Bioinformatics, 2023, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Sequence Analysis

M-Ionic: Prediction of metal ion binding sites from sequence using residue embeddings

Aditi Shenoy[1†], Yogesh Kalakoti[2†], Durai Sundar[2,3], Arne Elofsson[1*]
[1] Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, 171 21 Solna, Sweden
[2] Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India
[3] Yardi School of Artificial Intelligence, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India

*To whom correspondence should be addressed.

[†]The authors wish it to be known that the first two authors should be regarded as Joint First Authors

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Understanding metal-protein interaction can provide structural and functional insights into cellular processes. As the number of protein sequences increases, developing fast yet precise computational approaches to predict and annotate metal binding sites becomes imperative. Here, we have developed M-Ionic, a sequence-based method to predict which metals a protein binds and the binding residues. Since the predictions use only residue embeddings from a pre-trained protein language model, quick predictions can be made for the ten most frequent metal ions (Ca^{2+}, Co^{2+}, Cu^{2+}, Mg^{2+}, Mn^{2+}, Po_4^{3-}, So_4^{2-}, Zn^{2+}, Fe^{2+}, Fe^{3+}).

Availability: Google Colab Notebook at https://bit.ly/40FrRbK

Implementation: https://github.com/TeamSundar/m-ionic

Contact: arne@bioinfo.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 . Introduction

Proteins containing metal cofactors (metalloproteins) are involved in various cellular biochemical processes including protein folding (Palm-Espling *et al.*, 2012), enzymatic reactions (Andreini *et al.*, 2009), and signalling (Maret, 2017). Understanding metal-binding specificity can provide information about the functions and mechanisms of proteins. Of all proteins, about one-third of them are known to interact with metal ions (Tainer et al., 1991, Andreini *et al.*, 2009) , but as per Aptekmann *et al.*, 2022, only ~14% is annotated as metal binding in the UniProt database (The UniProt Consortium, 2017). Experimental methods to identify metal-binding sites are labour-intensive and time-consuming owing to several factors involving size and versatility of metal-ions. Multiple attempts to develop *in-silico* alternatives for metal-ion binding-site prediction have been previously proposed. These can be broadly categorised in three categories, namely (i) sequence-based, (ii) structure-based and (iii) both sequence/structure-based, largely based on the datatype employed for training the estimator. Structure-based methods such as MIB2 (Lu *et al.*, 2022) considers the 3-dimensional neighbourhood surrounding the metal ion. While sequence-based methods like LMetalSite (Yuan *et al.*, 2022), MebiPred (Aptekmann *et al.*, 2022) and IonSeq (Hu *et al.*, 2016) makes use of the protein sequence for training the estimator, IonCon (Hu *et al.*, 2016) employs both sequence and structural information. However, these methods are limited by the availability of template structures, sequence embedding protocols and coverage of important metal-ions. There have already been similar attempts to employ the information contained in embeddings extracted from various protein language models (PLMs) such as ESM-1b (Rives *et al.*, 2021) and ProtTrans (Elnaggar *et al.*, 2020) for identifying subcellular localisation (Stärk *et al.*, 2021), post-translational modification (Pokharel *et al.*, 2022) and ligand binding sites (Littmann *et al.*, 2021). In this study, residue-level PLM embeddings were employed to estimate (i) the likelihood of a given metal ion binding to a protein (protein level prediction) and (ii) the propensity of individual protein residues to interact with the given metal-ion (residue level prediction). Taken together, M-ionic, a method which provides binding-site predictions for the ten most frequently occurring metal ion groups is presented. Even when M-ionic uses just single residue embedding, it obtains a comparable performance to state-of-the-art methods. This indicates the effectiveness of residue-level embeddings and developed workflow is sufficient to estimate metal-ion binding sites.

2. Methods

The latest release from the BioLip dataset (Yang *et al*., 2013) was downloaded (Apr 01, 2022). Metal-ion groups having a minimum of 1000 samples were selected (i.e., Ca^{2+}, Co^{2+}, Cu^{2+}, Mg^{2+}, Mn^{2+}, $Po4^{3-}$, $So4^{2-}$, Zn^{2+}, Fe^{2+}, Fe^{3+}, Na^{+}). Protein chains with >1000 amino acids and structures with >3Å resolution were excluded from the analysis (Table S1). Sequences were clustered using CD-HIT (Li and Godzik, 2006) at 100% sequence identity to remove identical sequences with contradictory labels. Next, sequence homology was reduced to 20% using BlastClust (Altschul *et al*., 1990) with the parameters -p T -L .9 -b T -S 20. The 6203 clusters obtained were shuffled and split into six partitions. One partition was reserved as the independent test set, and the other 5 clusters were used for training the estimator under a five-fold cross-validation strategy.

