Electronic Supplementary Information (ESI)

Mining Peptides for Mining Solutions: Evaluation of Calcium-Binding Peptides for Rare Earth Element Separations

Farid F. Khoury[a], Bradley S. Heater[b], Daniel R. Marzolf[b], Sameera Abeyrathna[a], Jonathan W. Picking[b], Piyush Kumar[a], Steven A. Higgins[b], Randy Jones[b], Alan T. Lewis, Jr.[b], Katarzyna H. Kucharzyk[b,*], and Scott Banta[a,*]

[a] Department of Chemical Engineering, Columbia University, New York, NY, 10027

[b] Battelle Memorial Research Institute, Columbus, OH, 43201

Authors contributed equally

[*] Corresponding authors

E-mail: sbanta@columbia.edu

E-mail: kucharzyk@battelle.org

Table of Contents

Materials & Methods

Chemicals. Q5 High-Fidelity DNA Polymerase, NEBuilder HiFi DNA Assembly mix, restriction enzymes (*Kpn*I and *Xho*I), 5-alpha and BL21(DE3) competent *E. coli* cells were purchased from New England Biolabs. Terrific Broth and 0.5 M EDTA solutions were purchased from Research Products International. Isopropyl-β-D-thiogalactopyranoside (IPTG), ampicillin, and D-biotin were purchased from GoldBio. Pierce™ High-Capacity Ni-IMAC Resin and Invitrogen NuPAGE™ Bis-Tris SDS-PAGE gels and Novex™ protein standards were purchased from Thermo Scientific. Prepacked columns MBPTrap HP, StrepTrap XT, HiLoad 16/600 pg 75, and PD-10 desalting columns were purchased from Cytiva. Protein concentrators Amicon Ultra filters were purchased from Millipore Sigma. HaloLink resin was purchased from Promega. Empty chromatography columns were purchased from Diba Industries. Custom DNA oligos were purchased from Integrated DNA Technologies. All other chemicals were purchased from Sigma-Aldrich unless specified otherwise.

Bioinformatics. Proteins sequences of microbes affiliated with soil environment were accessed using ProGenomes (v3) database[1], and aligned with a profile of known calcium/lanthanide binding proteins to identify potential candidates. Profile hiden Markov models (HMMs) were constructed from seed sequences for lanthanide/calcium binding domains accessed through InterPro[2] using hmmbuild[3] (v3.3.2), and protein candidates were aligned against the profile HMM using hmmsearch (E-value cutoff 1E-5). The resultant alignment outputs were parsed with custom python scripts and the resulting files were filtered to remove sequences with detected domains smaller than one standard deviation of the mean domain length. The length-filtered domain sequences were then clustered at 90% amino acid identity using cd-hit[4] (v4.8.1, default settings, -c 0.9) and aligned using mafft[5] (v7.453) with default settings, and the five most representative

sequences in the alignment selected using trimAl[6] (v1.4.rev22). Overall, between three and five representative microbial protein sequences from the proGenomes database were selected for each category of calcium-binding domains for further analysis.

Plasmid construction & cloning. *E. coli* codon-optimized genes for the proteins of interest were purchased pre-assembled in a pET-28a(+) vector from Twist Biosciences. To generate the Förster resonance energy transfer (FRET) constructs with the cyan fluorescent protein (CFP) and the enhanced yellow fluorescent protein (YFP), DNA digestion enzymes *Kpn*I and *Xho*I were used to flank out the RTX domain from a previously described CFP-RTX-EYFP construct in pET-20b(+) vector[7]; The genes encoding proteins of interest were amplified by PCR with the appropriate overhangs, and assembled using NEBuilder® HiFi DNA Assembly. Maltose binding proteins with self-cleaving intein domain fusions (MBP-I-protein) were constructed by amplifying both the vector and the insert via PCR with the appropriate overhangs for ligation using NEBuilder® HiFi DNA Assembly. The MBP-I-protein vector was amplified from a previously described construct with the RTX domain (MBP-I-RTX)[8]. Assembly products were transformed into 5-alpha cells, which were subsequently confirmed by Sanger Sequencing. The required oligos were designed using NEBuilder and purchased from IDT.

Protein expression and purification. CFP-Protein-EYFP fusions were expressed in BL21(DE3) *E. coli* cells, as previously described[7]. Briefly, a sterilized Terrific Broth media flask (1L) supplemented with 100 mg L^{-1} ampicillin was inoculated with 5 mL of saturated culture. The culture was incubated in a biological shaker at 37ºC until the mid-log phase was reached (OD_{600nm} 0.4-0.6). The protein expression was induced using 0.04 mM IPTG and carried for 16-20 hours at 20ºC. The cells were harvested by centrifugation at 5,000 g for 15 minutes at 4ºC, resuspended in 40 ml of Ni-NTA washing buffer (20 mM Tris/HCl, 200 mM NaCl, 5 mM imidazole, pH 7.4),

lysed via sonication, and centrifuged at 10,000g for 1 hour at 4ºC. Clarified cell lysates containing the his-tagged fusion proteins were purified on Ni-NTA resin according to the manufacturer's protocols. The purified proteins were buffered exchanged into storage buffer (50 mM Tris, 50 mM KCl, pH 7.4) using PD-10 desalting columns.

RTX, HEW5, A0A7, and CaM(III, IV), for ITC and XO measurements, were expressed fused to a self-cleaving intein domain. Briefly, MBP-I-proteins were expressed in NEB 5-alpha *E.coli* cells. Cultures (1L) were inoculated with 5 mL of saturated cell cultures in sterilized Terrific Broth supplemented with 100 mg L^{-1} ampicillin and 2 g L^{-1} glucose. The cell culture was incubated in a biological shaker at 37ºC until the mid-log growth phase was reached (OD_{600nm} 0.4-0.6). The protein expression was induced with 0.3 mM IPTG, and the cell cultures were further incubated for 4-6 hours at 37ºC. The cells were harvested by centrifugation at 5,000 g for 15 minutes, resuspended in 100 mL of MBP binding buffer (20 mM Tris/HCl, 200 mM NaCl, 1 mM EDTA, pH 7.4), lysed via sonication, and clarified by centrifugation at 10,000 g for 1 hour. The clarified cell lysate was purified using two 5 mL MBPTrap HP connected in series (10 mL column) on an ÄKTA pure™ protein purification system. Briefly, the column was equilibrated with 10 CVs of MBP binding buffer, 50 mL of the clarified cell lysate was injected onto the column using a 50 mL superloop, the column was washed with MBP binding buffer for 25 CVs or until a stable UV_{280nm} below 3.0 mAU was reached, and the protein was eluted by 3 CVs of MBP elution buffer (20 mM Tris/HCl, 200 mM NaCl, 1 mM EDTA, 10 mM mltose, pH 7.4). The eluted fusion proteins were concentrated using 50 kDa MWCO Amicon® ultra centrifugal filters, diluted to 50 mL in intein cleaving buffer (137 mM NaCl, 2.7 mM KCl, 8.1 mM Na_2HPO_4, 1.76 mM KH_2PO_4, 40 mM bis-Tris, 2 mM EDTA, pH 6), and incubated at 37ºC for 16-20 hours. The cleaved proteins were concentrated using a 10 (ten) kDa MWCO Amicon® ultra centrifugal filters and separated from

the maltose binding protein by size exclusion chromatography using a HiLoad 16/600 Superdex 75 pg column. Protein purities were confirmed using gel electrophoreses (SDS-PAGE). The proteins were buffer exchanged into MES saline buffer (50 mM MES, 50 mM NaCl, pH 6.0) using PD-10 desalting columns.

K3T(VV), K3T(VN), HJH0 were expressed with an N-terminus strep tag in a pET-28a(+) vector. A 5 mL saturated culture of BL21(DE3) *E. coli* cells harboring each plasmid was used to inoculate a sterilized Terrific Broth (1L) supplemented with 50 mg L^{-1} kanamycin. The cell culture was grown to the mid-log growth phase (OD_{600nm} 0.4-0.6) in a shaking incubator at 37ºC. The culture was transferred to a 20ºC shaking biological incubator and induced with 0.3 mM IPTG. The culture was incubated for an additional 16-20 hours. The cells were harvested by centrifugation at 5,000 g for 15 minutes, resuspended in Strep binding buffer (100 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, pH 8), lysed via sonication, and clarified by centrifugation at 10,000 g for 1 hour. The clarified cell lysate was injected onto a 5-ml StrepTrap XT column and eluted using 10 CVs of Strep elution buffer (100 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 50 mM Biotin, pH 8).

pET-28a(+): (Halo-A0A7), pET-28a(+): (Halo-HEW5), and pET-28a(+): (Halo-RTX) plasmids were transformed into BL21(DE3) *E. coli* cells. A saturated culture was used to inoculate a sterilized Terrific Broth (1L) supplemented with 50 mg L^{-1} kanamycin. The cell culture was grown to the mid-log growth phase (OD_{600nm} 0.4-0.6) in a shaking incubator at 37ºC, and the expression was induced with 0.3 mM IPTG for 4-6 hours at 37ºC or 16-20 hours at 20ºC. The cells were harvested by centrifugation at 5,000 g for 15 minutes, resuspended in Halotag immobilization buffer (100 mM Tris/HCl, 150 mM NaCl, pH 7.6), lysed via sonication, and clarified by centrifugation at 10,000 g for 1 hour. The clarified cell lysate was used directly to immobilize the protein on the HaloLink™ resin by adding 10 mg target protein per mL of HaloLink™ beads,

according to the manufacturer's instructions. The resin was packed into chromatography columns with 6.6 mm inner diameter.

Förster resonance energy transfer (FRET). Emission spectra (460nm-550nm) of CFP-protein-EYFP samples (1 μM in 50 mM Tris-HCl buffer pH 7.4) titrated with Ln^{3+} ions were recorded in a black 96-well plate following CFP excitation at 420 nm using a SpectraMax M2 spectrophotometer (Molecular Devices). Ion-induced conformational change of the protein reduces the distance between the two fluorophores promoting FRET, which is characterized by a decrease in CFP (donor fluorophore) emissions at 480 nm and an increase in EYFP (acceptor fluorophore) emissions at 530 nm. FRET efficiency is calculated as the ratio of acceptor fluorophore emission to the sum of donor and acceptor emissions. The apparent disassociation consonant ($k_{d,\ apparent}$) is determined by fitting relative FRET efficiencies to the Hill equation.

Circular dichroism (CD) spectrophotometry. Protein samples (60 μM) were prepared in low MES buffer (1 mM MES, 50 mM KCl, pH 6.0) and titrated with $CaCl_2$ or $NdCl_3$. CD spectra scans (250 nm-190 nm) were performed at a rate of 1 nm per second in 0.1 mm pathlength quartz cuvette using Chirascan V100 Circular Dichroism Spectrometer at 25ºC. The reported spectra are an average of three scans per sample.

Determination of high affinity binding sites using xylenol orange (XO). One mL of M4 buffer (20 mM MES, 20 mM potassium acetate, 100 mM KCl, pH 6.1) supplemented with 10 μM XO and 5 μM protein was placed in a 1-cm pathlength disposable cuvette (BRAND). $YbCl_3$ (1 mM in M4 buffer) was titrated into the solution and mixed using a P200 pipette. The absorbance was recorded at 575 nm using SpectraMax QuickDrop UV-Vis Spectrophotometer (Molecular Devices), and the titration continued until a plateau was reached.[9, 10] The total volume of the titrant

did not exceed 3% of the initial volume, and the absorbance was corrected for the change in volume. The difference in metal ion concentration needed to reach 10% saturation of XO, in the presence and absence of the protein, was used to determine the number of high affinity (sub-micromolar) lanthanide binding sites per protein.

Binding sites analysis. AlphaFold3[11] was used to predict protein structures containing docked calcium ions from primary amino acid sequences. The coordination sphere between the metal and the protein was visualized, and details such as binding geometry and bond lengths were extracted using MolStar[12]. The binding site was represented as a spherical cavity with a radius equivalent to the average metal coordination bond length.

Isothermal titration calorimetry (ITC). The heat of interaction between the protein and lanthanides was measured via MicroCal AutoITC 200. The results were analyzed using MicroCal Origin software supplied with the equipment to determine the mechanisms of binding, the stoichiometries, and the binding affinities. In each experiment, 30 μM protein is transferred to the ITC cell and titrated via a syringe with 3 mM $LnCl_3$ at 25ºC; 29 injections delivering a total of 40 μL of ligand per experiment were performed: the first 12 injections delivered 0.5 μL each, while the subsequent 17 injections delivered 2 μL each. Injections delivered the ligand to the ITC cell at a rate of 0.5 μL/second, with 150 seconds spacing between injections. The proteins were buffer exchanged into MES saline buffer (50 mM MES, 50 mM NaCl, pH 6.0) using PD-10 desalting columns prior to ITC experiments. Lanthanide stock solutions (100 mM) were prepared in 1 mM HCl solution, and diluted to 3 mM in MES saline buffer. Control experiments without protein were preformed to account for the heat of dilution and mixing.

Separation experiments. The binding capacity of the packed column was determined by injecting 20 μmoles of lanthanum on the column, the column was washed with 30 mL of 20 mM sodium acetate pH 5.5, followed by elution of bound ions with 30 mL of 20 mM sodium acetate pH 3.0. 1 mL fractions were collected and μmoles in each fraction were determined using the Arsenazo assay as described previously.[13] The column binding capacity (Bt) was calculated as the sum of μmoles of ions in the elution step.

REEs separation chromatography was performed on an ÄKTA Explorer FPLC system using a 7-mL column containing immobilized protein. The separation was carried out at a flow rate of 4 mL/min. A five-component REE mixture (Dy, Y, Sm, La, Nd) or a two-component mixture (La, Nd) was prepared by diluting concentrated working stocks in water into a 50 mM MES/50 mM NaCl buffer at pH 6.0, achieving a final total REE concentration of 0.75 mM. Initially, the column was equilibrated with 30 mL of 50 mM MES/50 mM NaCl buffer (pH 6.0). Following equilibration, 5 mL of the REE sample was applied to the column using the ÄKTA P960 sample pump. To ensure the complete application of the sample, an additional 30 mL of the same buffer was passed through the sample pump tubing. The column was then washed with an additional 20 mL of 50 mM MES/50 mM NaCl buffer. Single fractions were collected during the sample application and column washing steps to assess whether the REEs bound to the column. Elution was performed using an isocratic gradient, where 50 mM MES/50 mM NaCl buffer (pH 3.0) was flowed over the column. Fractions of 3 mL were collected over a total elution volume of 150 mL.

For the simulated leachate mixture, 0.5 mL of an impurity/REE mixture (Mg – 10.76 mg/L; Ca – 71.33 mg/L; Fe – 6.01 mg/L; Sr – 3.08 mg/L; Y – 0.63 mg/L; La – 194.59 mg/L; Ce – 64.39 mg/L; Pr – 24.41 mg/L; Nd – 62.46 mg/L; Sm – 4.14 mg/L; Eu – 0.65 mg/L; Gd – 3.35 mg/L; Tb – 0.24 mg/L; Dy 0.20 mg/L) was loaded onto the column in water, the column was washed with 50 mL of 20 mM sodium acetate pH 5.5, then a linear gradient was performed in 20 mM sodium acetate

from pH 4.75 to 3.5, followed by a 50 mL wash with 20 mM sodium acetate pH 3.0 to remove all bound REEs. 10 mL fractions were collected throughout (except washes which were collected using 25 mL fractions) and quantified by ICP-MS.

