

Aberrant coordination geometries discovered in most abundant metalloproteins

Sen Yao^{1,2,4,5,6}, Robert M. Flight^{4,5,6}, Eric C. Rouchka^{1,2,3}, and Hunter N.B. Moseley^{4,5,6,7*}

¹School of Interdisciplinary and Graduate Studies, ²Department of Computer Engineering and Computer Science, ³KBRIN Bioinformatics Core University of Louisville, Louisville, KY 40292, USA

⁴Department of Molecular and Cellular Biochemistry, ⁵Markey Cancer Center, ⁶Center for Environmental and Systems Biochemistry

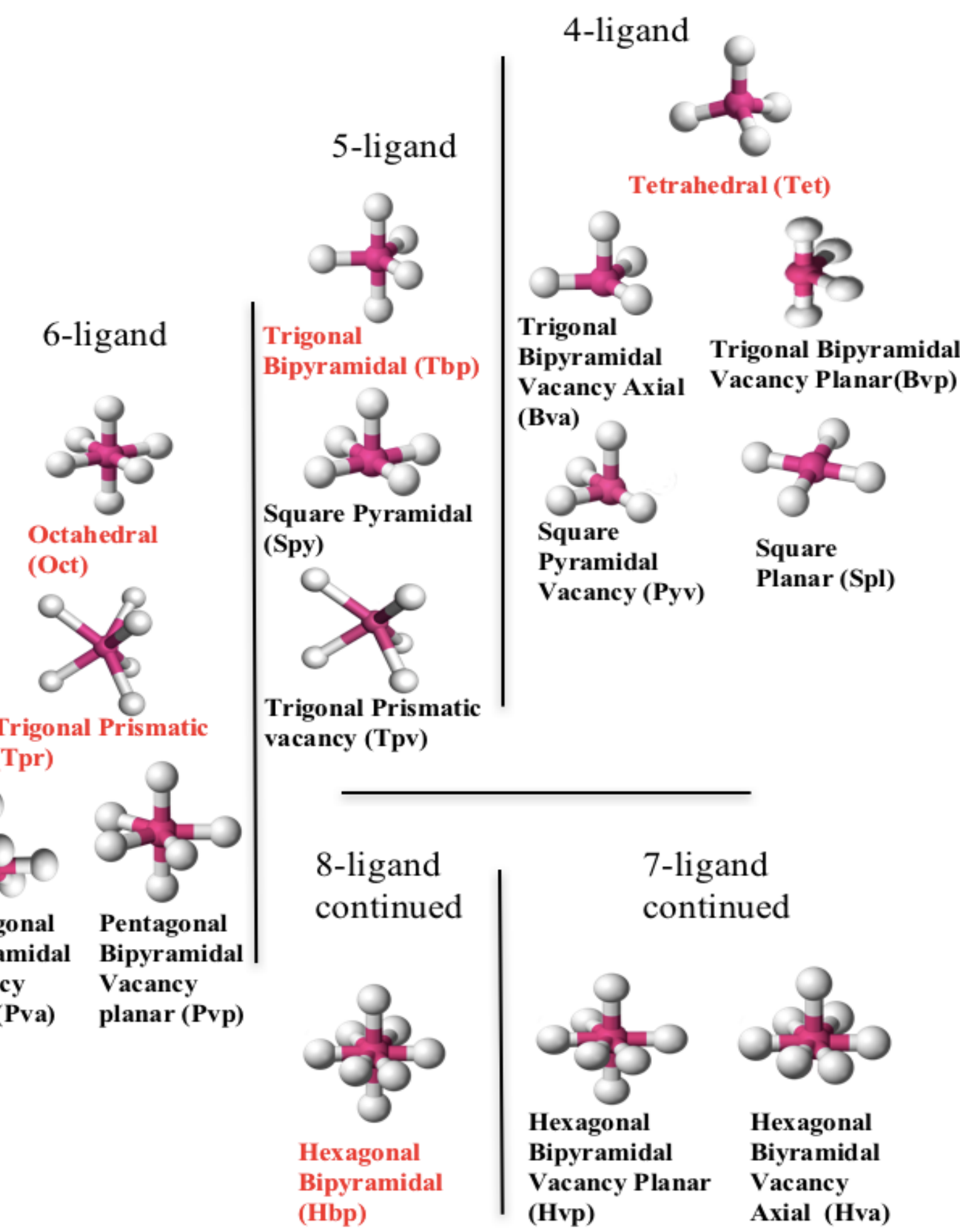
⁷Institute for Biomedical Informatics, University of Kentucky, Lexington KY 40356, USA



