

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

Edward Snell – Hauptman-Woodward Medical Research Institute



Metals in Biology

From the MetalsPDB Website out of 150,149 models examined in the Protein Data Bank, 57,494 of those models have metals (over 38%)

The top six are:

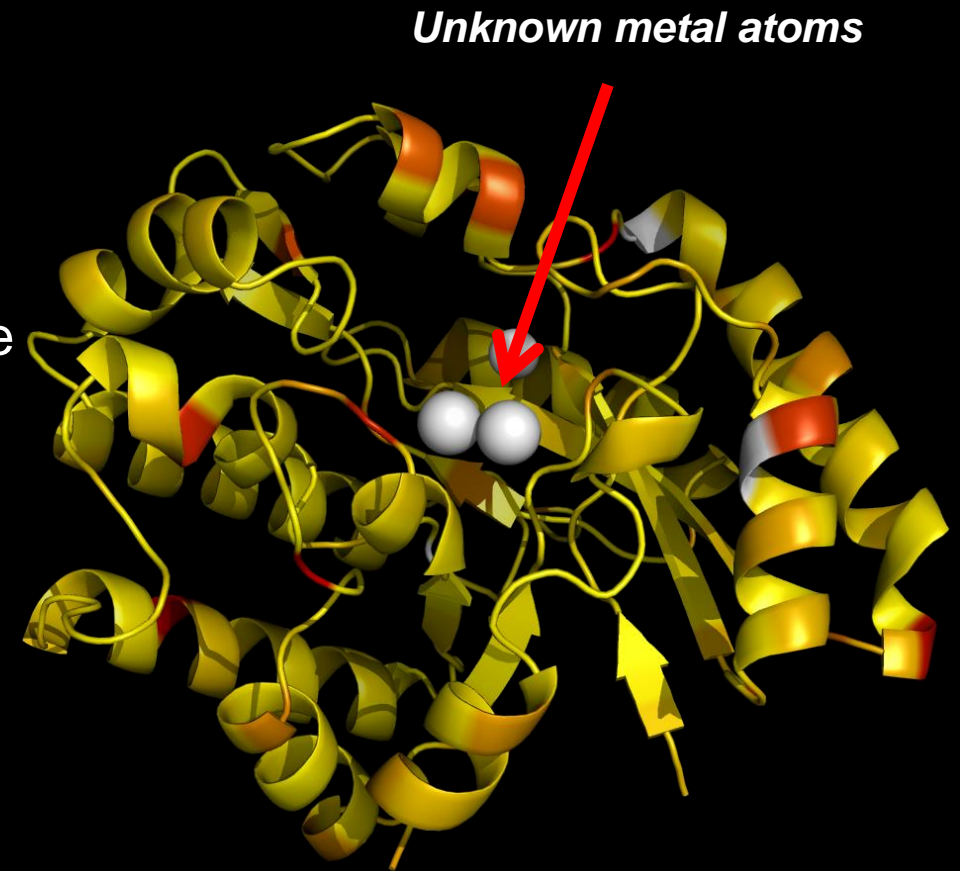
- Zn (~26%)
- Mg (~25%)
- Ca (~19%)
- Fe (~15%)
- Na (~14%)
- Mn (~6%).

Over 86% of the sites are mono-nuclear, ~10% bi-nuclear, just over 2% tri-nuclear and the remainder greater.



Metalloproteins

- Many proteins (especially enzymes) contain small numbers of metal atoms
 - Often critical for mechanism
 - Also common for them to help in maintaining structure
 - Present in 30-40% of all proteins
- X-ray crystallography measures electron density
 - But ... cannot accurately determine atomic number unless anomalous techniques are used
 - The metal is often inferred indirectly from knowing initial conditions, looking to homologous structures, examining geometry, and or molecular modeling



PROBLEM



The majority of metalloprotein structural models produced by metalloprotein groups are accurate (not the problem)

However, the majority of metalloprotein structural models are not produced by metalloprotein groups (the problem)

How do we know there is a problem?

- In 1994, 25% of the models in the protein data bank also had the experimental data deposited.
- By 1996, that reached over 50%.
- In 2008 it became mandatory to deposit the experimental data supporting any model that was produced.
- With the experimental data we can visualize the problem by looking at the data in addition to the model.

Two maps are typically used produced with the observed structure factors, F_o , and the calculated structure factors F_c .

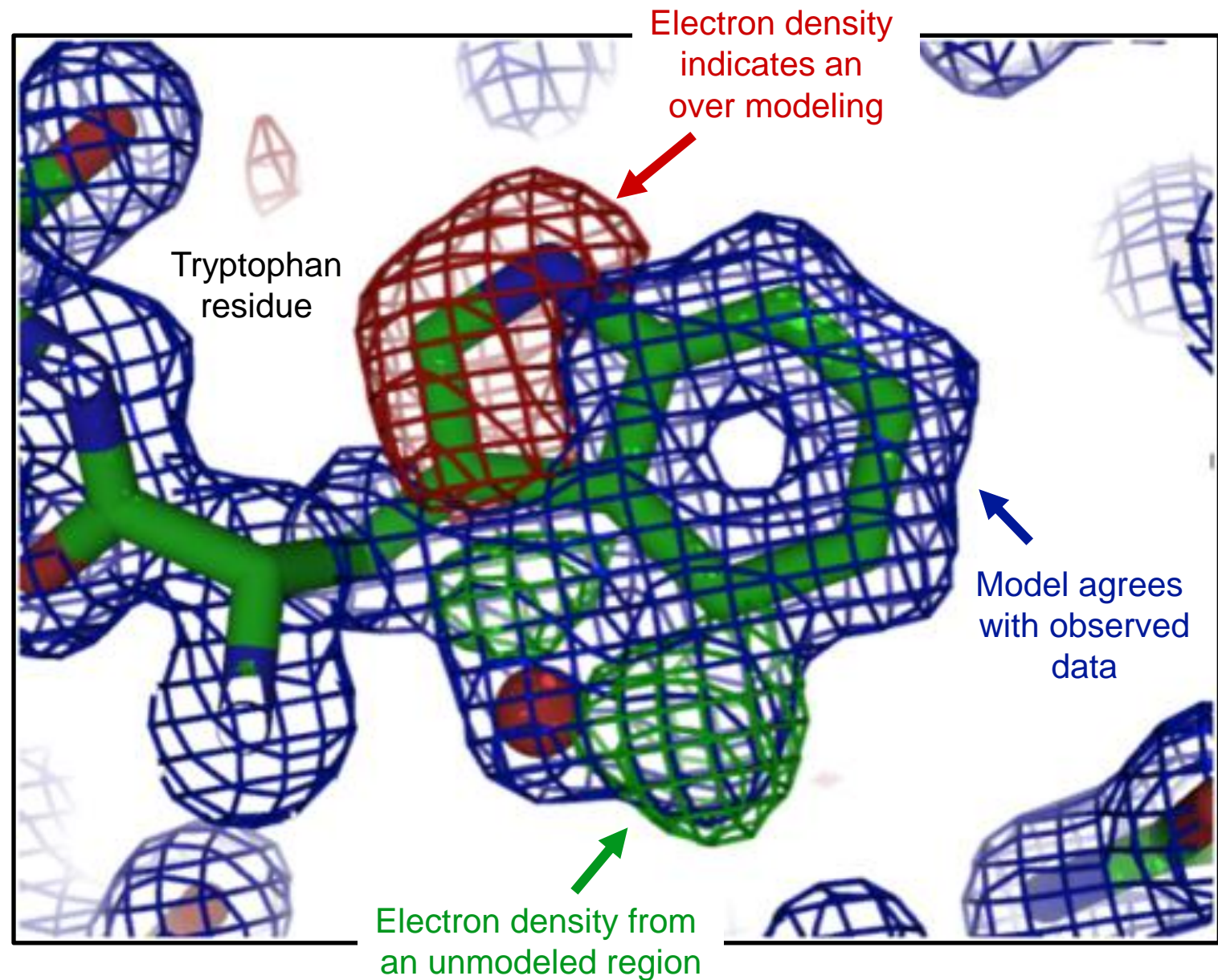
These maps are:

- The $2F_o$ map (usually displayed in blue)
- The $F_o - F_c$ map (positive in green, negative in red)

The X-ray scattering by any atom is proportional to the square of its atomic number

Where the model is wrong, the $F_o - F_c$ map will have:

- Positive density (green) if there are atoms that have not been modeled or have too low an atomic number
- Negative density (red) if there is a modeled atom that is not there, if it has an occupancy less than the model, or if it is a lighter atom.



How do we make use of this?

We built a suite of routines called Alchemy (*a kind of thinking that leads to a way of understanding* – Marcel Duchamp, 1887-1968)

1. Read the structural model and experimental scattering factor data
2. Calculate $2F_o - F_c$ and $F_o - F_c$ map for the model and data (if present, calculate the anomalous map).
3. Integrate the difference data in a sphere around atoms of interest and calculate the real space Z-score.
4. Automagically produce images of the maps and model around the metal environment.
5. Tabulate the results.

In the existing data, is there a problem and how big is it?


- Give Alchemy a list of every metalloprotein structure deposited in the PDB.
- Let it run for about 1 month (mostly image generation).
- Tabulate the results.





Examples of the Alchemy output

Examples
from metalloprotein
Groups (even they
have problem days)



The *PDB_REDO* server for macromolecular structure model optimizationRobbie P. Joosten,^{a*} Fei Long,^b Garib N. Murshudov^b and Anastassis Perrakis^{a*}^aDivision of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, and ^bStructural Studies Division, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0HQ, England. *Correspondence e-mail: r.joosten@nki.nl, a.perrakis@nki.nlReceived 4 March 2014
Accepted 25 April 2014

Edited by E. N. Baker, University of Auckland, New Zealand

Keywords: *PDB_REDO*; validation; model optimization

The refinement and validation of a crystallographic structure model is the last step before the coordinates and the associated data are submitted to the Protein Data Bank (PDB). The success of the refinement procedure is typically assessed by validating the models against geometrical criteria and the diffraction data, and is an important step in ensuring the quality of the PDB public archive [Read *et al.* (2011), *Structure*, **19**, 1395–1412]. The *PDB_REDO* procedure aims for 'constructive validation', aspiring to consistent and optimal refinement parameterization and pro-active model rebuilding, not only correcting errors but striving for optimal interpretation of the electron density. A web server for *PDB_REDO* has been implemented, allowing thorough, consistent and fully automated optimization of the refinement procedure in *REFMAC* and partial model rebuilding. The goal of the web server is to help practising crystallographers to improve their model prior to submission to the PDB. For this, additional steps were implemented in the *PDB_REDO* pipeline, both in the refinement procedure, *e.g.* testing of resolution limits and *k*-fold cross-validation for small test sets, and as new validation criteria, *e.g.* the density-fit metrics implemented in *EDSTATS* and ligand validation as implemented in *YASARA*. Innovative ways to present the refinement and validation results to the user are also described, which together with auto-generated *Coot* scripts can guide users to subsequent model inspection and improvement. It is demonstrated that using the server can lead to substantial improvement of structure models before they are submitted to the PDB.

1. Introduction

Crystallographic structure elucidation is a stepwise process with many decision points, and is therefore complex and labour-intensive. Over the years, this process has become more and more streamlined by automation. The crystallographic process, starting from the diffraction experiment itself, has greatly benefitted from faster computers and advances in crystallographic software. Automated pipelines are available for data reduction (*e.g.* Otwinowski & Minor, 1997; Vonrhein *et al.*, 2011; Krug *et al.*, 2012; Monaco *et al.*, 2013; Winter *et al.*, 2013), experimental phasing (*e.g.* Panjikar *et al.*, 2005; Terwilliger *et al.*, 2009; Pannu *et al.*, 2011), molecular replacement (*e.g.* Keegan & Winn, 2007; Long *et al.*, 2008; McCoy *et al.*, 2007), density-map tracing and model building (*e.g.* Perrakis *et al.*, 1999; Ioerger & Sacchettini, 2002; Cowtan, 2006; Terwilliger *et al.*, 2008) and combinations thereof (*e.g.* Brunzelle *et al.*, 2003; Holton & Alber, 2004; Kroemer *et al.*, 2004).

