1   **Computation-guided optimization of split protein systems**

2

3   Taylor B. Dolberg[1,2,9], Anthony T. Meger[3,4,9], Jonathan D. Boucher[2,5], William K. Corcoran[2,5], Elizabeth E.

4   Schauer[1,2], Alexis N. Prybutok[1,2], Srivatsan Raman[3,4,8*], Joshua N. Leonard[1,2,5,6,7*]

5

6

7   [1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208,

8   United States

9   [2]Center for Synthetic Biology, Northwestern University, Evanston, Illinois 60208, United States

10   [3]Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

11   [4]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706,

12   United States

13   [5]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois 60208, United

14   States

15   [6]Chemistry of Life Processes Institute, Northwestern University, Evanston, Illinois 60208, United States

16   [7]Member, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Evanston, Illinois

17   60208, United States

18   [8]Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

19   [9]These authors contributed equally to this work

20

21   **Contact Information**

22   *Corresponding authors: Srivatsan Raman (sraman4@wisc.edu) and Joshua N. Leonard (j-

23   leonard@northwestern.edu)

24

25 **ABSTRACT**

26 Splitting bioactive proteins, such as enzymes or fluorescent reporters, into conditionally reconstituting

27 fragments is a powerful strategy for building tools to study and control biochemical systems. However, split

28 proteins often exhibit a high propensity to reconstitute even in the absence of the conditional trigger, which

29 limits their utility. Current approaches for tuning reconstitution propensity are laborious, context-specific, or

30 often ineffective. Here, we report a computational design-driven strategy that is grounded in fundamental

31 protein biophysics and which guides the experimental evaluation of a focused, sparse set of mutants—

32 which vary in the degree of interfacial destabilization while preserving features such as stability and catalytic

33 activity—to identify an optimal functional window. We validate our method by solving two distinct split

34 protein design challenges, generating both broad insights and new technology platforms. This method will

35 streamline the generation and use of split protein systems for diverse applications.

36

37 **KEYWORDS**: synthetic biology, split proteins, computational protein design, protein engineering

38

39 **INTRODUCTION**

40

41 Split proteins and conditional reconstitution systems are powerful tools for interrogating biology and

42 controlling cell behavior.[1-4] These systems work by splitting a protein into two fragments to disrupt the

43 protein's function. Each fragment is then fused to a partner domain such that the split protein is

44 reconstituted, and its function is restored only when the partner domains interact. This modular strategy

45 may be applied to diverse functional proteins to control bioluminescence[5, 6], fluorescence[7], proteolytic

46 cleavage[8-10] and transcription[11, 12]. As a result, conditionally-reconstituted split proteins have been

47 employed in a variety of applications including probing and discovering new protein-protein interactions[13-

48 16], studying post-translational modifications[17], imposing small molecule-regulated control over enzymatic

49 activity[18, 19], and rewiring cellular signaling[9, 20].

50

51 Despite their utility in certain contexts, broader application of split protein systems is largely limited by the

52 spontaneous reconstitution of fragments, resulting in high background activity (**Fig. 1a**). Splitting a protein

53 tends to expose its hydrophobic core, creating highly unfavorable interactions between the core and

54 solvent. Reconstitution is driven by a strong inherent preference to desolvate by recombining the fragments.

55 Evaluating alternative splitting sites can vary reconstitution propensity, but this approach often only partially

56 ameliorates the problem because changing splitting sites may not significantly affect underlying

57 hydrophobic forces. Therefore, it is necessary to identify variants with a reconstitution propensity that

58 precludes spontaneous reconstitution but enables reconstitution under desired conditions. Variants with a

59 range of reconstitution propensities can be generated by random mutagenesis and screened for the desired

60 property. However, high-throughput screening is not readily available for all split protein systems, and low-

61 throughput clonal testing of variants can be laborious and suffer from inefficient exploration of sequence

62 space. Even when screens generate improved variants, it may be difficult to interpret why only certain

63 mutants were successful, and as a result, generalizable rules cannot be transferred to guide the tuning of

64 new split protein systems. Furthermore, split protein systems tuned by mutagenesis exhibit performance

65 characteristics determined by (and limited to) the conditions used in the initial screens, again posing a

66 barrier to general applications.

3

67

68   Here, we report a general strategy based on fundamental principles of protein biophysics for optimizing

69   split protein systems which we term **S**plit **P**rotein **O**ptimization by **R**econstitution **T**uning, or SPORT. We

70   use computational mutagenesis with the Rosetta macromolecular modeling suite to guide limited

71   experimental screening and thereby discretely map the sequence-energy landscape of the split interface.

72   This allows us to determine optimal interaction energies that maximize the performance of the split protein

73   system. We demonstrate proof-of-concept by optimizing a split protease system for conditional

74   reconstitution in two different contexts: membrane-embedded and cytosolic. Our approach generates

75   simple design rules that may be extended to tune other split protein systems for distinct design goals and

76   can be implemented by most research laboratories. This work demonstrates a new method for efficiently

77   engineering split protein systems, which will streamline the generation and expand the use of split protein

78   systems for diverse applications.

79

80   **RESULTS**

81

82   **Formulation of the design challenge and strategy**

83

84   As a first step toward developing a strategy that addresses the challenge of designing split proteins, we

85   selected a model system based on the well-studied Tobacco Etch Virus protease (TEVp)—we sought to

86   tune the reconstitution propensity of split TEVp. For this purpose, we modified a synthetic receptor system

87   that we previously reported (Modular Extracellular Split Architecture, or MESA)[21] to serve as a reporter of

88   conditional split TEVp reconstitution. In this testbed, ligand binding-induced dimerization of a membrane

89   receptor reconstitutes an intracellular split TEVp, which then autolytically liberates a sequestered

90   transcription factor to drive reporter gene expression (**Fig. 1b**). Our initial evaluation demonstrated that the

91   canonical split TEVp (split between residues 118/119)[22] showed high propensity to reconstitute, resulting

92   in high background from ligand-independent signaling (**Supplementary Fig. 1**). As the original screens

93   used to identify this split were performed in a soluble rather than membrane-bound context, these data

94   suggest that tethering split TEVp to a membrane may promote reconstitution. Furthermore, we determined

95   that this problem is not limited to the canonical split site, as other TEVp partitioning also yielded poor

96   performance (**Supplementary Fig. 2**). Given these observations, we formulated a design goal: rationally

97   mutate split TEVp to optimize two key MESA performance characteristics—minimal reporter gene

98   expression in the absence of ligand and a substantial fold increase in reporter expression upon ligand

99   addition.

100

101   **Biophysical principles underlying SPORT**

102

103   We developed SPORT, a computation-guided workflow to rationally design split protein interfaces to

104   optimize reconstitution propensity (**Fig. 1c**). SPORT employs Rosetta, a state-of-the-art software package

105   for protein design.[23] Given a protein with a predetermined split site, our first step was to identify key

106   interfacial residues to target for mutagenesis. Residues with large differences in solvent-accessible surface

107   area (SASA)—when comparing intact protein and split fragments—were classified as buried residues.

108   These buried residues are ideal targets for mutagenesis as they likely contribute substantially to the driving

109   force for spontaneous reconstitution. For each buried residue, we performed a comprehensive, *in silico*

110   mutational scan to evaluate the energy perturbation of all possible single-point mutations on the interaction

111   energy across the split protein interface ($\Delta\Delta G_{Interfacial}$) and total stability of the mutated protein ($\Delta\Delta G_{Total}$)

112   relative to the parent. The degree of disruption is a critical design consideration. Insufficient disruption may

113   retain high background activity while excessive disruption may impair catalytic activity due to loss of overall

114   protein stability. Therefore, the interface must be carefully tuned so that the driving force provided by ligand

115   binding-induced dimerization promotes reconstitution. This "Goldilocks zone" likely differs for each

116   individual protein and perhaps depends upon context, and this zone is difficult to define *a priori*. Therefore,

117   our strategy was to identify the Goldilocks zone for a given protein by choosing mutations that span the

118   range of $\Delta\Delta G$ values. We hypothesized that a limited test set of mutants would direct subsequent

119   mutagenesis efforts by predicting desirable mutant combinations from a vast amount of sequence space.

