# Systematic enhancement of protein crystallization efficiency by bulk lysine-to-arginine (KR) substitution

Nooriel E. Banayan [1†], Blaine J. Loughlin [1†], Shikha Singh [1],
Farhad Forouhar [1], Guanqi Lu [1], Kam-Ho Wong [1§], Matthew Neky [1],
Henry S. Hunt [2], Larry B. Bateman Jr. [3], Angel Tamez [3],
Samuel K. Handelman [1§], W. Nicholson Price [1§], & John F. Hunt [1*]

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

[2] Physics Department, Stanford University, Stanford, CA 94305, USA; and

[3] Accendero Software, P.O. Box 2826, Idaho Falls, ID 83404, USA.

[†] These authors contributed equally to the work reported in this paper.

[*] To whom correspondence may be addressed:
JFH – jfh21@columbia.edu, (646)-270-5012 *voice*, (212)-865-8246 *FAX*.

[§] Current addresses:
SKH, 307 E. Merrill St, Indianapolis, IN 46225; MN, Mailman School of Public Health, Columbia University, 722 W. 168th Street, New York, NY 10032; WNP, University of Michigan Law School, 625 S. State St, Ann Arbor, MI 48109; KHW, 401 N Middletown Rd, Pearl River, NY 10965.

## Bulk K-to-R substitution enhances crystallization

**Keywords: protein crystallization / protein thermodynamics / protein engineering / homology analysis / circular dichroism spectroscopy / protein solubility / x-ray crystallography.**

**Structural genomics consortia established that protein crystallization is the primary obstacle to structure determination using x-ray crystallography. We previously demonstrated that crystallization propensity is systematically related to primary sequence, and we subsequently performed computational analyses showing that arginine is the most overrepresented amino acid in crystal-packing interfaces in the Protein Data Bank. Given the similar physicochemical characteristics of arginine and lysine, we hypothesized that multiple lysine-to-arginine (KR) substitutions should improve crystallization. To test this hypothesis, we developed software that ranks lysine sites in a target protein based on the redundancy-corrected KR substitution frequency in homologs. We demonstrate that three unrelated single-domain proteins can tolerate 5-11 KR substitutions with at most minor destabilization and that these substitutions consistently enhance crystallization propensity. This approach rapidly produced a 1.9 Å crystal structure of a human protein domain refractory to crystallization with its native sequence. Structures from bulk-KR-substituted domains show the engineered arginine residues frequently make high-quality hydrogen-bonds across crystal-packing interfaces. We thus demonstrate that bulk KR substitution represents a rational and efficient method for probabilistic engineering of protein surface properties to improve protein crystallization.**

More than 50 years after the solution of the first protein crystal structure [1-3], protein crystallization remains a hit-or-miss proposition. Synergistic developments in crystallographic methods[4-9], synchrotron beamlines[10-13], and high-speed computing have made structure solution and refinement routine, even for massive complexes, as long as high-quality crystals are available. However, there has been comparatively little progress in improving methods for protein crystallization. Structural genomics consortia have systematically confirmed that most naturally occurring proteins do not readily yield high-quality crystals suitable for x-ray structure determination and that crystallization is the major obstacle to the determination of protein structures using diffraction methods [14-16]. While numerous methods have been developed that have some efficacy in improving protein crystallization properties [17-27], none work with sufficiently high efficiency to have been applied with significant frequency by practicing crystallographers. We therefore set out to develop efficient methods for rational engineering of protein surface properties to improve crystallization propensity. The first phase of our research identified a large number of local primary sequence patterns, which we called crystallization epitopes, that are strongly overrepresented in crystal-packing interfaces [28]. We demonstrated that introducing these epitopes individually into proteins generally increases their crystallization propensity and that introducing multiple such epitopes progressively increases crystallization propensity. The cumulative nature of the observed improvements suggested that multiple simultaneous mutations could potentially produce definitive improvements in crystallization propensity in a single protein construct based on large-scale probabilistic engineering of protein surface properties. We herein present an efficient method to achieve this goal while preserving protein stability, solubility, and function.

Our efforts to develop rational methods to improve protein crystallization properties are grounded in sequence and structural analyses of historical crystallization results and associated thermodynamics studies. Our published analyses of large-scale experimental studies showed that the surface properties of proteins, and particularly the entropy of the exposed sidechains, are a major determinant of protein crystallization propensity [16]. These studies demonstrated that overall thermodynamic stability is not a major determinant of protein crystallization propensity. They identified a number of primary sequence properties that correlate with successful crystal structure determination, including strong anticorrelations with predicted backbone disorder and surface sidechain entropy and weak positive correlations the fractional content of several individual amino acids [16]. In follow-up studies, we analyzed 87,683 crystal structures from the Protein Data Bank (PDB) and identified contiguous amino acid patterns strongly overrepresented in crystal packing interfaces [28]. This analysis also generated data on the relative overrepresentation of individual amino acids in crystal-packing interfaces segregated by secondary structure (**Fig. 1**), and these data suggested the streamlined approach reported in this paper that enhances protein crystallization propensity based on multiple simultaneous surface mutations.

Our computational analysis of crystal-packing interactions in the PDB showed a substantially higher probability for arginine to mediate inter-molecular packing contacts than lysine (**Fig. 1**), consistent with our expectations based on earlier analyses of correlations between primary sequence features and protein crystallization propensity [16]. The observation that arginine mediates crystal-packing contacts more frequently than lysine is particularly notable because the entropy of the arginine sidechain is generally estimated to be somewhat higher than that of lysine [29-32], and higher surface sidechain entropy opposes crystallization [16, 17, 22-24]. Therefore, the more frequent occurrence of arginine compared to lysine in crystal-packing contacts suggests that the guanidino group on arginine is substantially “stickier”, in terms of intermolecular interaction free energy, than the primary amine on lysine. This inference is consistent with the well-known properties of the guanidinium ion as a protein denaturant [33-35], a property that is not shared by primary amines. Therefore, we hypothesized that introducing multiple arginine-to-lysine (KR) substitutions in a protein would enhance crystallization propensity and that, given the very similar physicochemical properties of arginine and lysine in terms of size and polarity, multiple simultaneous substitutions would be tolerated without significantly impairing thermodynamic stability.

