Supporting information for:

An Atypical Mechanism of Split Intein Molecular Recognition and Folding

Adam J. Stevens[1#], Giridhar Sekar[2#], Josef. A. Gramespacher[1], David Cowburn[2*], and Tom W. Muir[1*]

[#]These authors contributed equally to this work

[1]Department of Chemistry, Princeton University, Frick Laboratory, Princeton, New Jersey 08544, United States

[2]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, United States

*To whom correspondence should be addressed: cowburn@cowburnlab.org; muir@princeton.edu

Supporting Information Table of Contents **Page #**

Materials

Oligonucleotides and synthetic genes were purchased from Integrated DNA Technologies (Coralville, IA). Pfu Ultra II Hotsart fusion polymerase for cloning was purchased from Agilent (La Jolla, CA). All restriction enzymes and 2x Gibson Assembly Master Mix were purchased from New England Biolabs (Ipswich, MA). High-competency cells used for cloning and protein expression were generated from One Shot Bl21 (DE3) chemically competent *E. coli* and sub-cloning efficiency DH5α competent cells purchased from Invitrogen (Carlsbad, CA). DNA purification kits were purchased from Qiagen (Valencia, CA). All plasmids were sequenced by GENEWIZ (South Plainfield, NJ). Luria Bertani (LB) media, and all buffering salts were purchased from Fisher Scientific (Pittsburgh, PA). Dimethylformamide (DMF), dichloromethane (DCM), Coomassie brilliant blue, triisopropylsilane (TIS), β-mercaptoethanol (BME), DL-dithiothreitol (DTT), sodium 2-mercaptoethanesulfonate (MESNa), 5(6)-carboxyfluorescein, and thermolysin were purchased from Sigma-Aldrich (Milwaukee, WI). Tris (2-carboxyethyl) phosphine hydrochloride (TCEP) and isopropyl-β-D-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). Roche Complete Protease Inhibitors were used for protein purification (Roche, Branchburg, NJ). Nickel-nitrilotriacetic acid (Ni-NTA) resin was purchased from Thermo scientific (Rockford, IL). Fmoc amino acids were purchased from Novabiochem (Darmstadt, Germany) or Bachem (Torrance, CA). *O*-(Benzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate (HBTU) was purchased from Genscript (Piscataway, NJ). Trifluoroacetic acid (TFA) was purchased from Halocarbon (North

Augusta, SC). MES-SDS running buffer was purchased from Boston Bioproducts (Ashland, MA).

Equipment

Analytical reverse phase high performance liquid chromatography (RP-HPLC) was carried out on Hewlett-Packard 1100 and 1200 series instruments equipped with a C18 Vydac column (5 μm, 4.6 x 150 mm). All HPLC runs used the following solvents at a flow rate of 1 mL/min: 0.1 % TFA (trifluoroacetic acid) in water (solvent A) and 90 % acetonitrile in water with 0.1 % TFA (solvent B). All peptides and proteins were analyzed using the gradient: 0% B for 2 min followed by 0-73% B for 30 min. Electrospray ionization mass spectrometric analysis (ESI-MS) was carried out on a Bruker Daltonics MicroTOF-Q II mass spectrometer. Size-exclusion chromatography (SEC) was performed on an AKTA FPLC system (GE Healthcare) with a Superdex S75 16/60 column (125 mL column volume) for preparative runs and a Superdex S75 10/300 column for analytical runs. Gels were imaged with a LI-COR Odyssey Infrared Imager. Circular dichroism experiments were carried out on a Chirascan Circular Dichroism spectrometer (Applied Photophysics). Cell lysis was carried out using a S-450D Branson Digital Sonifier. NMR experiments were carried out on a Bruker 900, 800, 600 and 500 MHz spectrometers with 5 mm TCI triple resonance cryoprobes. Steady state fluorescence measurements were performed on a Horiba Flourmax 4 fluorimeter. Stopped flow anisotropy measurements were performed on an Applied Photophysics SX20 stopped-flow spectrometer.

Consensus Protein Design

Homologues of AceL TerL were identified through a BLAST[1] search of metagenomic data in the NCBI[2] (nucleotide collection) and JGI[3] databases using the TerL DNA sequence. This led to the identification of TerL N- and C-inteins with high sequence identity to AceL (Table S1). Because we could not match the cognate N- and C- inteins, the split inteins were treated as two distinct datasets and analyzed separately. MSAs of these split inteins were then generated in Jalview[4], and the consensus sequence was determined. At some positions in the N-intein, additional residues from the alignment corresponding to loops not present in AceL were included in the consensus sequence.

Cloning of Recombinant DNA

Synthetic genes were purchased and introduced into pET-30 expression vectors using Gibson assembly. Targeted mutations were introduced using inverse PCR with Pfu Ultra II HF Polymerase. The identity of all recombinant plasmids was confirmed through sequencing and the corresponding protein sequences are reported in Table S9.

Expression and Purification of Inteins for Splicing Assay

Expression and purification of the inteins was carried out as previously described.[5-7] The expressed N-intein constructs contained the following architecture: His_6-SUMO-MBP-EFE-Int^N, where "His_6" is a 6x polyhistidine affinity tag, "SUMO" is the ubiquitin-like protein SMT3, "MBP" is maltose binding protein, "EFE" is the wild type -1, -2, and -3 N-extein sequence of TerL inteins, and Int^N is the N-intein. The expressed C-intein constructs contained the following architecture: His_6-SUMO-Int^C-CEFL-GFP, where "Int^C" is the C-intein, "CEFL" is the +1, +2, +3, and

+4 C-extein residues of TerL inteins, and "GFP" is green fluorescent protein. For the screen of extein dependence, constructs corresponding to each indicated point mutation in the "EFE" or "CEFL" extein sequences were utilized.

E. *coli* BL21(DE3) cells were transformed with an MBP-Int^{N} or Int^{C}-GFP intein plasmid and grown at 37 ˚C in 1 L of LB containing 50 μg/mL of kanamycin. Once the culture reached an OD_{600}=0.6, 0.5 mM IPTG was added to induce expression (0.5 mM final concentration, 18 h at 18 ˚C). For test expression of the SUMO-Cat^{C} constructs, expression tests were also carried out at 37 ˚C for 3 hours upon addition of IPTG. Following expression, the cells were pelleted via centrifugation (5,000 rcf, 30 min) and stored at -80 ˚C.

The cell pellet was then resuspended in 30 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, pH 8.0) containing a protease inhibitor cocktail. The cells were lysed by sonication (35% amplitude, 8 x 20 s pulses on / 30 s off) and then pelleted by centrifugation (35,000 rcf, 30 min). The supernatant was incubated with 4 mL of Ni-NTA resin for 30 min at 4 ˚C to bind the His-tagged inteins. The slurry was then loaded onto a fritted column, the flow through was collected, and the column was washed with 20 mL of lysis buffer. The protein was then eluted from the column with 20 mL of elution buffer (lysis buffer + 250 mM imidazole).

The eluted protein was dialyzed into lysis buffer while being treated with 10 mM TCEP and Ulp1 protease overnight at 4 ˚C to cleave the His_6-SUMO expression tag. The dialyzed protein was then incubated with 4 mL Ni-NTA resin for 30 min at 4 ˚C, after which it was applied to a fritted column with the flow through collected

together with a 10 mL wash of lysis buffer. The protein was then treated with 10 mM TCEP, concentrated to 2 mL, and purified over an S75 16/60 gel filtration column using degassed splicing buffer (100 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, pH 7.2) as the mobile phase. Fractions were analyzed by analytical RP-HPLC and ESI-MS (Figure S6, table S8), and either immediately utilized in the splicing assay or stored long term in glycerol (20% v/v) after being flash-frozen in liquid N_2.

Splicing Assays

Splicing assays were carried out as adapted from a previously described protocol.[8] Briefly, N- and C-inteins (4 μM Int^N, 4μM Int^C) were individually preincubated in splicing buffer (100 mM sodium phosphates, 150 mM NaCl, 1 mM EDTA, pH 7.2) with 2 mM TCEP for 15 min. Splicing reactions were carried out at indicated temperatures and concentrations of urea. For the extein characterization, the Cat^C-GFP and MBP-Cat^N proteins containing the indicating extein mutations were spliced with their cognate wild type N- or C- intein at 30 ˚C. Splicing of Cat and AceL* in the presence of urea was carried out at 30 ˚C. Splicing was initiated by mixing equal volumes of N- and C- inteins with aliquots removed at the indicated times and quenched by the 1:1 addition of 4X loading dye (160 mM Tris, 40% glycerol, 4% SDS, 0.08% Bromophenol Blue, 8 % BME). Samples were analyzed by SDS-PAGE gel electrophoresis (12 % bis-tris, 60 min, 150 v) and quantified by densitometry (Figures S7, S8).

Kinetic analysis of *trans*-splicing reactions

To determine the splicing rates of *trans*-splicing reactions, the data was fit to the first order rate equation using GraphPad Prism software.

$$[P](t) = [P]_{max} \cdot (1 - e^{-kt})$$

Where $[P]$ is the normalized intensity of product, $[P]_{max}$ is the reaction plateau, and k is the rate constant (s^{-1}). The mean and standard error for each value are reported (n = 3).

Expression of Inteins for Structural Studies

Construct optimization was required in order to isolate Cat^C with minimal extein sequence for structural characterization. Compared to $AceL^{*C}$ and GOS^C, SUMO-Cat^C had increased yields during recombinant expression in *E. coli* (18 ˚C, 16 h or 37 ˚C for 3 h) (Figure S1). However, removal of the SUMO expression tag resulted in Cat^C aggregating upon cleavage (possibly due its neutral charge at physiological pH, pI = 7.2). Charged residues were therefore appended immediately flanking Cat^C to improve the solubility of the protein in solution, specifically an N-terminal FLAG epitope tag and "CESRGK" C-extein sequence (SUMO-Flag-Cat^C). The Cat^N construct utilized in these structural studies was expressed as a SUMO fusion (SUMO-Cat^N) and contains the minimal "EFE" N-extein following SUMO cleavage. In addition, inactivating C1A and N134A mutations were included in the constructs to prevent splicing during structural analysis of the associated complex. Expression

and purification of these Cat^N and Cat^C constructs for structural study were carried out as described above for the proteins utilized for splicing.