120   Each of these propositions was tested using experimental case studies.

121

122   **Validating SPORT by tuning a membrane-tethered split TEV protease**

123

124   To investigate and validate SPORT, we applied our design workflow to the split TEVp MESA system. We

125   first assessed the per-residue change in solvent accessible surface area (ΔSASA) between the intact form

126   and split fragments (**Fig. 2a**). In total, 130 of the 218 residues showed increased SASA in the isolated

127   fragments. We excluded from this set the catalytic triad and 27 residues lying within a 6 Å coordination

128   sphere around the catalytic triad to avoid perturbing the catalytic function of reconstituted TEVp. Of the

129   remaining 100 positions, we chose the 15 positions with the largest ΔSASA (9 in N-terminal and 6 in C-

130   terminal halves of split TEVp) as candidates for mutagenesis. Next, we evaluated the energy perturbation

131   of all possible single-point mutations (285 in total) at these positions using Rosetta. As expected, few

132   mutations were predicted to increase stability, and a vast majority were destabilizing (**Fig. 2b,**

133   **Supplementary Figure 3**). Many positions exhibited a variety of stabilizing, benign, and destabilizing point

134   mutations that were consistent with structure. For instance, bulky sidechain substitutions (W, F, Y, R, K and

135   H) at position 103 resulted in many steric clashes with neighboring residues (**Fig. 2b** right panel) and

136   subsequently conferred large decreases in predicted stability.

137

138   Guided by these predictions, we next experimentally characterized 20 single-mutant split TEV variants that

139   span a wide range of $\Delta\Delta G_{Interfacial}$ (3.1 to 16.1 Rosetta Energy Units, or REU) and $\Delta\Delta G_{Total}$ (-1.9 to 30 REU)

140   energies (**Fig. 2c**). We observed high background signaling (i.e., reporter expression) for disruptions up to

141    $\Delta\Delta G_{Interfacial}$ ~6.6 REU, which suggested that destabilization was insufficient. However, four out of ten

142   variants with $\Delta\Delta G_{Interfacial}$ >10 REU exhibited reduced background signaling and substantial ligand-induced

143   activation (**Fig. 2c**). The remaining six were completely inactive (or "dead"); they induced no signaling under

144   any conditions. Energy-based partitioning across different phenotypes (inducible, not inducible, and dead)

145   became evident when comparing $\Delta\Delta G_{Interfacial}$ and $\Delta\Delta G_{Total}$ of all 20 single mutant split TEV variants (**Fig.**

146   **2d**). Variants with high background activity due to insufficient destabilization fell in the region where

147    $\Delta\Delta G_{Interfacial}$ < 10 and $\Delta\Delta G_{Total}$ < 10 REU. Variants with the dead phenotype had $\Delta\Delta G_{Interfacial}$ > 10 and

148    $\Delta\Delta G_{Total}$ > 10 REU; since these mutants were expressed, as confirmed by Western blot (**Supplementary**

149   **Fig. 4**), the lack of signaling suggested that these mutations directly preclude reconstitution. Most of the

150   inducible variants (three out of four) were observed in the energy window where $\Delta\Delta G_{Interfacial}$ > 10 and

6

151 $\Delta\Delta G_{Total}$ < 10 REU, which may represent the Goldilocks zone we hypothesized to exist. An additional region

152 contained a mixture of inducible and dead phenotypes. By inspection of these results, we then proposed a

153 model for broadly classifying experimental phenotypes based on energy partitions (**Fig. 2e**).

154

155 **SPORT predicts outcomes of combining mutations**

156

157 We next evaluated whether our proposed classifier model—developed based upon observations with single

158 mutants—could predict the phenotypes of combined NTEVp and CTEVp mutations, including both double

159 (two mutations on one chain) and paired (one mutation on each chain) mutants derived by combining the

160 initial 14 single non-dead mutations tested (**Fig. 3a**). Of the 67 possible double and paired mutants tested,

161 28 were predicted to be inducible. We experimentally tested 14 of these and found that 10 exhibited

162 inducible signaling as predicted, one was dead, and three were not inducible; this yields an observed

163 accuracy of 0.71 for inducible predictions (**Fig. 3b,c**). Interestingly, three of the prediction failures fell at the

164 low end of the range of predicted changes in interfacial energy, suggesting potential opportunities for

165 refining the classifier model. We also noted an interesting trend in our $\Delta\Delta G$ (total and interfacial)

166 calculations—for the sixty-seven mutants tested, the calculated $\Delta\Delta G$ for the double and paired mutants

167 were nearly identical to the sums of the $\Delta\Delta G$ calculated for the associated single mutants (**Supplementary**

168 **Fig. 5**). Thus, for subsequent analyses of combined mutants, we simply added the effects of single mutants

169 in our calculations.

170

171 We next investigated how variations in expression level might impact the inducibility of the mutants. We

172 used Western blot analysis to normalize and vary chain expression levels by adjusting DNA doses

173 (**Supplementary Fig. 6**). Notably, these constructs remained inducible across the entire range of

174 expression levels tested, suggesting that the biophysical mechanism of optimized split protein reconstitution

175 is robust to variations in the expression levels and ratio of the membrane-bound split TEVp fragments.

176 However, we observed that the performance of these constructs (i.e., fold induction of signaling upon ligand

177 addition) could be substantially improved through tuning expression such that protein levels of each

178 fragment are comparable (**Fig. 3d, Supplementary Fig. 7**). Taken together, these results suggest that a

7

179   classifier calibrated with a limited set of experimental observations spanning the full range of ΔΔG can

180   predict function of new mutants with high accuracy in a manner that is independent of expression level of

181   the construct.

182

183   **SPORT predicts phenotypes of novel mutations and combinations**

184

185   All mutations previously tested were derived from predictions based on our computational method. Next,

186   we wanted to test a broad set of mutants (outside the calibration set) to investigate the accuracy of our

187   classification scheme. Therefore, we experimentally characterized additional single mutants (not included

188   in the original calibration set) and all combinations of paired mutations derived from both the original and

189   this expanded mutation set (omitting dead constructs); mutants were selected to explore the boundaries of

190   the energy landscape classifier model and were expected to reflect a wide range of induced and uninduced

191   reporter expression levels and ratios (**Figs. 2e**, **3a**). We also sought to investigate whether the phenotypic

192   partitioning demonstrated for inducibility (**Fig. 3b,c**) is extensible to the other phenotype classes (i.e., dead

193   and not inducible). This expanded set paired 10 N-terminal mutations with 16 C-terminal mutations. In

194   general, we observed that variants with larger $\Delta\Delta G_{Interfacial}$ energies had lower reporter expression levels in

195   both the background (OFF) and ligand-induced (ON) states (**Fig. 4a**, left and middle panels). Thus, the cost

196   of lowering the OFF state is to also lower the ON state, but these reductions are not always proportional.

197   This is evident by the diversity of calculated fold inductions (**Fig. 4a**, right panel). Only one variant,

198   H75P/W198E, exhibited a significant decrease in the OFF state and an increase in the ON state relative to

199   wild type (WT). However, the variants with highest fold inductions, such as H75S/I163P (17.3 fold induction)

200   and H75T/I163P (9.92 fold induction), exhibited significantly lower OFF and ON states than did the WT,

201   reflecting a tradeoff between desirable performance characteristics. Overall, we observed moderate

202   agreement between the actual and predicted phenotypes for these novel variants and combinations (**Fig.**

203   **4b**). The classifier model was most accurate for predicting the not inducible phenotype (25 of 29, 86%).