Introduction

Metalloproteins play crucial biochemical roles in our body and are essential across all domains of life [1]. The structural environment around a metal ion, especially the coordination geometry (CG), is both sequentially and functionally relevant. Studies of the metalloprotein's CG will greatly help alleviate the imbalance between the ample sequence data available and the insufficient knowledge on protein functions. Current methodologies in characterizing metalloproteins' CG consider only previously reported CG (canonical CG) models based primarily on non-biological chemical context [2,3]. Exceptions to these canonical CG models can greatly hamper the ability to characterize metalloproteins both structurally and functionally.

Figure 1. The canonical coordination geometries (CGs) of metalloproteins. The magenta ball represents the metal ion, and the white balls represent coordination ligands. For each row, a major CG (red) is followed by its associated minor CGs (black), which can be viewed as missing ligands from the major one. The abbreviations are in parenthesis. From the left to the right, the CGs are separated by lines to have 8, 7, 6, 5, and 4 ligands respectively.



Results

Table 1. The number of top 10 metalloproteins in wwPDB as of Feb 2015.

Metal	Number of PDB entries	Number of total metal sites	Metal	Number of PDB entries	Number of total metal
Zn	9360	26788	K	1673	5306
Mg	9145	53896	Cu	1134	4397
Ca	7762	24335	Ni	935	2252
Fe	6359	27514	Co	915	2087
Na	4888	16527	...		
Mn	2266	8138	Total	47,527	187,587

Table 2. Ligand counts and error rates by metal.

Metal	Number of metal clusters	Number of usable metal sites (>3-ligand)	Number of unusable metal sites (<=3-ligand)	4-ligand	5-ligand	6-ligand	7-ligand	8-ligand	9-ligand	Total	Estimated Ligand Detection Error rate	Non-redundant set
Zn	572	21,257	4,959	11,380	2,365	750	2	-	-	14,497	0.000443	4,800
Mg	691	29,859	23,346	3,595	2,941	5,674	69	2	-	12,281	0.002113	2,813
Ca	196	21,057	3,082	918	1,490	4,485	5,399	1,258	18	13,568	0.001760	4,080
Fe	11,287	14,990	1,237	1,071	3,929	5,804	2	-	-	10,806	0.000057	2,370
Na	240	11,475	4,812	703	1,557	1,840	186	17	-	4,303	0.000000	1,184
										Overall	0.001128	

^aHighest coordination number considered valid for the given metal.
^bCoordination numbers considered erroneous and thus used in ligand detection error estimation.