To evaluate whether the proposed method can distinguish between (metal) binding and non-binding proteins, a negative set (i.e., protein chains that do not bind to any metal ion) was created. Protein bio-assemblies from PDB with less than six protein chains and chains with no metals in the HETATM records were filtered (n=3810). After clustering at 100% identity, 3224 unique sequences were used as the negative set.

ESM-1b (esm2_t33_650M_UR50D) and ProtT5 (ProtT5-XL-UniRef50) residue-level embeddings were generated for all sequences. The input multiple sequence alignments (MSA) to the esm_msa1b_t12_100M_UR50S model were generated using HHblits (Remmert *et al*., 2012) with the default parameters. The input to the network was the residue-level embeddings (Figure S1). The model generated an 11-dimensional prediction vector with ten dimensions corresponding to individual metal-ions and one exclusive dimension for designating non-binding status of the residue. The performance of M-Ionic was benchmarked against other metal-binding predictors (e.g. Lippi et al., 2008; Yu et al., 2013; Hu *et al*., 2016; Yuan *et al*., 2022) using the independent validation dataset from MIonSite (Qiao and Xie, 2019).

3. Results and Discussion

Performance at the protein level: The independent test set and negative set (described in Methods) were used to compare the performances of M-Ionic, mebi-pred (Aptekmann *et al.*, 2021) and LMetalSite (Yuan *et al.*, 2022) in predicting whether a protein binds to a metal and which metals interact with the protein. We observed that M-Ionic could distinguish (metal) binding and non-binding proteins better than the other methods, with a 64% increase in MCC (Table S2). Receiver operating characteristic curves (Figure 1a) were employed as a metric to evaluate if the methods could correctly identify the metals involved with protein binding. M-Ionic showed comparable performance with mebi-pred for most ions. However, LMetalSite had a higher false positive rate and reported poor performance distinguishing metal ions for the four most frequent ions (Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}). It can be inferred that the drop in performance could be due to how LMetalSite was trained. Since the training set did not contain any non-metal-binding examples, and the datasets had more than a 20% sequence identity, it is quite possible that the model might have been over-trained.

Performance at the residue level: Besides predicting the metals that bind to the protein of interest, knowing which residues are involved in metal binding is essential. Figure 1b shows that LMetalSite and M-Ionic have similar performance in predicting Zn^{2+}, Ca^{2+}, Mg^{2+}, and Mn^{2+}.

LMetalSite and M-Ionic differ in their inputs and how the network is trained. LMetalSite batches proteins of different lengths, pads them together and trains along the length of the proteins. M-Ionic, however, is trained on the embedding dimension (1280), i.e., ignoring information from neighbouring residues. In addition to predicting six more ions, M-Ionic shows that a single residue embedding is sufficient to predict the metal binding probability for that residue.

The network was retrained with batch size 1 (esm2_t33_650M_UR50D_batch1_CV) and by scrambling the order of residues within a batch of 128 residues (esm2_t33_650M_UR50D_batch128_scrambled) to test whether there was interference from neighbouring residues during training. No difference between the models was observed (Figure S2). We did not see an improvement in performance using (ProtT5-XL-UniRef50) residue-level embeddings (Figure S2). The results from the independent benchmark set from MIonSite are reported in (Table S3).

Impact of MSAs: The same model architecture was retrained with MSA embeddings (esm_msa1b_t12_100M_UR50S) as inputs to test whether using explicit evolutionary information could improve the residue-level prediction performance. However, no improvement was observed (see esm_msa1b_t12_100M_UR50S_batch128 in Figure S2).

4. Conclusions

At the protein level (i.e., the ability to identify which metals a protein binds to), the performance of M-Ionic was comparable to mebi-pred. At the residue level (i.e., the ability to predict which residues are involved in metal-ion-specific binding), M-Ionic generated binding-site predictions (with MCC scores comparable with LMetalSite) for six additional metal ions (i.e. Co^{2+}, Cu^{2+}, Po_4^{3-}, So_4^{2-}, Fe^{2+}, Fe^{3+}). Further, it was observed single residue embeddings contained sufficient information about the metal-binding capacity of the residue. An easy-to-use notebook to generate quick predictions with a probability of binding ten metal ions for each residue in the protein of interest has also been made available.