204   Many inducible phenotype predictions were also confirmed (31 of 52, 60%). This success rate is impressive

205   given that phenotypic classification boundaries were set roughly based upon the sparse calibration set (**Fig.**

8

206   **2e**). Interestingly, this analysis also indicates that the energy landscape calculated by SPORT correlates

207   with each phenotype to differing degrees.

208

209   In order to gain additional insight into how choice of calibration set and sample size may impact the accuracy

210   of SPORT predictions, we performed retrospective bootstrapping analysis of the data presented in **Figure**

211   **4** (see **Supplementary Note 2** for full details). Experimental data were stratified by the energy landscape

212   and partitioned randomly into calibration and prediction subsets. Logistic regression modeling was applied

213   to evaluate the accuracy of classification under various conditions (**Supplementary Fig. 8**). We observed

214   robust prediction accuracy using multiple unique calibration sets and using sample sizes ranging from 4 to

215   28. This outcome suggests that our ability to generate a general classification model was not dependent

216   upon the specific calibration data we used in our initial characterization experiment (**Fig. 2**), and that a

217   relatively small set of calibration data drawn from a distribution like that included in **Figure 4** would be

218   sufficient to generate a general classification model.

219

220   **Extension of SPORT predictions to new design goals**

221

222   A major limitation to current approaches for employing split proteins is that often a variant selected to

223   perform well in one context fails in a different context. To investigate whether the SPORT design method is

224   generalizable beyond our initial model system, we developed a distinct model system. This new system

225   employs split TEVp in a soluble form, where we hypothesized that a different reconstitution propensity

226   would be required compared to the membrane-bound model system. To generate such a soluble test

227   system (**Fig. 5a**), ligand binding domains were fused to split TEVp domains along with a soluble

228   transcription factor flanked by protease cleavage sites and nuclear export signals (NES); thus, TEVp-

229   mediated cleavage removes the NES from the transcription factor to enable nuclear localization and

230   reporter expression. We first developed and tested a panel of soluble transcription factors that could

231   implement this mechanism. This evaluation included varying the number of NES elements, their placement

232   at N and/or C terminus of the transcription factor, and the P1' residue of the TEVp cleavage sequence

233   which governs cleavage kinetics[24] (**Fig. 5b**). Several soluble transcription factor constructs exhibited the

234   desired phenotype of low signaling in the absence of TEVp and high signaling when co-expressed with full

235   TEVp; construct TF10 was selected for evaluating split TEVp variants.

236

237   Using our soluble split TEVp test system, we evaluated a panel of 10 single TEVp mutants and 10 paired

238   TEVp mutants spanning a range of interfacial energies (**Fig. 5c**). The construct based upon WT split TEVp

239   exhibited a substantial fold induction, which is consistent with the fact that this split protein was identified

240   by screens performed in the soluble phase[22]. However, the WT construct also yielded high background

241   signaling in the absence of ligand, indicating an opportunity to improve performance. For modest increases

242   in $\Delta\Delta G_{Interfacial}$ (~0–3.6 REU), background signaling persisted. For intermediate increases in $\Delta\Delta G_{Interfacial}$ (~6-

243   10 REU), a mixture of phenotypes was observed, including dead constructs and those with both

244   substantially reduced background signaling and substantial fold inductions. For large increase in

245    $\Delta\Delta G_{Interfacial}$ > ~10 REU, constructs were generally weakly-inducible or dead. Thus, a focused evaluation of

246   20 design variants, guided by SPORT, yielded ~5 variants exhibiting improved performance compared to

247   the WT construct. Altogether, these observations demonstrate that the SPORT method and associated

248   energy landscape concept can be employed to efficiently solve distinct split protein optimization challenges.

249

250   **DISCUSSION**

251

252   In this study, we developed and validated what is—to our knowledge—the first computational strategy for

253   tuning split protein reconstitution propensity. Although the split TEVp MESA used as our first model system

254   would have been deemed infeasible using standard evaluations of split proteins (**Supplementary Figs. 1-**

255   **2**), application of SPORT to tune this system yielded multiple high-performing new synthetic receptor

256   scaffolds (**Fig. 3d**). We show that unlike the classical MESA receptors we have characterized in prior work[21,]

257   [25, 26], split TEVp MESA tuned by SPORT exhibit excellent performance characteristics (i.e., low-background

258   and high fold-induction) in a manner that is robust to variations in both biosensor expression level and the

259   ratio at which biosensor chains are expressed (**Supplementary Figure 7**). This property is of great practical

260   utility, as it precludes the need to carefully tune the implementation of each biosensor.

261

10

262   Several important insights emerged from this study. First, our approach demonstrated that testing a sparse

263   set of mutants along the energy landscape is an effective strategy to choose optimal interfacial energies to

264   promote conditional reconstitution. Second, multiple point mutations with similar energies exhibited similar

265   performance, which suggests reconstitution propensity depends on the energy of destabilization but is

266   agnostic to specific mutations. Third, the concept of a Goldilocks zone is likely generalizable to different

267   proteins and application contexts, but the optimal energy window may have to be adjusted on a case-by-

268   case basis. We find that membrane-bound split proteins must be destabilized to a greater degree than

269   soluble split proteins in order to avoid spontaneous reconstitution. Altogether, these results show that split

270   protein systems can be engineered based on fundamental principles of protein biophysics, which obviates

271   the need for exhaustive screening and generates rules applicable to new candidate proteins.

272

273   There are several interesting opportunities for extending and improving SPORT in future work. First,

274   although our analysis showed that SPORT can be used to identify mutations that confer specific energy

275   changes, it does not yet enable *a priori* prediction of where the Goldilocks zone will fall for new applications.

276   It is possible that subsequent analysis of many case studies could identify trends that enable such

277   predictions and thus harness SPORT to further focus experimental investigations. An additional opportunity

278   is pairing SPORT with a multiparameteric optimization framework for exploring pareto-optimal tradeoffs

279   between performance characteristics; for example, in our model system, there seems to be such a tradeoff

280   between low background in the ligand-free state and high signaling in the ligand-induced state. Finally, the

281   SPORT algorithm itself may be refined to better avoid false positives (e.g., dead mutants that share a region

282   of the energy landscape with inducible variants). Altogether, our findings suggest many opportunities for

283   expanding the utility of split proteins for many new applications and highlight the impact of SPORT-guided

284   development of novel biochemical and synthetic biology tools.

285

286

287   **METHODS**

288

289   **General DNA assembly**

11

290   Plasmid cloning was performed using standard molecular biology techniques of PCR and restriction

291   enzyme cloning with Phusion DNA Polymerase (NEB), restriction enzymes (NEB; Thermo Fisher), T4 DNA

292   Ligase (NEB), and Antarctic Phosphatase (NEB). Development of the tTA-responsive YFP reporter plasmid

293   was described previously[21]. Plasmids were transformed into chemically competent TOP10 *E. coli* (Thermo

294   Fisher) and grown at 37°C. Plasmid maps are provided as GenBank files (**Supplementary Data 1**).

295

Plasmid preparation

297   Plasmid DNA used for transfection was prepared using the PEG precipitation method, which was previously

298   described in detail.[27]

299

Cell culture

301   HEK293FT cells (Life Technologies/Thermo) were maintained at 37°C incubator and 5% CO_2. Cells were

302   cultured in DMEM (Gibco 31600-091) with 10% FBS, 6 mM L-glutamine (2 mM from Gibco 31600-091 and

303   4 mM from additional Gibco 25030-081), penicillin (100 U/µL), and streptomycin (100 µg/mL) (Gibco

304   15140122).

