Atomic density distributions in proteins: structural and functional implications

Sotirios Touliopoulos, Nicholas M. Glykos

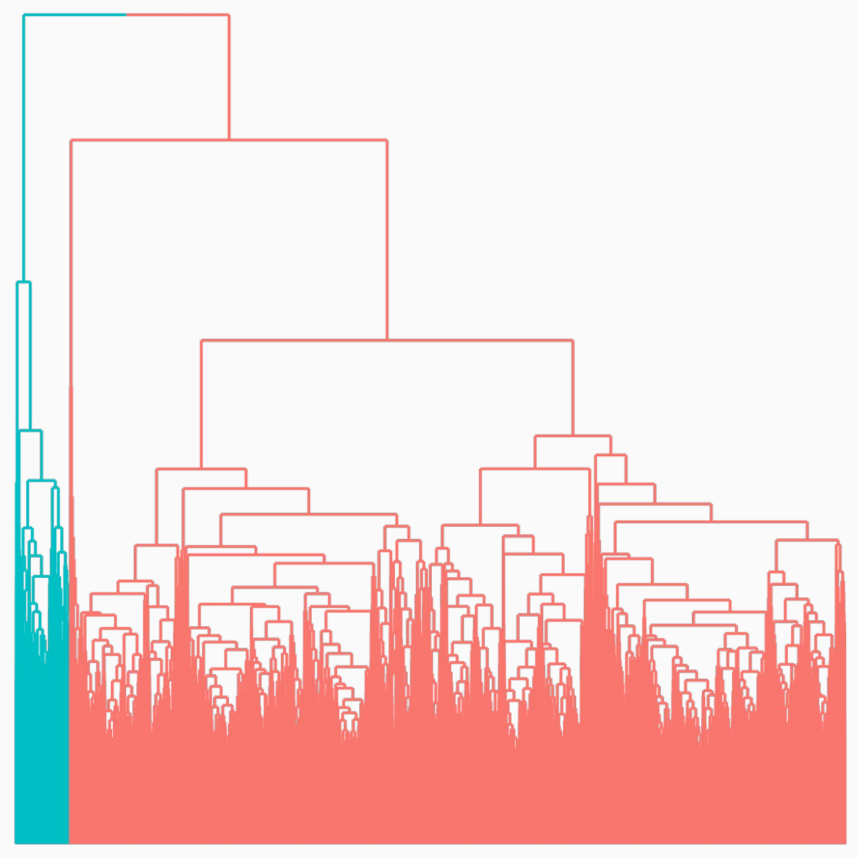
Structural and Computational Biology Laboratory, Dept. of Molecular Biology and Genetics, DUTH
email: s.v.touliopoulos@gmail.com, glykos@mbg.duth.gr