Author contributions

AS prepared the datasets. YK developed training architecture and optimised model parameters. AS and YK analysed the results. AS wrote the initial draft of the manuscript. All authors contributed to the manuscript.

Acknowledgements

We thank Petras Kundrotas, Gabriele Pozzati, Ho-Yeung Chim, Wensi Zhu and Sarah Danielsson for their valuable discussions.

Availability and Implementation:

An easy to use Google Colab Notebook is freely availabe at https://bit.ly/40FrRbK source code is available from https://github.com/TeamSundar/m-ionic

Funding

AE was funded by the Vetenskapsrådet Grant No. 2021-03979 and Knut and Alice Wallenberg foundation. The computations/data handling was enabled by the supercomputing resource Berzelius provided by the National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation and SNIC, grant No: SNIC 2021/5-297 and Berzelius-2021-29.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Andreini,C. *et al.* (2009) Metalloproteomes: A Bioinformatic Approach. *Acc. Chem. Res.*, **42**, 1471–1479.

Aptekmann,A.A. *et al.* (2022) mebipred: identifying metal-binding potential in protein sequence. *Bioinformatics*, **38**, 3532–3540.

Aptekmann,A.A. *et al.* (2021) mebipred: identifying metal-binding potential in protein sequence.

Elnaggar,A. *et al.* (2020) ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing Bioinformatics.

Hu,X. *et al.* (2016) Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics*, **32**, 3260–3269.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 1–7.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lippi,M. *et al.* (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094–2095.

Littmann,M. *et al.* (2021) Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.*, **11**, 23916.

Lu,C.-H. *et al.* (2022) MIB2: metal ion-binding site prediction and modeling server. *Bioinformatics*, **38**, 4428–4429.

Maret,W. (2017) Zinc in Cellular Regulation: The Nature and Significance of "Zinc Signals". *Int. J. Mol. Sci.* , 2285.

Palm-Espling,M.E. *et al.* (2012) Role of metal in folding and stability of copper proteins in vitro. *Biochim. Biophys. Acta BBA - Mol. Cell Res.*, **1823**, 1594–

1603.
Pokharel,S. *et al.* (2022) Improving protein succinylation sites prediction using embeddings from protein language model. *Sci. Rep.*, **12**, 16933.
Qiao,L. and Xie,D. (2019) MIonSite: Ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. *Anal. Biochem.*, **566**, 75–88.
Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
Rives,A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.*, **118**.
Stärk,H. *et al.* (2021) Light Attention Predicts Protein Location from the Language of Life Bioinformatics.
Tainer,J.A. *et al.* (1991) Metal-binding sites in proteins. *Curr. Opin. Biotechnol.*, **2**, 582–591.
The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
Yang,J. *et al.* (2013) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
Yu,D.-J. *et al.* (2013) Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 994–1008.
Yuan,Q. *et al.* (2022) Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Brief. Bioinform.*, **23**, bbac444.