305

Transfection

307   Transfections were performed in 24 well plates seeded at 1.5×10^5 cell in 0.5 mL of DMEM media. At 6-8

308   hours post-seeding, cells were transfected using calcium phosphate method with a total DNA content of 1-

309   2 ug DNA per mL of media, using DNA prepared by PEG precipitation. All experiments included blue

310   fluorescent protein (BFP) as a control to assess transfection efficiency. The exact DNA amounts added to

311   the mix per well are as follows, unless otherwise stated in figure captions: 25 ng of each TEVp chain, 200

312   ng of BFP control, 200 ng of YFP reporter plasmid, and 150 ng of pcDNA plasmid. This mixture was added

313   dropwise to an equal-volume solution of 2x HEPES-Buffered Saline (280 mM NaCl, 0.5 M HEPES, 1.5 mM

314   Na_2HPO_4) and gently pipetted up and down four times. After 2.5 minutes, the solution was mixed vigorously

315   by pipetting ten times and 100 µL of this mixture was added dropwise to each well of the plated cells, and

316   the plates were swirled gently. For functional experiments, 12 hours post-transfection, media containing 0.1

317   µM rapamycin analog (Takara AP21967) or 0.1% ethanol as a control was added to cells. At 24-30 hours

318   post-media change, cells were harvested for flow cytometry with Trypsin-EDTA, which was then quenched

319   with medium, and the resulting cell solution was added to at least 2 volumes of FACS buffer (PBS pH 7.4

320   with 2–5 mM EDTA and 0.1% BSA). Cells were spun at 150 x g for 5 min, FACS buffer was decanted, and

321   fresh FACS buffer was added. All experiments were performed in biological triplicate.

322

323   **Flow Cytometry**

324   Approximately 10^4 live cells from each transfected well of the 24-well plate were analyzed using a BD LSR

325   Fortessa Special Order Research Product (Robert H. Lurie Cancer Center Flow Cytometry Core) running

326   FACSDiva software. Samples were analyzed using FlowJo v10 software (FlowJo, LLC). The HEK293FT

327   cell population was identified by FSC-A vs. SSC-A gating, and singlets were identified by FSC-A vs. FSC-

328   H gating. A control sample of cells—generated by transfecting cells with a mass of pcDNA (empty vector)

329   equivalent to the mass of DNA used in other samples in the experiment—was used to distinguish

330   transfected and non-transfected cells. For the single-cell subpopulation of the pcDNA-only sample, a gate

331   was made to identify cells that were positive for the constitutively driven blue fluorescent protein (BFP) used

332   as a transfection control in other samples such that the gate included no more than 1% of the non-

333   fluorescent cells. The mean fluorescence intensity (MFI) of the single-cell transfected population was

334   calculated and exported for further analysis. BD LSR Fortessa settings used were as follows: BFP was

335   collected in the Pacific Blue channel (405 nm excitation, 450/50 nm filer) and EYFP was collected in the

336   FITC channel (488 nm excitation, 505 LP and 530/30 nm filter). To quantify reporter expression, the FITC

337   channel MFI was averaged across three biological replicates. Cell autofluorescence was subtracted and

338   MFI was converted to Mean Equivalents of Fluorescein (MEFLs) using the coefficient determined by the

339   calibration curve of UltraRainbow Calibration Particles (Spherotech URCP-100-2H) run in each individual

340   experiment. Standard error was propagated through all calculations.

341

342   **Western Blotting**

343   Western blots were performed to evaluate protein expression and normalize total expression of each TEVp

344   chain. A 3X-FLAG tagged NanoLuciferase as a normalization control, and images were analyzed using

345   ImageJ software. A detailed western blot protocol was previously described.[27]

13

346

Solvent-accessible Surface Area

The structure of TEVp was obtained from the Research Crystallography for Structural Bioinformatics

(RCSB) PDB (ID code: 1LVM). Per-residue solvent-accessible surface areas (SASA) were computed using

GROMACS v2018.1, which utilizes the double cubic lattice method (DCLM) described by Eisenhaber et

al.[28] The change in solvent accessible area was computed as

$$\Delta SASA = SASA_{fragments} - SASA_{reconstituted}$$

where structures of the N and C-terminal fragments were isolated from the crystal structure.

354

Computational Interface Scanning

All modeling calculations were performed using the *Rosetta* molecular modeling suite v3.9. Single-point

mutants were generated using the standard Relax application, which enables local conformational sampling

to minimize energy (**Supplementary Note 1** includes full details). The total energy (ΔG_{Total}) of each mutant

was computed as the average of 100 relaxed models. The energy perturbation to total energy was

computed as

$$\Delta\Delta G_{Total} = \Delta G_{Total}^{Mutant} - \Delta G_{Total}^{WT}$$

The Rosetta Scripts application with the InterfaceAnalyzeMover was applied to each relaxed model to

compute the average residue-residue interaction energies between the N- and C-terminal fragments

(**Supplementary Note 1** includes the full details). The interfacial energy was computed as the pair-wise

sum of all short-range interaction energies as shown by

$$\Delta G_{Interfacial} = \sum_i \sum_j Energy_{i-j}^{SR}$$

where i and j denote the sets of residues within each fragment. The energy perturbation of each mutation

to the interfacial energy was then computed as

$$\Delta\Delta G_{Interfacial} = \Delta G_{Interfacial}^{Mutant} - \Delta G_{Interfacial}^{WT}$$

370

Phenotype Classifier

14

372   Experimentally characterized variants were assigned class labels (not-inducible, inducible and dead) based

373   on reporter expression levels in the ligand-absent and ligand-induced states. Variants with ≥1.2 fold higher

374   reporter expression in the ligand-induced state relative to the ligand-absent state were labeled as inducible.

375   For variants with expression levels <5% of wild-type (WT) sequence in the ligand-induced state and <1.2

376   fold activation were classified as functionally dead. The remaining variants were assigned the not-inducible

377   class label.

378

379   **STATISTICAL ANALYSIS**

380

381   Statistical details for each experiment are included the figure legends. The data shown reflect the mean

382   across these biological replicates of the mean fluorescence intensity (MFI) of approximately 2,000–3,000

383   single, transfected cells. Error bars represent the SEM (standard error of the mean).   For statistical

384   analyses, two-tailed Student's t-tests were used to evaluate whether a significant difference exists between

385   two groups of samples, and the reported comparisons meet the two requirements of this test: (1) the values

386   compared are expected to be derived from a normal distribution, and (2) the variance of each group is

387   expected to be comparable to that of the comparison group since the same transfection methodologies and

388   data collection methods were used for all samples that were compared. A p value of ≤ 0.05 was considered

389   to be statistically significant.

390

391   **DATA AVAILABILITY**

392

393   Data reported in composite figures (**Figs. 2a,b,d**, **3a, 4b, Supplementary Fig. 4**) are included as **Source**

394   **Data**.

395

396   **CODE AVAILABILITY**

397

398   Rosetta details and script are provided in **Supplementary Note 1**.

399

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

T.B.D., A.T.M., S.R, and J.N.L conceptualized the project. T.B.D., J.D.B, W.K.C., E.E.S, created reagents, designed and performed experiments, and analyzed the data. A.N.P. assisted in analyzing and visualizing the data. A.T.M., developed the computational model and code. T.B.D., A.T.M., S.R., and J.N.L. drafted the manuscript, T.B.D., A.T.M., and A.N.P. created the figures. J.N.L. and S.R. supervised the work. All authors edited and approved the final manuscript.

COMPETING INTERESTS

J.N.L is a co-inventor on a patent that covers the MESA technology used in this manuscript (US Patent 9,732,392 B2).

REFERENCES

1. Romei, M.G. & Boxer, S.G. Split Green Fluorescent Proteins: Scope, Limitations, and Outlook. *Annual Review of Biophysics* **48**, 19-44 (2019).

