Systematic conformation-to-phenotype mapping via limited deep-sequencing of proteins

Eugene Serebryany[1] *, Victor Y. Zhao[1], Kibum Park[1], Amir Bitran[1], Sunia A. Trauger[2], Bogdan Budnik[2], and Eugene I. Shakhnovich[1] *

[1]Department of Chemistry and Chemical Biology and [2]Center for Mass Spectrometry, Harvard University, Cambridge, MA

Summary

Non-native conformations drive protein misfolding diseases, complicate bioengineering efforts, and fuel molecular evolution. No current experimental technique is well-suited for elucidating them and their phenotypic effects. Especially intractable are the transient conformations populated by intrinsically disordered proteins. We describe an approach to systematically discover, stabilize, and purify native and non-native conformations, generated *in vitro* or *in vivo*, and directly link conformations to molecular, organismal, or evolutionary phenotypes. This approach involves high-throughput disulfide scanning (HTDS) of the entire protein. To reveal which disulfides trap which chromatographically resolvable conformers, we devised a limited deep-sequencing method for proteins that identifies the exact sequence positions of both cysteines simultaneously within each polypeptide molecule. HTDS of the abundant *E. coli* periplasmic chaperone HdeA revealed distinct classes of disordered hydrophobic conformers with variable cytotoxicity depending on where the backbone was cross-linked. HTDS can bridge conformational and phenotypic landscapes for many proteins that function in disulfide-permissive environments.

Introduction

Crystallography, NMR, and cryoEM have revealed high-resolution native structures of some 50,000 proteins. However, intrinsically disordered, fold-switching, and kinetically trapped proteins can populate multiple physiologically relevant conformers (Ascenzi and Gianni, 2013; Borgia et al., 2015; Brockwell and Radford, 2007; Dishman and Volkman, 2018; Gershenson et al., 2020; Porter and Looger, 2018; Uversky, 2019) with distinct functions and phenotypic effects (Datta et al., 2008; Gautier et al., 2020; Nussinov et al., 2019). *In-vivo* conformations can also differ from *in-vitro* ones (Guin and Gruebele, 2019; Hingorani and Gierasch, 2014; Smith et al., 2016). Misfolding often leads to loss of fitness (Geiler-Samerotte et al., 2011; Wu et al., 2022). Notably, some. However, mapping conformational landscapes to phenotypic or fitness landscapes *in vivo* is extremely challenging because non-native (including disordered) conformations can rarely be purified or crystallized and do not leave clear signatures in the multiple sequence alignments on which computational methods like AlphaFold and RosettaFold rely. Deep mutational scanning can reveal complex *in-vivo* mutational fitness landscapes in sequence space (Jones et al., 2020; Mayor et al., 2016; Sarkisyan et al., 2016; Wu et al., 2022), but it cannot probe conformational space. The challenge is especially acute for intrinsically disordered proteins (IDPs), which may fold up *in vivo* (Leuenberger et al., 2017; Metskas and Rhoades, 2020) or adopt a plethora of transient backbone conformations (**Figure 1a**).

Disulfide bonds encode 3D conformational constraints directly in a protein's sequence. Disulfides easily form in many environments *in vivo* and, thanks to the action of disulfide isomerases, generally do not perturb protein conformations but stabilize them (Kosuri et al., 2012). Disulfide scanning mutagenesis (determining which double-Cys variants of a protein can form intramolecular disulfides) is a well-established technique for testing

structural models – e.g., elucidating the angle and register between two α-helices (Butler and Falke, 1998; Krshnan et al., 2016; Molnar et al., 2014; Taguchi et al., 2018). However, the number of double-Cys variants scales as the square of polypeptide length, so scanning an entire protein has never been practical. We report a high-throughput disulfide scanning (HTDS) methodology that surmounts this key limitation, providing a qualitatively new capability: mapping a protein's conformational landscape onto phenotypic landscapes.

To our knowledge, this study is the first implementation of HTDS even at the DNA level (via amplicon deep sequencing of pooled double-Cys mutant libraries). However, only single-molecule deep sequencing of entire proteins can reveal which double-Cys variants in a pooled library formed disulfides and which did not. Despite impressive recent progress (Alfaro et al., 2021), it is not yet feasible. We therefore devised our own protein deep-sequencing method that fulfills the more limited requirements of HTDS: precisely determining unknown positions of two Cys residues within the same polypeptide molecule at the same time. We achieved this by site-specific polypeptide backbone cleavage at Cys positions via cyanylation-aminolysis chemistry (Jacobson et al., 1973; Wu and Watson, 1997), generating "middle" peptides whose termini, assignable by MS/MS, are the two Cys positions.

As proof of concept, we applied HTDS to a highly abundant *E. coli* periplasmic chaperone, HdeA. One of the few known IDPs in bacteria (Link et al., 1997; Liu et al., 2004), HdeA is a holdase that inhibits aggregation of periplasmic proteins as the bacterium passes through stomach acid on its way to the gut of a new host (Gajiwala and Burley, 2000; Hong et al., 2005; Stull et al., 2018). Its functional (low-pH) conformation is largely disordered (Hong et al., 2005; Tapley et al., 2009), but at neutral pH it folds into an inactive dimer, with four α-helices and a disordered N-terminal region in each subunit (Gajiwala and Burley, 2000; Yu et al., 2017). That HdeA folds up when inactive suggests the unfolded state may have a fitness cost, although an *in-vivo* crosslinking study questioned whether the inactive state is stably folded (Fu et al., 2019). A conserved disulfide links the only two Cys residues (18 and 66) in the native sequence; its reduction disorders the structure and inhibits chaperone activity *in vitro* (Tapley et al., 2009; Zhai et al., 2016), though not entirely (Aguirre-Cardenas et al., 2021). *In-vivo* consequences of a lost or shifted disulfide have not been investigated; we now report that dose-dependent cytotoxicity typically results.

We report relative cytotoxicity of 1,453 double-Cys variants of HdeA on the C18S/C66S ("noC") background, expressed with the native periplasm-targeting sequence; redox state and hydrophobicity of hundreds of variants at the protein level; and low-throughput biophysical characterization for select variants. We hypothesized that introducing disulfides consistent with WT folded structure into the disordered noC might rescue structure and mitigate cytotoxicity. Surprisingly, very few disulfides rescued the toxicity of noC, but many disulfides greatly enhanced it. Toxicity was associated with increased hydrophobicity *in vitro* and protein misfolding stress *in vivo* with DnaK overexpression prior to cell lysis. Since HdeA has no known function at neutral pH, our observations provide clear evidence of misfolding-induced gain of toxicity. We conclude that the WT HdeA conformation is dependent on its 18-66 disulfide, while a well-defined subset of ectopic disulfides favors disordered yet highly toxic backbone conformations. Our methodology can be directly applied to many other IDPs and paves the way for applications to larger well-folded proteins.

Results

Disulfide-scanning an entire protein is now feasible

Our HTDS workflow consists of two branches (**Figure 1b**): DNA deep sequencing to obtain a fitness/toxicity landscape and protein deep-sequencing to identify disulfide-forming Cys pairs or compare variant abundances in distinct chromatographic fractions. A chemical thiol blocker distinguished variants with free vs. disulfide-forming Cys residues, and subsequent reduction of disulfides and cyanylation of the newly free Cys residues allowed site-

specific backbone aminolysis for LC/MS/MS (**Figure 1c**). Subsequent sections (see **Figures 4 and 7**) illustrate the possibilities this novel method offers.

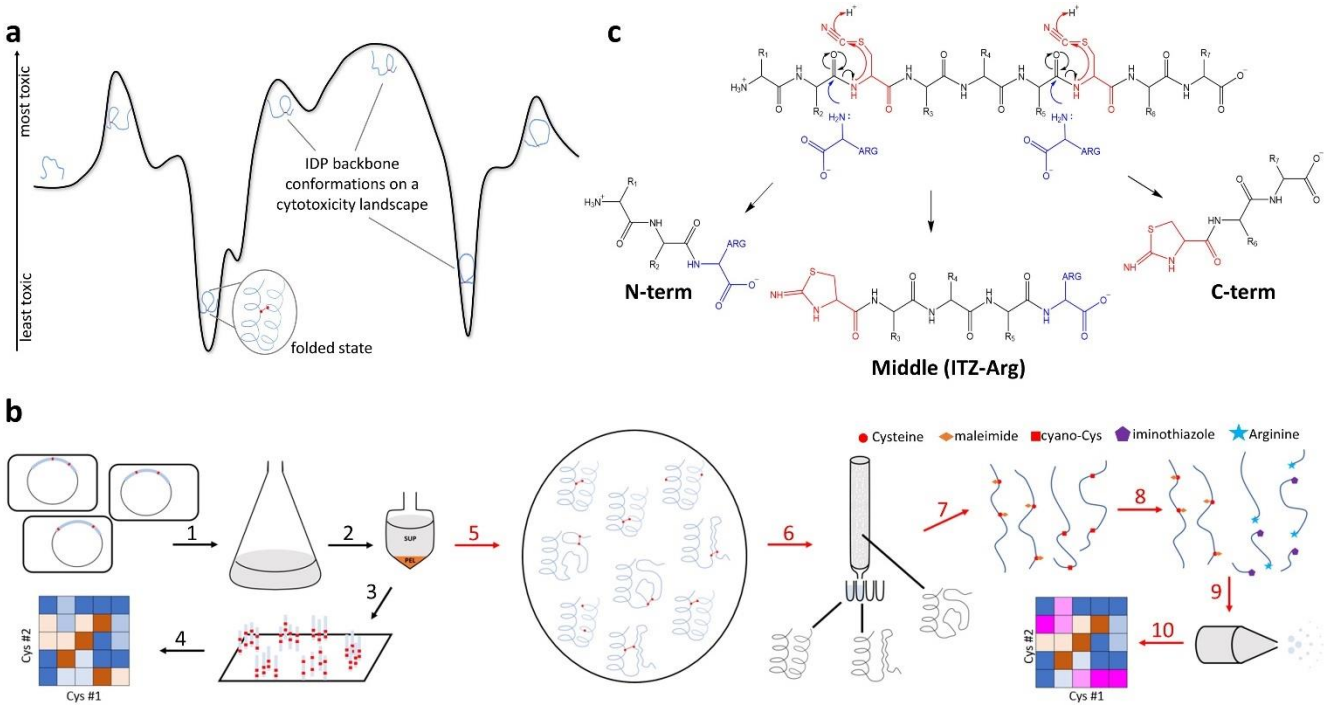

Figure 1: The concept and implementation of conformation-to-phenotype mapping by HTDS. (a) In an IDP, distinct backbone conformations may produce distinct molecular and organismal phenotypes, such as cytotoxicity. **(b)** Overview of HTDS, starting from bacterial transformation with a pooled double-Cys variant library of the target gene. *Black arrows* indicate phenotype-mapping steps: (1) pooled culture with induction of protein expression; (2) pellet/supernatant separation, during which plasmids that encode lysogenic variants are released into the supernatant; (3) pooled Illumina paired-end sequencing of indexed barcoded amplicons from the pellet and supernatant fractions; (4) resulting phenotypic landscape based on variant abundance ratios in the DNA library fractions. *Red arrows* indicate conformation-mapping steps: (5) extraction and partial purification of the protein library, where some variants form disulfides, and some disulfides stabilize distinct conformations; (6) chromatographic fractionation of the library, e.g., by surface hydrophobicity; (7) denaturation of fractionated proteins with thiol blocker (*orange kites*) to passivate non-bonded Cys residues, then reduction and cyanylation of disulfide-bonded Cys residues (*red squares*); (8) aminolysis, e.g., by free arginine, to yield three fragments for each polypeptide: N-terminal (which incorporates the nucleophile at its C-terminus), C-terminal (which forms an iminothiazole at its N-terminus), and the middle peptide (which does both); (9) precise identification of variants from middle peptides by LC/MS/MS; (10) resulting conformational (or other biophysical) landscape based on variant abundance ratios in the protein library fractions. **(c)** Cyanylation-aminolysis is currently the key enabling chemistry for HTDS. At pH 9, the amine of free Arg attacks the carbonyl immediately upstream of cyano-Cys, a good leaving group thanks to internal cyclization to iminothiazole (ITZ).

Ectopic disulfides make HdeA cytotoxic

We expressed a multi-Cys scanning library (~½ double-Cys variants) from the pCK302 plasmid, which has a well-repressed and rhamnose-titratable rhaBAD promoter (Kelly et al., 2016), in the non-rhamnose-metabolizing *E. coli* strain BW25113 Δ*hdea*. To facilitate initial quality checks and subsequent MS/MS-based sequencing, the library was enriched for variant 40/65 and all variants containing Cys66, including WT (see Methods). Eleven variants (most chosen randomly, some as controls) were grown in 96-well plates with varying [rhamnose]. We observed pronounced, dose-dependent dips in the growth curves of most HdeA mutants, but not superfolder GFP or WT HdeA, at the onset of stationary phase (**Figure 2a**), indicating that many double-Cys HdeA variants were cytotoxic. All cultures in **Figure 2a** were inoculated identically and measured in parallel. Using tenfold smaller

inocula altered the shapes of the growth curves, but the variant-dependent toxicity remained, and the rank order of variants was similar (**Figure SI 1**).

