SUPPLEMENTARY INFORMATION

for

Understanding the physical properties controlling protein crystallization based on analysis of large-scale experimental data

W. Nicholson Price II [1,2], Yang Chen [1,2], Samuel K. Handelman [1,2], Helen Neely [1,2], Philip Manor [1,2], Richard Karlin [1,2], Rajesh Nair [1,3], Jinfeng Liu [1,3], Michael Baran [1,4], John Everett[1,4], Saichiu N. Tong [1,4], Farhad Forouhar [1,2], Swarup S. Swaminathan[1,2], Thomas Acton [1,4], Rong Xiao [1,4], Joseph R. Luft [1,5], Angela Lauricella [1,5], George T. DeTitta [1,5], Burkhard Rost [1,3], Gaetano T. Montelione [1,4,6], and John F. Hunt [1,2]*

[1] Northeast Structural Genomics Consortium;

[2] Department of Biological Sciences, 702A Fairchild Center, MC2434, Columbia University, New York, NY 10027;

[3]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032;

[4] Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, New Jersey 08854;

[5] Hauptman-Woodward Institute, 700 Ellicott Street Buffalo NY 14203;

and

[6] Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey 08854.

***** Corresponding author: (212)-854-5443 voice; (212)-865-8246 FAX; jfhunt@biology.columbia.edu.

Running Title: Physical properties controlling protein crystallization

Supplementary Notes

Complexities in evaluating the relationship between protein stability and crystallization propensity. A longstanding strategy to obtain a crystal structure after screening of a mesophilic protein has failed is to switch to an orthologous protein from a hyperthermophilic organism based on the premise that it will have greater stability and therefore crystallize more readily. While this strategy has succeeded in many individual cases, switching to an orthologous protein from a different mesophilic organism also yields a crystal structure in many cases. Therefore, a properly controlled study is required to determine whether the hyperthermophile strategy is statistically justified and whether hyperthermophilic proteins crystallize significantly more avidly than mesophilic homologues.

The hyper-thermostable proteins in our thermal denaturation dataset (12 with T_m's $\geq 80^\circ$ C) come from both mesophilic and thermophilic organisms and yielded crystal structures at a higher frequency (67%) than those with melting temperatures between 30 and 80 ˚C (37%) (Fig. 1). Logistic regression analyses (summarized in the legend to Fig. 1) suggest that this difference is likely to be significant. However, even if this difference turns out to be reproducible and hyper-thermostable proteins do indeed crystallize better, it does not imply that thermodynamic stability *per se* is a mechanistically important parameter influencing crystallization. There are systematic biases in the structural properties of hyperthermophilic proteins which could promote crystallization for reasons not directly related to their affect on global thermodynamic stability. For example, hyperthermophilic proteins have a significantly higher frequency of cooperative salt-bridging / hydrogen-bonding (H-bonding) networks, which alter the balance of offsetting enthalpic and entropic forces determining the net change in free energy upon folding[1, 2]. By immobilizing high entropy sidechains (*i.e.*, lys, arg, and glu), such cooperative interaction networks will enable them to participate in crystal packing interaction without further loss of entropy. The sequence analysis results presented in the main body of this paper demonstrate that the entropy of surface-exposed sidechains is a major determinant of protein crystallization propensity. To the extent that hyperthermophilic proteins tend to have reduced surface sidechain entropy in the native conformational state, or to have systematic biases in other physical properties influencing crystallization (*e.g.*, backbone disorder of solvent-exposed loops), they could crystallize more avidly for reasons not directly determined by their global thermodynamic stability. Broader and deeper analyses will be required to achieve a reliable understanding of the crystallization properties of hyper-thermostable proteins.

Another widely applied strategy to obtain a crystal structure from an otherwise refractory protein is to add a high-affinity ligand, which will thermodynamically stabilize the native conformational state of the protein[3], or even to add components to the buffer that increase the net thermodynamic stability of the protein[4, 5]. While both of these strategies improve crystallization results for specific proteins, there are also explanations for these results that are not directly related to the effects of the additives on global protein stability. Surface loops involved in ligand binding are often partially disordered in the absence of the cognate ligand, which represents a general strategy to decouple binding specificity from affinity[6]. Ligand binding to such sites has been demonstrated to reduce the conformational flexibility of the constituent loops in many proteins (see Forouhar *et al.*[7] and Wang and Palmer[8] for two specific examples of this ubiquitous

phenomenon). Stabilizing salts can similarly bind to and immobilize flexible surface loops. Even cosmotropic solutes like glycerol, which inhibit backbone exposure to solvent and thereby promote greater compactness[9, 10], are likely to reduce the dynamics of surface loops. Therefore, stabilizing ligands and solutes could also promote crystallization based on their effect on surface entropy rather than global thermodynamic stability. Once again, additional analyses will be required to clarify their mechanism of action in promoting protein crystallization.

Detailed analysis of proteolytic susceptibility data. While several measures of overall protease resistance do not correlate with crystallization outcome, the size of the dominant protected fragment shows a significant positive correlation with crystallization success (Supplementary Fig. 1). Strikingly, proteins that were completely digested crystallized essentially as well as proteins whose dominant resistant fragment is close to the size of the intact protein, consistent with the observation that low stability proteins crystallize well (Fig. 1). However, proteins giving a dominant fragment from 10-40% of the size of the intact protein did not crystallize at all in our dataset. A smaller size should correlate with having a larger number of dynamic loops at internal positions in the protein sequence. The datamining analyses presented in the main text show a significant anti-correlation between fraction of predicted disordered residues and crystallization success. Therefore, the experimental observation that the dominant proteolytic fragment obtained under these conditions is $\leq 40\%$ the size of the intact protein is likely to indicate that disordered internal loops are a barrier to crystallization of a given target.

Correlations between sequence parameters and thermodynamics properties. Our large scale experimental analyses allow evaluation of potential correlations between thermodynamic and sequence properties of biochemically well behaved proteins (Supplementary Figs. 2-3 and Supplementary Table 1). Notably, our data on primarily bacterial proteins do not support Uversky's conclusion[11] that specific combinations of hydrophobicity and net charge reliably identify natively unfolded proteins. Nearly all NESG proteins predicted by this metric to be unfolded are observed to be folded, while nearly all the proteins observed to be unfolded are predicted by this metric to be folded (Supplementary Fig. 4). Moreover, mean hydrophobicity (GRAVY[12] – GRand AVerage of hydropathY) shows little correlation with protein stability in our dataset (Supplementary Fig. 10), a controversial issue in previous literature[13-15]. Protein T_m and $\Delta G_{unfolding}$ are strongly and significantly positively correlated (Supplementary Fig. 2). Overall protease resistance is marginally correlated with T_m and uncorrelated with $\Delta G_{unfolding}$ – probably attributable to complexities of the proteolysis process, especially the roles of surface loops and progressive destabilization after initial cleavage of the polypeptide backbone. Complex correlations between proteolytic susceptibility and protein sequence content are described in the legend to Supplementary Figure 1.

Crystallization propensity as an intrinsic property of individual proteins. Proteins failing to give a hit in initial crystallization screening can certainly yield a crystal structure if further effort is invested in the crystallization process. Leading strategies to achieve this goal involve adding stabilizing ligands and solvents (as described in the first section above in the Supplementary

Notes), reengineering the termini of the protein construct, or switching to an orthologous proteins from another organism. All of these strategies directly or indirectly alter the surface properties of the protein and are consistent with crystallization propensity being an intrinsic property of the protein construct significantly determined by the prevalence of well ordered surface epitopes capable of mediating inter-protein packing interactions. Nonetheless, in many cases, manipulation of solution conditions has been used to obtain high quality protein crystals from a construct that was refractory to crystallization in initial screening. Beyond the addition of stabilizing ligands, which are likely to directly alter protein surface entropy, McPherson *et al.* have described an elegant solution-centered strategy to promote protein crystallization that focuses on the addition of small molecule reagents likely to promote non-covalent protein interactions by sticking simultaneously to two protein surfaces across a packing interface[16]. Such small molecules presumably promote strong interaction between surfaces epitopes that are not themselves chemically complementary. Moreover, in many historical cases, simply extending the scope of brute-force random crystallization screening has successfully yielded a crystal structure. These results reinforce the obvious conclusion that the solution environment plays a critical role in the protein crystallization process and can even be the critical determining factor in crystallizing specific proteins. However, the data presented in Supplementary Figure 5 demonstrate that the probability of obtaining a crystal structure increases as the number of conditions that must be screened before obtaining a crystal hit decreases. These results statistically validate the longstanding dogma among practicing crystallographers that the probability of getting a crystal structure is reduced for each screen that has been conducted on a protein construct without getting a hit. Moreover, they demonstrate that crystallization propensity can indeed be considered an intrinsic property of a protein construct. A construct refractory to crystallizing can still yield a crystal structure based on brute-force or strategically clever manipulation of crystallization conditions. However, the difficulty in finding a suitable crystallization condition and the probability in succeeding in crystal structure determination is demonstrated by the data in Supplementary Figure 6 to be a property of the protein construct.

Brief explanation of logistic regression and the meaning of "predictive value". Logistic regression transforms the binary success/failure outcome into a continuous variable, *i.e.*, the log of the ratio of success to failure in bins across the range of the independent variable. Following this transformation, it functions equivalently to standard regression procedures. Although logistic regression constrains the predictive effect of the continuous variable to follow a simple monotonic functional form, it provides a statistically rigorous estimate of the probability that the value of the variable influences outcome. The different sequence parameters considered in our work vary considerably in both their mean magnitudes (from fractional amino acid content to protein chain length) and relative ranges. These variations prevent direct comparison of regression slopes to evaluate either mechanistic importance or practical influence determining outcome. However, multiplying the regression slope by the standard deviation of the distribution of parameter values in the dataset provides a measure of the predictive value of the parameter. This measure (labeled "SD*Slope" in table headings) provides a practical assessment of the relative importance of each parameter in influencing outcome by scaling for both its magnitude and range of variation.

Limitations of the current analysis in detecting the influence of rare amino acids. The fractional content of either trp or his residues shows a strong slope in logistical regression against success in crystal structure determination (Fig. 3b), roughly equivalent to the slope observed for gly, indicating that these amino acids could potentially have an equivalently strong effect in promoting protein crystallization. However, as opposed to the result for gly, the uncertainty in slope for both trp and his is large enough that their positive correlation with crystallization success is not statistically significant within our dataset. The larger uncertainty is attributable to their low overall frequency of occurrence compared that of the amino acids showing statistically significant correlations (*i.e.*, 1.0% for trp and 2.4% for his *vs.* 6.8% for gly, as shown in Supplementary Table 3). Therefore, the size of the current dataset limits the ability to reliably establish the effects of rare amino acids on protein crystallization propensity. In this context, it is not possible to make reliable inferences from the current analysis concerning the influence of cys, his, pro, or trp on protein crystallization propensity. Note that even the most common amino acid pair occurs occur at a lower frequency than individual his residues, meaning that reliable quantitation of the effects of more complex sequence epitopes will require a substantially larger dataset or an alternative analytical approach (unless their influence on crystallization is substantially more powerful than observed for the most strongly correlated individual amino acids).

Weak anti-correlation of some high entropy sidechains with crystallization propensity. Arg, gln, and met all have high sidechain entropy. Consistent with our hypothesis regarding the inhibitory effect of high entropy surface-exposed sidechains on crystallization propensity, the fractional content of each of these residues is anti-correlated with crystallization success (Fig. 3). However, these trends are not statistically significant within our dataset. The failure of met's anti-correlation to be significant is likely to be attributable to its low abundance in proteins, especially in surface-exposed positions (Supplementary Table 3). Given the complex interplay of factors influencing protein crystallization, the effects of rare amino acids classes cannot be reliably determined from a dataset of the current size. (See the next section for further discussion of this point.) However, arg and gln occur at high frequencies similar to those of lys and glu at predicted exposed sites in the 679 proteins under analysis (Supplementary Table 3), and they would likely have given a stronger signal if they had a similarly strong effect in opposing successful crystallization. These observations make it likely that they do not oppose crystallization as strongly as lys and glu, residues with similarly high entropy sidechains.

We hypothesize that the weak effect of arg and gln in opposing crystallization in our dataset is attributable to the functional groups on their sidechains having a favorable interaction tendency that offsets the entropic cost of immobilizing them in crystal packing contacts. The functional groups on both of these amino acids are chemically similar to the most widely used protein denaturants. The positively charged guanidino group in arg differs from the guanidinium ion only by the substitution of a single nitrogen-hydrogen bond by a nitrogen-carbon bond, while the amide group in gln matches two of the three functional groups in the urea molecule. The well known properties of guanidinium and urea in denaturing proteins establishes that these chemical structures interact strongly with proteins in aqueous solutions. Notably primary amines

and small carboxylic acids, compounds are would be most similar to the functional groups in lys and glu, do not have similar denaturation effects. These qualitative chemical observations support our inference that the functional groups on arg and gln are likely to have favorable interaction tendencies that offset their unfavorable sidechain entropy properties. This inference is further supported by a preliminary analysis of all crystal packing contacts in the PDB which shows a high content of arg, gln, and gly in interprotein interfaces (Naumov, Price, Handelman, and Hunt, unpublished results).

This reasoning suggests that asn, which shares a functional group with gln but has lower sidechain entropy, should be more favorably correlated with crystallization propensity than gln, but this trend is not observed in our dataset (Fig. 3). The failure of asn to correlate more positively with crystallization than gln could be attributable to some stereochemical complexity, *e.g.*, a strong tendency for its functional group to make intra-protein interactions that limits its availability to mediate inter-protein contacts. More research will be required to understand the role of asn in influencing protein crystallization. This residue was previously observed to correlate with overall NESG pipeline success in an analysis of a small portion of the proteins considered here that conflated solubility and crystallization effects[17]. We assume that this previously observed correlation was either attributable to the positive influence of asn on solubility or an artifact of the limited size of the dataset used for that analysis.

Localization of gly residues promoting successful crystallization. We used secondary-structure prediction by PHD/PROF[18] to gain more insight into the structural properties of the predicted buried gly residues that promote successful protein crystallization (Supplementary Table 4). Gly predicted to be located in loops greater than 15 residues in length negatively correlate with crystallization success, while those predicted to be in other locations positively correlate (*i.e.*, in α-helices, β-sheets, loops 1-5 residues in length, and loops 6-15 residues in length). However, the only correlation to be statistically significant in our dataset is with loops 6-15 residues in length, where the positively correlated population strongly segregates into the predicted buried class (Supplementary Table 4). This categorization seems inconsistent because loops are overwhelmingly surface-localized. Manual inspection of 50 randomly chosen predicted "buried" loop glycines in PDB structures showed that 33 had at least some surface exposure (see examples in Supplementary Fig. 14). Two others were actually buried in an interprotein packing interface.

This effect could be attributable to several different factors. For the proteins in our dataset yielding structures, we explicitly checked the accuracy of the surface exposure predictions by comparison to the values calculated by DSSP[19], the program that generated the surface-exposure data used to train PHD/PROF[18]. The accuracy of the predictions were slightly worse for gly than other amino acids but consistent with the published rates above 70%. However, our qualitative visual analyses included backbone atoms, while the DSSP surface exposure metric was optimized to analyze exposure of the amino acid sidechain. The fact that gly has no sidechain could account for complex behavior in classifying it as exposed or buried. However, PHD/PROF[20] was trained on the asymmetric units of crystal structures, in which residues mediating crystal packing contacts will have some tendency to be buried in the

interfaces between non-crystallographic symmetry mates. Therefore, it is possible that PHD/PROF could have a bias leading to surface-exposed residues preferentially mediating interprotein contacts being categorized as buried. In this context, we note that predicted buried phe residues correlate positively with successful crystallization, even though predicted exposed phe residues correlate more strongly. This highly hydrophobic and relatively low entropy residue seems likely to preferentially mediate crystal-packing contacts when in a surface-exposed location. The positive correlation observed for the predicted buried population suggests that there may also be some bias in classifying certain phe residues. Finally, the results presented below using genomic distributions to evaluate parameters correlated with successful crystal structure determination of human proteins, several additional predicted buried residue populations are positively correlated with successful crystal structure determination (Supplementary Fig. 20). Additional investigations will be required to determine whether PHD/PROF has a systematic tendency to classify residues mediating crystal packing contacts as buried.