2. Shekhawat, S.S. & Ghosh, I. Split-protein systems: beyond binary protein-protein interactions. *Curr Opin Chem Biol* **15**, 789-797 (2011).

3. Wehr, M.C. & Rossner, M.J. Split protein biosensor assays in molecular pharmacological studies. *Drug Discovery Today* **21**, 415-429 (2016).

4. Muller, J. & Johnsson, N. Split-ubiquitin and the split-protein sensors: chessman for the endgame. *Chembiochem* **9**, 2029-2038 (2008).

5. Paulmurugan, R. & Gambhir, S.S. Monitoring protein-protein interactions using split synthetic renilla luciferase protein-fragment-assisted complementation. *Anal Chem* **75**, 1584-1589 (2003).

6. Dixon, A.S. et al. NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells. *ACS chemical biology* **11**, 400-408 (2016).

7. Ozawa, T., Kaihara, A., Sato, M., Tachihara, K. & Umezawa, Y. Split luciferase as an optical probe for detecting protein-protein interactions in mammalian cells based on protein splicing. *Analytical Chemistry* **73**, 2516-2521 (2001).

8. Gray, D.C., Mahrus, S. & Wells, J.A. Activation of specific apoptotic caspases with an engineered small-molecule-activated protease. *Cell* **142**, 637-646 (2010).

9. Gao, X.J., Chong, L.S., Kim, M.S. & Elowitz, M.B. Programmable protein circuits in living cells. *Science* **361**, 1252-1258 (2018).

10. Fink, T. et al. Design of fast proteolysis-based signaling and logic circuits in mammalian cells. *Nature chemical biology* **15**, 115-122 (2019).

11. Zetsche, B., Volz, S.E. & Zhang, F. A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nature biotechnology* **33**, 139-142 (2015).

12. Nihongaki, Y., Otabe, T., Ueda, Y. & Sato, M. A split CRISPR–Cpf1 platform for inducible genome editing and gene activation. *Nature chemical biology* **15**, 882-888 (2019).

17

452   13.   Paulmurugan, R., Umezawa, Y. & Gambhir, S.S. Noninvasive imaging of protein-protein
453   interactions in living subjects by using reporter protein complementation and reconstitution
454   strategies. *Proc Natl Acad Sci U S A* **99**, 15608-15613 (2002).

455   14.   Fetchko, M. & Stagljar, I. Application of the split-ubiquitin membrane yeast two-hybrid system to
456   investigate membrane protein interactions. *Methods* **32**, 349-362 (2004).

457   15.   Pandey, N., Nobles, C.L., Zechiedrich, L., Maresso, A.W. & Silberg, J.J. Combining random gene
458   fission and rational gene fusion to discover near-infrared fluorescent protein fragments that report
459   on protein-protein interactions. *ACS Synth Biol* **4**, 615-624 (2015).

460   16.   Jones, K.A. et al. Development of a Split Esterase for Protein–Protein Interaction-Dependent Small-
461   Molecule Activation. *ACS Central Science* (2019).

462   17.   Wehr, M.C., Reinecke, L., Botvinnik, A. & Rossner, M.J. Analysis of transient phosphorylation-
463   dependent protein-protein interactions in living mammalian cells using split-TEV. *BMC Biotechnol*
464   **8**, 55 (2008).

465   18.   Camacho-Soto, K., Castillo-Montoya, J., Tye, B. & Ghosh, I. Ligand-gated split-kinases. *J Am*
466   *Chem Soc* **136**, 3995-4002 (2014).

467   19.   Camacho-Soto, K., Castillo-Montoya, J., Tye, B., Ogunleye, L.O. & Ghosh, I. Small molecule gated
468   split-tyrosine phosphatases and orthogonal split-tyrosine kinases. *J Am Chem Soc* **136**, 17078-
469   17086 (2014).

470   20.   Fink, T. et al. Design of fast proteolysis-based signaling and logic circuits in mammalian cells. *Nat*
471   *Chem Biol* **15**, 115-122 (2019).

472   21.   Daringer, N.M., Dudek, R.M., Schwarz, K.A. & Leonard, J.N. Modular extracellular sensor
473   architecture for engineering mammalian cell-based devices. *ACS Synth Biol* **3**, 892-902 (2014).

474   22.   Wehr, M.C. et al. Monitoring regulated protein-protein interactions using split TEV. *Nature Methods*
475   **3**, 985-993 (2006).

476   23.   Crescitelli, R. et al. Distinct RNA profiles in subpopulations of extracellular vesicles: apoptotic
477   bodies, microvesicles and exosomes. *J Extracell Vesicles* **2** (2013).

478   24.   Kapust, R.B., Tozser, J., Copeland, T.D. & Waugh, D.S. The P1' specificity of tobacco etch virus
479   protease. *Biochem Biophys Res Commun* **294**, 949-955 (2002).

480 25.   Hartfield, R.M., Schwarz, K.A., Muldoon, J.J., Bagheri, N. & Leonard, J.N. Multiplexing Engineered

481   Receptors for Multiparametric Evaluation of Environmental Ligands. *Acs Synthetic Biology* **6**, 2042-

482   2055 (2017).

483 26.   Schwarz, K.A., Daringer, N.M., Dolberg, T.B. & Leonard, J.N. Rewiring human cellular input-output

484   using modular extracellular sensors. *Nat Chem Biol* **13**, 202-209 (2017).

485 27.   Donahue, P.S., Draut, J.W., Muldoon, J.J., Edelstein, H.I., Bagheri, N., Leonard, J.N. COMET: A

486   toolkit for compsoing customizable genetic programs in mammalian cells. *bioRxiv* (2019).

487 28.   Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. & Scharf, M. The double cubic lattice method:

488   Efficient approaches to numerical integration of surface area and volume and to dot surface

489   contouring of molecular assemblies. *Journal of Computational Chemistry* **16**, 273-284 (1995).

490

491

492

493 **FIGURES**

494
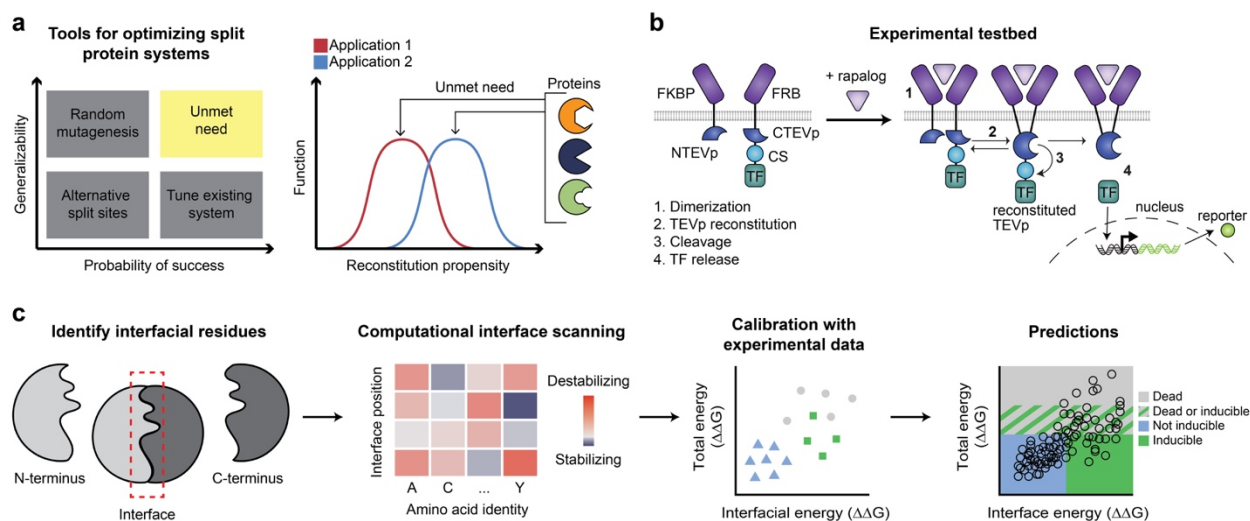
495

496 **Fig. 1 Design-driven strategy for tuning split protein systems. a,** Current methods for optimizing split

497 proteins are limited (left); an ideal tool would enable adapting split proteins for multiple applications, each

498 of which may require distinct reconstitution propensities (right). **b,** This cartoon illustrates the experimental

499 testbed used here; ligand binding-induced chain dimerization results in split TEVp reconstitution, trans-

500 cleavage, and release of a previously sequestered transcription factor to drive reporter expression. **c,** Split

501 Protein Optimization by Reconstitution Tuning (SPORT) workflow: identify important residues at the split

502 interface which are mutable, identify mutations that alter the total and interfacial energy, and use an

503 application-specific model, trained on limited experiments, to identify those mutations that are predicted to

504 yield a desired functional phenotype.