Supplementary Tables

Table S1: Calcium and lanthanide binding domains used as seed sequences for bioinformatic search.

Protein	Domain	Pfam ID	Reference
Calmodulin / Lanmodulin	EF-hand	PF09068	[9, 14]
Synaptotagmin-1	C2	PF00168	[15, 16]
Bγ crystallins	Cystall_2	PF09076	[17, 18]
Thrombospondin	TSP3 or TT3R	PF18884	[19]
Psychrophilic alkaline protease (PAP)	PPE_SVP	PF12484	[20, 21]
YokF	Excalibur	PF05901	[14]
EGF	Epidermal growth factor	PF07645	[22]
XoxF5	PQQ-dependent dehydrogenase	PF01011	[23-27]

Table S2: PCR primers for cloning.

Oligo	Sequence (overlap/ANNEAL)
CFP-A0A7-EYFP.FWD	5’ gctgtacaagatgccggtaccaAGTACGTCTGAAGATCAG 3’
CFP-A0A7-EYFP.REV	5’ cctcgcccttgctcaccatctcgagTGATGAGGAACTGCTATC 3’
CFP-HEW5-EYFP.FWD	5’ gctgtacaagatgccggtaccaATGGGATCTGGTCCTAGTTC 3’
CFP-HEW5-EYFP.REV	5’ cctcgcccttgctcaccatctcgagGGAGCCATCATGTACAAC 3’
CFP-HJH0-EYFP.FWD	5’ gctgtacaagatgccggtaccaATGTCAGATTCAGAGGTGAAC 3’
CFP-HJH0-EYFP.REV	5’ cctcgcccttgctcaccatctcgagACGCTCACAGACAATCCC 3’
CFP-K3T(VN)-EYFP.FWD	5’ gctgtacaagatgccggtaccGTCCGTTGCCTCGCAGAC 3’
CFP-K3T(VN)-EYFP.REV	5’ cctcgcccttgctcaccatctcgagATTTGTTACCCCATCGTTATCGC 3’
CFP-K3T(VV)-EYFP.FWD	5’ gctgtacaagatgccggtacctTCCGTTGCCTCGCAGACTG 3’
CFP-K3T(VV)-EYFP.REV	5’ cctcgcccttgctcaccatctcgagAACAAATTCGCAAGGATCCGTAG G 3’
CFP-MbIBP-EYFP.FWD	5’ gctgtacaagatgccggtaccgATGAACGTTTCGCAAAGCAATTC 3’
CFP-MbIBP-EYFP.REV	5’ cctcgcccttgctcaccatctcgagGTAATCCGCGAAGTCGGC 3’
CFP- CaM(III,IV)-EYFP.FWD	5’ gctgtacaagatgccggtaccaATGAAAGACACGGATTCG 3’
CFP- CaM(III,IV)-EYFP.REV	5’ cctcgcccttgctcaccatctcgagTTTAGCCGTCATCATTTG 3’
MBP-I-A0A7.vFWD	5’ ctgatagcagttcctcatcaTAAAAGCTTGGCACTGGC 3’
MBP-I-A0A7.vREV	5’ CGGTACCGGCATGTTGTG 3’
MBP-I-A0A7.iFWD	5’ tacacaacatgccggtaccgAGTACGTCTGAAGATCAG 3’

MBP-I-A0A7.iREV	5’ TGATGAGGAACTGCTATC 3’
MBP-I-HEW5.vFWD	5’ TAAAAGCTTGGCACTGGC 3’
MBP-I-HEW5.vREV	5’ CGGTACCGGCATGTTGTG 3’
MBP-I-HEW5.iFWD	5’ tacacaacatgccggtaccgAGTGGGCCAAGTTCAACC 3’
MBP-I-HEW5.iREV	5’ cggccagtgccaagcttttagtggtggtggtggtggtgACCGTCATGAACAACGGTG 3’
MBP-I-CaM(III,IV).vFWD	5’ TAAAAGCTTGGCACTGGC 3’
MBP-I-CaM(III,IV).vREV	5’ CGGTACCGGCATGTTGTG 3’
MBP-I-CaM(III,IV).iFWD	5’ tacacaacatgccggtaccgATGAAAGACACGGATTCG 3’
MBP-I-CaM(III,IV).iREV	5’ cggccagtgccaagcttttagtggtggtggtggtggtgTTTAGCCGTCATCATTTG 3’

Table S3: Proteins primary sequences.

Construct	Amino acid sequence
CFP-RTX-EYFP	MVSKGEELFTGVVPILVELDGDVNGHRFSVSGEGEGDATYGKLTLKFICT TGKLPVPWPTLVTTLTWGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTI FFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYISH NVYITADKQKNGIKAHFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDN HYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKMPVP GSARDDVLIGDAGANVLNGLAGNDVLSGGAGDDVLLGDEGSDLLSGDA GNDDLFGGQGDDTYLFGVGYGHDTIYESGGGHDTIRINAGADQLWFAR QGNDLEIRILGTDDALTVHDWYRDADHRVEIIHAANQAVDQAGIEKLVE AMAQYPD LEMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFI CTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQE RTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNY NSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVL LPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKGEHH HHHH
CFP-A0A7-EYFP	CFP-MPVP STSEDQYYDTNYDGQVDTVVTDTDGNGVYDAAVYDTDGNGVADTVAY DSDENGVVDTVGFDYNEDGVVDEVVTDYNEDGYADSSSSS LE-EYFP-GEHHHHHH
CFP-HEW5-EYFP	CFP-MPVP MGSGPSSTEYDADGDGYVDTRESDTDGDGYVDTIETDTDGDGWVDTVA TDTDGDGYIDTVATDTDGDGYADVVETDTDGDGYTDEVAYDADGDGYI DTVEADTDGDGYTDTVVHDGS LE-EYFP-GEHHHHHH
CFP-HJH0-EYFP	CFP-MPVP GTDPHFGTCGEANANGYGDYVSGVDPEYDWYTDRDSDGIVCER LE-EYFP-GEHHHHHH
CFP-K3T(VN)-EYFP	CFP-MPVP SVASQTVPTSTVWNDLDCDNDGVTNGDEIANGTDPLNPDTDGDGVTDG DEIIDGTDPTDPCEFVVASQTLPTSILLRDPVCDNDGVSNGDEIANGTDPL

	NPDTDGDGVTDGDEIIDGTDPTDPCEFVVASQTVPTSTVWNDLDCDNDGVTN LE-EYFP-GEHHHHHH
CFP-K3T(VV)-EYFP	CFP-MPVP SVASQTVPTSTVWNDLDCDNDGVTNGDEIANGTDPLNPDTDGDGVTDGDEIIDGTDPTDPCEFV LE-EYFP-GEHHHHHH
CFP-CaM(III,IV)-EYFP	CFP-MPVP SMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK LE-EYFP-GEHHHHHH
MBP-Intein-↑RTX	MKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNNNNLGIEGR ISEFALAEGTRIFDPVTGTTHRIEDVVGGRKPIHVVAAAKDGTLRARPVVSWFDQGTRDVIGLRIAGGAILWATPDHKVLTEYGWRAAGELRKGDRVAQPRRFDGFGDSAPIPARVQALADALDDKFLHDMLAEELRYSVIREVLPTRRARTFGLEVEELHTLVAEGVVVHN↑MPVP GSARDDVLIGDAGANVLNGLAGNDVLSGGAGDDVLLGDEGSDLLSGDAGNDDLFGGQGDDTYLFGVGYGHDTIYESGGGHDTIRINAGADQLWFARQGNDLEIRILGTDDALTVHDWYRDADHRVEIIHAANQAVDQAGIEKLVEAMAQYPD
MBP-Intein-↑A0A7	MBP-NSSSNNNNNNNNNNLGIEGR Intein↑MPVP STSEDQYYDTNYDGQVDTVVTDTDGNGVYDAAVYDTDGNGVADTVAYDSDENGVVDTVGFDYNEDGVVDEVVTDYNEDGYADSSSSS

MBP-Intein-↑HEW5	MBP-NSSSNNNNNNNNNLGIEGR Intein↑MPVP MGSGPSSTEYDADGDGYVDTRESDTDGDGYVDTIETDTDGDGWVDTVATDTDGDGYIDTVATDTDGDGYADVVETDTDGDGYTDEVAYDADGDGYIDTVEADTDGDGYTDTVVHDGS
MBP-Intein-↑CaM(III, IV)	MBP-NSSSNNNNNNNNNLGIEGR Intein↑MPVP MKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK
Streptag-Sumo-K3T(VN)	MGSSWSHPQFEKGS MSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGS VASQTVPTSTVWNDLDCDNDGVTNGDEIANGTDPLNPDTDGDGVTDGDEIIDGTDPTDPCEFVVASQTLPTSILLRDPVCDNDGVSNGDEIANGTDPLNPDTDGDGVTDGDEIIDGTDPTDPCEFVVASQTVPTSTVWNDLDCDNDGVTN
Streptag-Sumo-K3T(VV)	MGSSWSHPQFEKGS MSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGS VASQTVPTSTVWNDLDCDNDGVTNGDEIANGTDPLNPDTDGDGVTDGDEIIDGTDPTDPCEFV
Streptag-Sumo-HJH0	MGSSWSHPQFEKGS MSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGS GTDPHFGTCGEANANGYGDYVSGVDPEYDWYTDRDSDGIVCER

Table S4: p-values from one-way ANOVA testing of the statistical difference of the means of the calculated high affinity sites per binding domain via the XO competition assay. A p-value less than 0.05 indicates statistically significant differences in number of high affinity binding sites between the two tested groups.

Group 1	Group 2	p-value
K3T(VN)	A0A7	0.16
K3T(VN)	HEW5	3E-07
K3T(VN)	K3T(VV)	2E-08
K3T(VN)	CaM(III, IV)	1E-11
K3T(VN)	RTX	8E-12
K3T(VN)	HJH0	2E-12
K3T(VN)	Buffer	6E-12
A0A7	HEW5	2E-05
A0A7	K3T(VV)	7E-07
A0A7	CaM(III, IV)	2E-10
A0A7	RTX	1E-10
A0A7	HJH0	2E-11
A0A7	Buffer	7E-11
HEW5	K3T(VV)	0.43
HEW5	CaM(III, IV)	1E-06
HEW5	RTX	5E-07
HEW5	HJH0	5E-08
HEW5	Buffer	3E-07
K3T(VV)	CaM(III, IV)	3E-05
K3T(VV)	RTX	1E-05
K3T(VV)	HJH0	9E-07

K3T(VV)	Buffer	7E-06
CaM(III, IV)	RTX	1.00
CaM(III, IV)	HJH0	0.38
CaM(III, IV)	Buffer	0.99
RTX	HJH0	0.64
RTX	Buffer	1.00
HJH0	Buffer	0.84

Table S5: The apparent disassociation constant and the Hill coefficient determined for each domain for four lanthanides spanning light to heavy REEs. The errors represent the 95% confidence interval of the fit.

Lanthanide	$CeCl_3$		$NdCl_3$		$DyCl_3$		$YbCl_3$	
Protein	Kd	n	Kd	n	Kd	n	Kd	n
A0A7	120 ± 10	1.9 ± 0.2	90 ± 6	2.2 ± 0.3	81 ± 4	2.0 ± 0.2	68 ± 4	2.1 ± 0.2
CaM (III, IV)	39 ± 5	2.7 ± 0.8	50 ± 8	2.5 ± 0.9	78 ± 11	2.8 ± 1.0	67 ± 9	2.2 ± 0.7
HEW5	15 ± 4	1.4 ± 0.5	28 ± 7	1.1 ± 0.3	50 ± 9	1.2 ± 0.3	47 ± 6	1.4 ± 0.2
HJH0	65 ± 4	2.5 ± 0.4	59 ± 4	2.5 ± 0.4	58 ± 3	2.3 ± 0.3	60 ± 4	2.4 ± 0.3
K3T(VN)	85 ± 7	1.8 ± 0.2	51 ± 4	2.7 ± 0.5	61 ± 5	2.4 ± 0.4	57 ± 4	2.3 ± 0.3
K3T(VV)	71 ± 7	1.8 ± 0.3	74 ± 4	1.7 ± 0.2	55 ± 5	1.6 ± 0.2	69 ± 3	2.0 ± 0.1
RTX	110 ± 10	2.1 ± 0.5	69 ± 8	2.3 ± 0.6	41 ± 6	2.1 ± 0.7	36 ± 5	2.6 ± 1.0

Table S6: Disassociation constants for A0A7, HEW5, and RTX determined via ITC. The disassociation constants for calcium chloride were estimated as the ion concentration needed to induce conformational change by CD. The errors represent 95% confidence intervals.

