Fig. S1 Suppl. Mat.

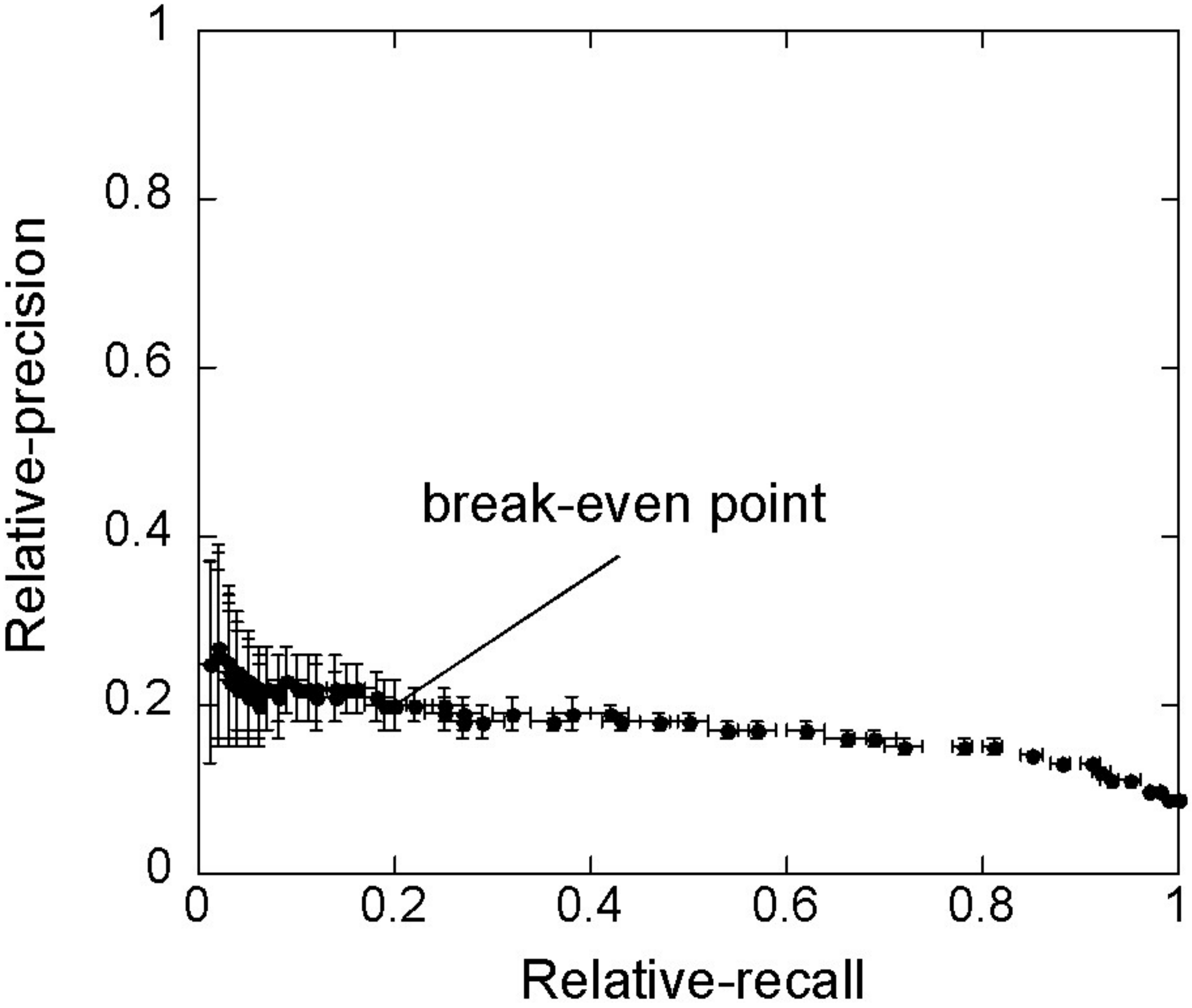

Fig. S1. Random baseline calculated using a counter predictor. In principle, the higher the number of His and Cys in a protein the higher the chance for MetalDetector to predict a metal binding residue. If this were true also for proteins, i.e. if proteins with more His and Cys were more likely to bind metals, we could think to predict metal binding sites by simply counting the overall number of His and Cys in a protein. Since, in our dataset, we do see a correlation between the overall number of His and Cys residues in a protein and its identification