All 11 HdeA variants were well expressed at high induction (though less well than sGFP), particularly when starter cultures were grown in pH 8 MOPS-buffered LB (**Figure 2b**). Expression in unbuffered LB was more variable (**Figure SI 2**) but still tunable by varying [rhamnose] (**Figure 2c**). Faster migration of WT (18/66), 11/73, 33/86, and 18/86 relative to noC, 32/66, 40/65, etc., in non-reducing SDS-PAGE (**Figure 2b**) was consistent with compaction of the denatured state by the longer-range disulfides. Variant 32/66 accumulated less than WT at low [rhamnose] but more than WT at high [rhamnose] (**Figure 2d**), suggesting it may be degraded at the lower expression levels. Comparing total expression cultures to identical volumes of clarified supernatants of the same cultures yielded two unexpected observations (**Figure 2b**). First, HdeA was found mostly in the supernatant (unlike sGFP). We are not aware of prior studies of whether HdeA is exported from the cell. All constructs contained the periplasmic targeting sequence of native *E. coli* HdeA. Only the HdeA band was visible in the WT HdeA supernatant, so export of this protein was not via cell lysis. By contrast, the many protein bands in supernatants of HdeA mutants indicated extensive cell lysis, which explains the dips in their growth curves (**Figure 2a**). Second, the lysogenic variants triggered strong overexpression of an endogenous 70 kD protein (marked on the gel in **Figure 2b**). Trypsinization and LC/MS/MS of this band assigned it as DnaK (with 125 peptide spectrum matches (PSM) for DnaK, compared to 29 PSM for the second-most abundant protein, which was trypsin). Quantifying the band intensities (**Figure 2e**) suggested a DnaK/HdeA ratio converging to ~0.25 for both noC and 32/66. We conclude that cytotoxic HdeA variants triggered a protein-misfolding response in their host cells.

The die-off in stationary phase was typically followed by recovery (**Figure 2a**). Further investigation (**Figures SI 3 and SI 4**) revealed that the recovery was driven by cells that had turned off the toxic variant's expression. Sanger sequencing of miniprepped plasmids from the recovered cultures confirmed that the plasmids were retained, including the HdeA mutations. More detailed study of the expression shut-off mechanism is beyond the scope of this work; here, we focused on the initial die-off as a convenient readout of fitness costs (cytotoxicity) associated HdeA double-Cys variants.

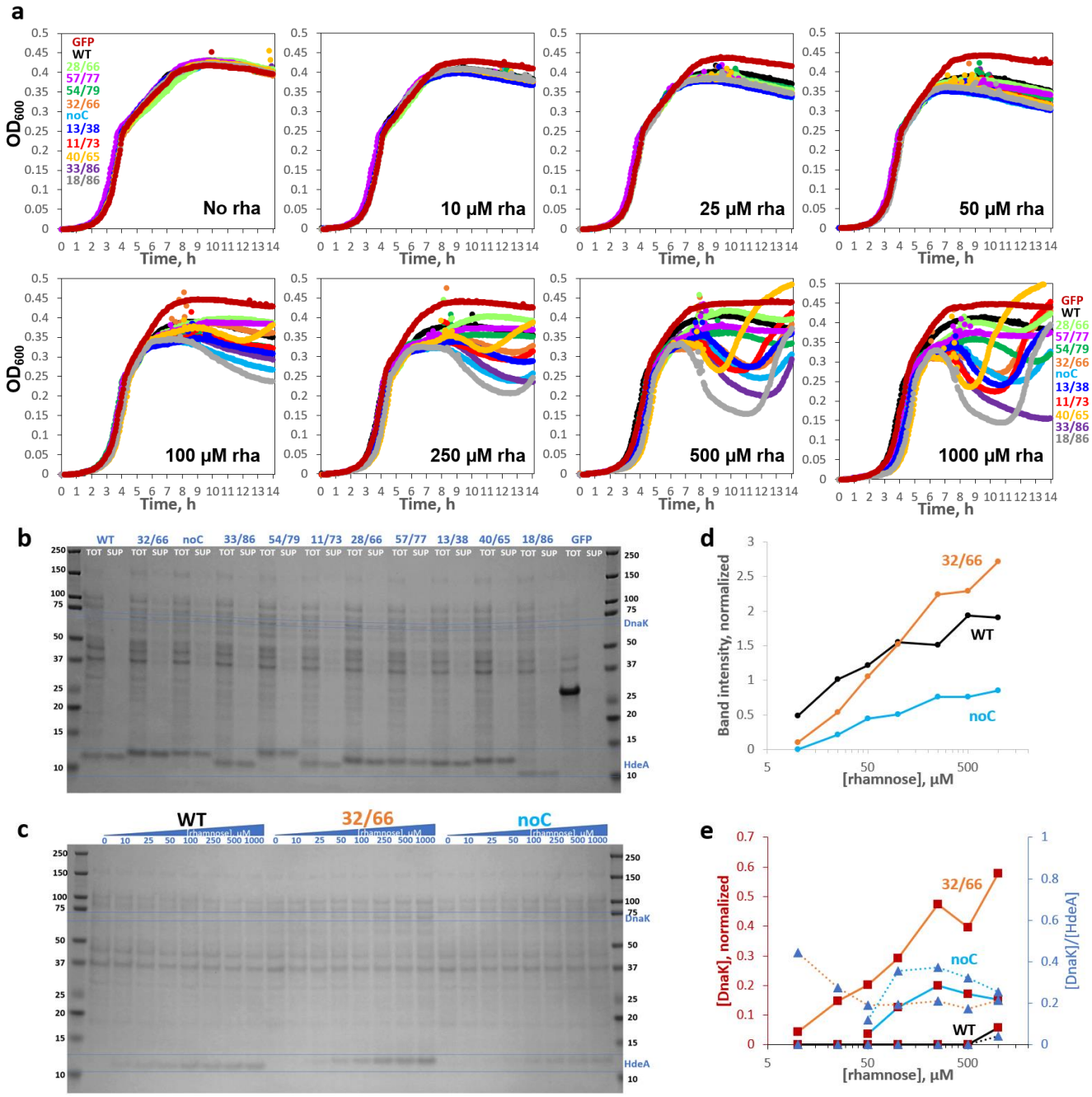

Figure 2: HdeA variants exhibit varying levels of toxicity characterized by DnaK overexpression and cell lysis. (**a**) Growth curves of HdeA variants in standard LB broth with varying [inducer], measured in parallel in the same 96-well plate, with all wells for each variant started from the same inoculum, showed variant- and [inducer]-dependent drops in OD_{600}. WT HdeA and sGFP did not. (**b**) Non-reducing SDS-PAGE of end-point samples from 2500 mM rhamnose cultures showed clear expression of all constructs, albeit weaker than sGFP. Variants with more sequence-distal Cys pairs migrated lower on the gel, consistent with greater compaction of the unfolded state by longer-range disulfides (note rightward skew of the gel). Juxtaposing identical volumes of centrifuged cell culture supernatant ("SUP") and total culture ("TOT") samples showed 86±11% (mean±S.D.) of HdeA in SUP across variants. All mutants showed many cytoplasmic protein bands in SUP, indicating cell lysis, and DnaK overexpression (confirmed by MS/MS). DnaK's prominence in SUP indicated that it did not prevent lysis. (**c**) Non-reducing SDS-PAGE of the WT, 32/66, and noC constructs (TOT) as a function of [rhamnose]. (**d**) Quantitation of the HdeA bands in **c**, internally normalized to the abundant 37 kD band, showed a crossover between WT and 32/66 abundance. (**e**) Quantitation of the DnaK bands in panel **c** for 32/66 (*orange line*) and noC (*light-blue line*); observed [DnaK]/[HdeA] ratios converged to ~0.25.

In total, 3,423 variants had at least 50 full paired-end Illumina reads on average across pellet ("PEL") and supernatant ("SUP") samples of four replicate cultures at 9.0 h post-induction (see Methods). Of these, 1,453 were double-Cys variants that passed the quality check for outliers. 76 of 85 theoretically possible single-Cys variants were likewise detected. Variant noC comprised 3.15 ± 0.01% (mean ± S.E.M.) of PEL reads and 2.80 ± 0.02% of SUP reads – a final SUP/PEL abundance ratio of 0.89. All single-Cys variants combined comprised 24.73 ± 0.02% of PEL and 23.76 ± 0.10% of SUP reads. All double-Cys variants combined comprised 46.03 ± 0.07% of PEL and 47.72 ± 0.17% of SUP. The balance was mostly sequences with three Cys or a non-Cys mutation.

HdeA has no native function at neutral pH, hence no loss-of-function mutations. Yet, many double-Cys variants clearly exacted a fitness cost due to protein misfolding stress and cell lysis. To build a cytotoxicity (lysogenicity) landscape for the 1,453 double-Cys variants, we defined a variant's raw toxicity *R* as its SUP/PEL ratio and epistatic fitness cost *E* as *R* corrected for single-Cys effects to isolate the effect of Cys-Cys interaction – i.e., of the putative disulfide (for both *R* and *E*, higher means more toxic):

$$E = \frac{R_{doubleCys}}{R_{Cys\#1} R_{Cys\#2}} \quad ,$$

Gratifyingly, either measure put WT's toxicity in the bottom 1%. In **Figure 3a** and throughout this text, we express *E* for each variant as the number of S.E.M. away from noC (for which *R* = *E* = 0.89) to reflect both the amplitude and the confidence of variant fitness costs. However, using the more conventional log ratio epistasis (Rollins et al., 2019) yields qualitatively the same map (**Figure SI 5a**). *R* was lower than *E* for many variants with a Cys in the natively disordered N-terminal region due to single-Cys effects (**Figure SI 5b**).

Our initial hypothesis was that disulfides consistent with the WT tertiary structure would rescue misfolding-associated cytotoxicity, but the observed landscape differed markedly from the WT's contact map (**Figure 3b**): most double-Cys variants were neutral or toxic. Cumulative average of *E* did show a clear signature of reduced toxicity for variants whose Cys residues had short native C_α-C_α distances or were close to their WT (18, 66) sequence positions (**Figure 3c**). Yet, the effect was surprisingly small, the cumulative averages barely rising above noC. Of the 44 variants with native C_α-C_α distances < 4 Å above WT (averaged for the two subunits in PDB 5WYO), the least-toxic variant was WT itself (18/66, at +5.5 SEM relative to noC), followed by 17/66 (+2.4 SEM), 13/73 (+1.8 SEM), 17/70 and 17/73 (both at +1.3 SEM), 13/75 (+0.8 SEM), 65/81 (+0.5 SEM), and 17/72 (+0.1 SEM). The remaining 38 variants in this set were all more toxic than noC, including 18/65 (-1.5 SEM) and 18/67 (-1.0 SEM). Nor did shifting the disulfide by one helical turn restore fitness: 18/70 was at -0.5 SEM to noC, and 22/66 at -3.4 SEM. Meanwhile, swathes of double-Cys sequence space were highly cytotoxic, most notably a rough rectangle at positions 27-52 (Cys#1) X 63-66 (Cys#2) that included the previously identified toxic variants 32/66 and 40/65. Thus, in terms of phenotype, the native state of HdeA appears to be highly brittle: moving the disulfide from its native location – even when the new location remains consistent with the WT 3D structure – is not well tolerated by the organism.