The *PDB_REDO* pipeline (Joosten *et al.*, 2012) focuses on automating the final steps of the crystallographic process, *i.e.*



Ian J. Tickle

Astex Pharmaceuticals, 436 Science Park,
Milton Road, Cambridge CB4 0QA, England

Correspondence e-mail: ian.tickle@astx.com

Statistical quality indicators for electron-density maps

The commonly used validation metrics for the local agreement of a structure model with the observed electron density, namely the real-space *R* (RSR) and the real-space correlation coefficient (RSCC), are reviewed. It is argued that the primary goal of all validation techniques is to verify the accuracy of the model, since precision is an inherent property of the crystal and the data. It is demonstrated that the principal weakness of both of the above metrics is their inability to distinguish the accuracy of the model from its precision. Furthermore, neither of these metrics in their usual implementation indicate the statistical significance of the result. The statistical properties of electron-density maps are reviewed and an improved alternative likelihood-based metric is suggested. This leads naturally to a χ^2 significance test of the difference density using the real-space difference density *Z* score (RSZD). This is a metric purely of the local model accuracy, as required for effective model validation and structure optimization by practising crystallographers prior to submission of a structure model to the PDB. A new real-space observed density *Z* score (RSZO) is also proposed; this is a metric purely of the model precision, as a substitute for other precision metrics such as the *B* factor.

Received 10 June 2011
Accepted 2 September 2011

1. Background

Global metrics of accuracy of the structure model (such as R_{free}) do not identify local errors in a model. A better metric of local accuracy of the model is consistency with the electron density in real space. This assumes that the electron density itself, and therefore the phases from which it is derived, are accurate. This is a reasonable assumption because density-based validation is normally performed near the completion of refinement when the model is mostly correct and only a small number of minor errors remain to be resolved.

2. Outline

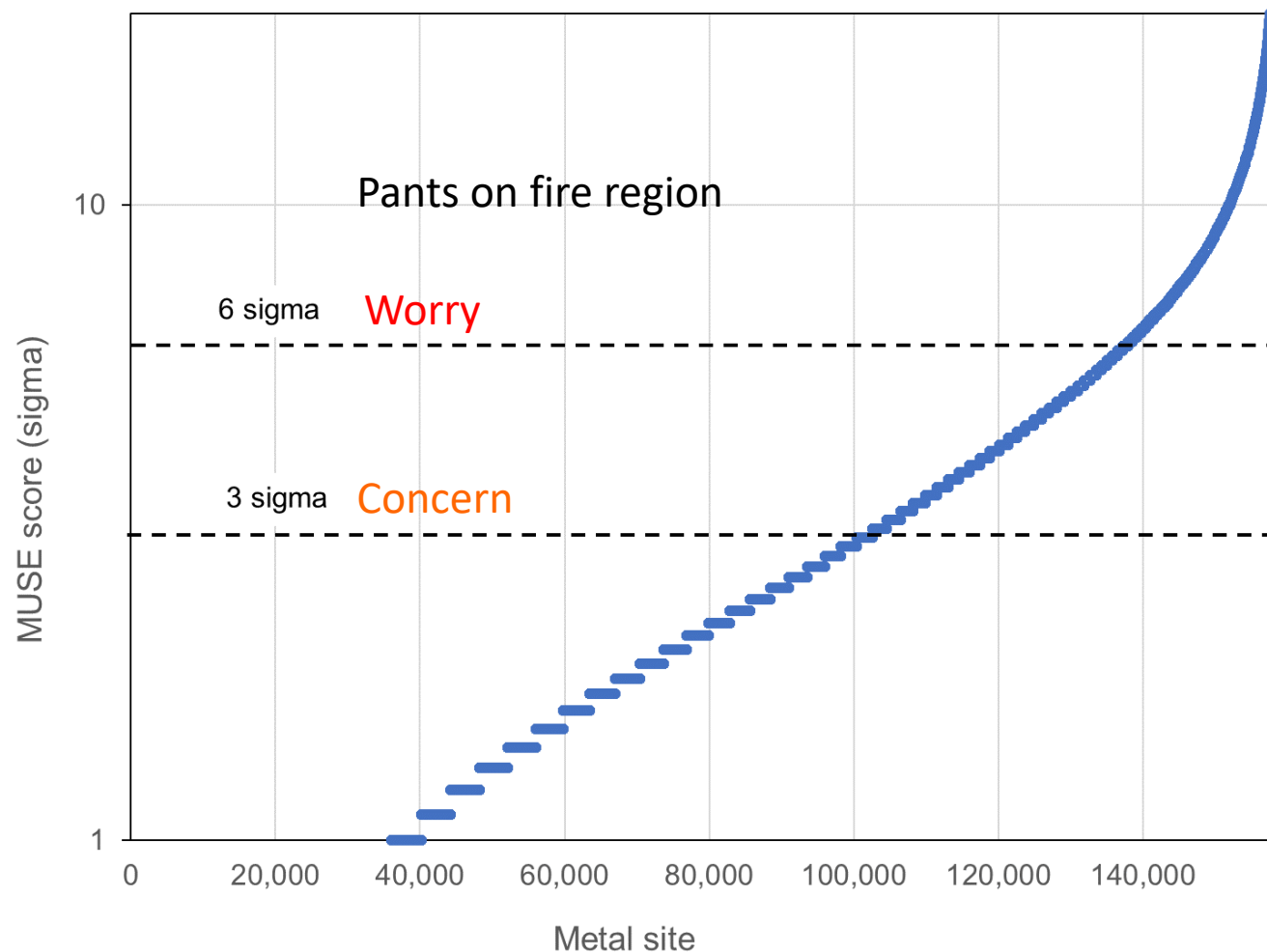
2.1. Review existing real-space electron-density metrics

- Real-space *R* (RSR).
- Real-space correlation coefficient (RSCC).
- Why *both* these metrics are sub-optimal as validation metrics.
- What are the characteristics of an optimal metric?

2.2. Other issues related to current implementations of RSR and RSCC

The sensitivity of any real-space metric of electron density depends critically on the following.

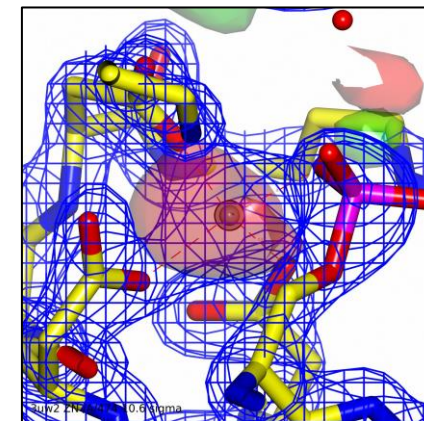
All metal sites in the PDB



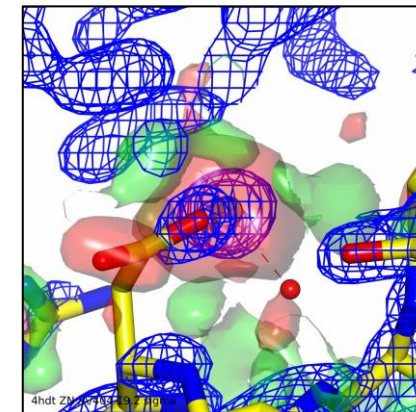
35,287 individual PDB models. 158,791 metal sites

MUSE (Metal Uncertainty ScoreE) = $\text{Max}(|\Delta\rho|)/\sigma(\Delta\rho)$ = real space difference density score

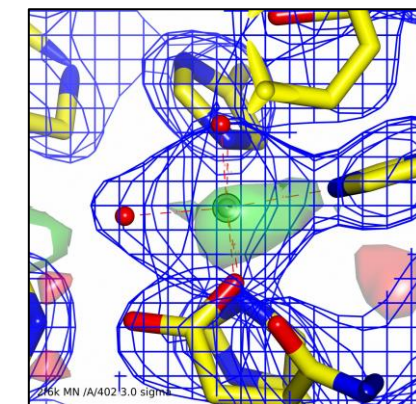
6,607 (4%) > 10 sigma
Just plain wrong!



20,764 (13%) > 6 sigma
Worry, need to be corrected



56,109 (35%) > 3 sigma
Concern and should be verified



Sodium

18,341 Na sites
24% > 3 sigma
8% > 6 sigma
2% >10 sigma

Calcium

31,429 Ca sites
43% > 3 sigma
18% > 6 sigma
7% >10 sigma

Nickel

2,749 Ni sites
56% > 3 sigma
29% > 6 sigma
12% >10 sigma

Zinc

31,429 Zn sites
43% > 3 sigma
18% > 6 sigma
7% >10 sigma

Magnesium

56,860 Mg sites
35% > 3 sigma
12% > 6 sigma
3% >10 sigma

Manganese

9,765 Mn sites
35% > 3 sigma
12% > 6 sigma
4% >10 sigma

Cobalt

1,544 Co sites
52% > 3 sigma
25% > 6 sigma
10% >10 sigma

Similar numerical trends are seen with all metal sites.

The worst cases are Ni and Co.