We herein report the results of biophysical studies that support the validity of this hypothesis. We developed a computer program that automates selection of sites for KR mutagenesis based on the frequency of such substitutions in naturally occurring homologs, which should avoid sites where lysine is critical for function or structural stability. We furthermore characterized the effects of introducing multiple simultaneous KR mutations on the thermodynamic stability, solubility and crystallization propensity of three unrelated test proteins, one of which crystallizes readily and two of which are recalcitrant to crystallization with their native sequences. These studies demonstrate that introducing multiple KR mutations into a protein, which we call Bulk KR substitution, is a simple and effective method to improve crystallization propensity. Physicochemical analyses have thus guided the development of an efficient method for large-scale probabilistic engineering of protein surface properties to improve crystallization, which was historically considered a stochastic phenomenon refractory to rational experimental manipulation.

# Results

***KR mutation site-selection algorithm and software.*** Sites for Bulk KR substitution are ranked based on the frequency of these substitutions in naturally evolved sequences in a phylogenetic alignment. This procedure is fully automated in Python code available at Github (https://github.com/huntmolecularbiophysicslab/pxengineering) that can be run interactively via a webserver that can be accessed at http://www.pxengineering.org. The algorithm implemented by the program ranks sites based on a redundancy-compensated estimate (explained below) of the frequency of KR substitutions in homologous sequences, which are divided into mutually exclusive bins with progressively lower levels of overall percent identity relative to the target sequence. The first bin includes sequences with greater than or equal to 90% identity and less than 99% identity (to avoid mutant variants of the target sequence), and subsequent mutually exclusive bins reduce the lower identity level in 10% steps down to a minimum of 30%. The algorithm steps through these bins in order from highest to lowest percent identity, selecting sites in each bin in inverse order of their redundancy-compensated count of lysine-to-arginine substitutions down to a minimum user-adjustable threshold count. This threshold is imposed to avoid selecting a site based on an arginine substitution in a single sequence that could potentially be inaccurate or in a small number of very closely related sequences that could potentially share a function-impairing or stability-impairing mutation; it defaults to a value of 1.1, which ensures observation of a lysine-to-arginine substitution in at least two sequences with no more than ~93% identity to one another.

The software provides graphical displays of summary parameters characterizing the amino acid distribution in the homologs in each of the percent-identity bins at every lysine site in the target sequence (**Fig. 2**), as well as a graphical display of the overall sequence diversity in each of the bins (**Fig. ED1**). The displayed summary parameters are the Shannon entropy of the amino-acid frequency distribution, the frequency of all residues other than lysine, the arginine-to-lysine ratio, the total count of sequences with an arginine residue at the site, and two different estimates of that count after compensation for redundancy between those sequences. Both redundancy-compensation calculations use the same heuristic estimate of the degree of mutational resampling between pairs of sequences, which is described in the **Methods** section along with explanations of the details of the two algorithms. In brief, the first redundancy-reduced count evaluates all sequences using a calculation that has rigorously correct behavior in the cases of full redundancy and full independence between the sequences but is otherwise approximate. The second count provides a rigorous probabilistic estimate of the number of independent observations in the seven most remotely related sequence pairs in the bin having arginine at that site. Extending this calculation to more sequences is computationally prohibitive, but the estimate based on a limited set of the most diverged homologs provides a highly effective method to ensure that multiple independently determined protein sequences have an arginine residue at the lysine site in the target protein, which is the essential goal of the redundancy-compensation calculations. This second calculation is used for the automated site-ranking algorithm described above.

The program additionally provides a ranking of sites for introducing aspartate-to-glutamate and asparagine-to-glutamine mutations together with a record of which of those sites have potential ionic interaction or H-bonding partners in the target sequence that would tend to reduce the entropy of the longer sidechains in beta-sheet (i±2) or α-helical (i±3,i±4) secondary structures [36-38]. (The rationale behind this approach is described in the ***Discussion*** section below.) Lysine, arginine, and histidine are considered potential ionic interaction partners for glutamate and H-bonding partners for glutamine, while asparagine, glutamine, serine, and threonine are considered potential H-bonding partners for both with the addition of aspartate and glutamate for glutamine.

***Test protein selection and expression.*** We chose to test the Bulk KR substitution approach using three proteins with different crystallization properties. The hPDIa domain is a human drug target [39, 40] that represents the first of four domains in the endoplasmic-reticulum-resident human Protein Disulfide Isomerase (hPDI) protein. The hPDIa domain had never successfully been crystallized on its own, but its structure was known from a relatively low-resolution crystal structure of a much longer multi-domain construct containing hPDIa [41], which enables evaluation of the impact of Bulk KR substitutions on its structure, as reported below. *E. coli* RNaseH is difficult to crystallize in the absence of ligands stabilizing active site structure but has had its crystal structure determined by groups studying its enzymological mechanism and folding [42-47]. MA_2137, an S-adenosyl-methione-dependent RNA methyltransferase from *Methanosarcina acetivorans*, crystallizes well in the presence of S-adenosyl-homocysteine (SAH), the product of the methyltransferase reaction that it catalyzes. We included this last protein because we previously demonstrated that increasing hit count in high-throughput crystallization screening is strongly correlated with the probability of successful crystal-structure determination [16], which implies quantification of hit count for a protein that crystallizes relatively easily is an effective assay for crystallization propensity. KR mutations were introduced into the D65R mutant of MA_2137 because we had previously demonstrated that this single mutation improves crystallization of this protein, and we wanted to determine whether bulk KR substitutions could improve it even further.