For use in NMR spectroscopy, expression of the isotopically enriched Cat proteins was carried out as previously described.[5,7] The intein plasmids were used to transform BL-21 (DE3) cells, and the cells were grown overnight in 5 mL LB starter cultures (37 ˚C, 18 h). The starter cultures were then spun down (4,000 rcf, 5 min). The supernatant was discarded, and the cells were then resuspended and grown in 1L of M9 medium supplemented with ^{13}C-glucose and $^{15}NH_4Cl$ as the sole carbon and nitrogen sources (50 μg/mL kanamycin, 37 ˚C). Once the cells reached OD_{600} = 0.6, expression was induced with the addition of IPTG (0.5 mM, 18 h, 18 ˚C). Following expression, the cells were spun down by centrifugation (5,000 rcf, 30 min) and stored at -80 ˚C. Purification was carried out with the general method described above for intein constructs. The masses of the purified proteins correspond to an isotopic labeling efficiency of 99% for both the Cat^N and Cat^C proteins.

NMR Spectroscopy

NMR experiments were performed using Cat^N and Cat^C in free form and in complex. NMR samples were prepared by buffer exchanging purified protein to 20 mM sodium phosphate 150 mM NaCl, 2 mM TCEP (pH 6.8, 37 °C). The uniformly labeled ^{15}N, ^{13}C, ^{1}H proteins were concentrated to final concentrations of ~300-600 μM respectively. For the HSQC experiments of the complex reported in figures 3a, 3b, the isotopically labeled intein fragments were mixed with the complementary

unlabeled intein solution in a ratio of 1:1.5 and concentrated to a final concentration similar to the free protein and measured directly. For structure determination isotopically labeled intein fragments were mixed at a Cat^N:Cat^C ratio of 1.5:1. The complex was further purified by size exclusion chromatography to remove the free forms.

Experiments were performed at field strengths of 600, 700, 800 or 900 MHz and Non-Uniform Sampling (NUS) acquisition was employed as appropriate. NMR spectra were processed using Bruker Topspin 3.0 or NMR Pipe software and NUS spectra were reconstructed by compressed sensing using qMDD.[9,10].

Chemical shift assignment

Backbone chemical shifts were assigned using HNCO, HN(CA)CO, HNCACB, CBCA(CO)NH triple resonance experiments. Side chain assignments were obtained from H(CC)(CO)NH, (H)CC(CO)NH, H(C)CH-TOCY and (H)CCH-TOCSY experiments. Aromatic assignments were obtained from CT-^{13}C-resolved [^{1}H,^{1}H]-NOESY (mixing time = 100 ms), (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE experiments. CcpNmr Analysis software was used for manual chemical shift assignment and other data analysis.[11] Chemical shift values have been validated and deposited to the Biological Magnetic Resonance Bank (BMRB No : 30480). Random coil chemical shifts were calculated using CcpNmr analysis.[12]

Spin relaxation measurements

Spin-spin relaxation (R_2) rates of ^{15}N spins (mixing times of 0, 17, 34, 51, 85, 119, 170, 255, 340, 510, 680 ms) and [^{15}N-^{1}H] NOE experiments were measured at a field strength of 600 MHz.

Structure determination

Dihedral angle restrains were calculated from chemical shifts using TALOS software.[13] NOE cross peaks were picked from ^{15}N-resolved [^{1}H,^{1}H]-NOESY (mixing time = 80 ms), ^{13}C-resolved-[^{1}H,^{1}H]-NOESY (mixing time = 80 ms), CT-^{13}C-resolved aromatic [^{1}H,^{1}H]-NOESY experiments (mixing time = 100 ms) and assigned automatically using ARIA and CNS softwares.[14,15] Assignment and structure calculation was done in 8 cycles, calculating 20 structures in each step. The assigned NOEs were verified manually and violation analysis was done. The verified NOE peak lists were used to generate distance restraints. 3,283 unambiguous restraints, 206 ambiguous restraints and 180 dihedral angle restraints were used to finally calculate 256 structures. 20 least energy structures were selected and water refinement was performed. Structures have been validated and deposited to the Protein Data Bank (PDB ID : 6DSL).

Circular Dichroism (CD)

Cat^N, Cat^C, and 1:1 complex of Cat^N and Cat^C were dialyzed into CD buffer (25 mM sodium phosphate, 50 mM NaF, 1 mM DTT, pH 7.2). CD spectra were measured at 25 °C in a 1 mm pathlength cuvette (10 μM sample concentration).

Analytical Size Exclusion Chromatography (SEC)

Analytical SEC experiments were run on an S75 10/300 column at 4 °C in splicing buffer (25 mM sodium phosphate, 150 mM NaCl, 1 mM DTT, pH 7.2. For all runs, UV absorbance was monitored at 214 nm. Samples were injected with a sample volume of 500 µL (25 µM) and eluted with a flow rate of 0.5 mL/min.

Limited Proteolysis

EFE-Cat^N, Flag-Cat^C, and 1:1 complex of EFE-Cat^N and Flag-Cat^C were dialyzed into thermolysin buffer (50 mM Tris HCl, 100 mM NaCl, 2 mM MgSO4, 2 mM CaCl2, 1 mM DTT, pH 7.4) and diluted to a concentration of 10 µM. Thermolysin powder (Sigma) dissolved to 0.4 mg/mL in thermolysin buffer was then prepared and added to each solution (1:50 v/v). At the indicated times, aliquots were removed and quenched with the 1:3 addition of 8 M Guanidine HCL 4% TFA. The samples were then analyzed by RP-HPLC and ESI-MS. Masses from each peak were compared to predicted cleavage products of the inteins from ProteinProspector (UCSF).

Production of Inteins for Binding Experiments

The fluorescein labeled Cat^N (Fl-Cat^N) peptide was synthesized by standard 9-fluorenylmethyl-oxycarbonyl (Fmoc) solid phase peptide synthesis (SPPS). After coupling the last amino acid in the peptide, the N-terminus was capped with 5(6)-carboxyfluorescein. The synthesized Fl-Cat^N peptide was purified by preparative RP-HPLC and characterized by analytical RP-HPLC and ESI-MS. The C-intein expressed for the binding experiments was SUMO-Flag-Cat^C construct detailed above. Instead of carrying out an Ulp1 digestion, the expressed SUMO-Flag-Cat^C

protein was purified directly over the S75 16/60 gel filtration column following Ni-NTA enrichment.

Steady State Fluorescence Anisotropy

Equilibrium measurements were performed using 500 pM Fl- Cat^N with given concentrations of SUMO-Flag-Cat^C (0 pM - 2,500 pM) in low salt (50 mM sodium phosphate, 100 mM NaCl, 1mM DTT, 1mM EDTA, pH 7.0) and high salt (50 mM sodium phosphate, 500 mM NaCl, 1mM DTT, 1mM EDTA, pH 7.0) buffers. Proteins were diluted from stock solutions to desired concentrations and incubated at 25 °C for 30 min. Samples were transferred to a cuvette of 1 cm path-length and the fluorescence anisotropy was measured immediately. Constants in the one site binding equation were obtained using non-linear least squares curve fitting method in MATLAB. For both the high and low salt conditions, the constants obtained from these fits (Table S6) fall below the concentration of Cat^N used for the measurements. We therefore report the K_d as < 500 pM, as we were unable to measure fluorescence anisotropy at lower concentrations of Cat^N.

Stopped flow fluorescence anisotropy

The stopped flow syringes were loaded with Fl-Cat^N and SUMO-Flag-Cat^C protein solutions so as to obtain final concentrations of 100 nM Cat^N and reported concentrations of Cat^C (200, 325, 500, 750, 1000 nM). Change in anisotropy values were measured in low salt and high salt buffers for a duration of 50 s. The change in anisotropy over time was fit to a double exponential kinetic model previously reported using non-linear least squares curve fitting method in MATLAB to obtain

kinetic constants of binding (k_{obs1} and k_{obs2}) for each concentration.[16] The k_{obs1} and k_{obs2} values were then plotted as a function of Cat^{C} concentration, fit to a line, and the slope of the line was interpreted as k_{on}.