Potassium

6,348 K sites
32% > 3 sigma
11% > 6 sigma
2% >10 sigma

Iron

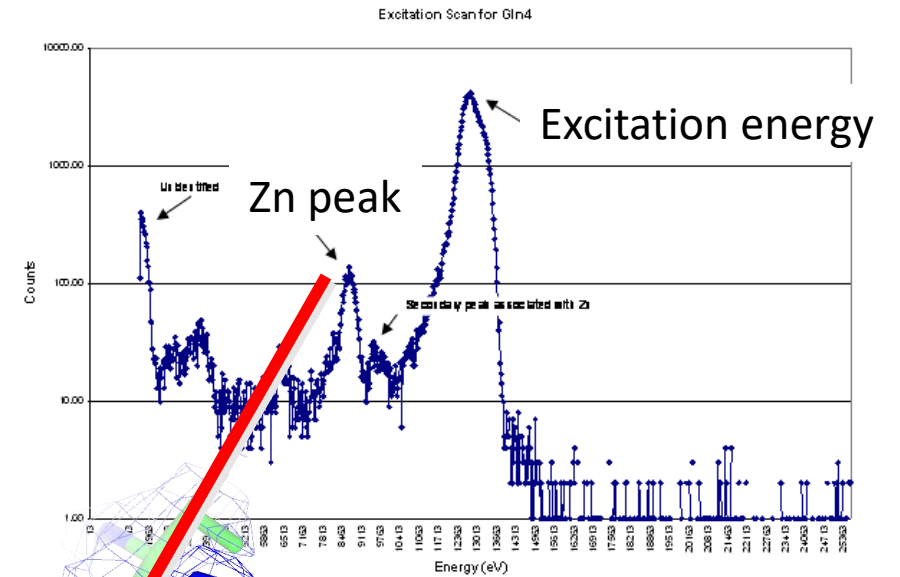
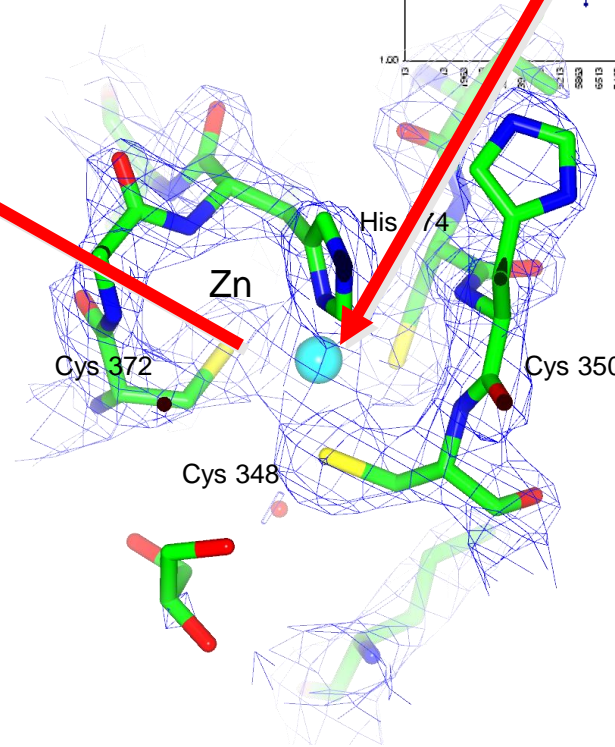
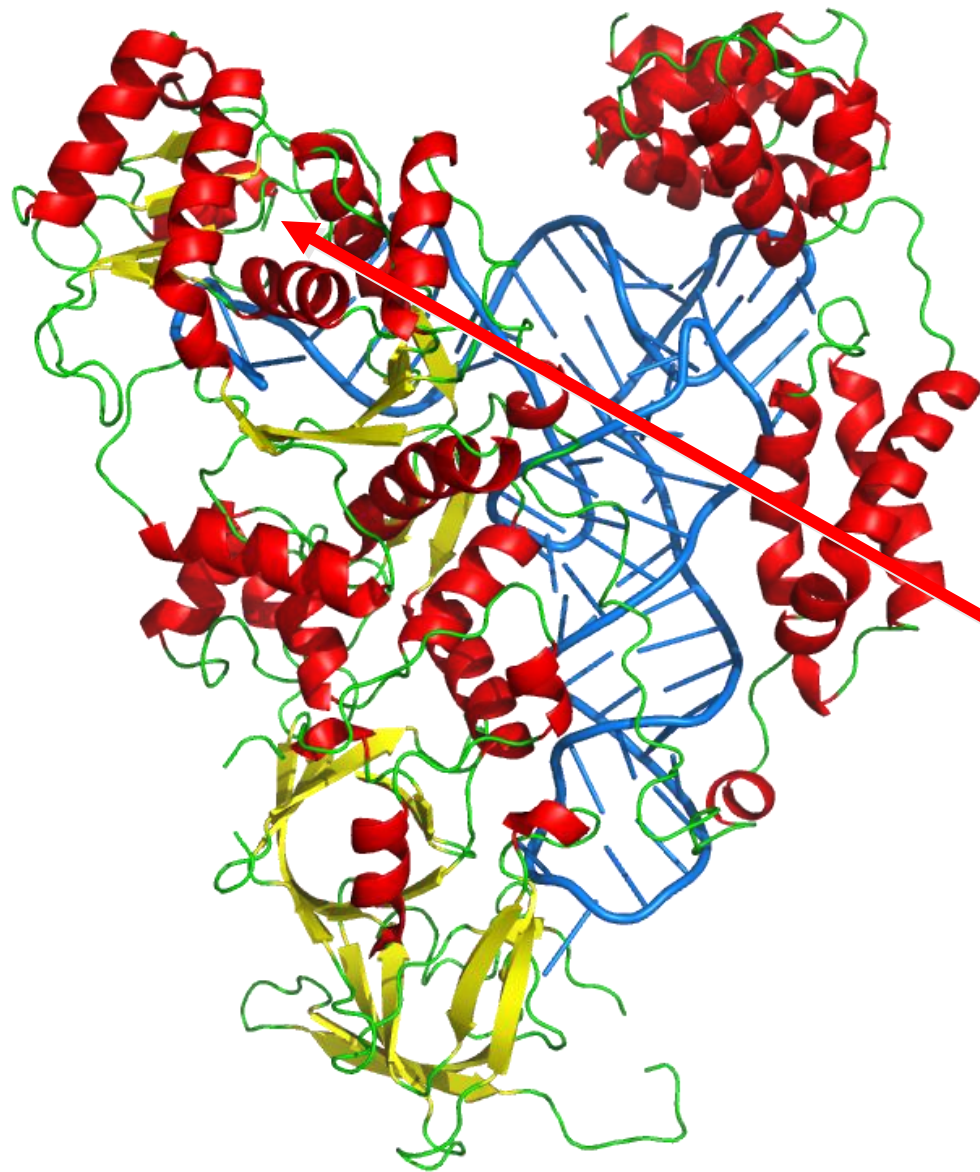
3,179 Fe sites
40% > 3 sigma
16% > 6 sigma
6% >10 sigma

Copper

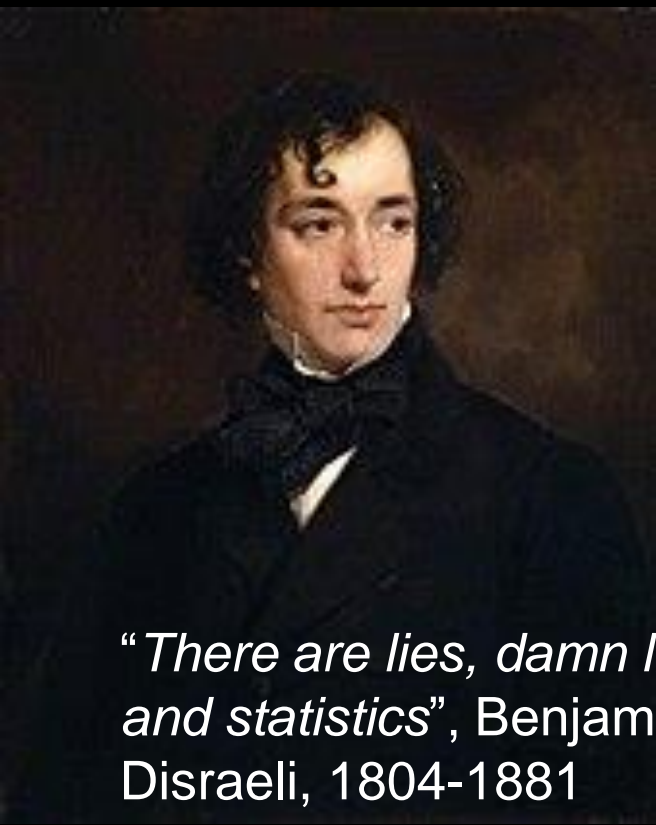
3,212 Cu sites
42% > 3 sigma
16% > 6 sigma
5% >10 sigma

The trends are worrying.

A good experiment can identify the metal

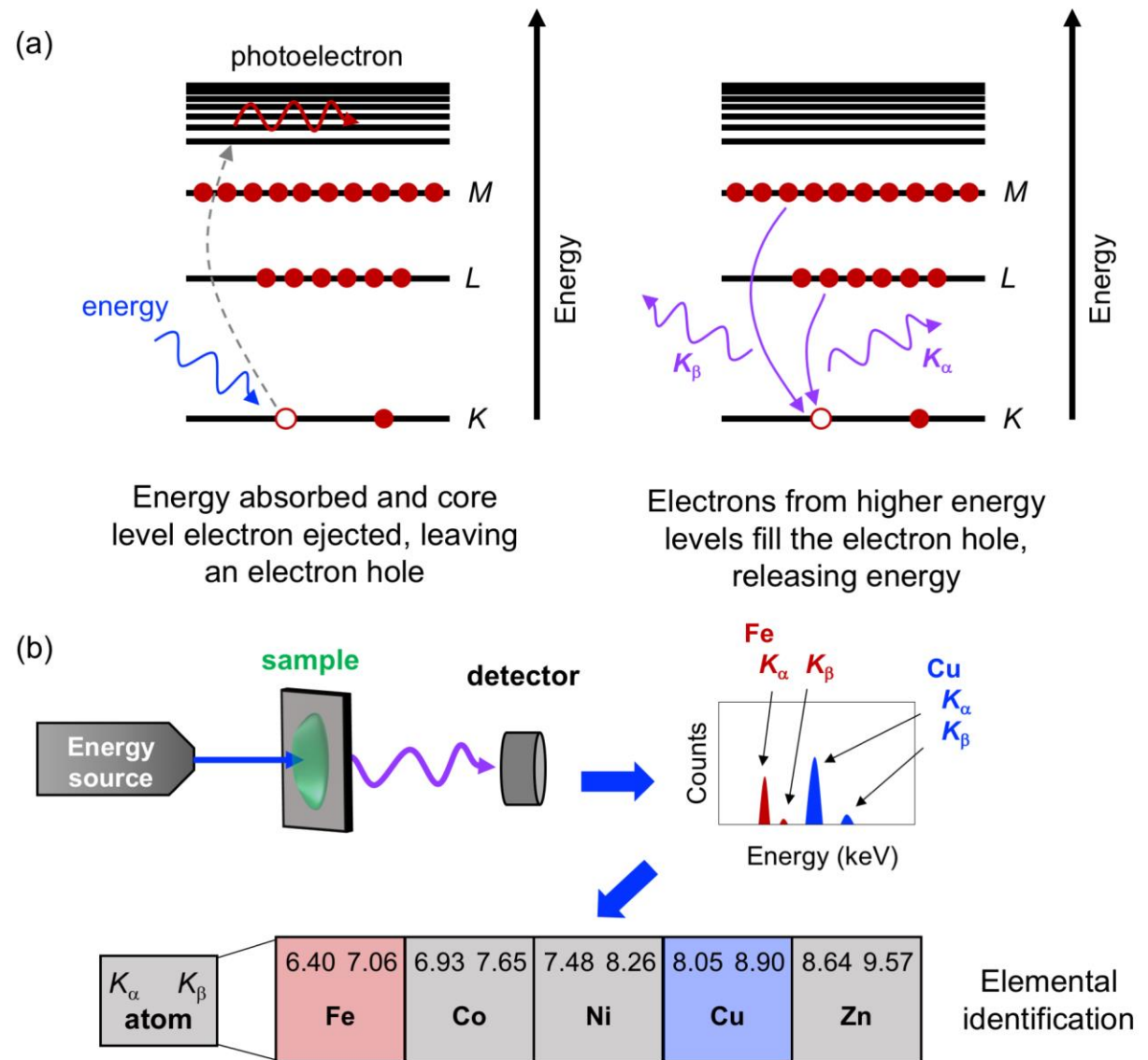


Measurements before or during data collection can identify if a metal is present and what that metal is.



“There are lies, damn lies, and statistics”, Benjamin Disraeli, 1804-1881

Computational analysis of data from others is great ... but let's do our own experiment



Use an atomic technique to directly measure the metal

Graphic from Sarah Bowman



PIXE

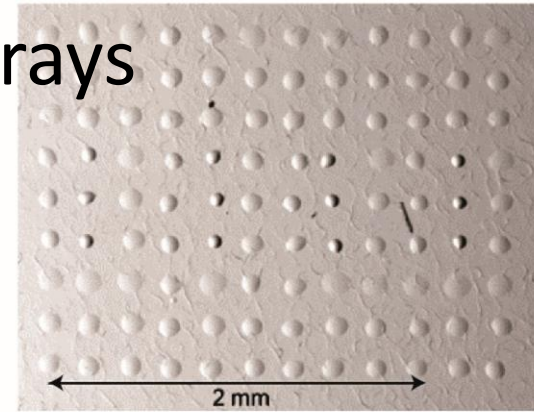
(not PIXIE)



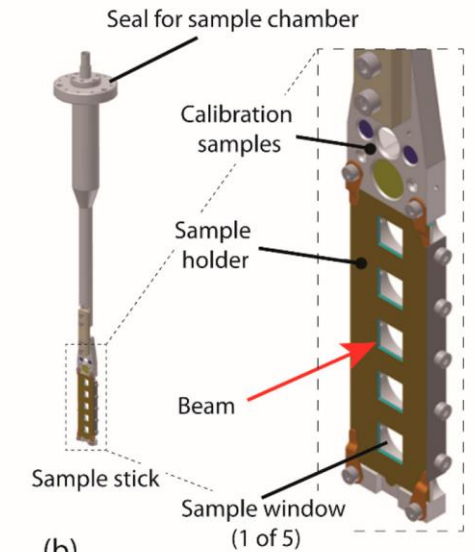
PIXE - Particle Induced Emission of X-rays

An atomic technique independent of the state of the sample

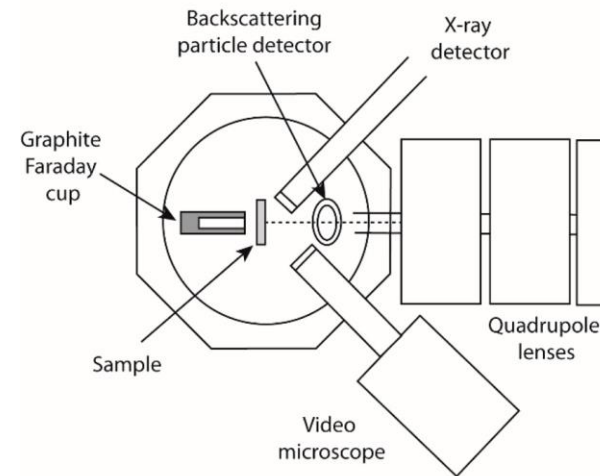
The Surrey Ion Beam Centre in the UK provides the only proton microprobe facility in the world with developed capability for high-throughput protein analysis.



