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*]

[1]School of Interdisciplinary and Graduate Studies, [2]Department of Computer Engineering and Computer Science, [3]KBRIN Bioinformatics Core
University of Louisville, Louisville, KY 40292, USA
[4]Department of Molecular and Cellular Biochemistry, [5]Markey Cancer Center, [6]Center for Environmental and Systems Biochemistry
[7]Institute 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.

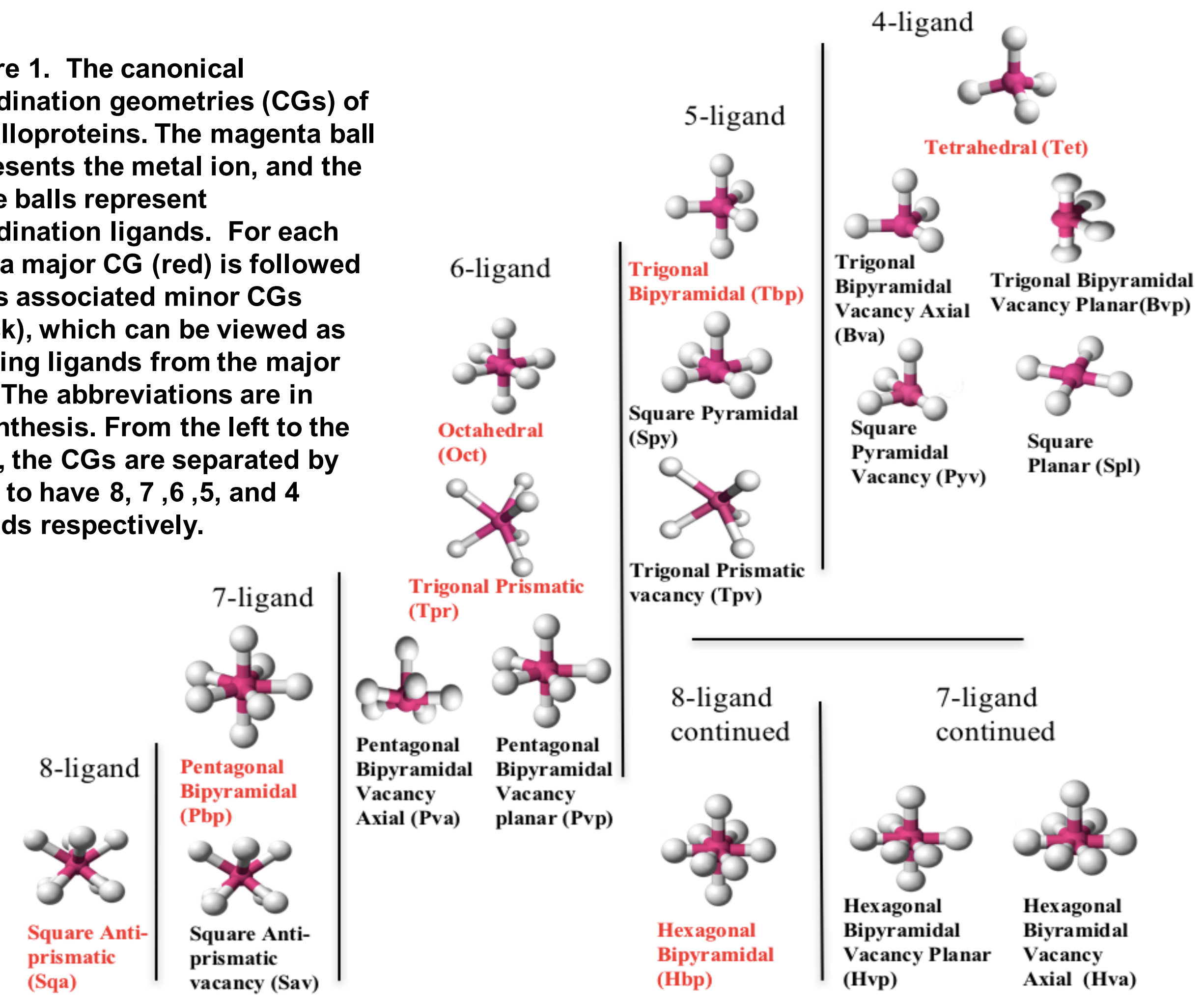
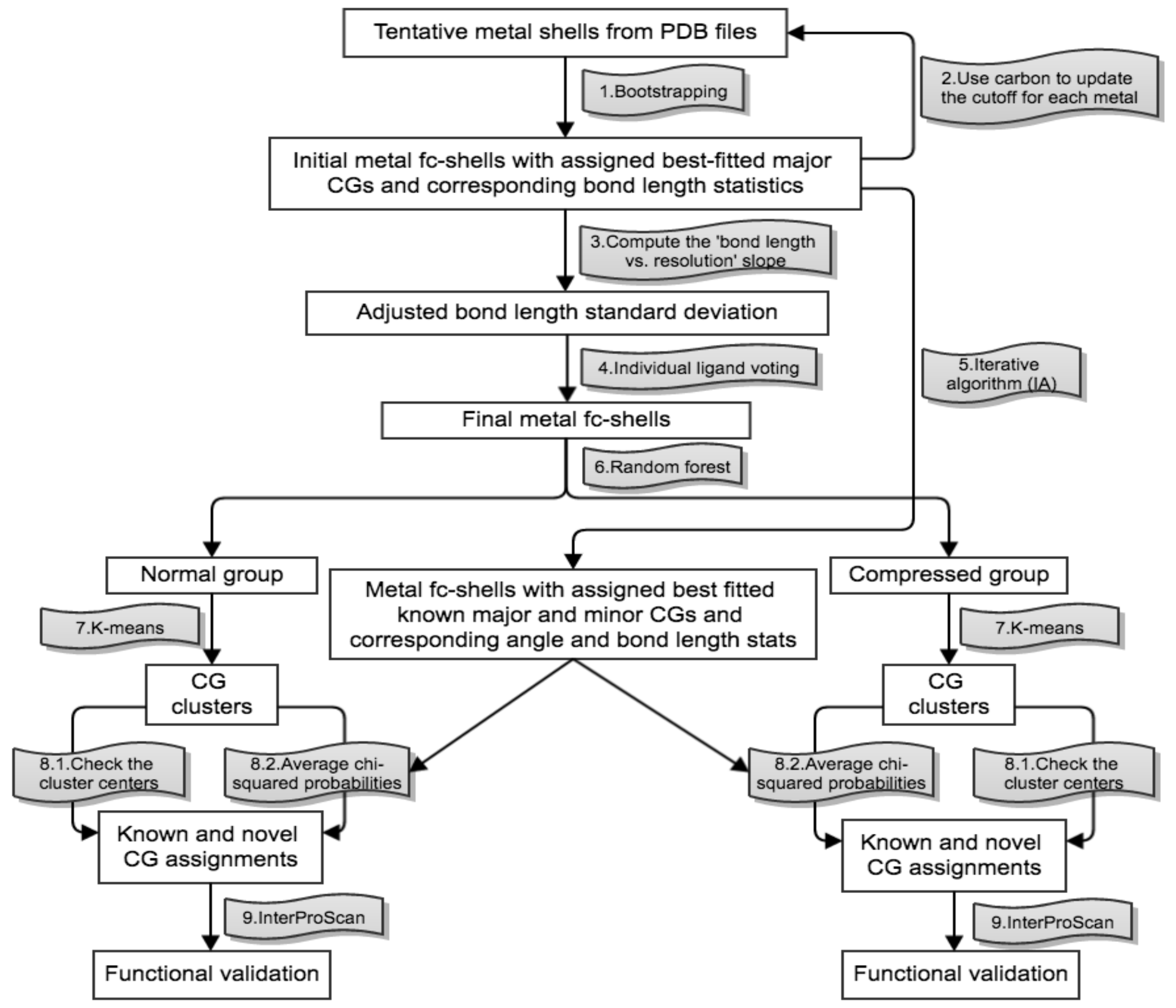