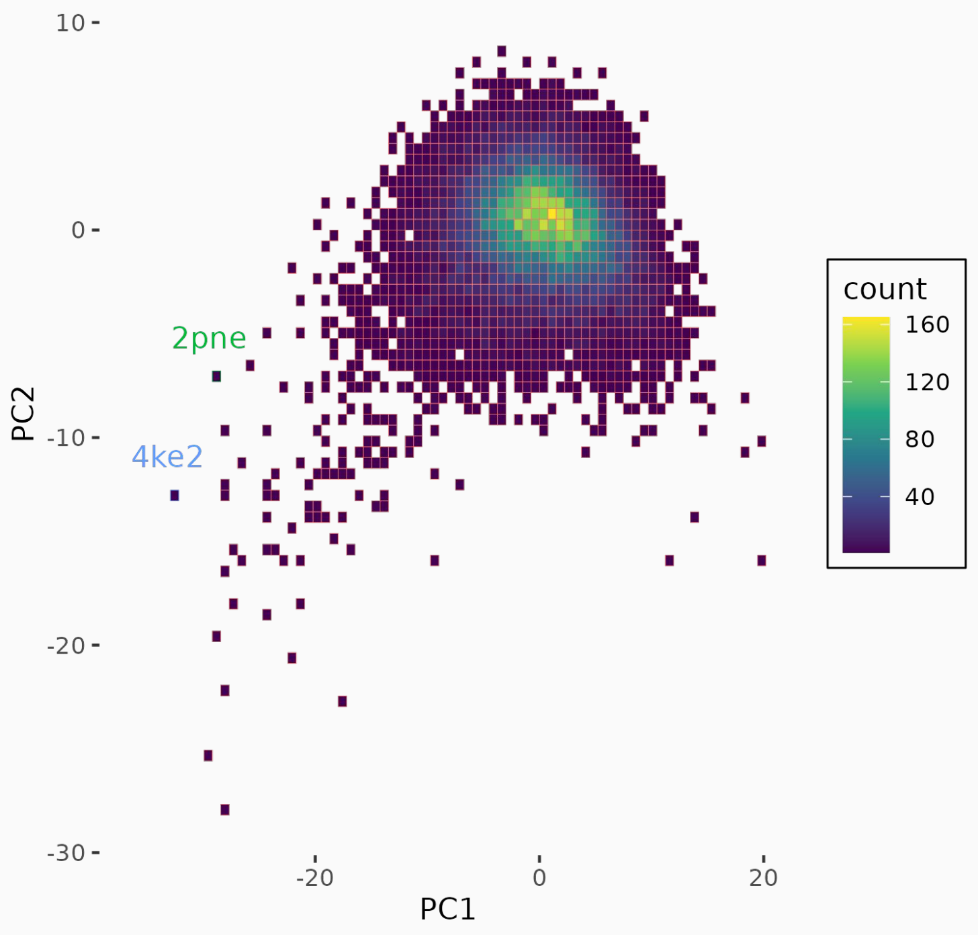
Introduction

Atomic density is the number of atoms per Å^3 (atoms / Å^3). It is calculated by dividing the sum of atomic masses of atoms inside a hypothetical sphere, with the volume of the sphere. Each sphere comes with a certain radius in Angstrom values (Å), with the atom's position as the center. Atomic density in a protein structure is a measure of proximity between protein's atoms. A protein's atomic density distribution shows how well packed is a structure and it may include information on potentially identifying proteins with special folding patterns. In this preliminary report we examine atomic density distributions derived from 21.300 protein structures and show that statistically significant differences between those distributions are present. Several protein structures deviate significantly and systematically from the average behaviour and —not unexpectedly— include proteins with characteristic structures.

Clustering & Dimension reduction

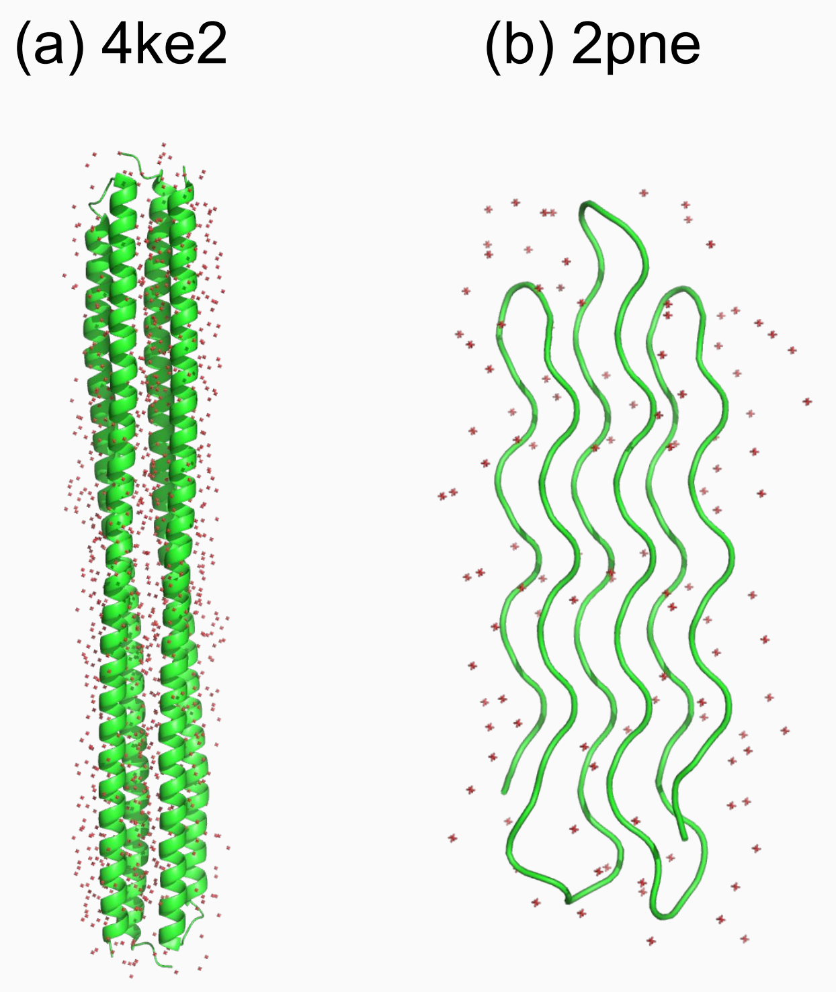
Hierarchical clustering of the raw data showed that a far distinct cluster with uncommon distributions exists (colored in cyan). Extraction of 2 structures from this cluster (PDB ids: 4ke2, 2pne), followed by Principal Component Analysis (PCA) of the raw data validated that these structures are outliers. They appear as separate data points in the 2D density plot.