Analyses of crystallization propensity in whole proteomes. To evaluate the generalizability of P_{XS} and explore the interplay between potentially conflicting factors influencing successive steps in the structure determination process, all non-redundant sequences from the human or *E. coli* proteomes were analyzed, excluding those predicted to have a signal peptide or transmembrane α-helix. Identical filtering methods were applied to the sequences of protein constructs from those organisms that have yielded crystal structures deposited into the PDB, and logistic regressions were performed comparing parameter values in this "In-PDB" set to the corresponding proteome.

With one notable exception (described in the next section), the individual parameter trends observed in the NESG training and validation datasets are recapitulated in these empirical proteome-wide datasets, with most being statistically significant but reduced in strength (Supplementary Figs. 16-18). Notably, $\mathbf{P_{XS}}$ values from the set of successfully crystallized sequences are significantly higher than those from the corresponding proteome (Supplementary Fig. 16), indicating that the metric has some predictive value for genomic sequences in spite of the complexities cited above. The weaker predictive strengths of the metric and the underlying sequence parameters are likely to derive at least in part from offsetting influences of some parameters on protein solubility *vs.* crystallization (unpublished results). Several parameters opposing successful crystallization are known to enhance protein solubility (fractions of lys and glu), while one that promotes successful crystallization is known to cause severe solubility problems in some proteins[21] (fraction phe).

Human but not bacterial proteins contain a high population of disordered residues with low <SCE>. Surprisingly, human In-PDB sequences show an inverse trend in <SCE> compared to the corresponding proteome distribution (Supplementary Fig. 17), *i.e.*, higher in the In-PDB set even though this parameter is strongly skewed towards lower values for successfully crystallized proteins in the predominantly bacterial NESG dataset. Analyzing individual amino acid frequencies shows that most of those that are significantly predictive in the NESG dataset display qualitatively equivalent trends in the human genomic dataset (Supplementary Fig. 18), including

those whose influence was inferred to be attributable to SCE effects. These results demonstrate that the same physical and mechanistic effects influence crystallization of human proteins, probably including exposed <SCE>.

We inferred that human proteins are likely to have a large content of crystallization-inhibiting sequences with low <SCE> that are uncommon in bacterial proteins. Re-analyzing human proteome sequences after removing predicted disordered regions substantially shifts the <SCE> distribution towards higher values, giving a distribution that looks similar to that in *E. coli* proteins either with or without predicted disordered regions removed (Supplementary Fig. 17). This analysis demonstrates that human but not *E. coli* proteins have a very high prevalence of low SCE residues in disordered sequences (Supplementary Fig. 19), especially gly and pro residues which have zero SCE. While gly strongly promotes crystallization when located in ordered loops (Supplementary Table 4), it is likely to have the opposite effect in disordered regions. Indeed, multiple logistic regression on human In-PDB sequences *vs.* complete proteome confirms that both gly and pro content in predicted disordered regions strongly oppose crystal structure determination (data not shown).

$P_{XS\text{-}C\text{-}Hs}$: a conflated solubility / crystallization propensity measure for human proteins. The distinct sequence properties of human proteins, especially the prevalence of low SCE residues in crystallization-inhibiting disordered sequences, are likely to limit the reliability of the $\mathbf{P_{XS}}$ metric in predicting their crystallization behavior from genomic sequence. Therefore, we applied equivalent statistical tools to the differences in In-PDB *vs.* genomic sequence distributions to develop a crystallization prediction metric specific for human genomic sequences that we call $\mathbf{P_{C\text{-}XS\text{-}}}_{\boldsymbol{Hs}}$ (Supplementary Fig. 20). This metric conflates expression, solubility, and crystallization effects, which means that it is uninformative relative to crystallization mechanism. However, it is designed to provide practical guidance in obtaining crystal structures from human protein domains. In-PDB and proteome sequences were randomly assigned in a 4:1 ratio to training or validation sets. As described more fully in the Methods section, individual logistic regressions against In-PDB status were run on 92 sequence characteristics. Factors that correlated with success in single-parameter regressions at the Bonferroni-corrected significance level of 0.00054 (*i.e.*, 0.05/92) were combined in order of significance by forward stepwise regression, with a $p<0.05$ threshold for inclusion in the final multiple regression. The final conflated crystal structure solution metric for sequences from the *Homo sapiens* genome ($\mathbf{P_{C\text{-}XS\text{-}}}_{\boldsymbol{Hs}}$) provides a highly significant ordering of the training set (N = 22,190, $p<10^{-300}$), with strong calibration (insignificant Hosmer-Lemeshow lack of fit[22], $P = 0.411$) and discriminatory power (area under the ROC of 0.882). It performs almost as well on the test set (N = 5,457, $P = 1.61\text{x}10^{-67}$), with insignificant lack of fit ($P = 0.319$) and high ROC area (0.871). The webserver performing the standard $\mathbf{P_{XS}}$ calculation also performs this calculation (http://www.nesg.org/PXS/).

More detailed discussion of overall results and implications. This paper provides large-scale experimental validation of the hypothesis that surface properties are a principal determinant of

protein crystallization propensity. Previous engineering studies (*e.g.*, implementing the "surface entropy reduction" strategy) have demonstrated that manipulation of the surface properties of individual proteins can improve their crystallization[23-27]. The studies reported in this paper not only provide rigorous statistical validation of the specific assumptions behind this strategy but also extend the underlying concept. They demonstrate that the mean sidechain entropy (<SCE>) of surface-exposed residues is a highly significant determinant of protein crystallization propensity (Fig. 4b). Previously reported epitope engineering work has focused primarily on eliminating lysine, glutamate, and glutamine (three of the four highest entropy residues)[23-27], while the studies reported here show that reducing <SCE> of all surface-exposed residues correlates with increased probability of producing high quality crystals.

Moreover, the detailed statistical analyses presented above suggest that this surface entropy effect is responsible for other sequence correlations with crystallization propensity that were previously observed[17, 28-31] but not mechanistically explained. While increasing pI or mean hydrophobicity reduces crystallization probability[28, 30], these parameters both correlate with higher <SCE>, and their statistical significance in predicting crystallization propensity is lost when considered simultaneously with <SCE> (Table 1D and Supplementary Fig. 13). Combining the statistical dominance of <SCE> with the significant anti-correlation between crystallization propensity and predicted protein backbone disorder indicates that the prevalence of well-ordered surface epitopes is a major determinant of protein crystallization propensity, presumably because of the potential of such epitopes to mediate stereochemically specific interprotein packing interactions. Given our observation that thermodynamic stability does not have a major influence on crystallization outcome (Fig. 1), we hypothesize that the prevalence of such epitopes is the dominant determinant of protein crystallization propensity. On this basis, we propose that future research should focus on deeper understanding of the crystal-packing potential of linear sequence epitopes with the goal of generating more sophisticated sequence-based probabilistic assessments of crystallization propensity to apply to target selection and crystallization epitope engineering.

While the $\mathbf{P_{XS}}$ metric reported above represents an initial step towards this goal, the size of our existing experimental database fundamentally limits us to considering average sequence properties plus single amino acid effects. The content of many individual amino acids in specific secondary-structure / exposure classes is too low in this dataset to reliably assess their effect on crystallization propensity (Supplementary Table 4), as is the frequency of the most prevalent amino acid pairs (as discussed above). Ongoing expansion in the size of well curated crystallization databases will help detect more complex sequence correlations. However, other approaches will probably be required to detect correlations with longer linear sequence epitopes located in specific secondary structures, which are likely to be most powerful in predicting and engineering protein crystallization properties. Comprehensive characterization of the packing interactions observed in existing crystal structures could provide relevant insight.

In the meantime, the results reported in this paper provide potential hints regarding the nature of some favorable crystallization epitopes. The positive effects of gly, ala, and phe residues on crystallization propensity, which remain statistically significant even after taking into account <SCE> (Table 1D and Supplementary Fig. 13), suggests that these residues may have a strong tendencies to mediate stereochemically specific interprotein packing interactions. Ala does not make a significant independent contribution in multiple logistic regression analysis

(Table 1D), possibly due to the functional form of its influence not being well modeled by the logistic regression equation (Supplementary Fig. 8a), but it is predictive of outcome in the analysis of sets of proteins varying in their ala content while having equivalent distributions of predicted exposed <SCE> (Supplementary Fig. 13b).

Gly may be particularly effective in mediating low-affinity but stereochemically specific interprotein interactions because of enhanced exposure of the amphiphilic polypeptide backbone in the absence of a sidechain. Alternatively, its flexibility may permit "induced fit" conformational changes to optimize interaction geometry at packing sites, in which case the key gly residues are likely to be proximal to but not directly participating in interprotein interfaces. Further analyses will be required to clarify the physiochemical mechanism by which gly promotes the formation of high quality protein crystals. Ala may similarly be effective in mediating interprotein packing interactions by providing increased access to the backbone for hydrogen bond formation. Derewenda's protein engineering results support this hypothesis, since the ala introduced by mutation has been involved in backbone hydrogen bond formation within crystal packing contacts in multiple cases [27, 32].

In contrast to gly and ala, phe is strongly hydrophobic and therefore an excellent candidates to mediate hydrophobic interprotein packing interactions. However, surface-exposed phe's have been demonstrated to severely reduce the solubility of some proteins by promoting non-specific self-association[21], one form of aggregation, which we demonstrate to strongly reduce the frequency of successful protein crystallization (Fig. 2b). Therefore, phe is likely to have both positive and negative effects on crystallization propensity, depending on its exact stereochemical context as well as the other biophysical properties of individual protein.

These considerations highlight a fundamental complexity regarding protein crystallization. Obtaining a highly soluble protein preparation is the essential starting point for effective crystallization. However, the fact that a crystal represents an insoluble phase of the protein means that there is an inherent physiochemical discrepancy between these requirements. Both low solubility (*i.e.*, amorphous precipitation at low concentration) and crystallization are driven by low-affinity, non-physiological interprotein interactions. The requirements for crystallization are substantially more stringent, because multiple orientation-controlling interactions must be made simultaneously consistent with the lattice geometry, but the individual interprotein contacts stabilizing a lattice are fundamentally similar to those driving amorphous precipitation. Therefore, the sticky surface epitopes mediating the stereochemically specific interprotein interactions required to obtain a good crystal are likely to have some tendency to promote non-specific self-association, aggregation, and amorphous precipitation in the protein stock. The fact that the same protein features can simultaneously promote and compete with high quality crystallization represents a fundamental conceptual and technical conundrum. Successful crystallization requires striking a potentially elusive balance between factors promoting protein solubility and factors promoting controlled interprotein interaction.

The data presented above provide many specific examples of the physiochemical tradeoffs influencing this process. High entropy charged sidechains clearly promote solubility while opposing crystallization. The reduced predictive value of the crystallization-promoting sequence features in analyzing whole-genome results (Figs. S12-S14) is probably attributable to the pervasiveness of such offsetting effects on protein solubility and crystallizability.

Supplemental Information (continued)

Well-ordered, surface-exposed gly residues may provide an advantageous solution to the crystallization conundrum by forming epitopes capable of mediating stereochemically specific, low-affinity interprotein contacts while having a comparatively weaker tendency to promote promiscuous surface interactions. Strongly hydrophobic surface features, like phe sidechains, are likely to mediate such promiscuous surface interactions promoting protein aggregation in addition to promoting stereochemically specific interprotein contacts during crystallization. Therefore, different kinds of surface features are likely to have different effects on the delicate balance between promotion of non-specific interactions and stereochemically specific interprotein packing interactions. Further research will be required to critically evaluate the mechanistic hypotheses underlying these inferences as well as to elucidate more complex protein sequence features that optimally balance the physiochemical factors promoting the formation of high quality protein crystals.

Supplemental Information (continued)

Supplementary Methods

Protein expression, purification, and analysis. Proteins were expressed, purified, concentrated to 5-12 mg/ml, and flash-frozen in small aliquots as previously described[33]. All proteins contained short 8-residue hexa-histidine purification tags at their N- or C-termini and were metabolically labeled with selenomethionine. Matrix-assisted laser-desorption mass spectrometry was used to verify construct molecular weight. All proteins were ≥95% pure based on visual inspection of Coomasie Blue stained SDS-PAGE gels. The distribution of hydrodynamic species in the protein stock was assayed using static light-scattering and refractive index detectors (Wyatt, Inc., Santa Barbara, CA) to monitor the effluent from analytical gel filtration chromatography in 100 mM NaCl, 0.025% (w/v) NaN_3, 100 mM Tris-Cl, pH 7.5, on a Shodex 802.5 column (Showa Denko, Tokyo, Japan). Protein samples were flash frozen in liquid nitrogen in small aliquots prior to crystallization or biophysical characterization. Oligomeric state was inferred from the molecular weight determined by Debye analysis of the light-scattering data.

Target selection and classification. The 679 training and 200 validation protein sequences were culled from the SPINE database [34, 35], and included all proteins which passed aggregation screening for the selected time periods. Proteins with transmembrane α-helices predicted by TMMHMM[36] or >20% low complexity sequence were excluded from the pipeline. Proteins were coded successful if they yielded a PDB structure deposition and failures otherwise. While the reasons for failure in crystal structure solution could be complex, most are related to problems with the strength and consistency of the interprotein packing interactions in the lattice. Thus, we inferred that proteins giving diffraction that was not suitable for structure determination might have sub-optimal physical properties, and we decided to include this small sub-population among those in the failure set. Retrospective analysis of the sequence properties of the subset of proteins that give diffracting but unsolvable crystals shows that several key parameters have trends that are more similar to proteins that fail to give diffracting crystals than to proteins that give crystal structures (Supplementary Fig. 7). The consistency of the trends in this small subset with those in the larger failure set demonstrates that its inclusion is unlikely to significantly influence results.

Creation of sets with matched predicted exposed <SCE> distributions to evaluate for independent sequence effects. To determine whether some sequence parameters have an independent influence on crystallization propensity beyond their contribution to exposed sidechain entropy, the 679 proteins in the development/training set were sorted based on their predicted exposed <SCE> and then divided into 100 narrow and equally spaced bins based on this parameter value. Each of the 71 predicted exposed <SCE> bins containing at least two proteins was then sorted based on a single covarying sequence parameter. This secondary sorting was done separately for each sequence parameter being evaluated for a potentially independent effect (*i.e.*, individual amino acid frequencies and charge parameters). The two proteins in each bin with the highest and lowest values of that parameter were then assigned

respectively to the high and low value sets for that parameter. These two protein sets have nearly identical distributions of predicted exposed <SCE> but strongly separated distributions of the parameter of interest (as shown in Fig. 20). The fraction of proteins yielding a PDB deposition was calculated for each of these two sets, and the *P*-value for the probability of both sets coming from the same parent distribution was calculated using the Student's T-test.

Proteome-wide calculations. Complete proteome sequences were taken from RefSeq22[37] for humans and from Genbank[38] for *E. coli K12*. Proteins with one or more transmembrane helices predicted by the program TMHMM[36] were excluded from analysis. PDB sequences were extracted from the SEQRES field of each file in the February 15, 2008 release of the database based on the source organism identified in the header remark. The fraction of proteins in the PDB in each parameter bin was calculated as the number of PDB sequences in a particular bin divided by the number of predicted soluble proteome sequences in that bin, ignoring complexities related to domain structure and construct optimization. Rolling $\mathbf{P_{XS}}$ averages were calculated using bins of 2000 proteins starting at the lowest observed parameter value.

Conflated human crystallization metric ($P_{C\text{-}XS\text{-}Hs}$). Proteome and PDB sequences were randomly assigned at a 4:1 ratio to training or validation sets. Single logistic regressions were run to evaluate potential correlations between logical status (PDB *vs.* proteome sequence) and 92 different continuous variables calculated from the protein sequence: the count and fraction of each amino acid predicted to be buried or exposed by PHD/PROF[18], fractional and whole values of net charge, the absolute value of net charge, the number and fraction of charged residues, GRAVY, <SCE>, predicted exposed <SCE>, the fraction of the backbone predicted to be disordered by DISOPRED2[39], and sequence length. The combined model was built by stepwise forward logistic regression. Each variable with a *P*-value below a Bonferroni-adjusted cutoff of 0.00054 (0.05/92) in a single logistic regression was added to a null model in order of decreasing statistical significance and retained if its *P*-value within the model was less than 0.05. If the addition of a new variable caused a previous variable to become statistically insignificant, the variable which had a larger effect on the overall model *P*-value was retained. Each variable in the resulting model was individually removed to check for improvement in chi-squared *P*-value or Akaike's Information Criterion (AIC)[40]. All such trials worsened the model *P*-value and AIC, so all variables in the stepwise model were retained.