4-ligand
Tetrahedral (Tet)
Trigonal Bipyramidal Vacancy Axial (Bva)
Trigonal Bipyramidal Vacancy Planar(Bvp)
Square Pyramidal Vacancy (Pyv)
Square Planar (Spl)
5-ligand
Trigonal Bipyramidal (Tbp)
Square Pyramidal (Spy)
Trigonal Prismatic vacancy (Tpv)
6-ligand
Octahedral (Oct)
Trigonal Prismatic (Tpr)
Pentagonal Bipyramidal Vacancy Axial (Pva)
Pentagonal Bipyramidal Vacancy planar (Pvp)
7-ligand
Pentagonal Bipyramidal (Pbp)
Square Anti-prismatic vacancy (Sav)
8-ligand
Square Anti-prismatic (Sqa)
8-ligand continued
Hexagonal Bipyramidal (Hbp)
7-ligand continued
Hexagonal Bipyramidal Vacancy Planar (Hvp)
Hexagonal Biyramidal Vacancy Axial (Hva)

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.

♦ Methods

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


Tentative metal shells from PDB files

1.Bootstrapping

2.Use carbon to update the cutoff for each metal

Initial metal fc-shells with assigned best-fitted major CGs and corresponding bond length statistics

3.Compute the 'bond length vs. resolution' slope

Adjusted bond length standard deviation

4.Individual ligand voting

5.Iterative algorithm (IA)

Final metal fc-shells

6.Random forest

Normal group

Metal fc-shells with assigned best fitted known major and minor CGs and corresponding angle and bond length stats

Compressed group

7.K-means

CG clusters

8.1.Check the cluster centers

8.2.Average chi-squared probabilities

Known and novel CG assignments

9.InterProScan

Functional validation

Figure 2. The overall workflow.

♦ 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	[a]750	[b]2	-	-	14,497	0.000443	4,800
Mg	691	29,859	23,346	3,595	2,941	[a]5,674	[b]69	[b]2	-	12,281	0.002113	2,813
Ca	196	21,057	3,082	918	1,490	4,485	5,399	[a]1,258	[b]18	13,568	0.001760	4,080
Fe	11,287	14,990	1,237	1,071	3,929	[a]5,804	[b]2	-	-	10,806	0.000057	2,370
Na	240	11,475	4,812	703	1,557	1,840	186	[a]17	-	4,303	0.000000	1,184
										Overall	0.001128	

[a]Highest coordination number considered valid for the given metal.
[b]Coordination numbers considered erroneous and thus used in ligand detection error estimation.