Uncommon Structures

Examination of the 2 outlier protein structures from the distinct cluster with Pymol software [4], revealed some early implications of atomic density. The first structure (a) is an alanine-rich 4-helices bundle protein that retains ~400 water molecules in its core (PDB id: 4ke2). The other structure (b) is a glycine-rich protein, made up of six antiparallel left-handed polyproline type II helices (PDB id: 2pne). Both molecules are antifreeze proteins. Antifreeze proteins are a class of proteins that adsorb to the surface of ice crystals to prevent their growth. They are critical to the existence of life at sub-zero temperatures [5]. These proteins are also very compact, as depicted in their distributions.



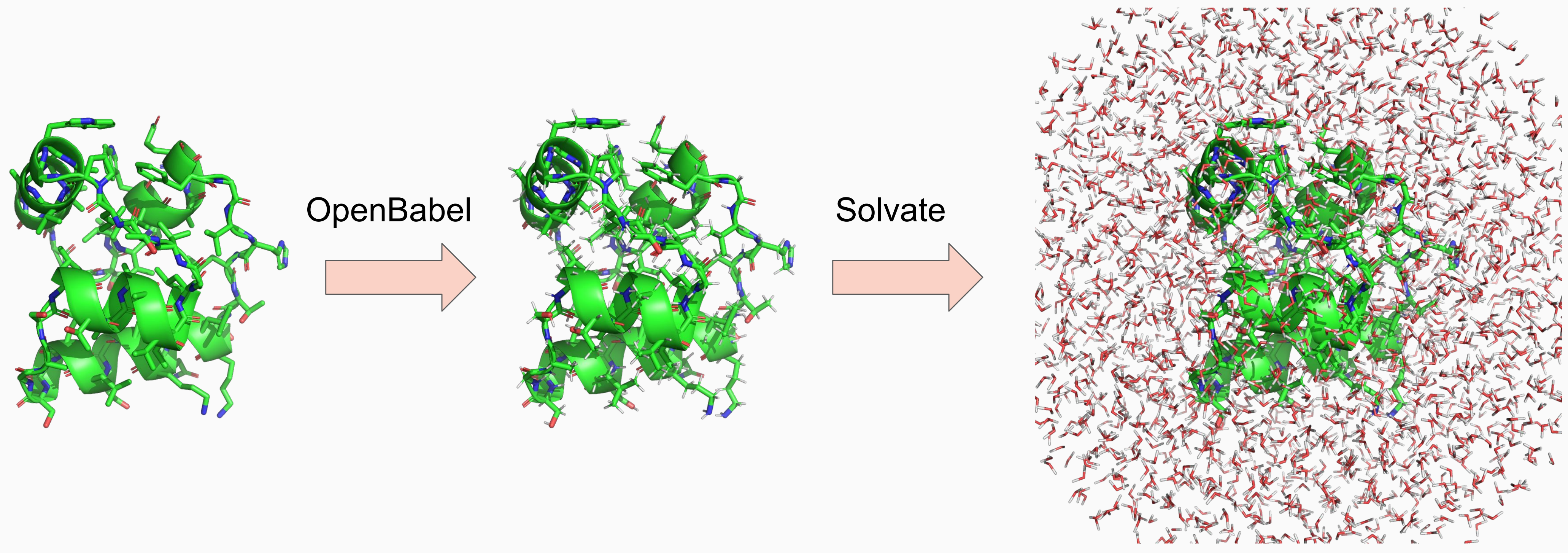
Future Work

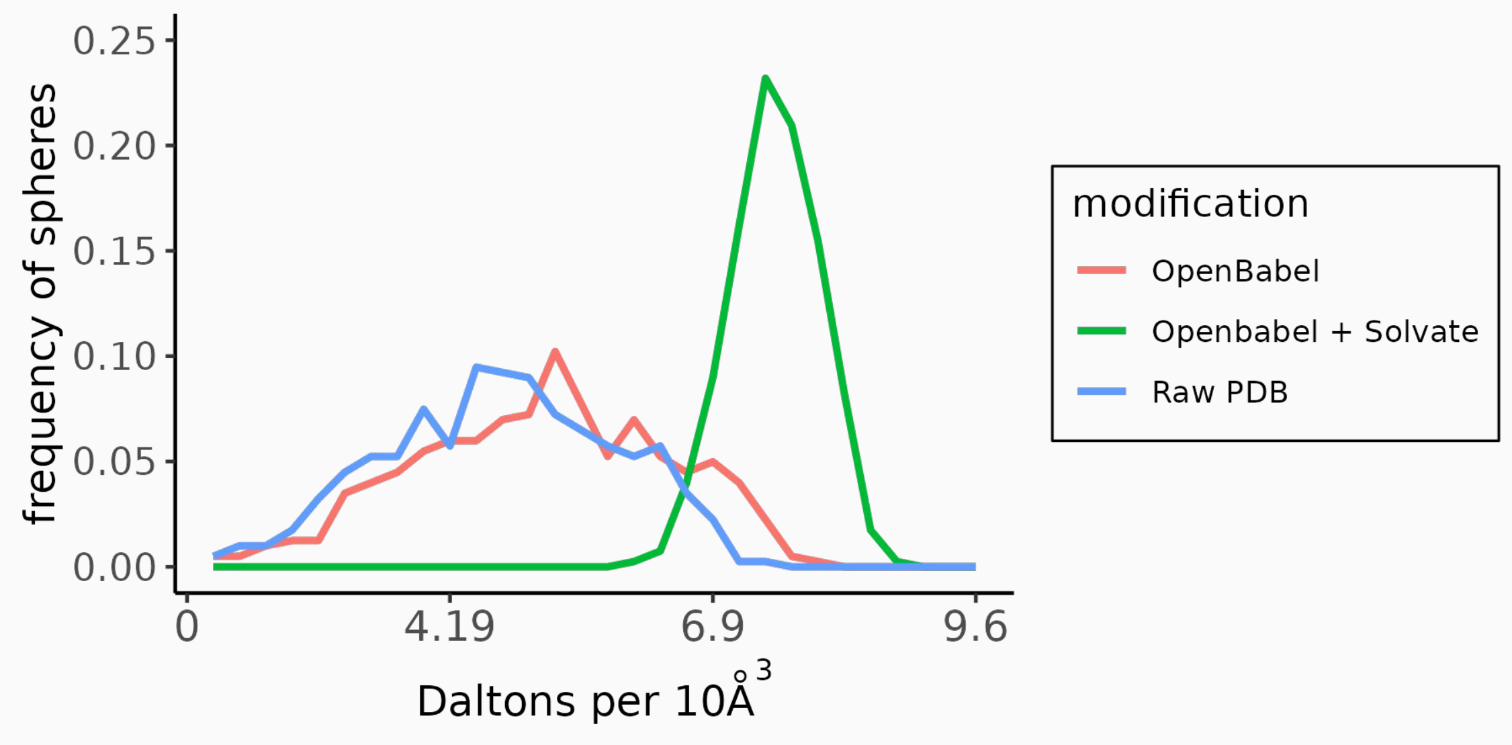
Further examination of outliers derived from different clustering algorithms (hierarchical, k-means and hdbscan) and use of Gene Ontology (GO) terms will give more insights into the structural and functional implications of atomic density. Comparison of proteins clusters derived from different radius and use of other dimension reduction algorithms such as UMAP will remove bias and enhance validity to the results.

PDB files modification

PISCES [1] culling server was used to obtain a diverse structural and functional protein sample. Structures from this sample were downloaded from the Protein Data Bank (PDB) using a File Transfer Protocol.

Missing hydrogen atoms from the PDB files were added using the OpenBabel [2] software. A solvation shell was created around each structure using the Solvate [3] program, to simulate the in-vitro conditions.