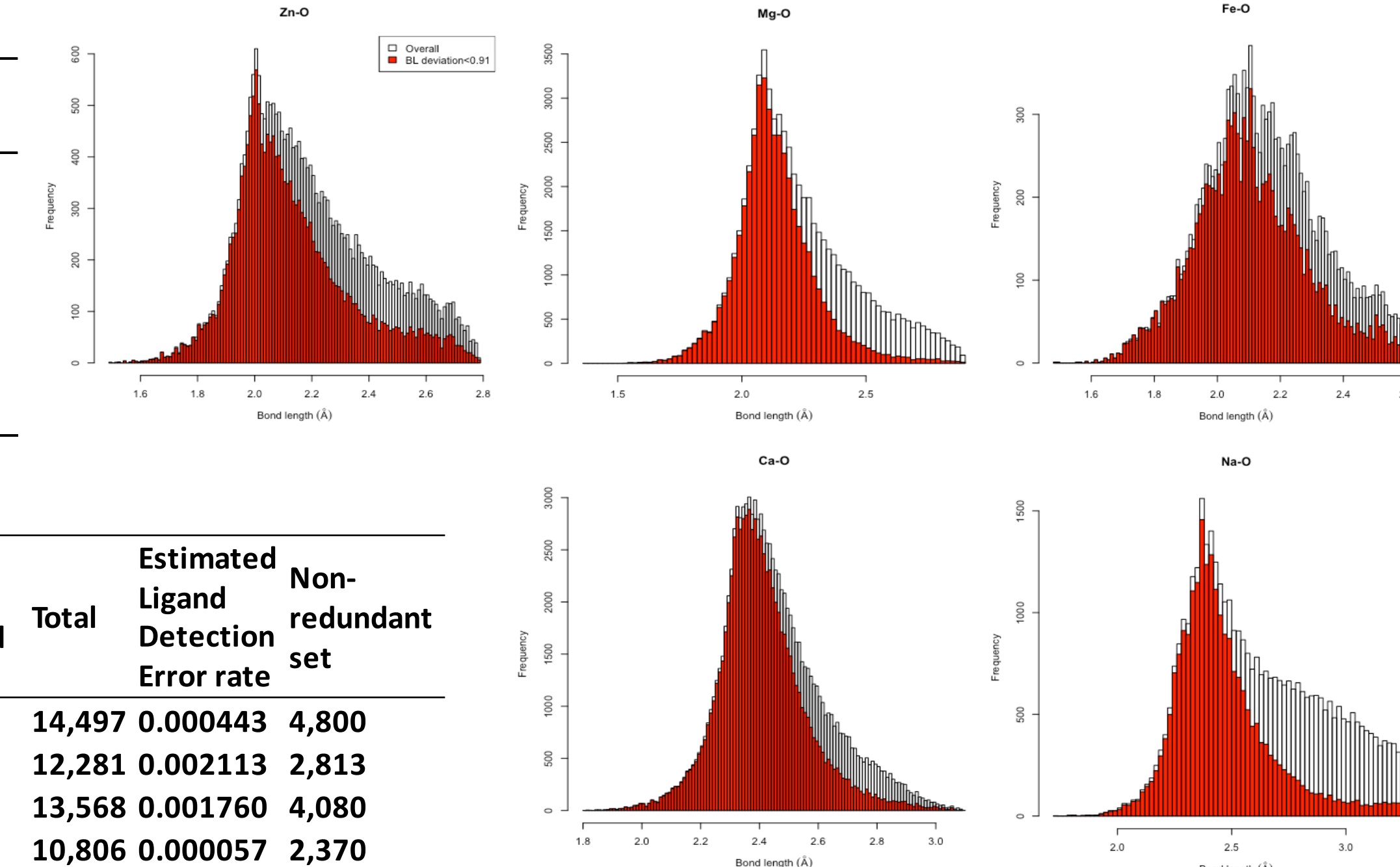


Figure 3. Metal-O bond length distribution after average bond length deviation filtering. The average deviation filter can detect potential misassigned metal ion, and removes the skewed long tails.