Chemical denaturation experiments. Crystallization stocks were diluted to 1 mg/ml in a buffer containing 50 mM NaCl, 10 mM $NaPO_4$, pH 8.0. Baseline CD spectra were measured from 190 to 300 nm in an Aviv Model 202 spectropolarimeter to evaluate folding status prior to titration. An auto-titrator was used to mix this solution with an equivalent solution containing 1 mg/ml of the same protein in the same buffer plus 8M guanidine•GdnHCl. The titration was monitored based on $\Theta_{222\,nm}$ to a final concentration of 6MGdn•HCl. Free energy of unfolding in the absence of denaturant (ΔG^0) was calculated by non-linear least-squares fitting of the equation $\Delta G_u([Gdn•HCl]) = \Delta G^0 + m[Gdn•HCl])$, with $\Delta G_u([Gdn•HCl])$ substituted by RT•ln([Folded]/[Unfolded]). The ratio of folded-to-unfolded protein concentration was inferred

at each data point from the observed $\Theta_{222\,nm}$ relative to baselines linearly extrapolated from pre- and post-denaturation regions of the titration. Proteins which showed no baseline before the unfolding transition were considered partially unfolded in the absence of denaturant.

Thermal denaturation experiments. Crystallization stocks were diluted 1:40 to a final concentration of ~0.25 mg/ml in a buffer containing 150 mM NaCl, 100 mM HEPES, pH 7.5, plus 5X SYPRO Orange dye (Invitrogen). Proteins were heated from 25˚ to 95˚ C at 1 degree/min in optically clear PCR tubes in a Stratagene Mx3005P realtime-PCR machine. Dye fluorescence was measured each minute. Denaturation was detected based on a sigmoidal increase in fluourescence emission intensity caused by binding of the dye to hydrophobic regions of the protein[41]. T_m's were identified as the temperature at which the first derivative of the fluorescence intensity was highest. Proteins which showed no baseline prior to the thermal unfolding transition were considered partially unfolded in their stock solutions. Proteins with high initial fluorescence or no visible transition had CD spectra measured in order to evaluate their folding status.

Limited proteolysis experiments. Crystallization stocks were diluted to 1 mg/ml in 50 mM NaCl, 10 mM Tris, pH 7.5 with either 0.005 mg/ml Proteinase K, 0.02 mg/ml trypsin, or no protease. Reactions were incubated at 25 degrees for 40 minutes, stopped by addition of 5 mM PMSF, and incubated for an additional 2 minutes before final quenching with SDS sample buffer and flash freezing. All proteins were digested and run on gels in parallel with two fiducial proteins, with protease conditions adjusted to produce close to total digestion of one (*B. cereus* agmatine deiminase, NESG ID BcR51) and negligible digestion of the other (*E. coli* ElbB, NESG ID ER105). A computer interface was developed for display of scanned gels and blind scoring of five characteristics for each individual reaction: number of visible bands, percent of protein remaining in the intact band, percent of protein remaining in the most intensely stained or "dominant" band (whatever its size), percent of protein remaining integrated across all visible bands, and size of the dominant band relative to the intact band. Two independent scorers agreed very closely (Pearson = 0.89), and the final score was taken as the average of the scorers and the average of the two proteases (because no significant differences were observed in single-protease metrics). No significant change in correlation strength or probability was observed when the scores were normalized to those of the fiducial control proteins run in parallel, most likely due to the consistent conditions used for all proteolysis reactions. Therefore, non-normalized scores are reported.

Bis-ANS binding experiments. The protein crystallization stock was diluted to a final concentration of 15 μM in a fluorescence cuvette containing 1.25 ml of 0.93 μM B*is*-ANS, 50 mM NaCl, 10 mM $NaPO_4$. Fluorescence spectra were acquired at room temperature in a PTI QuantaMaster C-61 spectrofluorimeter using an excitation wavelength of 375 nm. Spectra from 400-600 nm were taken every 30 seconds until stabilized, up to a maximum of 5 minutes.

Supplemental References

1. Karshikoff, A. & Ladenstein, R. Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads. *Trends in biochemical sciences* **26**, 550-556 (2001).
2. Kumar, S. & Nussinov, R. How do thermophilic proteins deal with heat? *Cell Mol Life Sci* **58**, 1216-1233 (2001).
3. Vedadi, M. et al. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15835-15840 (2006).
4. Meining, W., Scheuring, J., Fischer, M. & Weinkauf, S. Cloning, purification, crystallization and preliminary crystallographic analysis of SecA from Enterococcus faecalis. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **62**, 583-585 (2006).
5. Sousa, R. Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization. *Acta crystallographica* **51**, 271-277 (1995).
6. Endrizzi, J.A., Beernink, P.T., Alber, T. & Schachman, H.K. Binding of bisubstrate analog promotes large structural changes in the unregulated catalytic trimer of aspartate transcarbamoylase: implications for allosteric regulation. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5077-5082 (2000).
7. Forouhar, F. et al. Crystal structures of two bacterial 3-hydroxy-3-methylglutaryl-CoA lyases suggest a common catalytic mechanism among a family of TIM barrel metalloenzymes cleaving carbon-carbon bonds. *The Journal of biological chemistry* **281**, 7533-7545 (2006).
8. Wang, C., Karpowich, N., Hunt, J.F., Rance, M. & Palmer, A.G. Dynamics of ATP-binding cassette contribute to allosteric control, nucleotide binding and energy transduction in ABC transporters. *Journal of molecular biology* **342**, 525-537 (2004).
9. Gekko, K. & Timasheff, S.N. Mechanism of protein stabilization by glycerol: preferential hydration in glycerol-water mixtures. *Biochemistry* **20**, 4667-4676 (1981).
10. Gekko, K. & Timasheff, S.N. Thermodynamic and kinetic examination of protein stabilization by glycerol. *Biochemistry* **20**, 4677-4686 (1981).
11. Uversky, V.N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739-756 (2002).
12. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105-132 (1982).
13. Bigelow, C.C. On the average hydrophobicity of proteins and the relation between it and protein structure. *J Theor Biol* **16**, 187-211 (1967).
14. Goldsack, D.E. Relation of the hydrophobicity index to the thermal stability of homologous proteins. *Biopolymers* **9**, 247-252 (1970).

15. Ponnuswamy, P.K. Hydrophobic characteristics of folded proteins. *Prog Biophys Mol Biol* **59**, 57-103 (1993).

16. McPherson, A. & Cudney, B. Searching for silver bullets: an alternative strategy for crystallizing macromolecules. *Journal of structural biology* **156**, 387-406 (2006).

17. Goh, C.S. et al. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *Journal of molecular biology* **336**, 115-130 (2004).

18. Rost, B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods in enzymology* **266**, 525-539 (1996).

19. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).

20. Rost, B., Yachdav, G. & Liu, J. The PredictProtein server. *Nucleic Acids Res* **32**, W321-326 (2004).

21. Lewis, H.A. et al. Impact of the deltaF508 mutation in first nucleotide-binding domain of human cystic fibrosis transmembrane conductance regulator on domain folding and structure. *The Journal of biological chemistry* **280**, 1346-1353 (2005).

22. Hosmer, D.W. & Lemeshow, S. Applied logistic regression. (Wiley, New York; 1989).

23. Cooper, D.R. et al. Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr* **63**, 636-645 (2007).

24. Derewenda, Z.S. The use of recombinant methods and molecular engineering in protein crystallization. *Methods* **34**, 354-363 (2004).

25. Derewenda, Z.S. Rational protein crystallization by mutational surface engineering. *Structure* **12**, 529-535 (2004).

26. Derewenda, Z.S. & Vekilov, P.G. Entropy and surface engineering in protein crystallization. *Acta Crystallogr D Biol Crystallogr* **62**, 116-124 (2006).

27. Longenecker, K.L., Garrard, S.M., Sheffield, P.J. & Derewenda, Z.S. Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. *Acta Crystallogr D Biol Crystallogr* **57**, 679-688 (2001).

28. Canaves, J.M., Page, R., Wilson, I.A. & Stevens, R.C. Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. *Journal of molecular biology* **344**, 977-991 (2004).

29. Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K. & Markley, J.L. Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins* **59**, 444-453 (2005).

30. Overton, I.M. & Barton, G.J. A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett* **580**, 4005-4009 (2006).

31. Slabinski, L. et al. The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* **16**, 2472-2482 (2007).

32. Mateja, A. et al. The impact of Glu-->Ala and Glu-->Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 A resolution. *Acta Crystallogr D Biol Crystallogr* **58**, 1983-1991 (2002).

33. Acton, T.B. et al. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods in enzymology* **394**, 210-243 (2005).

34. Bertone, P. et al. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29**, 2884-2898 (2001).

35. Goh, C.S. et al. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31**, 2833-2838 (2003).

36. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-580 (2001).

37. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).

38. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36**, D13-21 (2008).

39. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138-2139 (2004).

40. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723 (1974).

41. Niesen, F.H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature protocols* **2**, 2212-2221 (2007).

42. Creamer, T.P. Side-chain conformational entropy in protein unfolded states. *Proteins* **40**, 443-450 (2000).

43. Rost, B. in The Proteomics Protocols Handbook. (ed. J.E. Walker) 875-901 (Humana Press, Totowa; 2005).

44. Delano, W.L. 2002).

Supplementary Table 1: ***Correlations between stability measurements***.[1]

Variable 1	Variable 2	Pearson	*P*-value
$\Delta G_{unfolding}$	T_m	-0.748	**0.0000000696**
$\Delta G_{unfolding}$	Percent Protein Intact	-0.264	0.144
$\Delta G_{unfolding}$	Percent Protein Remaining	-0.0442	0.810
$\Delta G_{unfolding}$	Percent Size: Dominant Fragment	-0.066	0.718
$\Delta G_{unfolding}$	Number of Bands	0.0905	0.622
$\Delta G_{unfolding}$	Percent Protein: Dominant Fragment	-0.193	0.291
T_m	Percent Protein Intact	0.304	0.0636
T_m	Percent Protein Remaining	0.247	0.135
T_m	Percent Size: Dominant Fragment	-0.0473	0.778
T_m	Number of Bands	0.0181	0.914
T_m	Percent Protein: Dominant Fragment	0.332	0.042
Fraction (K+R)	Percent Protein Intact (PK)	-0.0463	0.627
Fraction (K+R)	Percent Protein Intact (T)	-0.145	0.125
Fraction (K+R)	Percent Protein Remaining (PK)	0.328	**0.000397**
Fraction (K+R)	Percent Protein Remaining (T)	0.107	0.258
Fraction (K+R)	Percent Size: Dominant Fragment (PK)	0.132	0.163
Fraction (K+R)	Percent Size: Dominant Fragment (T)	-0.0384	0.687
Fraction (K+R)	Number of Bands (PK)	0.164	0.0831
Fraction (K+R)	Number of Bands (T)	-0.0253	0.790
Fraction (K+R)	Percent Protein: Dominant Fragment (PK)	0.0835	0.379
Fraction (K+R)	Percent Protein: Dominant Fragment (T)	-0.0708	0.456
Fraction Disordered	Percent Protein Intact	-0.0222	0.825
Fraction Disordered	Percent Protein Remaining	-0.0305	0.761
Fraction Disordered	Percent Size: Dominant Fragment	-0.0208	0.836
Fraction Disordered	Number of Bands	-0.0698	0.486
Fraction Disordered	Percent Protein: Dominant Fragment	-0.00324	0.973

[1] Pearson correlation coefficients and associated *P*-values were calculated for pairs of possibly correlated variables including a variety of metrics describing the large scale limited proteolysis results. *P*-values below 0.01 are shown in boldface type. $\Delta G_{unfolding}$ was measured by chemical denaturation with guanidinium hydrochloride monitored by circular dichroism at 222 nm. Thermal melting temperature (T_m) was measured by monitoring the fluorescence enhancement of the hydrophobic reporter dye SYPRO Orange in a realtime PCR machine[41]. Limited proteolysis assays were conducted with either proteinase K or trypsin, and the data gels were scored blind by two independent evaluators based on the following parameters: percent of intact protein remaining, percent of initial protein remaining summed across all significant bands, relative size of the dominant fragment (as percent height of the band in the gel relative to the undigested protein – not molecular weight), number of significant bands, and percent of original protein remaining in the dominant (most intense) band. All limited proteolysis variables listed are the average of proteinase K and trypsin scores, except correlations with fraction arginine and lysine ("K+R" under Variable 1) which are given separately for digestions with proteinase K ("PK" under Variable 2) or trypsin ("T" under Variable 2). A significant correlation was observed between melting temperature and $\Delta G_{unfolding}$ by chemical denaturation. A significant and surprising correlation was also observed between the fraction of lysine and arginine in the protein chain and the percent of protein remaining across all bands following proteinase K digestion. The fraction disordered was predicted by the program DISOPRED2[39] using a 5% false positive rate threshold.

Supplementary Table 2: ***Logistic regressions of charge variables.***[1]

Regression Variable	Slope	SD*Slope	*P*-value
Number of Positive Residues	-0.00529	-0.071	0.454
Number of Negative Residues	-0.000367	-0.0055	0.952
Number of Charged Residues	-0.00133	-0.037	0.692
Net Charge	-0.0182	-0.12	0.175
Absolute Net Charge	0.00270	0.014	0.876
Fraction Positive Residues	-10.7	-0.34	**0.000665**
Fraction Negative Residues	-7.57	-0.24	**0.0144**
Fraction Charged Residues	-7.36	-0.37	**0.000249**
Fractional Net Charge	-2.13	-0.081	0.376
Fractional Absolute Net Charge	-1.47	-0.042	0.650
eSCE & Frac. Pos./Neg. Residues			**0.0000192**
eSCE	-5.58	-0.441	**0.0028**
Fraction Positive Residues	-1.37	-0.044	0.742
Fraction Negative Residues	-0.488	-0.015	0.894

[1] Logistic regressions were calculated to determine the relationship between various charge measurements and the probability of crystal structure solution for 679 NESG pipeline proteins. We call the product of the slope and the variable's standard deviation the "predictive value", the significance of which is explained below in the SI. Variables describing electrostatic charge content are only significantly predictive after normalization to protein length (*i.e.*, as fractional values). However, the multiple logistic regression presented in the bottom section of the table, which analyzes mean side chain entropy (<SCE>) of predicted exposed residues together with the fractions of positive and negative residues, demonstrates that the predictive signal from the charged residues is lost when considered simultaneously with SCE. This result suggests that the latter term is mechanistically dominant.

Supplementary Table 3: ***Amino acid frequencies in the dataset of 679 NESG proteins and GRAVY/SCE values.*** [1]

Residue	Total fraction	Fraction in predicted buried class	Fraction in predicted exposed class	Fraction of predicted buried class	Fraction of predicted exposed class	GRAVY [2]	SCE [3] (kcal/mol)
Ala	0.076	0.037	0.039	0.092	0.072	1.8	0
Arg	0.049	0.004	0.045	0.01	0.076	-4.5	2.13
Asn	0.039	0.0045	0.035	0.01	0.054	-3.5	1.22
Asp	0.06	0.0035	0.056	0.0078	0.096	-3.5	0.86
Cys	0.013	0.0099	0.0026	0.027	0.0052	2.5	0.58
Gln	0.04	0.0035	0.036	0.008	0.058	-3.5	1.79
Glu	0.082	0.0029	0.079	0.006	0.125	-3.5	1.39
Gly	0.064	0.023	0.041	0.054	0.074	-0.4	0
His	0.024	0.0078	0.016	0.019	0.026	-3.2	1.27
Ile	0.064	0.052	0.012	0.138	0.017	4.5	0.77
Leu	0.093	0.073	0.02	0.196	0.033	3.8	0.54
Lys	0.066	0.0012	0.065	0.0027	0.098	-3.9	2.06
Met	0.027	0.015	0.012	0.039	0.019	1.9	1.6
Phe	0.038	0.031	0.0068	0.084	0.013	2.8	0.73
Pro	0.042	0.0087	0.033	0.02	0.06	-1.6	0
Ser	0.055	0.013	0.042	0.033	0.069	-0.8	0.6
Thr	0.05	0.014	0.037	0.033	0.057	-0.7	0.48
Trp	0.012	0.009	0.0026	0.024	0.004	-0.9	1.17
Tyr	0.032	0.021	0.011	0.055	0.015	-1.3	0.72
Val	0.069	0.052	0.017	0.14	0.027	4.2	0.41

[1] For each amino acid, its fractional content in the entire dataset is given in column 2, while the division of this fraction among the PHD/PROF-predicted buried and exposed classes is given in columns 3-4. Columns 4-5 give the factional content of each amino acid among all the residues predicted to be buried or exposed, respectively. The metrics in columns 2-4 are used for all analyses based on fractional amino acid content presented in this paper, but the normalization applied in columns 5-6 is used in calculating $\mathbf{P}_{\mathbf{XS}}$.