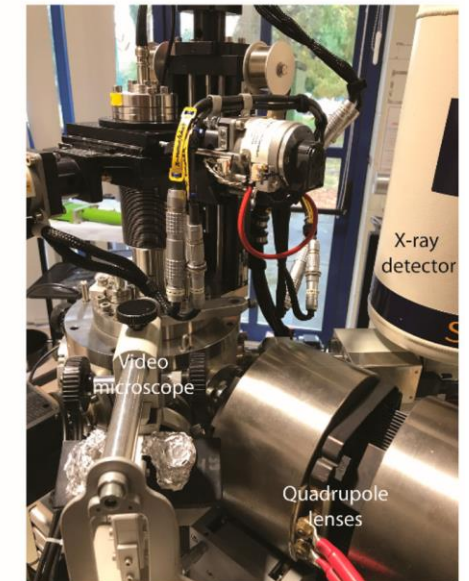
(a)



(b)



(c)



(d)

Particle Induced X-ray Emission (PIXE) (and energy dispersive X-ray emission spectroscopy (EDX)) require only pico- or nanoliter sample volumes.

The PIXE experimental setup

High-Throughput PIXE as an Essential Quantitative Assay for Accurate Metalloprotein Structural Analysis: Development and Application

Geoffrey W. Grime,[†] Oliver B. Zeldin,[‡] Mary E. Snell,[§] Edward D. Lowe,[‡] John F. Hunt,[⊥] Gaetano T. Montelione,[#] Liang Tong,[⊥] Edward H. Snell,^{*,§,||} and Elspeth F. Garman^{*,‡}

[†]Ion Beam Centre, Advanced Technology Institute, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom

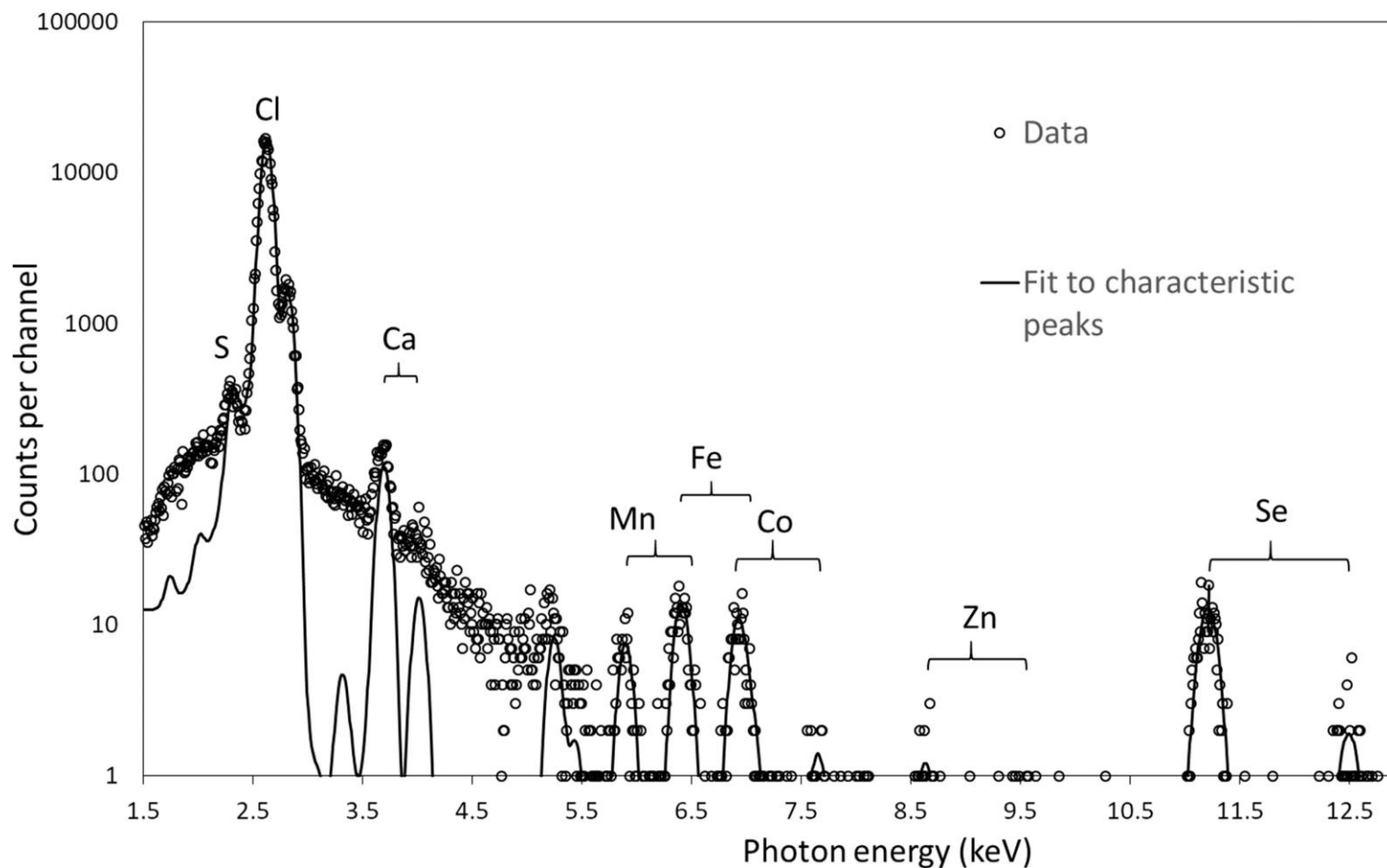
[‡]Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom

[§]Hauptman-Woodward Medical Research Institute, 700 Ellicott St., Buffalo, New York 14203, United States

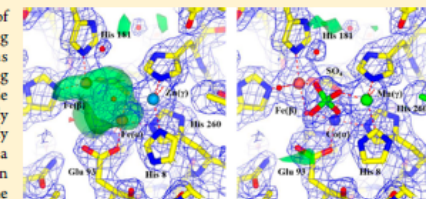
^{||}Materials Design and Innovation, SUNY Buffalo, 700 Ellicott St., Buffalo, New York 14203, United States

[⊥]Department of Biological Sciences, Columbia University, New York, New York 10027, United States

[#]Department of Chemistry and Chemical Biology, Center for Biotechnology and Interdisciplinary Sciences, Rensselaer Polytechnic Institute, Troy New York 12180 United States



ABSTRACT: Metalloproteins comprise over one-third of proteins, with approximately half of all enzymes requiring metal to function. Accurate identification of these metal atoms and their environment is a prerequisite to understanding biological mechanism. Using ion beam analysis through particle induced X-ray emission (PIXE), we have quantitatively identified the metal atoms in 30 previously structurally characterized proteins using minimal sample volume and a high-throughput approach. Over half of these metals had been misidentified in the deposited structural models. Some of the PIXE detected metals not seen in the models were explainable as artifacts from promiscuous crystallization reagents. For others, using the correct metal improved the structural models. For multinuclear sites, anomalous diffraction signals enabled the positioning of the correct metals to reveal previously obscured biological information. PIXE is insensitive to the chemical environment, but coupled with experimental diffraction data deposited alongside the structural model it enables validation and potential remediation of metalloprotein models, improving structural and, more importantly, mechanistic knowledge.



1. INTRODUCTION

Metals are important in biology with over one-third of all proteins having one or more metal ligands playing a key structural or catalytic role¹ critical for the progression of many diseases and attractive for therapeutic intervention.² The correct identity and accurate stoichiometry of these ligands are vital biophysical data for characterizing proteins and understanding mechanism, but there is currently no widely accepted standard metal assay. If the structural model is known, circumstantial evidence from the model is used, but this has been shown to be unreliable, particularly at low resolutions.³ For models determined by X-ray crystallography or similar resolution techniques, the choice of metal made at the refinement stage affects the restraints, biasing the final structure. There are sophisticated techniques that make use of anomalous signals in the structure factors which allow for element identification^{4,5} independent of particle induced X-ray emission (PIXE) data. However, these require a knowledge of the expected elemental species and use multiple refinements, comparing the models produced for the different species, or

the use of multiple incident X-ray wavelengths. In the absence of structural information is available, bioinformatic approaches can be used,⁶ but experimental measurement is not part of routine characterization protocols.

To identify and quantitate the metal content of proteins, the limit of detection (LoD) required is a single metal atom in a 100 kDa macromolecule or 500 ppm by dry weight. Current methods for metal identification (reviewed by Hare et al.⁷) include wet assays,⁸ mass spectrometry⁹ and X-ray¹⁰ or electron based characterization¹¹ and imaging,¹² and various combinations of these.^{13,14} The former rely on detecting chemical compounds bound to metal atoms but can only detect one species at a time. Mass spectrometry is sensitive, but results can be degraded by partial occupancies, glycosylation, or post-translational modifications. X-ray absorption spectroscopy (XAS) requires a high sample volume and has stringent experimental requirements.¹⁵ Electron induced X-ray emission