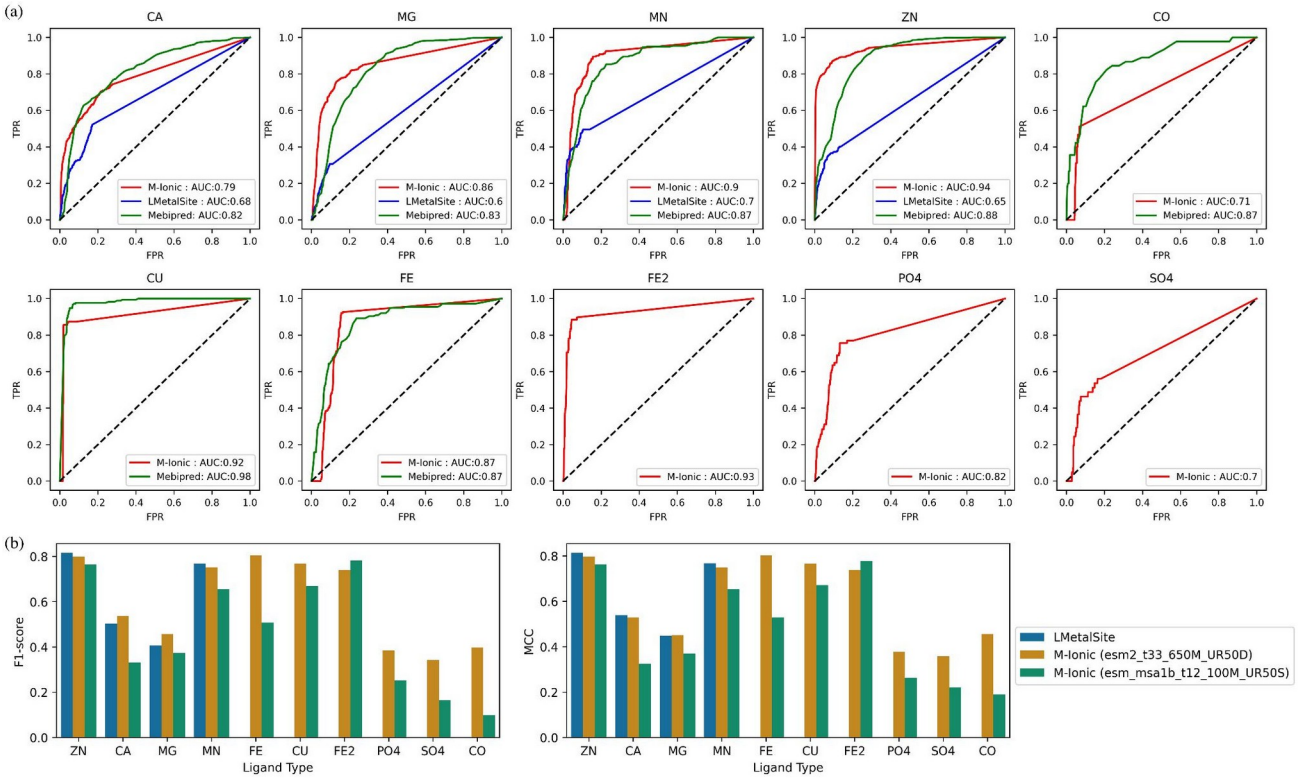


Fig. 1. (a) Protein-level: Comparison of ROC curves for the performance of each ion type for M-Ionic (this study) trained on esm2_t33_650M_UR50D embeddings, LMetalSite (Yuan *et.al.*, 2022) and mebi-pred (Aptekmann *et al.* 2022), (b) Residue-level: MCC scores of performance of M-Ionic (trained on esm2_t33_650M_UR50D and esm_msa1b_t12_100M_UR50S embeddings) and LMetalSite (Yuan *et.al.*, 2022) on homology reduced independent test set

Supplementary file

M-Ionic: Prediction of metal ion binding sites from sequence using residue embeddings

Aditi Shenoy[1†], Yogesh Kalakoti[2†], Durai Sundar[2,3], Arne Elofsson[1*]

[1] Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, 171 21 Solna, Sweden
[2] Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India
[3] Yardi School of Artificial Intelligence, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India

*To whom correspondence should be addressed.
[†]The authors wish it to be known that the first two authors should be regarded as Joint First Authors

Email addresses:

AS: aditi.shenoy@scilifelab.se
YK: yogesh.kalakoti@dbeb.iitd.ac.in
DS: sundar@dbeb.iitd.ac.in
AE: arne@bioinfo.se

Figures

Fig. S1. Overall workflow for M-Ionic.