Lanthanide	$k_{d, A0A7}$ (μM)	$k_{d, HEW5}$ (μM)	$k_{d, RTX}$ (μM)
$LaCl_3$	17 ± 2	5.2 ± 1.7	40 ± 4
$CeCl_3$	11 ± 2	4.9 ± 1.5	35 ± 9
$PrCl_3$	11 ± 2	4.6 ± 1.3	14 ± 5
$NdCl_3$	15 ± 2	6.2 ± 1.5	12 ± 3
$EuCl_3$	29 ± 6	6.5 ± 1.6	8.0 ± 1.0
$TbCl_3$	36 ± 6	7.2 ± 1.9	9.1 ± 1.6
$DyCl_3$	62 ± 13	9.6 ± 2.9	10 ± 2
$YbCl_3$	67 ± 16	39 ± 12	11 ± 2
$LuCl_3$	47 ± 5	36 ± 10	13 ± 2
$CaCl_2$	≈2000	≈750	≈250

Table S7: Molar composition of synthetic mixture simulating the low-cerium concentration stream of hydrochloric acid leachate of bastnäsite. The composition is representative of Mountain Pass rare earth concentrate with depleted cerium.[28, 29]

Light REEs					
$LaCl_3$	$CeCl_3$	$PrCl_3$	$NdCl_3$		
29%	7.3%	3.7%	9.0%		
Heavy REEs					
$SmCl_3$	$EuCl_3$	$GdCl_3$	$TbCl_3$	$DyCl_3$	YCl_3
0.6%	0.1%	0.4%	0.03%	0.02%	0.2%
Non-REEs					
$MgCl_2$	$CaCl_2$	$FeCl_2$	$SrCl_2$		
9.4%	37%	2.3%	0.7%		

Supplementary Figures

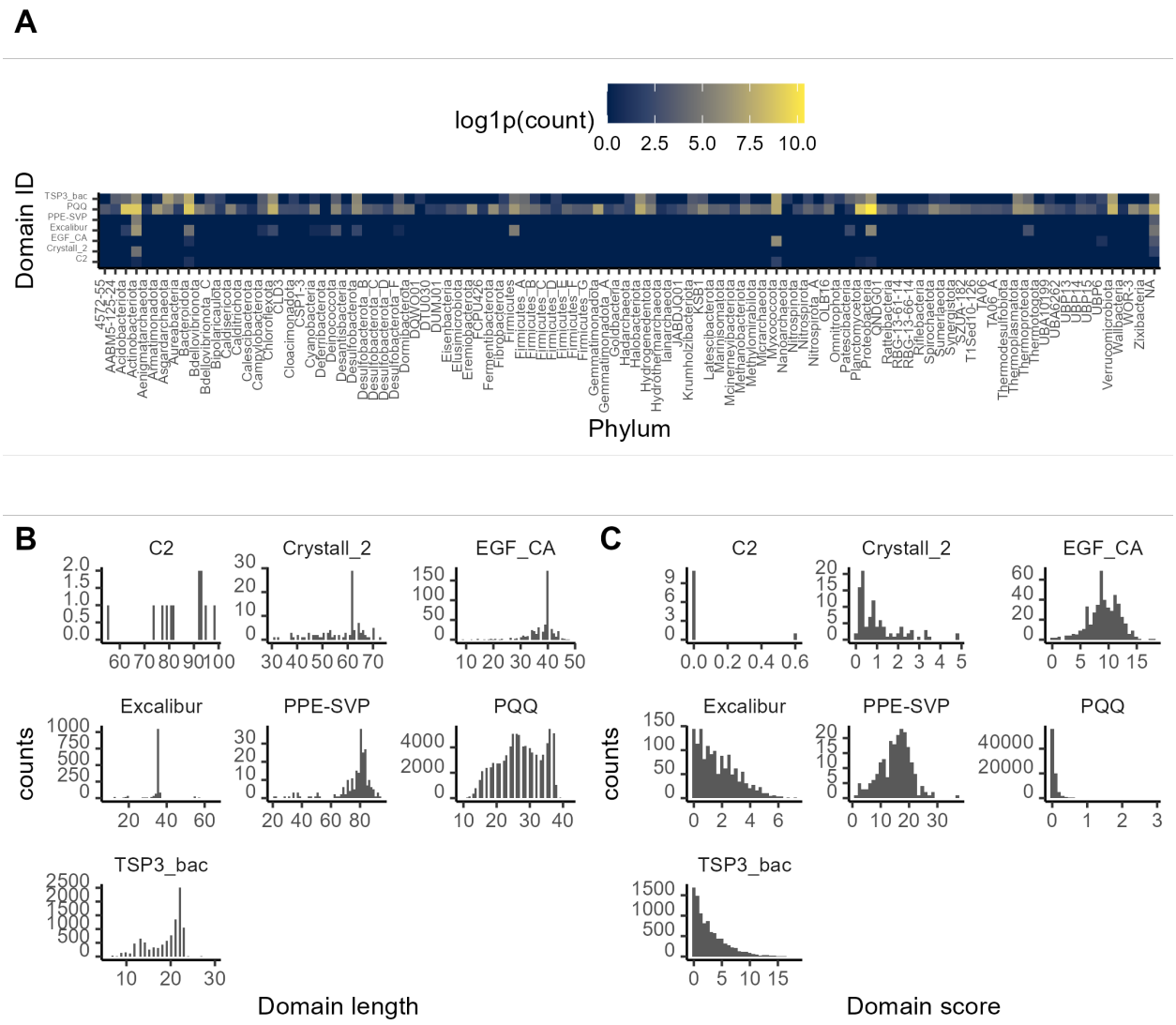

Figure S1: A) Protein sequence counts (natural logarithm + 1) of select calcium-binding domains detected within bacterial phyla. B) Distribution of calcium-binding domain lengths detected. C) Distribution of calcium-binding domain scores observed for proGenome protein sequences using profile HMM searches.

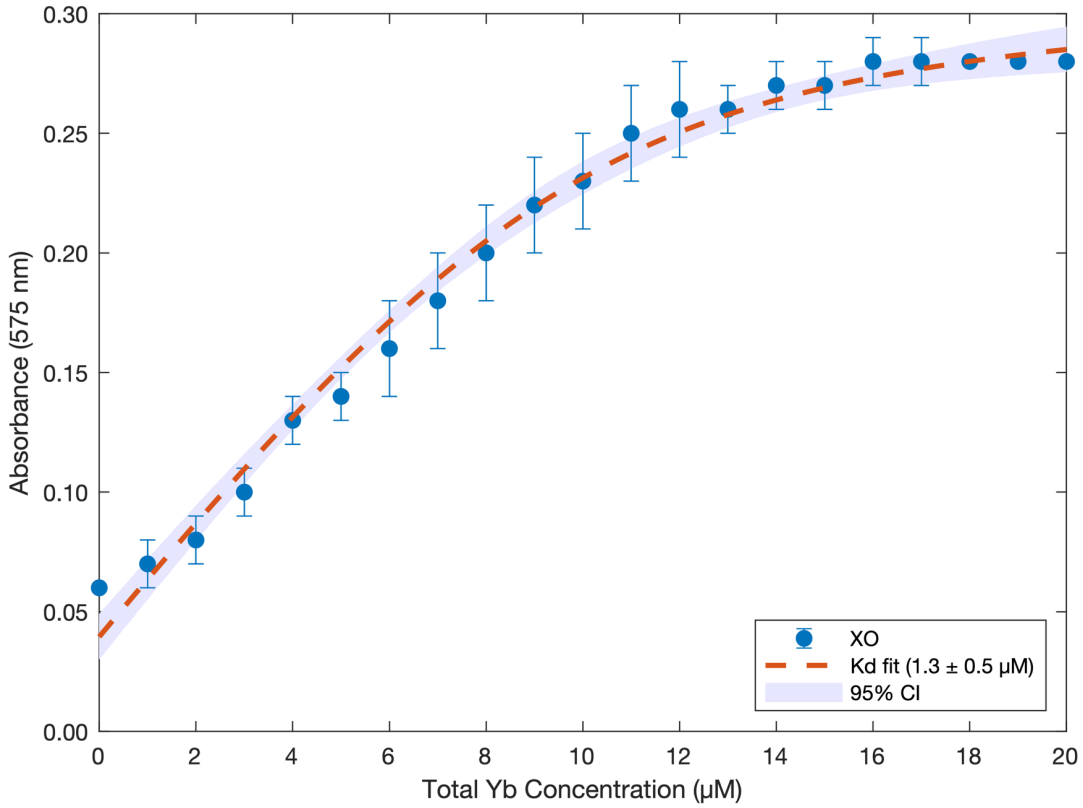

Figure S2: Estimated dissociation constant (Kd) of xylenol orange with Yb. The Kd was determined by fitting absorbance data to the quadratic binding equation, yielding a value of 1.3 ± 0.5 μM. Error bars represent the standard deviation from three independent trials.

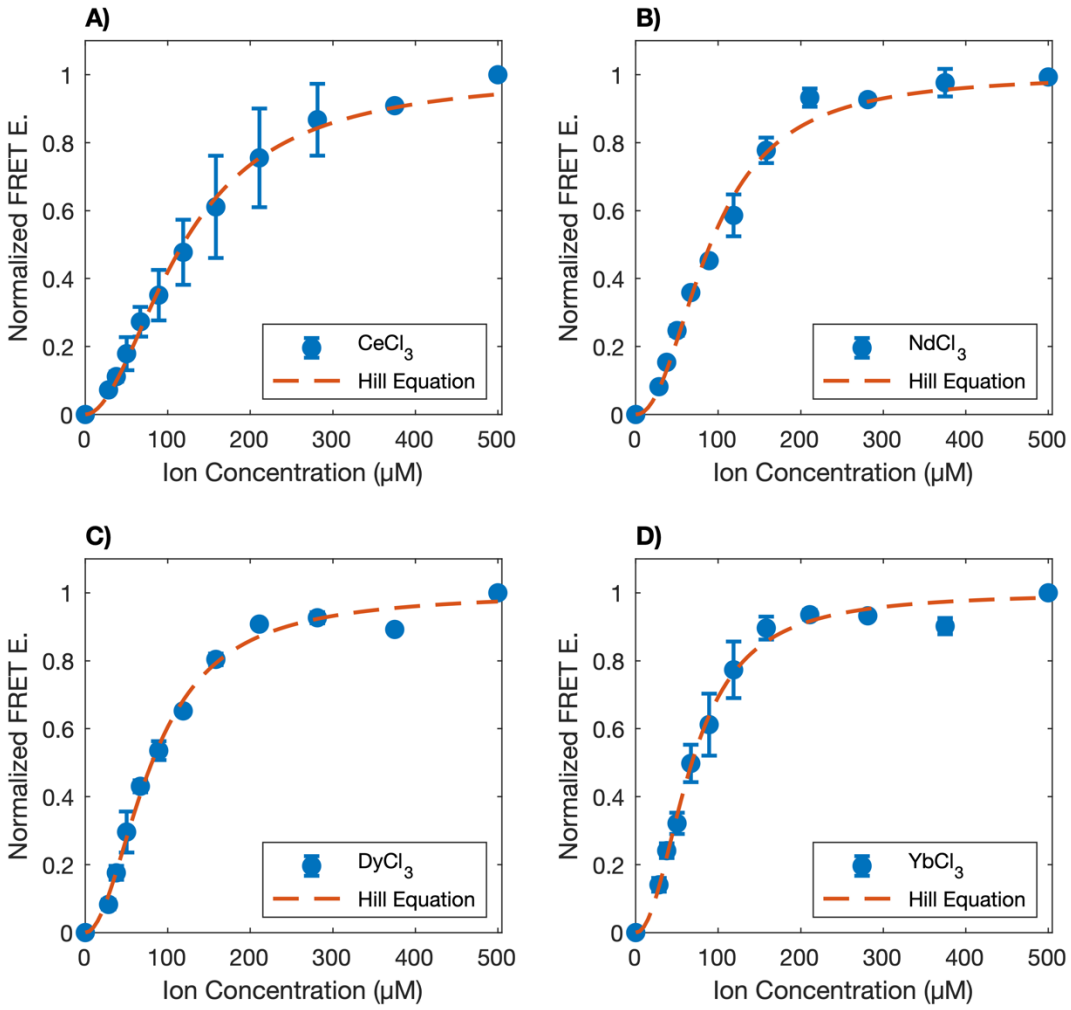

Figure S3: Normalized FRET efficiencies for A0A7 (CFP-A0A7-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants ($k_{d,\ apparent}$) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

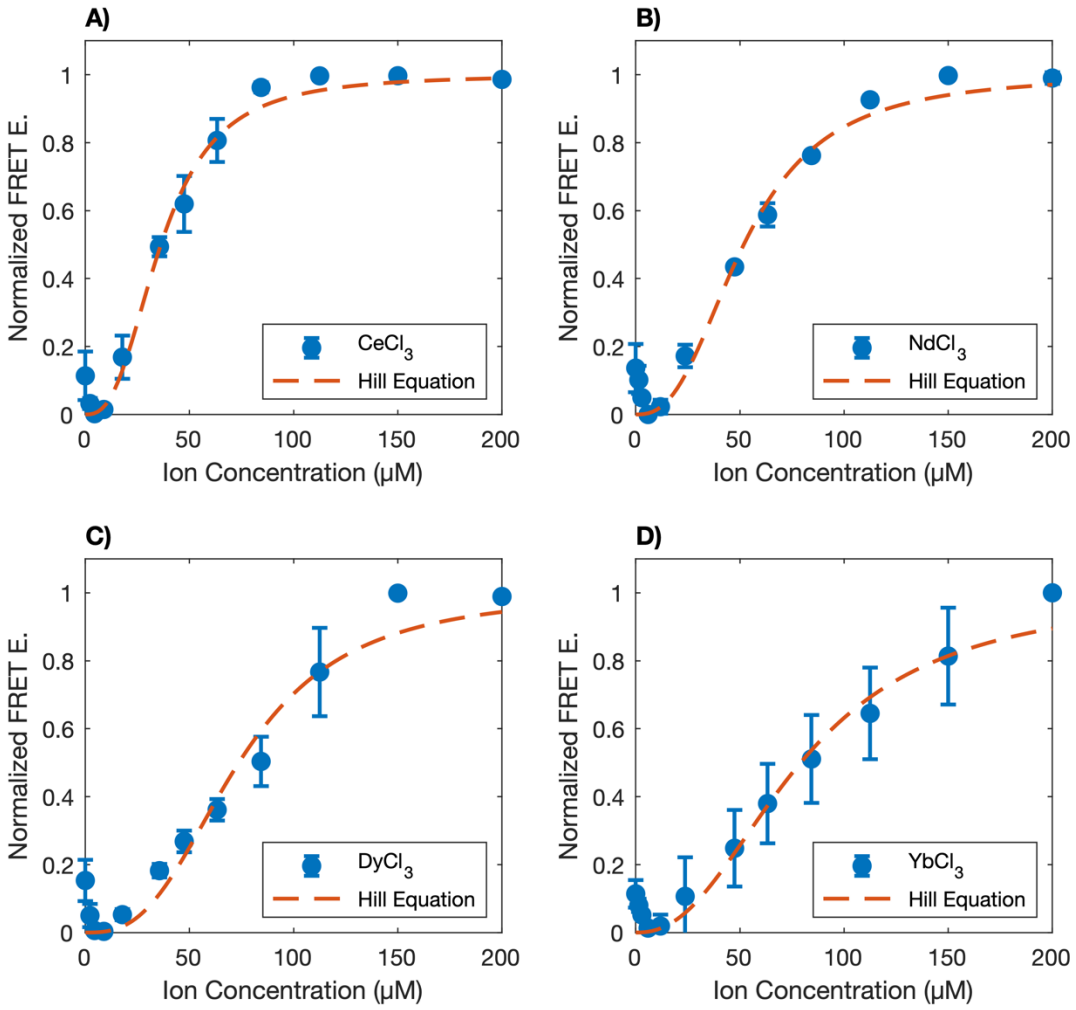

Figure S4: Normalized FRET efficiencies for CaM(III,IV) (CFP-CaM(III,IV)-EYFP) titrated with A) CeCl3, B) NdCl3, C) DyCl3, and D) YbCl3. The results were fitted to the Hill equation, and the estimated apparent disassociation constants (kd, apparent) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