as a metalloprotein by HT-XAS, it is interesting to compare the performance of a simple His and Cys counter predictor to that one of MetalDetector. At 10% HT-XAS-recall, the simple counter predictor achieves 22% HT-XAS-precision, to be compared to 42-60% for MetalDetector at the same recall. HT-XAS-precision of the counter predictor at break-even point is instead 20% (32-45% for MetalDetector).

Fig. S2 Suppl. Mat. HT-XAS setup at beamline X3B-NSLS.

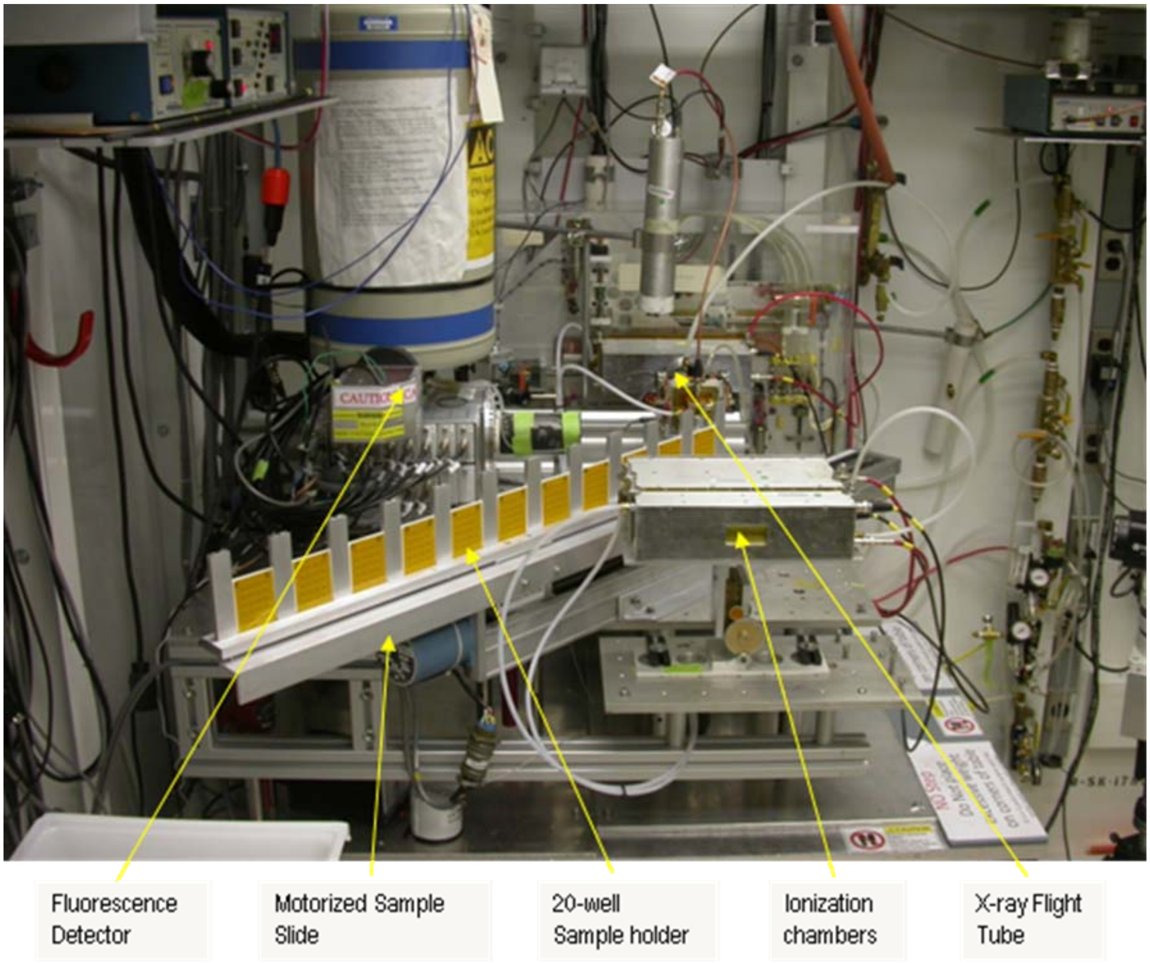

Table S1 Suppl. Mat. Zn Metalloproteins identified from NYSGXRC PSI-2 proteins and annotation of closely related genes.