Received: August 24, 2019

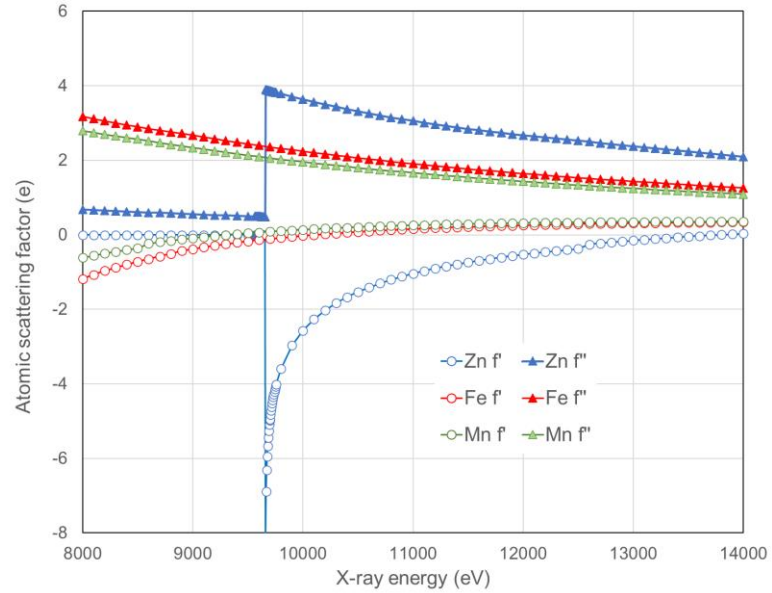
Published: December 3, 2019

Grime et al., *J. Am. Chem. Soc.* 2020, 142, 185–197

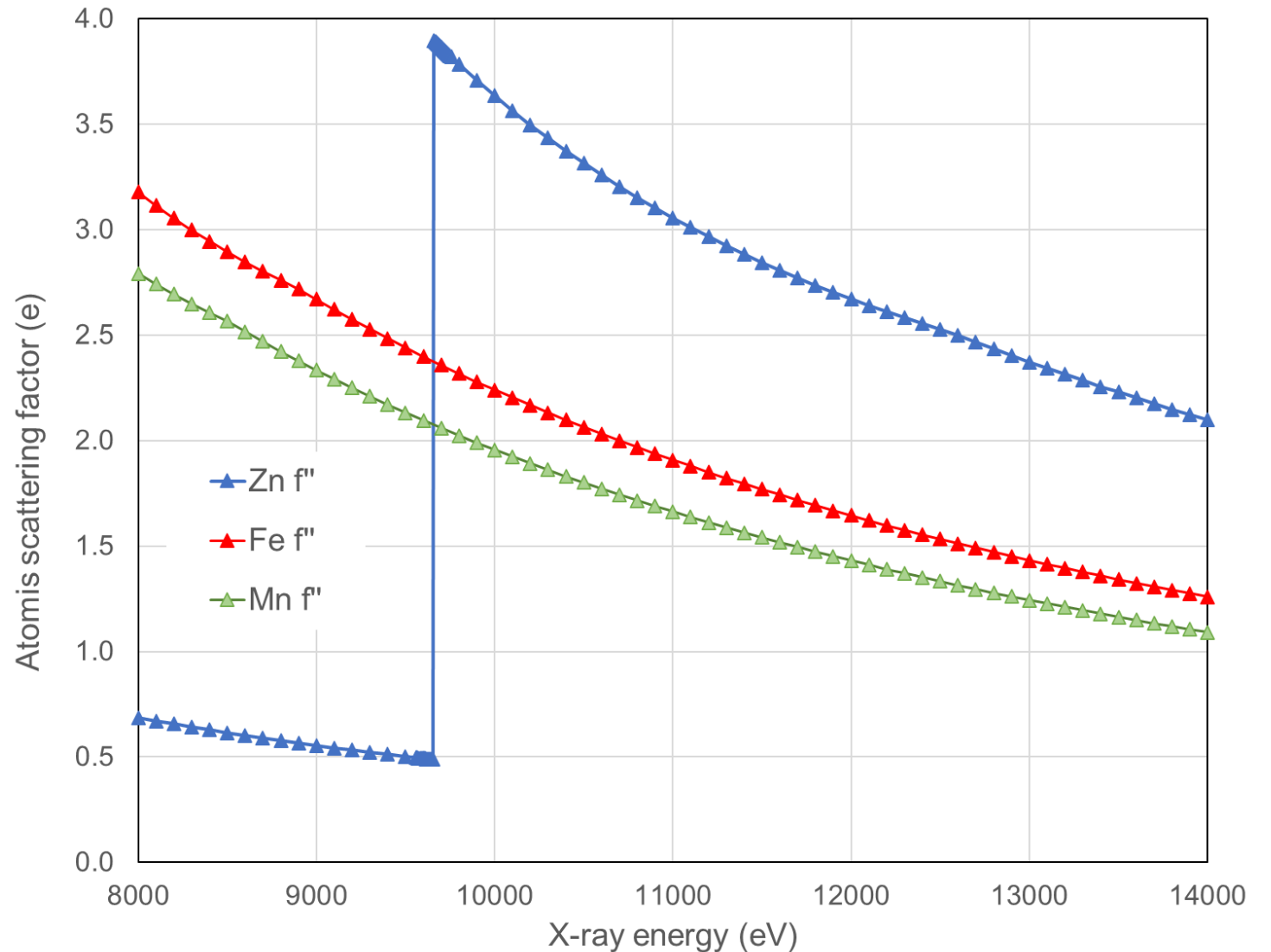
DOI: (10.1021/jacs.9b09186)

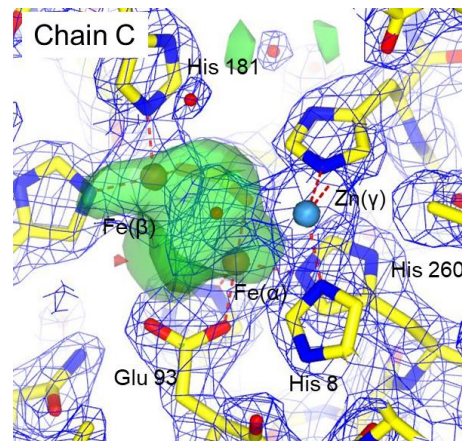
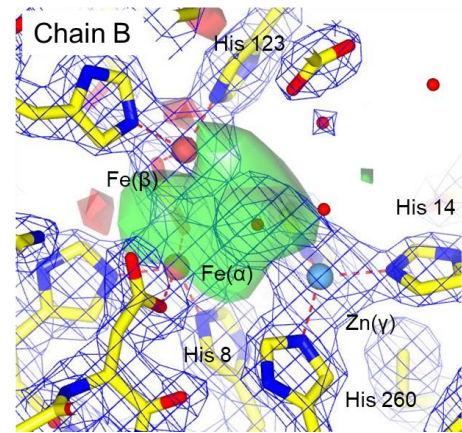
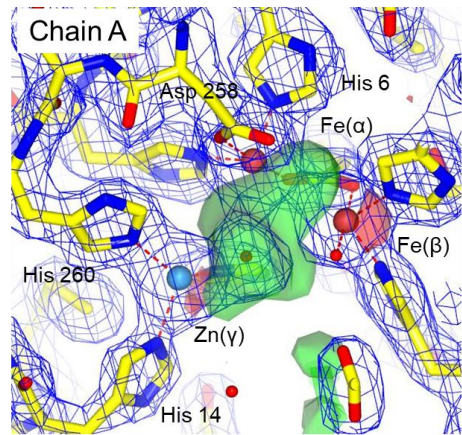
After identifying the metals combine with X-ray
crystallographic data

Different atoms have different scattering factors as a function of energy (wavelength)



And these scattering factors can be calculated as a function of energy





Incorrectly modeled metal

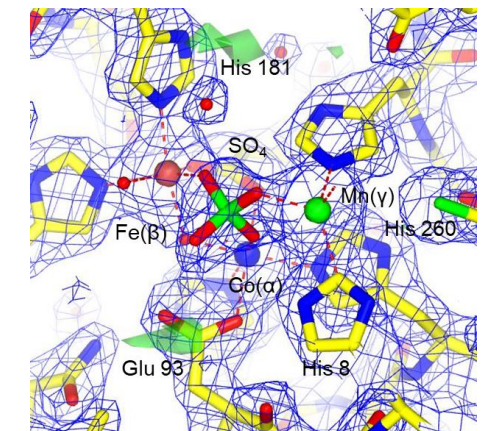
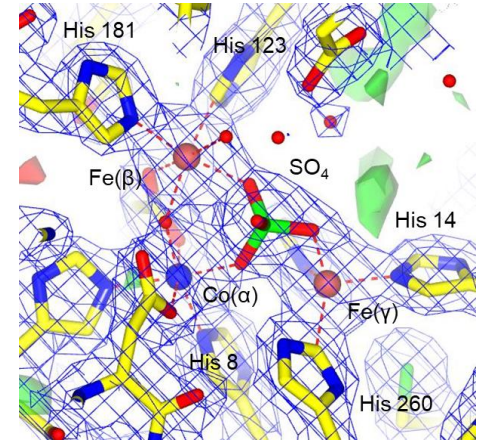
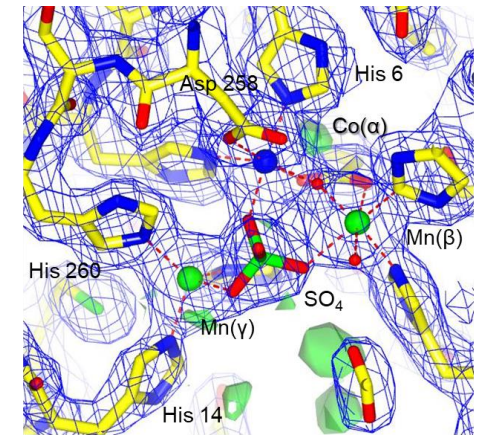
Calculated

	N	f'' at 0.979 Å	f'' normalized to					
			Se	Zn	Co	Fe	Mn	Ca
Se	34	3.843	1.00					
Zn	30	2.478	0.64	1.00				
Co	27	1.715	0.45	0.69	1.00			
Fe	26	1.500	0.39	0.55	0.87	1.00		
Mn	25	1.303	0.34	0.53	0.76	0.87	1.00	
Ca	20	0.565	0.15	0.23	0.33	0.38	0.43	1.00

Measured

	Position	f''	Normalized f''	Old metal	New metal
3DCP					
Chain A	α	9.59	1.00	Fe	Co
	β	6.83	0.71	Fe	Mn
	γ	6.81	0.71	Zn	Mn
Chain B	α	8.42	1.00	Fe	Co
	β	7.50	0.89	Fe	Fe
	γ	7.6	0.90	Zn	Fe
Chain C	α	9.11	1.00	Fe	Co
	β	7.70	0.85	Fe	Fe
	γ	6.28	0.69	Zn	Mn