[2] Hydropathy index values used for calculation of GRand AVerage of hYdropathy (GRAVY)[12].

[3] Monte Carlo sidechain entropy (SCE) values at 37° C as given by Creamer[42] and used for the calculations presented in this paper.

Supplemental Information (continued)

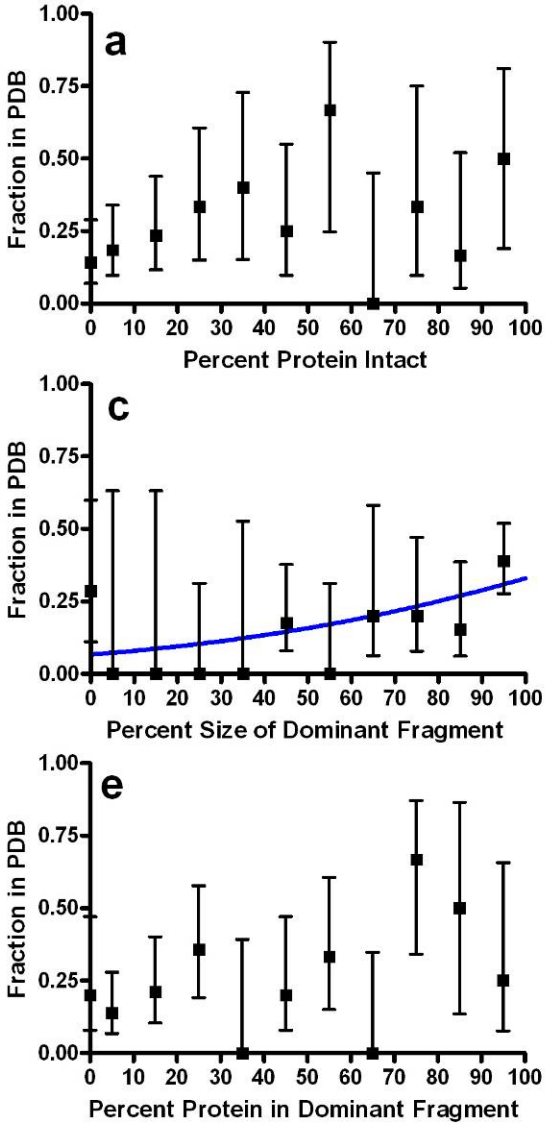
Supplementary Table 4: ***Amino acid correlations in PHD/PROF-predicted secondary structure classes.***[1]

	All	Buried	Exposed		All	Buried	Exposed
	8.8	9.6	6.6		-10.4	26	-9.8
Alanine	**0.0019**	0.025	0.057	**Glutamate**	**0.0029**	0.2	**0.0043**
Helix	0.17	1.3	1.6	Helix	-10.66	19.6	-11.3
	0.96	0.82	0.77		**0.0042**	0.542	**0.003**
Sheet	29	35	34.0	Sheet	7.69	65.1	5.8
	0.00019	**0.00049**	0.027		0.29	0.096	0.43
Loop(1-5)	18	17	17.7	Loop(1-5)	-0.51	94.9	-2.1
	0.085	0.41	0.13		0.96	0.20	0.83
Loop(6-15)	16	33	11.1	Loop(6-15)	2.10	-33.6	2.4
	0.034	0.041	0.18		0.76	0.64	0.72
Loop(>15)	11	38	7.5	Loop(>15)	-13.02	-126.6	-12.6
	0.24	0.12	0.48		0.18	0.37	0.20
	All	Buried	Exposed		All	Buried	Exposed
	10.7	18	2.9		-9.7	-13.5	-9.4
Glycine	**0.0046**	**0.00077**	0.53	**Lysine**	**0.0018**	0.69	**0.0028**
Helix	5.7	25.9	-10.6	Helix	-11.99	-3.9	-12.2
	0.5	0.042	0.40		**0.010**	0.941	**0.010**
Sheet	17	27.9	-0.5	Sheet	-6.76	41.2	-7.2
	0.082	0.025	0.98		0.37	0.571	0.35
Loop(1-5)	9.2	16.4	6.9	Loop(1-5)	-12.11	-210.6	-11.1
	0.18	0.25	0.36		0.23	0.20	0.27
Loop(6-15)	15	39.0	8.2	Loop(6-15)	-4.36	-27.6	-4.2
	0.0088	**0.00087**	0.23		0.59	0.79	0.60
Loop(>15)	-8.8	-25.5	-8.1	Loop(>15)	-98.46	-98.5	-7.0
	0.32	0.32	0.44		0.60	0.60	0.44
	All	Buried	Exposed				
	12.9	9.7	23.9				
Phenylalanine	**0.014**	0.087	0.038				
Helix	-0.51	-2.5	16.4				
	0.95	0.762	0.49				
Sheet	18.10	16.5	42.0				
	0.013	0.042	0.05				
Loop(1-5)	16.02	16.7	18.3				
	0.36	0.45	0.55				
Loop(6-15)	14.10	20.8	-3.0				
	0.25	0.15	0.91				
Loop(>15)	3.70	-16.3	53.5				
	0.83	0.48	0.11				

Total	**Slope/*P***	**Count**
Helix	-1.05	49918
	0.036	
Sheet	1.46	26232
	0.032	
Loop(1-5)	0.445	18753
	0.77	
Loop(6-15)	1.44	30890
	0.14	
Loop(>15)	-0.297	14521
	0.71	

[1] For each significantly predictive amino acid, separate logistic regressions were calculated for the prevalence of the amino in a particular PHD/PROF[18, 20, 43] predicted secondary structure, surface exposure category, or both, as a fraction of the entire protein chain length. Each cell in the table shows the slope of the logistic regression (in blue) above the *P*-value for that regression (in maroon). *P*-values at or below 0.01 are shown in boldface type. The predictive effect of alanine localizes to sheets for poorly understood reasons, but becomes statistically insignificant when combined with other features in the final metric. Glycine's effect is localized to residues predicted to be buried in loops from 6-15 residues in length; visual inspection shows these residues to be primarily surface exposed (Supplementary Fig. 14). The frequency of phenylalanine in the dataset is too low to give statistically reliable results after segregation into PHD/PROF classes. Lysine and glutamic acid are both localized to predicted exposed positions in α-helices, consistent the surface entropy hypothesis; both are redundant with SCE in combined

regressions. The sub-table at bottom right displays logistic slopes and *P*-values for the fractional content of PHD/PROF-predicted secondary structure classes along with the total count of amino acids in each class (in all proteins in the training dataset combined together).



f

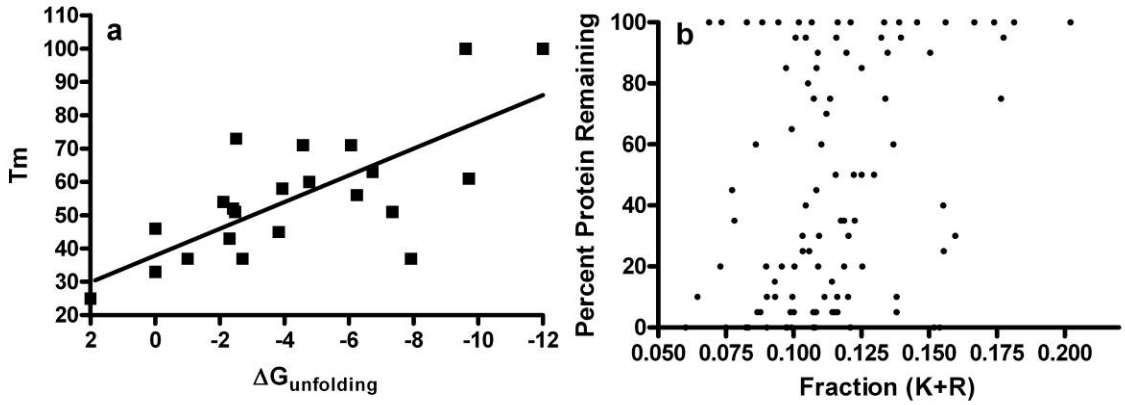
Regression Var.	Slope	SD*Slope	P val.
Prot. Intact	0.010	0.30	0.076
Prot. Remaining	0.007	0.24	0.14
Size: Dom. Frag.	0.019	0.61	**0.015**
Size (excluding 0)	0.035	0.93	**0.0012**
Number of Bands	-0.063	-0.11	0.32
Prot. in Dom. Frag.	0.013	0.0056	0.10

Supplementary Figure 1: ***Relationships between limited proteolysis results and protein crystallization outcome.*** A representative set of 114 NESG pipeline proteins were subjected to limited proteolysis using proteinase K and trypsin (in two separate reactions). Two independent evaluators scored the SDS-PAGE gels of the proteolysis products according to the criteria indicated below based on Coomasie Blue staining intensity. Scores were averaged between the evaluators and the two different proteases. Boxes indicate the fraction of proteins in the PDB for bins of each scored characteristic, while whiskers (error-bars) indicate 95% confidence limits calculated from counting statistics using the numbers in each bin. The blue line in panel **c** traces the only significant logistic relationship. **(a)** Percent of protein remaining intact (*i.e.*, at the undigested position). **(b)** Percent of starting protein mass (*i.e.*, staining) remaining in all visible bands summed together. **(c)** Size of the single most intensely stained or "dominant" fragment as a percentage of the intact protein size, based on height in the gel (not molecular weight). **(d)**

Number of visible Coomasie-stained bands. **(e)** Percent of starting protein mass (*i.e.*, staining) remaining in the single most intensely stained or "dominant" fragment. **(f)** Slopes and *P*-values for the logistic regressions, with *P*-values below 0.05 shown in boldface type.

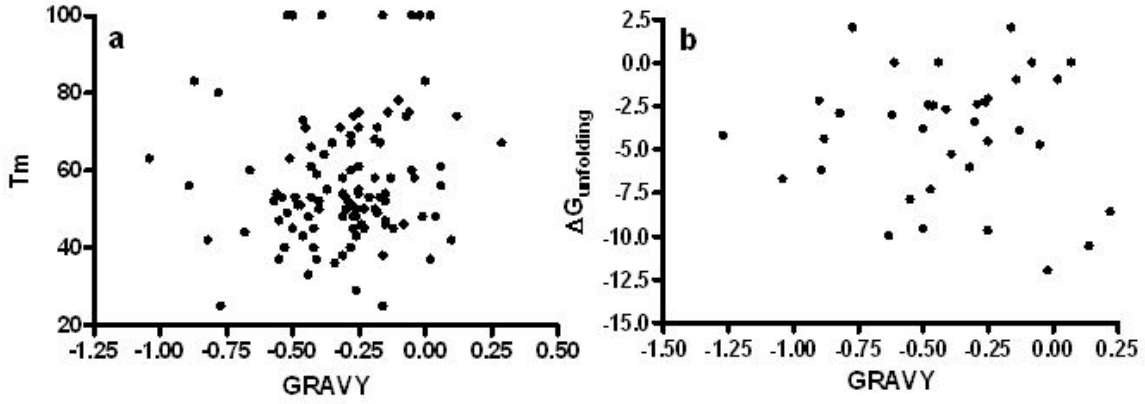
The data show a statistically significant relationship between the size of the dominant protected fragment and the likelihood of solving a crystal structure (panel **c)**. This correlation is likely to reflect the influence of disordered protein segments in inhibiting protein crystallization (Fig. 4d and Table 1B). Note that while no structures were produced by the proteins in this dataset showing protease-resistant fragments when the dominant protected fragment was less than less than 40% of the size of the intact protein, a good rate of crystal structure determination was observed for the completely digested proteins (*i.e.*, those in the 0% bin). Because of this pattern, the statistical significance and predictive value of the dominant protected fragment size are both increased when the 0% bin containing completely digested proteins is removed from logistic regression analysis (panel **f)**. Completely digested proteins are likely to have low thermodynamic stability, which does not reduce crystallization propensity relative to other folded mesophilic proteins (Fig. 1). However, those showing dominant protected fragments less than 40% of the size of the intact protein are likely to have higher stability, at least in one domain, but also have multiple internal sites accessible to protease. These proteins are likely to have a significant content of flexible surface loops and/or disordered backbone segments, which inhibit effective crystallization (Fig. 4d and Table 1B). Based on these arguments, except in the 0% bin, the size of the dominant proteolytically protected fragment is likely to be a measure the extent of flexible and disordered polypeptide segments rather than the overall stability of the protein.

Supplemental Information (continued)



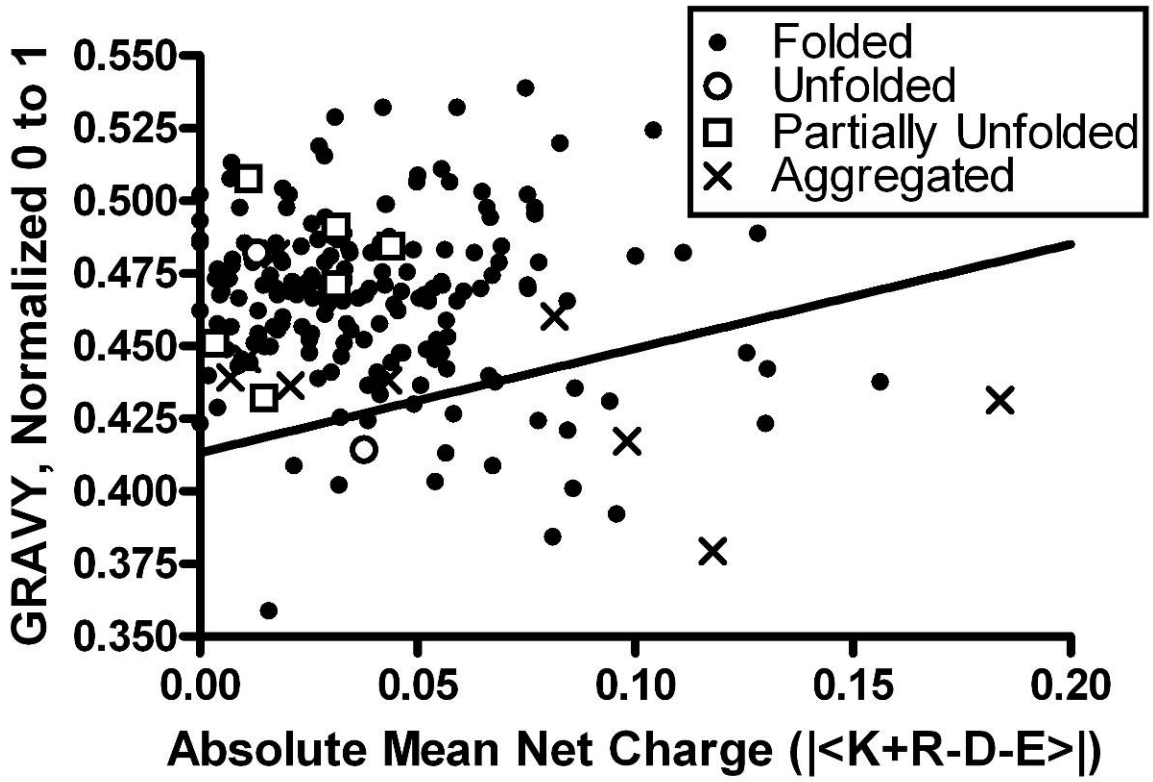
Supplementary Figure 2: ***Correlations between different stability measurements and between protease sensitivity and sequence composition***. **(a)** 23 NESG pipeline proteins were evaluated in both chemical denaturation studies and thermal unfolding studies. The plot shows a significant correlation between $\Delta G_{unfolding}$ and T_m (N = 23, *P* = 0.000027). **(b)** 114 NESG pipeline proteins underwent limited proteolysis by proteinase K. The plot shows the significant correlation between the fraction of protein remaining summed across all bands after proteinase K digestion and the fraction of arginine lysine and arginine residues ("K+R") in the protein chain (N = 114, *P* = 0.00059).