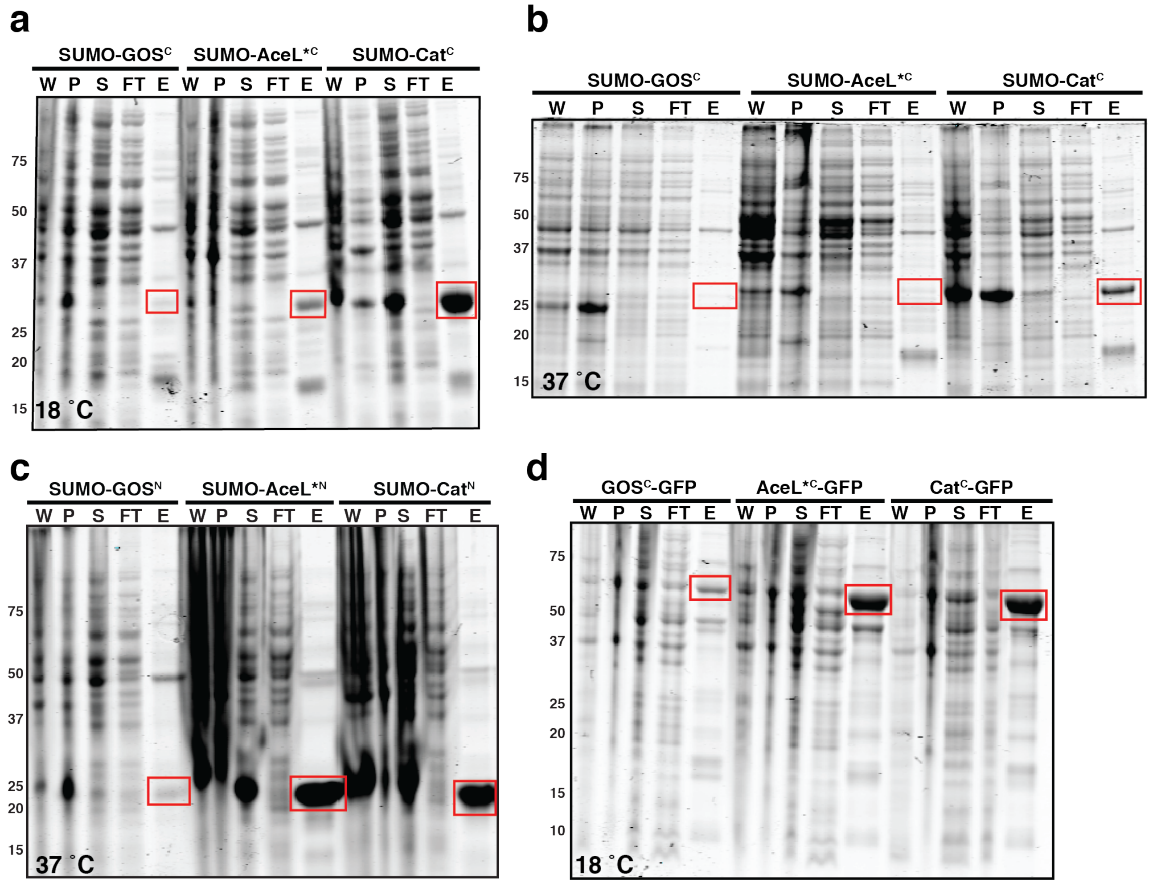


Figure S1: Expression of Atypical Split Inteins. Lanes correspond to (W) the whole cell lysate, (P) the inclusion body pellet, (S) the soluble fraction of the lysate, (FT) flow through of the soluble lysate batch bound to Ni-NTA affinity beads, (E) a 3 CV elution of 250 mM imidazole. (a) Purification of SUMO-GOS^{C}, SUMO-AceL*^{C}, and SUMO-Cat^{C} from *E. coli* expression (18 °C, 16 h). (b) Purification of SUMO-GOS^{C}, SUMO-AceL*^{C}-Sumo, and SUMO-Cat^{C} from *E. coli* expression (37 °C, 3 hours). (c) Purification of SUMO-GOS^{N}, SUMO-AceL*^{N}-Sumo, and SUMO-Cat^{N} from *E. coli* expression (37 °C, 3 hours). (d) Purification of GOS^{C}-GFP, AceL*^{C}-GFP, and Cat^{C}-GFP from *E. coli* expression (18 °C, 16 hours).

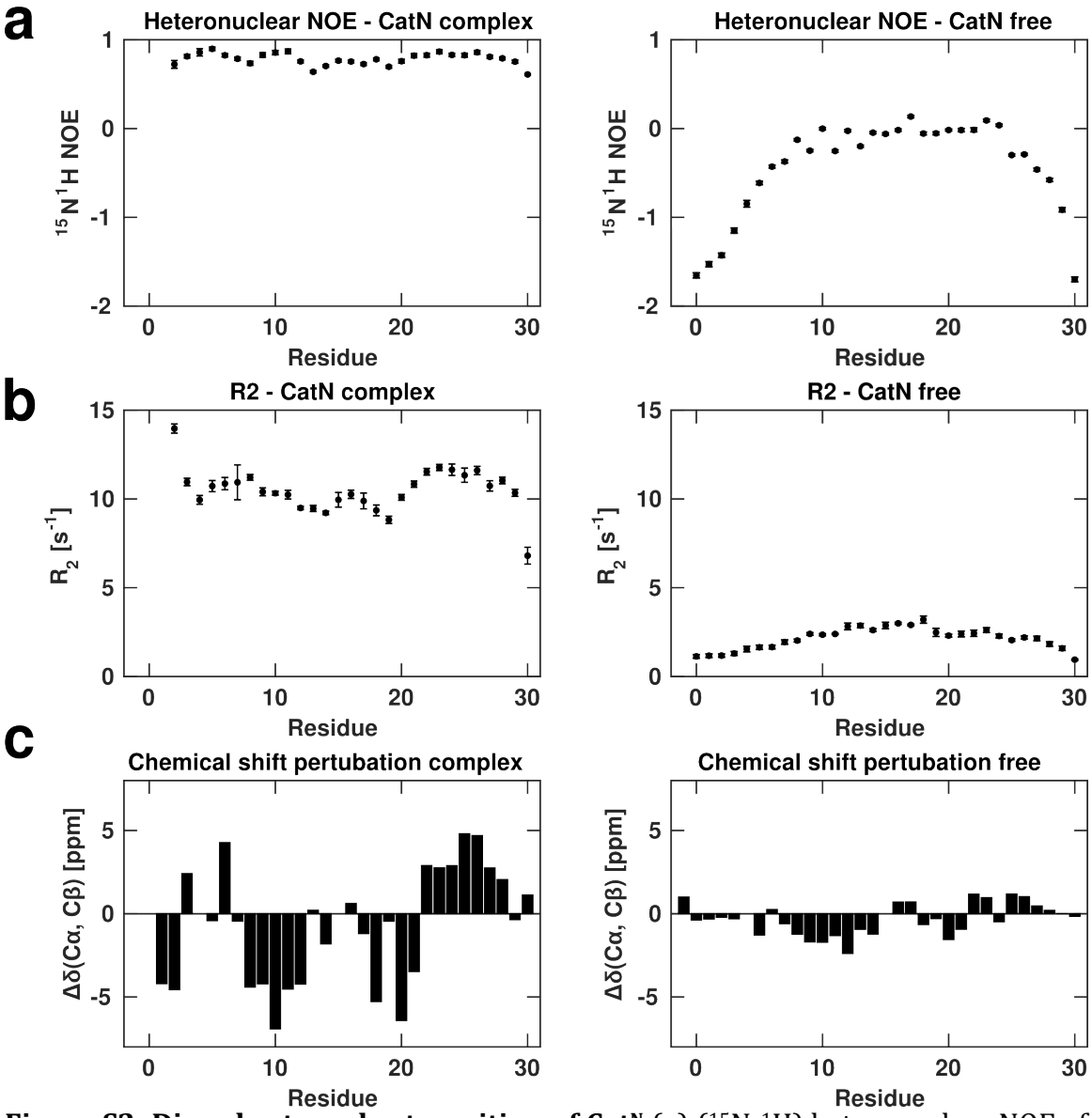


Figure S2: Disorder to order transition of Cat^{N} (a) (^{15}N-^{1}H) heteronuclear NOE of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). (b) Spin-spin relaxation rate of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). (c) Perturbation of Cα and Cβ chemical shifts of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). $\Delta\delta(C\alpha,C\beta) = (\delta C\beta - \delta C\alpha)_{Observed} - (\delta C\beta - \delta C\alpha)_{Random\ Coil}$.