We introduced 2 to 13 KR mutations into these proteins (**Table 1**), and we first examined the expression and solubility levels of the full set of mutant constructs when expressed from a pET plasmid using T7 RNA polymerase in *E. coli*, which yields high-level expression of the three parental proteins in the form of efficiently purified monodisperse monomers. The largest number of KR mutations tested preserved high-yield protein production in a monodisperse state for both hPDIa and MA_2137-D65R (*i.e.*, the hPDIa-9KR and MA_2137-D65R-11KR constructs). The RNaseH-2KR and RNaseH-5KR constructs similarly preserved high-yield protein production in a monodisperse state. However, the RNaseH-7KR construct yielded polydisperse protein that co-purified with the Hsp33 molecular chaperone protein [48, 49], while the RNaseH-11KR was completely insoluble even though it expressed at a high level (data not shown). The stability studies presented in the next section confirm earlier research [50, 51] showing that RNaseH has a low thermal melting temperature ($T_m$) of ~45 ˚C, making it marginally stable, which likely explains its tolerance for fewer KR mutations than the other target proteins.

***KR mutations are generally only minimally destabilizing.*** The thermal stabilities of all the successfully purified bulk KR construct were characterized using circular dichroism spectroscopy. These assays show a variable but generally very small degree of destabilization by KR mutations (**Fig. 3** and **Table 1**). RNase-5KR shows an approximately unaltered $T_m$ compared to the wild-type protein, demonstrating that KR mutations can have a completely neutral effect on stability. PDIa-9KR shows an ~8˚ reduction compared to the 68 ˚C $T_m$ of the wild-type domain, while MA_2137-D65R-11KR shows an ~6˚ reduction compared to the 69 ˚C $T_m$ of the parental protein. Considering the entire set of mutant proteins in our study that could be purified, which includes 25 different KR mutations (**Table 1**), there is on average a 0.54 +/- 0.30˚ reduction in $T_m$ per KR mutation. Therefore, KR mutations are generally very well tolerated, although large sets of mutations tend to produce modest reductions in protein stability [52] that can reduce soluble protein yield *in vivo* when the stability of the wild-type (WT) protein is relatively low.

***Bulk KR mutations enhance crystallization propensity and yield strongly diffracting crystals.*** The purified protein constructs harboring the largest number of KR mutations (*i.e.*, PDIa-9KR and MA_2137-D65R-11KR) along with matched controls were screened for crystallization at the National Crystallization Center at the Hauptman-Woodward Institute (HWI) using their automated, high-throughput 1536-condition screen. This well-documented [53-58] microbatch-under-oil screen was employed for initial crystallization screening by the Northeast Structural Genomics Consortium[59-62] (www.nesg.org), which used it to generate 664 crystal structures deposited in the PDB. Neither the WT or 5KR construct of RNaseH yielded any crystallization hits in a screen intentionally conducted without any ligands stabilizing active site structure in order to provide the most exacting test of protein crystallization propensity; the lack of success for this protein was potentially influenced by the high 15 mg/ml protein concentration used for screening, which produced pervasive amorphous precipitation in the screen at the earliest observation times. However, the hPDIa-9KR and MA_2137-D65R-11KR constructs both yielded significantly more crystallization hits than the control proteins. MA_2137-D65R-11KR yielded hits under twice as many conditions as the MA_2137-D65R control protein, while hPDIa-9KR yielded 9 high quality hits compared to no hits at all for the WT construct (**Fig. 4** and **Table 1**). A small number of hit conditions for each protein were chosen for optimization, which very rapidly yielded 1.9 Å structures for both Bulk KR constructs based on a single session of remote synchrotron diffraction screening and data collection (**Figs. 5 & ED2** and **Table ED1**). Therefore, for both target proteins, crystallization screening only had to be conducted on the soluble construct harboring the largest number of Bulk KR mutations in order to rapidly obtain high-quality crystal structures.

***Bulk KR mutations frequently make H-bonds in crystal-packing interfaces without perturbing protein structure.*** The 1.9 Å crystal structures of our Bulk KR substitution (**Fig. 5** and **Table ED1**) constructs show 0.32-0.33 Å root-mean-square deviations for their backbone Cα atoms compared to the references structures (**Fig. ED2**) (*i.e.*, the much larger multidomain hPDI(abb'xa') construct for hPDIa because the isolated domain has never successfully been crystallized before and the parental MA_2137-D65R construct for MA_2137-D65R-11KR). The observed deviations are close to the expected coordinate error in well-refined crystal structures in the operative resolution range [63], indicating our Bulk KR substitution method does not significantly perturb protein conformation for either of our targets. Detailed analyses of the intermolecular interactions in our crystal structures demonstrates that the engineered arginine sidechains make extensive crystal packing contacts, substantially exceeding the number of van der Waals contacts and especially H-bonds made by the native arginine sidechains in the same constructs and greatly exceeding the number of both kinds of contacts made by lysine sidechains in the parental constructs (**Table 2** and **Fig. 5**). The larger number of crystal-packing contacts made by the engineered *vs.* native arginine residues could potentially reflect greater sequestration of the native residues in local surface interactions reducing the probability of reaching across a packing interface to make an energetically stabilizing interaction with a neighboring molecule in the crystal lattice. More extensive experimentation will be required to evaluate this possibility and to establish the statistical robustness of the trends documented in **Table 2**, but they nonetheless support the premise underlying our Bulk KR substitution strategy, which was based on the substantially stronger overrepresentation of arginine *vs.* lysine in crystal-packing interfaces in our large-scale analysis of crystal structures previously deposited in the PDB (**Fig. 1**).