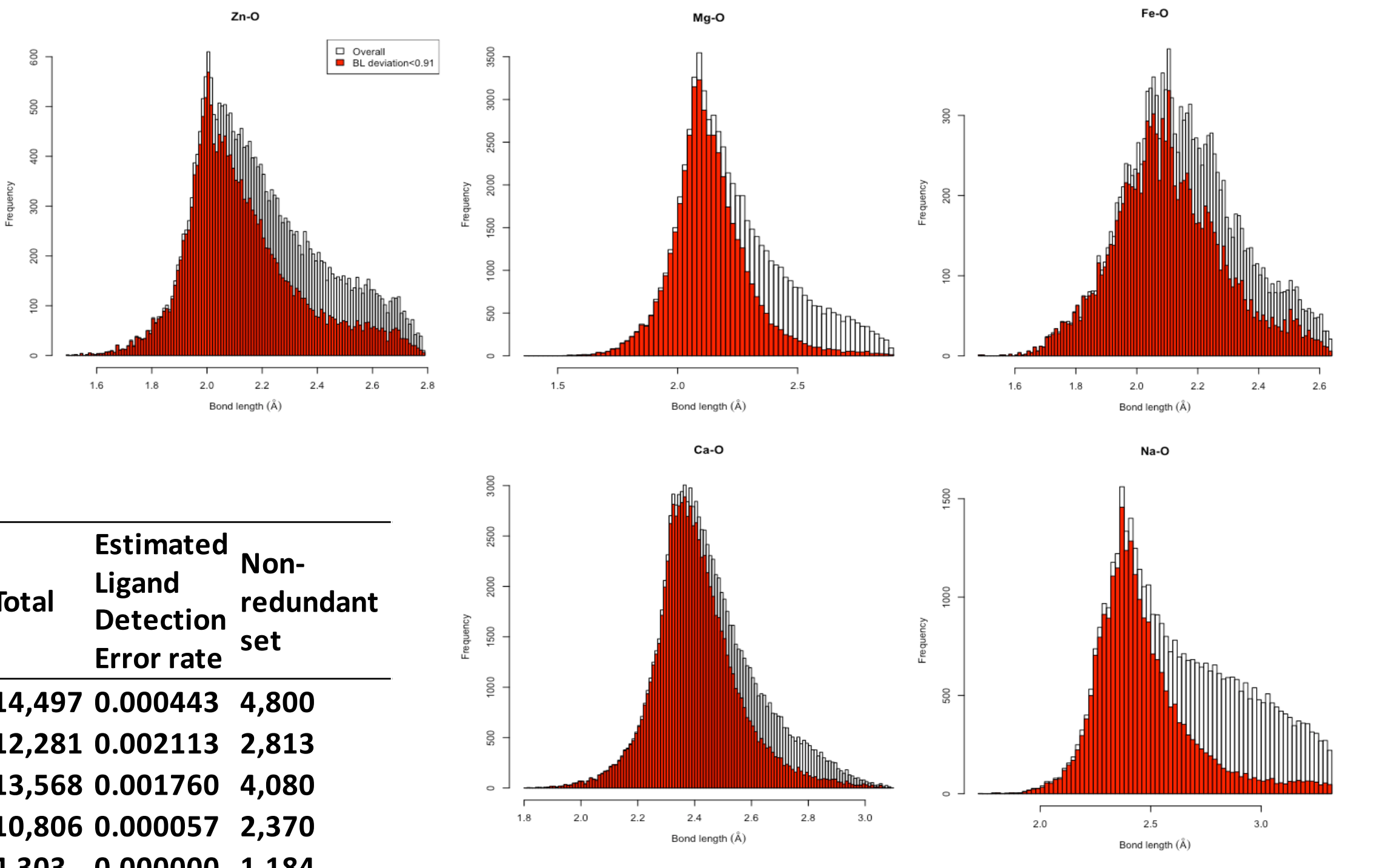


Figure 3. Metal-O bond length distribution after average bond length deviation filtering. The average deviation filter can detect potential misassigned metal ion, and removes the skewed long tails.