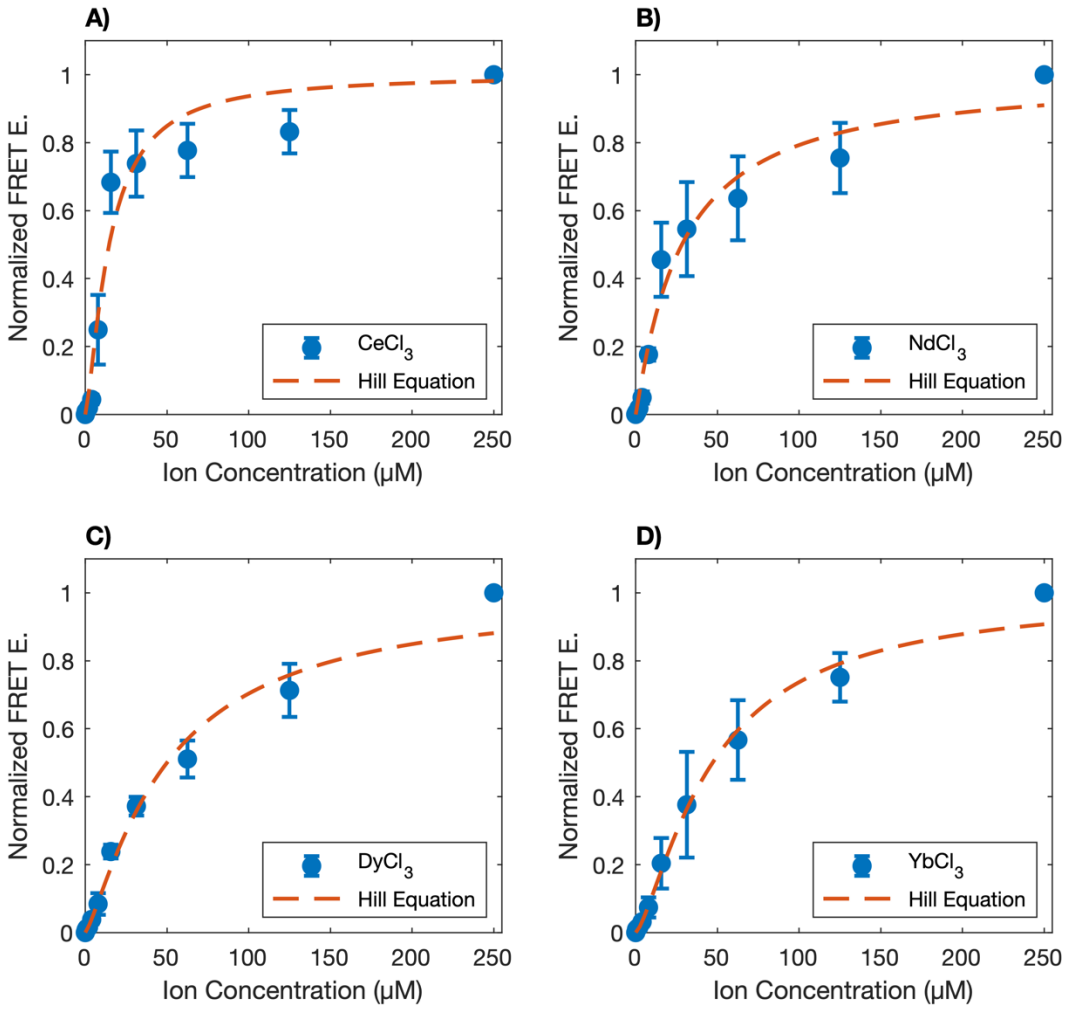

Figure S5: Normalized FRET efficiencies for HEW5 (CFP-HEW5-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants ($k_{d,\ apparent}$) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

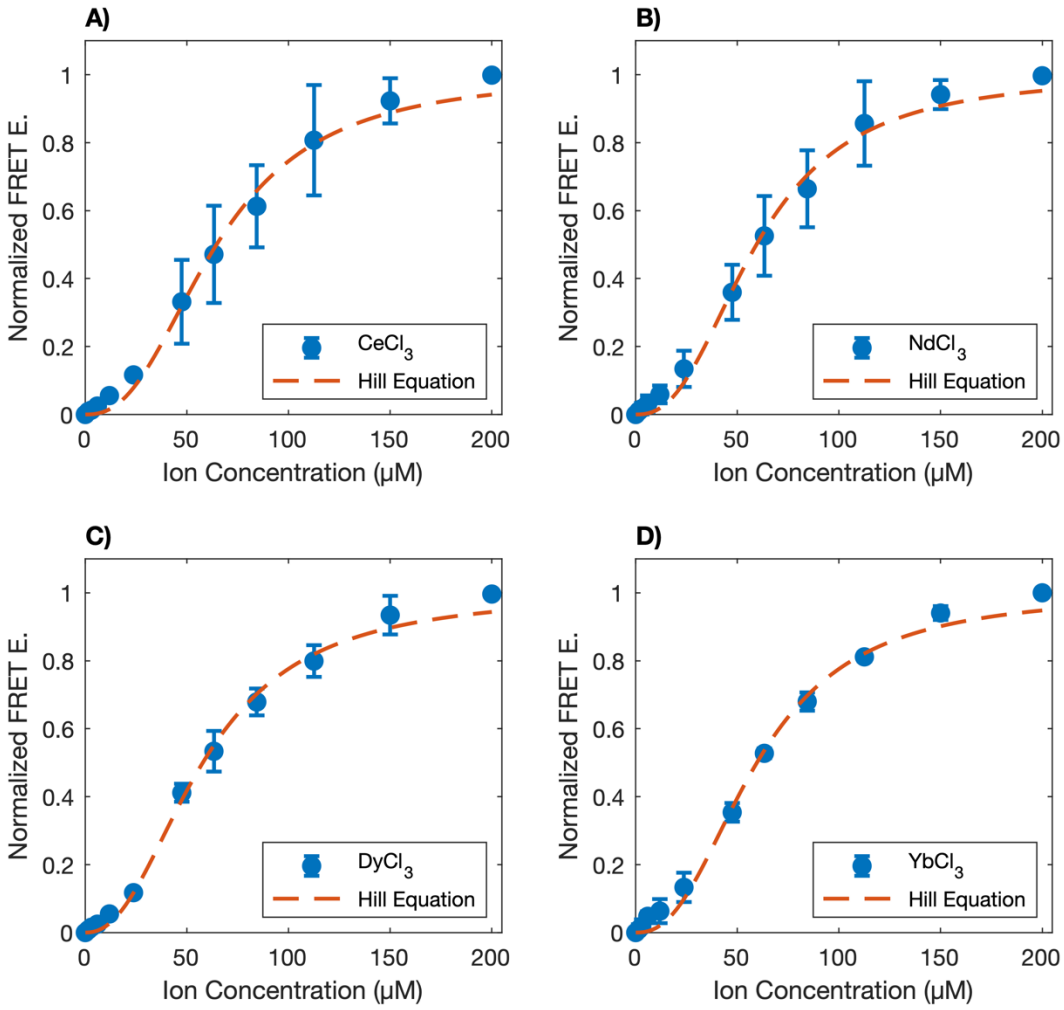

Figure S6: Normalized FRET efficiencies for HJH0 (CFP-HJH0-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants ($k_{d,\ apparent}$) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

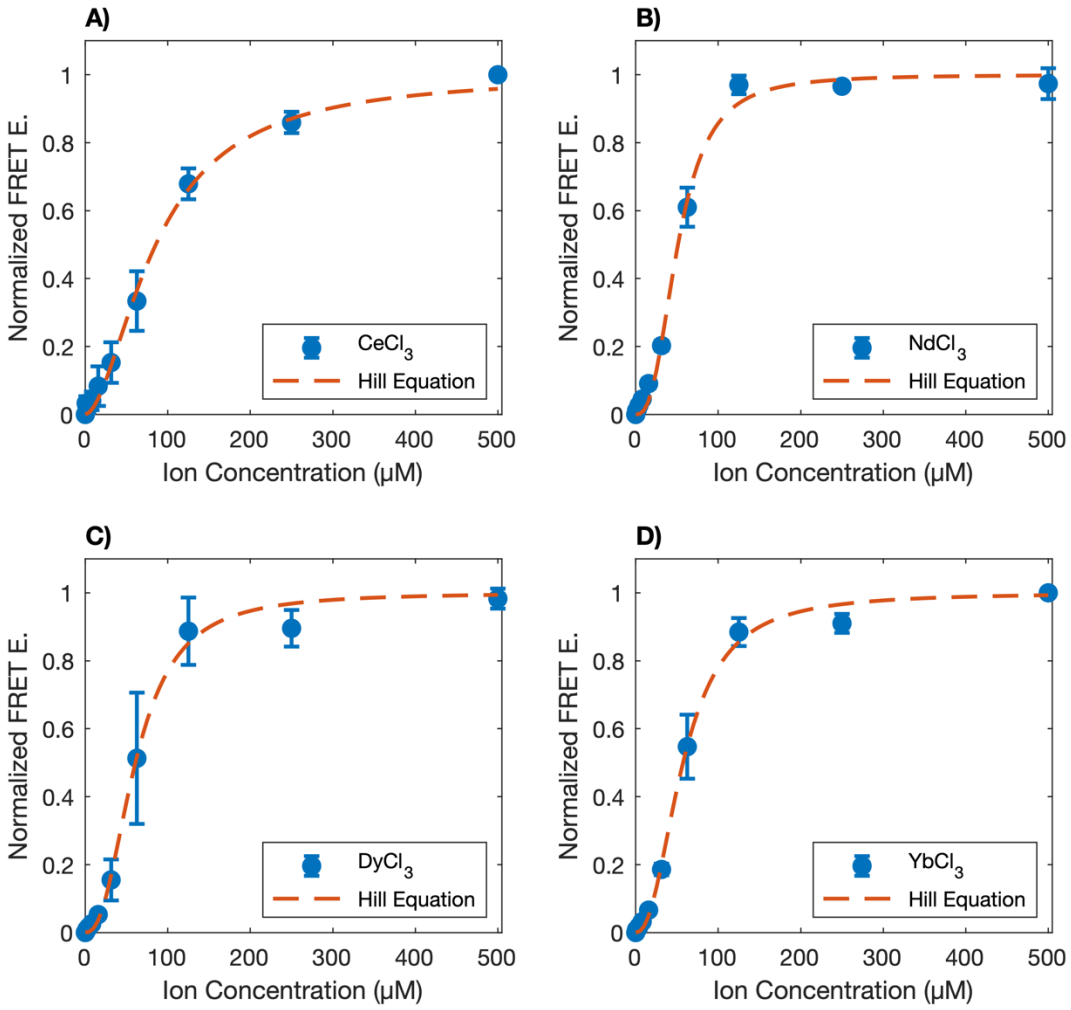

Figure S7: Normalized FRET efficiencies for K3T(VN) (CFP-K3T(VN)-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants ($k_{d,\ apparent}$) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

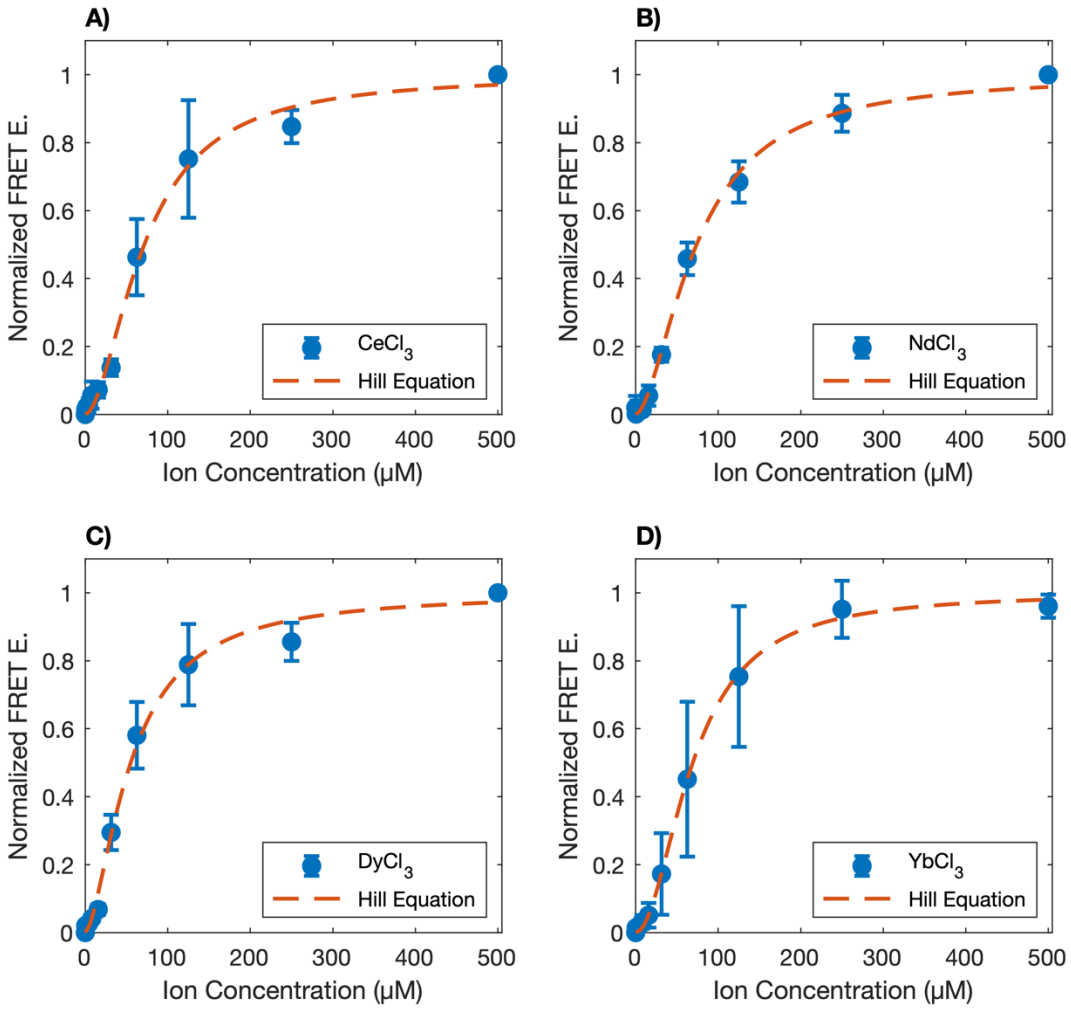

Figure S8: Normalized FRET efficiencies for K3T(VV) (CFP-K3T(VV)-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants (kd, apparent) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

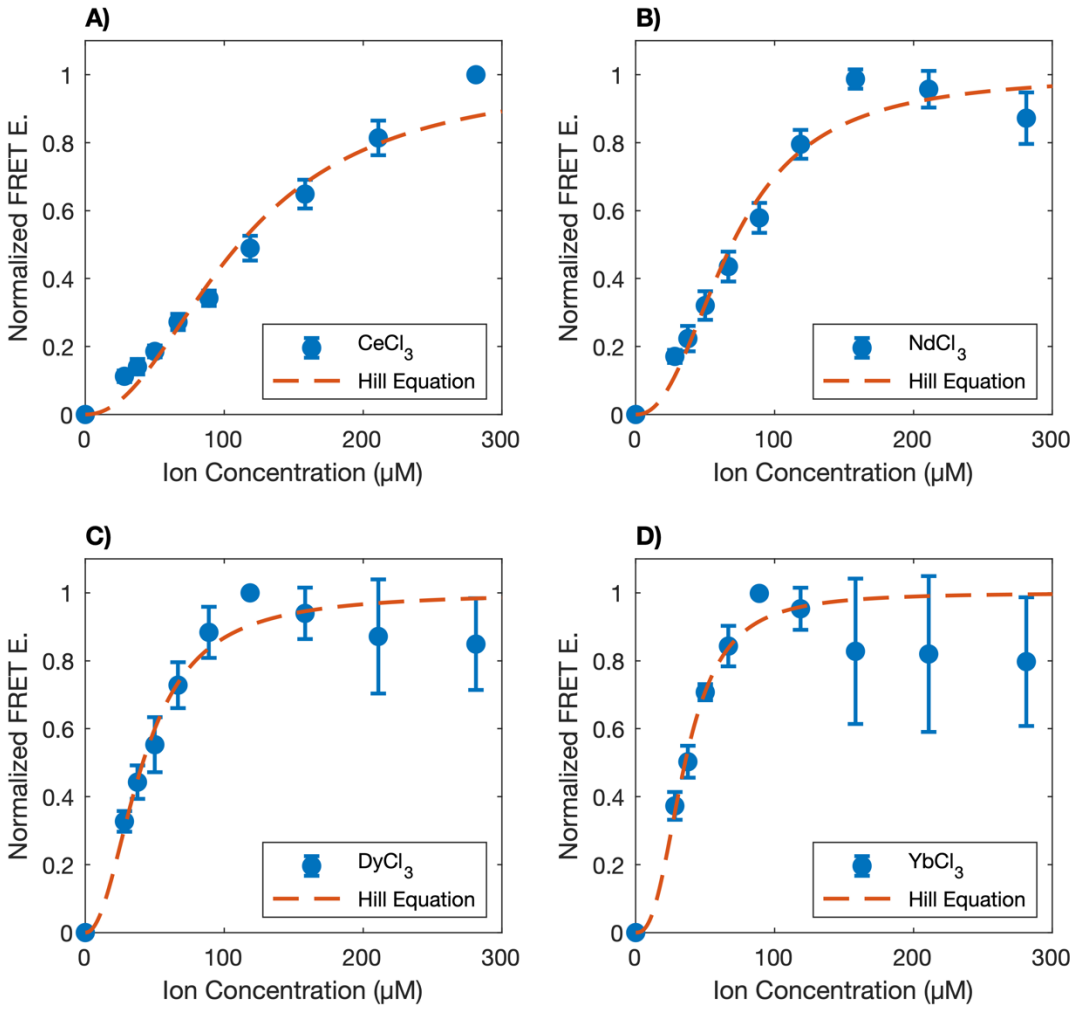

Figure S9: Normalized FRET efficiencies for RTX (CFP-RTX-EYFP) titrated with A) $CeCl_3$, B) $NdCl_3$, C) $DyCl_3$, and D) $YbCl_3$. The results were fitted to the Hill equation, and the estimated apparent disassociation constants ($k_{d,\ apparent}$) and the Hill coefficients (n) are reported in Table S4. The error bars represent the standard deviations of three independent trials.