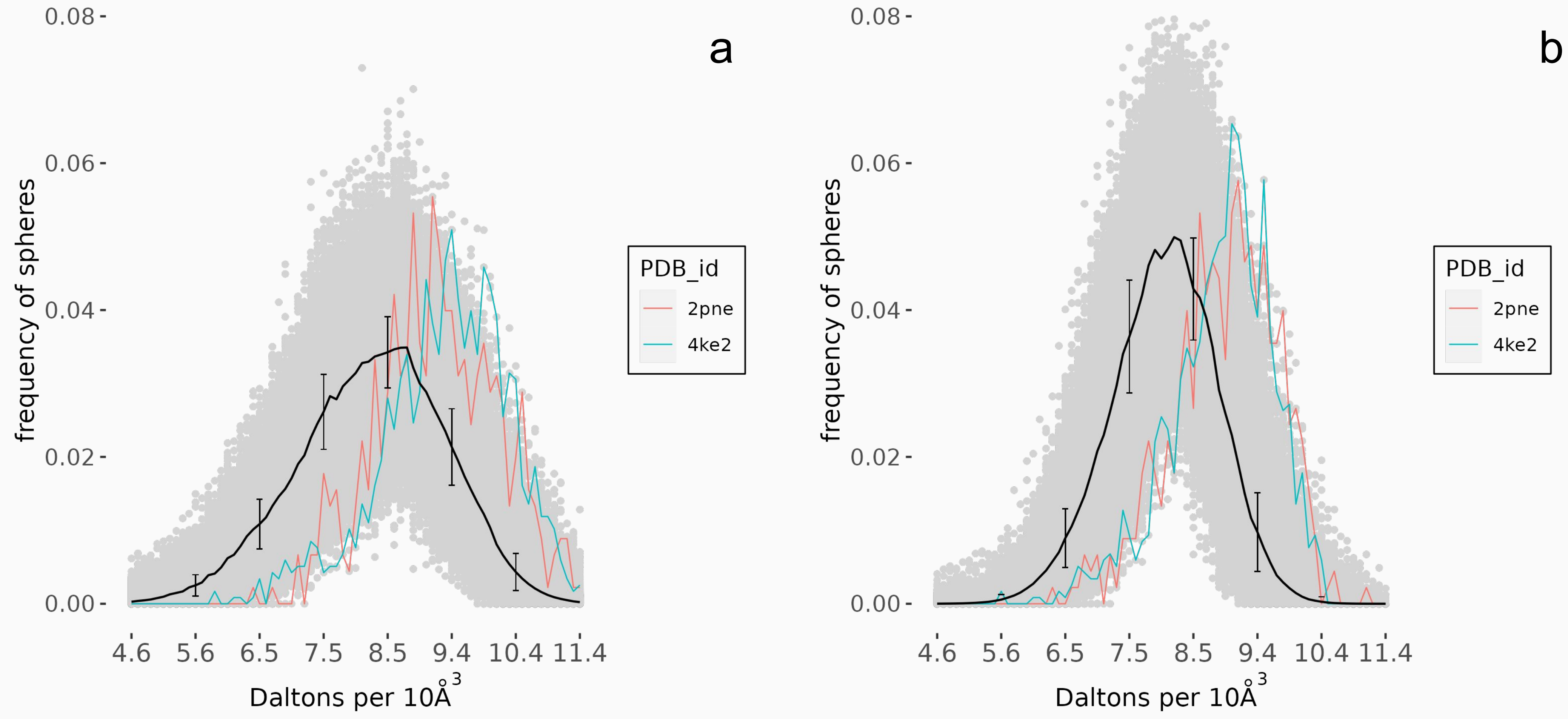




Atomic density distributions for the raw versus the modified PDB files, shows that both hydrogen atoms and water molecules when added to a PDB file, contribute to higher atomic density values and removal of the low ones. This way, distributions are closer to what is observed in-vitro.

Raw data

Raw data scatter plots from distributions of 2 different radius: 5Å (a) and 6Å (b) are shown. The black line stands for the mean value and the black vertical bars stand for the standard deviation. 2 uncommon distributions from outlier structures are also plotted separately in their own line plot.



A slight left shift in the raw data of the 6Å radius is observed, compared to the 5Å. The increase of radius also narrows the edges of the distribution and results in a more bell-curved distribution (gaussian shaped).

References

[1] Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003; doi: 10.1093/bioinformatics/btg224.
[2] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminformatics*. 2011; doi: 10.1186/1758-2946-3-33.
[3] Solvate | https://www.mpinat.mpg.de/grubmueller/solvate
[4] PyMol | pymol.org. https://pymol.org/
[5] Pentelute BL, Gates ZP, Tereshko V, Dashnau JL, Vanderkooi JM, Kossiakoff AA, et al.. X-ray Structure of Snow Flea Antifreeze Protein Determined by Racemic Crystallization of Synthetic Protein Enantiomers. *J Am Chem Soc*. 2008; doi: 10.1021/ja8013538.