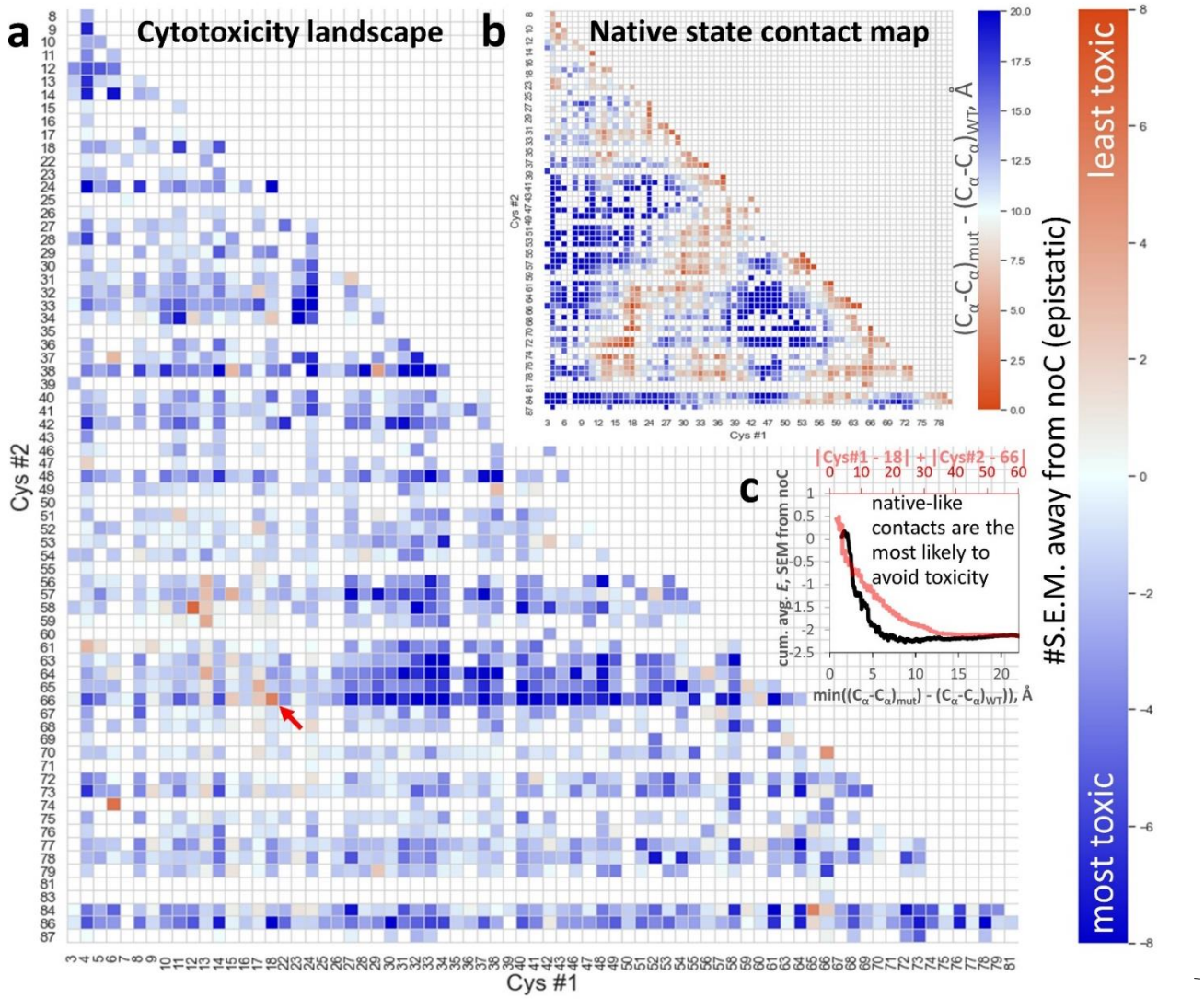

Figure 3: Cytotoxicity landscape of 1,453 double-Cys variants. **(a)** A fitness landscape was constructed using pairwise fitness epistasis for all 2-Cys variants that could be quantified in pooled paired-end sequencing of amplicons generated from pellet and supernatant samples of quadruplicate cultures of the library grown with 1000 mM rhamnose. Data are reported here as the difference of means of *E* for the given variant and noC, divided by the S.E.M. of the variant. Positive difference (*orange*) indicates less toxicity (cell lysis) relative to noC; negative difference (*blue*) indicates worse toxicity than noC. Very few variants were observed to be less toxic than noC. The WT gene (18/66, indicated by *red arrow*) ranked #3 out of 1,453 by this metric. **(b)** A pairwise contact map of C_α-C_α distances with respect to WT, derived by averaging such distances in chain A and chain B of PDB ID 5WYO. **(c)** Cumulative average epistatic fitness (expressed as in panel **a**) as a function of the pairwise native-state C_α-C_α distance (*black*) or sequence distance (*red*) from WT for all double-Cys variants with Cys at least 12 peptide bonds apart showed lower cytotoxicity for Cys pairs with either native-like C_α-C_α distances or native-like sequence positions.

However, strong epistasis suggests but does not prove disulfide bonding; weak epistasis may result from disulfides that stabilize benign conformations or simply fail to form. Proper interpretation of the DNA-level experiments requires knowing whether a given pair of Cys forms a disulfide *in vivo*. Therefore, we extracted the pooled protein library from the periplasm of the expressing cells, partly purified it (see Methods), and determined disulfide bonding status experimentally for all variants with peptides detectable at <10% false discovery rate (FDR). The library was split into treatment and reference samples: in the former, free thiols were blocked by N-ethyl maleimide (NEM), before reduction and cyanylation (as in **Figure 1b**); in the latter, NEM was not used, so every Cys was cyanylated. We confirmed by SDS-PAGE that NEM blocking of Cys residues prevents backbone

cleavage at those sites (**Figure SI 6**). Wide variation in DNA-level abundance in the library (see above) allowed only ~11% of the double-Cys variants to be detected by LC/MS/MS (165 out of 1,453). Almost all detected variants did form disulfides *in vivo* (**Figure 4a**), and the rest were mostly of low abundance and could have been missed by chance (**Figure 4d,e,f**). Disulfide bonding was not always all-or-none, consistent with the gel in **Figure 2b**, where the highly toxic 18/86 variant had a prominent downshifted band (expected given its long-range disulfide) but also a faint band at the "noC" position. The average estimated *in vivo* disulfide bonding propensity across all variants was just below that of WT HdeA, which is fully disulfide-bonded (**Figure 4b,c**).

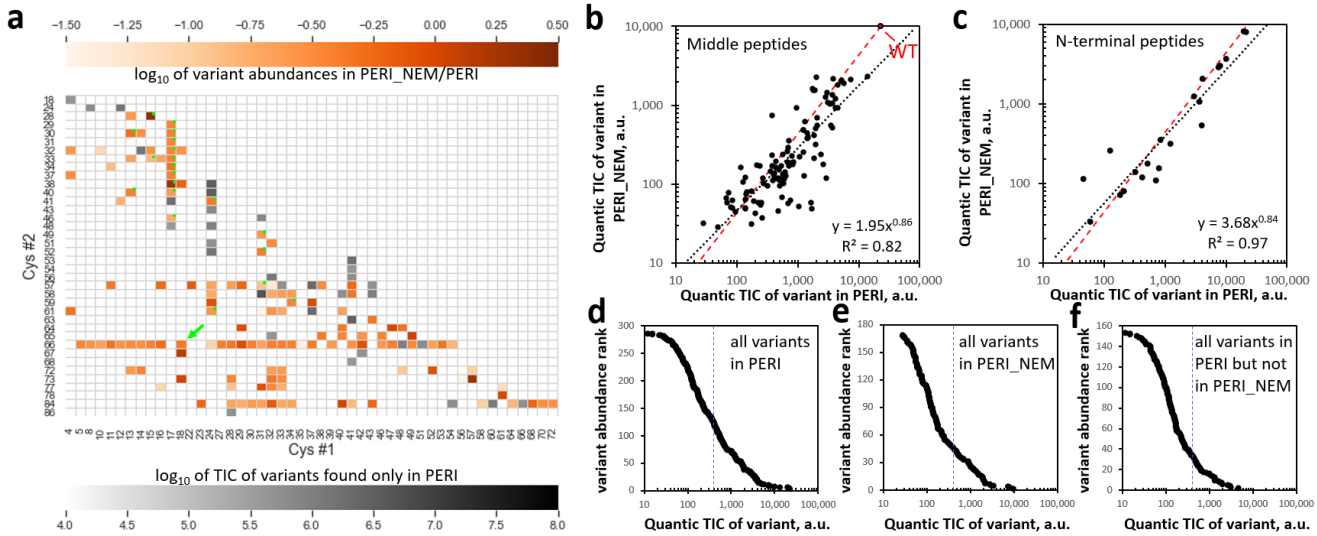

Figure 4: Disulfide-bonding propensities of 165 double-Cys variants. (a) Heatmap showing the abundance ratios (by total ion current) for variants found in NEM-treated ("PERI_NEM") and untreated ("PERI") periplasmic samples. Abundance of the most abundant variants found only without NEM is in gray. The *green arrow* indicates WT. Variants identified exclusively from N-terminal missed-cyanylation peptides are marked with *green dots*. **(b)** Abundances in PERI and PERI_NEM (quantified by total ion current of the middle peptides, using Quantic) were strongly correlated, with a linear-regression trendline (*black dashed line*) close to the ratio measured for WT (*red dashed line*), so the vast majority of variants formed disulfides. **(c)** Same as **b** but for variants identified from N-terminal missed-cyanylation peptides only. **(d)** Log-linear plot of all variants in PERI ranked by abundance revealed a sigmoidal distribution of abundance. **(e)** Same as **d** but for PERI_NEM. **(f)** Same as **d** but for variants found in PERI and not in PERI_NEM, showing that most were low-abundance variants (weak total ion currents). A blue dashed line at 400 a.u. is shown in all three plots (**d-f**) for reference.

Absent or ectopic disulfides promote dissociation of folded HdeA dimers to disordered monomers

To investigate the structural basis of toxicity, we used atomistic Monte-Carlo protein unfolding (MCPU) simulations with a knowledge-based potential; this freely available software was developed in our lab and shown to be effective in folding up small proteins (Yang et al., 2007) and predicting stability effects of mutations (Tian et al., 2015), effects of non-native disulfides on aggregation (Serebryany et al., 2016), and protein free energy landscapes (Bitran et al., 2020) (version used here can be found at https://github.com/proteins247/dbfold). We chose 50 double-Cys variants and noC for multiplexed temperature-replica exchange simulations. Variants were selected based on preliminary experiments, spanning various levels of toxicity and regions of sequence space. Simulations began from WT homodimer (PDB ID 5WYO). Intramolecular disulfides were modeled as strong flat-bottomed harmonic restraints (Serebryany et al., 2016). Sampled structures from all simulated variants were clustered together by intermolecular contact maps using DBSCAN (Ester et al., 1996) to find conformations with similar intermolecular interactions.

Structures from all variants fell into just three main clusters, plus numerous smaller clusters and non-clustering structures (**Figure 5a**). The largest cluster, "dissociated," comprised structures with few or no intermolecular contacts between the subunits. The second-largest, "C-term," comprised non-native dimeric structures, in which the subunits' C-termini unexpectedly bound each other as antiparallel helices. The third-

largest, "WT-like," largely preserved the native homodimer interface. Most of the smaller clusters (963 of 1,029) had 28 or fewer structures each and thus could be unique conformations from a single coordinate replica of a single HdeA variant. A total of 7,484 structures were assigned as non-clustering.

Figure 5b shows how the cluster populations changed with simulation temperature. As expected, the dissociated cluster grew larger at higher temperatures. Interestingly, dissociated and non-clustering structures were fewest at intermediate Monte-Carlo temperatures where the C-term cluster was largest; we speculate that when both subunits lose their native structure, they may form the C-term structure before higher temperatures melt it, too. Representative main-clusters structures are shown in **Figure 5c**. The WT-like structure of 32/66 shows that certain non-native disulfides do not preclude native-like conformations, despite causing cytotoxicity. Many minor-cluster conformers were also observed in 32/66, however, and may together constitute a molten globule. Many "C-term" structures, like the 40/65 structure shown, had helices 3 and 4 of a subunit aligned into a single, long helix, with the rest apparently molten. The hydrophobic residues on helix 4, which were originally buried in the core, formed the new intersubunit interface. This kind of structure has not been experimentally observed, but it might explain some natively "forbidden" intermolecular crosslinks previously found *in vivo* (Fu et al., 2019).

Strikingly, many non-native disulfides forced dissociation of the dimer interface, even when neither Cys was in the main interface helix (**Figure 5d**). Dissociation did not occur in noC under the same simulation conditions (**Figure 5d**, *"None"*). The sum total of dissociated, minor-cluster, or non-clustering conformers likewise often exceeded that of noC (**Figure 5d**). This counterintuitive observation of disulfides promoting disorder, together with the observation that disulfides formed *in vivo* whether or not they were consistent with the native state (**Figure 4**), could help explain why most double-Cys variants were more cytotoxic than noC (**Figure 3**).

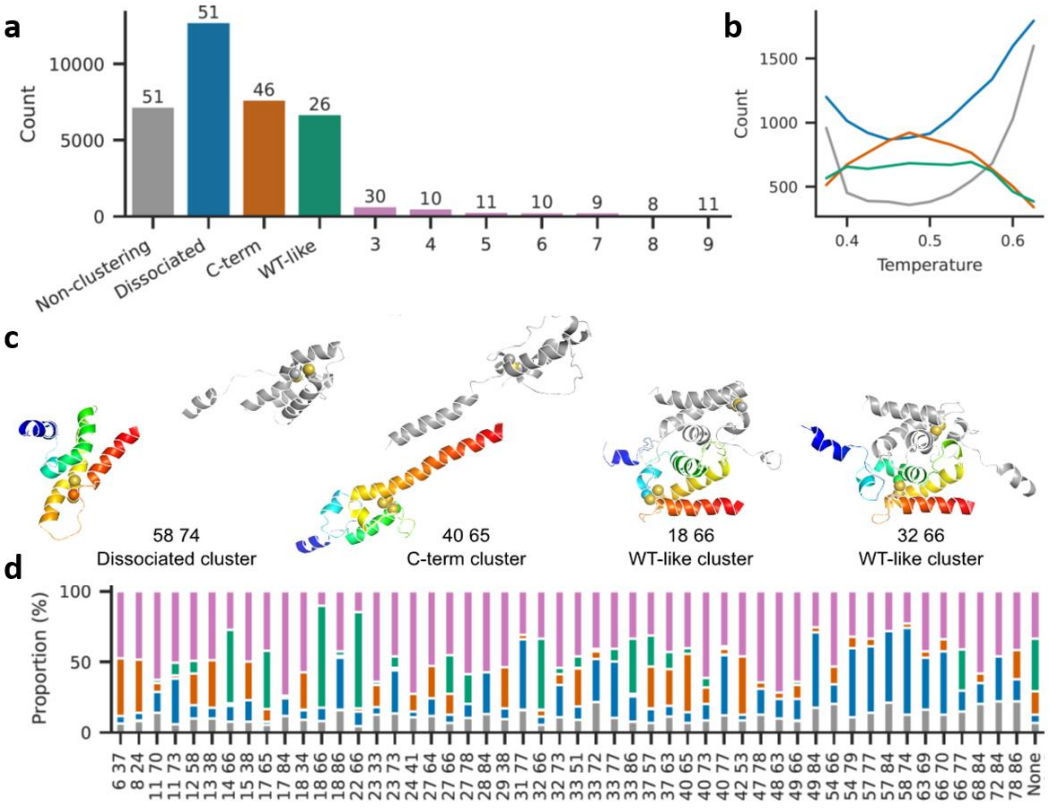

Figure 5: Many non-native disulfides cause dissociation of HdeA dimers and melting of monomers in atomistic simulations. (a) Combined sizes of top 10 clusters from all variant simulations; the number above each bar indicates how many variants had structures in that cluster. Total structures: 61600. Total clusters: 1032. Total structures belonging to none of the three main clusters: 29455. **(b)** Variation in cluster sizes with simulation temperature (colors as in **a**). **(c)** Representative structures from the three main clusters (numbers indicate Cys positions). The A subunit is in *rainbow colors* while the B subunit is gray. Cysteine residues are shown as spheres. Simulations included a 13-residue Gly-Ser linker between the subunits, here omitted for clarity. **(d)** Proportional cluster sizes for each simulated variant (colors as in **a**). "None" signifies the noC variant.