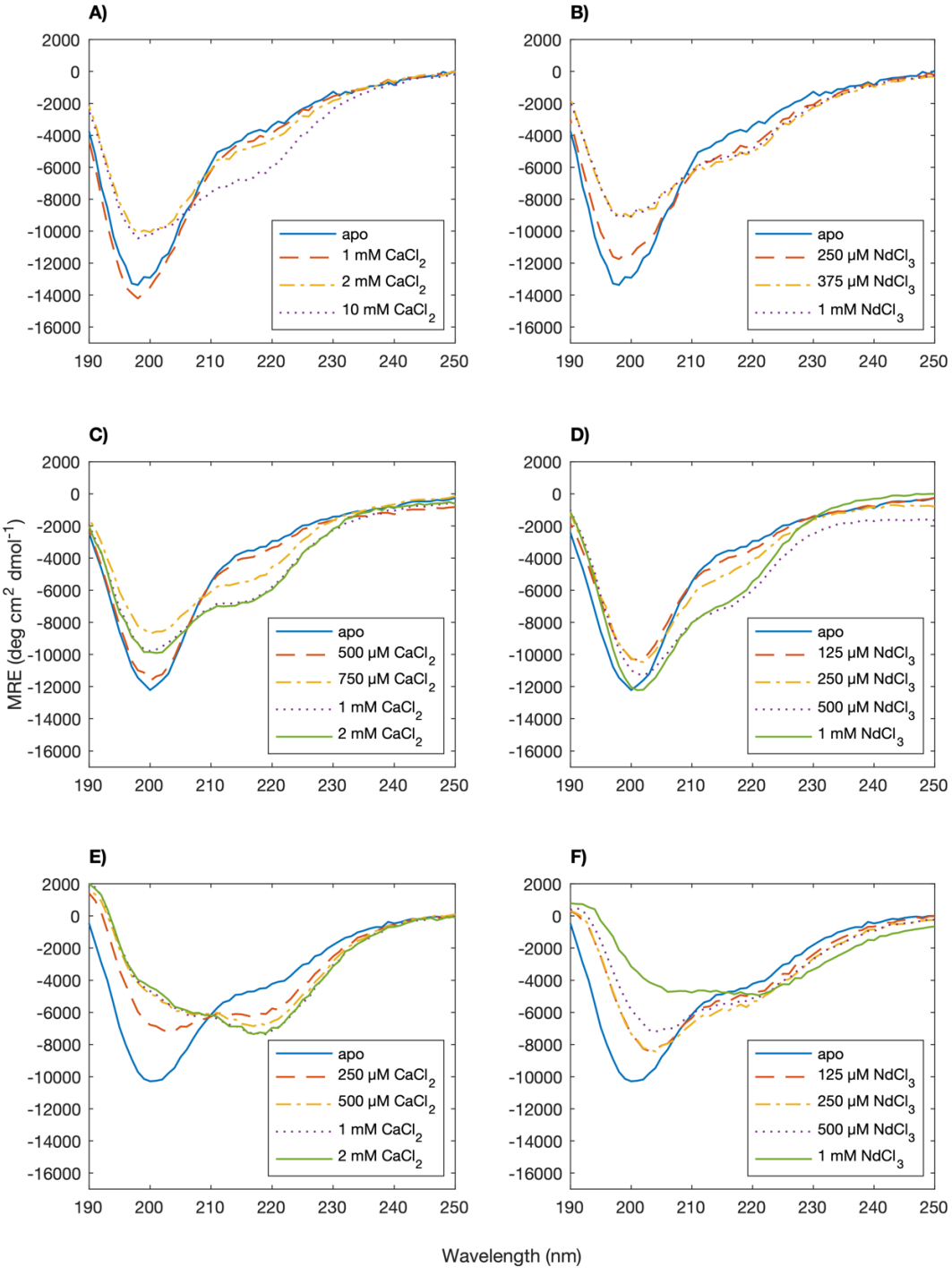

Figure S10: CD spectra of down-selected domains (60 µM) titrated with $CaCl_2$ and $NdCl_3$. A) A0A7-$CaCl_2$, B) A0A7-$NdCl_3$, C) HEW5-$CaCl_2$, D) HEW5-$NdCl_3$, E) RTX-$CaCl_2$, F) RTX-$NdCl_3$. The reported spectra are the means of three scans.

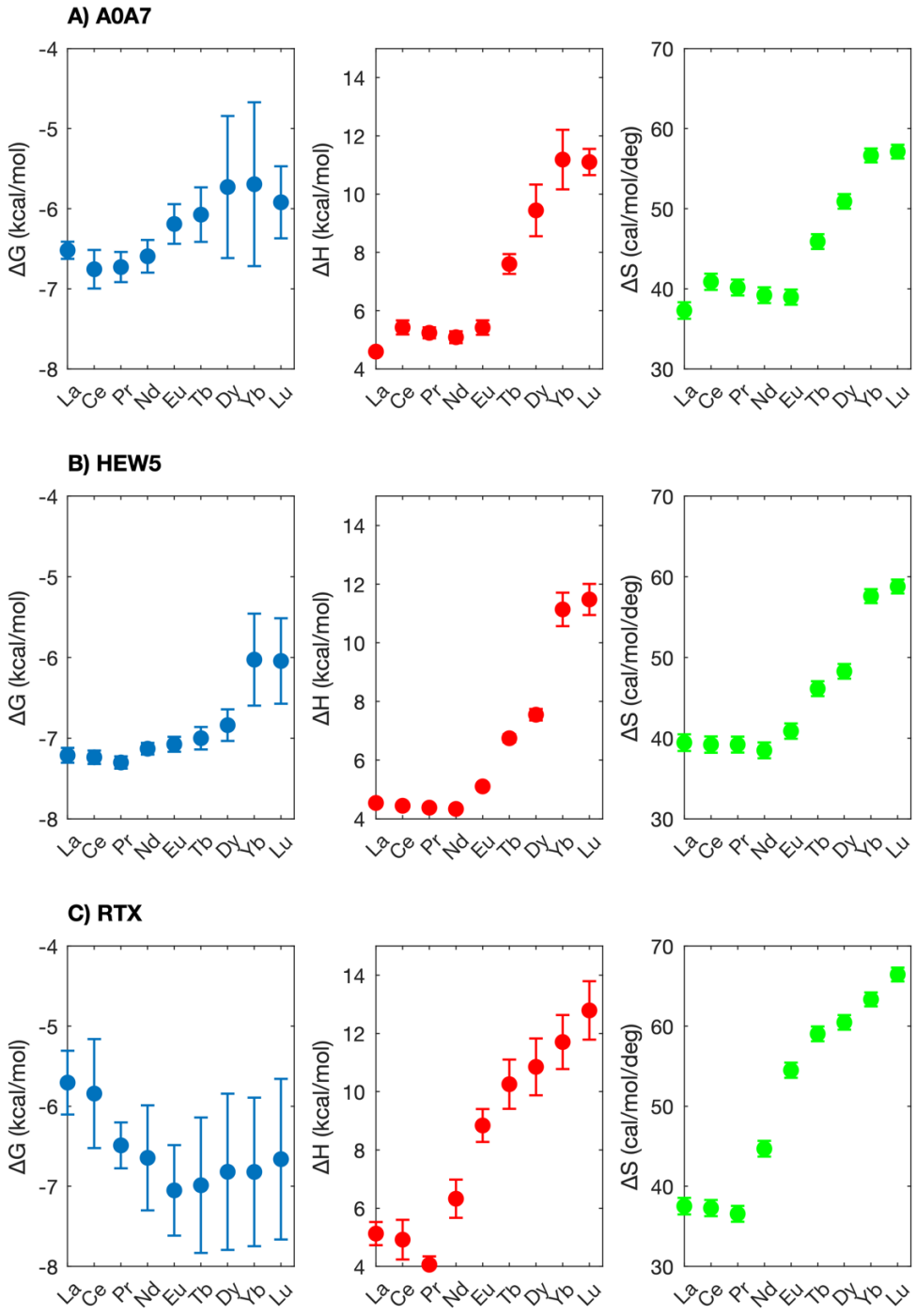

Figure S11: Summary of measured thermodynamic properties for protein-lanthanide interactions for A) A0A7, B) HEW5, and C) RTX.

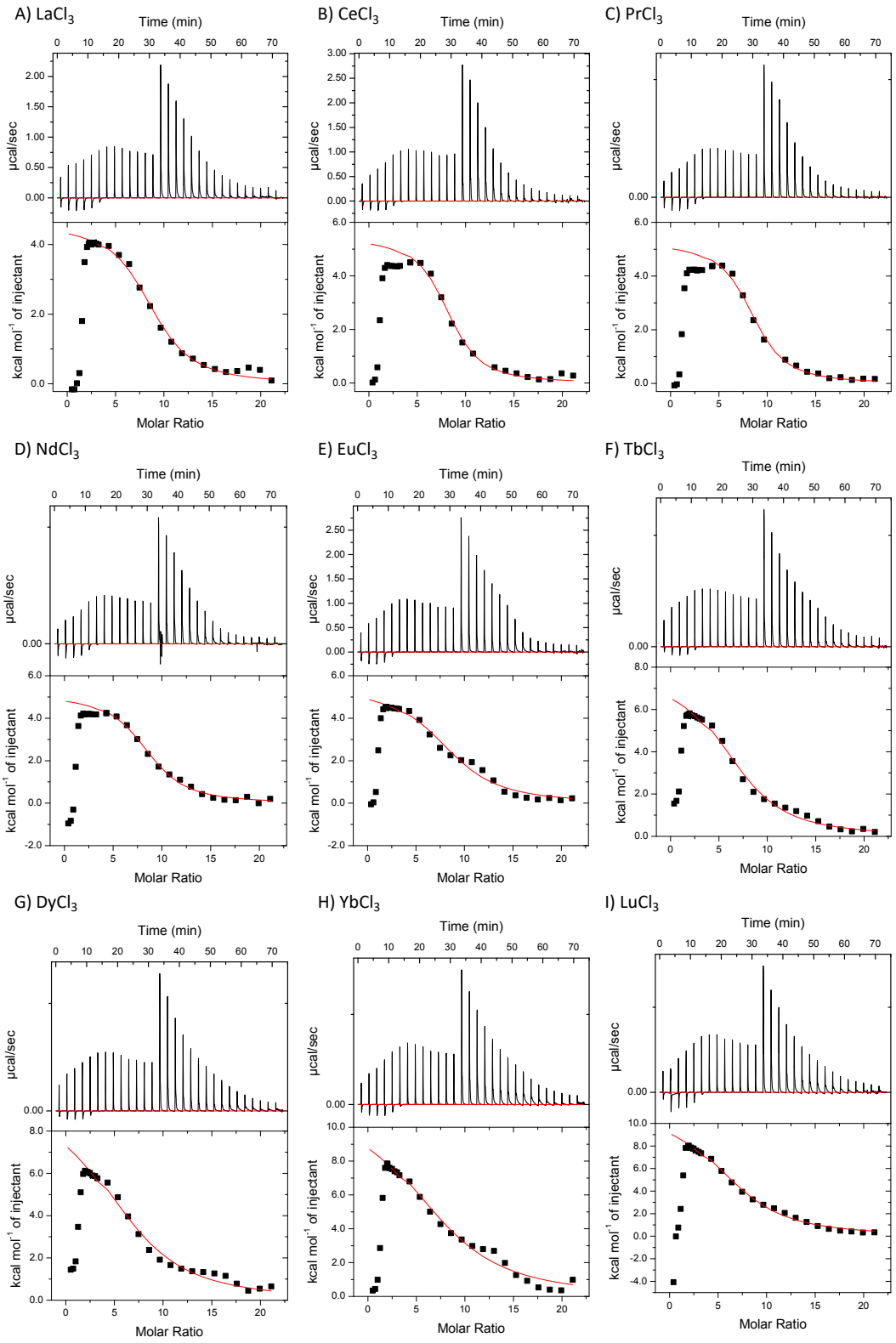

Figure S12: ITC isotherm for A0A7 titrated with A) $LaCl_3$, B) $CeCl_3$, C) $PrCl_3$, D) $NdCl_3$, E) $EuCl_3$, F) $TbCl_3$, G) $DyCl_3$, H) $YbCl_3$, and I) $LuCl_3$.