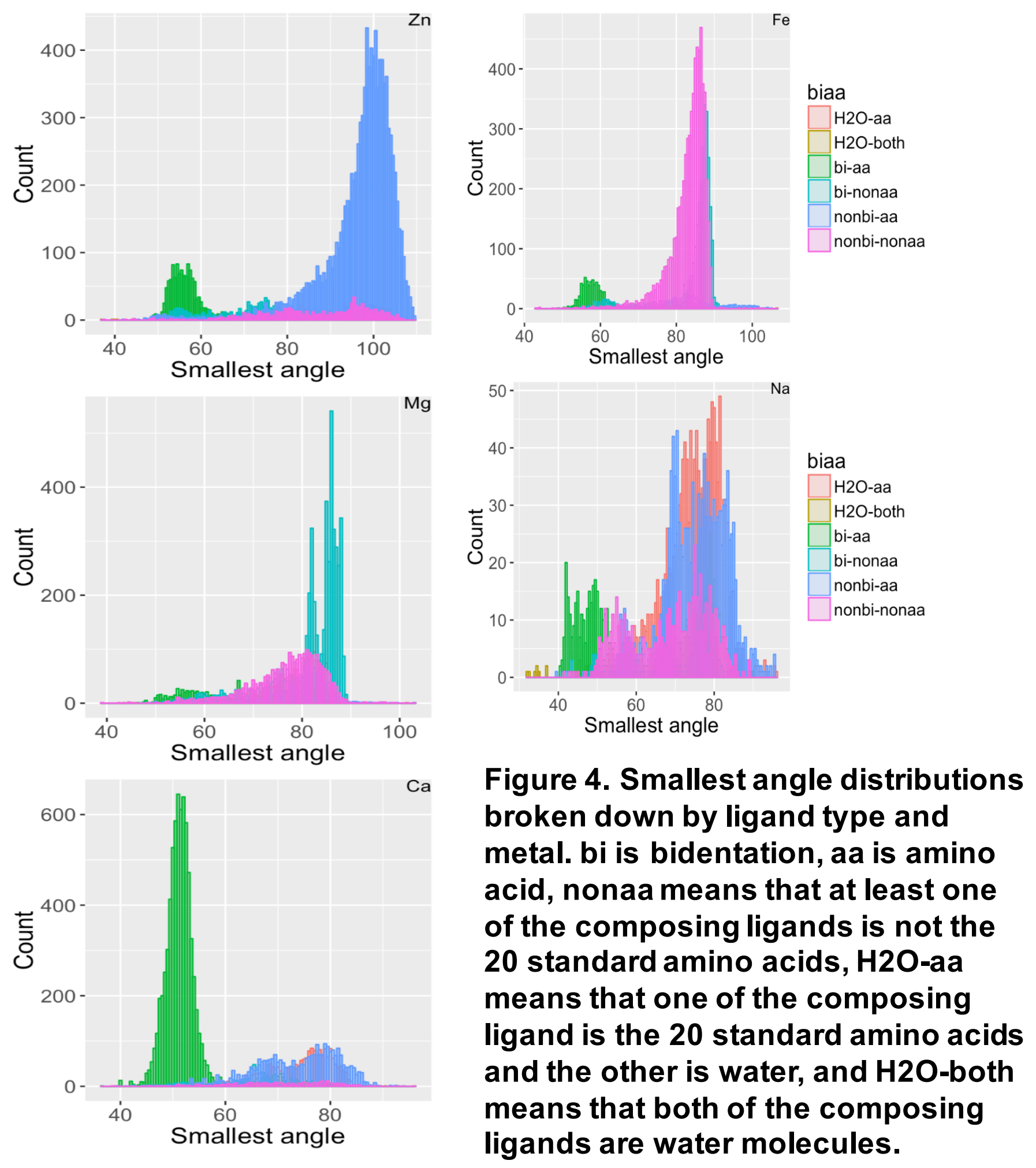