ID	Metal	NMA	Length	Protein Annotation	Clusters of orthologs	BLAST-PDB	Related PDB
10073b	Zn	0.5	339	AC: Q45450, REP_1; OS: *Bacillus subtilis*	Ev=e-54, GDP-3,6-dideoxy-L-galactose biosynthesis protein	--	--
9465e	Zn	0.5	312	AC: gi|27383234, NUDIX family; OS:*Bradyrhizobium japonicum*	Ev=4e-60, NADH pyrophosphohydrolases containing a Zn-finger	*NADH pyrophosphatase from *Escherichia coli*	PDB Id: 2GB5; 23% identity; Metal ion=Zn
11319p	Zn	0.9	326	AC: YP_354808.1, TRAP-T family transporter; OS:*Rhodobacter sphaeroides*	Ev=3e-160, KDPG and KHG aldolase	putative periplasmic glutamate/glutamine-binding protein from *Thermus thermophilus*	PDB Id: 1US5; 20% identity; Metal ion=No
9482b	Zn	0.5	339	AC: gi|109898866, alcohol dehydrogenase GroES-like protein; OS: *Pseudoalteromonas atlantica*	Ev=1e-50, zinc-containing alcohol dehydrogenase	Ev=4e-41, threonine dehydrogenase from *Pyrococcus horikoshii*	PDB Id: 2DFV; Identities=33%; Metal ion=Zn
9453d	Zn	0.5	337	AC:BAB88741, 4-oxalomesaconate hydratase; OS: *Sphingomonas paucimobilis*	Ev=0.0, metallo-dependent hydrolase	Ev=2e-120, 4-oxalomesaconate hydratase from *Rhodopseudomonas palustris*	PDB ID: 2GWG; Identities=60%; Metal ion=Zn
9550a	Zn	0.8	254	AC:gi|15025762, PHP family hydrolase; OS:*Clostridium acetobutylicum*	Ev=e-70, histidinol phosphate phosphatase HisJ family	Ev=4e-9, Histidinol Phosphate Phosphatase from *Thermus thermophilus*	PDB ID:2YXO; Identities=21%, Metal ions=Zn, Fe
12087a	Zn,Ni	0.9,0.5	285	Ac:AAM99938.1, hypothetical protein; OS: *Streptococcus agalactiae*	Ev=2e-131, NAD-dependent protein deacetylases, SIR2 family	Transcriptional regulatory protein, SIR2 family, from Archaeoglobus fulgidus	PDB ID: 1ICI; 16% identity; Metal ion=Zn
9242a	Zn	0.5	441	AC:gi|29375795, amidohydrolas; OS: *Enterococcus faecalis*	E=0.0, adenine deaminase	A hypothetical protein TM0936 from *Thermotoga maritima*	PDB ID: 1P1M; 26% identity; Metal ion=Ni
9321a	Zn	1	261	AC: gi|15896575,	Ev=e-153, metallo-	Isoaspartyl	PDB ID: 2AQO;