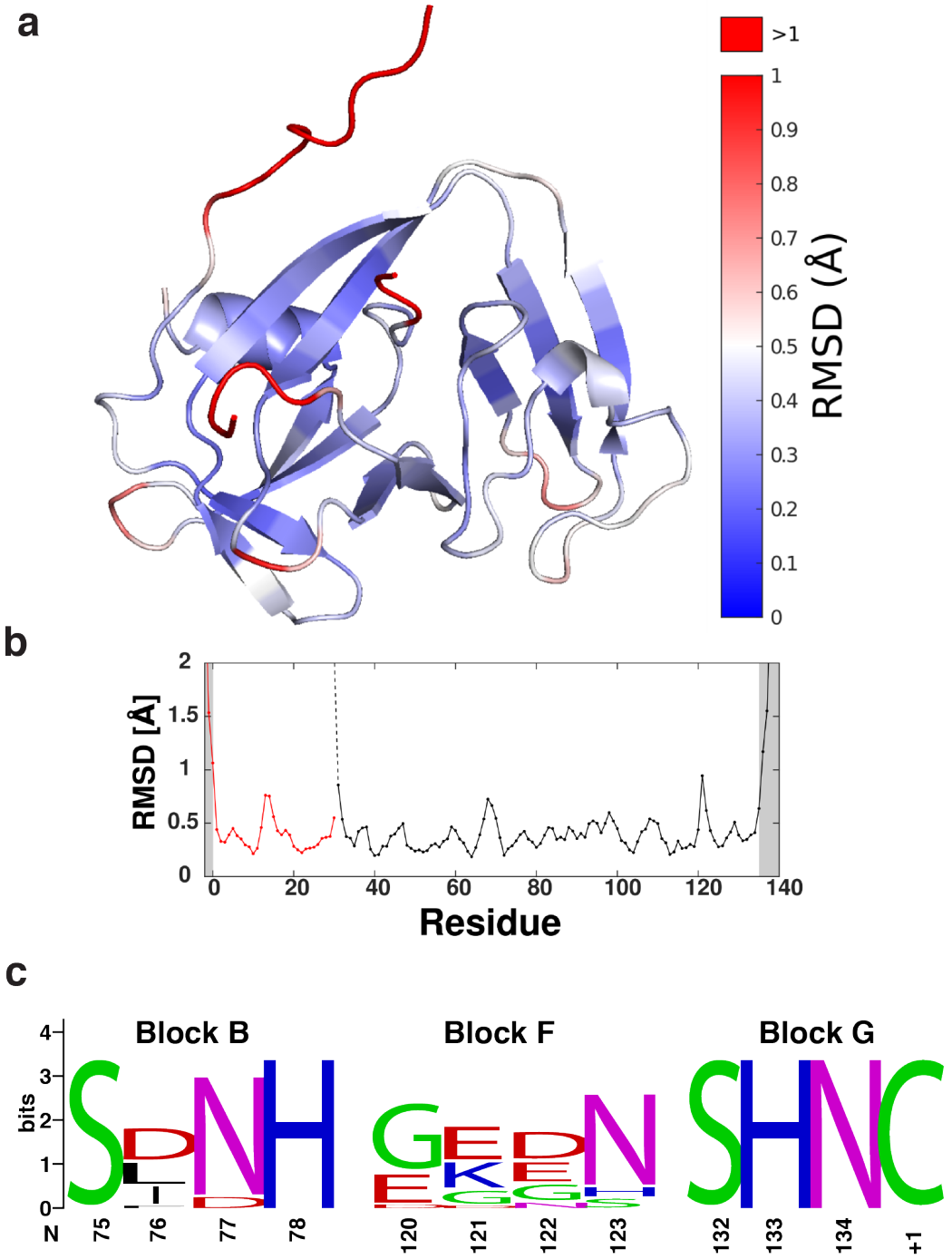


Figure S3: Structure of Cat Complex (a) Average per residue Root Mean Square Deviation (RMSD) from average structure for 20 least energy conformers of Cat^N-Cat^C complex obtained in NMR structure calculation. (b) Average per residue RMSD plotted against residue number for Cat^N (red) - Cat^C (black) complex. Extein regions are colored grey and the solubility tag used with Cat^C is shown as dashed lines. (c) Sequence logo of the Block B loop (left) Block F loop (middle) and C-terminal Block G (right) generated from an alignment of TerL intein homologues (Table S1).

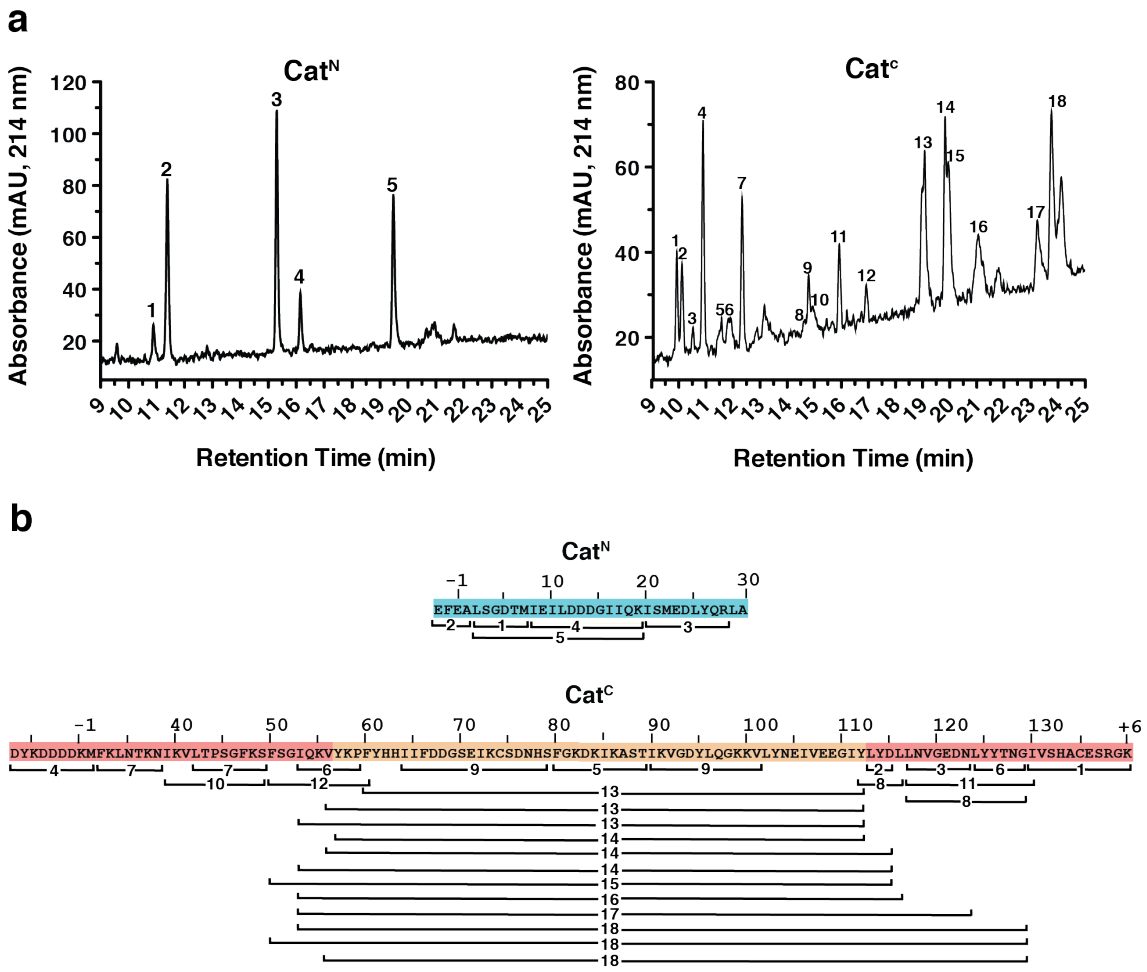


Figure S4: RP-HPLC analysis of limited Proteolysis of Cat fragments. (a) RP-HPLC from the Cat^N (left) and Cat^C (right) proteolysis experiment (t = 30 min) with numbered samples corresponding to the ESI-MS data in Table S4. (b) Primary sequence of the Cat^N and Cat^C inteins used in the limited proteolysis experiment with the proteolysis fragments detected indicated below as brackets. The number of each bracket corresponds to the RP-HPLC peak in panel a.

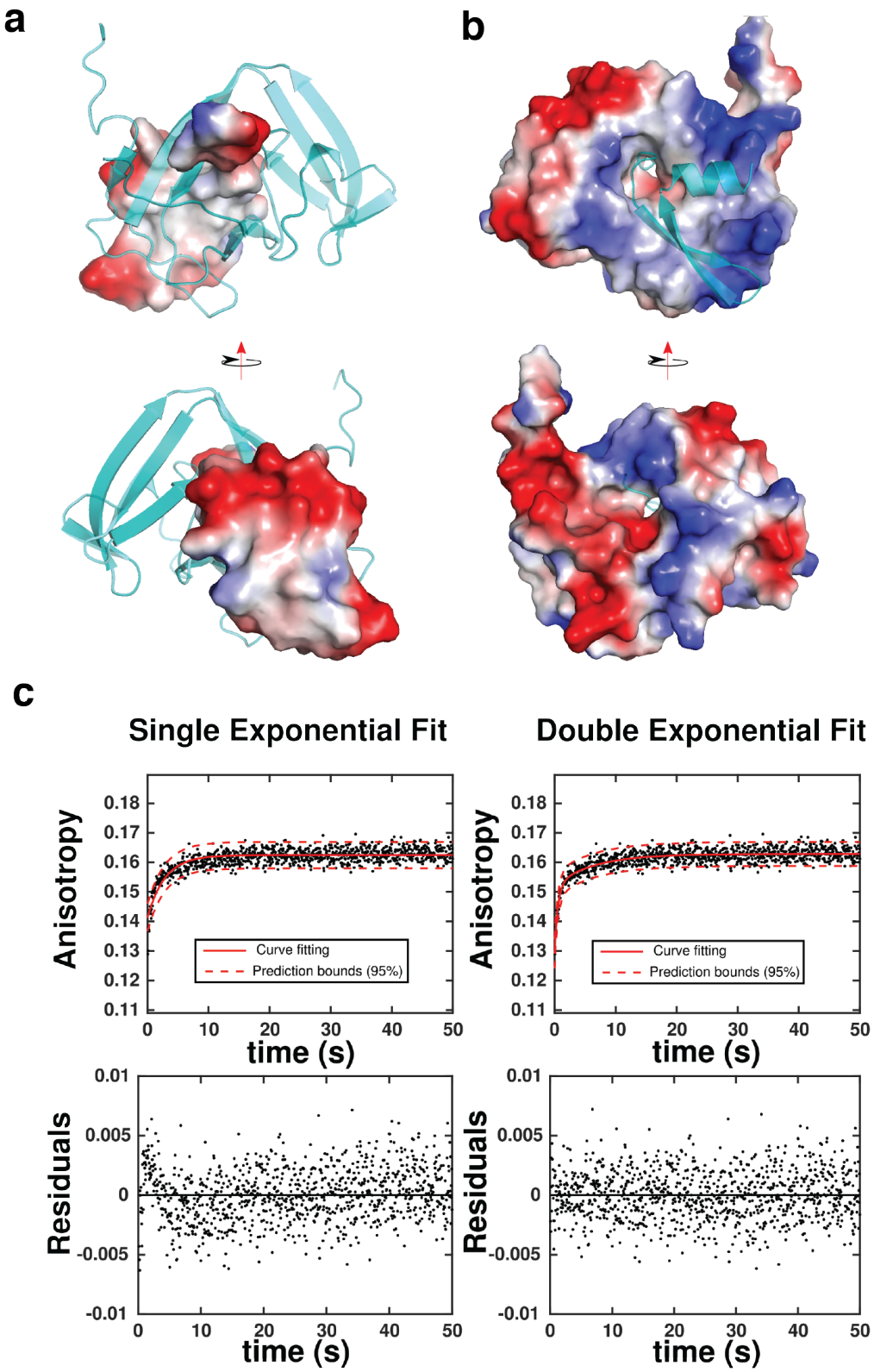


Figure S5: Electrostatic surface of Cat. (a) Electrostatic surface potential of Cat^N with electronegative regions colored in red, electropositive regions colored in blue, and neutral regions colored in white. Cat^C is depicted as a cartoon (cyan). (b)

Electrostatic surface potential of Cat^C with electronegative regions colored in red, electropositive regions colored in blue, and neutral regions colored in white. Cat^N is depicted as a cartoon (cyan). (c) Representative data and fits for kinetic binding experiments. Top: Single (left) and double (right) exponential models for the nonlinear least squares fitting of stopped flow anisotropy measurements of Fl-Cat^N upon mixing with SUMO-Cat^C. Bottom: Residual values obtained between experimental and predicted values are plotted for the single (left) and double (right) exponential fits.