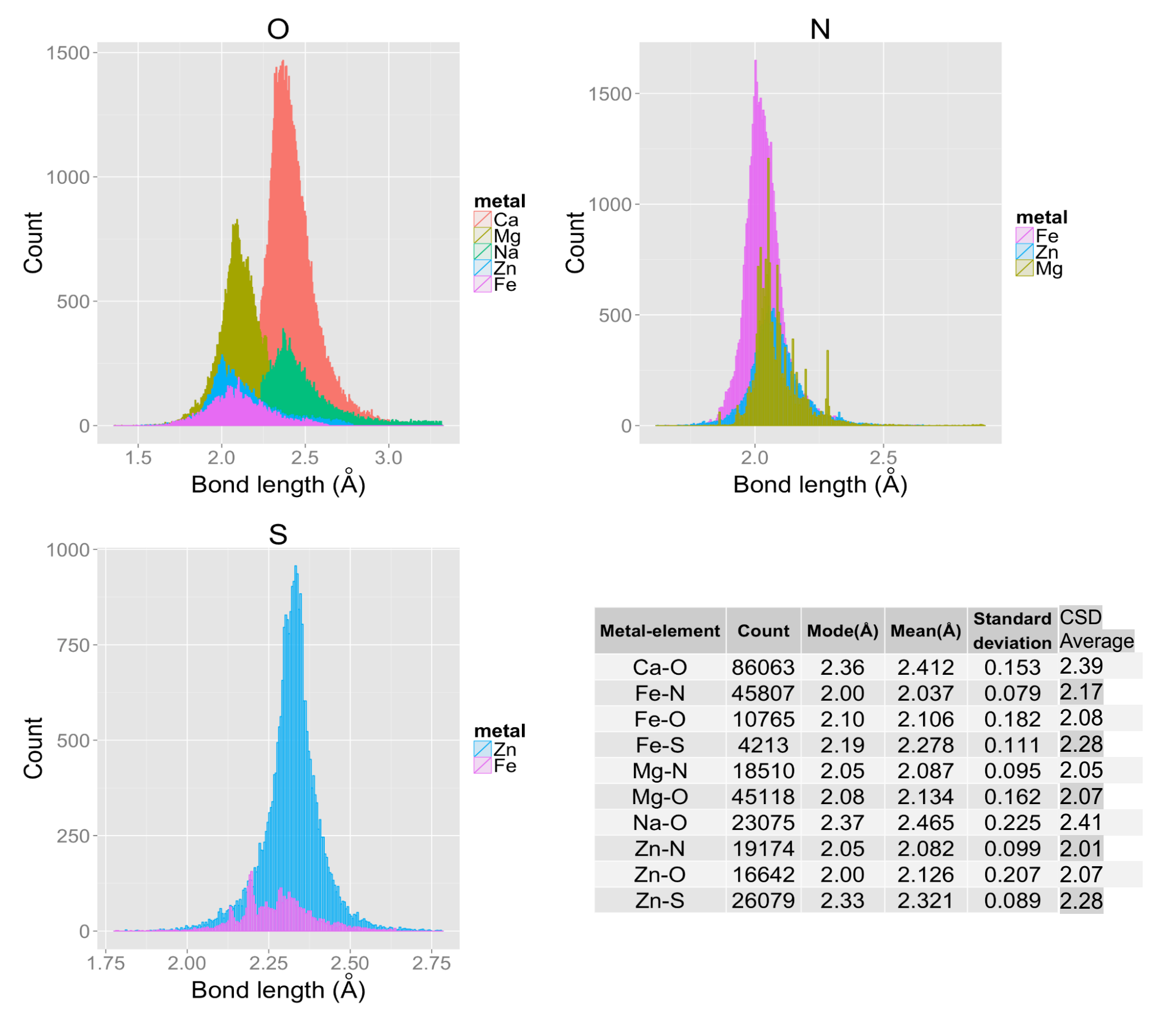