To validate these insights, we purified WT, noC, 32/66, and 40/65 and characterized them experimentally. The three mutants had quenched and red-shifted intrinsic tryptophan fluorescence spectra compared to WT (**Figure 6a**), indicating high solvent exposure of the two Trp residues (Trp16 and Trp 84). Gel filtration indicated similar hydrodynamic radii for WT and variants, with noC slightly more expanded and 32/66 and 40/65 slightly more compact. Yet, multiangle light scattering clearly indicated all three mutants were monomeric, unlike WT at the same concentration (**Figure 6b**). The WT yielded a circular dichroism spectrum consistent with the PDB structure (**Figure 6c**), but the mutants' spectra indicated much less helicity and predominantly random coil. This was even true of noC, unlike **Figure 5d**. Perhaps the need to connect subunits by a long linker in our simulations resulted in an overestimate of dimer stability; HdeA's net negative charge (-4.1 at pH 7 for noC) likely provides an extra driving force for dimer dissociation that MCPU's statistical potential does not fully take into account.

Despite their very similar CD spectra, 32/66 was significantly more hydrophobic than noC (**Figure 6d**). This increased hydrophobicity was completely eliminated by reduction (**Figure 6e**), so it was entirely attributable to the non-native disulfide bond. (WT's partially buried disulfide was not reduced under these conditions.) Hydrophobic interaction chromatography (HIC) easily separated 32/66 from WT due to the large difference in hydrophobicity (**Figure 6f**).

We additionally collected CD spectra of eight other double-Cys variants, of varying toxicity and from distinct regions of the sequence: 31/66 and 33/66 (to test whether a twist of the interface helix could rescue 32/66); 31/48 and 41/63 for comparison; low-toxicity variants 12/58, 13/59, and 6/74; and the near-neutral 17/68, whose Cys are close to the native 18/66 positions. CD, intrinsic fluorescence, and bisANS fluorescence spectra of all these variants (**Figure SI 7**) were similar to those in **Figure 6**. Thus, the native conformation appears to be highly fragile: it melts when the native disulfide is lost or even shifted by a few sequence positions.

To map predicted free energy landscapes of dissociated monomers with non-native disulfides, we used replica-exchange MCPU simulations of the monomers with umbrella biasing (by percent native contacts) and obtained unbiased statistics using MBAR in DBFOLD (Shirts and Chodera, 2008). We chose variants 32/66 and 18/32, which are highly hydrophobic (**Figure 7**) yet differ in their toxicity (**Figure 3**), and noC and WT monomers for reference. Fraction of native contacts Q and C_α root-mean squared deviation (RMSD) were defined with respect to monomeric equilibrated WT starting structure. Free energy landscapes (**Figure 6g**) showed a single native-like basin for WT but two basins for noC. The non-native basin was much more populated in the highly-toxic 32/66 variant than in the mildly-toxic 18/32 along the RMSD coordinate (**Figure 6i**), though not along the Q coordinate (**Figure 6h**). These findings suggest that 32/66 merely stabilizes a non-native cytotoxic conformational basin already present in the underlying free energy landscape. We also computed average equilibrium surface hydrophobicity of monomeric noC, 32/66, and 18/32 vs. dimeric WT simulated the same way (**Figure 6j**). The mutants were more hydrophobic than WT, and slightly more than noC, consistent with the experimental trend (**Figure 6d,e**).

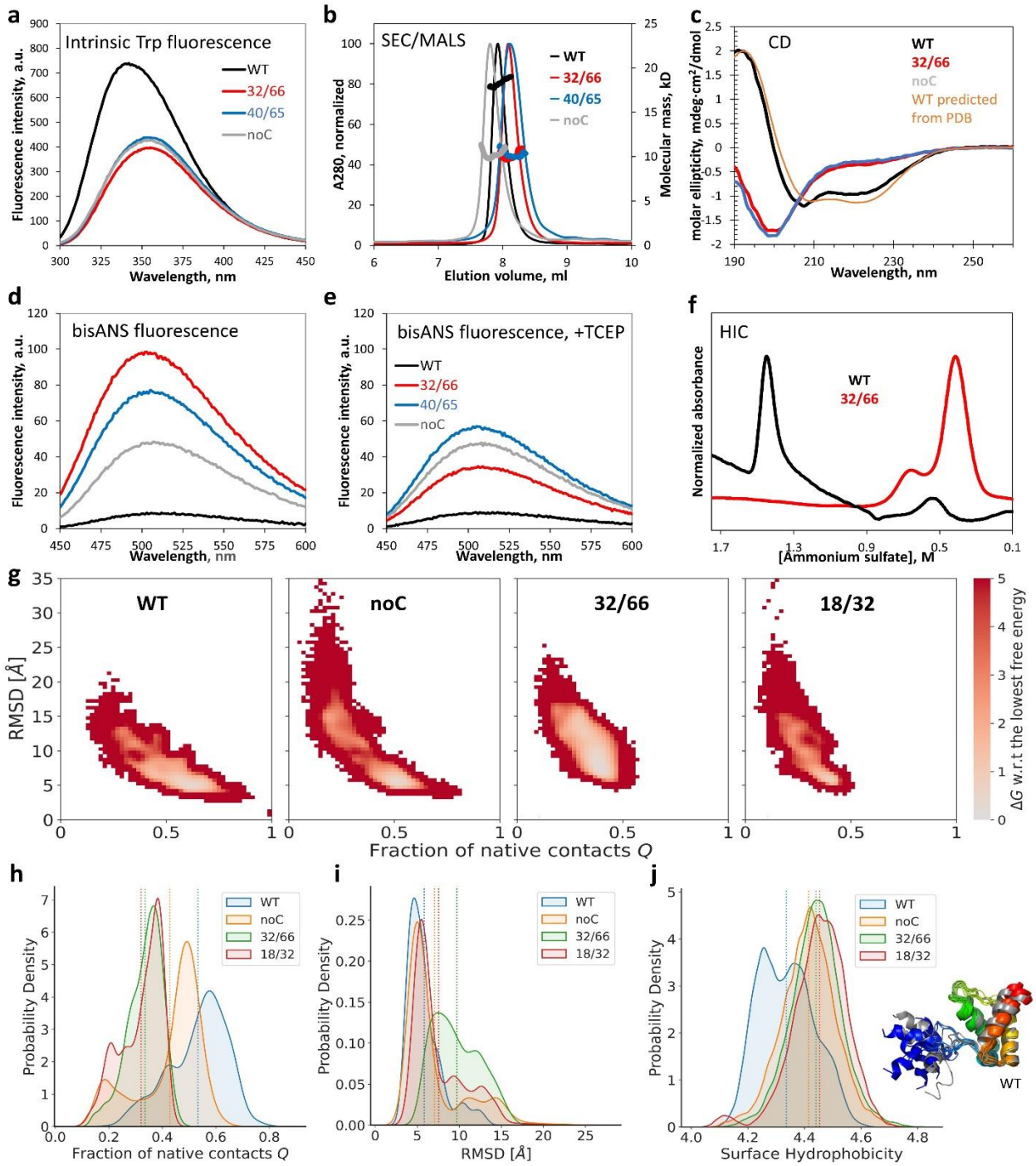

Figure 6: Biophysical characterization indicates HdeA mutants are monomeric, highly disordered, and hydrophobic. (a) Intrinsic tryptophan fluorescence was quenched and red-shifted in 32/66, noC, and 40/65 compared to WT. **(b)** SEC/MALS revealed that, despite similar elution positions, only WT was dimeric; all mutants were monomers. **(c)** CD of WT matched predictions from PDB ID 5WYO by the PDB2CD tool (https://pdb2cd.cryst.bbk.ac.uk), but 32/66 and noC clearly did not. **(d)** The hydrophobicity probe bisANS bound much more strongly to the mutants (same samples as in **a**), especially 32/66. **(e)** Upon reduction of the disulfides, 32/66 and 40/65 bisANS fluorescence resembled noC; WT was resistant to reduction. **(f)** Folded (WT) and disordered (32/66) were easily separated by HIC. (**g**) Calculated free energy landscapes of monomeric WT, noC, and two double-Cys variants from MCPU simulations; the monomeric WT starting structure is defined as Q = 1. (**h**) The distribution of Q-values (fraction of native contacts) for mutant and WT HdeA. (**i**) The corresponding distribution of RMSD values relative to the starting WT structure. (**j**) The distribution of surface hydrophobicity values for WT dimer and mutant monomers; although it appears bimodal for WT, representative structures of the WT subunits (*rainbow- vs.*

gray-colored ribbon diagrams) did not appreciably differ outside the N-terminal disordered region (*blue/gray*), which was often helical in our simulations. *Dashed lines* in panels **h-j** are distribution averages.

Cytotoxicity correlates with hydrophobicity even among intrinsically disordered core variants

To investigate how cytotoxicity relates to biophysical properties, we purified two protein library batches (expressed 2.5 months apart) from culture supernatants (see Methods). A sample from each was deep-sequenced as above. Using the empirical 10% FDR and analyzing only middle peptides, we identified 81 variants in batch 1 and 115 in batch 2. The overlap was 56 variants. Quantification using Quantic (Comet pipeline in TPP6) showed very good batch/batch agreement (**Figure SI 8**). The two batches were combined and fractionated by HIC on an ammonium sulfate gradient. In HIC, hydrophilic proteins elute first (at higher [ammonium sulfate]), followed by increasingly hydrophobic ones. A total of 339 variants were detected as above in at least one of eight HIC fractions (**Figure 7a**): 199 from middle peptides only, 104 from N-terminal missed-cyanylation peptides only, and 36 from both. (The 10% FDR threshold was applied to each fraction separately.) The low overlap was expected because MS/MS fragmentation typically fails for long peptides (>50-60 residues), and the maximum peptide length in Comet is 63 (**SI files 1,2**); thus, e.g., WT could never be identified from N-terminal missed-cyanylation because that peptide would be 66 residues long.

Due to technical limitations, especially variable DNA-level abundances in the library (see Discussion), just 39 variants were detected in at least four HIC fractions each, yielding interpretable elution profiles. Of these, 17 were classed as "core" variants, i.e., neither Cys residue was in the N-terminal flexible region (defined here as residues 1-17). They were identified mostly from middle peptides, and most of the others from N-terminal ones. We clustered the core variants into four types of elution profiles by hydrophobicity, defined as the centroid of abundances across the HIC fractions where the variant was found. WT and 70/84 were the least hydrophobic in this set, and the least toxic (**Figure 7b**). Moderately hydrophobic variants (**Figure 7c**) were typically moderately toxic. More hydrophobic variants (**Figure 7d**) were also more toxic, except 18/32, which resembled the "non-core" variants in **Figure 7f**, many containing Cys17: hydrophobic yet only modestly toxic. The most toxic variants had anomalous saddle-shaped elution profiles (**Figure 7e**), suggesting transient burial of their hydrophobics, either via folding or (more likely) via interaction with other proteins or even transient aggregation. **Table SI 2** lists the fitted hydrophobicity and measured cytotoxicity values of all 39 variants.

Most variants in **Figure 7** were detected also in **Figure 4** and confirmed to form disulfides. The exceptions were 37/61 and 31/58: mildly-hydrophobic, and non-disulfide-forming (**Figure 4a**), their toxicity was close to noC (-2.2 and -1.6 SEM, respectively). By contrast, 31/57 formed the disulfide at least some of the time per **Figure 4a** and was at -5.7 SEM from noC. Variant 24/49 (-1.5 SEM) was not detected in **Figure 4a** but was likely non-disulfide-bonding like the nearby 24/48 and 24/51. So, core variants with noC-like phenotypes had noC-like structures (i.e., no disulfide). The only hydrophilic mutant, 70/84, did form the disulfide but had noC-like toxicity (-1.0 SEM). All variants in **Figure 7f** also formed disulfides, so disulfides near either terminus may be less toxic.

We calculated the PSM-weighted average *E* of each HIC fraction and found a strong, apparently sigmoidal hydrophobicity-toxicity correlation for all 235 variants identified from middle peptides but none at all for the 104 variants identified from N-terminal peptides only (**Figure 7g**). Plotting hydrophobicity vs. toxicity of the 39 most-abundant variants, without any PSM-weighting, revealed a statistically significant ($p = 0.02$) correlation for the core variants, identified from either type of peptide (**Figure 7h**), which became even stronger ($p = 0.003$) after correcting for centroid shifts due to saddle-shaped HIC profiles (**Figure 7i**). The non-core variants showed no correlation (**Figure 7j**). Thus, overall hydrophobicity was necessary but not sufficient for high cytotoxicity; it mattered which parts of the polypeptide backbone were crosslinked.