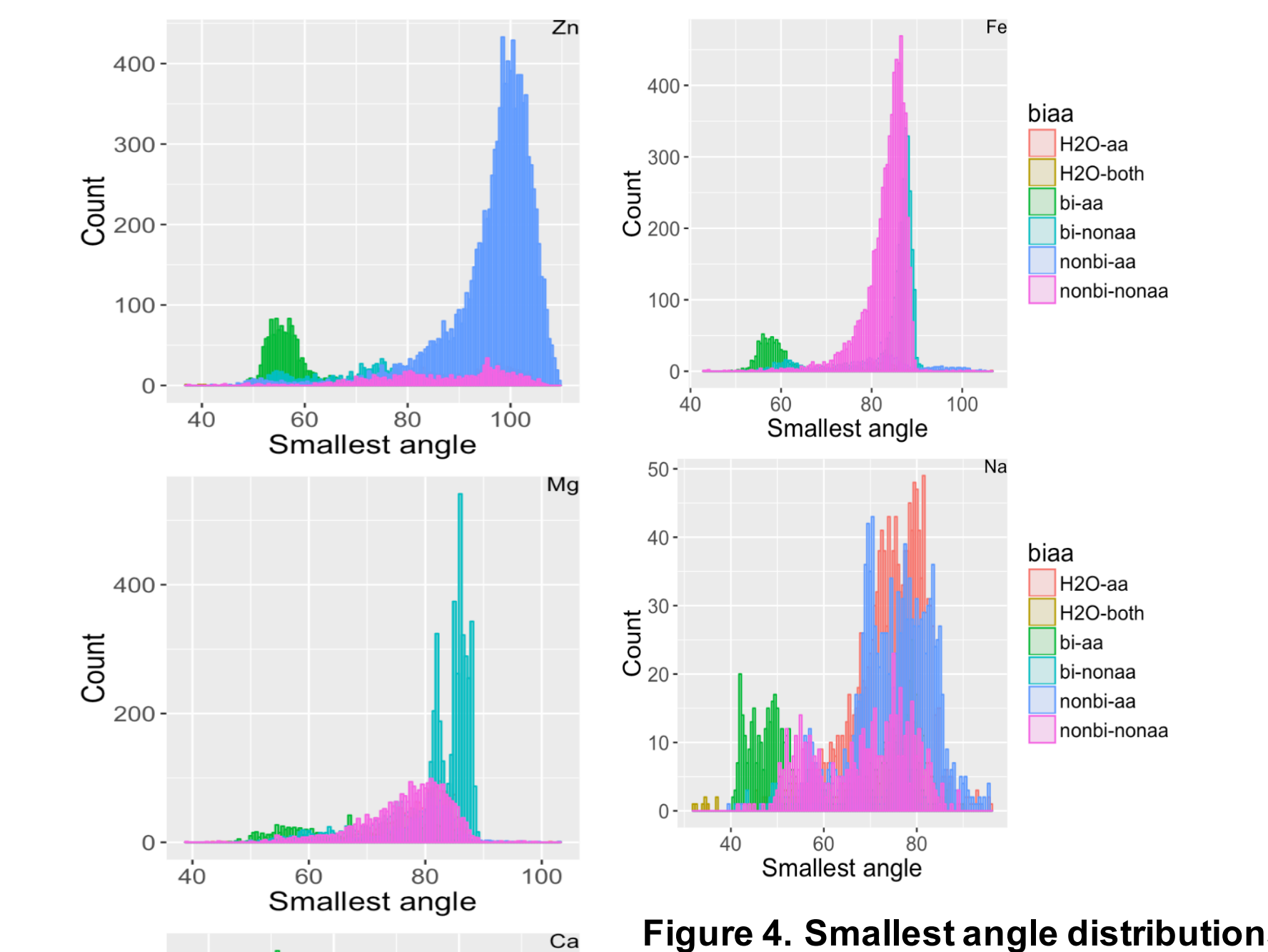


Figure 4. Smallest angle distributions broken down by ligand type and metal. bi is bidentation, aa is amino acid, nonaa means that at least one of the composing ligands is not the 20 standard amino acids, H2O-aa means that one of the composing ligand is the 20 standard amino acids and the other is water, and H2O-both means that both of the composing ligands are water molecules.