				amidohydrolase; OS: *Clostridium acetobutylicum*	dependent hydrolase superfamily	dipeptidase from *Escherichia coli*	15% identity; Metal ion=Zn
9265h	Zn	0.8	437	AC:gi|38111844, hypothetical protein; OS: *Magnaporthe grisea*	Ev=0.0, mandelate racemase/muconate lactonizing enzyme	Ev=0.0, L-rhamnonate dehydratase from *Gibberella zeae*	PDB ID: 3FXG; identities=77%; Metal ion=Mg
11099b	Zn	0.6	521	AC: CAH08783.1, putative exported hexosaminidase; OS: *Bacteroides fragilis*	Ev=0.0, beta-hexosaminidase	Ev=3e-64, beta-hexosaminidase from *Paenibacillus Sp.*	PDB ID: 3GH4; Identities=31%; Metal ion=no
11092n	Zn	0.6	583	AC:AAO78028.1, beta-galactosidase; OS: *Bacteroides thetaiotaomicron*	Ev=0.0, beta-galactosidase	Ev=2e-19, putative beta-galactosidase from *Bacteroides fragilis*	PDB ID: 3CMG; Identities=23%, Metal ion=no
10519c	Zn	0.7	159	AC:Q4Z915, hypothetical protein; OS: *Staphylococcus phage Twort*	No close related protein.	--	--
9265i	Zn	1	438	AC:gi|40743575, hypothetical protein; OS: *Aspergillus nidulans*	Ev=0.0, mandelate racemase/muconate lactonizing enzyme	Ev=0.0, L-rhamnonate dehydratase from *Gibberella zeae*	PDB ID: 3FXG; identities=78%; Metal ion=Mg
10178e	Zn	0.5	138	AC:O26737, hypothetical protein; OS: *Methanothermobacter thermautotrophicus*	Ev=e-24, excinuclease ABC subunit C	--	--
9328a	Zn	1	249	AC:gi|11499354, hypothetical protein AF1765; OS: *Archaeoglobus fulgidus*	Ev=e-136, metal-dependent hydrolase, TatD-related deoxyribonuclease	Ev=e-136, Tatd-like protein (Af1765) from *Archaeoglobus fulgidus*	PDB ID:3GUW, Identities=97%,; Metal ion=Zn
9252d	Zn,Mn	0.8,0.5	408	AC:gi|13474288; guanine deaminase; OS: *Mesorhizobium loti*	Ev=0.0, metal-dependent hydrolase, guanine deaminase	Ev=2e-79, Guanine Deaminase from *Bradyrhizobium japonicum*	PDB ID: 2OOD, Identities=38%; Metal ion=Zn
13851a	Zn	0.5	239	AC:AAV79486.1, transcriptional	Ev=e-137, transcriptional	Ev=4e-29, phosphate transport	PDB ID: 1T8B; Identities=32%;

				regulator PhoU; OS: *Salmonella typhimurium*	regulator PhoU	system regulatory protein PhoU from *Streptococcus pneumoniae*	Metal ion=Zn
10453f	Zn	1	796	AC:Q8Y675, PriA protein; OS: *Listeria monocytogenes*	Ev=0.0, primosome assembly protein PriA	Ev=3e-05, a DNA helicase domain	PDB ID: 1D9X; Identities=31% (108 residues); Metal ion=Zn
9256a	Zn	1.6	418	AC:gi|15805850,putative hydrolase; OS: *Deinococcus radiodurans*	Ev=0.0, metal –dependent hydrolase, amidohydrolase	Ev=0.0; amidohydrolase Dr_0824 from *Deinococcus radiodurans*	PDB ID: 2IMR, Identities=100%; Metal ion=Zn
10068b	Zn	0.8	163	AC:Q9X179, Hypothetical protein; OS: *Thermotoga maritima*	Ev=2e-10, putative 3'-5' exonuclease	Ev=0.012, Rad50 zinc-hook in DNA recombination and repair.	PDB ID:1L8D, Identities=31% (92 residues), Metal ion=no
10114c	Zn	1	238	AC:Q9JZR1, Cytidine and deoxycytidylate deaminase family protein; OS: *Neisseria meningitidis*	Ev=2e-55, zinc-binding CMP/dCMP deaminase	Ev=2e-31, tRNA adenosine deaminase TadA from *Escherichia coli*	PDB ID: 1Z3A, Identities=44% (146 residues), Metal ion=Zn
9247a	Zn	0.8	426	AC:gi|10173106, putative amidohydrolase; OS: *Bacillus halodurans*	Ev=0.0, amidohydrolase	EV=0.0, putative amidohydrolase BH0493 from *Bacillus halodurans*	PDB ID:2QEE, Identities=100%, Metal ion=Zn, Mg.
10203c	Zn	0.8	316	AC:O30168, Hypothetical protein; OS: *Archaeoglobus fulgidus*	Ev=2e-123, CRISPR-associated autoregulator DevR family	--	--
9431b	Zn	0.7	261	AC:gi|44324736, amidohydrolase; OS: *Bacillus halodurans*	Ev=5e-128, metallo-dependent gydrolase, TatD-related deoxyribonuclease	Ev=3e-46, YJJV, TATD homolog from *Escherichia coli* k12	PDB ID: 1ZZM, Identities=40%; Metal ion=Zn
9304c	Zn	0.8	358	AC:gi|66807941, amidohydrolase; OS: *Dictyostelium discoideum*	Ev=2e-128, metallo-dependent hydrolase, aminocarboxymuconate semialdehyde	Ev=5e-65, alpha-Amino-beta-Carboxymuconate-epsilon-semialdehyde-	PDB ID: 2HBV; Identities=38%; Metal ion=Zn, Mg

