The Mythical Metal – Insights on the Accuracy of Metal Identification in Structural Biology

Edward H. Snell[1,2], Elspeth F. Garman[3], Geoffrey Grime[4], Aina Cohen[5] and Sarah Bowman[1]

1 Hauptman-Woodward Medical Research Institute, 700 Ellicott St., Buffalo, New York 14203, United States, 2Department of Materials Design and Innovation, SUNY University at Buffalo, 700 Ellicott St., Buffalo, New York 14203, United States, 3 Department of Biochemistry, South Parks Road, Oxford, OX1 3QU, UK, 4 Ion Beam Centre, Advanced Technology Institute, University of Surrey, Guildford, Surrey GU2 7XH, UK., 5 Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory2575 Sand Hill Road, Menlo Park, CA 94025, USA

The Protein Data Bank (PDB) contains information more than 200,000 proteins. Some 25-50% of all proteins contain one or more metals or metal co-factors playing a key structural or catalytic or role. Metalloproteins can be difficult targets for structural studies because the specific metals required for enzymatic activity are not always those with the highest binding affinity. To incorporate uncompetitive metals into an enzyme active site, cells have evolved sophisticated machinery, including metal transporters and chaperones, which are often bypassed during protein production in the laboratory. The problem of metal promiscuity is compounded by the difficulty to accurately identify incorporated metals. In many cases, the small sample volumes available confound the issue.

Using a small subset of 32 metalloproteins with models deposited in the Protein Data Bank (PDB) we observed that up to 50% of models do not accurately reflect the metal in the sample protein. In one of these cases studied in detail, we used the experimental and theoretical anomalous signal ratios to position the 3 different metals detected by PIXE and re-refined the PDB deposited structure. This revealed a previously missed ligandin the active site and therefore new biological information. Further studies on an additional 50 different metalloproteins demonstrated that this trend is common. A computational analysis performed on the entire PDB as of December 2021, comparing the modelled metal to the difference electron density from deposited experimental data, suggests that this error is more the norm than the exception with 33% of ~160,000 metal sites showing concern. A metal incorrectly modelled can lead to misunderstandings of mechanism and function.

While many laboratories specializing in metalloprotein studies take extreme care to characterize the metal present, the problem is significant, the majority of metalloprotein models do not come from metalloprotein laboratories. In 2021, there were 1,465,753,416 models downloaded from the PDB and conservatively, ~300 million of those were metalloproteins. The PDB curators estimate that over 99% of downloads are not by structural biologists and the statistical data show that only ~4% are downloaded with their experimental data. Many researchers who use these models, both inside and outside the field of structural biology, may be unaware of these errors or even how to detect them. With the growth of machine learning techniques, any large scale error in metal identification in the training set will propagate into predicted models.

In this presentation we will describe the findings above and discuss the development of sensitive X-ray techniques using small sample volumes on existing crystallography beamlines that build on the PIXE studies and computational analysis and provide the capability for routine metal identification with minimal sample volume.