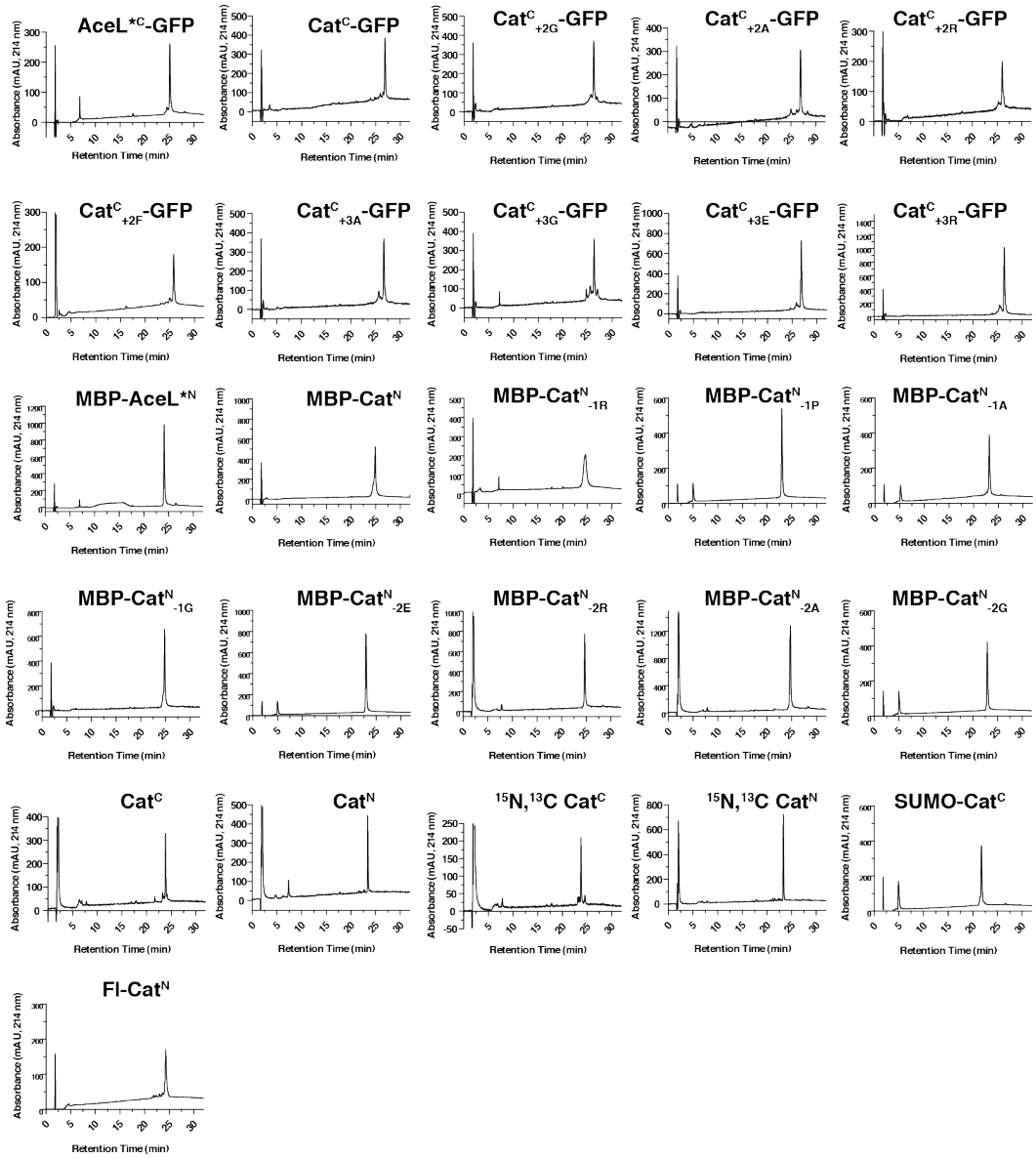
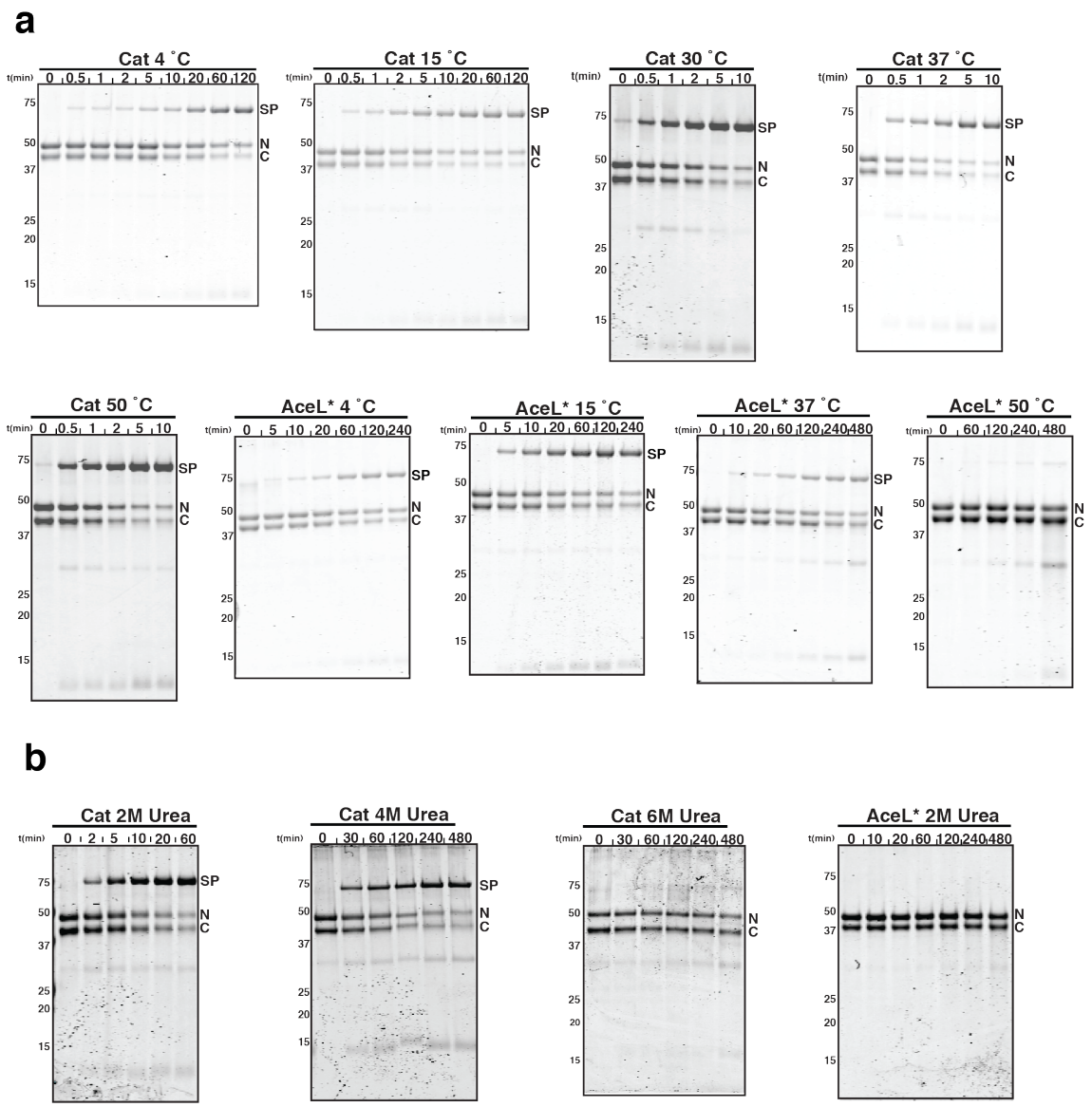


Figure S6: RP-HPLC analysis of inteins utilized in this study. The masses corresponding to each RP-HPLC chromatogram are reported in Table S8.


a
Cat 4 ˚C
t(min) 0 0.5 1 2 5 10 20 60 120
Cat 15 ˚C
t(min) 0 0.5 1 2 5 10 20 60 120
Cat 30 ˚C
t(min) 0 0.5 1 2 5 10
Cat 37 ˚C
t(min) 0 0.5 1 2 5 10
Cat 50 ˚C
t(min) 0 0.5 1 2 5 10
AceL* 4 ˚C
t(min) 0 5 10 20 60 120 240
AceL* 15 ˚C
t(min) 0 5 10 20 60 120 240
AceL* 37 ˚C
t(min) 0 10 20 60 120 240 480
AceL* 50 ˚C
t(min) 0 60 120 240 480
75
50
37
25
20
15
SP
N
C
b
Cat 2M Urea
t(min) 0 2 5 10 20 60
Cat 4M Urea
t(min) 0 30 60 120 240 480
Cat 6M Urea
t(min) 0 30 60 120 240 480
AceL* 2M Urea
t(min) 0 10 20 60 120 240 480