505

21
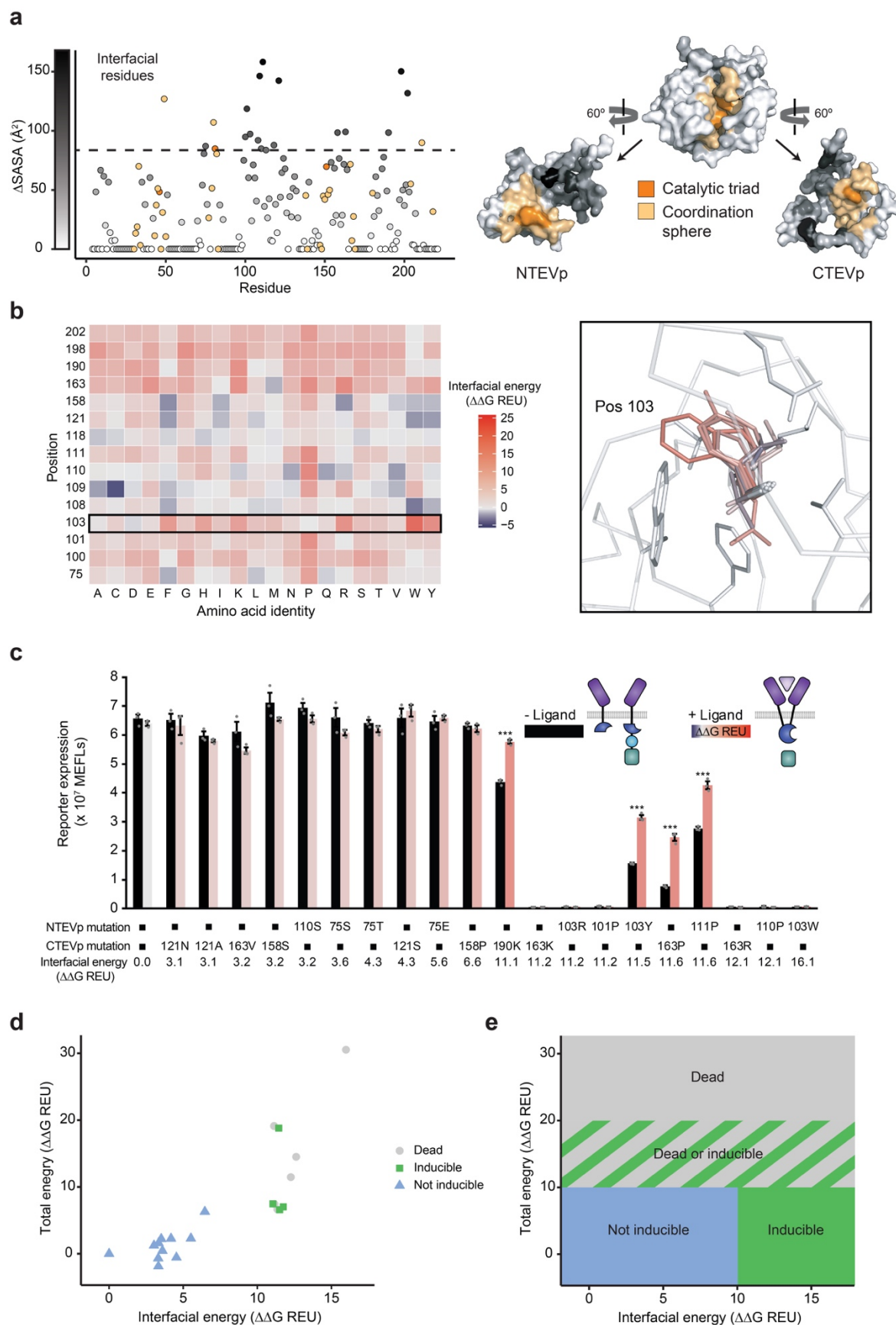
507   **Fig. 2 Computation guided method development and experimental analysis. a,** Left, characterization

508   of the solvent accessible surface area (SASA) of each residue of 118/119 split TEVp. Right, 3D depiction

509   of 118/119 split TEVp, showing the catalytic triad (orange), coordination sphere (yellow), and ΔSASA

510   (greyscale). **b,** Mutational scanning of high ΔSASA residues (left) and example of all possible mutations of

511   residue 103 (right), with change in interfacial energy indicated by color. **c,** Experimental analysis of TEVp

512   mutations predicted to span a range of interfacial energies. Error bars depict S.E.M. (*$p \leq 0.05$, ***$p \leq$

513   0.001). **d,** Experimental phenotypes observed in **c** were plotted on an energy landscape and annotated as

514   indicated by color (reporter expression normalized to WT < 0.05 is "dead", fold induction < 1.2 is "not

515   inducible", fold induction ≥ 1.2 is "inducible"). **e,** Proposed model for predicting zones of functional

516   phenotypes based upon total and interfacial energy; the boundaries were proposed based upon

517   observations using the initial 20 mutants tested in **c**.