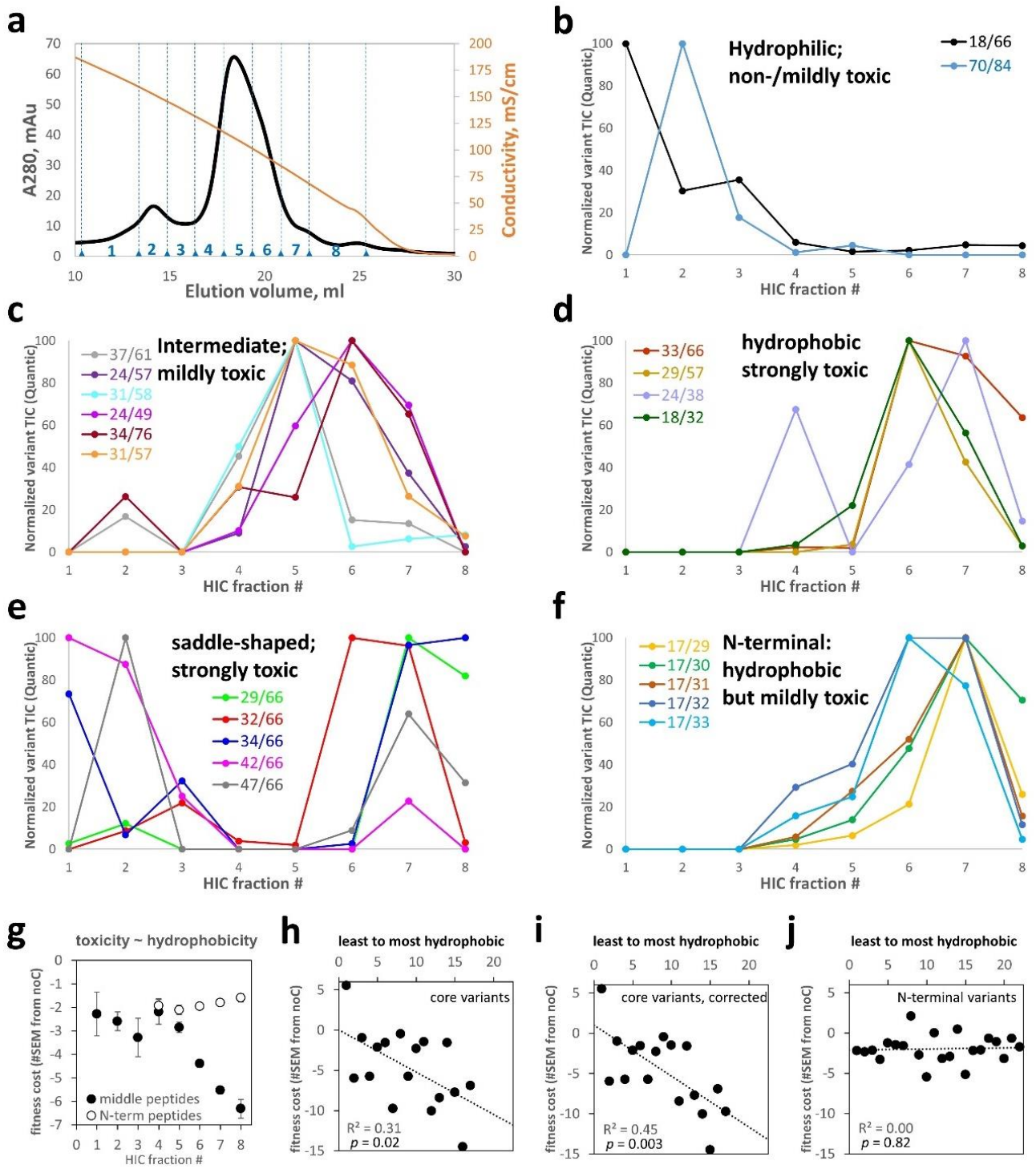

Figure 7: HIC of a pooled HdeA variant library reveals a correlation between hydrophobicity and toxicity. (a) HIC elution trace of the total HdeA variant library on an ammonium sulfate gradient reported by conductivity (*orange curve*). Eight fractions were collected, numbered as indicated. Elution profiles of all 17 "core" variants are shown, grouped by similarity of elution profiles and centroids of elution: **(b)** below 4.7, **(c)** 4.7-6.0, **(d)** 6.0-6.8. **(e)** Some toxic core variants had unusual saddle-shaped elution profiles. **(f)** Most non-core variants were identified from N-terminal peptides, many with the missed cyanylation at position 17; those variants had high hydrophobicity despite modest toxicity. **(g)** Hydrophobicity and toxicity (PSM-averaged *E* value) were strongly correlated for variants identified from middle peptides but not from N-terminal peptides. **(h)** Differences between core and N-terminal variants accounted for this divergence, especially after correcting for centroid shifts due to saddle-shaped elution profiles (by ignoring data from the first 3 HIC fractions) **(i). (j)** By contrast, non-core variants – identified from either set of peptides – showed no hydrophobicity-toxicity correlation.

Apparently non-toxic variants have greatly reduced expression levels

The WT gene had the third-lowest epistatic fitness cost among the 1,453 variants in **Figure 3**. The two variants with even lower apparent cost were 12/58 and 6/74. The former also had a clear island of non-toxicity around it in sequence space. We examined the growth curves of 12/58 and 6/74, along with the near-neutral 17/68 and 57/77 for comparison (**Figure SI 9**) and, indeed, observed less die-off during stationary phase than the noC variant, with 12/58 and 6/74 curves quite similar to WT. Four highly toxic variants served as "positive controls" for toxicity and, indeed, showed greater stationary-phase die-offs than noC. These low-throughput finding further validate the high-throughput assay using pooled cultures.

The Cys pairs in 12/58 and 6/74 are too far apart for native-state intramolecular crosslinks, even considering N-terminal flexibility, and non-reducing SDS-PAGE (**Figure SI 10**) showed no evidence of extensive intersubunit crosslinks, either. Isotopically resolved intact-protein mass spectrometry confirmed intramolecular disulfide bonding (-2 Da. shift) in 12/58, along with WT and 32/66 (**Figure SI 11**). Variant 6/74 could not be reliably measured due to extensive protein degradation. Indeed, protein degradation was the most likely explanation for its lack of toxicity. At moderate induction, 6/74, 12/58, and even 17/68 accumulated to a much lower level than other variants (**Figure SI 10**) and caused much less DnaK overexpression or cell lysis. At very high induction, however (2.5 mM rhamnose, exceeding the 1 mM used for the high-throughput assay), expression, DnaK, and cell lysis all returned. It appears, then, that while some disulfides stabilize cytotoxic backbone conformations, others stabilize degradation-prone conformations and thereby avoid cytotoxicity, except at the highest [inducer].

Discussion

We have demonstrated the feasibility of disulfide-scanning an entire protein in one experiment: at the DNA level using amplicon deep-sequencing and at the protein level by devising and implementing a novel method for limited deep-sequencing of proteins. HTDS can probe protein conformational landscapes in a model-free manner and map them to molecular and organismal phenotypes. Our approach is widely accessible: the total cost of all DNA and protein sequencing reported in this study was just over $1,000. HTDS will be even more powerful when full single-molecule protein deep-sequencing becomes available.

We found direct experimental evidence for misfolding toxicity in HdeA, a conditional IDP and one of the most abundant *E. coli* proteins. Removing the native 18-66 disulfide was cytotoxic but shifting it by even one residue (to 18-67 or 18-65) or one helical turn (22-66 or 18-70) was even more so. Only a few near-native disulfides (such as 17/66, 17/65, 15/66) provided significant (>2 SEM) rescue of noC toxicity. This brittleness is surprising given the high thermostability of the native WT structure (Foit et al., 2013). HdeA may have evolved to be brittle because function requires on-cue denaturation. Ectopic disulfides like 32/66 raised both hydrophobicity and cytotoxicity above noC, and reducing them restored noC-like hydrophobicity. Thus, a specific subset of crosslinks may stabilize especially cytotoxic conformers that are otherwise transient (the resulting landscape resembles **Figure 1a**). Sharp phenotypic peaks or valleys on the backbone conformational landscape may be common in IDPs: e.g., a recent study found that the 17/28 internal disulfide in amyloid-β_{42} led to formation of highly cytotoxic oligomers, but 16/29 or 18/27 did not (Matsushima et al.).

The detailed molecular mechanism of HdeA cytotoxicity remains a topic for future research. We have shown that it involves protein misfolding stress, evidenced by strong DnaK overexpression, and ultimately cell lysis. Protein misfolding is frequently cytotoxic; proposed explanations include direct disruption of cell membranes, aberrant protein-protein interactions, or overloading of protein quality control systems (Folger and Wang, 2021; Geiler-Samerotte et al., 2011). Elucidating HdeA toxicity mechanisms could uncover new bacterial physiology, especially given our unexpected observation that WT HdeA is efficiently exported without rupturing the cell.

This study's main technical innovation is protein-level deep-sequencing of libraries of double-Cys variants to identify their redox state or enrichment in chromatographic fractions. Yet, library coverage was ~10X lower for proteins than DNA, mainly due to the much lower practical dynamic range of MS compared to Illumina sequencing.

Variant abundances at the DNA level varied by >1000X in our plasmid library – partly by design and partly because of the exponential PCR amplification steps of megaprimer-based mutagenesis. The Cys66-containing sub-library that was enriched at the DNA level had much better sequence coverage (**Figure 4a**), with 56 variants detected by Illumina and 31 by LC/MS/MS. As high-throughput DNA synthesis advances, *de novo* synthesis of double-Cys libraries of desired composition will overcome the variability due to PCR amplification while also improving control of library composition (e.g., 100% double-Cys variants instead of ~47% in this study).

HTDS, like chemical crosslinking MS (XL/MS), may be used for protein structure determination when combined with *in-silico* simulations (MacCallum et al., 2015; Orban-Nemeth et al., 2018). Low-throughput disulfide scanning can already distinguish structural models or clarify specific tertiary structure elements (Butler and Falke, 1998; Krshnan et al., 2016; Molnar et al., 2014; Taguchi et al., 2018). We expect that HTDS will allow model-free structure determination, at moderate resolution but *in vivo*, including for non-native and perhaps even aggregated conformers. If many IDPs are indeed ordered *in vivo* (Leuenberger et al., 2017), HTDS could help elucidate those structures. HTDS has distinct advantages and limitations relative to XL/MS for structure determination. Disulfides easily form *in vivo* without external perturbation, even in cytoplasmic proteins in strains such as SHuffle® (Lobstein et al., 2012). Being genetically encoded, disulfides can reveal heritable and evolutionary effects of protein conformations. HTDS as implemented here reports only intramolecular crosslinks, unambiguously identified from the termini of linear peptides, which greatly simplifies analysis. Finally, the disulfide bond's combination of small size and reversibility is unmatched by chemical crosslinkers. On the other hand, HTDS cannot be applied to entire proteomes, as limited proteolysis/MS has already been applied (Leuenberger et al., 2017) and XL/MS could be, in principle. An oxidizing environment is required, and some mutations could interfere with folding. Finally, HTDS in its current form cannot be applied to natively disulfide-rich proteins; only full single-molecule protein sequencing, without cleavage, will likely overcome this limitation. Finally, XL/MS is better for mapping quaternary structure.

Importantly, HTDS can not only reveal native or non-native conformations but also stabilize them. This is critical for applications such as drug screening or vaccine design. Stabilizing the wrong conformation of the antigen in the RSV vaccine led to tragic consequences (Killikelly et al., 2016). Low-throughput disulfide engineering later enabled stabilization of the correct conformation (McLellan et al., 2013). The same approach has been applied to the SARS-CoV-2 spike protein, resulting in some useful prefusion stabilization (Riley et al., 2021; Xiong et al., 2020) but also a very high failure rate (Hsieh et al., 2020). In those studies, the sites for disulfide engineering were chosen by some combination of structural intuition and computational modeling, and <100 variants were screened. HTDS could allow screening of many thousands of variants even before an atomistic structure is available and may reveal useful disulfides with unexpected allosteric effects. In other cases, identifying disulfides that trap a non-native cytotoxic conformation – as we report here for *E. coli* HdeA – could enable screening for drugs that bind and stabilize that conformation even in the wild-type protein. Such compounds could become a novel class of antibiotics, with induced misfolding of key pathogenic proteins as their mode of action.

Methods

Generation of 2-Cys scanning libraries

A linear dsDNA fragment encoding the C18S/C66S variant of *E. coli* HdeA ("noC"), with the native signal sequence for periplasmic export, was purchased from GeneUniversal. Except for the two point mutations, the sequence was entirely native, with no codon optimization. This fragment was cloned into the pCK302 vector(Kelly et al., 2016) in place of the superfolder GFP gene, using the FastCloning method.(Li et al., 2011) The vector was obtained via Addgene. In subsequent validation experiments, all indicated individual variants were synthesized by GeneUniversal or Thermo Scientific and cloned into the same vector in the same way. A single-Cys scanning library covering all 83 positions from Gln4 to Lys86 (numbered based on the mature HdeA protein, without signal sequence) was purchased from GeneUniversal and cloned into the pCK302 vector using FastCloning. Limited T5 exonuclease digestion(Xia et al., 2019) was used to increase cloning efficiency, resulting in ~200 transformant clones.

Mutagenesis of the single-Cys scanning library ("CSL") was carried out using the megaprimer method.(Tyagi et al., 2004) First, short mutagenic forward primers, covering 51 positions from Ala3 to Lys87, were

purchased from IDT DNA. These 51 positions were chosen to avoid mutations to Pro, Gly, Trp, Phe, Leu, Ile, Glu, or Asp codons at this stage. The noC gene was PCR-amplified in 51 separate reactions using those forward primers and a fixed reverse primer downstream of the gene, in addition to a primer-less negative control. The primer sequences are listed in **Table SI1**. The resulting megaprimers were directly applied, without further purification, to a new set of 51 PCR reactions, this time with CSL as the template instead of noC. The pooled PCR products were incubated at a 10:1 ratio with DpnI enzyme (New England Biolabs) at room temperature for 5 min, then at 37 °C for 50 min. Self-ligation of the long linear PCR fragments was achieved by *in vivo* homologous recombination, assisted by limited T5 exonuclease digestion, as above, prior to transformation to DH5α competent *E. coli* cells (New England Biolabs), to create a multi-Cys scanning library ("MCSL") in the pCK302 vector. Separate digestions and transformations were carried out for targeted libraries having either or the two native Cys residues, Cys18 or Cys66, on the CSL background (termed "C18CSL" and "C66CSL"), as well as for the negative-control library. After recovery in SOC medium for 1.5 h, samples representing 2.5% of the volume of each library were plated on LB-agar plates containing ampicillin, resulting in 200-300 colonies each for MCSL, C18CSL, and C66CSL, and no colonies for the negative control, indicating ~10,000 transformants per library. The remaining 97.5% of each transformation mixture was inoculated directly into SuperBroth (Teknova) containing ampicillin, to a total volume of 4 ml, cultured overnight at 37 °C with shaking. Plasmids were prepared using the Qiagen miniprep kit with two preps per culture.