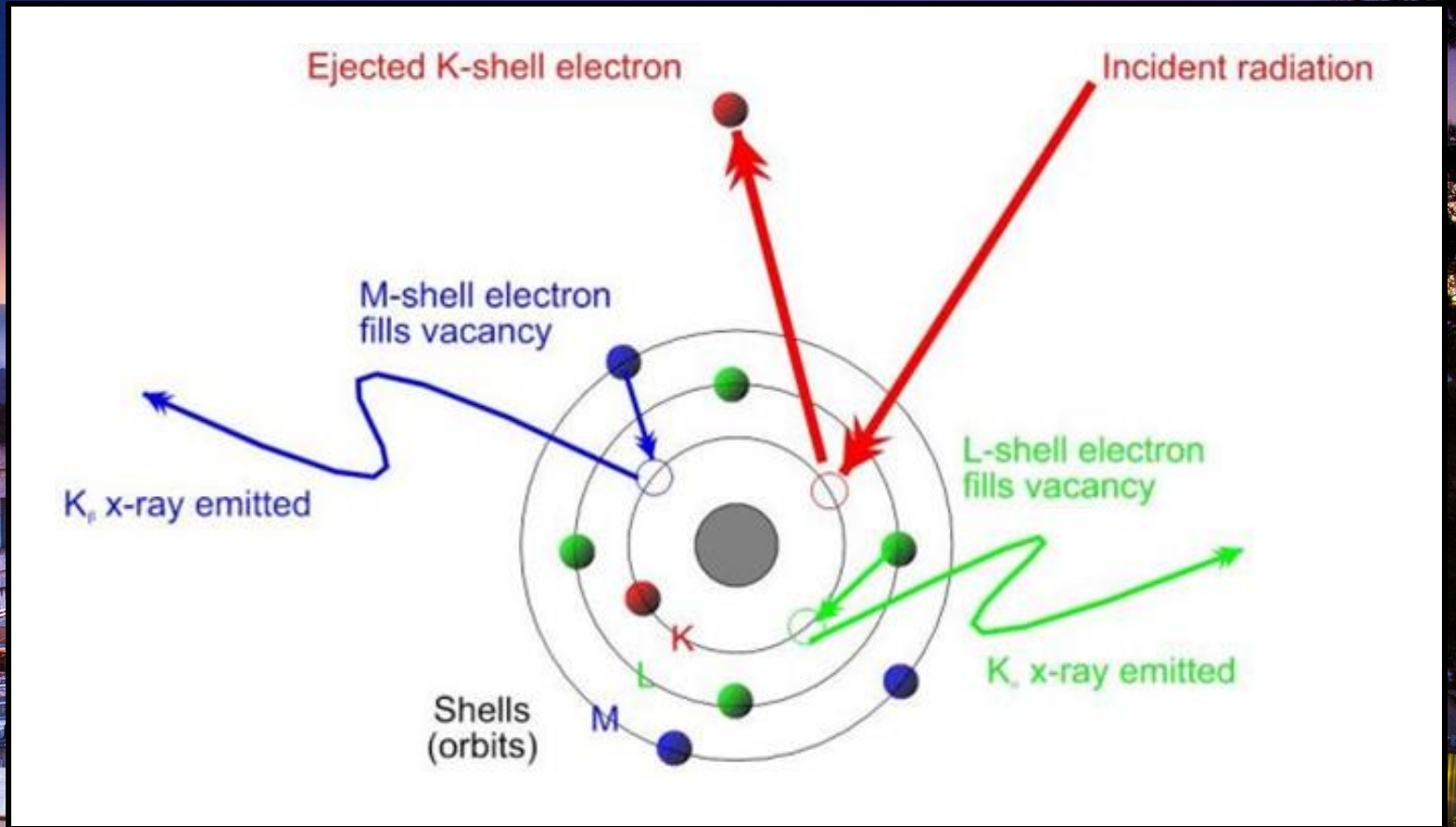
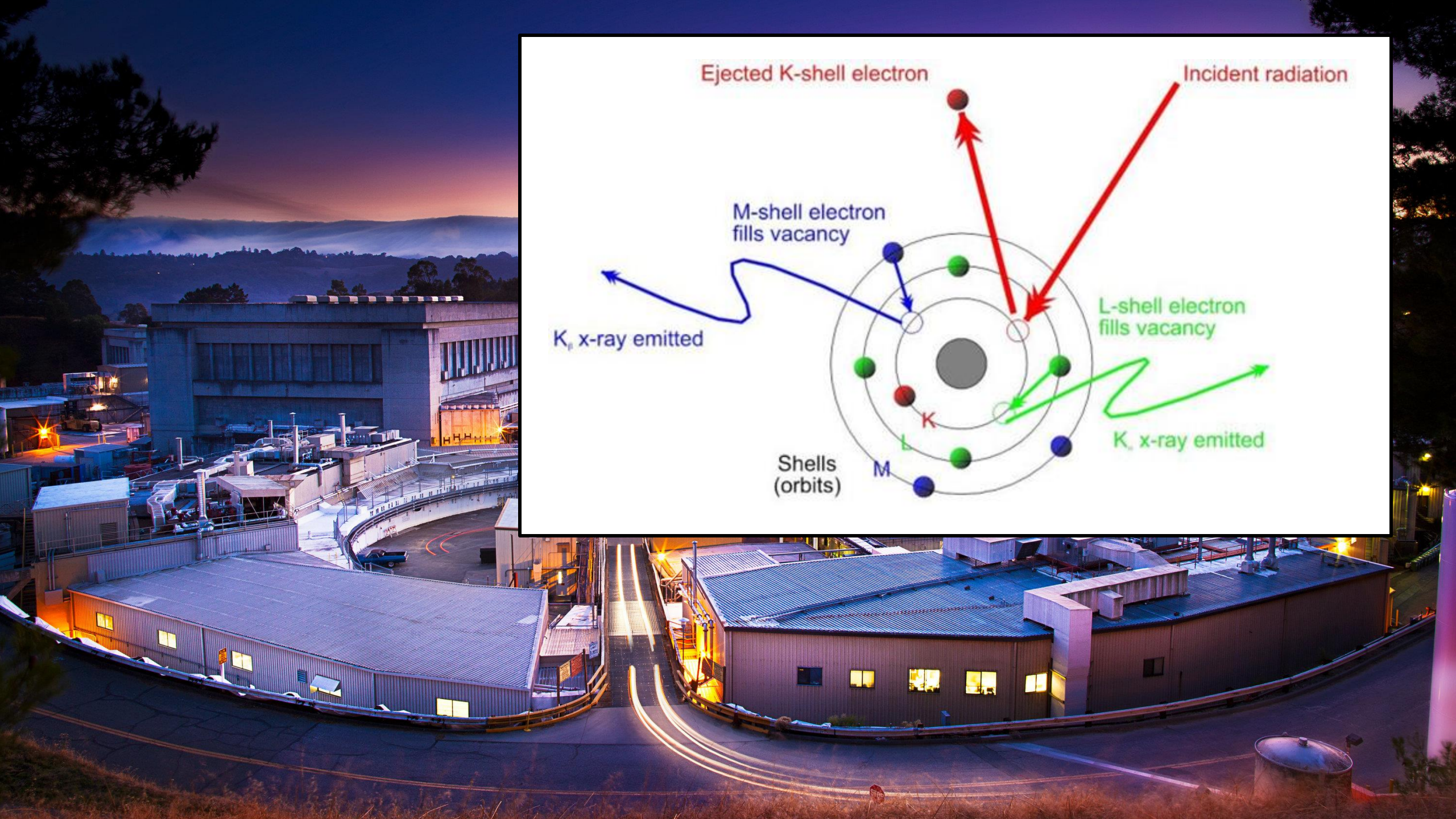
The correct metals can be assigned using the identification and stoichiometry from PIXE with the difference electron density from X-ray crystallography to position them – revealed a ligand important for mechanism.



Correctly modeled metal

There are currently ~3 facilities worldwide that could perform PIXE experiments on proteins
(and only one that does)

There are ~160 synchrotron macromolecular crystallography beamlines worldwide that could do X-ray fluorescence methods



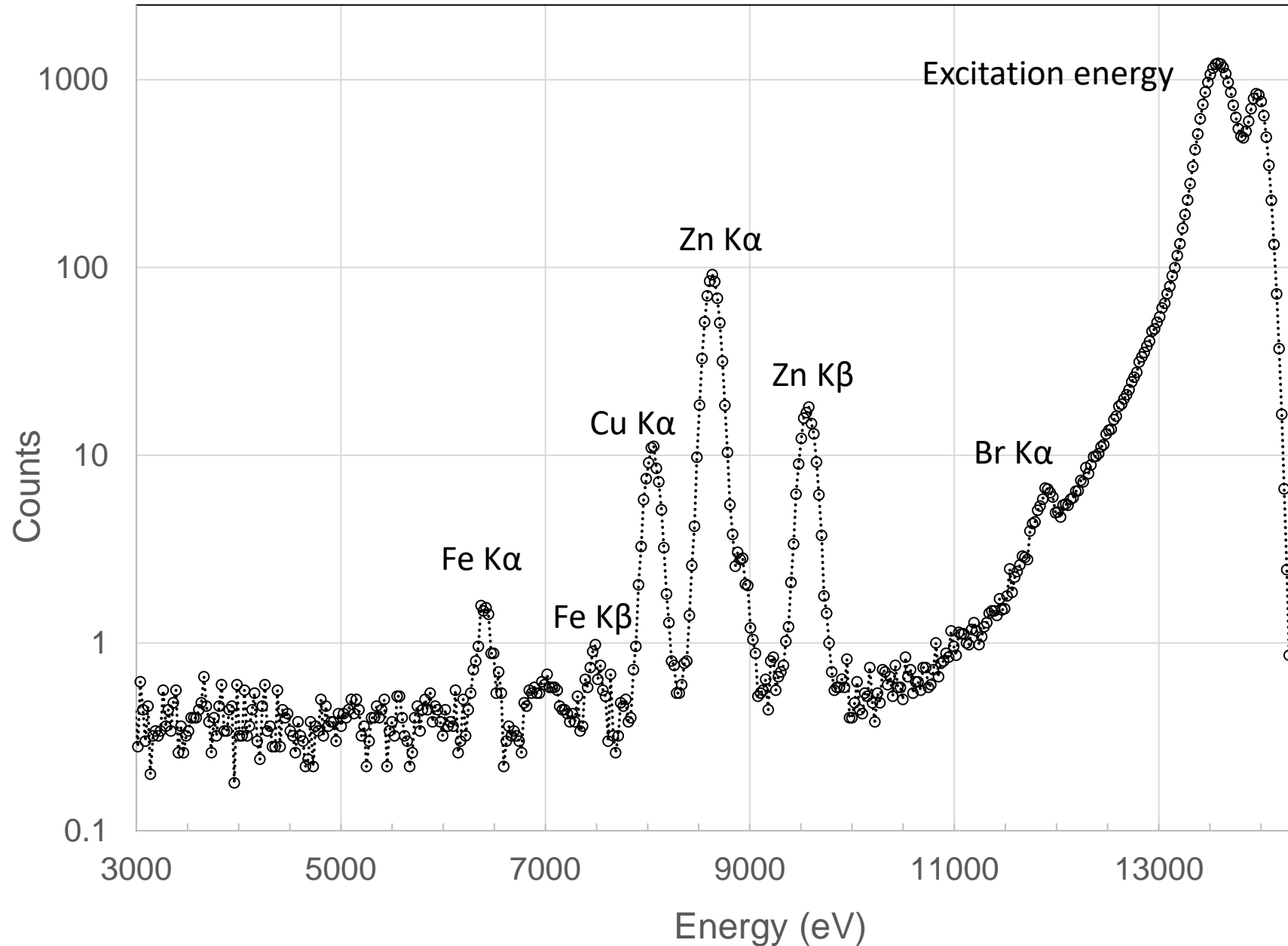
Beamline 12-2 SSRL



X-ray
fluorescence
studies

With Aina Cohen
and Sarah
Bowman





Fluorescence scan of a NIST trace standard containing

- 500 $\mu\text{g/ml}$ Zn
- 100 $\mu\text{g/ml}$ Cu
- 30 $\mu\text{g/ml}$ Fe
- 5 $\mu\text{g/ml}$ Mn

Note Ni $K\alpha$ and Br $K\alpha$ signals at 7478 eV and 11924 eV.

The Cu $K\beta$ signal can be seen in the higher energy tail of the Zn $K\alpha$ signal.