Supplemental Information (continued)



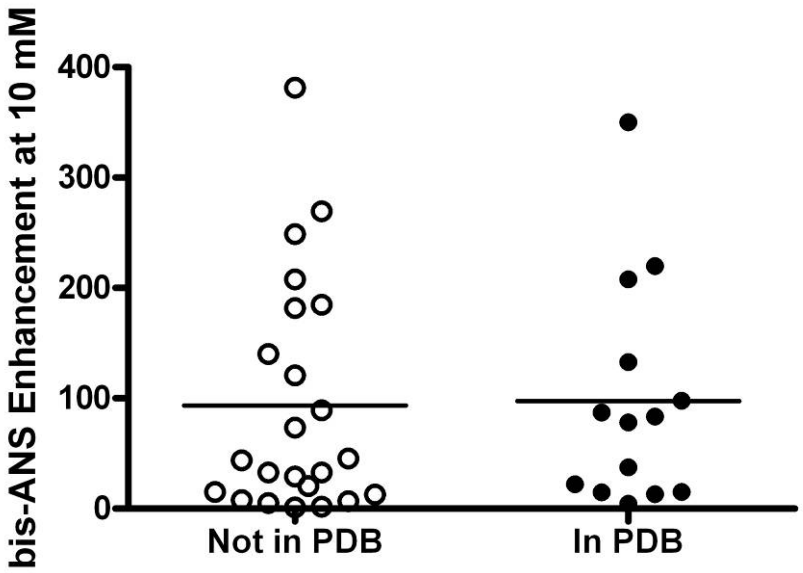
Supplementary Figure 3: ***Lack of correlation between mean hydrophobicity and protein stability***. Thermal denaturation studies were conducted on 117 NESG pipeline proteins, while chemical denaturation studies using guanidinium-HCl were conducted on 36 to estimate $\Delta G_{unfolding}$. **(a)** The plot of GRAVY (mean hydrophobicity) *vs.* thermal melting temperature (T_m) shows no significant correlation (Pearso n =0.12, N = 117, *P* = 0.19). **(b)** The plot of GRAVY *vs.* $\Delta G_{unfolding}$ shows no significant correlation (Pearson = -0.059, N = 36, *P* = 0.73).

Supplemental Information (continued)

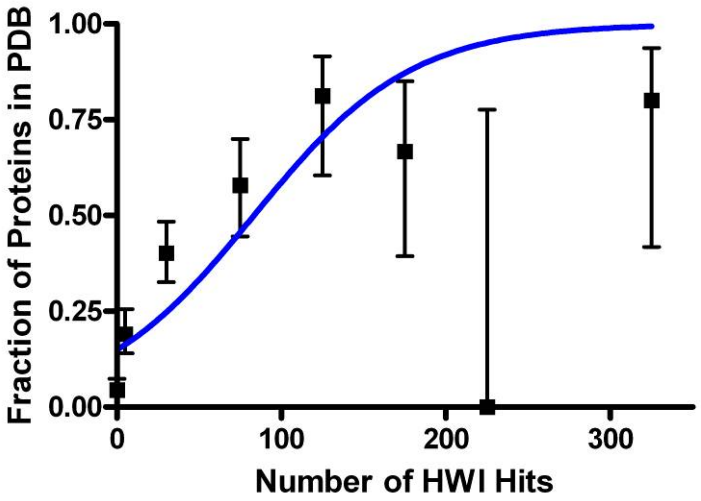


Supplementary Figure 4: ***Evaluation of the relationship between folding status and mean net charge / hydrophobicity***. Symbols encoding the folding status of each protein are positioned according to their mean net charge on the abscissa and mean hydrophobicity (GRAVY[12]) on the ordinate. Proteins classified as unfolded had a minimal CD signal and exhibited no transition in either thermal or chemical denaturation experiments, while those classified as partially unfolded exhibited no pre-transition baseline and a T_m under 35°C in thermal denaturation experiments or no pre-transition baseline in chemical denaturation experiments. Aggregated proteins were identified based on analytical gel filtration / static light scattering analysis. The black line has been posited by Uversky to segregate stably folded proteins (above the line) from natively unfolded proteins (below the line)[11]. A Fisher's exact test on the data shown here does not support that hypothesis ($p = 1$).

Supplemental Information (continued)

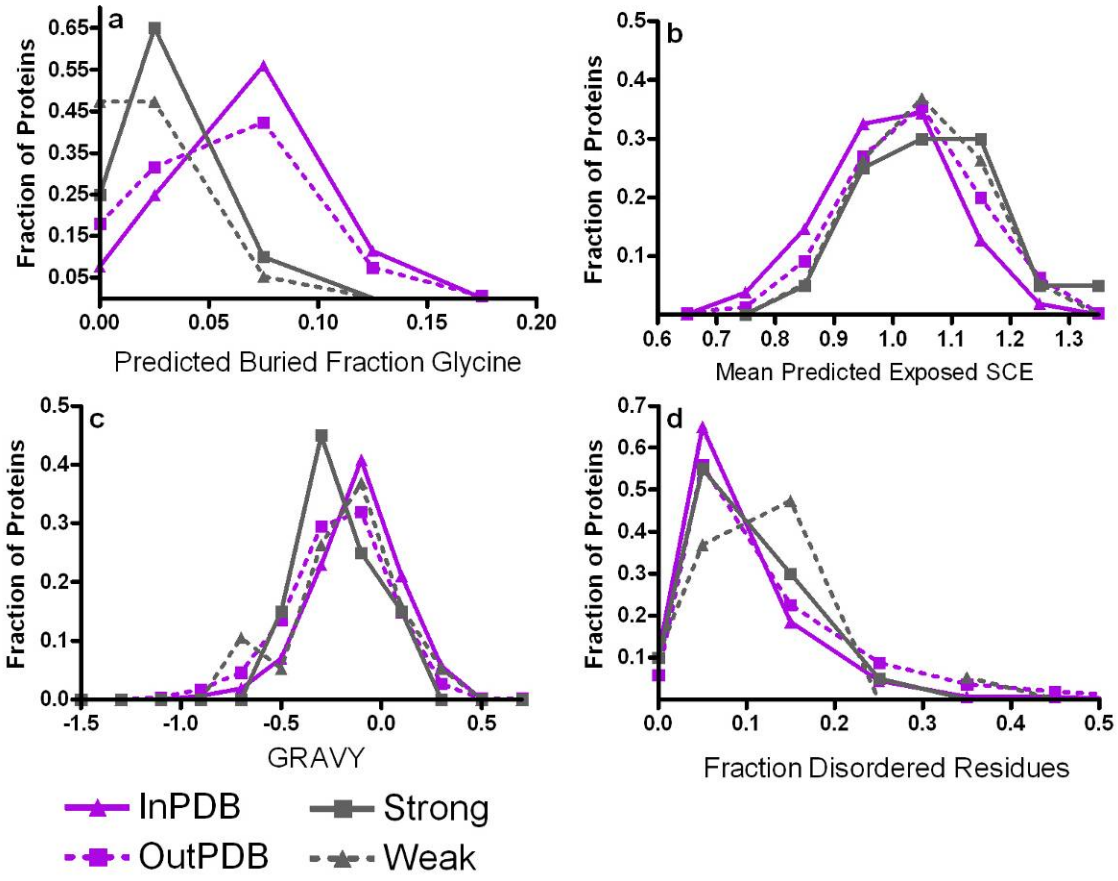


Supplementary Figure 5: ***Fluorescence enhancement of the hydrophobic reporter dye bis-ANS is not correlated with crystallization success.*** Experiments were conducted on 38 NESG pipeline proteins, which were added to a final concentration of 15 μM to a cuvette containing 0.93 μM *bis*-ANS. The dye was excited at 395 nm while fluorescence emission spectra were acquired from 400-600 nm. The enhancement was calculated as the ratio of dye fluorescence in the presence of protein to that in its absence at the peak wavelength of the emission spectrum in the presence of protein. No significant difference is observed in enhancement for proteins that yielded a crystal structure compared to those that did not ($p_{T\text{-}test}$=0.91). Among the limited set of proteins for which both stability and *bis*-ANS measurements were performed, enhancement did not correlate significantly with $\Delta G_{unfolding}$ (N = 9, P = 0.23) or T_m (N = 12, P = 0.68) according to logistic regression analysis. All proteins determined to be unfolded in their crystallization stocks exhibited greater than 100-fold fluorescence enhancement, but not all proteins with more than 100-fold enhancement were unfolded.



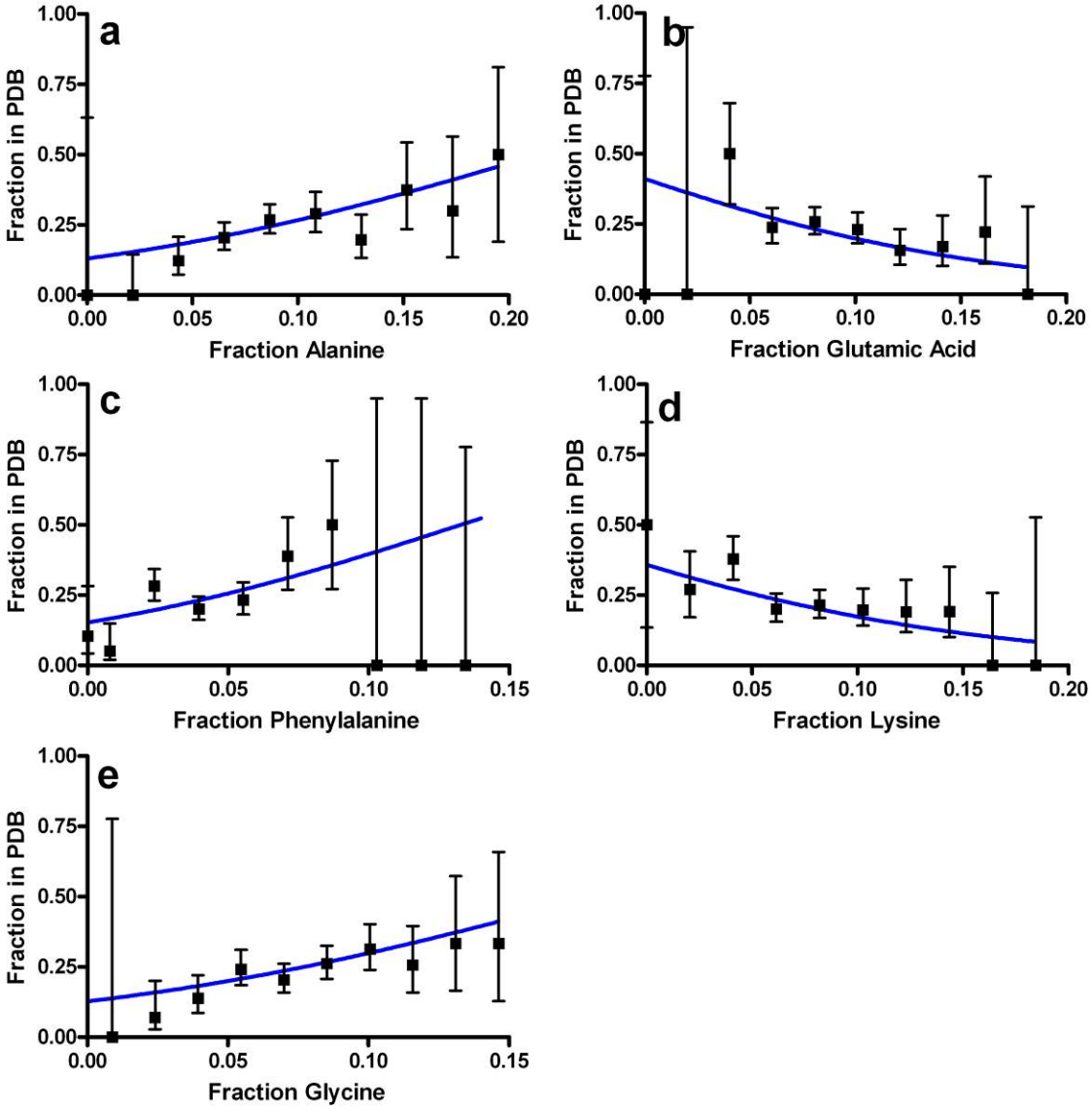
Supplementary Figure 6: ***Crystallization screen hit count is correlated with eventual structure solution***. The number of unvalidated crystal hits in the Hauptman-Woodward Institute's microbatch-under-oil 1536-condition screen was correlated with the probability of eventually depositing a crystal structure in the PDB. This analysis included 522 of the 679 biochemically tractable NESG pipeline proteins used for other datamining analyses. The boxes indicate the fraction of proteins for which crystal structures were successfully determined binned by the number of HWI hits. The whiskers (error-bars) indicate 95% confidence limits calculated from counting statistics using the numbers in each individual bin. A strong logistic relationship was observed (N = 522, P = 8.4x10^{-19}). Of the 10 crystal structures solved for proteins that failed to give any crystal hits in this screen, initial lead crystals were identified for eight through vapor-diffusion screening in the absence of added ligand while the final two only gave crystals leads after screening in the presence of physiological ligands. We hypothesize that proteins with multiple potential contact sites are more likely to form crystals in many conditions and are also inherently more likely to form high quality crystals suitable for structure determination. A larger number of hits does imply a higher probability of obtaining a crystal suitable for structure determination even if all hits have an equal probability of achieving the requisite quality. Further analysis will be necessary to understand the mechanistic details underlying the observed relationship.

Supplemental Information (continued)

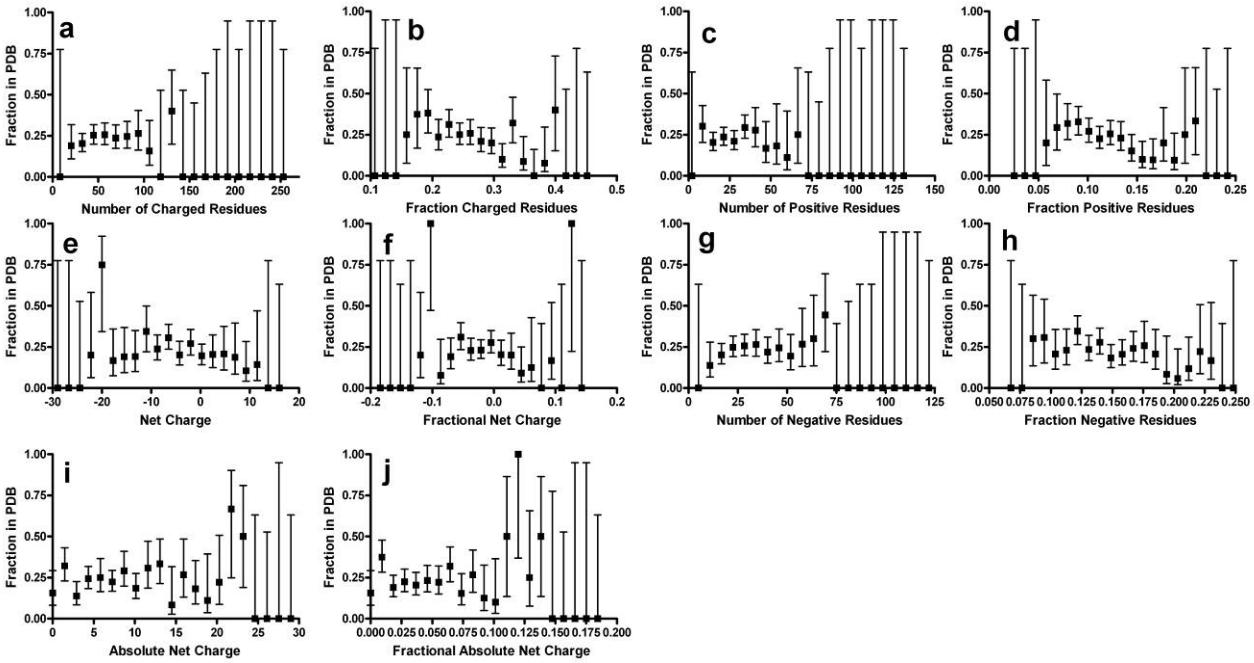


Supplementary Figure 7: ***Sequence parameter distributions in proteins giving unsolved crystals.*** Histograms of parameter distributions are shown for the 679 NESG pipeline proteins in the training set which produced crystal structures (InPDB, solid purple) and which failed to produce crystal structures (OutPDB, dotted purple) as well as for the subset of 39 of those proteins that produced diffracting crystals of insufficient quality for successful structure determination (Strong and Weak, solid and dotted gray, respectively). Strong indicates that diffraction was observed at 4 Å or better at the synchrotron, while weak indicates that diffraction was observed but only at lower resolution. Strongly diffracting crystals that failed to yield structures generally have high mosaicity or anisotropy. Distributions are shown for: **(a)** fraction of glycine in predicted buried residues; **(b)** predicted exposed <SCE>; **(c)** GRAVY; and **(d)** fraction of predicted disordered residues. Two-tailed unpaired T-tests for differing means showed no difference between the strong and weak sets, so they were combined for comparison with the InPDB and OutPDB sets. The combined set of diffracting crystals of insufficient quality to support structure determination is significantly different from the InPDB set in its fraction of glycine in predicted buried residues (P = 6.1x10^{-15}), predicted exposed <SCE> (P = 0.00039), and GRAVY (P = 0.040). This set is also significantly different from the OutPDB set in its fraction of glycine in predicted buried residues (P = 2.9x10^{-10}). Note that the set of diffracting crystals that failed to yield structures has a lower frequency of glycine in predicted buried residues than either the InPDB or OutPDB sets. These results support the premise that proteins yielding poor or pathological crystals are more similar to those that do not yield

diffracting crystals than to those that provide high quality crystals suitable for structure determination.