On average, the well-ordered engineered arginine sidechains in our Bulk KR structures make 1.08 H-bonds each to a neighboring protein molecule in the crystal lattice, compared to 0.48 each for the native arginines (**Table 2**). In comparison, the well-ordered lysine sidechains make an average of 0.10 H-bonds each to a neighboring protein molecule in the native structures and none in our Bulk KR structures. These results support our hypothesis for the physicochemical basis of the greater overrepresentation of arginine compared to lysine in crystal-packing interfaces in the PDB, which is that that guanidino group in arginine is substantially more efficacious than the primary amine group in lysine in mediating energetically stabilizing H-bonds in the relevant stereochemical contexts (**Fig. 5**).

The number of van der Waals contacts per ordered sidechain follow a similar trend. Engineered and native arginines, respectively, make on average 4.62 and 3.43 contacts each, while lysines in the native structures make on average 0.28 contacts each, and lysines in the engineered structures making none (**Table 2**). The greater number of intermolecular van der Waals contacts made by the arginine sidechains could potentially be influenced by their greater H-bonding propensity leading to more frequent occurrence in crystal-packing interfaces, but additional research will be required to determine the relative energetic contributions of their van der Waals *vs.* H-bonding interactions to lattice stabilization. Notably, the number of backbone H-bonding and van der Waals interactions make by arginine *vs.* lysine residues in our reference and engineered structures do not show any clear trends (**Table 2**).

***Influence of Bulk KR mutations on protein solubility in PEG3350 solutions.*** Thermodynamic solubility assays using polyethylene glycol 3350 (PEG3350) to induce protein precipitation [64-66] assess the relative free energy of the hydrated state of individual protein molecules compared to the most favorable self-associated state under conditions of constant ionic strength but reduced water activity (effective concentration). In practice, these assays monitor optical density at 280 nm in the supernatant of solutions containing different concentrations of protein in the presence of increasing concentrations of PEG3350 after centrifugation to remove large particulate molecular assemblies, so they effectively measure the equilibrium concentration of protein that remains soluble as water activity is reduced. The observed results depend intrinsically on the free energy of the self-associated state, which varies significantly for different proteins in different solvent environments and can include crystalline phases and liquid-liquid phase separated (LLPS) phases in addition to heterogeneous amorphously precipitated phases. This factor can complicate interpretation of thermodynamic solubility assays, but they nonetheless provide insight into physicochemical properties that ultimately control protein crystallization behavior.

PEG3350 precipitation assays on WT hPDIa and the 9KR mutant show no significant difference in their behavior in PEG3350 precipitation assays (**Fig. ED3a**), indicating that the Bulk KR substitutions in this protein domain do not alter its thermodynamic solubility under these conditions even though they enable crystallization and high-resolution structure determination of a domain that does not crystallize at all with its native sequence. Even when harboring the 9KR mutations, hPDIa crystallizes only under a very small fraction (0.6%) of the solution conditions explored in high-throughput crystallization screening (**Fig. 4**) while showing amorphous precipitation in many of them (data not shown). Therefore, our thermodynamic solubility assays on the hPDIa constructs are likely measuring the free energy of the hydrated state of individual protein molecules compared to amorphously precipitated phases, and they demonstrate that the physicochemical properties controlling the formation of such phases are likely different from those controlling protein crystallization behavior.

PEG3350 precipitation assays on our MA_2137 constructs demonstrate more complex phase behavior (**Fig. ED3b-c**) likely reflecting different physical forms of self-association under assay conditions. Notably, the WT and D65R mutant could not be precipitated by the highest 35% (v/v) concentration of PEG3350 that was assayed (**Fig. ED3b**). These protein constructs instead showed some tendency to exhibit a small increase in optical density at low PEG3350 concentration, likely reflecting light scattering due some form of protein self-association in a low-density state that does not sediment during low-speed centrifugation. During crystallization screening, these constructs showed clear evidence of liquid-liquid phase separation LLPS without any apparent amorphous precipitation in many reaction conditions (**Fig. ED4**). Therefore, the inability to precipitate these constructs at high PEG3350 likely concentration likely reflects LLPS being energetically more favorable for this protein under conditions of low water activity than amorphous precipitation. Our crystal structures of MA_2137 constructs show clear and well-ordered electron density for every residue in this 202-residue protein except for the C-terminal hexahistidine tag that was added to enable purification using NiNTA affinity chromatography and a 12-residue internal loop that is disordered in the MA_2137-D65R-11KR structure, although well ordered by $Ca^{++}$ ions from the mother liquor in the structure of the parental MA_2137-D65R construct (**Fig. ED2b**). Furthermore, our CD thermal melting data demonstrate that the protein is very stably folded (**Fig. 3**). Therefore, our solubility data (**Fig. ED3b**) combined with our crystallization screening data (**Fig. ED4**) suggest that MA_2137 undergoes LLPS in an essentially fully folded conformational state.

In contrast to the behavior of the WT and the D65R constructs, the 5KR and 11KR constructs of MA_2137 show precipitation at the highest PEG3350 concentrations used in our solubility assays, with the 11KR construct showing stronger precipitation than the 5KR construct (**Fig. ED3c**). These results indicate the free energy of these MA_2137 constructs is lower in the precipitated state than in the LLPS state under conditions of very low water activity, reflecting a reduction in thermodynamic solubility. However, these constructs both crystallize extremely promiscuously, with the 5KR and 11KR constructs yielding crystallization hits in ~8% and ~15% of screened conditions, respectively (**Fig. 4**). These results raise the possibility that the precipitate formed by the 5KR and 11KR constructs at very high PEG3350 concentration could be in a microcrystalline state rather than amorphously precipitated state due to the high efficacy of the Bulk KR mutations in promoting crystallization. Further research will be needed to determine whether the reduced solubility of the 5KR and 11KR constructs reflects stabilization of crystalline states or amorphously precipitated states of MA_2137-D65R.

