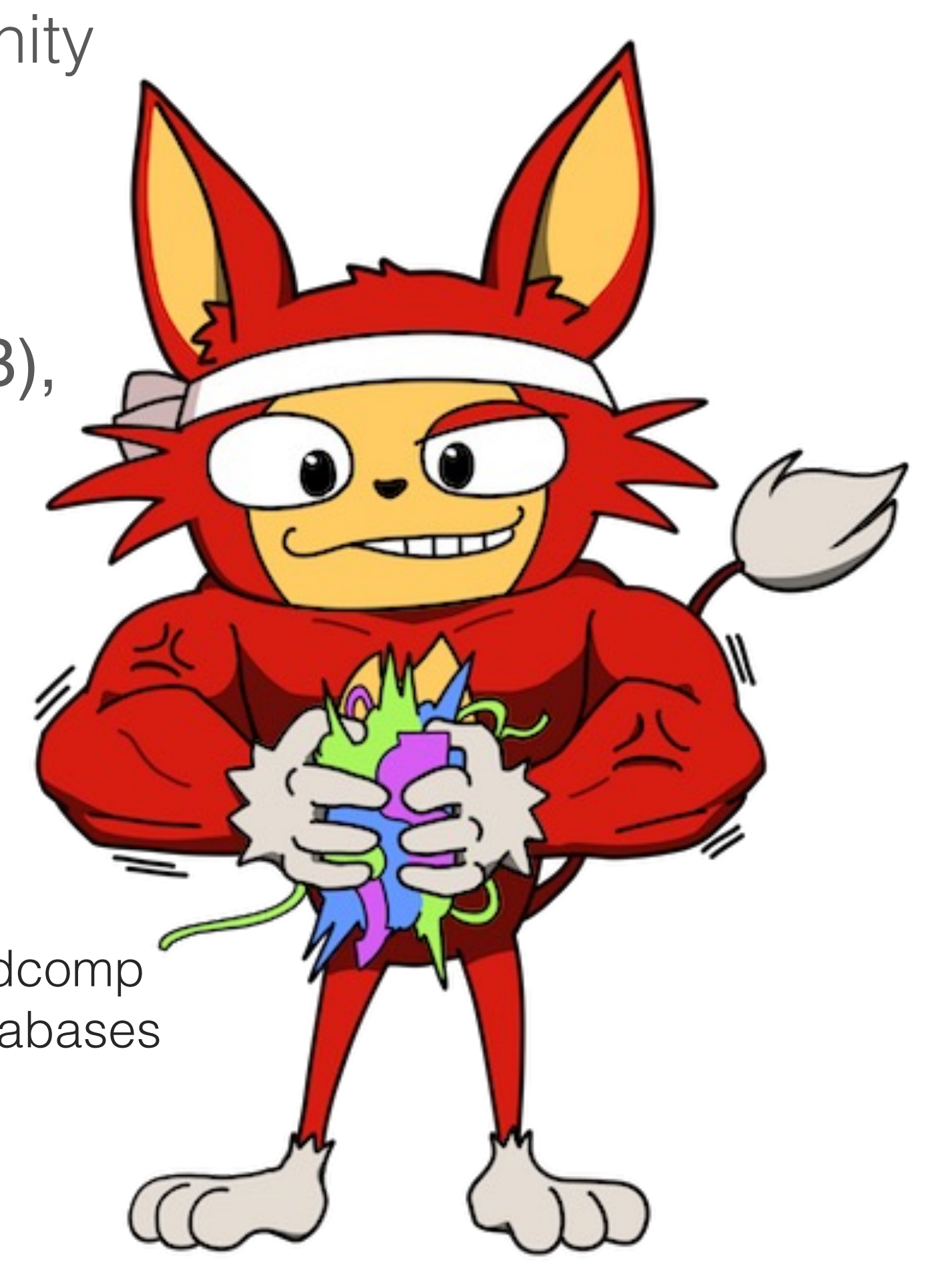
pip install foldcomp

Foldcomp is compatible with Foldseek and additionally, has been augmented by community contributions, such as a PyMol plugin and a dataset wrapper in Graphein.

We provide the compressed database of AlphaFold database (1.1TB), ESMatlas (1.8TB), SwissProt (2.9GB), and recently released AlphaFold2 cluster representatives (2.2GB) at https://foldcomp.steineggerlab.workers.dev.

Foldcomp Github repository

Foldcomp Databases



References

AlphaFold database – Varadi et al., *Nucleic acids research* (2022).
ESMatlas – Lin et al., *Science* (2023).
Foldseek - van Kempen et al., *Nature biotechnology* (2023).
MMseqs2 - Steinegger and Söding., *Nature biotechnology* (2017).
NeRF - Parsons et al., *Journal of computational chemistry* (2005).
Foldcomp is published at https://doi.org/10.1093/bioinformatics/btad153.