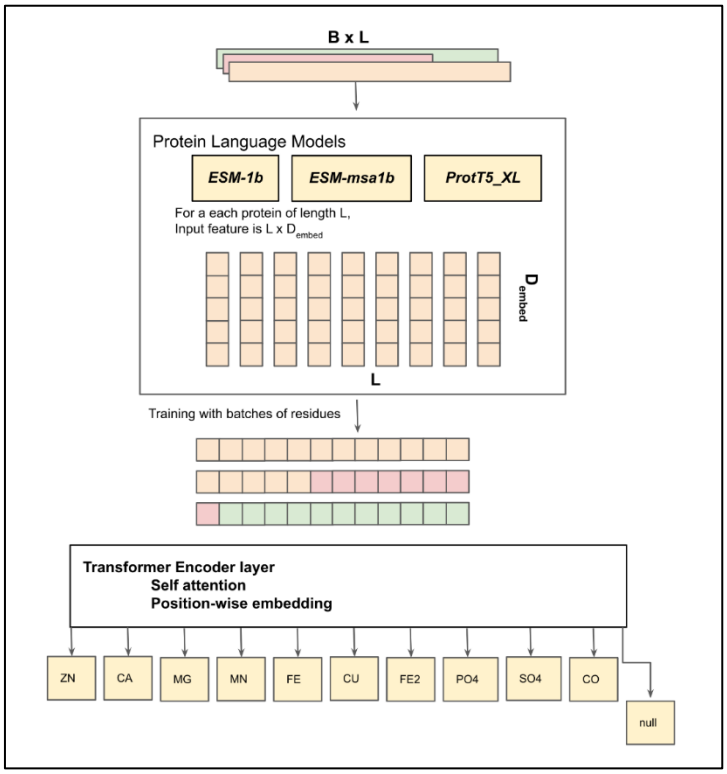
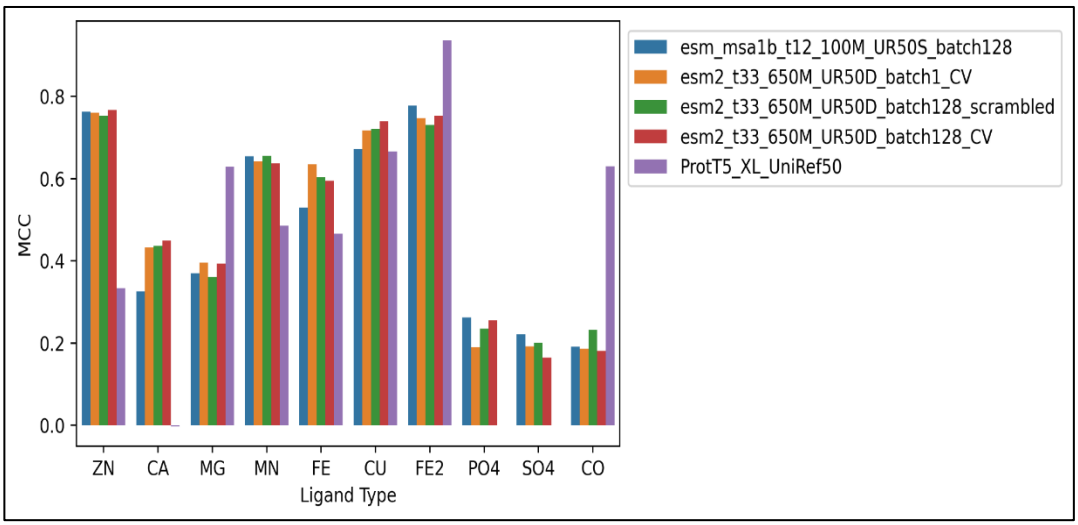


Fig. S2. MCC scores for the various models trained (a) esm_msa1b_t12_100M_UR50S_batch128 where esm_msa1b_t12_100M_UR50S is trained with batch size 128 (b) esm2_t33_650M_UR50D_batch1_CV is trained with esm2_t33_650M_UR50D model with batch size 1 (c) esm2_t33_650M_UR50D_batch128_scrambled trained on esm2_t33_650M_UR50D model with batch size 128 but with the residues within each batch were shuffled randomly along the length dimension such that the positions of the residues within the sequence were changed but the embedding of each residue was untouched (d) esm2_t33_650M_UR50D_batch128_CV is trained with esm2_t33_650M_UR50D model with batch size 128 as in original M-Ionic implementation (e) ProtT5_XL_UniRef50 trained with ProtT5_XL_UniRef50 model with batch size 128.



Tables

Table S1. Summary of dataset employed for training the M-ionic framework.

Ions	Number of protein entries	Number of ions bound with protein receptors	Number of binding residues (P)	Number of non-binding residues (N)	Ratio of binding to non-binding residues (P)/(N)
CA	38258	39147	131250	6041662	0.022
CO	2018	33127	6533	369360	0.018
CU	5617	27348	16257	948823	0.017
FE2	3054	10444	9168	601517	0.015
FE	6366	6038	20311	1227819	0.017
MG	34958	5315	78542	6959299	0.011
MN	11650	2850	34315	2254077	0.015
PO4	2759	2513	12595	605193	0.021
SO4	2365	2291	9840	432458	0.023
ZN	48593	1952	126278	7704568	0.016

Table S2. Performance on distinguishing metal binding and non-binding proteins based independent test set and negative set

Methods	Average precision (%)	F1 score	MCC
mebi-pred	56.5	0.433	0.206
LMetalSite	67.6	0.773	0.482
M-Ionic	86.8	0.892	0.792