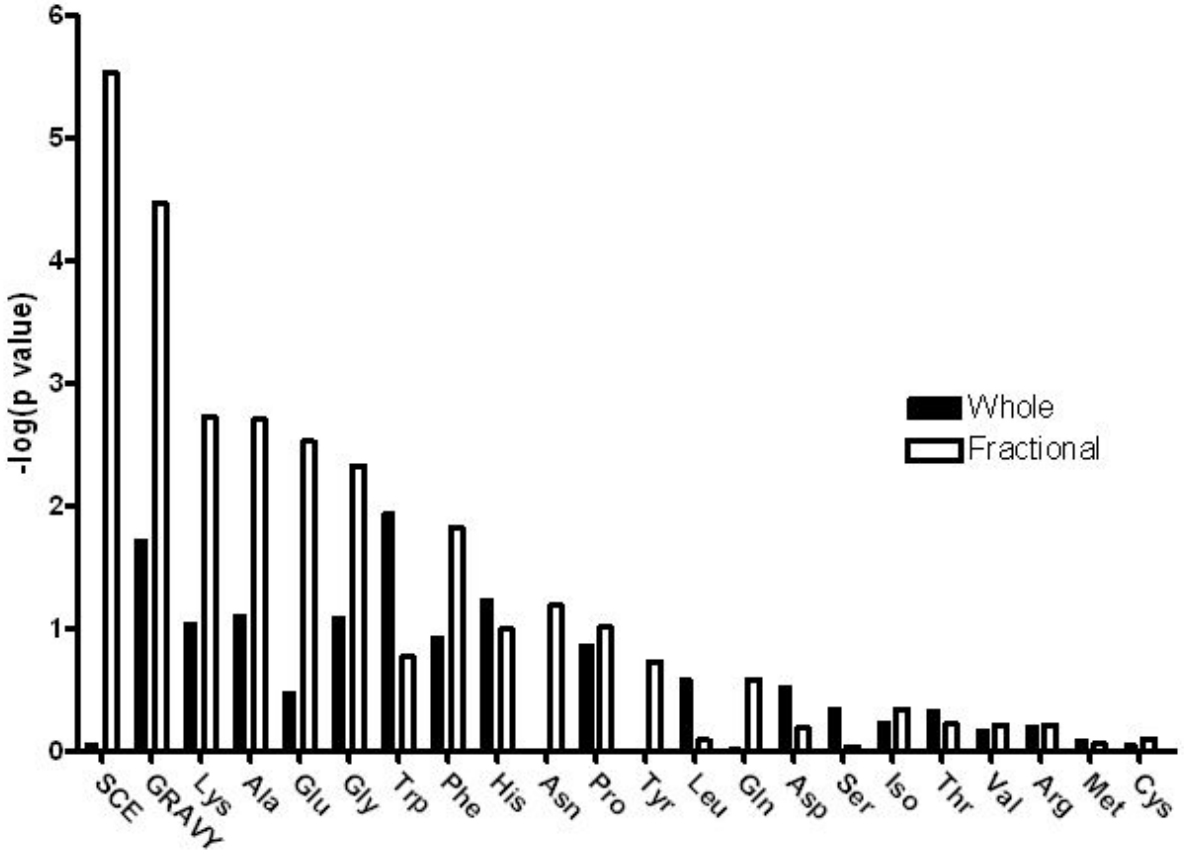


Supplementary Figure 8: ***Effects of significantly predictive individual amino acids on success in crystal structure determination***. Box-and-whisker plots show data for the 679 NESG pipeline proteins comprising the training set for the analyses presented in the paper. The boxes indicate the fraction of proteins for which crystal structures were successfully determined binned by fractional content of each amino acid. The whiskers (error-bars) indicate 95% confidence limits calculated from counting statistics using the numbers in each individual bin. The blue lines trace the logistic relationships for **(a)** ala (P = 0.0019), **(b)** glu (P = 0.0029), **(c)** phe (P = 0.015), **(d)** lys (P = 0.0018), and **(e)** gly (P = 0.0046).

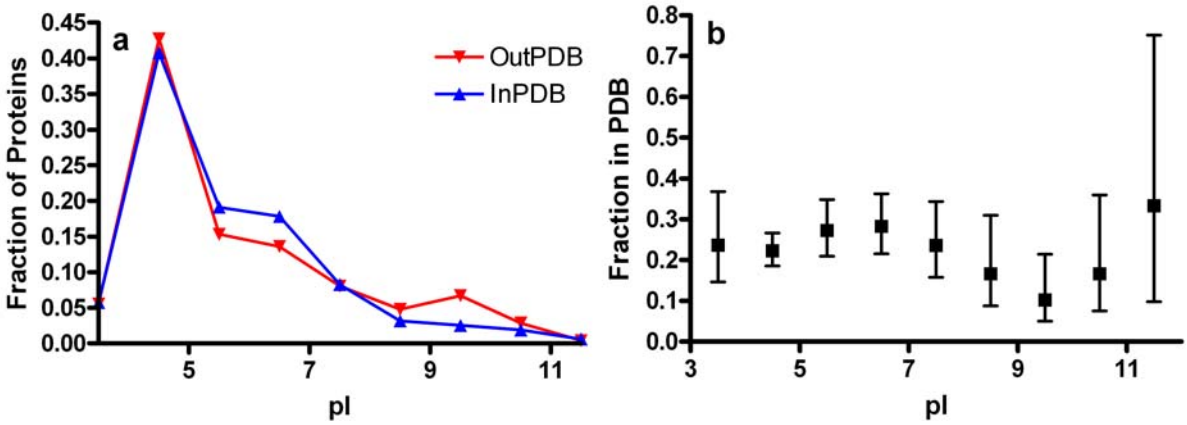


Supplementary Figure 9: ***Relationship between protein electrostatic charge variables and success in crystal structure solution***. Box-and-whisker plots show the fraction of proteins for with crystal structures were deposited in the PDB and associated 95% confidence limits (calculated from counting statistics using the numbers in each bin) for the 679 NESG pipeline proteins comprising the training set. Binning is based on the whole (not length-normalized) or fractional (length-normalized) electrostatic charge variable indicated on the abscissa of each plot. Logistic regression slopes, predictive values, and *P*-values from these analyses are presented in Supplementary Table 2 above. **(a)** Predicted number of electrostatically charged residues in each protein at neutral pH, calculated as the sum of the number of arg, lys, asp, and glu residues. **(b)** Predicted fractional content of electrostatically charged residues (*i.e.*, the data from panel **a** normalized by protein chain-length). **(c)** Predicted net electrostatic charge of each protein at neutral pH, calculated as the number of arg and lys residues minus the number of asp and glu residues. **(d)** Predicted net electrostatic charge from panel **c** normalized by protein chain-length. **(e)** Absolute value of the predicted net electrostatic charge of the protein at neutral pH (*i.e.*, the absolute values of the numbers used to generate panel **c)**. **(f)** Absolute value of the predicted net electrostatic charge from panel **e** normalized by protein chain-length. **(g)** Number of positively charged residues (arg plus lys). **(h)** Number of positively charged residues from panel **G** normalized by protein chain-length. **(i)** Number of negatively charged residues (asp plus glu). **(j)** Number of negatively charged residues from panel **i** normalized by protein chain-length.

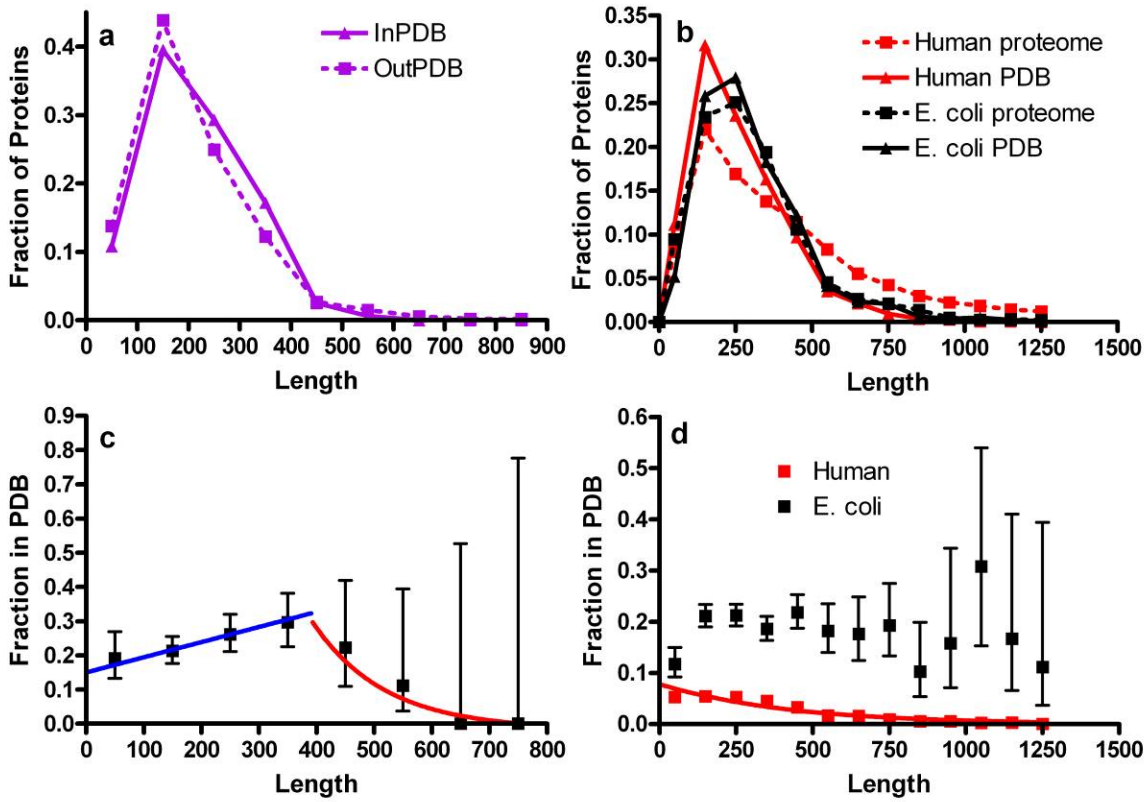
Supplemental Information (continued)



Supplementary Figure 10: ***Comparison of whole and fractional predictors***. Negative log *P*-values are shown from logistic regressions of whole (unnormalized) and fractional (length-normalized) parameters correlated against crystal structure solution for 679 NESG pipeline proteins. Parameters are shown in descending order of the statistical significance of the fractional predictor. Fractional values are more significant than whole values for every significant predictor.

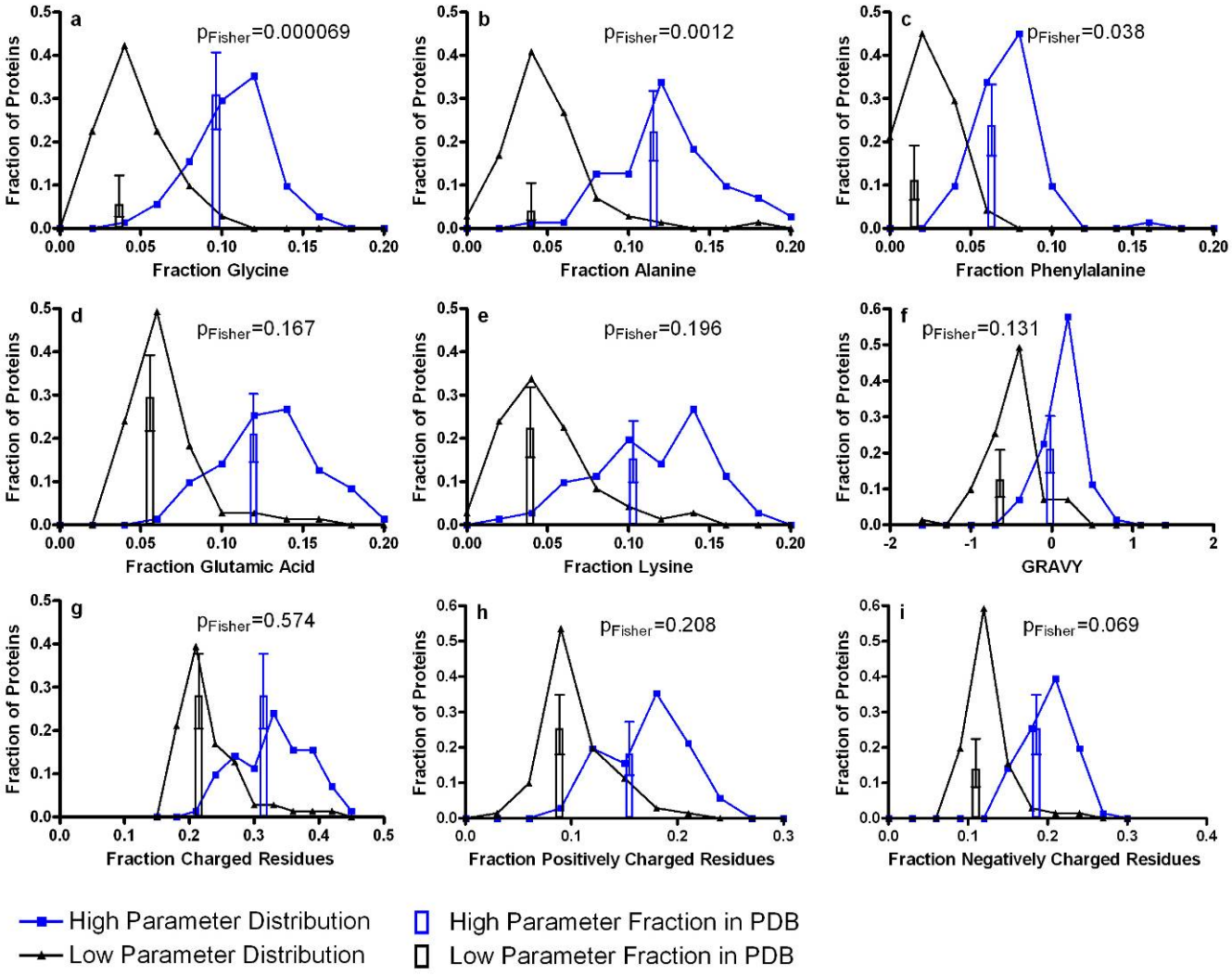


Supplementary Figure 11: ***Relationship between protein isoelectric point (pI) and crystal structure solution***. Data is presented for the 679 NESG pipeline proteins comprising the training set. **(a)** Histograms showing pI distributions for proteins that yielded a crystal structure (In PDB) or that did not (Out PDB). **(b)** A box-and-whiskers plot binned by pI, showing the fraction in PDB and 95% confidence limits calculated from counting statistics using the numbers in each bin.