A total of 10 individual clones from MCSL were Sanger-sequenced, along with two clones each from the targeted CSLs. The MCSL clones included two 1-Cys variants, six 2-Cys variants, and two 3-Cys variants; the C66CSL clones were two 2-Cys variants; and the C18CSL clones were one 1-Cys variant (containing C18) and one variant with a run-on duplication of the primer.

For all further experiments, a combined library ("MCSL++") was generated to enrich for the native Cys residues and ensure reasonable abundance of variants that could serve as markers for the DNA-level and protein-level sequencing of the pooled library cultures. Specifically, 830 ng of the MCSL library was combined with 130 ng of the C66CSL library and 10 ng each of the cytotoxic 40/65 variant and the non-toxic WT.

Bacterial fitness assays

All fitness assays were carried out in the non-rhamnose-metabolizing *E. coli* BW25113-derived Δ*hdea* strain from the Keio knockout collection. The strain was plated, and one clone selected for further work. The cells were made chemically competent by a modified version of manganese-assisted permeabilization.(Untergasser, 2008) Filtered, chilled Inoue solution was prepared from 25 ml MilliQ water, 0.5 g KCl, 0.25 g $MnCl_2 \cdot 4H2O$, 0.055g $CaCl_2 \cdot 4H2O$, and with 10 mM (final) PIPES buffer pH 6.7. Two cell pellets grow in SuperBroth medium (Teknova) were washed with 8 ml of this solution each, then resuspended in a combined 4 ml volume of the same solution, always on ice. Room-temperature DMSO (0.3 ml) was then added, along with 1.7 ml 50% glycerol if cells were to be stored at -70 °C. Otherwise, the cells were transformed immediately. All constructs used in the fitness assay were then transformed into this strain of competent cells. In the case of MCSL++, 1% of the transformant mixture was plated, resulting in 96 colonies. The full library therefore contained ~10,000 transformants, which was deemed sufficient.

Thawed polyclonal glycerol stocks were diluted 1:100 to fresh LB broth with 100 µg/ml ampicillin and 50 µg/ml kanamycin and allowed to recover for 1-2 h at 37 °C with shaking. Induction cultures in the same medium containing various amounts of L-rhamnose in 96-well plate format were then inoculated 1:100 or 1:1000 (as indicated) from these recovered cultures. The 96-well plates were incubated for 14 h at 37 °C with 567 cpm linear shaking (3 mm amplitude) in BioTek Epoch plate readers with optical density monitored at 600 nm.

Amplicon library preparation and deep sequencing

Replicate time point samples from pooled cultures of the MCSL++ transformant library were centrifuged for 15 minutes at 3,000*g* in 15-ml conical tubes or 14-ml round-bottom tubes to fully clarify them. The supernatants were then carefully withdrawn by pipetting and both pellets and supernatants iced or frozen thereafter. Plasmids

were extracted from the pelleted cells using a GeneJet (Thermo) miniprep kit according to the manufacturer's instructions. Amplicons with Nextera adapters for Illumina sequencing were generated from the pelleted cells ("PEL") using 10 ng of purified plasmid as the template for each 25 µl reaction and 25 PCR cycles with the Q5 Master Mix (New England Biolabs). Amplicons from the clarified culture supernatant ("SUP") were generated the same way, using 1 µl of the SUP sample as the template. PCR cleanup was carried out on each amplicon sample using the Monarch PCR cleanup kit (New England Biolabs). The amplicons were then dual-indexed for paired-end read Illumina sequencing using Nextera primers N701 and N702 for the PEL and SUP of the 5.5 h time point and N705 and N706 for the PEL and SUP of the 9.0 h time point, respectively, and primers S517, S502, S503, and S504 for samples from the four replicate cultures in each case. Per manufacturer instructions, 8 PCR cycles were used, with 100 ng template per 50 µl reaction with Q5 Master Mix. Success of each reaction was confirmed by running 5 µl of each indexed PCR produced on an agarose gel with ethidium bromide staining, which showed bands of equal intensity for all. The samples were then pooled into 8-plex libraries by time point (containing equal volumes of PCR products for all replicates of both PEL and SUP amplicons), each pool PCR-cleaned again as above, and the resulting indexed 8-plex amplicon libraries sequenced by Novogene in a single split NovaSeq lane using paired-end 155 base-pair reads. The reads were merged using AmpliMERGE(2021) and the reads filtered, trimmed, and variants counted (with all synonymous reads combined) using Enrich2.(Rubin et al., 2017) For the 9.0 h time point, 1.8-2.2 million reads per replicate were identified, and for the 5.5 h time point, 1.2-1.7 million reads per time point. The 5.5 h data were not used further due to very high variability: a large number of outliers, particularly in the SUP sample, with one of the four replicates showing >10X higher abundance than the others. This may be attributable to the very low amount of DNA in the culture supernatants at this early time point. Therefore, we focused on the 9.0 h time point, where such cases were few. Only variants with at least 50 reads on average across all conditions were included in further analysis. Variants containing outliers were filtered out by requiring that the standard error of the SUP/PEL abundance ratio be no more than half as large as the ratio itself. The SUP/PEL ratio was used as the raw measure of toxicity (fitness cost), since protein variants that caused cell lysis were the ones whose plasmids were disproportionately enriched in the supernatant of the pooled cell cultures.

Calculation of epistatic fitness of the 2-Cys variants

Ratio epistatic fitness(Rollins et al., 2019) for each double-Cys variant was calculated by dividing its SUP/PEL abundance ratio by the product of the SUP/PEL ratios for the two single-Cys variants at those positions. For the noC variant, toxicity was defined as just the raw SUP/PEL ratio, which was 0.89 for the 9.0 h time point samples. Since variant abundances in the MCSL++ library varied by at least three orders of magnitude at the DNA level (partly by design), the standard errors also varied, tending to be greater for the less-abundant variants. To account for this, we chose to plot the fitness-cost landscape of Figure 3 in terms of the number of standard errors by which the mean of epistatic fitness differed from the mean of noC (i.e., from 0.89). The resulting map did not differ qualitatively from the more traditional measure of the logarithm of ratio epistasis (shown in Figure **SI 5**). Heatmaps were generated using Matplotlib(Hunter, 2007) and Seaborn(Waskom, 2021) in Python 3.

Protein expression

Small-scale (typically 5 ml) LB cultures 100 µg/ml ampicillin and 50 µg/ml kanamycin were inoculated 1:100 with thawed glycerol stocks of the transformed *E. coli* BW25113 Δ*hdea* and allowed to recover for 1-2 h. In the case of MCSL++ cultures, the initial inoculate was 1:1000. Larger batches (0.5 L) of the same medium were then inoculated 1:1000 from these recovery cultures and incubated in 2L flasks at 37 °C with 250-300 rpm shaking for 14-16 h in the presence of 1-3 mM L-rhamnose as the inducer. Protein expression was verified by SDS-PAGE of both total and supernatant fractions of samples taken immediately thereafter. Cultures were harvested in conical or square bottles by centrifugation in a swing-bucket rotor. The supernatants were further centrifuged in conical bottles (Celltreat) at maximum speed, treated with Complete EDTA-free protease inhibitor mixtures (1 crushed tab per 1-2 L medium), and passed through bottle-top 0.2-µm filters (VWR). Thus treated, culture supernatants were stored tightly capped at 4 °C in sterile bottles until further use. As long as the cultures remained clear, SDS-PAGE did not show any noticeable protein degradation even after several weeks.

Preparation of periplasmic extracts

Extraction buffer comprised 330 mM Tris pH 8, 1 M sucrose, and 2 mM EDTA. Immediately prior to extraction, one crushed tab of Complete EDTA-free protease inhibitor cocktail (Roche) was added to ice-cold extraction buffer. Pelleted cell cultures were gently resuspended on ice in this buffer using a cell scraper and allowed to incubate for 10 or 30 minutes (the shorter time reduced the amount of complete cell lysis when toxic variants were expressed and was therefore used for the MCSL++ library). The extracts were immediately clarified by centrifugation at 12,000 rpm in a microcentrifuge chilled in advance to 4 °C. Extracts were then stored at -70 °C.

Protein purification

Since the majority of overexpressed HdeA was typically found in the culture supernatant (see **Figure 2b** and **Figure SI 4**), all protein samples except for the periplasmic samples in **Figure 4** were purified from supernatants prepared as described above and concentrated ~50-100-fold in Centricon Plus70 concentrators (Millipore). The concentrated supernatants were filtered through 2x5ml tandem Sepharose Q columns (Cytiva) in pH 6.7 10 mM PIPES buffer; mixed 1:1 with 4 M ammonium sulfate (Teknova) and centrifuged for 10 min., then fractionated by hydrophobicity on a HiTrap Phenyl HP column (Cytiva) with 2-0 M ammonium sulfate gradient in 10 mM PIPES pH 6.7 buffer. Finally, the peak HIC fractions were concentrated as needed and fractionated by size exclusion on a Superdex 75 Increase 10x300 column (Cytiva) equilibrated in buffer suitable for the given experiment (10 mM sodium phosphates pH 7 for CD; the same with 150 mM NaCl for SEC/MALS; or 10 mM PIPES pH 6.7 with 150 mM NaCl for cyanylation/aminolysis).

Purification of the pooled libraries followed the same procedure, but with a step gradient in HIC elution (from 2M, then 1.5M, then 0M ammonium sulfate), due to the varying hydrophobicity of library components. This procedure resulted in ~90% purity for the MCSL++ library as determined by SDS-PAGE, which was deemed an acceptable compromise given the ability of mass spectrometry to identify peptides in complex mixtures.

Purification of periplasmic extracts was carried out exactly as above, except that NaCl was added to the periplasmic extracts to 150 mM final concentration prior to the Sepharose Q filtration step to minimize unwanted binding to the column.

Cys-specific cleavage of proteins by cyanylation/aminolysis

CDAP (Sigma) was prepared as a 100 mM stock in pure acetonitrile, the headspace of the vial purged with nitrogen, the vial lid taped, and the vial stored at -20 °C. Before being opened for use, the vial was allowed to equilibrate to room temperature to minimize condensation of water vapor from the air, and after use it was immediately purged, taped, and stored as before. We found that when this procedure was carefully followed the reagent retained most of its activity for multiple uses over the course of several months.

A near-saturated solution of 6 M L-Arginine at pH 9 was prepared by dissolving the powdered compound stepwise in a minimum amount of water while titrating the pH at each step with sodium hydroxide pellets and neat hydrochloric acid to ensure solubility. The solution was stored at room temperature protected from light and remained usable for several months.

Protein samples (10-100 µM concentration) were denatured in 4 M guanidinium chloride (Thermo), pH7, at 42 °C for 30-60 min. in LoBind microcentrifuge tubes (Eppendorf). For experiments identifying disulfide bonding propensities, this incubation was carried out in the presence of either 5 mM TCEP (for the total sample) or 10 mM N-ethyl maleimide (NEM, for the free thiol-blocked sample), with 10% v/v DMSO in each case (since N-ethyl maleimide stock was prepared in DMSO). Both samples were then fractionated by SEC on a Superdex75 Increase 10x300 column equilibrated in pH 7 sodium phosphates with 4 M guanidinium chloride, and the elution peaks (at ~10.0 ml for the TCEP-treated sample and ~10.4 ml for the NEM-treated sample) collected manually. Note that under the same conditions fully reduced WT HdeA eluted at 10.0 ml and non-reduced WT at 10.6 ml. This denaturing SEC step was expected to remove any disulfide-bridged dimers from the NEM-treated library. Prior to cyanylation, the SEC peak fraction of the NEM-treated library was reduced with TCEP for 45 min at 37 °C. For the proof-of-concept experiments on hydrophobicity, no SEC or thiol blocking was carried out to minimize sample loss. Instead, samples were reduced for 1 h with TCEP as above, then buffer-exchanged by Zeba desalting columns directly into the cyanylation buffer.

For cyanylation, all samples were buffer-exchanged to pH 3 buffer containing 20 mM sodium citrate and 4 M guanidinium chloride, either by ultrafiltration (using Pall Nanosep filters spun at 6000 rpm in a microcentrifuge) or by centrifugal desalting columns (Thermo Fisher Zeba 2 ml) and immediately used for cyanylation without the addition of more reducing agent. To cyanylate the Cys residues, CDAP reagent was added to 10 mM final concentration, and the mixture was incubated at room temperature for ~2 h. Each sample was then mixed 1:1 with the 6 M pH 9 L-arginine solution and incubated at room temperature overnight, protected from light. The aminolysis reaction was then quenched by addition of 2% v/v of 95% formic acid, bringing the mixture to pH ~4.