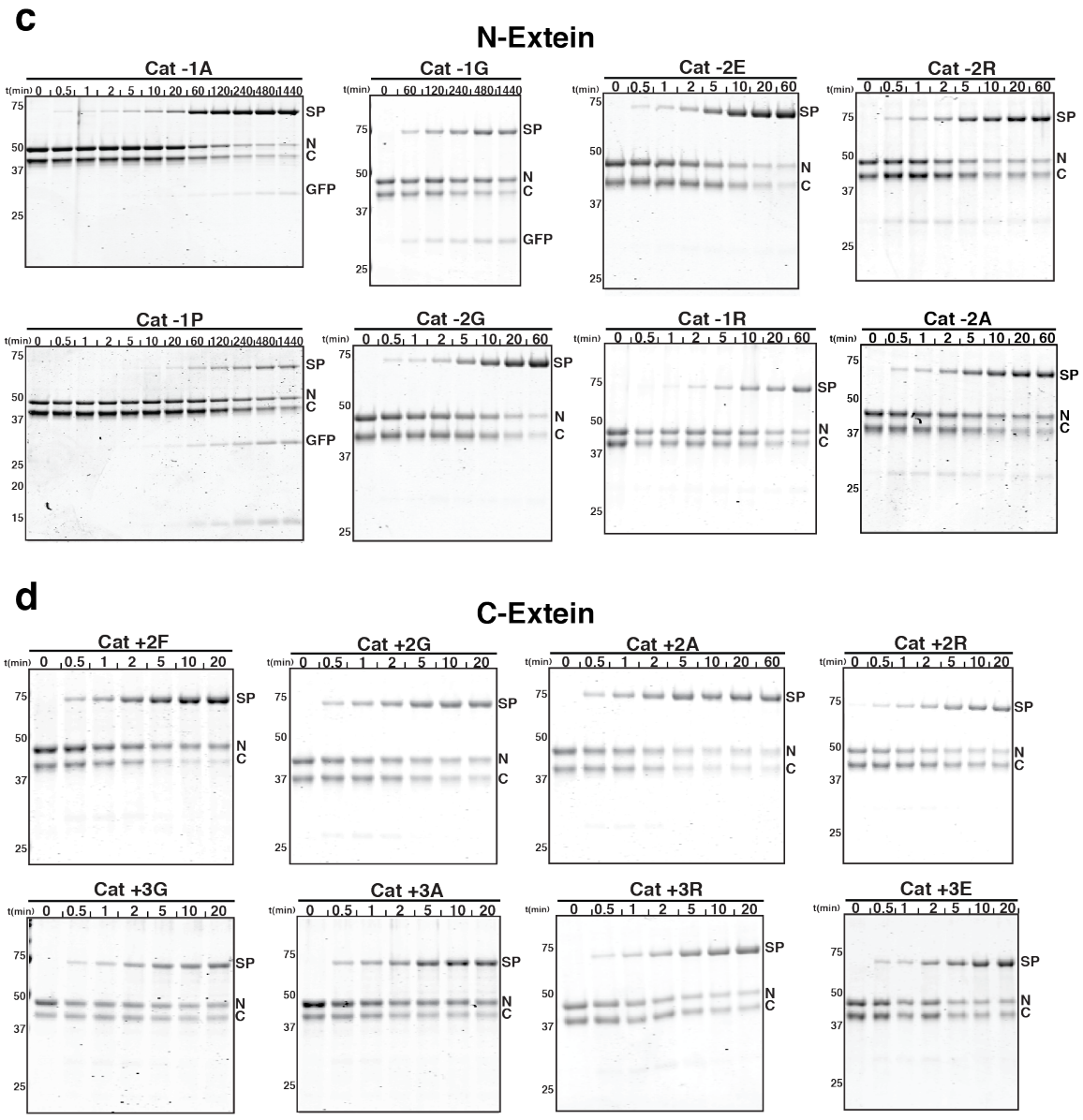


Figure S7: Representative splicing gels of protein *trans*-splicing reactions. (a) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat and AceL* at the indicated temperatures. Bands correspond to MBP-Int^{N} (N), Int^{C}-GFP (C) and the spliced product (SP) are indicated. (b) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat and AceL* at the indicated concentrations of urea. Bands corresponding to MBP-Int^{N} (N), Int^{C}-GFP (C) and the spliced product (SP) are indicated. (c) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat with the indicated -1 and -2 N-extein mutations (from the WT "FE" sequence). Bands corresponding to MBP-Cat^{N} (N), Cat^{C}-GFP (C) and the spliced product (SP) are indicated. C-terminal cleavage is observed for the -1A and -1P mutations and are indicated on the gel (GFP). (d) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat with the indicated +2 and +3 C-extein mutations (from the

WT "EF"). Bands corresponding to MBP-Cat^{N} (N), Cat^{C}-GFP (C) and the spliced product (SP) are indicated.

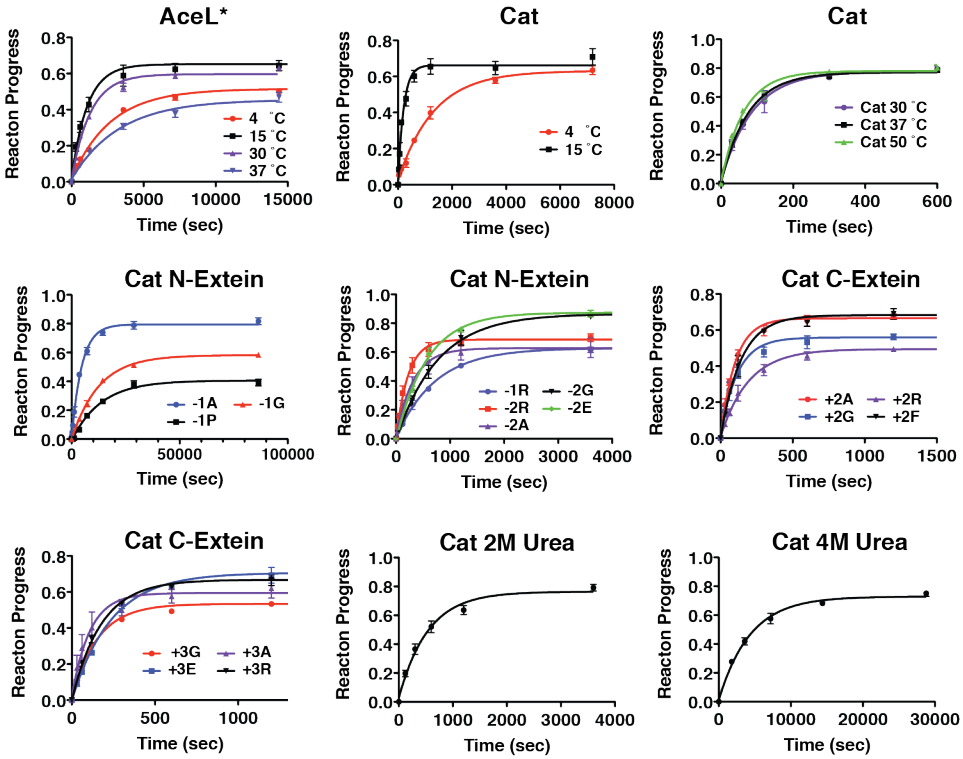


Figure S8: Reaction progress curves. Reaction progress curves are presented for the splicing reactions carried out in this study. The best-fit lines for each reaction are shown.

Supplementary Tables

Table S1: Identified TerL Inteins

C-intein	Source	Project #	Sequence
AceL	Mootz et al[17]		MFRTNTNNIKILSPNGFSNFNGIQKVERNLYQHIIFDD DTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLYNEL VNENIFLYDPINVEKESLYITNGVVSHN
AceL*	Mootz et al[17]		MFRTNTNNIKILGPNGFSNFIGIQKVERDQYQHIIFDD DTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLYNEL VNENIFLYDPINVEKESLYITNGVVSHN
Cep	NCBI	CEPX01183120	MFKTNTNNIKILSPDGFSNFNGIQKVKRKLYQHIIFEG GTEIKTSINHSFGKDKILARDIKVGDYLNNKKVLYNEL VNEKIFLYDPINVEKENLYITNDVVSHN
Mdt	NCBI	MDTC01246584	MYKVNNNIKVKTPTGFQSFSGIQKVFKPFYHWIIFDDG SEIKCSDNHSFGSEKIKASSLKLDDIIQGKKVLYNEIV EEGIYLYDLLDVGEENLYITNKIISHN
Meh	NCBI	MEHZ010888690	MSKTYEVLSPSGFVKFSGIQKVSRSKYRHFIFDDGAEL KCSLNHRFGKDEILASSLWPSSDLQGKNILYAEDVEED IDLYDLLNVGGGNLYYTNGLVSHN
Aac	NCBI	AACY020064168	MFKINKNIKVKTPDGFKDFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGKEKIKASTIKVDDILQEKKVLYNEIV EEGIYLYDLLDVGEDNLYYSNNIVSHN
Aac2	NCBI	AACY023445674	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGKEKIKASTIRVDDFLQGKKVVYNEIV EEGIYLYDLLDVGENNLYYSNNIISHN
Cen	NCBI	CENI01048299	MYKLNSSIKVKTPRGFKKFAGIQKVRKPVYQWIIFGDD SEIKCSLDHSFGEEQVKAHTIKTGDLLQHKEVVYSEIV EEPIDLYDLLEVEDGNLYNTNGVVSHN
Cep2	NCBI	CEPQ01016765	MYEVLSPSGFVKFSGVQKVSRSKYRHFIFDDGTEIKCS LDHRFGGLWDEDEILASSLNRGEYLQGKKILYVEDVEE QIDLYDLMNVDGGNLYYTNGLVSHN
Cep3	NCBI	CEPZ01087314	MSKTYEVLSPSGFVKFSGIQKVSHSKYRHFIFDDGTEL KCSFNHRFGKDEILASSLCRGSDLQGKKILYAEDVEED IDLYDLLNVGGGNLYYTNGLVSHN
Cep4	NCBI	CEPZ01013308	MYIRYQKTTSKTYEVLSPSGFVNFSGIQTVPHSKYRHF IFDDGTELKCSLNHRFDKDEILASSLWRGAELQGKQIL YAEDIEEDIDLYDLLNVGGGNLYYTNGLVSHN
Cep5	NCBI	CEPS01165861	MFTKYKILTPNGYESFDGVNRIKRDMYSHLIFSSGIEI RCSLNHPLYISKGDIIKSYELKIGDKVLSKNGWEIVTY NEIIEEPIYLYDIINSGKDHNYYTNDILSHN
Cep6	NCBI	CEPZ01055800	MISKNFTKYKILTPNGYESFDGVNRIKRDMYSHLIFSS GIEIRCSLNHPLYISKGDIIKSYELKIGDKVLSKNGWE IVTYNEIIEEPIYLYDIINSGKDHNYYTNDILSHN
Lak	JGI	Ga0169931	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGKEKIKASTIRVDDFLQGKKVVYNEIV EEGIYLYDLLDVGENNLYYSNDIISHN
Kab	JGI	Ga0172376	MFKLNKNIRVKTPSGFKSFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGEEQIKASSIKVDDFLQGKKVVYNEIV EEGIYLYDLLDVGEDNLYYSNDVVSHN
Chb	JGI	Ga0129336	MFKLNKNIKVKTPRGFKFFSGIQKVYKPYYHWIIFDDG SEIKCSDNHSFGKEKTKASTIKVDDFLQGKKVVYNEIV EEGIYLYDLLDVGEDNLYYSNEIISHN
Del	JGI	Ga0075462	MFKLNKNIKVKTPSGFKYFSGIQKVYKPFYHWIIFDDG TEIKCSDNHSFGKEQIKASMIKVDDFFQGKKVVYNEIV EEEIYLYDLLDVGEDNLYFSNGIISHN
Del2	JGI	Ga0075478	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGKEKIKASTIKVDDLLQGKKVVYNEIV EEGIYLYDLLDVGEDNLYYSNNLVSHN
Del3	JGI	Ga0075478	MFKLNKNITVKTPSGFKSFSGIQKVYKPFYHWIIFDDG SEIKCSDNHSFGEEQIKASMIKVDDFLQGKKVVYNEVV EEGVYLYDLLDVGEDNLYYSNNIISHN
AceL2	JGI	Ga0075117	MFRTNTDNIKILSPSGFSNFNGIQKVERDLYQHIIFDD KSEIKTSINHPFGKDKILARNIKVGDYLNSKKVLYNEL VAEKITLYDPINVEKENLYITNGVISHN
AceL3	JGI	Ga0075117	MFRTNTDNIKILSPSGFSNFNGIQKVERDLYQHIIFDD KSEIKTSINHPFGKDKILARNIKVGDYLNSKKVLYNEL VNEKITLYDPINVEKENLYITNGVISHN
N-intein	**Source**	**Project #**	**Sequence**