					decarboxylase	decarboxylase	
9231a	Zn	1	464	AC:gi|27378957, amidohydrolase; OS:*Bradyrhizobium japonicum*	Ev=0.0; metallo-dependent hydrolase, cytosine deaminase	Ev=5e-60, N-isopropylammelide isopropylaminohydrolase Atzc from *Pseudomonas sp.*	PDB ID: 2QT3; Identities=33%; Metal ion=Zn
10418c	Zn	1	182	AC:Q9KCA3, hypothetical protein BH1670; OS: *Bacillus halodurans*	Ev=e-50, Prephenate dehydrogenase	Ev=5e-51, protein Ba1542 from *Bacillus anthracis*	PDB ID:3DO9; Identities=55%; Metal ion=no
9218a	Zn,Mn	0.5,0.3	528	AC:gi|15615762, amidohydrolase; OS: *Bacillus halodurans*	Ev=7e-151, Metallo-dependent hydrolase	Ev=3e-45, an uncharacterized metal-dependent hydrolase from *Pyrococcus furiosus*	PDB ID: 3ETK; Identities=30%; Metal ion=Zn
10064f	Zn	0.6	356	AC:Q8KDK9, CBS domain protein; OS: *Chlorobium tepidum*	Ev=e-163, putative CBS domain and cyclic nucleotide-regulated nucleotidyltransferase	Glutamine Synthetase adenylyltransferase from *Escherichia coli*	PDB ID: 1V4A; 13% identity; Metal ion=No
10072b	Zn,Ni	0.7,0.8	467	AC:P37875,SpoVR; OS: *Bacillus subtilis*	Ev=0.0, stage V sporulation protein R, SpoVR	--	--
9236e	Zn	0.6	478	AC:gi|44264246, amidohydrolase; OS:unknown	Ev=0.0, hydroxydechloroatrazine ethylaminohydrolase	Ev=0.0, amidohydrolase from an evironmental sample of Sargasso sea	PDB ID: 3H4U; Identities=100%; Metal ion=Zn
10382a	Zn	0.6	747	AC:P50830, putative ATP-dependent helicase; OS: *Bacillus subtilis*	Ev=0.0; helicase family protein with metal-binding cysteine cluster	Ev=5e-22, Archaeal dna helicase	PDb ID: 2ZJ2; Identities=25% (436 residues); Metal ion=no
10409h	Zn	0.5	362	AC:Q3EDM6, hypothetical protein; OS: *Actinobacillus succinogenes*	Ev=6e-156, N-acetylglucosaminyl transferase;	GlcNAc transferase from *homo sapien*	PDB ID: 1W3B; 12% identity; Metal ion=Ca
11019r	Zn	0.8	290	AC:Q82G55, Periplasma Binding protein type 1	Ev=6e-166, LacI family transcriptional	Ev=e-27, transcriptional regulator-	PDB:1RZR; Identities=30%; Metal ion=Mg