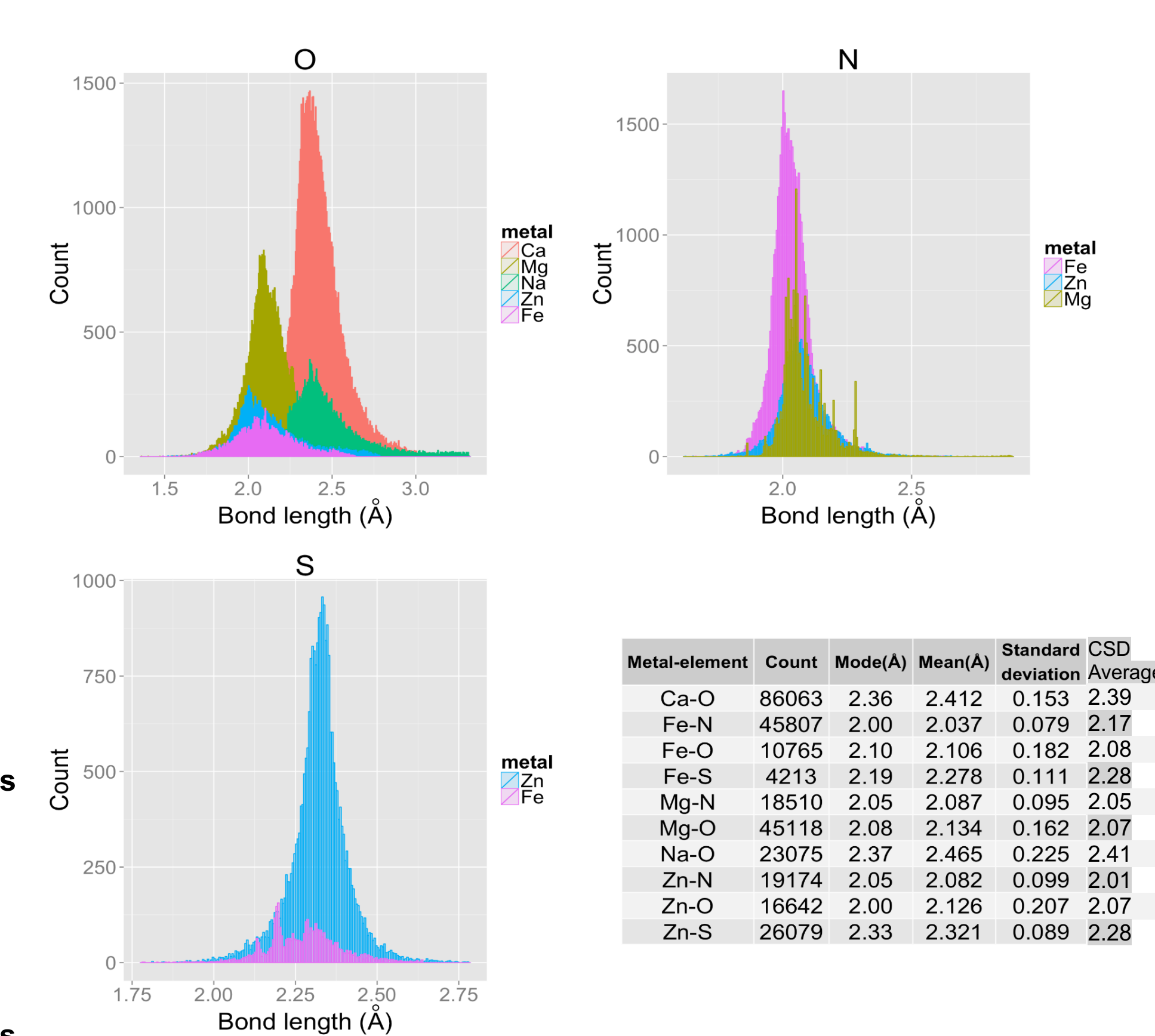


Figure 5. Bond length distributions and statistics of all bond types involving elements O, N, and S, which have greater than 5% occurrence.