Figure 4. Smallest angle distributions broken down by ligand type and metal. bi is bidentation, aa is amino acid, nonaa means that at least one of the composing ligands is not the 20 standard amino acids, H2O-aa means that one of the composing ligand is the 20 standard amino acids and the other is water,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



Metal-element	Count	Mode(Å)	Mean(Å)	Standard deviation	CSD Average
Ca-O	86063	2.36	2.412	0.153	2.39
Fe-N	45807	2.00	2.037	0.079	2.17
Fe-O	10765	2.10	2.106	0.182	2.08
Fe-S	4213	2.19	2.278	0.111	2.28
Mg-N	18510	2.05	2.087	0.095	2.05
Mg-O	45118	2.08	2.134	0.162	2.07
Na-O	23075	2.37	2.465	0.225	2.41
Zn-N	19174	2.05	2.082	0.099	2.01
Zn-O	16642	2.00	2.126	0.207	2.07
Zn-S	26079	2.33	2.321	0.089	2.28

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

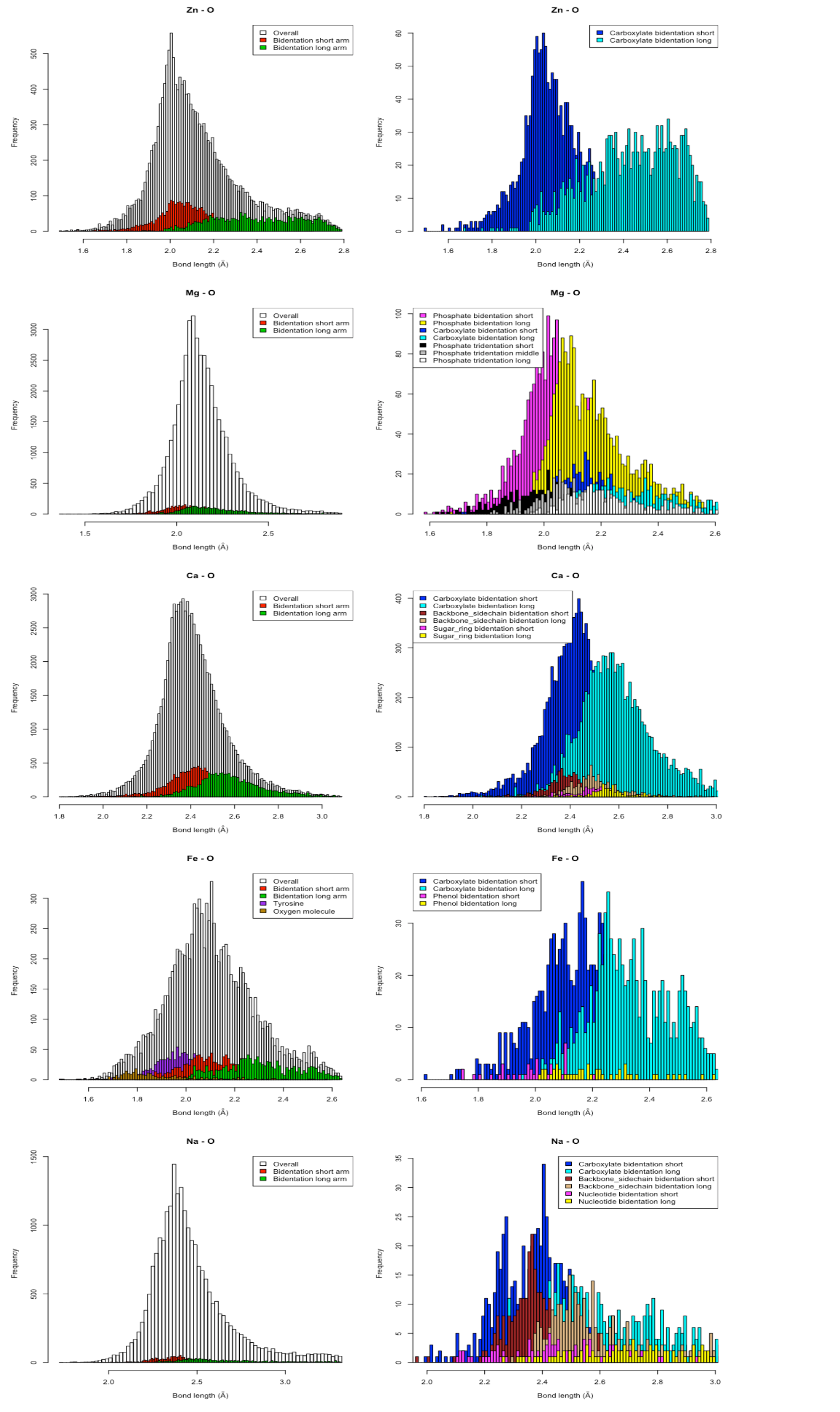


Figure 6. Chemical functional group and multidentation specific bond length modes. On the left is the overall bidentation short and long arms for each metal and some specific functional groups that contributing to the overall bond length histogram. On the right is a breakdown of the most abundant functional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