				superfamily; OS: *Streptomyces avermitilis*	regulator	phosphoprotein-dna complex	
10333e	Zn	0.5	913	AC:Q1G9P4, DNA polymerase III, alpha subunit (Gram-positive type) (PolC); OS: *Lactobacillus delbrueckii*	Ev=0.0, DNA-directed DNA polymerase	Ev=0.0, Dna polymerase Polc from *Geobacillus kaustophilus*	PDB ID: 3F2B; Identities=56%; Metal ion=Zn, Mg
11008p	Zn	0.6	329	AC:Q2AEH6, Periplasma Binding protein type 1 superfamily; OS: *Halothermothrix orenii*	Ev=9e-49, LacI family transcriptional regulator	Ev=e-46, transcriptional regulator-phosphoprotein-dna Complex from *Bacillus megaterium*	PDB:1RZR; Identities=31%; Metal ion=Mg
12026a	Zn	0.5	782	AC:NP_809262.1, hypothetical protein BT_034, DUF1680 family; OS: *Bacteroides thetaiotaomicron*	Ev=0.0, Acetyl-CoA carboxylase, biotin carboxylase	--	--
12018c	ZN	0.5	264	AC:NP_810652.1, excinuclease ABC subunit A; OS: *Bacteroides thetaiotaomicron*	Ev=2e-130, UvrABC SOS-repair system proteins	Ev=4e-70, Uvra2 from *Deinococcus radiodurans*	PDB ID: 2VF7; Identities=47%; Metal ion=Zn, Mg
9276d	Zn	0.5	392	AC:gi|15896575, amidohydrolase; OS: *Clostridium acetobutylicum*	Ev=0.0, metallo-dependent hydrolase	Ev=3e-6, uncharacterized protein Eah89906	PDB ID:3FEQ; Identities=22%; Metal ion=Zn

NMA, number of metal atoms per protein molecule; AC, accession number; OS, organism/species; Ev, E-value from Blast search.

*Some related PDBs with Ev not shown. The PDBs were found through building a PSSM (position-specific scoring matrix) by searching Genbank, then using the PSSM to search the PDB.

Supplementary Information

Extended X-ray Absorption Fine Structure (EXAFS) Data Collection and Analysis

EXAFS data were collected at the NSLS X3B beamline as K_{α} fluorescence spectra using a 13-element solid-state Ge detector array (Canberra), a nickel-coated mirror for harmonic rejection and a Si(111) double crystal sagittally focusing monochromator. A helium displex cryostat was used to maintain the sample temperature below 80 K to reduce the dynamic disorder in the samples. Energy calibration was performed by the simultaneous transmission measurement of the absorption spectrum of a Zn foil. The energy was scanned from 200 eV below (9459 eV) to 16 x k above (10635 eV) the Zn K-edge (defined as 9659 eV). From 200-20 eV below the edge, data were collected in 5 eV steps with a 1 s integration time. Around the edge region (from 20 eV below to 30 eV above), the step size was decreased to 0.2 eV and the integration increased to 3 s. From 30 eV above the edge until the end of the measurement, the step size was set to change as a function of k (0.05 x k) and the integration time was increased to 5 s. Data were taken over the course of a day (26 scans for each sample); comparison of data between initial and final scans showed no evidence of radiation damage. Careful examination of data from each channel was made to confirm the absence of artifacts before averaging; 8-11 channels were averaged for each scan. Data were processed and a first shell analysis performed using the IFEFFIT software (Newville 2001; Ravel and Newville 2005).

Table S2 Suppl. Mat. EXAFS fit results.

	9550a		9453d
ΔE_0	5.91 ± 2.54	ΔE_0	5.76 ± 4.16
R-factor	0.046	R-factor	0.006
R (4 neighbors)	2.01 ± 0.03	R (4 Amino Acids)	2.03 ± 0.04
		R (2 H_2O)	2.50 ± 0.03
σ^2 (4 neighbors)	0.006 ± 0.002	σ^2 (4 Amino Acids)	0.006 ± 0.002
		σ^2 (2 H_2O)	0.004 ± 0.006

ΔE_0: relative energy shifts used in the fits (eV); R-factor: goodness of fit parameter based on the misfit relative to the data size (Newville 2001); R: metal-ligand bond length (Å); σ^2: Debye-Waller disorder factor (Å^2).