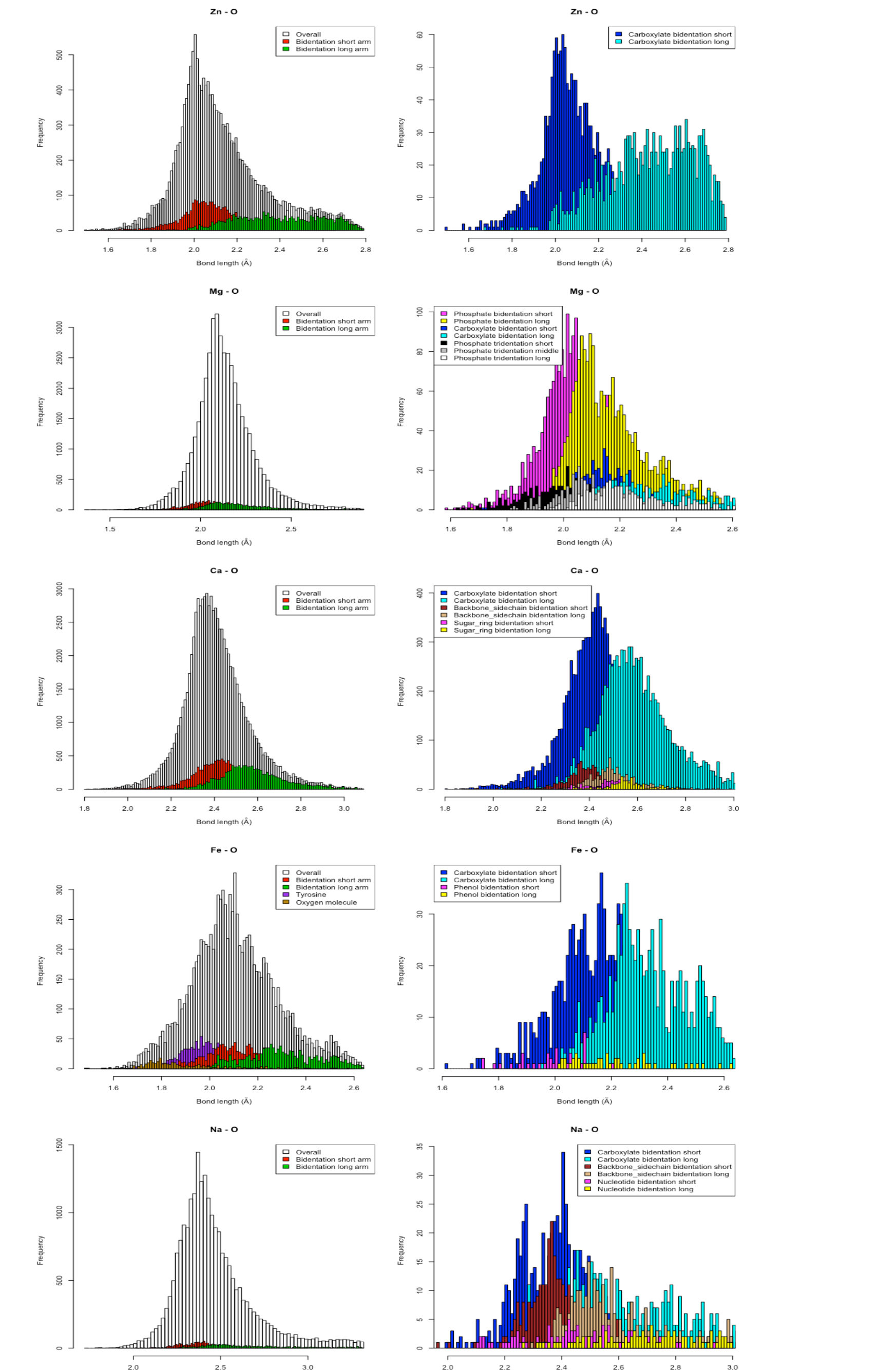


Figure 6. Chemical functional group and multidentation specific bond length modes. On the left is the overall bidentation short and long arms for each metal and some specific functional groups that contributing to the overall bond length histogram. On the right is a breakdown of the most abundant functional groups in the bidentation and multidentation, as they often exhibit distinct modes.

Conclusion

- Various filters were applied to ensure high quality of the data being analyzed, including removing of metal binding sites with abnormal normalized bond-length deviation and correcting standard deviations based on x-ray resolution. The ligand detection method is statistically rigorous, producing an estimated false positive rate of ~0.11% and an estimated false negative rate of ~1.2%.
- The detected bond length modes agree very well with studies based on Cambridge Structural Database (CSD) [7], and are specific to the functional group and multidentation.
- The existence of compressed angles are universal among the top five metals bound by proteins, but especially pronounced in Ca. They caused significant misclassification of metal binding sites into canonical CGs in all previous studies. Aberrant CG models were identified after separating the metal binding sites that have compressed vs. normal angles.
- The universal existence of aberrant CG clusters is further supported by the electron density maps.
- Distinct biochemical functions are associated with aberrant CGs versus non-aberrant CGs
- The ability to detect the structure-function correlation are greatly affected by the sample size (see Yao et al 2017 [5]). Lack of adequate data could be the reason that metal binding sites with compressed angles have never been treated separately as their own independent CGs before.

References

- Andreini et al., Acc Chem Res 2009; 42(10): 1471-1479.
- Alberts et al., Protein Sci 1998; 7(8):1700-1716.
- Andreini et al., Bioinformatics 2012; 28(12):1658-1660.
- Yao et al., Proteins 2015; 83(8): 1470-87.
- Yao et al., Proteins 2017; 85(5): 885-907.
- Yao et al., Proteins 2017; 85(5): 938-944.
- Harding, Acta Cryst 2006; D62, 678-682.

Future Directions

- Develop methods in assessing the quality of a metal binding sites in terms of its electron density fitness.
- Use information of functional group and multidentation bond length modes to improve the ligand detection method.
- Develop better representation for angle-space description of CGs, especially for 7- and 8-ligand CGs.
- Develop methods to better relate overlapping protein regions and improve association between functions and specific clusters.
- Generalize and apply the methodology to all other metal binding sites.