***Discussion.*** The results presented in this paper demonstrate the efficacy of a new method for probabilistic engineering of protein surface properties to enhance crystallization propensity based on substitution of multiple lysine (K) residues with arginine (R). The rationale behind this "Bulk KR" substitution method is that lysine and arginine have very similar physicochemical properties, but arginine shows substantially higher overrepresentation than lysine in a large-scale computational analysis we performed of crystal structures deposited in the PDB (**Fig. 1**). We have developed software to rank lysine sites for substitution based on the redundancy-corrected count of KR substitutions observed in homologous proteins with the highest level of sequence identity (**Fig. 2**), based on the rationale that biological evolution selects against destabilizing and function-impairing mutations. We demonstrate that mutations selected this way are only minimally destabilizing (**Fig. 3** and **Table 1**) and significantly enhance crystallization propensity for two of three test proteins (**Fig. 4**). The crystals yielded by our Bulk KR method diffract strongly and enabled efficient determination of a 1.9 Å crystal structure (**Table ED1**) for the hPDIa protein domain that does not crystallize at all with its native sequence (**Fig. 4** and **Table 1**). Our crystal structures of Bulk KR substituted proteins show no significant conformational or stereochemical differences *vs.* reference proteins (**Fig. ED2**), and the engineered arginine residues, like the native ones, making both van der Waals contacts and H-bonds in crystal-packing interfaces at substantially higher frequencies than either lysine residues or other residues (**Table 2**). These crystal structures were produced by the Bulk KR constructs harboring the highest number of substitutions, which were the only constructs for which any diffraction data were measured. These results support the efficacy of a streamlined pipeline for crystal structure determination in which solubility is tested for a set of constructs with an increasing number of KR mutations but purification and crystallization screening is only performed on the construct harboring the largest number of mutations. In summary, the biophysical results presented in this paper support bulk KR substitution being a rational and effective probabilistic strategy to engineer protein surface properties to enhance protein crystallization propensity.

Our Bulk KR method focuses on large-scale modification of protein surface properties while preserving physicochemical properties at every site. KR mutations have been explored in the past both for their ability to modulate protein stability and crystallization propensity. Bulk KR substitution in GFP was shown to greatly reduce the amount of soluble protein expressed *in vivo* in *E. coli* and also to reduce the fluorescence level of the protein that could be purified, although the mutations slowed the rate of unfolding by chemical denaturants [52]. However, this study only examined 14KR an 19KR mutations at sites selected based on diffuse criteria. Our studies show that KR mutations at 25 sites selected based on the frequency of substitution observed in homologs shows on average a 0.54 ˚C reduction in $T_m$ per KR mutation (**Fig. 3** and **Table 1**).

The previous studies of the influence of KR mutations on crystallization propensity were motivated by an earlier computational study using different stereochemical and statistical normalization methods to analyze the amino acids making crystal-packing interactions in a small set of 233 protein crystal structures [67]. This analysis produced very different conclusions when comparing results for all 20 amino acids (**Fig. ED5**), with the most salient difference being the conclusion that lysine, glutamate, and tryptophan are all disfavored in crystal-packing contacts, directly contradictory to the results of our computational analysis of 87,684 crystal structures (**Fig. 1**). Supporting the inaccuracy of this earlier conclusion and the validity of our analysis, we have successfully used introduction of both lysine and glutamate residues to improve protein crystallization (manuscript in preparation). The earlier computational analysis did conclude that arginine is favored in crystal-packing contacts, leading the authors to suggest that KR substitutions could improve protein crystallization propensity, but they did not perform any experiments testing this proposal. Two different groups subsequently tested this proposal using a small number of KR substitutions [24, 27]. Both concluded that KR substitution show limited efficacy in improving protein crystallization, but one of them was not able to obtain a diffracting crystal using a series of single-site KR substitutions [27]. The other group was able to obtain a crystal structure, but their overall conclusion was that lysine-to-alanine (KA) substitutions shows substantially greater efficacy in improving protein crystallization properties than KR substitutions[24]. Unfortunately, the KA method has had minimal impact improving protein crystallization based on structures submitted to the PDB during the ~20 years since it was proposed. One contributor to the lack of successful application of the KA method may be that the grossly different physicochemical properties of alanine compared to lysine can produce significant protein destabilization, which impedes introduction of multiple KA mutations. However, our computational analyses raise additional questions about the KA method because they indicate that alanine is underrepresented in crystal-packing interfaces while lysine is overrepresented (**Fig. 1**).

The conceptual foundation of our crystallization engineering methods is fundamentally different from that of these earlier studies. Our method focuses on probabilistic reengineering of protein surface properties via substitution of multiple amino acids with similar physicochemical properties but different propensity to make crystal-packing interactions based on large-scale computational analyses of previously determined crystal structures. The data presented in this paper support high efficacy for the method including importantly when applied in a streamlined fashion in which a series of mutant protein variants with increasing numbers of physiochemically conservative crystallization-enhancing mutations are evaluated for soluble protein expression and only the construct with greatest number of crystallization-enhancing is purified and subjected to crystallization screening (**Figs. 4-5**, **Tables 1-2**, and **Table ED1**).

Given the success of the Bulk KR method in improving protein crystallization behavior, our computational analysis of crystal-packing interactions in the PDB (**Fig. 1**) suggests several related strategies with promise to improve crystallization behavior based on the same conceptual approach. Aspartate and glutamate frequently substitute for one another in the course of evolution[68-70] due to their very similar physicochemical properties, but glutamate shows over 2-fold higher overrepresentation in crystal-packing interfaces (**Fig. 1**). A similar trend relative to crystal packing interactions is observed for asparagine and glutamine, which also have very similar physicochemical properties. These observations suggest bulk aspartate-to-glutamate (DE) and asparagine-to-glutamine (NQ) substitutions are also likely to improve crystallization propensity.