For target 9550a, a simple simulation based on a single average scattering distance for four neighboring atoms was tested against the 1M65 high-affinity Zn binding site model. Data were fit using multiple k-weights (1, 2 and 3) over a k-range of 2-11 Å^{-1} and an R range of 1.3-2.7 Å (Hanning windowing). Fitting results (Table 2 Suppl. Mat.) indicate that the experimental data are consistent with a tetrahedral arrangement of light element neighbors (Oxygen and Nitrogen) with an average distance in line with expected Zn-binding distances for these elements (~2 Å). The 1M65 high-affinity Zn-binding active site can be considered a reasonable initial model for the 9550a Zn-binding site.

For target 9453d, initial attempts at fitting the experimental data against the 2GWG Zn-binding site were unsuccessful, thus a modified model was constructed to decrease unreasonably long Zn-light element distances while maintaining general octahedral symmetry. A split-shell simulation based on a Zn-N/O (His, Glu) scattering distance with a degeneracy of four and a Zn-O (H_2O) scattering distance with a degeneracy of two was tested against the modified model. Data were fit using multiple k-weights (1, 2 and 3) over a k-range of 2-11 $Å^{-1}$ and an R range of 1.3-2.4 Å (Hanning windowing). Fitting results (Table 2 Suppl. Mat.) indicate that to a first shell approximation, this model is consistent with the experimental data and can be considered a reasonable model for the 9453d Zn-binding site.

Protein quality control by mass spectrometry MALDI-MS were performed on a Voyager, DE-RP reflecting time-of-flight mass spectrometer (PE Biosystems, Framingham MA) equipped with a 337 nm nitrogen laser operating in the linear, positive ion mode with an accelerating voltage of +25 kV and an extraction delay of 500 ns. A timed ion selector was used to deflect ions of low *m/z* (<5000) from the detector. Spectra were acquired by averaging data of approximately 150–200 laser shots to improve data quality and ion statistics. Mass spectra were calibrated using singly and doubly charged peaks of carbonic anhydrase (CA, 29 kDa) and bovine serum albumin (BSA, 66.6 kDa) for the proteins with mass less than 40 kDa and greater than 40 kDa, respectively. A modified "thin layer" method was used for MALDI-MS sample preparations.

HPLC-ESI-MS was used for accurate mass measurements of proteins (<100 ppm). The HPLC instrument (Agilent 1100) used in the study consisted of a degasser, binary pumping system, an auto sampler, a C-3 column (250 x 4.6 mm) and a variable wavelength UV-vis detector. 15 µg of each protein sample was injected using the auto sampler. Chromatography was carried out at an ambient temperature by flowing solvent A (5% acetonitrile, 0.1% formic acid) for 2.5 min to bind protein sample and remove buffer and salts, and then forming a linear gradient between solvent A-solvent B (95% acetonitrile, 0.1% TFA) for 4 min followed by re-equilibrating the column with solvent A for 0.5 min. The flow rate was 0.4mL/min and the UV detector was set at 280 nm. The HPLC was connected online with ESI-MS (API 150EX; Applied Biosystems, Foster City, CA). ESI-MS measurements were conducted at an ion spray voltage of 5500 V, source temperature of 300ºC, a focusing potential of 250 V and 1100 to 1900 m/z scan range. The molecular masses of proteins were estimated using BioAnalyst version 1.4.

The proteins having mass discrepancies greater than 100ppm using ESI-MS were further analyzed to confirm their identities using tandem mass spectrometry (MS/MS). For the LC-MS/MS analysis, the tryptic peptides were loaded onto a capillary C-18 LC column on-line with a Finnigan LCQ^{DECA} (ThermQuest, San Jose, CA, USA) ion-trap mass analyzer equipped with an ESI source. SEQUEST (Bioworks 2.0, ThermoFinnigan) was used to search the NCBI nonredundant protein database with the MS/MS data. Protein identification was performed for any potential reagent mix-ups, and sample information was corrected in the Laboratory Information Management System (LIMS), which served as a sample tracking database. The information on the targets can be accessed through PepcDB (http://pepcdb.sbkb.org/).