518
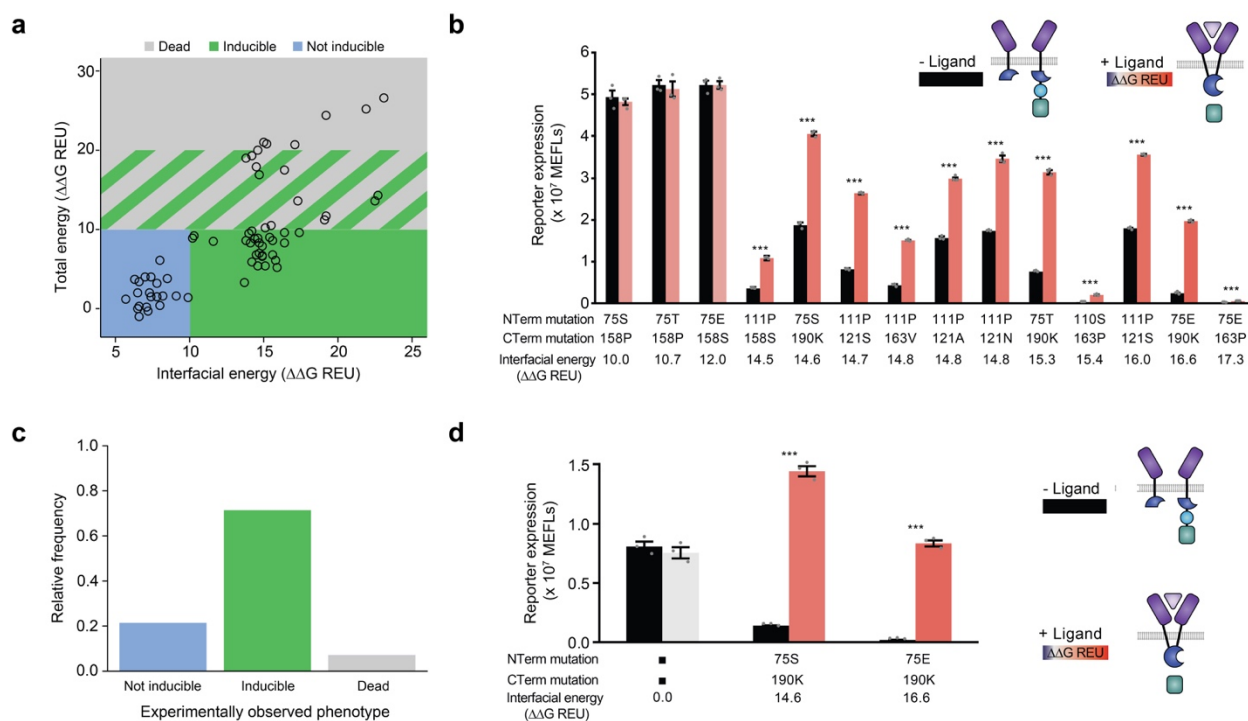
Fig. 3 Evaluation of model-predicted phenotypes for combined mutations. a, Computed energies and predicted phenotypes based on the classifier model—proposed in **Fig. 2e**—of all possible double and paired mutants constructed by combinatorial sampling of the initial single mutants tested (omitting dead mutations) in **Fig. 2c**. **b,** Experimental evaluation of selected mutants predicted to be inducible. **c,** Experimentally observed phenotypes for the fourteen mutants predicted to be inducible (from **b**), showing that the model predicts inducibility at a fairly high rate (10/14). **d,** Normalizing protein expression levels improves performance (fold induction) of selected mutants (from **b**), whereas WT function is not changed. Normalization was achieved using Western Blot analysis (**Supplementary Fig. L4)** to adjust DNA doses transfected (per well, N-terminal chains: 0.4 ng WT, 1 ng 75S, 1.4 ng 75E; C-terminal chains: 5 ng WT, 12 ng 190K). Error bars depict S.E.M. (*$p \leq 0.05$, ***$p \leq 0.001$).
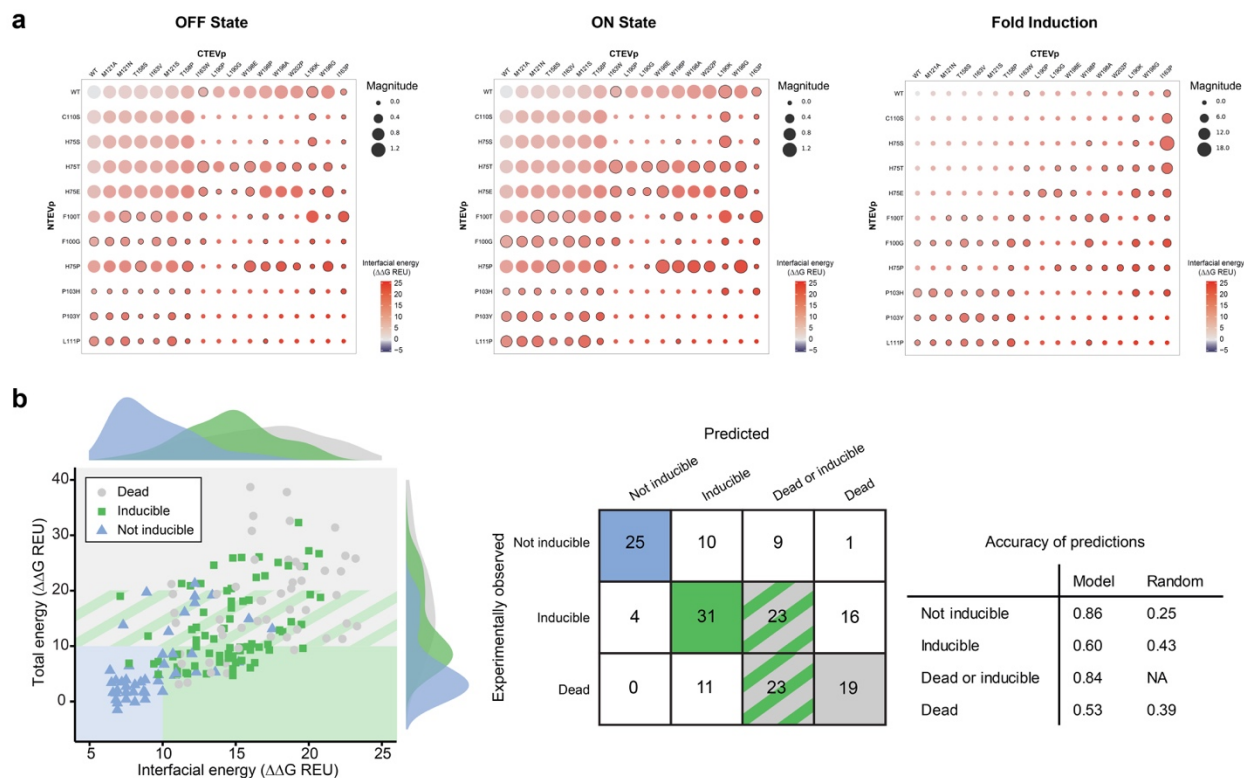
Fig. 4 Evaluation of model-predicted phenotypes for novel mutations and combinations. a, For each experimentally characterized construct, reporter output was quantified in the absence of ligand (OFF state) and following ligand addition (ON state), and the fold induction was calculated. Calculated interfacial energy for each construct is indicated by circle *color*, magnitude of reporter expression (or fold-induction) is indicated by circle *size*, and constructs with a fold induction ≥ 1.2 are denoted with a black border. Single mutants observed to be dead (**Fig. 2**) were not carried forward to this analysis. **b,** Left, experimentally observed phenotypes (data point color) were mapped onto the proposed classifier model from **Fig. 2e**, with observed frequency distributions shown as histograms. Right, evaluation of model prediction accuracy compared to random assignment of phenotypes.
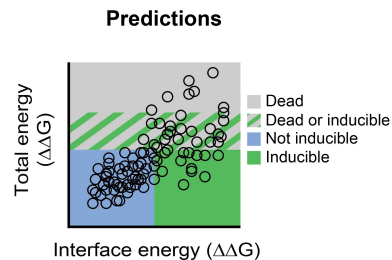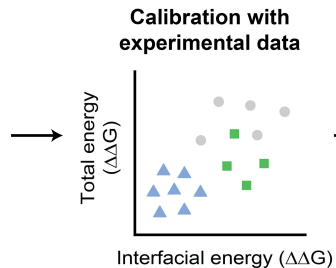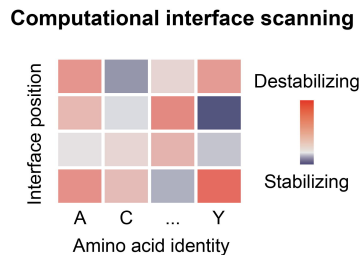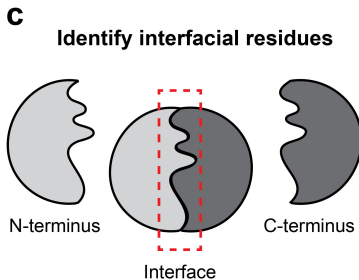
542
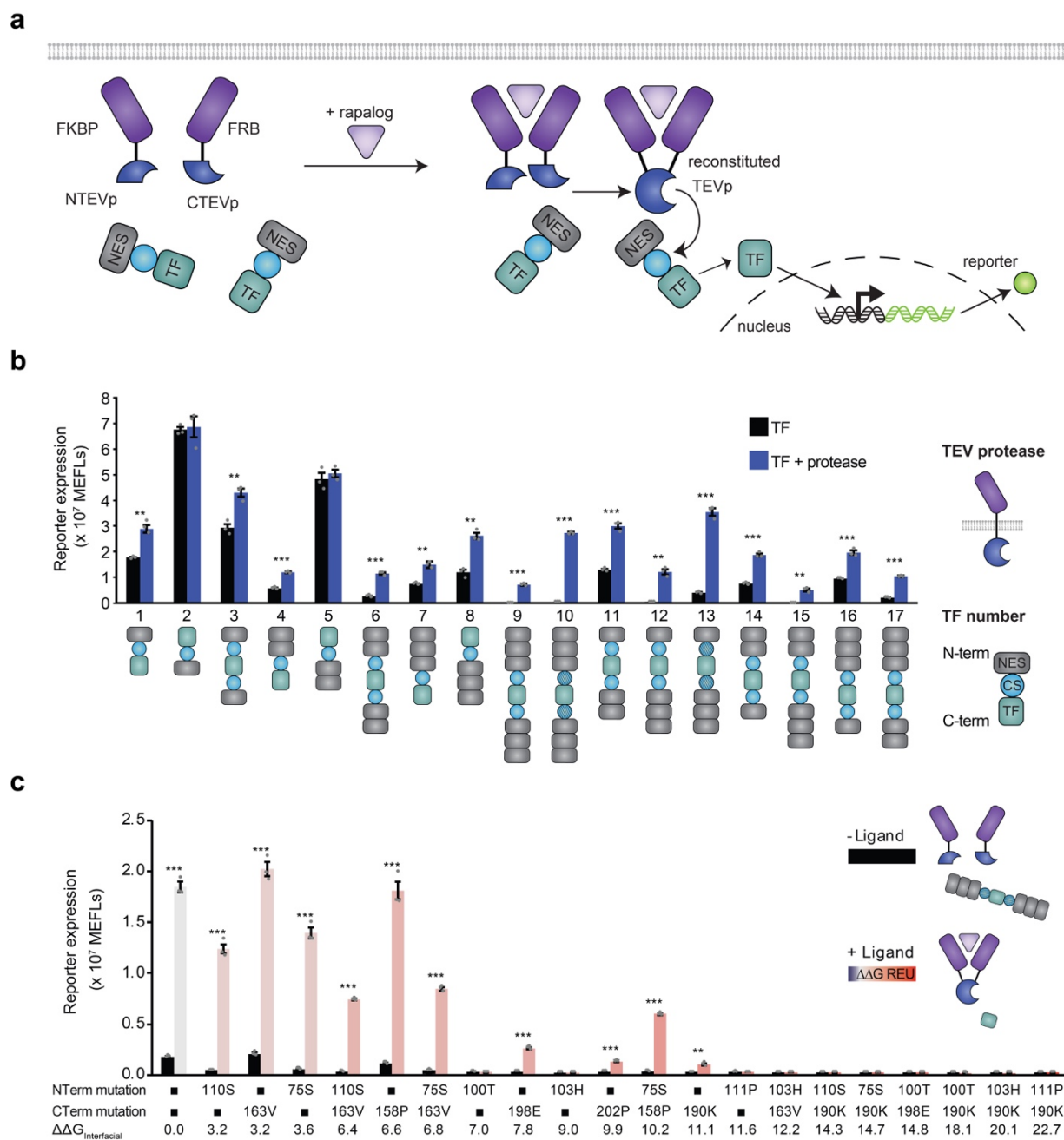
Fig. 5 Model guided design of a new split TEVp application in soluble context. a, This cartoon illustrates the soluble split TEVp testbed. Ligand-binding-induced dimerization mediates reconstitution of split TEVp, which then cleaves one or more nuclear export sequence (NES) elements from a soluble transcription factor, leading to nuclear import and reporter expression. **b,** Developing the testbed by evaluating engineered transcription factors (TF) for consistency with the mechanism proposed in **a**; shaded cleavage sequence (CS) domains indicate a G residue in the P1' position, unshaded CS domains indicate