Identification of pooled 2-Cys variants by mass spectrometry

A total of ~1 nmol of each cleaved sample was injected on a C18 HPLC column and fractionated with a 120-min gradient of 10-45% acetonitrile, followed by a 10-min washout gradient to 80% acetonitrile. All buffers contained 0.1% formic acid. Eluate from the inline C18 column was injected to a qExactive Plus mass spectrometer (Thermo) and top-5 MS/MS spectra collected with 1+ ions excluded from MS/MS.

FASTA libraries were generated containing only the "middle" peptides for all pairs of cleavage sites on the HdeA C18S/C66S background sequence. By definition, all these peptides began with a Cys and ended with an Arg (the nucleophile used for cleavage). Decoy libraries were generated using the De Bruyn algorithm in the Comet pipeline of the Trans Proteomic Pipeline v6.(Deutsch et al., 2015) All decoys were set to also begin with Cys and end with Arg, with only the sequence in-between those termini being scrambled. For additional searches using missed-cyanylation sites, separate FASTA libraries were generated for peptides spanning from the protein N-terminus to a C-terminal Arg or from Cys to the protein C-terminus, respectively. Decoys were likewise required to obey the same constraint on the termini as true-hit peptides. The C-terminal peptide library was not used further due to a very small number of missed-cyanylation peptides that could be identified in any samples. By contrast, the N-terminal missed-cyanylation library was useful due to the apparently reduced efficiency of cyanylating certain N-terminal sites, especially position 17.

The MS/MS output files were searched against these FASTA databases using Comet (version 6). A required modification of the peptide N-terminal Cys residue was added to the Comet parameter file to account for the conversion of Cys to iminothiazole. Mass tolerance was set to 1.1 Da. in the Comet parameter file, to account for incorrect calling of monoisotopic peaks for larger peptides. The Comet parameter file used for the searches are included in the SI. The output .pep.xml files were manually edited to change the value of the "enzyme name" field from "CDAP" to "nonspecific," then passed to the PeptideProphet tool with default parameters except for no accurate mass binning and using only the E-value as the discriminant. Finally, this output was passed to the Quantic tool with default parameters to quantify the total ion current for top-6 MS/MS peaks for each matched peptide spectrum regardless of nominal PeptideProphet probability. The output was exported to Excel.

The false-discovery rate as a function of Comet E-value was determined for each output file empirically. Since crosslinks between sequence near-neighbors are inherently less informative for HTDS, and, moreover, short peptides may not produce a sufficient number of MS/MS ions for Quantic, data were filtered to remove any peptides where the Cys residues were closer than 12 peptide bonds apart. Data were also filtered to remove any apparent peptide spectrum matches where the identified double-Cys variant was not found in the DNA-level deep sequencing dataset (i.e., the set of 1,453 variants whose toxicity values were deemed sufficiently reliable). The remaining peptides were ranked by ascending E-value and by whether they were decoys or true hits. The false-discovery threshold was set at 10%. In other words, the top N true-hit peptides with the smallest E-values were used for further analysis, such that no more than N/10 peptide spectra from decoy hits had E-values within this range. These N peptide spectrum matches were deemed usable. For evaluating the disulfide bonding propensities (**Figure 4**), the total ion current values for all usable peptide spectrum matches for a given variant were summed. For evaluating PSM-weighted averages of toxicity values by HIC fraction (**Figure 7g**), they were weighted by the number of peptide spectrum matches for that variant within that fraction.

Intact protein mass spectrometry

Electrospray-ionization isotopically resolved mass spectrometry of intact proteins was carried out as described.(Serebryany et al., 2018)
The protein samples were analyzed on a Bruker Impact II q-TOF mass spectrometer equipped with an Agilent 1290 HPLC. The separation and desalting was performed on an Agilent PLRP-S Column (1000A, 4.6 x 50 mm, 5 µm). Mobile phase A was 0.1% formic acid in water and mobile phase B was acetonitrile with 0.1% formic acid. A constant flow rate of 0.300 ml/min was used. Ten microliters of the protein solution was injected and washed on the column for the first 2 minutes at 0%B, diverting non-retained materials to waste. The protein was then eluted using a linear gradient from 0%B to 100%B over 8 minutes. The mobile phase composition was maintained at 100%B for 1 minutes and then returned to 0%B over 0.1 minute. The column was re-equilibrated to 0%B for the next 5.9 minutes. A plug of sodium formate was introduced at the end of the run, to perform internal m/z calibration to obtain accurate m/z values. The data were analyzed using Bruker Compass DataAnalysis™ software (Version 4.3, Build 110.102.1532, 64 bit). The charge state distribution for the protein produced by electrospray ionization was deconvoluted to neutral charge state using DataAnalysis implementation of the Maximum Entropy algorithm. Predicted isotope patterns were calculated at the resolving power of 50,000 and compared with isotopically resolved neutral mass spectra calculated using Maximum Entropy from the experimental charge state distribution.

Intrinsic Trp fluorescence
Fluorescence spectra were measured at room temperature in a Cary Eclipse fluorimeter (Varian/Agilent) with excitation at 290 nm (10 nm slit) and emission scanned from 300 to 450 nm (5 nm slit) and 700 V PMT voltage. Protein samples were prepared at 9-10 µM (identical within each set of 4 samples) in 10 mM pH 7 sodium phosphates buffer with 150 mM NaCl, unless otherwise indicated. Four samples were read in parallel using a multi-cuvette holder; no-protein blanks in the same cuvettes were subtracted. Five spectra per sample were averaged.

Bis-ANS fluorescence
To the same protein samples that were used for measuring intrinsic fluorescence, bisANS was added to 10 µM final concentration. Excitation wavelength was 400 nm (10 nm slit); emission was scanned 450-600 nm (5 nm slits) at room temperature. PMT voltage was kept at 700 V. For experiments with reducing agent, TCEP was added to 5 mM final concentration. Sample dilutions were no more than 5% from addition of the compounds.

Circular dichroism
Ellipticity was measured on a Jasco J-1500 instrument, at room temperature, in 10 mM pH 7 sodium phosphates buffer, in a 1 mm path length cuvette with 16-20 µM [protein] and 50 nm/min scanning rate, with 5 scans averaged per sample. Curves were smoothed by taking the moving average with a 3-nm sliding window.

Size-exclusion chromatography / Multiangle light scattering (SEC/MALS)
An SRT SEC-150 column (Sepax) was pre-equilibrated with 10 mM pH 7 sodium phosphates with 150 mM NaCl on an Infinity 2 1260 HPLC system (Agilent) with an inline degasser. The absolute refractive index value for the buffer was determined empirically. The inline UV detector, the Dawn Heleos II multiangle light scattering detector (Wyatt), and the OptilabTrEX RI detector (Wyatt) were aligned using a sample of bovine serum albumin. HdeA samples (20 µM, 100 µL) were injected and run at 0.5 ml/min for 45 min., with MALS detection every 0.5 s. Signals from the highest two and lowest two angle detectors were removed to reduce noise. Molecular weights were determined automatically using ASTRA 7 software, with “Normal” despiking and manually adjusted baselines where required.

Atomistic Monte-Carlo simulations
We used our previously established MCPU software, which uses a knowledge-based, statistical potential for atomistic modeling of protein unfolding intermediates,(Bitran et al., 2020; Serebryany et al., 2016; Yang et al., 2007) the MCPU version used in this study is available at https://github.com/proteins247/dbfold. To mimic the

effect of a disulfide linkage between two cysteines, we use a strong flat-bottomed harmonic potential as a restraint between the CB atoms of a pair of cysteine residues. For the restraint, the region of zero potential consists of CB-CB distances between 2.9 and 4.6 Å.(Hazes and Dijkstra, 1988) The restraint force constant was set to 200 $Å^{-2}$. An NMR structure of E. coli HdeA (PDB ID: 5WYO) was used as the initial structural model. A 13-residue linker consisting of alternating G and S residues was built to connect the C-terminus of the A chain to the N-terminus of the B chain in the homodimer structure. This linker is needed due to a limitation in the simulation software. For each double-Cys variant, Modeller (version 9.24)(Sali and Blundell, 1993) was used to mutate target residues to cysteine. Original cysteine residues 18 and 66 were mutated to serine as necessary. NAMD (version 2.13),(Phillips et al., 2020) with the CHARMM 22 forcefield,(MacKerell et al., 1998) was used to minimize the initial structure. Then, each minimized cysteine variant was simulated in a multiplexed temperature replica exchange MCPU simulation with flat-bottom harmonic restraints active between the two intramolecular cysteine pairs. Each simulation used 21 temperatures (between 0.350 and 0.850, inclusive, in steps of 0.025) with four replicas at each temperature. Simulations were run for 1.275 billion MC steps. A knowledge-based moveset was enabled for the first 300 million steps. The replica exchange interval was set to 10,000 steps, and simulation structures were saved every 1,000,000 steps.

Simulation data from the last 275 million steps were used for analysis. Simulation samples from between temperatures of 0.375 and 0.625 were analyzed. Simulation structures were clustered based on intermolecular (inter-subunit) contact maps. For each simulation sample, a 2D Boolean matrix corresponding to contacts between residues 11 to 89 of one subunit and residues 11 to 89 of the other subunit was constructed, with a contact defined as a CA-CA distance of less than 10 Å. The flattened matrix was then used as a feature vector for clustering. The Jaccard distance metric was used as the measure of dissimilarity, and clustering was performed using the DBSCAN algorithm,(Ester et al., 1996) as implemented in scikit-learn.(Pedregosa et al., 2011) For DBSCAN, epsilon was set to 0.57 and minimum points to 14. Clustering was performed on structures from all simulated variants. For computational tractability, the number of samples from simulations was reduced 10-fold (i.e. structures from every 10 million steps such that each replica provides 28 structures). Simulation observables such as distances, secondary structure, and radius of gyration were measured using the Python package MDTraj.(McGibbon et al., 2015) Expectation values for observables at each temperature were calculated using the pymbar package.(Shirts and Chodera, 2008)

For more in-depth simulations of the free energy landscapes of select variants, we used monomeric starting structures (subunit A of PDB ID 5WYO), since the mutants were shown experimentally to be monomeric, and applied umbrella biasing as well as replica exchange for enhanced sampling. The mutations were generated and monomeric variants minimized using MOE.(2022) From each minimized structure, we simulated the variant at 20 MCPU temperatures between 0.250 and 0.725 for 600 million MC steps (including knowledge-based moves for the first 50 million steps). The simulation structures along with their energy values were saved every 500,000 steps.

Data from the last 400 million steps were used for analysis. The pymbar package was used to calculate the free energy landscape as a function of the fraction of native contacts (Q) and RMSD. RMSD and native contacts fraction (using the definition from(Best, 2013)) were computed using the Python package MDTraj.(McGibbon et al., 2015) Surface hydrophobicity for each simulation frame was defined as the sum of Miyazawa–Jernigen hydrophobicity values of each residue(Ladiwala, 2006) weighted by the solvent accessible surface area (SASA) of that residue normalized to total SASA of the protein. By thermal averaging the surface hydrophobicity of each snapshot using the free energy difference calculated with the pymbar, the expectation value of the hydrophobicity was measured.

Acknowledgments
E. S. is grateful to Dr. Bharat Adkar and Dr. Joao Rodrigues for mentoring in the generation and cloning of DNA variant libraries and insightful discussions.

This work was supported by the National Institutes of Health grants F32GM126651 and K99GM141459 to E. S. and R35GM139571 to E. I. S.

(2021). AmpliMERGE (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplimerge: AmpliSAT: online tools for the analysis of amplicon sequencing data).
(2022). Molecular Operating Environment (MOE), 2020.09 (Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7).
Aguirre-Cardenas, M.I., Geddes-Buehre, D.H., and Crowhurst, K.A. (2021). Removal of disulfide from acid stress chaperone HdeA does not wholly eliminate structure or function at low pH. biorXiv.
Alfaro, J.A., Bohlander, P., Dai, M.J., Filius, M., Howard, C.J., van Kooten, X.F., Ohayon, S., Pomorski, A., Schmid, S., Aksimentiev, A.*, et al.* (2021). The emerging landscape of single-molecule protein sequencing technologies. Nature Methods *18*, 604-617.
Ascenzi, P., and Gianni, S. (2013). Functional Role of Transient Conformations: Rediscovering "Chronosteric Effects" Thirty Years Later. Iubmb Life *65*, 836-844.
Best, R.B., Hummer, G., Eaton, W. A. (2013). Native contacts determine protein folding mechanisms in atomistic simulations. Proc Natl Acad Sci U S A *110*, 17874–17879
Bitran, A., Jacobs, W.M., and Shakhnovich, E. (2020). Validation of DBFOLD: An efficient algorithm for computing folding pathways of complex proteins. Plos Computational Biology *16*.
Borgia, A., Kemplen, K.R., Borgia, M.B., Soranno, A., Shammas, S., Wunderlich, B., Nettels, D., Best, R.B., Clarke, J., and Schuler, B. (2015). Transient misfolding dominates multidomain protein folding. Nature Communications *6*.
Brockwell, D.J., and Radford, S.E. (2007). Intermediates: ubiquitous species on folding energy landscapes? Current Opinion in Structural Biology *17*, 30-37.
Butler, S.L., and Falke, J.J. (1998). Cysteine and disulfide scanning reveals two amphiphilic helices in the linker region of the aspartate chemoreceptor. Biochemistry *37*, 10746-10756.
Datta, S., Koutmos, M., Pattridge, K.A., Ludwig, M.L., and Matthews, R.G. (2008). A disulfide-stabilized conformer of methionine synthase reveals an unexpected role for the histidine ligand of the cobalamin cofactor. Proceedings of the National Academy of Sciences of the United States of America *105*, 4115-4120.
Deutsch, E.W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R.L. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clinical Applications *9*, 745-754.
Dishman, A.F., and Volkman, B.F. (2018). Unfolding the Mysteries of Protein Metamorphosis. ACS Chem Biol *13*, 1438-1446.
Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231.
Foit, L., George, J.S., Zhang, B.W., Brooks, C.L., and Bardwell, J.C.A. (2013). Chaperone activation by unfolding. Proceedings of the National Academy of Sciences of the United States of America *110*, E1254-E1262.