In the case of these substitutions, the higher entropy of the sidechain with greater crystal-packing propensity will tend to thermodynamically oppose immobilization in a crystal-packing interface, while this factor does not apply to bulk lysine-to-arginine substitution due to the very similar entropy of these sidechains. However, our computational analysis of crystal-packing interactions in the PDB shows that high-entropy sidechains mediating crystal-packing interactions tend to participate in salt-bridging and H-bonding interactions with nearby residues in the primary sequence, especially at ±3 and ±4 positions in α-helices and ±2 positions in β-strands [36-38] (manuscript in preparation). These interactions likely reduce the entropy of the sidechains in the isolated protein molecules, which will reduce or eliminate entropy loss due to immobilization in a crystal-packing interface. Therefore, bulk DE and NQ substitution seems likely to be most effective when residues with potential salt-bridging or H-bonding partners at ±3 and ±4 positions in α-helices or ±2 positions in β-strands are prioritized for substitution.

Future research will be required to assess the efficacy of these alternative bulk substitution methods and to establish the most efficient paradigm for combing KR, DE, and NQ mutations to maximize crystallization hit rate and crystal quality while minimizing the number of constructs needed to obtain a diversity of different crystal forms and a high-resolution crystal structure. Nonetheless, our bulk substitution methods focused on large-scale probabilistic remodeling of protein surface properties to enhance crystallization propensity already show significant efficacy for rational engineering of proteins to improve their crystallization properties.

***Acknowledgements.*** This work was supported by a grant from the US NIH-NIGMS to JFH (GM127883). We thank Accendro Inc. for assistance with web programming and G.T. Montelione, R. Xiao, and the other members of the Northeast Structural Genomics Consortium for long-term collaboration and advice.

***Author contributions.*** Conceptualization: JFH. Methodology: NEB, SS, SKH, WNP, HSH, & JFH. Investigation: NEB, BJL, SS, MN, & KHW. Visualization: NEB &JFH. Funding acquisition: JFH. Project administration: JFH. Supervision: JFH. Writing: NEB & JFH.

***Conflict of interest.*** JFH is a member of the Scientific Advisory Board of Nexomics Biosciences. The other authors declare no competing financial interests.

*Note: The* ***Supplementary Information*** *for this manuscript contains 3 figures and 1 table.*

# References

1. Kendrew, J.C. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666 (1958).
2. Kendrew, J.C. Structure and function in myoglobin and other proteins. *Federation proceedings* **18**, 740-751 (1959).
3. Kendrew, J.C. & Perutz, M.F. A comparative X-ray study of foetal and adult sheep haemoglobins. *Proceedings of the Royal Society of London* **194**, 375-398 (1948).
4. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019).
5. Terwilliger, T.C. Maximum-likelihood density modification using pattern recognition of structural motifs. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 1755-1762 (2001).
6. Otwinowski, Z. & Minor, W. Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307-326 (1997).
7. Sheldrick, G.M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D Biol Crystallogr* **66**, 479-485 (2010).
8. Liu, Q. & Hendrickson, W.A. Contemporary Use of Anomalous Diffraction in Biomolecular Structure Analysis. *Methods Mol Biol* **1607**, 377-399 (2017).
9. Hendrickson, W.A., Horton, J.R. & LeMaster, D.M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *The EMBO journal* **9**, 1665-1672 (1990).
10. Wilson, M.A. Mapping Enzyme Landscapes by Time-Resolved Crystallography with Synchrotron and X-Ray Free Electron Laser Light. *Annu Rev Biophys* **51**, 79-98 (2022).
11. Grimes, J.M. et al. Where is crystallography going? *Acta Crystallogr D Struct Biol* **74**, 152-166 (2018).
12. Sanishvili, R. & Fischetti, R.F. Applications of X-Ray Micro-Beam for Data Collection. *Methods Mol Biol* **1607**, 219-238 (2017).
13. Hendrickson, W.A. Synchrotron crystallography. *Trends Biochem Sci* **25**, 637-643 (2000).
14. 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. *J Mol Biol* **344**, 977-991 (2004).
15. Slabinski, L. et al. The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* **16**, 2472-2482 (2007).
16. Price, W.N., 2nd et al. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat Biotechnol* **27**, 51-57 (2009).
17. 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).

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

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

20. Derewenda, Z.S. & Godzik, A. The "Sticky Patch" Model of Crystallization and Modification of Proteins for Enhanced Crystallizability. *Methods Mol Biol* **1607**, 77-115 (2017).

21. Cieslik, M. & Derewenda, Z.S. The role of entropy and polarity in intermolecular contacts in protein crystals. *Acta crystallographica* **65**, 500-509 (2009).

22. Janda, I. et al. Harvesting the high-hanging fruit: the structure of the YdeN gene product from Bacillus subtilis at 1.8 angstroms resolution. *Acta Crystallogr D Biol Crystallogr* **60**, 1101-1107 (2004).

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

24. Czepas, J. et al. The impact of Lys-->Arg surface mutations on the crystallization of the globular domain of RhoGDI. *Acta Crystallogr D Biol Crystallogr* **60**, 275-280 (2004).

25. 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 crystallographica* **58**, 1983-1991 (2002).

26. 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 crystallographica* **57**, 679-688 (2001).

27. Anstrom, D.M., Colip, L., Moshofsky, B., Hatcher, E. & Remington, S.J. Systematic replacement of lysine with glutamine and alanine in Escherichia coli malate synthase G: effect on crystallization. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **61**, 1069-1074 (2005).

28. Naumov, V., Price, W.N., Handelman, S.K. & Hunt, J.F. (ed. U.P.a.T. Office) (The Trustess of Columbia University in the City of New York, United States; 2019).

29. Sternberg, M.J. & Chickos, J.S. Protein side-chain conformational entropy derived from fusion data--comparison with other empirical scales. *Protein engineering* **7**, 149-155 (1994).

30. DuBay, K.H. & Geissler, P.L. Calculation of proteins' total side-chain torsional entropy and its influence on protein-ligand interactions. *J Mol Biol* **391**, 484-497 (2009).

31. Bhowmick, A. & Head-Gordon, T. A monte carlo method for generating side chain structural ensembles. *Structure* **23**, 44-55 (2015).

32. Srinivasan, R. & Rose, G.D. A physical basis for protein secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14258-14263 (1999).