549   a M residue in this position.[24] **c,** Experimental analysis of single and paired mutants sampling a range of

550   interfacial energies (indicated by color and labeled), employing TF10 from **b**. Error bars depict S.E.M. (*p ≤

551   0.05, ***p ≤ 0.001).

552

a Tools for optimizing split protein systems

Generalizability

| Random mutagenesis | Unmet need |
| Alternative split sites | Tune existing system |

Probability of success

Application 1
Application 2

Function

Unmet need

Proteins

Reconstitution propensity

b Experimental testbed

FKBP   FRB

NTEVp   CTEVp
         CS
         TF

+ rapalog

1
2
3
4
reconstituted TEVp
TF
nucleus
reporter

1. Dimerization
2. TEVp reconstitution
3. Cleavage
4. TF release

c

Identify interfacial residues

N-terminus   Interface   C-terminus

Computational interface scanning

Interface position

Destabilizing

Stabilizing

A   C   ...   Y
Amino acid identity

Calibration with experimental data

Total energy (ΔΔG)

Interfacial energy (ΔΔG)

Predictions

Total energy (ΔΔG)

Interface energy (ΔΔG)

Dead
Dead or inducible
Not inducible
Inducible

a

Interfacial residues

NTEVp

CTEVp

Catalytic triad
Coordination sphere

b

Interfacial energy (ΔΔG REU)

Pos 103

c

- Ligand

+ Ligand

ΔΔG REU

NTEVp mutation	■		■		■	■	110S	75S	75T	■	75E	■	■	103R	101P	103Y	111P	■	110P	103W
CTEVp mutation	■	121N	121A	163V	158S	■	■	■	121S	■	158P	190K	163K	■	■	163P	■	163R	■	■
Interfacial energy (ΔΔG REU)	0.0	3.1	3.1	3.2	3.2	3.2	3.6	4.3	4.3	5.6	6.6	11.1	11.2	11.2	11.2	11.5	11.6	11.6	12.1	12.1	16.1

d

Dead
Inducible
Not inducible

e

Dead

Dead or inducible

Not inducible

Inducible

a

Dead / Inducible / Not inducible

Total energy (ΔΔG REU) vs Interfacial energy (ΔΔG REU)

b

Reporter expression (× 10⁷ MEFLs)

− Ligand / + Ligand / ΔΔG REU

NTerm mutation	75S	75T	75E	111P	75S	111P	111P	111P	111P	75T	110S	111P	75E	75E
CTerm mutation	158P	158P	158S	158S	190K	121S	163V	121A	121N	190K	163P	121S	190K	163P
Interfacial energy (ΔΔG REU)	10.0	10.7	12.0	14.5	14.6	14.7	14.8	14.8	14.8	15.3	15.4	16.0	16.6	17.3

c

Relative frequency vs Experimentally observed phenotype

Not inducible / Inducible / Dead

d

Reporter expression (× 10⁷ MEFLs)

− Ligand / + Ligand / ΔΔG REU

NTerm mutation	■	75S	75E
CTerm mutation	■	190K	190K
Interfacial energy (ΔΔG REU)	0.0	14.6	16.6

a

OFF State | ON State | Fold Induction

b

a

b

c

TEV protease

TF number

N-term — NES
CS
C-term — TF

- Ligand
+ Ligand
ΔΔG REU

NTerm mutation	■	110S	■	110S	■	75S	■	100T	■	103H	■	75S	■	111P	103H	110S	75S	100T	100T	103H	111P
CTerm mutation	■	■	163V	163V	■	158P	163V	■	■	■	198E	■	202P	158P	■	190K	■	163V	190K	190K	198E	190K	190K	190K
ΔΔG_Interfacial	0.0	3.2	3.2	3.6	6.4	6.6	6.8	7.0	7.8	9.0	9.9	10.2	11.1	11.6	12.2	14.3	14.7	14.8	18.1	20.1	22.7