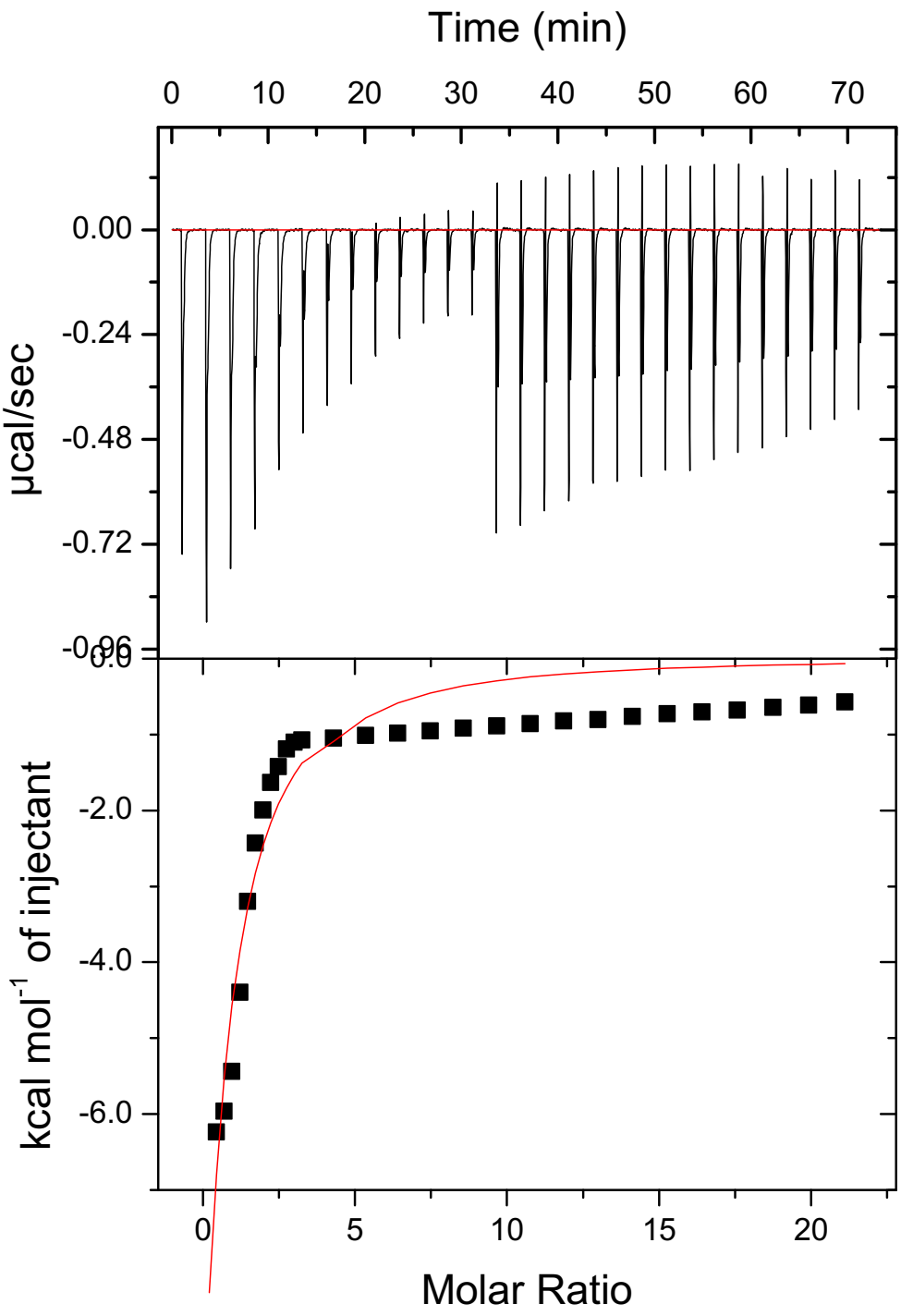

Figure S13: ITC isotherm for A0A7 titrated with $CaCl_2$.

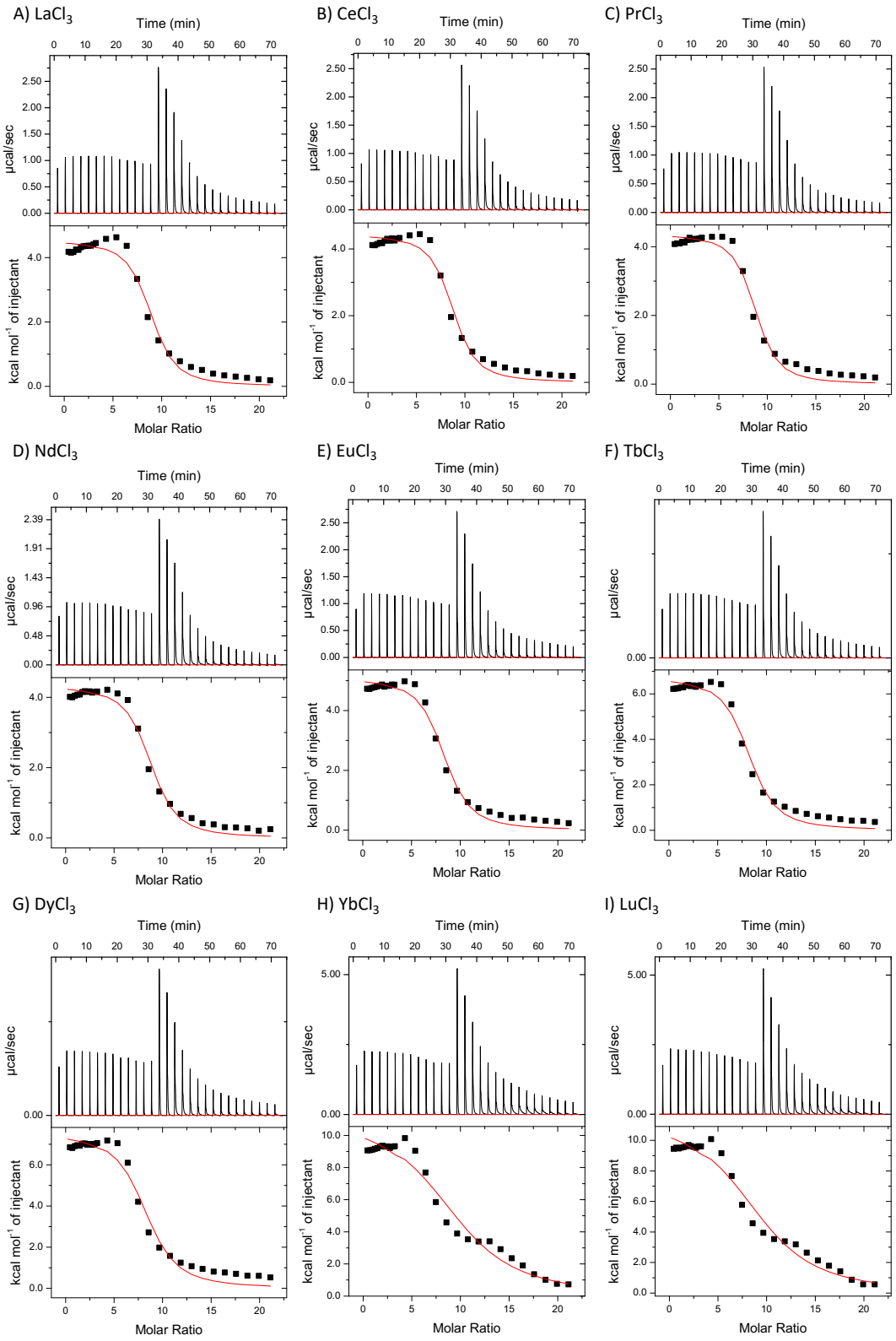

Figure S14: ITC isotherm for HEW5 titrated with A) $LaCl_3$, B) $CeCl_3$, C) $PrCl_3$, D) $NdCl_3$, E) $EuCl_3$, F) $TbCl_3$, G) $DyCl_3$, H) $YbCl_3$, and I) $LuCl_3$.

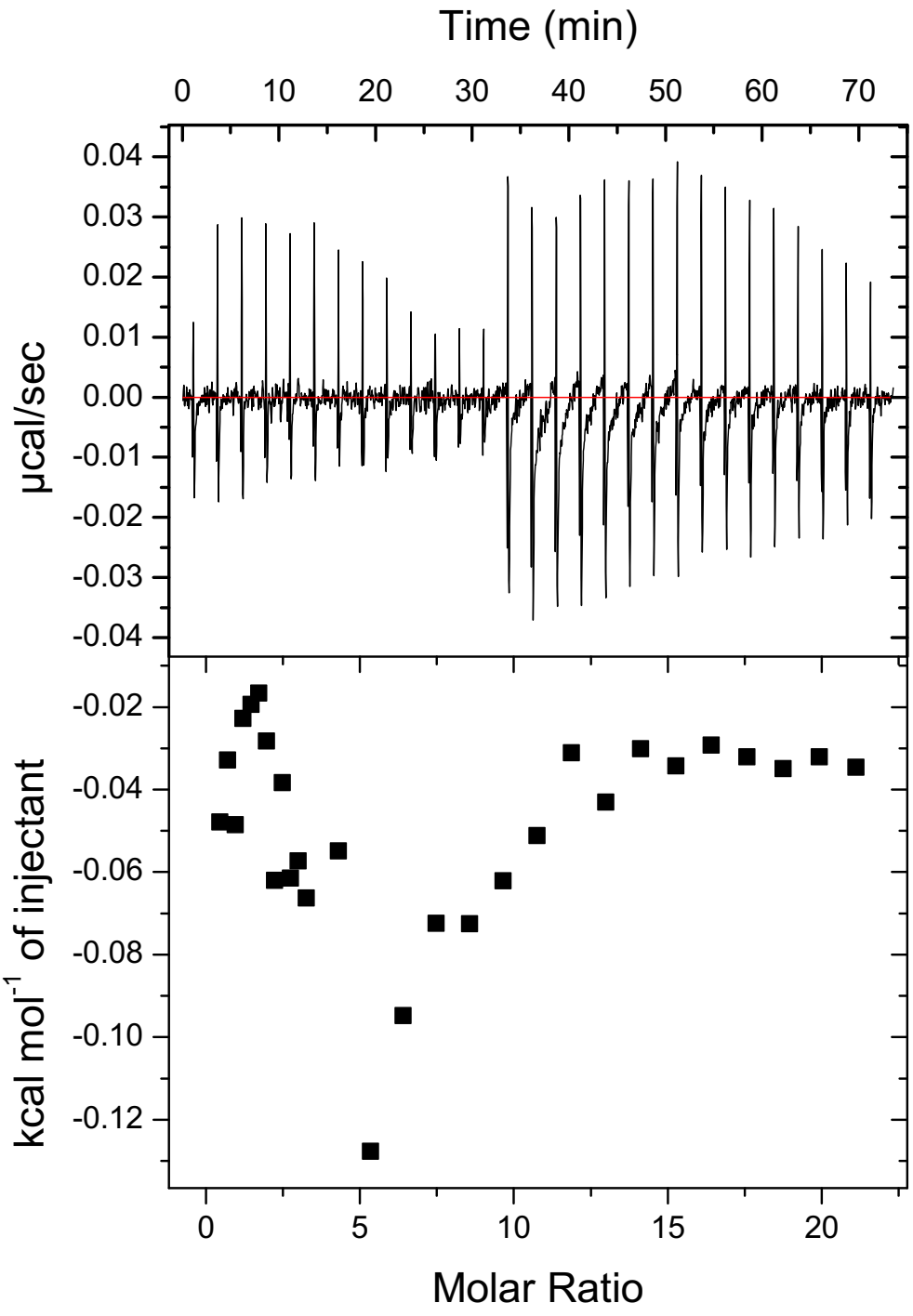

Figure S15: ITC isotherm for HEW5 titrated with $CaCl_2$.

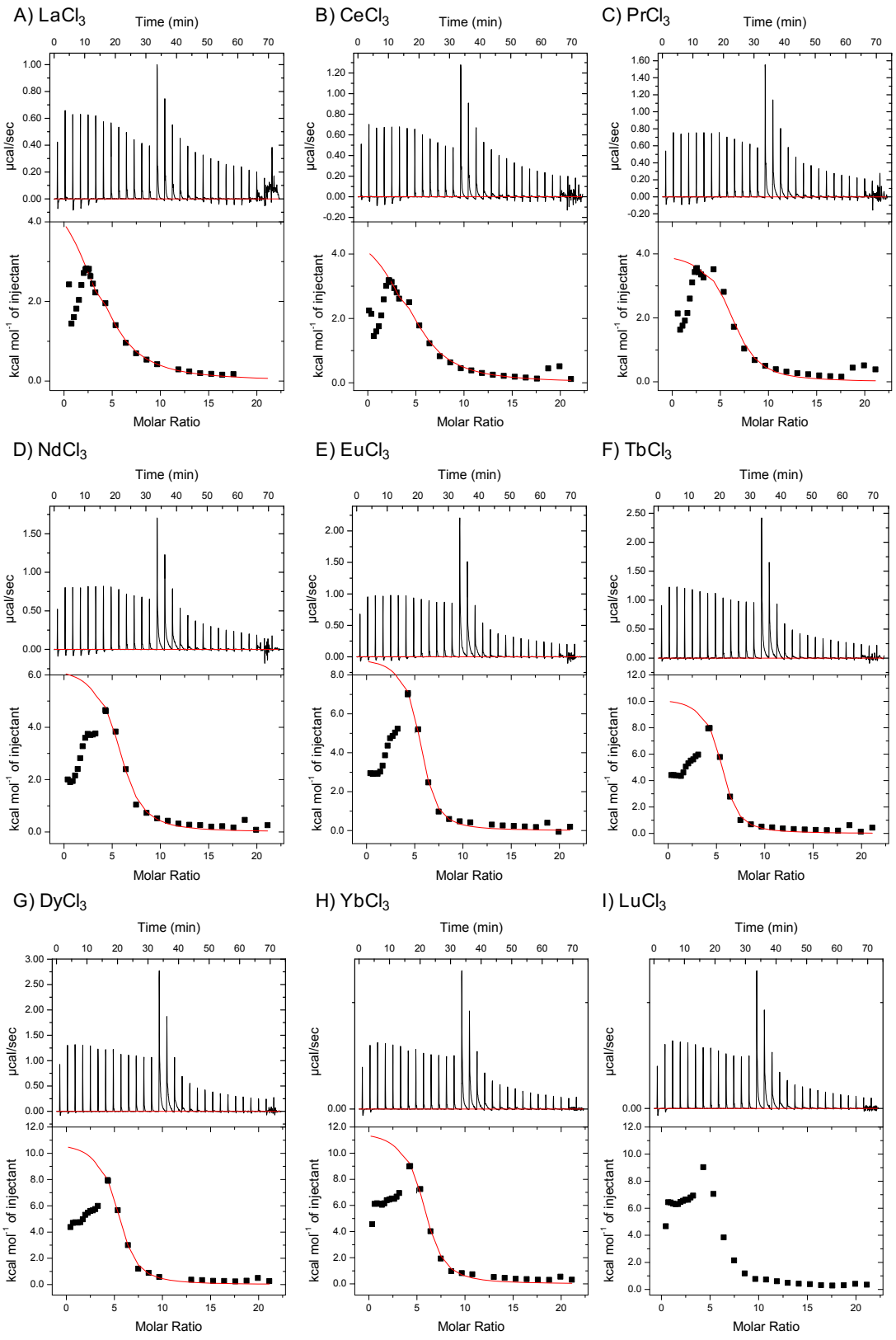

Figure S16: ITC isotherm for RTX titrated with A) $LaCl_3$, B) $CeCl_3$, C) $PrCl_3$, D) $NdCl_3$, E) $EuCl_3$, F) $TbCl_3$, G) $DyCl_3$, H) $YbCl_3$, and I) $LuCl_3$.