Folger, A., and Wang, Y.C. (2021). The Cytotoxicity and Clearance of Mutant Huntingtin and Other Misfolded Proteins. Cells *10*.
Fu, X., Wang, Y., Song, X., Shi, X., Shao, H., Liu, Y., Zhang, M., and Chang, Z. (2019). Subunit interactions as mediated by "non-interface" residues in living cells for multiple homo-oligomeric proteins. Biochem Biophys Res Commun *512*, 100-105.
Gajiwala, K.S., and Burley, S.K. (2000). HDEA, a periplasmic protein that supports acid resistance in pathogenic enteric bacteria. Journal of Molecular Biology *295*, 605-612.
Gautier, C., Troilo, F., Cordier, F., Malagrino, F., Toto, A., Visconti, L., Zhu, Y., Brunori, M., Wolff, N., and Gianni, S. (2020). Hidden kinetic traps in multidomain folding highlight the presence of a misfolded but functionally competent intermediate. Proc Natl Acad Sci U S A *117*, 19963-19969.
Geiler-Samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., Hartl, D.L., and Drummond, D.A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proceedings of the National Academy of Sciences of the United States of America *108*, 680-685.
Gershenson, A., Gosavi, S., Faccioli, P., and Wintrode, P.L. (2020). Successes and challenges in simulating the folding of large proteins. J Biol Chem *295*, 15-33.
Guin, D., and Gruebele, M. (2019). Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. Chem Rev *119*, 10691-10717.
Hazes, B., and Dijkstra, B.W. (1988). MODEL-BUILDING OF DISULFIDE BONDS IN PROTEINS WITH KNOWN 3-DIMENSIONAL STRUCTURE. Protein Engineering *2*, 119-125.
Hingorani, K.S., and Gierasch, L.M. (2014). Comparing protein folding in vitro and in vivo: foldability meets the fitness challenge. Current Opinion in Structural Biology *24*, 81-90.
Hong, W., Jiao, W., Hu, J., Zhang, J., Liu, C., Fu, X., Shen, D., Xia, B., and Chang, Z. (2005). Periplasmic protein HdeA exhibits chaperone-like activity exclusively within stomach pH range by transforming into disordered conformation. J Biol Chem *280*, 27029-27034.
Hsieh, C.L., Goldsmith, J.A., Schaub, J.M., DiVenere, A.M., Kuo, H.C., Javanmardi, K., Le, K.C., Wrapp, D., Lee, A.G., Liu, Y.T.*, et al.* (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. Science *369*, 1501-+.
Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering *9*, 90-95.
Jacobson, G.R., Schaffer, M.H., Stark, G.R., and Vanaman, T.C. (1973). Specific chemical cleavage in high-yield at amino peptide-bonds of cysteine and cystine residues. Journal of Biological Chemistry *248*, 6583-6591.
Jones, E.M., Lubock, N.B., Venkatakrishnan, A., Wang, J., Tseng, A.M., Paggi, J.M., Latorraca, N.R., Cancilla, D., Satyadi, M., Davis, J.E.*, et al.* (2020). Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. Elife *9*.
Kelly, C.L., Liu, Z.L., Yoshihara, A., Jenkinson, S.F., Wormald, M.R., Otero, J., Estevez, A., Kato, A., Marqvorsen, M.H.S., Fleet, G.W.J.*, et al.* (2016). Synthetic Chemical Inducers and Genetic Decoupling Enable Orthogonal Control of the rhaBAD Promoter. Acs Synthetic Biology *5*, 1136-1145.
Killikelly, A.M., Kanekiyo, M., and Graham, B.S. (2016). Pre-fusion F is absent on the surface of formalin-inactivated respiratory syncytial virus. Scientific Reports *6*.
Kosuri, P., Alegre-Cebollada, J., Feng, J., Kaplan, A., Ingles-Prieto, A., Badilla, C.L., Stockwell, B.R., Sanchez-Ruiz, J.M., Holmgren, A., and Fernandez, J.M. (2012). Protein folding drives disulfide formation. Cell *151*, 794-806.
Krshnan, L., Park, S., Im, W., Call, M.J., and Call, M.E. (2016). A conserved alpha beta transmembrane interface forms the core of a compact T-cell receptor-CD3 structure within the membrane. Proceedings of the National Academy of Sciences of the United States of America *113*, E6649-E6658.
Ladiwala, A., Xia, F., Luo, Q., Breneman, C.M., Cramer, S.M. (2006). Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models. Biotechnology and Bioengineering *93*, 836–850
Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P.J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science *355*, 12.

Li, C.K., Wen, A.Y., Shen, B.C., Lu, J., Huang, Y., and Chang, Y.C. (2011). FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. Bmc Biotechnology *11*.
Link, A.J., Robison, K., and Church, G.M. (1997). Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. Electrophoresis *18*, 1259-1313.
Liu, Y., Fu, X., Shen, J., Zhang, H., Hong, W., and Chang, Z. (2004). Periplasmic proteins of Escherichia coli are highly resistant to aggregation: reappraisal for roles of molecular chaperones in periplasm. Biochem Biophys Res Commun *316*, 795-801.
Lobstein, J., Emrich, C.A., Jeans, C., Faulkner, M., Riggs, P., and Berkmen, M. (2012). SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. Microbial Cell Factories *11*, 16.
MacCallum, J.L., Perez, A., and Dill, K.A. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proceedings of the National Academy of Sciences of the United States of America *112*, 6985-6990.
MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S.*, et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B *102*, 3586-3616.
Matsushima, Y., Irie, Y., Kageyama, Y., Bellier, J.P., Tooyama, I., Maki, T., Kume, T., Yanagita, R.C., and Irie, K. Structure Optimization of the Toxic Conformation Model of Amyloid beta 42 by Intramolecular Disulfide Bond Formation. Chembiochem.
Mayor, D., Barlow, K., Thompson, S., Barad, B.A., Bonny, A.R., Cario, C.L., Gaskins, G., Liu, Z.R., Deming, L., Axen, S.D.*, et al.* (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. Elife *5*.
McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernandez, C.X., Schwantes, C.R., Wang, L.P., Lane, T.J., and Pande, V.S. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophysical Journal *109*, 1528-1532.
McLellan, J.S., Chen, M., Joyce, M.G., Sastry, M., Stewart-Jones, G.B.E., Yang, Y.P., Zhang, B.S., Chen, L., Srivatsan, S., Zheng, A.Q.*, et al.* (2013). Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus. Science *342*, 592-598.
Metskas, L.A., and Rhoades, E. (2020). Single-Molecule FRET of Intrinsically Disordered Proteins. Annual Review of Physical Chemistry *71*, 391-414.
Molnar, K.S., Bonomi, M., Pellarin, R., Clinthorne, G.D., Gonzalez, G., Goldberg, S.D., Goulian, M., Sali, A., and DeGrado, W.F. (2014). Cys-Scanning Disulfide Cross linking and Bayesian Modeling Probe the Transmembrane Signaling Mechanism of the Histidine Kinase, PhoQ. Structure *22*, 1239-1251.
Nussinov, R., Tsai, C.J., and Jang, H. (2019). Protein ensembles link genotype to phenotype. PLoS Comput Biol *15*, e1006648.
Orban-Nemeth, Z., Beveridge, R., Hollenstein, D.M., Rampler, E., Stranzl, T., Hudecz, O., Doblmann, J., Schlogelhofer, P., and Mechtler, K. (2018). Structural prediction of protein models using distance restraints derived from cross-linking mass spectrometry data. Nature Protocols *13*, 478-494.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.*, et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825-2830.
Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Henin, J., Jiang, W.*, et al.* (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. Journal of Chemical Physics *153*.
Porter, L.L., and Looger, L.L. (2018). Extant fold-switching proteins are widespread. Proceedings of the National Academy of Sciences of the United States of America *115*, 5968-5973.
Riley, T.P., Chou, H.T., Hu, R.Z., Bzymek, K.P., Correia, A.R., Partin, A.C., Li, D.Q., Gong, D.Y., Wang, Z.L., Yu, X.C.*, et al.* (2021). Enhancing the Prefusion Conformational Stability of SARS-CoV-2 Spike Protein Through Structure-Guided Design. Frontiers in Immunology *12*.

Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., and Marks, D.S. (2019). Inferring protein 3D structure from deep mutation scans. Nature Genetics *51*, 1170-+.
Rubin, A.F., Gelman, H., Lucas, N., Bajjalieh, S.M., Papenfuss, A.T., Speed, T.P., and Fowler, D.M. (2017). A statistical framework for analyzing deep mutational scanning data. Genome Biology *18*.
Sali, A., and Blundell, T.L. (1993). COMPARATIVE PROTEIN MODELING BY SATISFACTION OF SPATIAL RESTRAINTS. Journal of Molecular Biology *234*, 779-815.
Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O.*, et al.* (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397-+.
Serebryany, E., Woodard, J.C., Adkar, B.V., Shabab, M., King, J.A., and Shakhnovich, E.I. (2016). An Internal Disulfide Locks a Misfolded Aggregation-prone Intermediate in Cataract-linked Mutants of Human gamma D-Crystallin. Journal of Biological Chemistry *291*, 19172-19183.
Serebryany, E., Yu, S.H., Trauger, S.A., Budnik, B., and Shakhnovich, E.I. (2018). Dynamic disulfide exchange in a crystallin protein in the human eye lens promotes cataract-associated aggregation. Journal of Biological Chemistry *293*, 17997-18009.
Shirts, M.R., and Chodera, J.D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. Journal of Chemical Physics *129*.
Smith, A.E., Zhou, L.Z., Gorensek, A.H., Senske, M., and Pielak, G.J. (2016). In-cell thermodynamics and a new role for protein surfaces. Proceedings of the National Academy of Sciences of the United States of America *113*, 1725-1730.
Stull, F., Hipp, H., Stockbridge, R.B., and Bardwell, J.C.A. (2018). In vivo chloride concentrations surge to proteotoxic levels during acid stress. Nature Chemical Biology *14*, 1051-+.
Taguchi, Y., Lu, L., Marrero-Winkens, C., Otaki, H., Nishida, N., and Schatzl, H.M. (2018). Disulfide-crosslink scanning reveals prion-induced conformational changes and prion strain-specific structures of the pathological prion protein PrP(Sc). J Biol Chem *293*, 12730-12740.
Tapley, T.L., Korner, J.L., Barge, M.T., Hupfeld, J., Schauerte, J.A., Gafni, A., Jakob, U., and Bardwell, J.C.A. (2009). Structural plasticity of an acid-activated chaperone allows promiscuous substrate binding. Proceedings of the National Academy of Sciences of the United States of America *106*, 5557-5562.
Tian, J., Woodard, J.C., Whitney, A., and Shakhnovich, E.I. (2015). Thermal Stabilization of Dihydrofolate Reductase Using Monte Carlo Unfolding Simulations and Its Functional Consequences. Plos Computational Biology *11*.
Tyagi, R., Lai, R., and Duggleby, R.G. (2004). A new approach to 'megaprimer' polymerase chain reaction mutagenesis without an intermediate gel purification step. Bmc Biotechnology *4*.
Untergasser, A. (2008). Preparation of Chemical Competent Cells. In Untergasser's Lab (http://www.untergasser.de/lab/protocols/competent_cells_chemical_v1_0.htm).
Uversky, V.N. (2019). Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Frontiers in Physics *7*, 18.
Waskom, M.L. (2021). seaborn: statistical data visualization. The Journal of Open Source Software *6*.
Wu, J., and Watson, J.T. (1997). A novel methodology for assignment of disulfide bond pairings in proteins. Protein Science *6*, 391-398.
Wu, Z.X., Cai, X.J., Zhang, X., Liu, Y., Tian, G.B., Yang, J.R., and Chen, X.S. (2022). Expression level is a major modifier of the fitness landscape of a protein coding gene. Nature Ecology & Evolution *6*, 103-+.
Xia, Y.Z., Li, K., Li, J.J., Wang, T.Q., Gu, L.C., and Xun, L.Y. (2019). T5 exonuclease-dependent assembly offers a low-cost method for efficient cloning and site-directed mutagenesis. Nucleic Acids Research *47*.
Xiong, X.L., Qu, K., Ciazynska, K.A., Hosmillo, M., Carter, A.P., Ebrahimi, S., Ke, Z.L., Scheres, S.H.W., Bergamaschi, L., Grice, G.L.*, et al.* (2020). A thermostable, closed SARS-CoV-2 spike protein trimer. Nature Structural & Molecular Biology *27*, 934-+.
Yang, J.S., Chen, W.W., Skolnick, J., and Shakhnovich, E.I. (2007). All-atom ab initio folding of a diverse set of proteins. Structure *15*, 53-63.

Yu, X.C., Yang, C.F., Ding, J.V., Niu, X.G., Hu, Y.F., and Jin, C.W. (2017). Characterizations of the Interactions between Escherichia coil Periplasmic Chaperone HdeA and Its Native Substrates during Acid Stress. Biochemistry *56*, 5748-5757.
Zhai, Z., Wu, Q., Zheng, W., Liu, M., Pielak, G.J., and Li, C. (2016). Roles of structural plasticity in chaperone HdeA activity are revealed by (19)F NMR. Chem Sci *7*, 2222-2228.