AceL	Mootz et al[17]		CVYGDTMVETEDGKIKIEDLYKRLA
AceL*	Mootz et al[17]		CVSGDTMVETEDGKIKIEDLYKRLA
AAC	NCBI	AACY023445674	CLGGDTIIEIQDDDGITQKISMEDLYERL
AAC2	NCBI	AACY020064168	CLGGDTEIEILDDNGIVQKTSMENLYERL
FUW	NCBI	FUWD010114546	CLGGETLIEIQDDNENISKVSMEDLYDRM
FUW2	NCBI	FUWD010387041	CVDGDTIVEIYDKKTKEEYCVKIKDLYDLI
FUW3	NCBI	FUWD012964875	CLSGDTQIEIKNVNDKIESVSMEELYERM
AAC3	NCBI	AACY020820060	CLSGDTMIEILDENGIPQKISMEDLYQR
MDT	NCBI	MDTB01192700	CVSGDTNIEIECEDGVETTTIKDLYDRM
CEP	NCBI	CEPX01183120	CVDGDTMVETEDGKIKIEDLYKKL
MEH	NCBI	MEHZ011579446	CVYGDTMVETEDGKIKIEDLYKKL
MDT2	NCBI	MDTC01246584	CVRGDTLVEVEKDDVISEMRIEDLYNRM
ABL	NCBI	ABLX01341501	CVGGNTLVEVEKDDIISEMRIEDLYNTM
SSF	JGI	Ga0102963	CLSGDTTIEILDVDGIPQKISMEDLYQRL
Del	JGI	Ga0075478_10047284	CLSGDTMIEILDESGIPQKISMKELYQRM
Del2	JGI	Ga0075478_10000264	CLDGNTSIEILDENNTIQKISMENLYKRL
Del3	JGI	Ga0070746	CLGGDTIIEIQDDDGITQKISMEDLYQRL
Del4	JGI	Ga0070752	CLDGGTSIEILDTNNITQKISLENLYERL
Del5	JGI	Ga0070749	CLSGDTSIEILDENNTIQKISMEDLYERL
Del6	JGI	Ga0070754	CLSGDTLIEIIDDDGNTQKISMEDLY
Del7	JGI	Ga0070751	CLSGDTLIEIIDDDGNTQKISMEDLYQ
Kab	JGI	Ga0172375	CLGGDTIIEIKDDDGITQKISMEDLYQRL

Table S2: Protein Splicing at Indicated Temperatures

Intein	Temp (˚C)	k_{splice} (s^{-1})	$t_{1/2}$ (s)
AceL*	4	$(3.70 \pm .26) \times 10^{-4}$	1873 ± 132
AceL*	15	$(9.17 \pm 1.2) \times 10^{-4}$	756 ± 100
AceL*	30	$(7.68 \pm .58) \times 10^{-4}$	902 ± 68
AceL*	37	$(3.03 \pm .30) \times 10^{-4}$	2287 ± 228
Cat	4	$(7.54 \pm .48) \times 10^{-4}$	919 ± 58
Cat	15	$(4.81 \pm .48) \times 10^{-3}$	144 ± 14
Cat	30	$(1.17 \pm .10) \times 10^{-2}$	59 ± 5
Cat	37	$(1.32 \pm .06) \times 10^{-2}$	52 ± 2
Cat	50	$(1.58 \pm .12) \times 10^{-2}$	44 ± 3

Table S3: Protein Splicing in Chaotropic Agents

Intein	[Urea] (M)	k_{splice} (s^{-1})	$t_{1/2}$ (s)
Cat	2	$(1.86 \pm .18) \times 10^{-3}$	373 ± 35
Cat	4	$(1.73 \pm 1.7) \times 10^{-4}$	3826 ± 368

Table S4: Masses from limited proteolysis

Cat^N			
Peak[a]	**$Mass_{obs}$ (Da)**	**$Mass_{exp}$ (Da)**	**Position**
1	623.27	623.27	2 to 7
2	495.21	495.21	-3 to 1
3	1154.56	1154.55	20 to 28
4	1371.75	1371.74	8 to 19
5	1976.00	1975.99	2 to 19
Cat^C			
Peak[a]	**$Mass_{obs}$ (Da)**	**$Mass_{exp}$ (Da)**	**Position**
1	1186.62	1186.6	130 to +6
2	410.2	410.19	113 to 115
3	760.35	760.35	117 to 123
4	1144.47	1144.45	-8 to 31
5	1094.65	1094.62	80 to 89
6	875.56	875.53	53 to 59
6	730.34	730.4	124 to 129
7	864.515	864.49	32 to 38
7	836.45	836.45	42 to 49
8	686.35	686.34	112 to 116
8	1471.67	1471.68	117 to 129
9	1779.8	1779.79	64 to 79
9	1347.8	1347.82	90 to 101
10	1176.71	1176.7	39 to 49
11	1584.76	1584.75	117 to 130
12	1313.66	1313.76	50 to 60
13	6093.1	6093.8	60 to 112
13	6580.3	6581.4	56 to 112
13	6950.6	6950.9	53 to 112
14	6483.2	6482.3	57 to 112
14	6972.5	6972.9	56 to 115
14	7341.7	7342.3	53 to 115
15	7632.9	7633.6	50 to 115
16	7455.8	7455.5	53 to 116
17	8197.1	8197.2	53 to 123
18	8908.4	8909	53 to 129
18	9199.6	9200.3	50 to 129
18	8538.2	8539.6	56 to 129

[a]The indicated peak number corresponds to the RP-HPLC traces in Figure S4.

Table S5: Steady State Binding Constants

	100 mM NaCl	500 mM NaCl
[Cat^C] (pM)	**Anisotropy**	**Anisotropy**
0	0.064 ± 0.001	0.084 ± 0.006
100	0.087 ± 0.009	0.103 ± 0.010
200	0.011 ± 0.010	0.120 ± 0.009
312.5	0.136 ± 0.013	0.142 ± 0.017
500	0.169 ± 0.004	0.168 ± 0.009
625	0.189 ± 0.008	0.185 ± 0.013
750	0.191 ± 0.006	0.185 ± 0.002
1000	0.195 ± 0.003	0.194 ± 0.005
1250	0.196 ± 0.008	0.199 ± 0.002
1875	0.203 ± 0.002	0.205 ± 0.005
2500	0.199 ± 0.003	0.203 ± 0.005
Fit	**K_d (pM)**	**K_d (pM)**
	33.87 ± 8.69	80.41 ± 15.41

Table S6: Kinetic Binding Constants

	100 mM NaCl		500 mM NaCl	
[Cat^C] (nM)	**k_{obs1} (s^{-1})**	**k_{obs2} (s^{-1})**	**k_{obs1} (s^{-1})**	**k_{obs2} (s^{-1})**
200	0.60 ± .07	0.08 ± 0.09	0.44 ± 0.04	0.05 ± 0.009
325	0.78 ± .08	0.07 ± 0.06	0.92 ± 0.12	0.09 ± 0.009
500	1.10 ± .12	0.10 ± 0.015	0.90 ± 0.08	0.10 ± 0.011
750	2.08 ± .16	0.15 ± 0.014	1.87 ± 0.22	0.16 ± 0.013
1000	2.74 ± .18	0.20 ± 0.010	2.32 ± 0.22	0.20 ± 0.020
	k_{on1} (M^{-1} s^{-1})	**k_{on2} (M^{-1} s^{-1})**	**k_{on1} (M^{-1} s^{-1})**	**k_{on2} (M^{-1} s^{-1})**
Fit	(2.80 ± .28) × 10^6	(0.16 ± .019) × 10^6	(2.34 ± 0.30) × 10^6	(0.18 ± 0.016) × 10^6