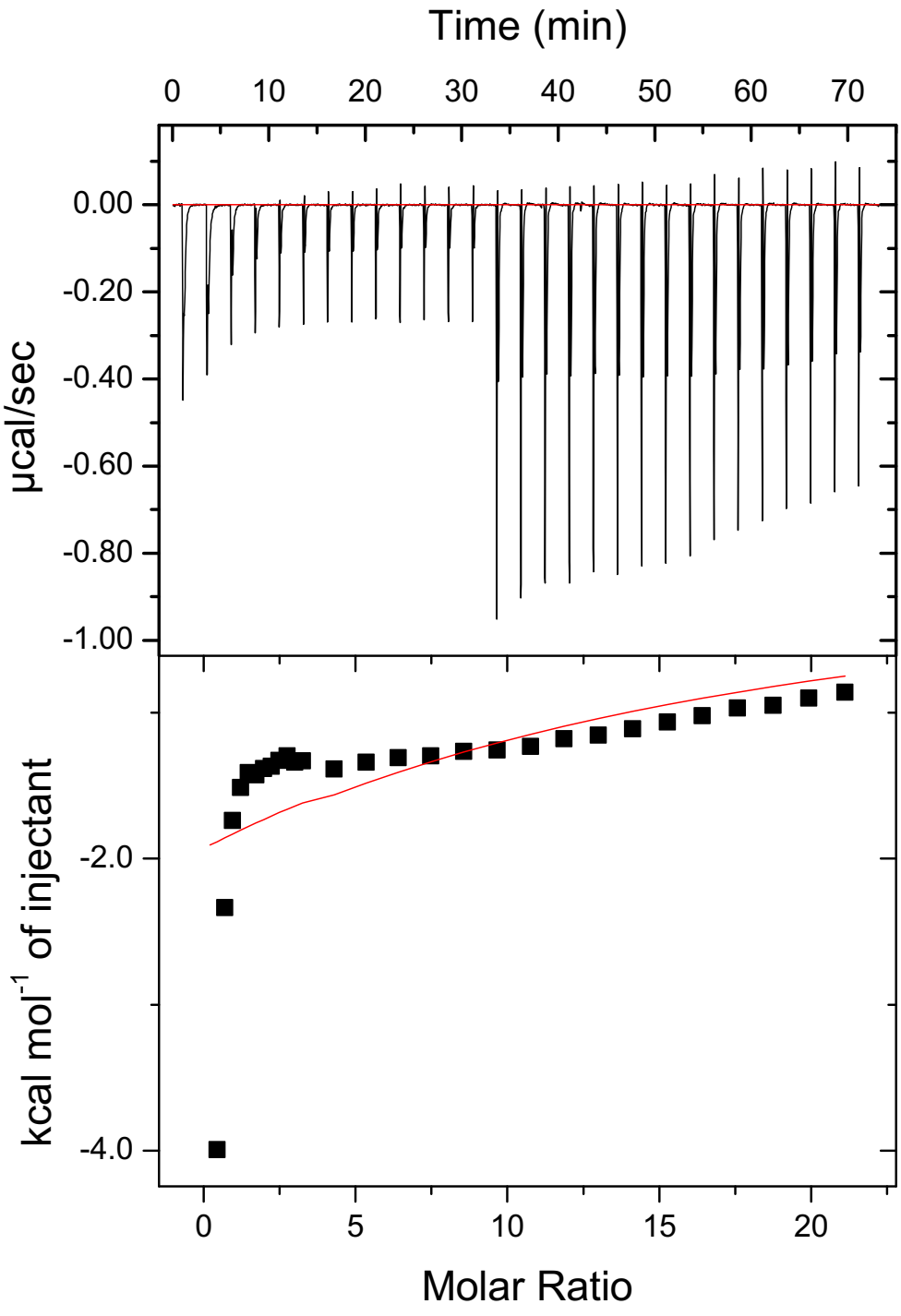

Figure S17: ITC isotherm for RTX titrated with $CaCl_2$.

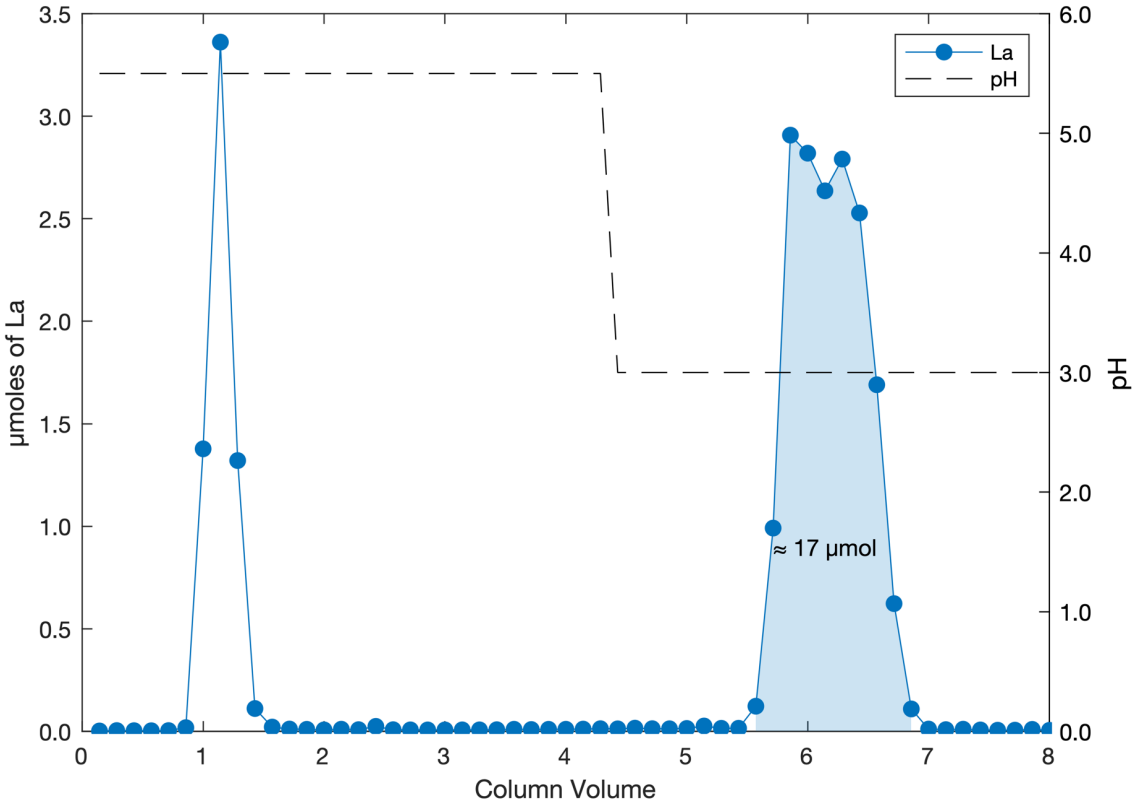

Figure S18: Binding capacity of HEW5 column determined via saturation to be nearly 17 µmoles of REE binding capacity. The shaded area represents the eluted fractions used in the calculation.

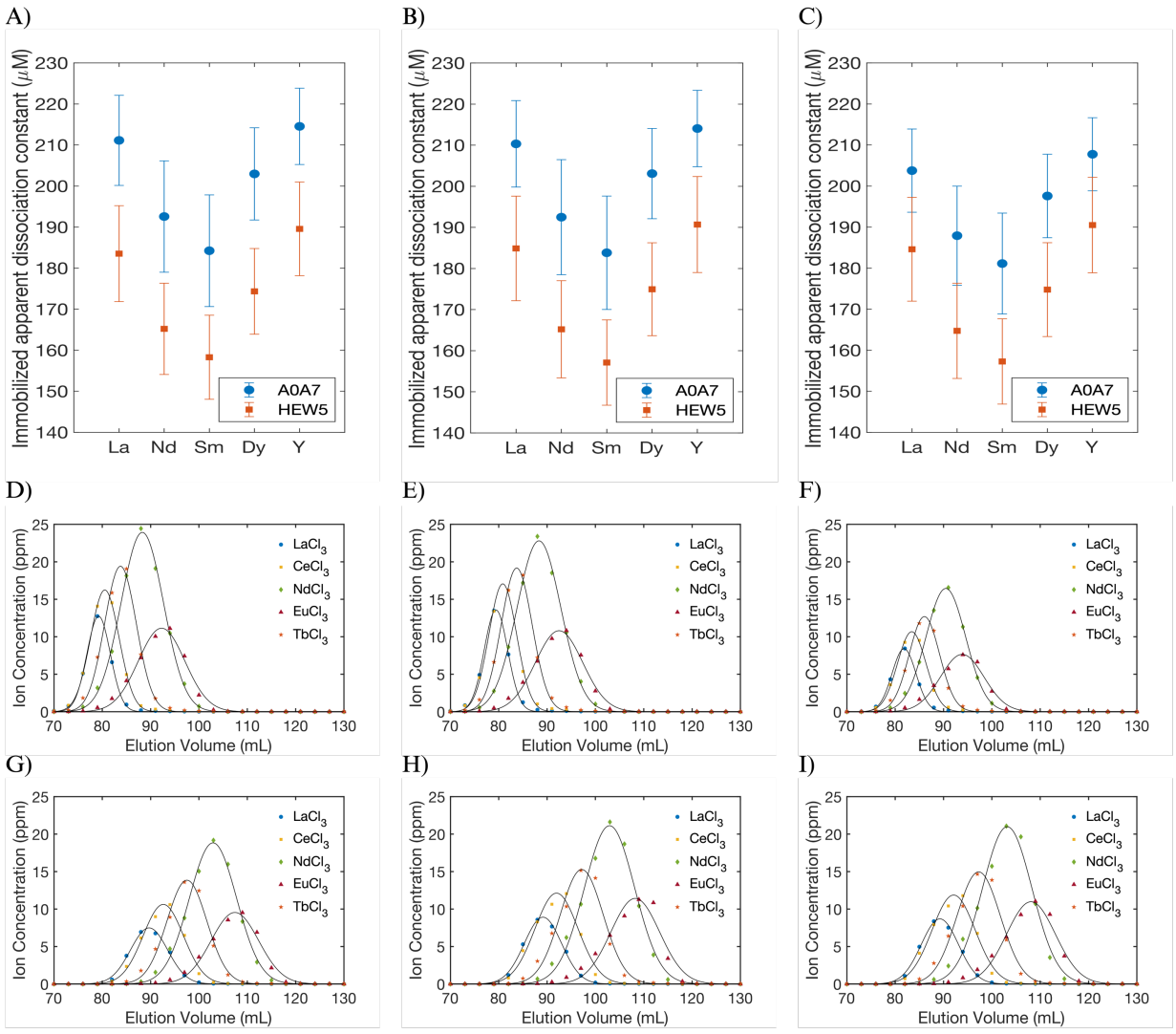

Figure S19: A-C) Three independent trials determining column-specific dissociation constants for a mixture of REEs under isocratic elution conditions (pH 3) for A0A7 and HEW5. The error bars represent the standard deviations. D-F) Elution profiles from A0A7 column for three trials. G-H) Elution profiles from HEW5 column for three trials.

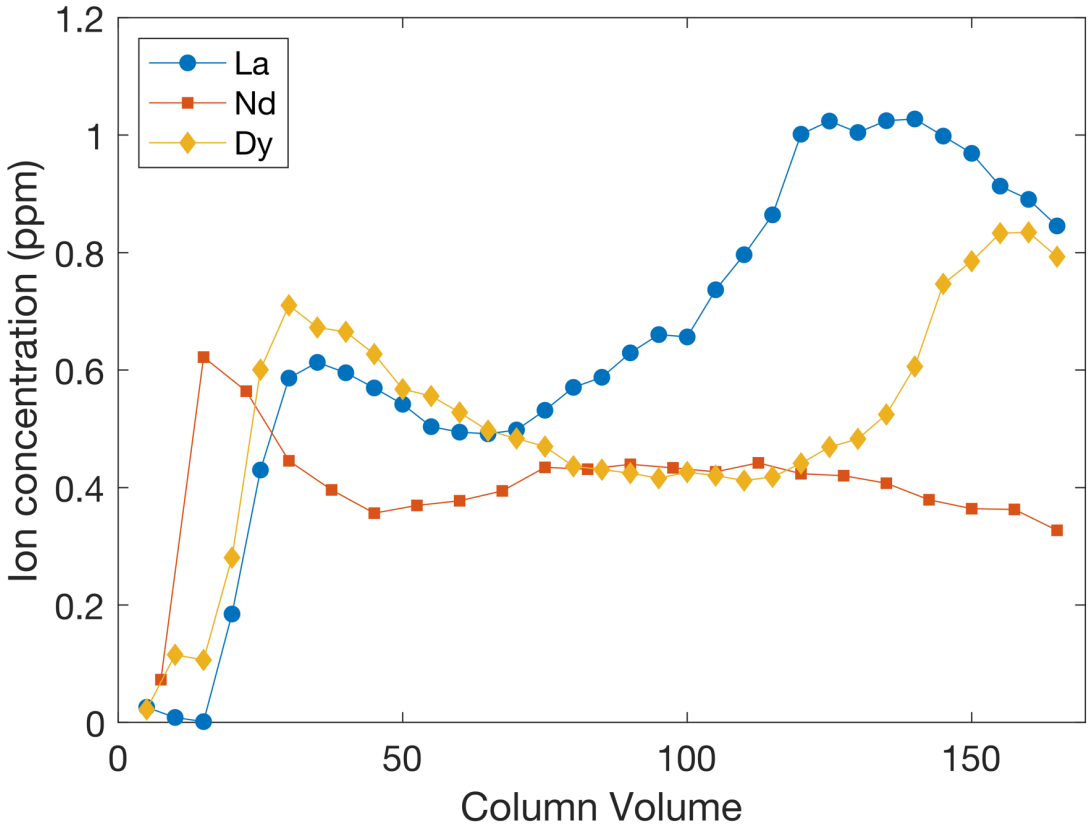

Figure S20: Elution profiles from Halo-RTX column for a mixture of REEs illustrating multipeak and broad elution profiles.