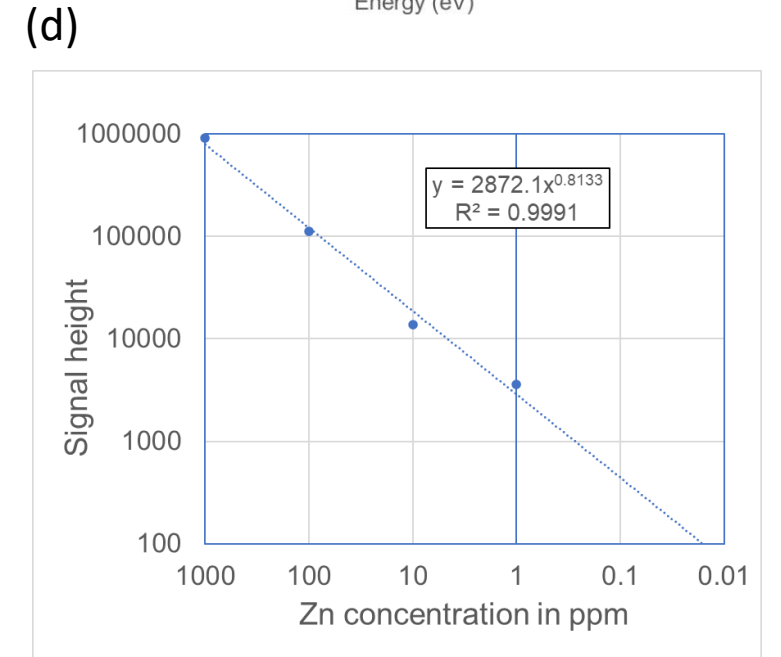
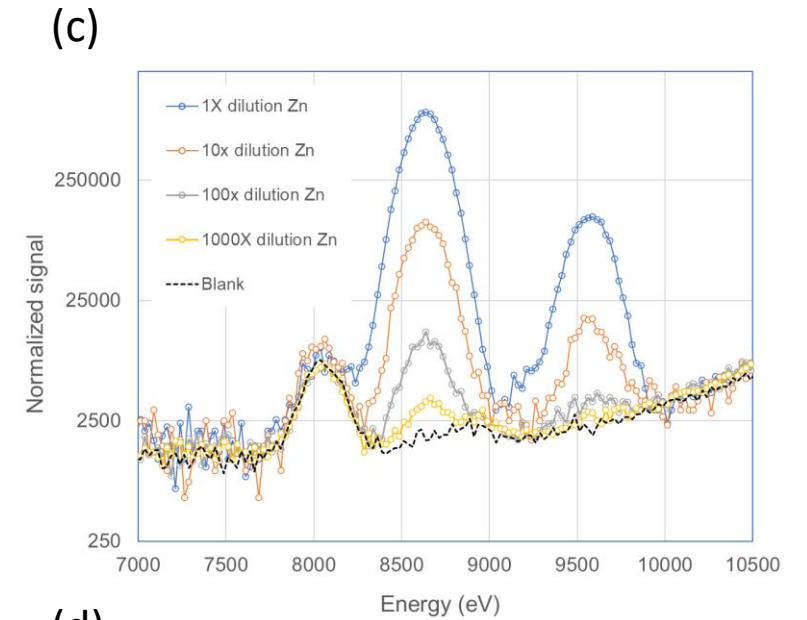
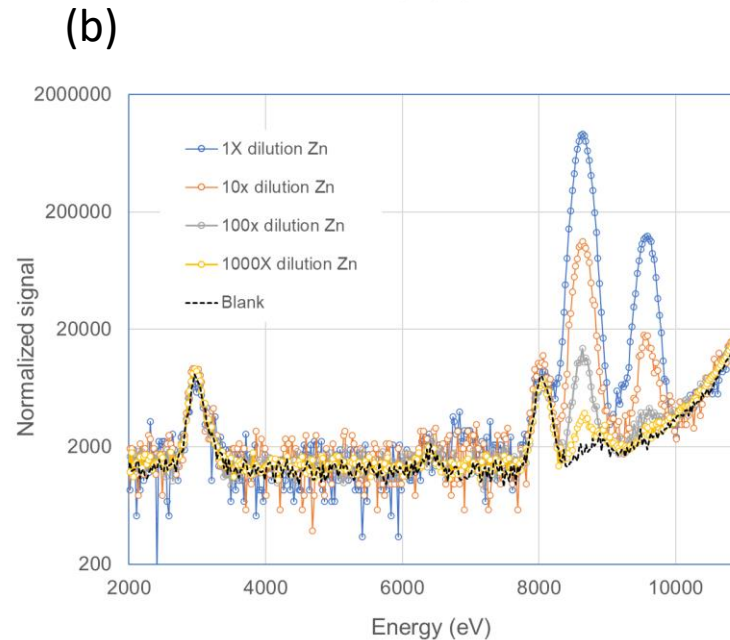
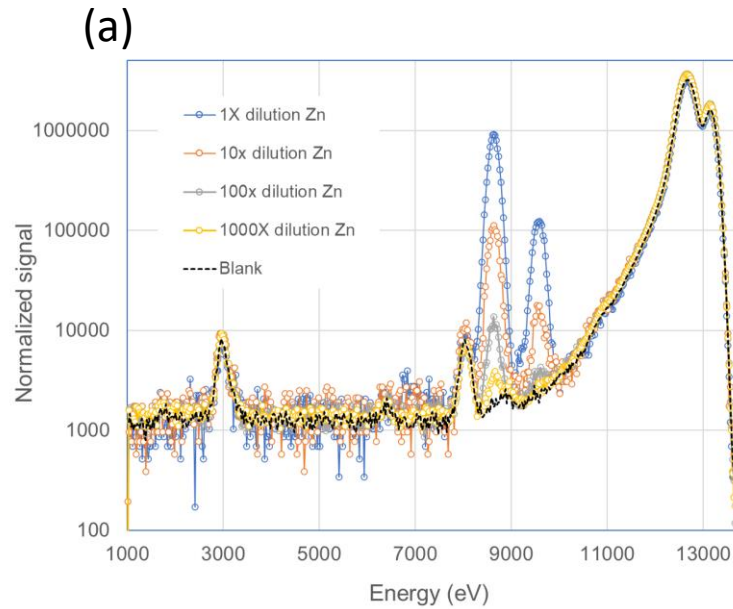
The step size of the measurements is 25 eV

Potential of technique

Zinc

Spectrums are shown with a Zn standard. Three plots are shown, (a) the full spectrum, (b) one enlarged to see a Rh peak from a mirror, and (c) another enlarged to show the Zn data.

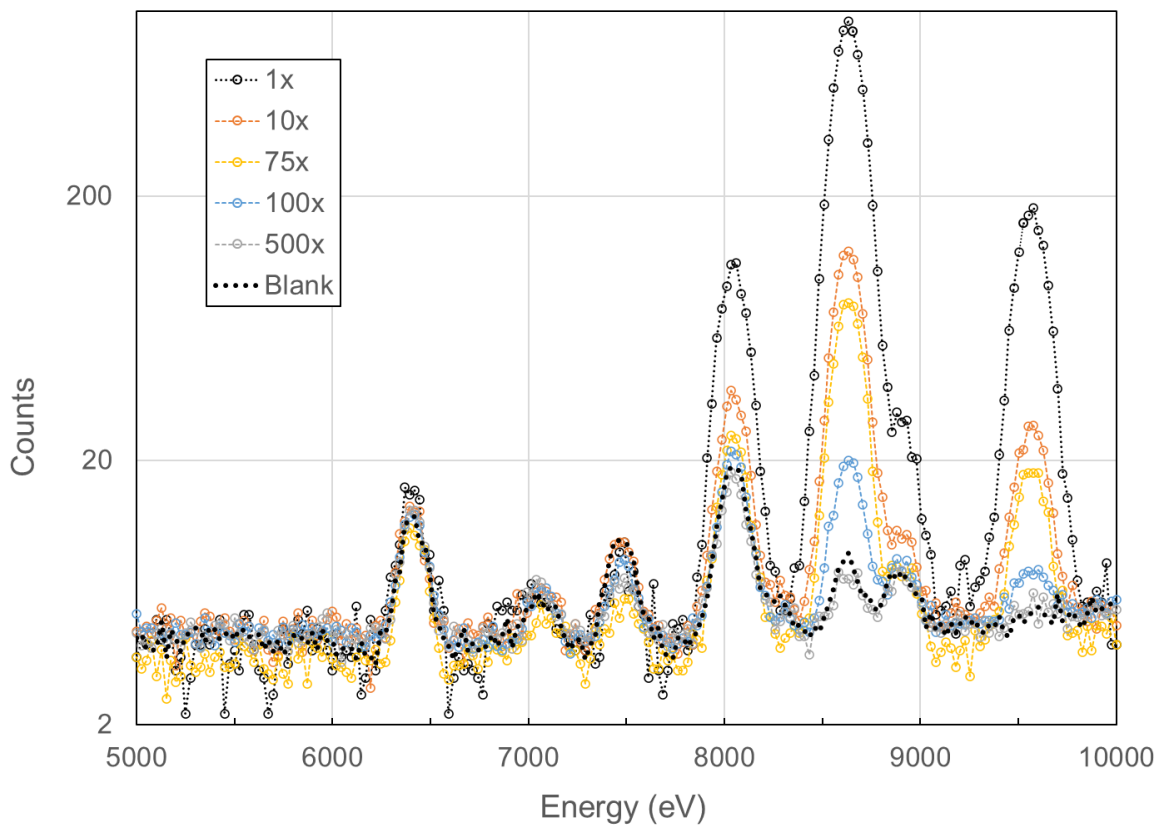
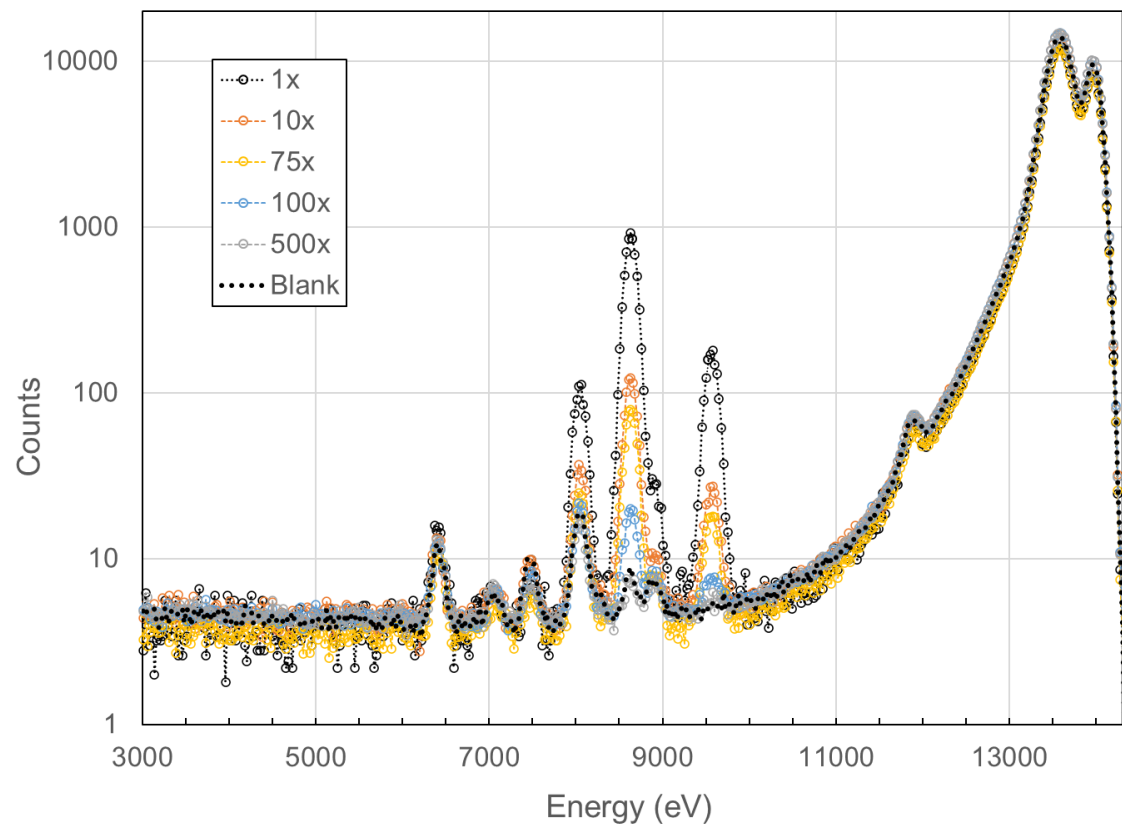
A fit to the peak height of the Zn K α (d) demonstrated **sensitivity down to 1 ppm.**