33. Wetlaufer, D.B. & Lovrien, R. Induction of Reversible Structural Changes in Proteins by Nonpolar Substances. *J Biol Chem* **239**, 596-603 (1964).

34. Bennion, B.J. & Daggett, V. The molecular basis for the chemical denaturation of proteins by urea. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5142-5147 (2003).

35. Scott, J.N., Nucci, N.V. & Vanderkooi, J.M. Changes in water structure induced by the guanidinium cation and implications for protein denaturation. *J Phys Chem A* **112**, 10939-10948 (2008).

36. Olson, C.A., Spek, E.J., Shi, Z., Vologodskii, A. & Kallenbach, N.R. Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins* **44**, 123-132 (2001).

37. Donald, J.E., Kulp, D.W. & DeGrado, W.F. Salt bridges: geometrically specific, designable interactions. *Proteins* **79**, 898-915 (2011).

38. Vener, M.V., Odinokov, A.V., Wehmeyer, C. & Sebastiani, D. The structure and IR signatures of the arginine-glutamate salt bridge. Insights from the classical MD simulations. *The Journal of chemical physics* **142**, 215106 (2015).

39. Khan, M.M., Simizu, S., Kawatani, M. & Osada, H. The potential of protein disulfide isomerase as a therapeutic drug target. *Oncol Res* **19**, 445-453 (2011).

40. Hoffstrom, B.G. et al. Inhibitors of protein disulfide isomerase suppress apoptosis induced by misfolded proteins. *Nature chemical biology* **6**, 900-906 (2010).

41. Wang, C. et al. Structural insights into the redox-regulated dynamic conformations of human protein disulfide isomerase. *Antioxid Redox Signal* **19**, 36-45 (2013).

42. Katayanagi, K. et al. Structural details of ribonuclease H from Escherichia coli as refined to an atomic resolution. *J Mol Biol* **223**, 1029-1052 (1992).

43. Yang, W., Hendrickson, W.A., Crouch, R.J. & Satow, Y. Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. *Science (New York, N.Y* **249**, 1398-1405 (1990).

44. Liao, Z. et al. Pivotal role of a conserved histidine in Escherichia coli ribonuclease HI as proposed by X-ray crystallography. *Acta Crystallogr D Struct Biol* **78**, 390-398 (2022).

45. Katayanagi, K. et al. Crystal structures of ribonuclease HI active site mutants from Escherichia coli. *J Biol Chem* **268**, 22092-22099 (1993).

46. Katayanagi, K., Okumura, M. & Morikawa, K. Crystal structure of Escherichia coli RNase HI in complex with Mg2+ at 2.8 A resolution: proof for a single Mg(2+)-binding site. *Proteins* **17**, 337-346 (1993).

47. Goedken, E.R. & Marqusee, S. Co-crystal of Escherichia coli RNase HI with Mn2+ ions reveals two divalent metals bound in the active site. *J Biol Chem* **276**, 7266-7271 (2001).

48. Moayed, F. et al. The Anti-Aggregation Holdase Hsp33 Promotes the Formation of Folded Protein Structures. *Biophys J* **118**, 85-95 (2020).

49. Graf, P.C. et al. Activation of the redox-regulated chaperone Hsp33 by domain unfolding. *J Biol Chem* **279**, 20529-20538 (2004).

50. Ishikawa, K., Nakamura, H., Morikawa, K. & Kanaya, S. Stabilization of Escherichia coli ribonuclease HI by cavity-filling mutations within a hydrophobic core. *Biochemistry* **32**, 6171-6178 (1993).

51. Goedken, E.R., Keck, J.L., Berger, J.M. & Marqusee, S. Divalent metal cofactor binding in the kinetic folding trajectory of Escherichia coli ribonuclease HI. *Protein Sci* **9**, 1914-1921 (2000).

52. Sokalingam, S., Raghunathan, G., Soundrarajan, N. & Lee, S.G. A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PLoS One* **7**, e40410 (2012).

53. Lynch, M.L., Snell, M.E., Potter, S.A., Snell, E.H. & Bowman, S.E.J. 20 years of crystal hits: progress and promise in ultrahigh-throughput crystallization screening. *Acta Crystallogr D Struct Biol* **79**, 198-205 (2023).

54. Budziszewski, G.R., Snell, M.E., Wright, T.R., Lynch, M.L. & Bowman, S.E.J. High-Throughput Screening to Obtain Crystal Hits for Protein Crystallography. *J Vis Exp* (2023).

55. Luft, J.R., Wolfley, J.R. & Snell, E.H. What's in a drop? Correlating observations and outcomes to guide macromolecular crystallization experiments. *Crystal growth & design* **11**, 651-663 (2011).

56. Luft, J.R., Snell, E.H. & Detitta, G.T. Lessons from high-throughput protein crystallization screening: 10 years of practical experience. *Expert opinion on drug discovery* **6**, 465-480 (2011).

57. Luft, J.R. et al. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J Struct Biol* **142**, 170-179 (2003).

58. Luft, J. et al. Macromolecular crystallization in a high throughput laboratory in the search phase. *J. Crys. Growth*, 591-595 (2001).

59. Everett, J.K. et al. A community resource of experimental data for NMR / X-ray crystal structure pairs. *Protein Sci* **25**, 30-45 (2016).

60. Boel, G. et al. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358-363 (2016).

61. Acton, T.B. et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol* **493**, 21-60 (2011).

62. Xiao, R. et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J Struct Biol* **172**, 21-33 (2010).

63. Cruickshank, D. The required precision of intensity measurements for single-crystal analysis. *Acta crystallographica* **13**, 774-777 (1960).

64. Kita, Y., Arakawa, T., Lin, T.Y. & Timasheff, S.N. Contribution of the surface free energy perturbation to protein-solvent interactions. *Biochemistry* **33**, 15178-15189 (1994).