Table S3. Benchmark on MIonSite benchmark set

Metal	Method	Sen (%)	Spe (%)	Acc (%)	MCC
Zn	MetalDetector	38.26	99.83	98.22	0.565
	S-SITE	45.14	97.88	96.51	0.387
	TargetS	41.70	99.80	98.29	0.588
	IonSeq	70.04	92.56	91.97	0.347
	MIB	40.12	99.07	97.53	0.451
	COACH	32.39	99.37	97.62	0.422
	IonCom	76.72	95.59	95.10	0.474
	MIonSite	70.65	99.68	98.92	0.771
	LMetalSite	78.72	99.81	99.37	0.836
	M-Ionic	76.64	99.50	99.04	0.76
Ca	MetalDetector	0.62	99.86	98.03	0.016
	S-SITE	10.17	99.51	97.87	0.159
	TargetS	19.71	99.7	98.23	0.322
	IonSeq	0.00	100	98.16	N/A
	MIB	17.61	99.17	97.68	0.213
	COACH	16.18	97.82	96.32	0.122
	IonCom	28.01	99.47	98.16	0.365
	MIonSite	42.53	99.71	98.66	0.552
	LMetalSite	36.02	99.81	98.97	0.506
	M-Ionic	44.41	99.27	98.45	0.45
Mg	MetalDetector	1.57	99.85	98.31	0.043
	S-SITE	35.02	97.28	96.31	0.227
	TargetS	15.16	99.84	98.52	0.298
	IonSeq	0	100	98.44	N/A
	MIB	22.21	99.33	98.12	0.268
	COACH	21.12	97.6	96.41	0.144
	IonCom	24.12	99.25	98.07	0.276
	MIonSite	24.35	99.69	98.51	0.362
	LMetalSite	24.88	99.87	99.03	0.410
	M-Ionic	29.01	99.59	98.57	0.38
Mn	MetalDetector	21.05	99.72	98.69	0.319
	S-SITE	69.47	98.38	98.01	0.493
	TargetS	28.42	99.75	98.82	0.408

	IonSeq	2.11	99.94	98.67	0.081
	MIB	47.75	99.52	98.84	0.514
	COACH	27.01	99.82	98.87	0.420
	IonCom	54.06	99.39	98.80	0.534
	MIonSite	57.27	99.40	98.84	0.558
	LMetalSite	67.78	99.81	99.41	0.740
	M-Ionic	58.46	99.8	99.29	0.68
Fe	MetalDetector	28.57	99.76	99.24	0.360
	S-SITE	90.48	98.78	98.71	0.560
	TargetS	28.57	99.55	99.03	0.296
	IonSeq	80.95	96.85	96.73	0.350
	MIB	52.63	97.83	97.49	0.277
	COACH	77.52	99.73	99.57	0.725
	IonCom	77.12	99.80	99.64	0.756
	MIonSite	82.90	99.75	99.63	0.765
	LMetalSite	95.00	99.93	99.90	0.927
	M-Ionic	80	99.78	99.62	0.77
Cu	MetalDetector	58.33	99.91	99.47	0.712
	S-SITE	25.00	97.41	96.64	0.138
	IonSeq	41.67	99.11	98.5	0.365
	MIB	83.33	98.03	97.88	0.503
	COACH	50.00	97.14	96.64	0.268
	IonCom	50.00	99.55	99.03	0.517
	MIonSite	48.02	99.63	99.08	0.523
	LMetalSite	81.82	100.0	99.82	0.904
	M-Ionic	54.55	100	99.56	0.74
Fe2	MetalDetector	33.33	99.86	99.04	0.496
	S-SITE	66.67	95.54	95.18	0.309
	IonSeq	97.8	99.28	99.26	0.782
	MIB	100	99.44	99.45	0.830
	COACH	66.67	98.05	97.66	0.437
	IonCom	100	99.44	99.45	0.830
	MIonSite	94.98	99.56	99.5	0.830
	LMetalSite	100.0	100.0	100.0	1.000
	M-Ionic	100	100	100	1
Co	MetalDetector	16.67	99.58	98.69	0.217

S-SITE	55.21	84.68	84.36	0.113
MIB	33.33	95.48	94.81	0.138
COACH	53.21	91.56	91.14	0.162
MIonSite	58.77	92.58	92.22	0.195
LMetalSite	22.22	100.0	99.17	0.469
M-Ionic	0	100	98.6	0