Acknowledgement

NSF 1252893, NIH 1U24DK097215-01A1, and P20GM103436 for KBRIN.

Methods

We have developed a less-biased method that directly handles potential exceptions without pre-assuming any canonical CG models[4-6].

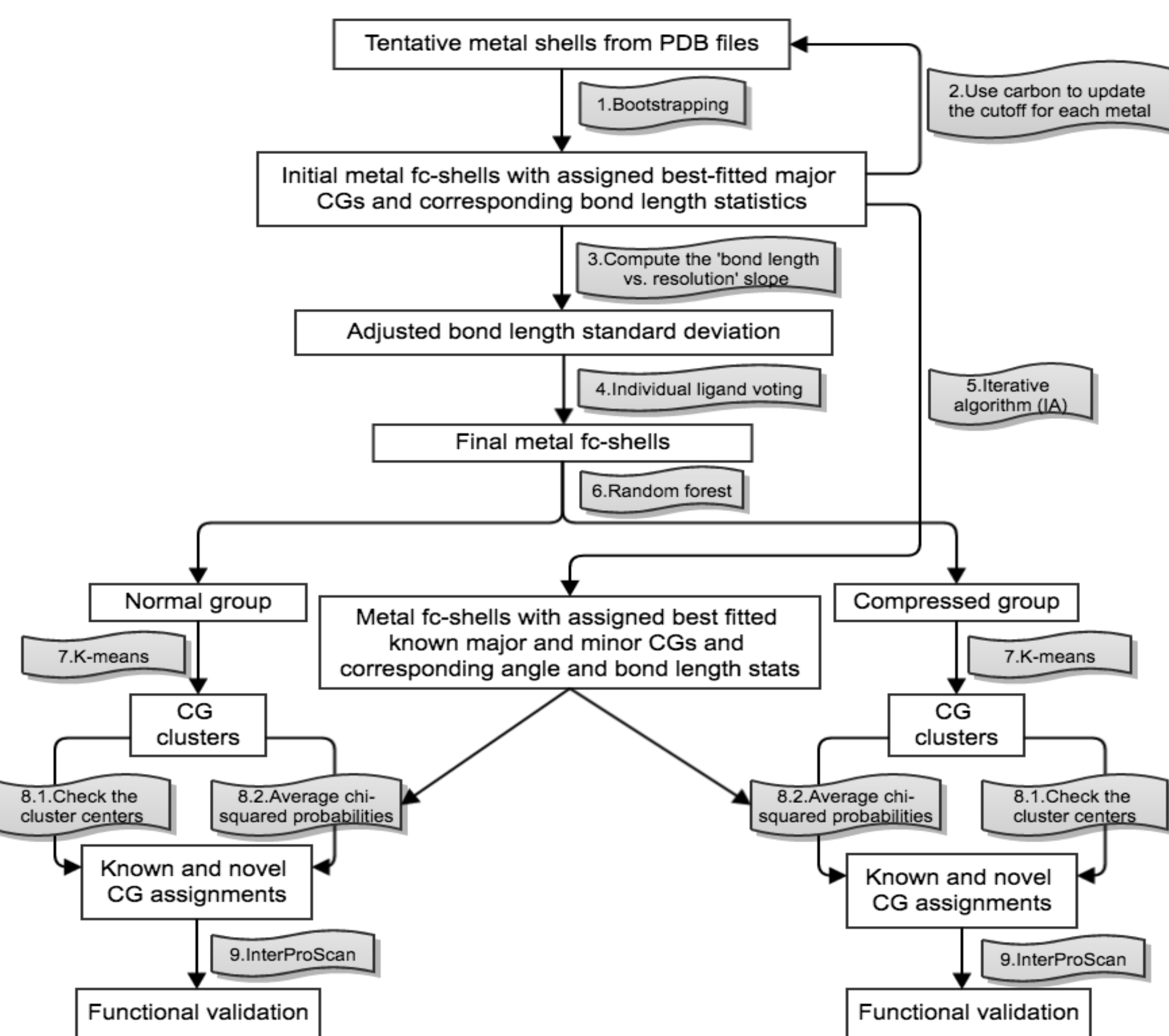


Figure 2. The overall workflow.