Mixture standard

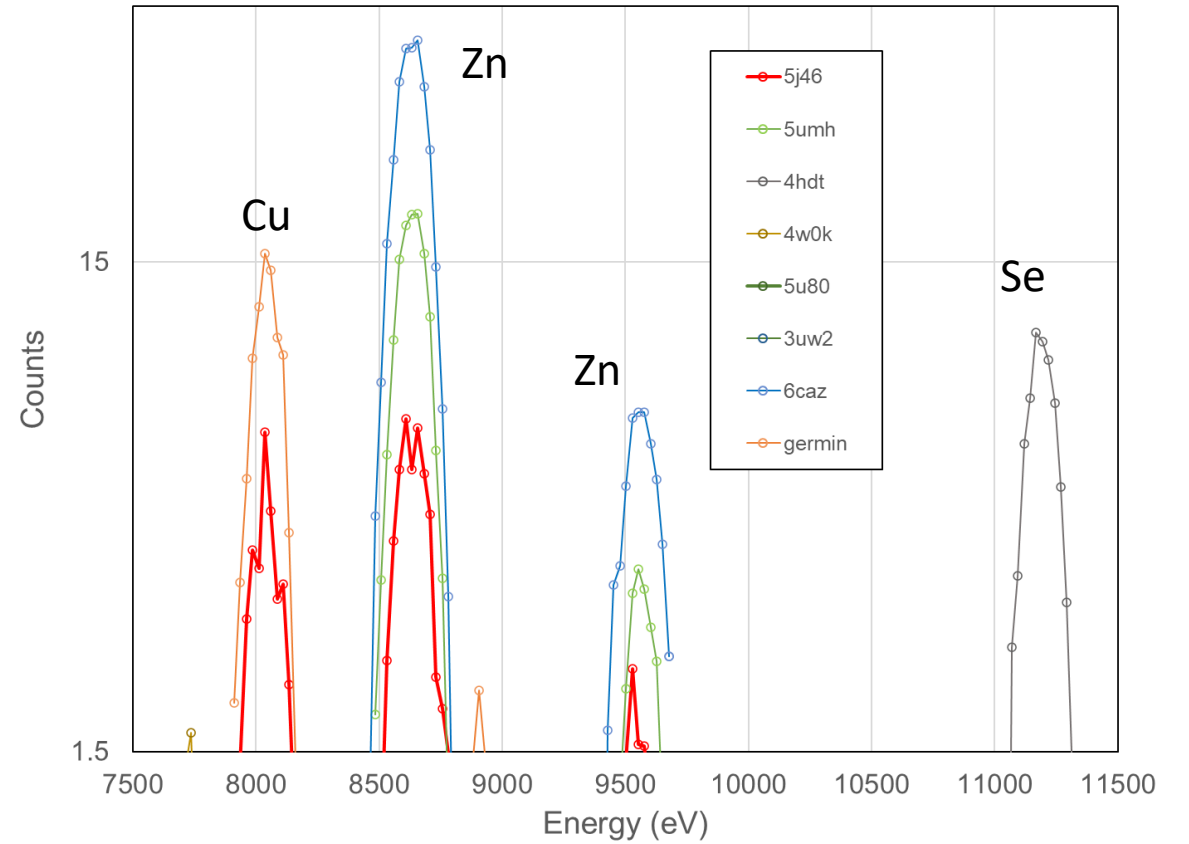
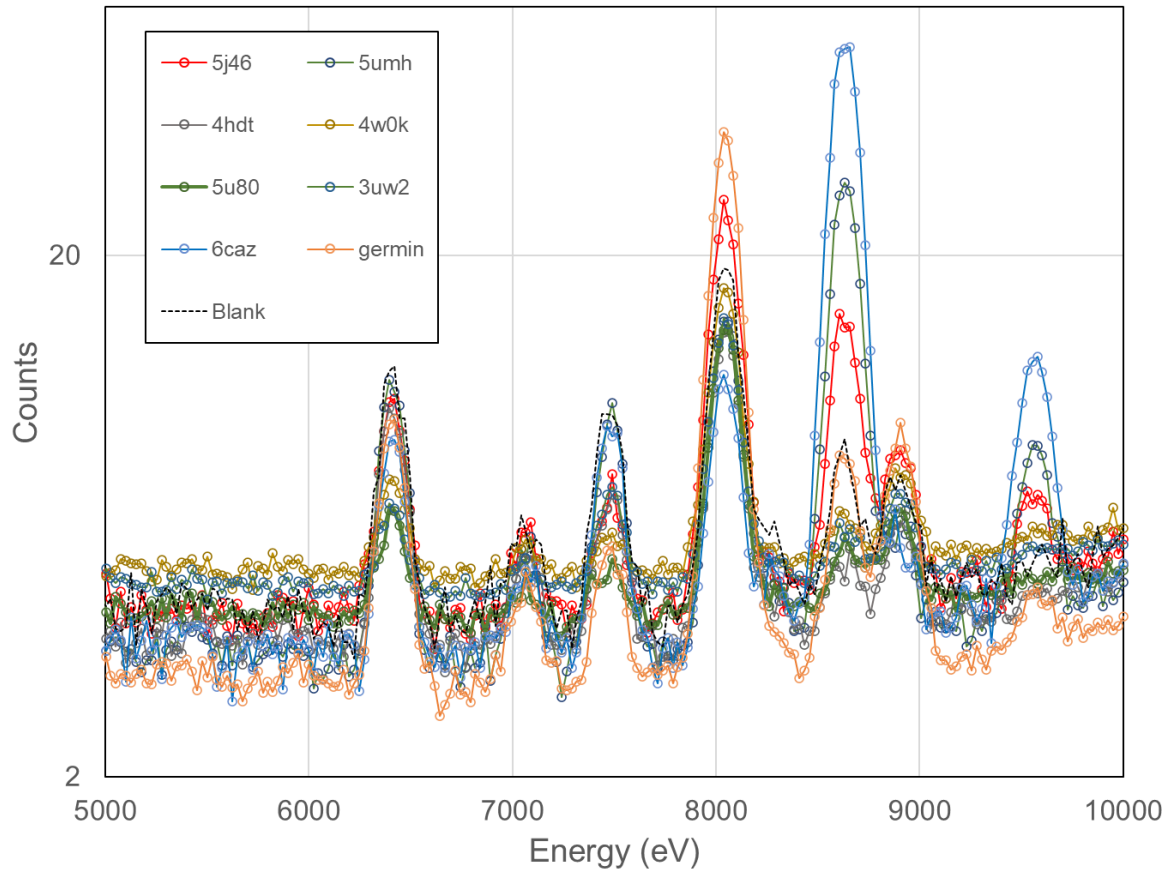
NIST standard		
	$\mu\text{g/ml}$	M
Zn	500	0.00765
Cu	100	0.00157
Fe	30	0.00054
Mn	5	0.00009

Sensitive to different metals



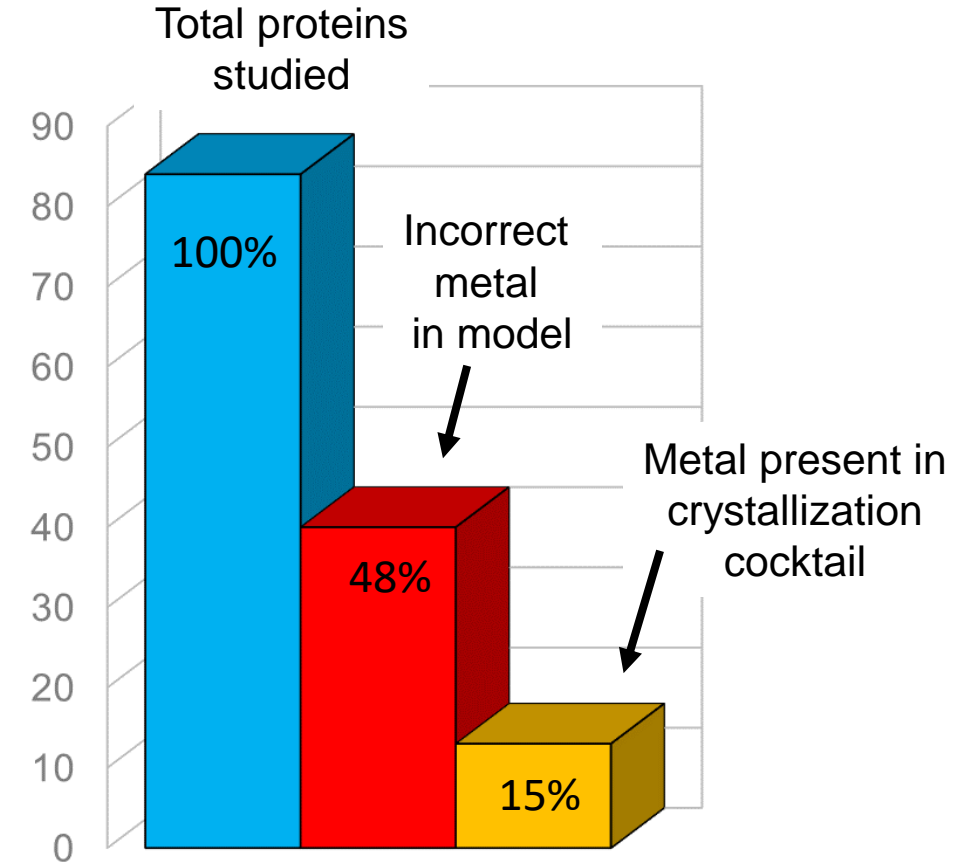
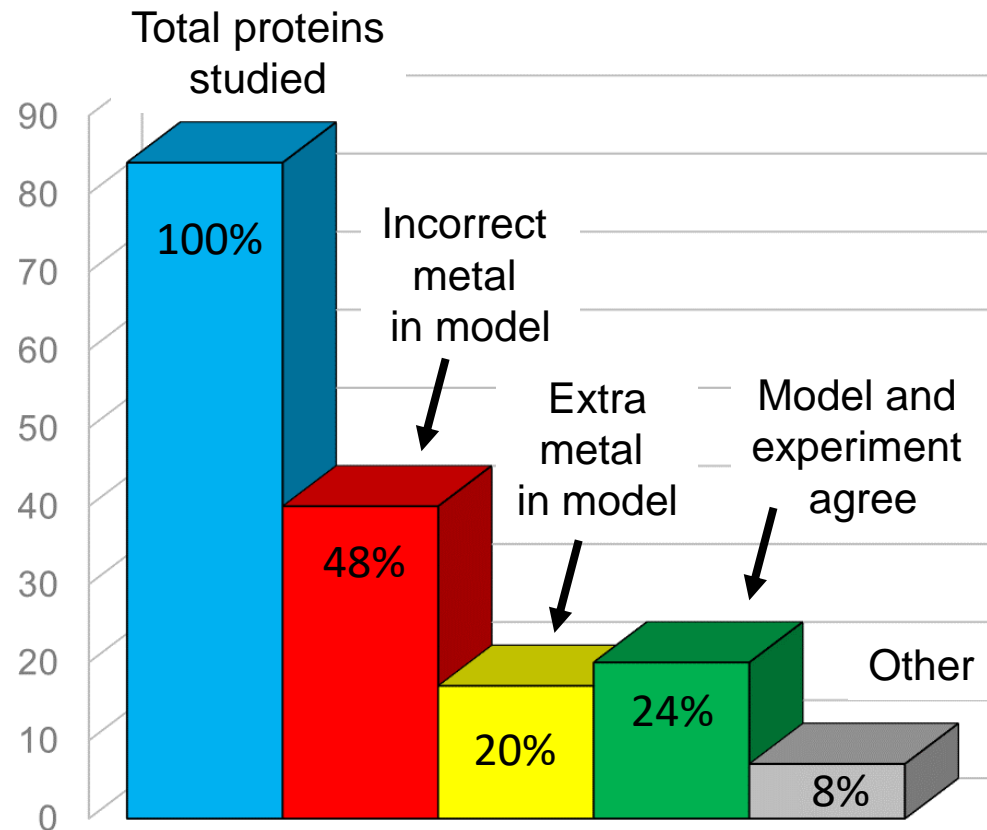
Protein samples

The protein samples measured are shown with a (a) full spectrum, (b) enlarged spectrum of interest.



Numbers from Experimental Data

PIXE and EDX study the protein sample, crystallographic structural models by necessity incorporate the crystallization conditions



48% of the models do not contain the experimentally detected metal. 15% of these can be explained by promiscuous metals in the crystallization conditions.

33% of metals are unexplained. This is in complete agreement with the 3-sigma cutoff of the Alchemy analysis of the 158,791 metal sites studied.

Take home message

- A computational study of over 150,000 metal sites and experimental data in the Protein Data Bank indicates that **one third of metal identities are suspect**.
- An experimental atomic-based study on 90 proteins is in complete agreement with this.
- There are ~4 million downloads of structures per day, 99% of those without experimental data.
- Over 200 data resources make use of this data.
- There are significant errors - **Caveat emptor**

Final thoughts

- All that glitters is not necessarily gold
- Always do an excitation scan after crystallographic data collection – all it costs is time (and not much).



Acknowledgements



Sarah Bowman (Hauptman-Woodward), Aina Cohen (Stanford Synchrotron Radiation Lightsource),
Catia Costa (University of Surrey – not pictured) Elspeth Garman (University of Oxford),
Geoff Grime (University of Surrey), Guy Montelione (Rensselaer Polytechnic Institute),
Elizabeth Snell (Hauptman-Woodward)

PDB
REDO



CCP4

α fill

Also, James Holton, Robbie Joosten, and the Seattle
Structural Genomics Center for Infectious Disease

Thank you and questions?



esnell@hwi.buffalo.edu



Services for academics, industry,
government and not-for-profits