Table S7: Protein splicing of Cat in varying Extein Contexts

[a]N-extein$_{-2,-1}$	[a]C-Extein$_{+1,+2,+3}$	k_{splice} (s^{-1})	$t_{1/2}$ (s)
FA	**CEF**	$(2.14 \pm .08) \times 10^{-4}$	3244 ± 116
FP	**CEF**	$(7.33 \pm .44) \times 10^{-5}$	9451 ± 575
FG	**CEF**	$(7.92 \pm .33) \times 10^{-5}$	8749 ± 364
FR	**CEF**	$(1.38 \pm .06) \times 10^{-3}$	504 ± 22
RE	**CEF**	$(4.53 \pm .33) \times 10^{-3}$	153 ± 11
AE	**CEF**	$(3.16 \pm .36) \times 10^{-3}$	220 ± 25
GE	**CEF**	$(1.30 \pm .06) \times 10^{-3}$	532 ± 23
EE	**CEF**	$(1.76 \pm .06) \times 10^{-3}$	394 ± 14
FE	**CAF**	$(9.75 \pm .60) \times 10^{-3}$	71 ± 4
FE	**CGF**	$(8.57 \pm .91) \times 10^{-3}$	80 ± 9
FE	**CRF**	$(5.16 \pm .42) \times 10^{-3}$	134 ± 11
FE	**CFF**	$(7.08 \pm .51) \times 10^{-3}$	98 ± 7
FE	**CEG**	$(6.47 \pm .40) \times 10^{-3}$	107 ± 7
FE	**CEE**	$(4.23 \pm .23) \times 10^{-3}$	164 ± 9
FE	**CEA**	$(9.20 \pm 1.42) \times 10^{-3}$	75 ± 12
FE	**CER**	$(5.65 \pm .25) \times 10^{-3}$	123 ± 5

[a]The position of mutation from the wild type extein sequence is highlighted in red.

Table S8: Masses of purified proteins

eGFP Proteins	Expected Mass (Da)	Observed Mass (Da)
AceL*^{C}-GFP	40424.6	40424.3
Cat^{C}-GFP	40277.7	40275.6
Cat$^{C}_{+2G}$-GFP	40205.5	40205.8
Cat$^{C}_{+2A}$-GFP	40219.5	40218.7
Cat$^{C}_{+2R}$-GFP	40304.6	40305.8
Cat$^{C}_{+2F}$-GFP	40295.6	40294.9
Cat$^{C}_{+3R}$-GFP	40286.6	40286.4
Cat$^{C}_{+3A}$-GFP	40201.5	40201.0
Cat$^{C}_{+3G}$-GFP	40187.5	40187.3
Cat$^{C}_{+3E}$-GFP	40259.5	40258.4
MBP Proteins	**Expected Mass (Da)**	**Observed Mass (Da)**
MBP-Cat^{N}	44094.0	44093.4
MBP-AceL*^{N}	43508.3	43508.1
MBP-Cat$^{N}_{-1R}$	44121.1	44120.2
MBP-Cat$^{N}_{-1A}$	43036.0	44034.5
MBP-Cat$^{N}_{-1G}$	44022.0	44022.4
MBP-Cat$^{N}_{-1P}$	44062.0	44062.8
MBP-Cat$^{N}_{-2R}$	44103.0	44101.9
MBP-Cat$^{N}_{-2A}$	44017.9	44017.2
MBP-Cat$^{N}_{-2E}$	44076.0	44076.4
MBP-Cat$^{N}_{-2G}$	44003.9	44002.8
Proteins for NMR	**Expected Mass (Da)**	**Observed Mass (Da)**
FLAG-Cat^{C}	13499.2	13499.9
Cat^{N}	3773.2	3773.0
^{15}N, ^{13}C FLAG-Cat^{C}	14257.6	14247.5
^{15}N, ^{13}C Cat^{N}	3974.7	3972.7
Proteins for Binding	**Expected Mass (Da)**	**Observed Mass (Da)**
Fl-Cat^{N}	4187.6	4186.7
SUMO-Flag-Cat^{C}	26766.0	26764.7

Table S9: Sequence of Proteins Utilized in this Study

Construct	[a]Sequence
SUMO-Cat^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**EFEALSGDTMIEILD DDGIIQKISMEDLYQRLA**
SUMO-AceL*^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**EFEAVSGDTMVETED GKIKIEDLYKRLA**
SUMO-GOS^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**EFEAISQESYINIEV NGKVETIKIGDLYKKLSFNERKFNE**
SUMO-Cat^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**MFKLNTKNIKVLTPS GFKSFSGIQKVYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDYLQGKKVLYNEIVEEGIY LYDLLNVGEDNLYYTNGIVSHACEFL**
[b]SUMO-Flag-Cat^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**DYKDDDDKMFKLNTK NIKVLTPSGFKSFSGIQKVYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDYLQGKKVLYN EIVEEGIYLYDLLNVGEDNLYYTNGIVSHACESRGK**
SUMO-AceL*^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**MFRTNTNNIKILGPN GFSNFIGIQKVERDQYQHIIFDDDTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLYNELVNENIF LYDPINVEKESLYITNGVVSHACEFL**
SUMO-GOS^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**MKLPESVVKNNINLK IETPYGFENFYGVNKIKKDKYIHLEFTNGEKLKCSLDHPLSTIDGIVKAKDLDKYTEVYTKFGGCFL KKSKVINESIELYDIVNSGLKHLYYSNNIISHACEFL**
AceL*^{C}-GFP	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**MFRTNTNNIKILGPN GFSNFIGIQKVERDQYQHIIFDDDTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLYNELVNENIF LYDPINVEKESLYITNGVVSHNCEFLMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYK TRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDG SVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKDYK DDDDK**
[c]Cat^{C}-GFP	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**MFKLNTKNIKVLTPS GFKSFSGIQKVYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDYLQGKKVLYNEIVEEGIY LYDLLNVGEDNLYYTNGIVSHNCEFLMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYK TRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDG SVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKDYK DDDDK**
MBP-AceL*^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**KIEEGKLVIWINGDK GYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPD KAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMF NLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAA FNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENY LLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINA ASGRQTVDEAPKDAQTNEFECVSGDTMVETEDGKIKIEDLYKRLA**
[c]MBP-Cat^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLR RLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG**KIEEGKLVIWINGDK GYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPD KAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMF NLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAA FNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENY LLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINA ASGRQTVDEAPKDAQTNEFECLSGDTMIEILDDDGIIQKISMEDLYQRLA**
Fl-Cat^{N}	**Fl-GEFEALSGDTMIEILDDDGIIQKISMEDLYQRLA**

[a]The sequences shown correspond to the complete protein expressed by the pET-30 expression vector. The sequence corresponding to the protein cleaved from the SUMO expression tag is shown in bold.

[b]The optimized Cat[C] intein construct with appended charged residues utilized for the structural studies
[c]The WT intein sequences are shown for both MBP-CatN and CatC-GFP. The highlighted red residues correspond to the positions of mutation for the extein activity screen

References

(1) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J Mol Biol* **1990**, *215*, 403.
(2) Tatusova, T.; Ciufo, S.; Fedorov, B.; O'Neill, K.; Tolstoy, I. *Nucleic Acids Res* **2014**, *42*, D553.
(3) Grigoriev, I. V.; Nordberg, H.; Shabalov, I.; Aerts, A.; Cantor, M.; Goodstein, D.; Kuo, A.; Minovitsky, S.; Nikitin, R.; Ohm, R. A.; Otillar, R.; Poliakov, A.; Ratnere, I.; Riley, R.; Smirnova, T.; Rokhsar, D.; Dubchak, I. *Nucleic Acids Res* **2012**, *40*, D26.
(4) Waterhouse, A. M.; Procter, J. B.; Martin, D. M.; Clamp, M.; Barton, G. J. *Bioinformatics* **2009**, *25*, 1189.
(5) Gutierrez, M.; Shah, B. D.; Gabrail, N. Y.; de Nully Brown, P.; Stone, R. M.; Garzon, R.; Savona, M.; Siegel, D. S.; Baz, R.; Mau-Sorensen, M. *Blood* **2013**, *122*, 90.
(6) Stevens, A. J.; Brown, Z. Z.; Shah, N. H.; Sekar, G.; Cowburn, D.; Muir, T. W. *J Am Chem Soc* **2016**, *138*, 2162.
(7) Stevens, A. J.; Sekar, G.; Shah, N. H.; Mostafavi, A. Z.; Cowburn, D.; Muir, T. W. *Proc Natl Acad Sci U S A* **2017**, *114*, 8538.
(8) Shah, N. H.; Dann, G. P.; Vila-Perello, M.; Liu, Z.; Muir, T. W. *J Am Chem Soc* **2012**, *134*, 11338.
(9) Jaravine, V.; Ibraghimov, I.; Orekhov, V. Y. *Nat Methods* **2006**, *3*, 605.
(10) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J Biomol NMR* **1995**, *6*, 277.
(11) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, M.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. *Proteins* **2005**, *59*, 687.
(12) Schwarzinger, S.; Kroon, G. J.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J Am Chem Soc* **2001**, *123*, 2970.
(13) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J Biomol NMR* **2009**, *44*, 213.
(14) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* **2003**, *19*, 315.
(15) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr D Biol Crystallogr* **1998**, *54*, 905.
(16) Shah, N. H.; Eryilmaz, E.; Cowburn, D.; Muir, T. W. *J Am Chem Soc* **2013**, *135*, 18673.
(17) Thiel, I. V.; Volkmann, G.; Pietrokovski, S.; Mootz, H. D. *Angew Chem Int Ed Engl* **2014**, *53*, 1306.