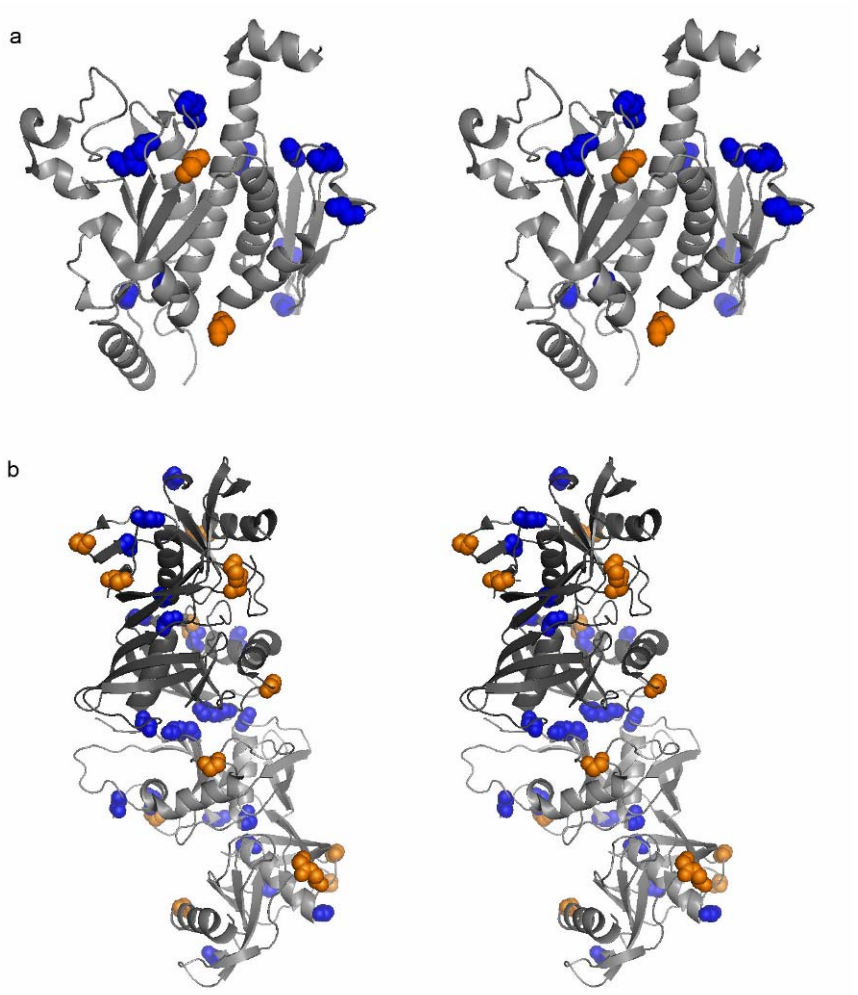
Supplementary Figure 12: ***Relationship between protein chain-length and success in crystal structure solution***. Data is shown in panels **a** and **c** for the 679 NESG pipeline proteins comprising our training set and in panels **b** and **d** for the predicted soluble proteomes from *H. sapiens* and *E. coli* compared to all PDB deposition of proteins from these species (culled at 30% sequence identity). **(a,b)** Histograms showing chain-length distributions for proteins that yielded a crystal structure (In PDB) or that did not (Out PDB). **(c,d)** Box-and-whiskers plots binned by chain-length showing the fraction in PDB and 95% confidence limits calculated from counting statistics using the numbers in each bin. The relationship appears bimodal for the NESG training set, with an inflection point at a protein chain-length of approximately 400 amino acids. The blue line in panel **c** traces the logistic regression result for NESG proteins under 400 amino acids in length ($p = 0.024$), while the red line traces that for proteins over 400 amino acids in length ($P = 0.11$). Length has no significant predictive effect for the *E. coli* proteome (black in panel **d)** when all proteins are considered at once, while increasing length shows a highly significant correlation with decreasing success in crystal structure determination for the human proteome ($P = 7.8 \times 10^{-52}$ – red in panel **d)**. Notably, the rate of PDB depositions from the human proteome appears uniform up to approximately 350 amino acids and declines steeply from 400-700 amino acids, matching the qualitative trend observed in the NESG training set.

Supplemental Information (continued)



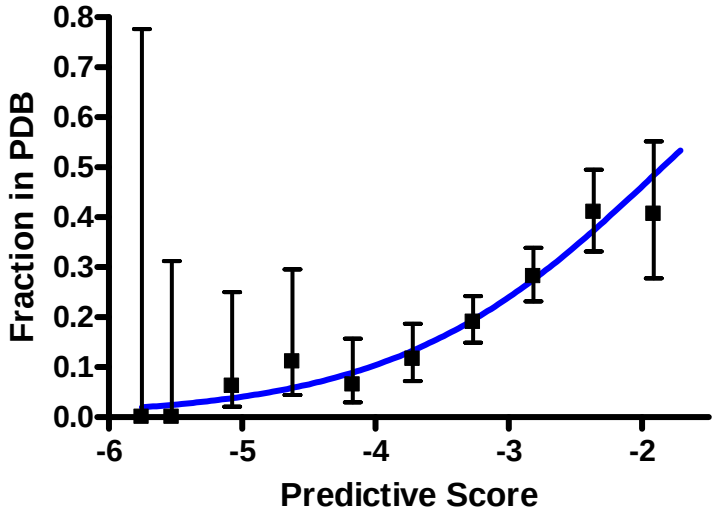
Supplementary Figure 13: ***Evaluation of the residual influence of individual amino acid frequencies on crystallization propensity in sets of proteins with matched predicted exposed <SCE> distributions.*** To examine whether specific amino acids have an influence on success in crystal structure determination beyond that associated with their SCE, protein sub-sets were generated with matched exposed <SCE> distributions but differing in content of specific amino acids. The 679 NESG proteins in the training set were separated into 100 equally spaced bins based on predicted exposed <SCE>, producing 71 bins containing at least two proteins. Each of these bins was sorted based on the fractional content of the amino acid to evaluated (*i.e.*, separately for each amino acid frequency), and the two proteins with the highest and lowest parameter values were assigned to the High and Low sets for that parameter. This procedure created two sets systematically differing in the frequency of the amino acid being evaluated while having nearly identical distributions of overall predicted exposed <SCE>. Each graph shows the resulting distributions of the amino acid frequency in the pair of sets created to evaluate whether it has any residual influence on crystallization propensity. Superimposed on each distribution is a bar showing the fraction of proteins in that set yielding a PDB deposition along with error bars representing 95% confidence limits calculated from counting statistics using the numbers in each bin. The Fisher's Exact Test was used to evaluate the statistical

significance of the observed difference in structure determination rate (*i.e.*, the *P*-value for the observed difference in structure determination rate to occur at random in populations of that size). Each analysis and graph includes two proteins in each of 71 bins (N = 142). The amino acids whose frequencies are positively correlated with success in structure determination in the entire training set (**(a)** gly**, (b)** ala**,** and **(c)** phe) remain statistically significant predictors of success even in sets of proteins with equivalent distributions of exposed <SCE>, indicating that they have more favorable properties in mediating high quality lattice packing contacts than accounted for by their low sidechain entropy. No other parameters are significantly predictive of success in protein sets with matched distributions of exposed <SCE> (*i.e.*, **(c)** glu, **(e)** lys, **(f)** GRAVY **(f)**, **(g)** fraction of charged residues (arg+lys+asp+glu), **(h)** fraction of positively charged residues (arg+lys), or **(i)** fraction of negatively charged residues (asp+glu)). Therefore, the influence of these parameters on crystallization propensity is primarily attributable to sidechain entropy and probably not any independent effect.



Supplementary Figure 14: ***Locations in representative crystal structures of glycines predicted by PHD/PROF to be in loops 6-15 residues in length.*** Gly residues predicted by PHD/PROF to be in loops of this length are displayed in space-filling representation in stereo ribbon diagrams generated by PYMOL[44]. PHD/PROF-predicted exposed and buried residues are colored orange and blue, respectively. The majority of the predicted buried residues are at least partially solvent-exposed in the protomer structures. **(a)** Chorismate synthase from *C. jejuni* (PDB id 1SQ1) has 23 glycines predicted to be in loops of the relevant size. Among the eight of these predicted to be exposed, five were not sufficiently well ordered to be modeled in the crystal structure, while the remainder were partially solvent-exposed. Among the 15 predicted to be buried, one was not sufficiently well ordered to be modeled, 11 were partially solvent-exposed, and only three appeared to be buried. **(b)** A dimer of the putative proline racemase from *B. melitensis* (PDB id 1TM0) is shown, which has 19 gly residues in each protomer predicted to be in loops of the relevant size. Among the eight of these predicted to be exposed, one was not sufficiently well ordered to be modeled in the crystal structure and seven were partially solvent-exposed. Among the 11 predicted to be buried, seven were partially solvent-exposed, two appeared to be buried in the protomer, and two were buried in the homodimer interface (shown at the center of the stereopair). These last two, which directly mediate the intersubunit packing interaction, would be solvent-exposed prior to dimer formation.

Supplemental Information (continued)

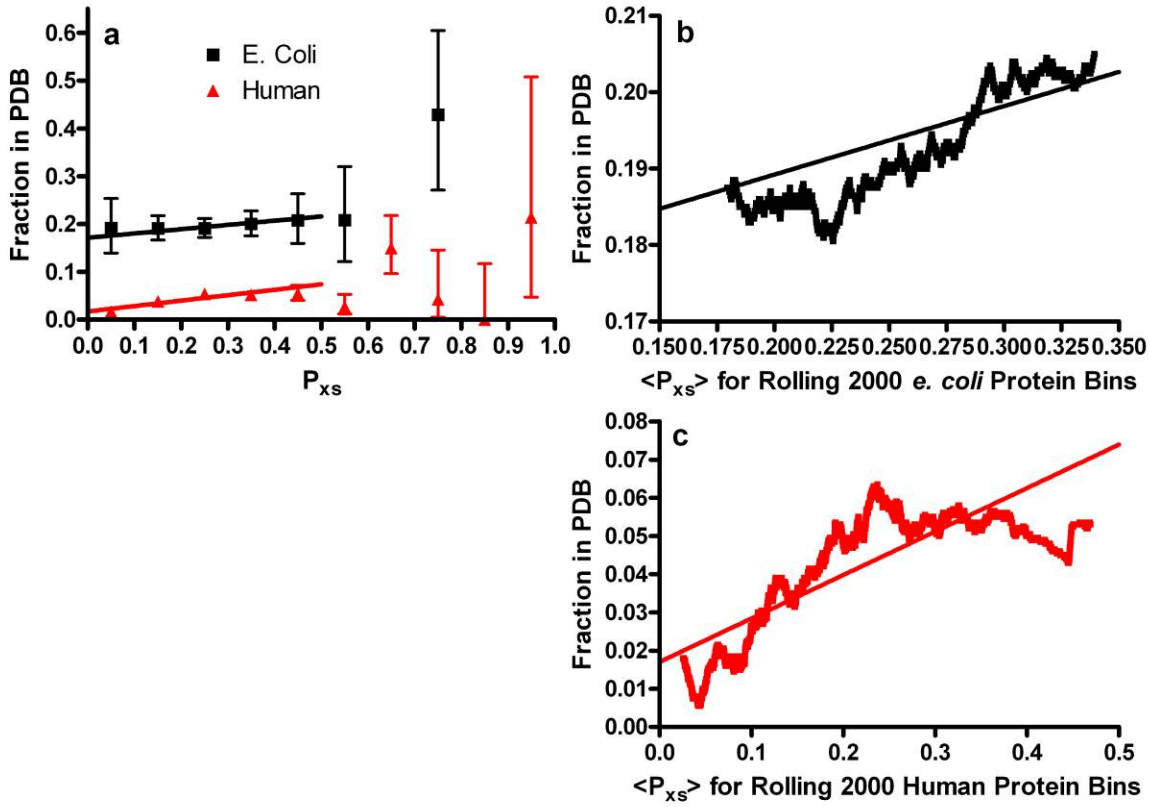


$$P_{XS} = e^{\Theta}/(1+e^{\Theta})$$

$$\Theta = -(1.85 - 3.77*\langle SCE\rangle_{exp} + 0.08*\%Gly_{bur} - 0.032*\%Diso + 0.14*\%Phe)$$

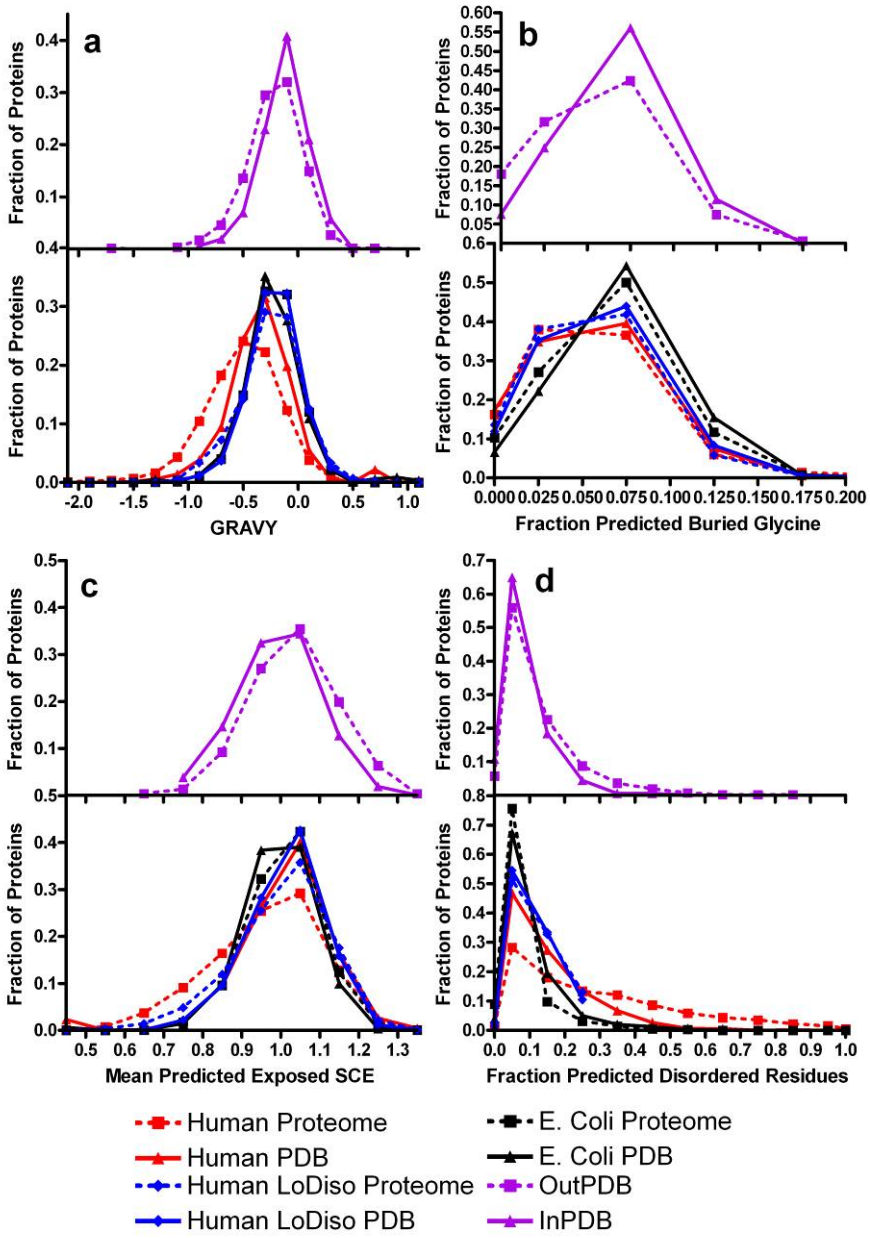
Supplementary Figure 15: ***Functional form of the P_{XS} predictive metric***. **(a)** The box-and-whiskers plot binned by P_{xs} value shows the fraction of the 679 NESG pipeline proteins in the training set yielding successful crystal structure determinations and 95% confidence limits calculated from counting statistics using the numbers in each bin. The blue line traces the result of the multiple logistic regression (N = 679, P = 0.0000000053). **(b)** The functional form of $\mathbf{P_{XS}}$ is shown. The contributing variables are mean side chain entropy of PHD/PROF-predicted exposed residues ($\langle SCE\rangle_{exp}$), percent glycine among predicted buried residues ($\%Gly_{bur}$), percent residues predicted to be disordered by DISOPRED2 (%Diso), and percent phenylalanine in the complete sequence (%F). A webserver performing this calculation on user-specified protein sequences is available on the internet for public use at http://www.nesg.org/PXS/.