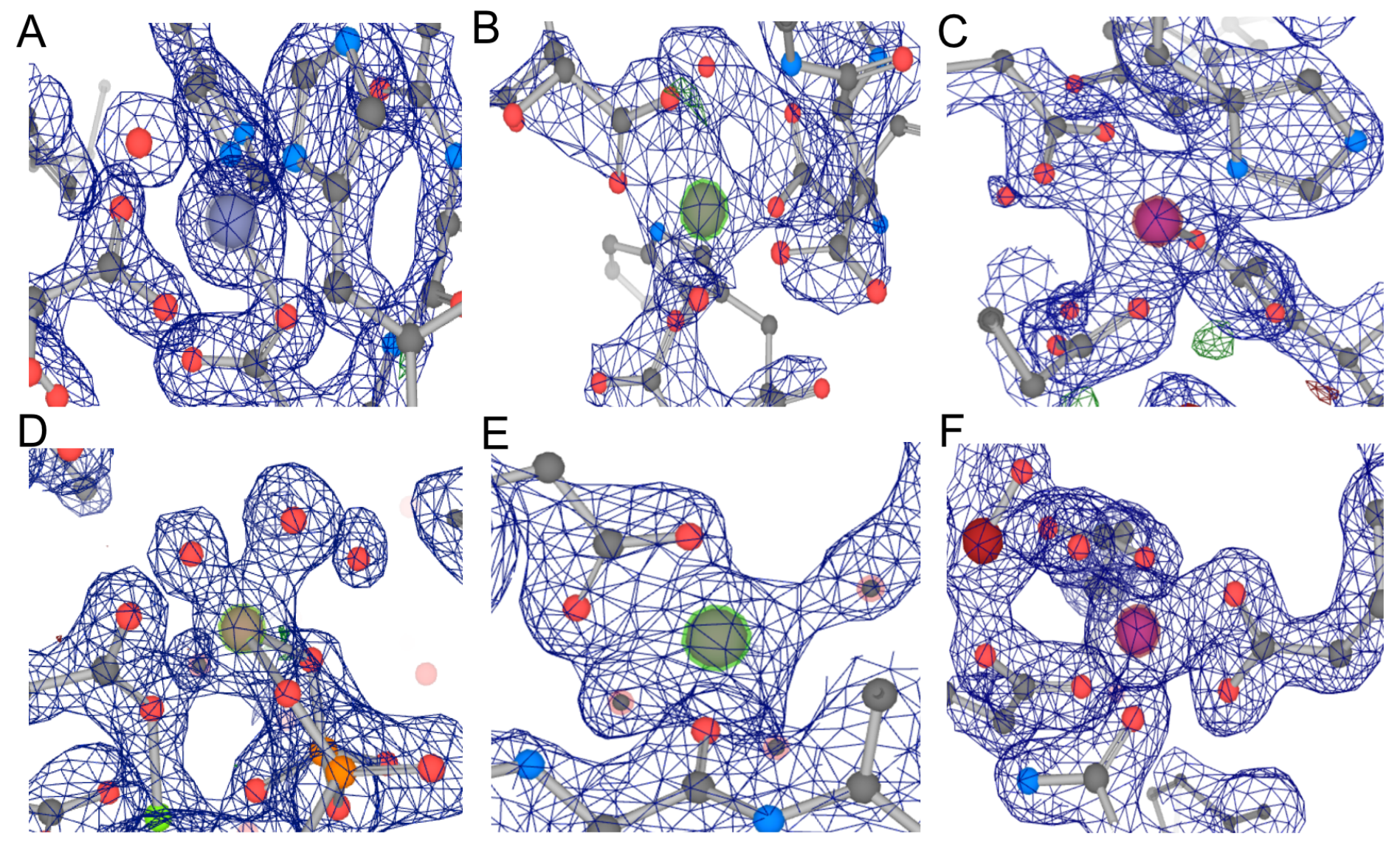


Figure 7. PDB structure and electron density maps of examples for highly aberrant CGs. Structures are shown in balls and sticks and featured by bidentated compressed angles. These structures are also supported by their fitness to the electron density maps. All structures were generated in LiteMol Viewer, with 2Fo–Fc at 1.5 σ and Fo – Fc at -3σ (red) and 3σ (green), except for panel E with 2Fo – Fc at 1.01 σ. Metal ions are put at the center of each subgraph with larger size, where Zn is represented as light blue, Fe as purple, and Mg and Ca as green. The cluster identifier, PDB metal site ID, and its resolutions are as follows: A, 5-ligand Zn, cluster 3, 2B13.B.401, resolution 1.55Å; B, 5-ligand, Ca, cluster 6, 3RYD.C.267, resolution 2.37Å; C, 5-ligand Fe, cluster 1, 4AM4.A.1161, resolution 1.68Å; D, 6-ligand Mg, cluster 8, 3ETH.A.402, resolution 1.60Å; E, 6-ligand Ca, cluster 3, 4P99.B.509 resolution 1.80Å; F, 6-ligand Fe, cluster 2, 2GYQ.B.404, resolution 1.40Å

Table 3: Enriched GO terms, having corrected p-value <= 0.05 and consistent enrichment patterns in combineLig and individual ligand enrichments.

			Normal				Compressed			
ID	Description	Type	P-adjust	Metal	%	Sig.	P-adjust	Metal	%	Sig.
GO:0044249	Cellular biosynthetic process	BP	6.595e-06	Zn	0.4264	TRUE	1.000e+00	Mg	0.3297	FALSE
GO:0072524	Pyridine-containing compound metabolic process	BP	4.483e-02	Mg	0.4430	TRUE	1.000e+00	Mg	0.5000	FALSE
GO:0034641	Cellular nitrogen compound metabolic process	BP	2.359e-08	Zn	0.4413	TRUE	1.000e+00	Mg	0.3211	FALSE
GO:0015077	Monovalent inorganic cation transmembrane transporter activity	MF	3.796e-02	Fe	0.4242	TRUE	1.000e+00	Ca	0.7500	FALSE
GO:0072593	Reactive oxygen species metabolic process	BP	9.952e-03	Fe	0.4821	TRUE	1.000e+00	Zn	1.0000	FALSE
GO:1901566	Organonitrogen compound biosynthetic process	BP	1.320e-03	Mg	0.3933	TRUE	1.000e+00	Mg	0.4865	FALSE
GO:0016053	Organic acid biosynthetic process	BP	3.090e-02	Mg	0.3543	TRUE	1.000e+00	Mg	0.3077	FALSE
GO:0044283	Small molecule biosynthetic process	BP	1.108e-03	Mg	0.4000	TRUE	1.000e+00	Mg	0.3333	FALSE
GO:0046394	Carboxylic acid biosynthetic process	BP	3.090e-02	Mg	0.3543	TRUE	1.000e+00	Mg	0.3077	FALSE
GO:0050794	Regulation of cellular process	BP	1.459e-04	Zn	0.4454	TRUE	1.000e+00	Zn	0.4545	FALSE
GO:0050896	Response to stimulus	BP	1.530e-04	Mg	0.3825	TRUE	1.000e+00	Ca	0.4860	FALSE
GO:0065008	Regulation of biological quality	BP	1.000e+00	Fe	0.2778	FALSE	6.807e-03	Fe	0.3878	TRUE
GO:0008484	Sulfuric ester hydrolase activity	MF	1.000e+00	Ca	0.3333	FALSE	2.187e-02	Ca	0.4286	TRUE
GO:0006811	Ion transport	BP	1.000e+00	Mg	0.3889	FALSE	6.392e-06	Ca	0.4364	TRUE

♦ References

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

♦ 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.

♦ 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.