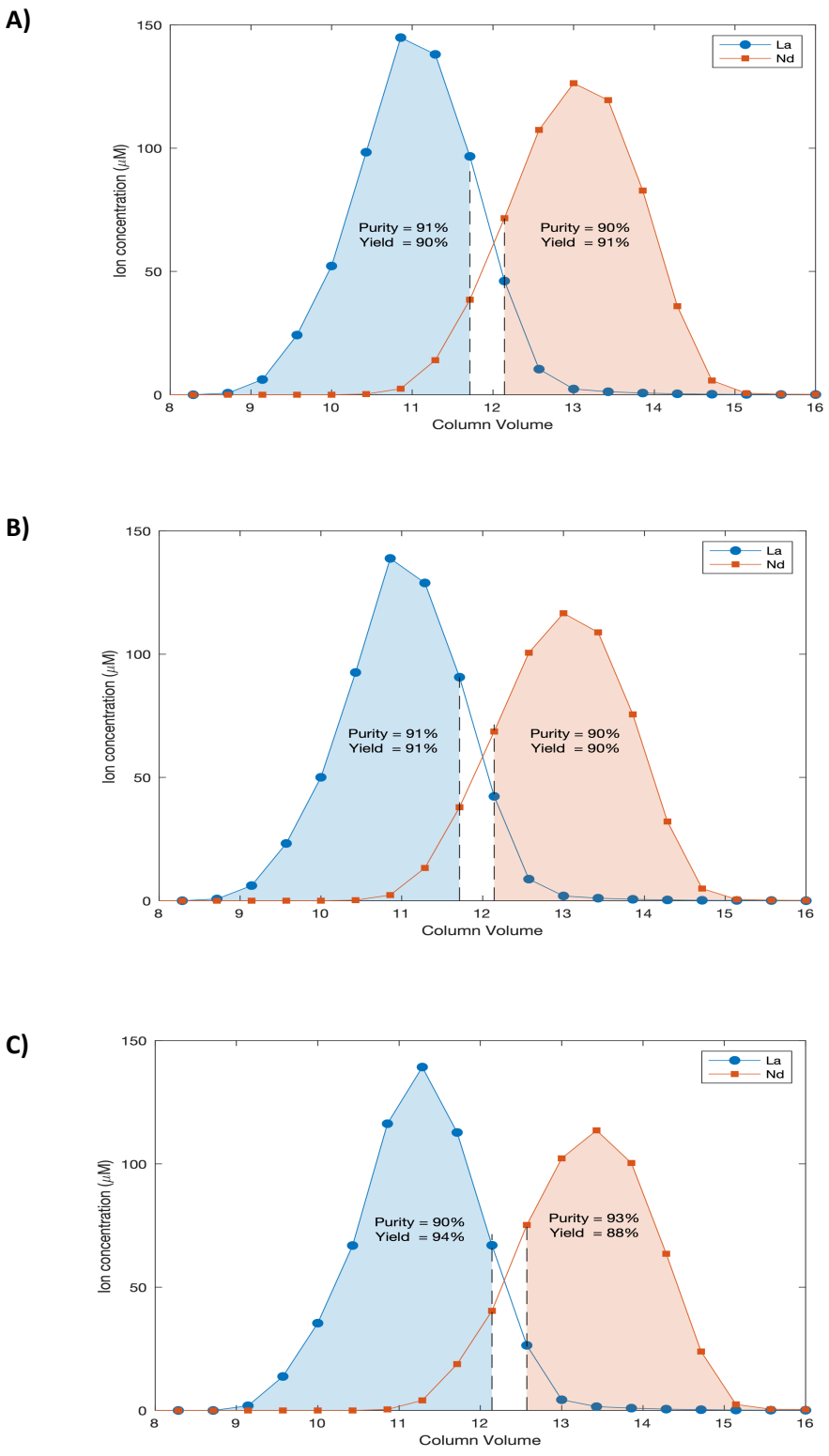

Figure S21: Single-stage separation of an equimolar mixture of lanthanum ($LaCl_3$) and neodymium ($NdCl_3$) using immobilized HEW5 (7-mL column volume). Panels A, B, and C represent three independent trials.

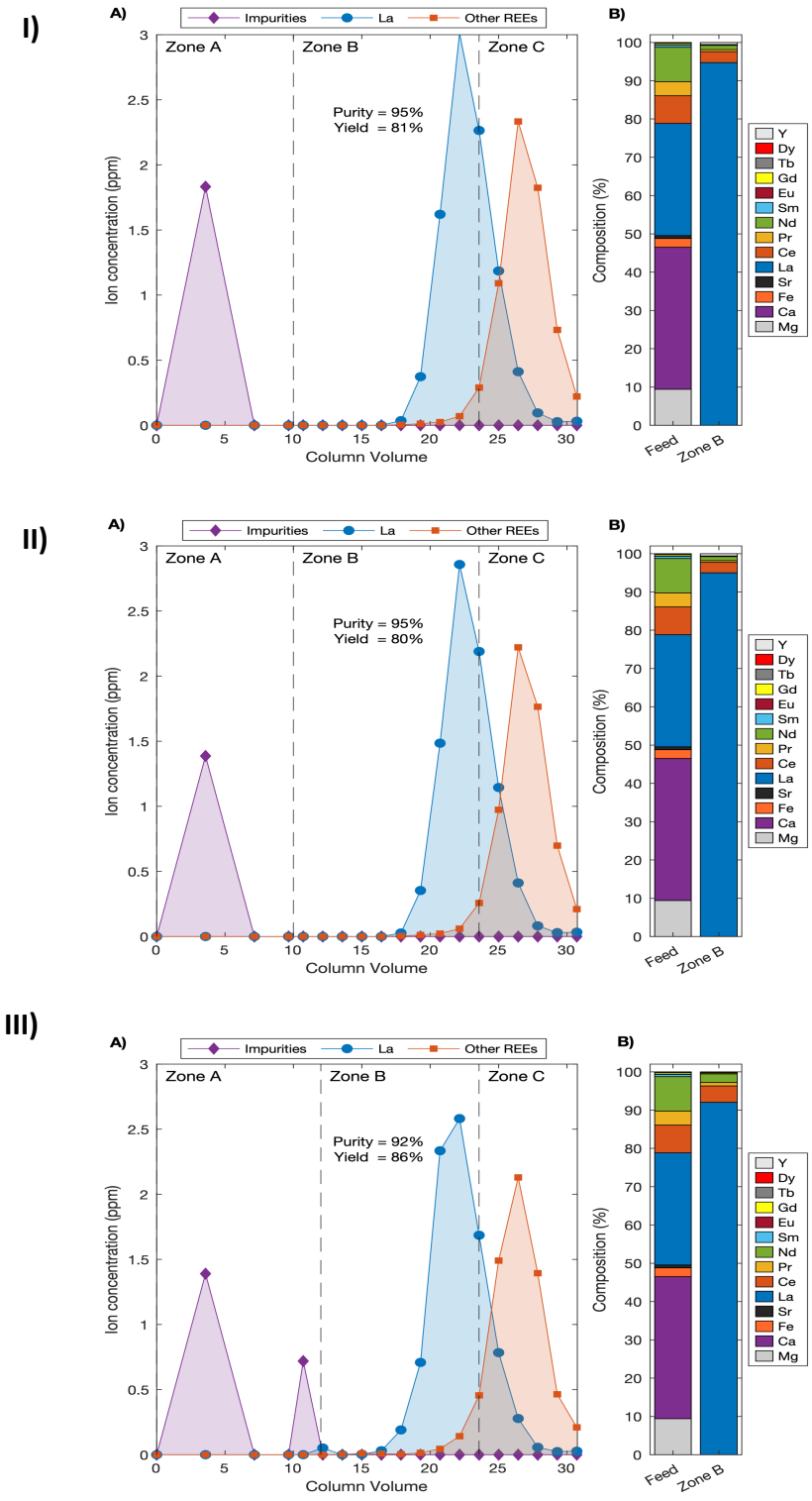

Figure S22: High yield and purity single stage-seperation of lanthanium from impurities and other REEs present in banasite leachate via immoblized HEW5. I-III) three indepdent seperation trials.

References

(1) Mende, D. R.; Letunic, I.; Huerta-Cepas, J.; Li, S. S.; Forslund, K.; Sunagawa, S.; Bork, P. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res* **2017**, *45* (D1), D529-D534. DOI: 10.1093/nar/gkw989.
(2) Mitchell, A.; Chang, H.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; et al. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **2015**, *43* (D1), D213-D221, Article. DOI: 10.1093/nar/gku1243.
(3) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **2011**, *7* (10), e1002195. DOI: 10.1371/journal.pcbi.1002195.
(4) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658-1659. DOI: 10.1093/bioinformatics/btl158.
(5) Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **2002**, *30* (14), 3059-3066. DOI: 10.1093/nar/gkf436.
(6) Capella-Gutiérrez, S.; Silla-Martínez, J. M.; Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25* (15), 1972-1973. DOI: 10.1093/bioinformatics/btp348.
(7) Szilvay, G. R.; Blenner, M. A.; Shur, O.; Cropek, D. M.; Banta, S. A FRET-based method for probing the conformational behavior of an intrinsically disordered repeat domain from Bordetella pertussis adenylate cyclase. *Biochemistry* **2009**, *48* (47), 11273-11282. DOI: 10.1021/bi901447j.
(8) Shur, O.; Banta, S. Rearranging and concatenating a native RTX domain to understand sequence modularity. *Protein Engineering Design & Selection* **2013**, *26* (3), 171-180, Article. DOI: 10.1093/protein/gzs092.
(9) Cotruvo, J.; Featherston, E.; Mattocks, J.; Ho, J.; Laremore, T. Lanmodulin: A Highly Selective Lanthanide-Binding Protein from a Lanthanide-Utilizing Bacterium. *Journal of the American Chemical Society* **2018**, *140* (44), 15056-15061, Article. DOI: 10.1021/jacs.8b09842.
(10) Featherston, E. R.; Mattocks, J. A.; Tirsch, J. L.; Cotruvo, J. A. Heterologous expression, purification, and characterization of proteins in the lanthanome. *Methods Enzymol* **2021**, *650*, 119-157. DOI: 10.1016/bs.mie.2021.02.004.
(11) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493-500. DOI: 10.1038/s41586-024-07487-w.
(12) Sehnal, D.; Bittrich, S.; Deshpande, M.; Svobodová, R.; Berka, K.; Bazgier, V.; Velankar, S.; Burley, S. K.; Koča, J.; Rose, A. S. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* **2021**, *49* (W1), W431-W437. DOI: 10.1093/nar/gkab314.
(13) Hogendoorn, C.; Roszczenko-Jasińska, P.; Martinez-Gomez, N. C.; de Graaff, J.; Grassl, P.; Pol, A.; Op den Camp, H. J. M.; Daumann, L. J. Facile Arsenazo III-Based Assay for Monitoring Rare Earth Element Depletion from Cultivation Media for Methanotrophic and Methylotrophic Bacteria. *Appl Environ Microbiol* **2018**, *84* (8). DOI: 10.1128/AEM.02887-17.
(14) Rigden, D. J.; Jedrzejas, M. J.; Galperin, M. Y. An extracellular calcium-binding domain in bacteria with a distant relationship to EF-hands. *FEMS Microbiol Lett* **2003**, *221* (1), 103-110. DOI: 10.1016/S0378-1097(03)00160-5.
(15) Nalefski, E. A.; Falke, J. J. The C2 domain calcium-binding motif: structural and functional diversity. *Protein Sci* **1996**, *5* (12), 2375-2390. DOI: 10.1002/pro.5560051201.
(16) Sutton, R. B.; Davletov, B. A.; Berghuis, A. M.; Südhof, T. C.; Sprang, S. R. Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold. *Cell* **1995**, *80* (6), 929-938. DOI: 10.1016/0092-8674(95)90296-1.

(17) Srivastava, S. S.; Mishra, A.; Krishnan, B.; Sharma, Y. Ca2+-binding motif of βγ-crystallins. *J Biol Chem* **2014**, *289* (16), 10958-10966. DOI: 10.1074/jbc.O113.539569.
(18) Suman, S. K.; Ravindra, D.; Sharma, Y.; Mishra, A. Association properties and unfolding of a βγ-crystallin domain of a Vibrio-specific protein. *PLoS One* **2013**, *8* (1), e53610. DOI: 10.1371/journal.pone.0053610.
(19) Dai, S.; Sun, C.; Tan, K.; Ye, S.; Zhang, R. Structure of thrombospondin type 3 repeats in bacterial outer membrane protein A reveals its intra-repeat disulfide bond-dependent calcium-binding capability. *Cell Calcium* **2017**, *66*, 78-89. DOI: 10.1016/j.ceca.2017.05.016.
(20) Aghajari, N.; Van Petegem, F.; Villeret, V.; Chessa, J. P.; Gerday, C.; Haser, R.; Van Beeumen, J. Crystal structures of a psychrophilic metalloprotease reveal new insights into catalysis by cold-adapted proteases. *Proteins* **2003**, *50* (4), 636-647. DOI: 10.1002/prot.10264.
(21) Chessa, J. P.; Petrescu, I.; Bentahir, M.; Van Beeumen, J.; Gerday, C. Purification, physico-chemical characterization and sequence of a heat labile alkaline metalloprotease isolated from a psychrophilic Pseudomonas species. *Biochim Biophys Acta* **2000**, *1479* (1-2), 265-274. DOI: 10.1016/s0167-4838(00)00018-2.
(22) Schürpf, T.; Chen, Q.; Liu, J. H.; Wang, R.; Springer, T. A.; Wang, J. H. The RGD finger of Del-1 is a unique structural feature critical for integrin binding. *FASEB J* **2012**, *26* (8), 3412-3420. DOI: 10.1096/fj.11-202036.
(23) Huang, J.; Yu, Z.; Groom, J.; Cheng, J. F.; Tarver, A.; Yoshikuni, Y.; Chistoserdova, L. Rare earth element alcohol dehydrogenases widely occur among globally distributed, numerically abundant and environmentally important microbes. *ISME J* **2019**, *13* (8), 2005-2017. DOI: 10.1038/s41396-019-0414-z.
(24) Keitel, T.; Diehl, A.; Knaute, T.; Stezowski, J. J.; Höhne, W.; Görisch, H. X-ray structure of the quinoprotein ethanol dehydrogenase from Pseudomonas aeruginosa: basis of substrate specificity. *J Mol Biol* **2000**, *297* (4), 961-974. DOI: 10.1006/jmbi.2000.3603.
(25) Keltjens, J. T.; Pol, A.; Reimann, J.; Op den Camp, H. J. PQQ-dependent methanol dehydrogenases: rare-earth elements make a difference. *Appl Microbiol Biotechnol* **2014**, *98* (14), 6163-6183. DOI: 10.1007/s00253-014-5766-8.
(26) Pol, A.; Barends, T. R.; Dietl, A.; Khadem, A. F.; Eygensteyn, J.; Jetten, M. S.; Op den Camp, H. J. Rare earth metals are essential for methanotrophic life in volcanic mudpots. *Environ Microbiol* **2014**, *16* (1), 255-264. DOI: 10.1111/1462-2920.12249.
(27) Wu, M. L.; Wessels, J. C.; Pol, A.; Op den Camp, H. J.; Jetten, M. S.; van Niftrik, L. XoxF-type methanol dehydrogenase from the anaerobic methanotroph "Candidatus Methylomirabilis oxyfera". *Appl Environ Microbiol* **2015**, *81* (4), 1442-1451. DOI: 10.1128/AEM.03292-14.
(28) Sun, X.; Li, Z.; Meng, D.; Huang, X.; Feng, Z.; Wang, M.; Yanyan, Z. Thermal decomposition behavior of Mountain Pass rare earth concentrate in air/CO2 atmosphere. *Journal of Rare Earths* **2024**. DOI: https://doi.org/10.1016/j.jre.2024.04.002.
(29) He, J.; Li, Y.; Xue, X.; Ru, H.; Huang, X.; Yang, H. Leaching of fluorine and rare earths from bastnaesite calcined with aluminum hydroxide and the recovery of fluorine as cryolite. *RSC Advances* **2017**, *7* (23), 14053-14059, 10.1039/C6RA28106K. DOI: 10.1039/C6RA28106K.