Supplemental Information (continued)



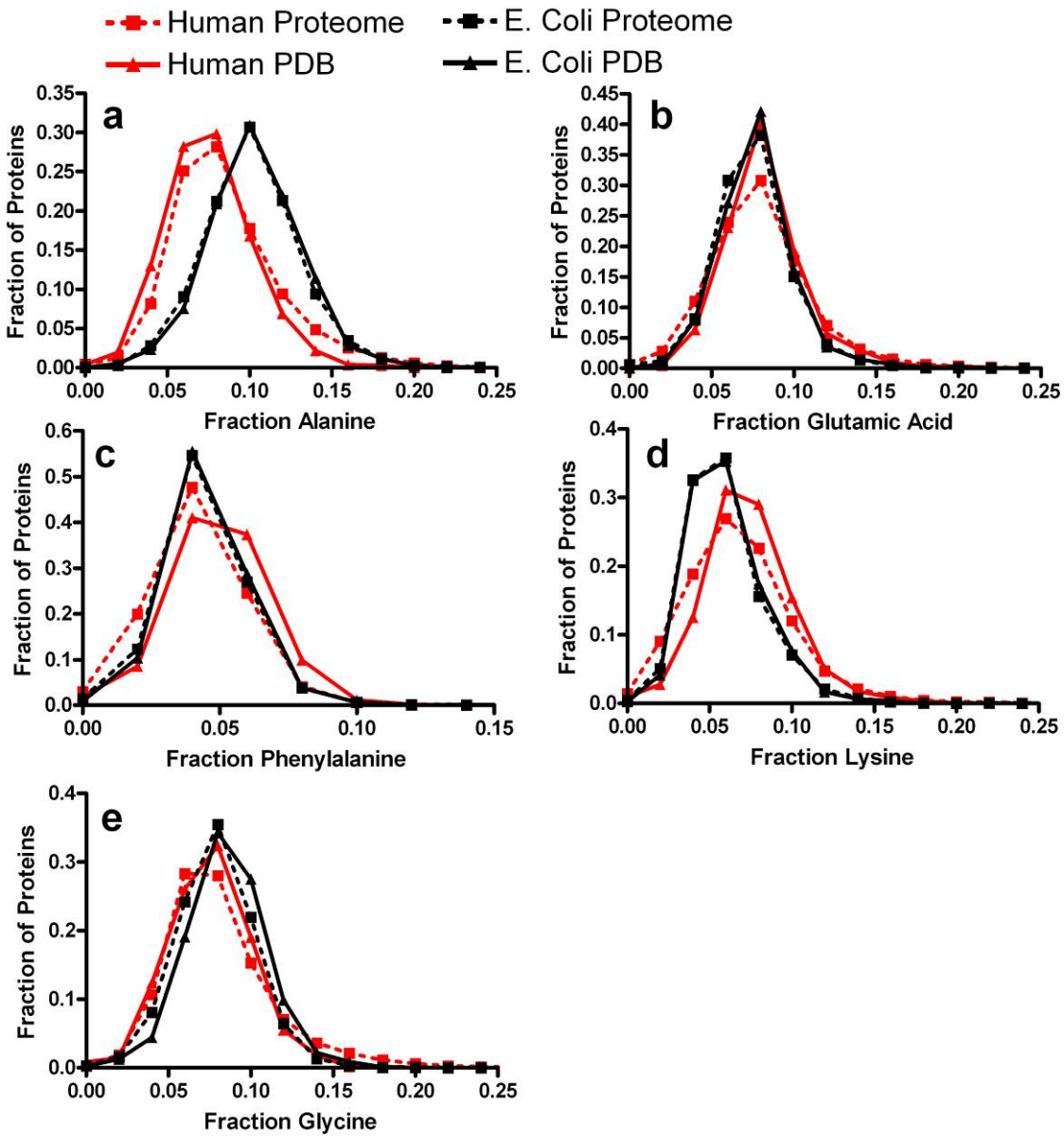
Supplementary Figure 16: ***Analyses of whole proteomes from E. coli and H. sapiens using P_{XS}.*** **(a)** Box-and-whisker plot binned by P_{xs} value showing the fraction of proteins from the predicted soluble proteomes of *E. coli* or *H. sapiens* with structures deposited in the PDB and 95% confidence limits calculated from counting statistics using the numbers in each bin. Fractions are calculated as the number of structures deposited in the PDB from each species in each bin divided by the number of genomic sequences in the same bin. Logistic regression lines are shown for *E. coli* in black (N = 3,962, P = 0.07) and humans in red (N = 27,652, $P = 6.7 \times 10^{-36}$). **(b)** The fraction of proteins deposited in the PDB is graphed as a function of the rolling average of 2000-protein bins from the *E. coli* (panel **b)** and human (panel **c)** predicted soluble proteomes. The same logistic regression lines are shown as in panel **a**.

Supplemental Information (continued)

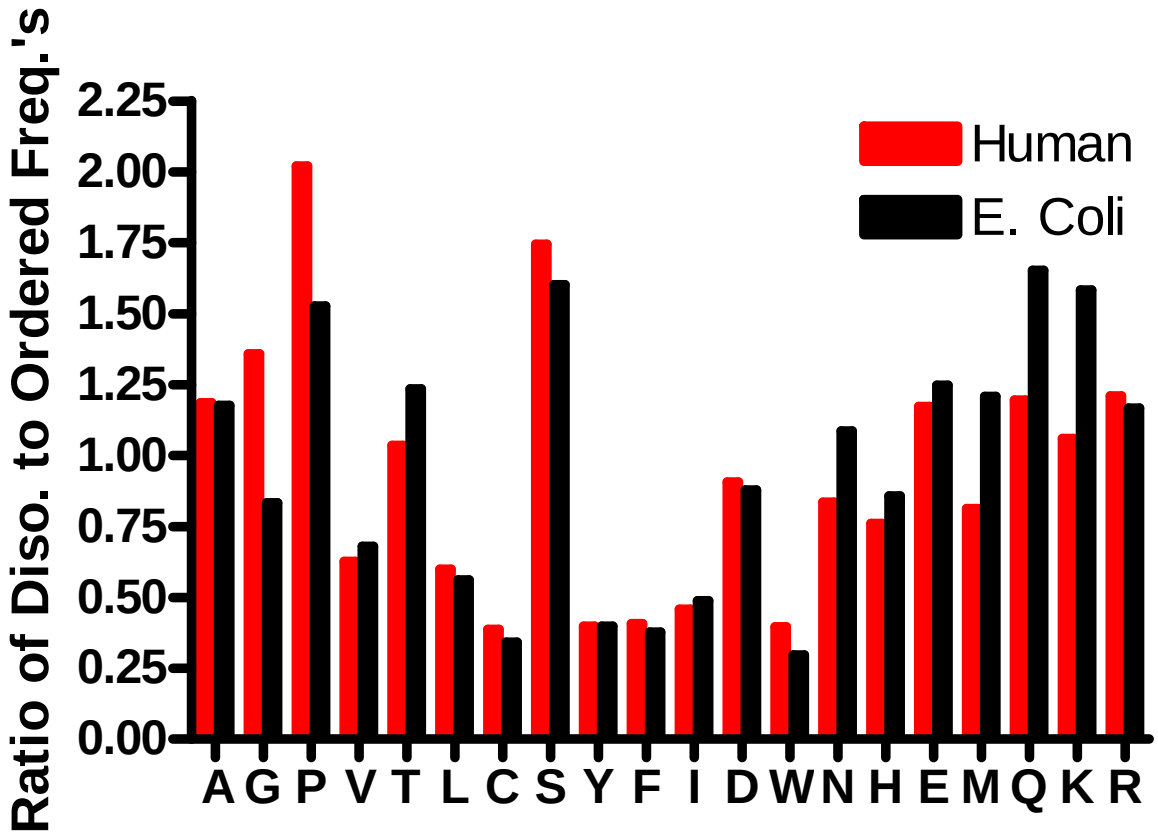


Supplementary Figure 17: ***Distributions of bulk sequence parameters in predicted soluble proteomes and corresponding sequences deposited in the PDB.*** Histograms of parameter distributions are shown for the 679 NESG pipeline proteins in the training set segregated by whether or not a crystal structure was successfully determined (In PDB represented by solid purple line *vs.* Out PDB represented by dotted purple line), for the complete *E. coli* (dotted black line) and *H. sapiens* (dotted red line) predicted soluble proteomes, and for the sets of proteins from these organisms whose crystal structures have been deposited in the PDB (culled at 30% sequence identity) (with the *E. coli* PDB represented by a solid black line and the *H. sapiens* PDB represented by a solid red line). The distributions labeled "LoDiso" are for the two *H. sapiens* datasets after removing all proteins with sequences predicted to be more than 25% disordered by DISOPRED2 (with the whole soluble proteome represented by a dotted blue line

and those deposited in the PDB by a solid blue line). **(a)** GRAVY (mean residue hydrophobicity). **(b)** Fractional gly content in PHD/PROF-predicted buried residues. **(c)** Mean sidechain entropy of PHD/PROF-predicted exposed residues. **(d)** Fraction of residues predicted to be disordered by DISOPRED2.

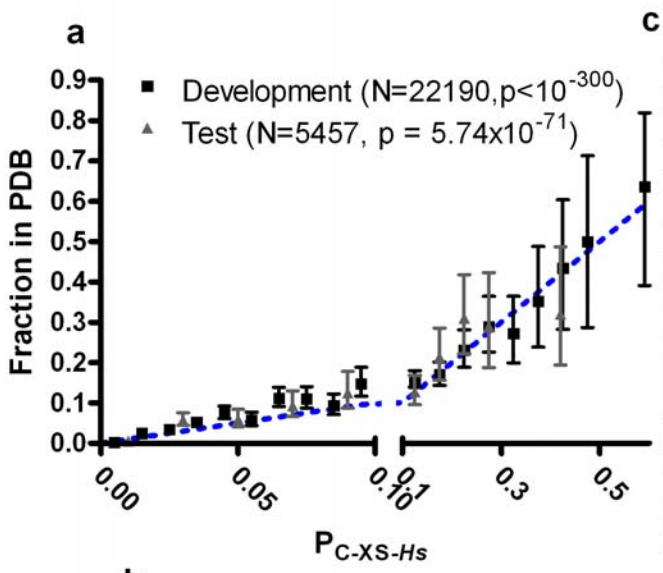


Supplementary Figure 18: ***Distributions of individual amino acid content in predicted soluble proteomes and corresponding sequences deposited in the PDB***. Histograms of amino acid content in the complete *E. coli* (dotted black line) and *H. sapiens* (dotted red line) predicted soluble proteomes, and for the sets of proteins from these organisms whose crystal structures have been deposited in the PDB (culled at 30% sequence identity) (with the *E. coli* PDB represented by a solid black line and the *H. sapiens* PDB represented by a solid red line). **(a)** Alanine. **(b)** Glutamic acid. **(c)** Phenylalanine. **(d)** Lysine. **(e)** Glycine.



Supplementary Figure 19: ***Enrichment of individual amino acids in disordered sequences from H. sapiens vs. E. coli***. The graph shows the ratio between the frequency of each amino acid in disordered *vs.* ordered sequences in the predicted soluble proteomes of *E. coli* (black bars) or *H. sapiens* (red bars). Disordered sequences are defined as continuous segments at least 10 residues in length predicted to be disordered by DISOPRED2, while ordered sequences are defined as all other segments. Gly is highly enriched in disordered sequences from human but not *E. coli* proteins, while pro is significantly enriched in both but more strongly in human proteins, and ser is highly enriched in both.

Supplemental Information (continued)



c

Variable	Location	Type	Slope	SD	SD*Slope	p
Constant			-2.3			8.00E-19
Tyr	Buried	Frac.	29.3	0.0117	0.343	7.97E-21
Val	Buried	Frac.	18.1	0.0201	0.364	5.62E-18
DISO	Total	Frac.	2.10	0.242	0.510	3.84E-11
Iso	Buried	Frac.	12.9	0.0185	0.239	1.43E-08
Phe	Buried	Whole	0.0436	14.1	0.615	4.65E-06
Leu	Buried	Whole	0.0173	38.7	0.671	3.31E-04
Asp	Exposed	Whole	0.0385	22.1	0.849	4.32E-04
Val	Exposed	Whole	0.0462	8.89	0.410	3.73E-03
His	Exposed	Frac.	10.2	0.0125	0.128	5.78E-03
Asp	Buried	Whole	0.0713	3.39	0.242	9.88E-03
His	Exposed	Whole	0.0333	8.00	0.267	1.78E-02
Tyr	Exposed	Frac.	8.34	0.00913	0.0761	1.83E-02
Arg	Buried	Frac.	-17.0	0.00551	-0.0937	4.66E-02
Trp	Exposed	Frac.	-18.2	0.00556	-0.101	3.00E-02
|Net Charge|	Total	Frac.	-6.58	0.0423	-0.279	3.18E-03
Lys	Exposed	Whole	-0.0208	31.8	-0.662	2.45E-03
Net Charge	Total	Frac.	-4.07	0.0563	-0.229	1.00E-03
|Net Charge|	Total	Whole	-0.0385	16.8	-0.645	7.52E-04
Pro	Buried	Whole	-0.0750	9.75	-0.731	6.36E-07
Lys	Buried	Whole	-0.253	1.73	-0.438	1.99E-07
Arg	Exposed	Whole	-0.0545	24.9	-1.36	2.98E-10
Cys	Exposed	Whole	-0.228	3.27	-0.744	1.70E-11
Ser	Exposed	Whole	-0.0593	35.0	-2.08	2.19E-15
Pro	Exposed	Frac.	-18.1	0.0412	-0.745	1.73E-21
Met	Exposed	Whole	-0.418	3.34	-1.40	4.69E-30

Supplementary Figure 20: ***Human conflated predictive metric***. The predicted soluble human proteome and redundancy-culled human PDB were randomly divided into two sets in a 4:1 ratio for training and testing the metric. Individual logistic regressions against In-PDB status were run on 92 sequence characteristics (including <SCE> and GRAVY, which are not included in the final model). Factors that correlated with success in single-parameter regressions at the Bonferroni-corrected significance level of 0.00054 (*i.e.*, 0.05/92) were combined in order of significance by forward stepwise regression, with a $p<0.05$ threshold for inclusion in the final multiple regression. The resulting metric conflates the experimental stages of expression, purification, crystallization, and structure solution. We call it $\mathbf{P_{C\text{-}XS\text{-}}}_{\boldsymbol{Hs}}$ because it predicts the conflated probability of determining a crystal structure for a sequence from the *Homo sapiens* genome. While not informative in terms of crystallization mechanism, this metric is design to aid selection of human targets for crystal structure determination and may aid in understanding competing trends across the crystallization process. The publicly available webserver that performs the standard $\mathbf{P_{XS}}$ calculation also performs this calculation (http://www.nesg.org/PXS/). **(a)** The fraction in PDB for bins of $\mathbf{P_{C\text{-}XS\text{-}}}_{\boldsymbol{Hs}}$ in the training and test sets are shown in black and gray, respectively, with 95% confidence limits calculated from counting statistics using the numbers in each bin. The multiple logistic regression model, shown by the dashed blue line, matches the training set (N = 22,190, $p<10^{-300}$) with an insignificant Hosmer-Lemeshow lack of fit[22] ($P = 0.411$), indicating good calibration, and an area under the ROC curve of 0.882, showing good discrimination. The $\mathbf{P_{C\text{-}XS\text{-}}}_{\boldsymbol{Hs}}$ metric predicts the test set nearly as well (N = 5,457, $P = 5.74\text{x}10^{-71}$, insignificant Hosmer-Lemeshow lack of fit ($P = 0.319$), and high ROC area of 0.871). **(b)** The functional form of the model. **(c)** Table showing the variables and coefficients used in the final model.