65. Bhat, R. & Timasheff, S.N. Steric exclusion is the principal source of the preferential hydration of proteins in the presence of polyethylene glycols. *Protein Sci* **1**, 1133-1143 (1992).

66. Arakawa, T. & Timasheff, S.N. Mechanism of poly(ethylene glycol) interaction with proteins. *Biochemistry* **24**, 6756-6762 (1985).

67. Dasgupta, S., Iyer, G.H., Bryant, S.H., Lawrence, C.E. & Bell, J.A. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins* **28**, 494-514 (1997).

68. Henikoff, S. & Henikoff, J.G. Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49-61 (1993).

69. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1293-1301 (2011).

70. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of residue coupling in protein families. *IEEE/ACM Trans Comput Biol Bioinform* **5**, 183-197 (2008).

71. Hwu, W.-m.W., Kirk, D.B. & El Hajj, I. in Programming Massively Parallel Processors (Fourth Edition). (eds. W.-m.W. Hwu, D.B. Kirk & I. El Hajj) 211-233 (Morgan Kaufmann, 2023).

72. Okuta, R., Unno, Y., Nishino, D., Hido, S. & Crissman (2017).

73. Terrific Broth. *Cold Spring Harbor Protocols* (2015).

74. Gill, S.C. & von Hippel, P.H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**, 319-326 (1989).

75. Kabsch, W. Automatic indexing of rotation diffraction patterns. *J.Appl.Crystallogr.* **21**, 67-71 (1988).

76. Kabsch, W. Evaluation of single-crystal x-ray diffraction data from a position-sensitive detector. *J.Appl.Crystallogr.* **21**, 916-924 (1988).

77. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr* **66**, 133-144 (2010).

78. Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132 (2010).

79. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* **66**, 22-25 (2010).

80. McRee, D.E. XtalView/Xfit-A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156-165 (1999).

81. Casanal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Sci* **29**, 1069-1078 (2020).

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

83. Snell, E.H. et al. The application and use of chemical space mapping to interpret crystallization screening results. *Acta Crystallogr D Biol Crystallogr* **64**, 1240-1249 (2008).

84. Pettersen, E.F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021).

# Methods

***GPU acceleration of sequence identity calculation.*** Our Python program accelerates calculation of the absolute percent identity between two sequences by parallelizing site-by-site comparison on a Graphical Processing Unit (GPU) chip using a custom reduction kernel [71] written using the CuPy library [72]. In brief, each pair of aligned amino acids in two sequences is sent to a separate GPU core. If the 1-character amino acid codes in both sequences match each other and neither is empty due to a gap in the sequence alignment, that core stores a value of 1 in its register, which is otherwise set to zero. The values in the registers for all sites are then summed using parallel reduction to count the number of identical amino acids in the sequence. Details can be found in the code at https://github.com/huntmolecularbiophysicslab/pxengineering.

***Prioritization of mutation sites based on redundancy-corrected counts of KR mutations observed in homologous proteins.*** In order to compensate for outright redundancy as well as inhomogeneous phylogenetic sampling in sequence databases, our software performs two redundancy-compensation calculations on the set of sequences in each percent-identity bin having an arginine substitution at a specific lysine site in the target protein. Both calculations use the same heuristic estimate for the probability of evolutionary resampling at an aligned site between two sequences *i* and *j* with an overall fraction of $f_{id}$ identical residues at all aligned sites:

$$P_{site-resampling}(f_{id})_{ij} = \left\{ \begin{matrix} 1.0 \;\; for\; f_{id} \leq f_{min} \\ \left(\frac{1 - f_{id}}{1 - f_{min}}\right) \;\; for\; f_{id} > f_{min} \end{matrix} \right\}.$$

We assume $f_{min} = 0.3$. One calculation gives a redundancy-reduced estimate $C_R$ of arginine counts using the following formula in which the summation is performed over all unique pairs of the $N$ sequences in the bin having an arginine substitution at one lysine site in the target protein:

$$C_R = \left(\frac{2}{N}\sum_{i<j} P_{site-resampling}(f_{id})_{ij}\right) + 1\,.$$

This calculation is extremely rapid and includes all homologs but is only rigorously accurate in the limiting cases of full redundancy or full independence. The second calculation gives a rigorous estimate of the expectation value for the number of independent observations of arginine based on combinations of the heuristic probabilities of being resampled or not being resampled for all unique pairs among the seven most diverged sequences in each bin that have arginine at a specific lysine site. These sequences are identified by taking the single sequence with the lowest percent identity to the target sequence and then progressively adding the sequence with the lowest average percent identity to those already selected. The details of the implementation can be found in the code at https://github.com/huntmolecularbiophysicslab/pxengineering.

***Protein expression and purification.*** Protein coding sequences harboring Bulk KR Substitutions were synthesized (Twist Bioscience, South San Francisco, CA) with a short C-terminal hexahistidine tag (LEHHHHHH), cloned into the T7 expression plasmid pET21_NESG (https://dnasu.org/DNASU/GetCloneDetail.do?cloneid=336944), and then transformed into BL21(DE3) Rosetta *E. coli* cells (MilliporeSigma, Burlington, MA). Protein expression was induced with 1mM IPTG for 4 hours at 30°C in Terrific Broth [73]. Cells were pelleted by centrifugation at 4,000 rpm for 25 min at 4˚C and resuspended on ice in 10 mM imidazole, 300 mM NaCl, 1 mM TCEP, 5% (w/v) glycerol, 50 mM $NaH_2PO_4$ , pH 7.5, prior to cell lysis by probe sonication. The supernatant following 15,000 rpm centrifugation at 4°C was mixed with Ni-NTA resin and incubated at 4 °C for 1 hour. The mixture was then transferred into a column and washed with the same buffer containing a higher 100 mM imidazole concentration prior to elution of the protein in 6 ml of the same buffer containing 250 mM imidazole. A 1mL aliquot of eluted protein was concentrated to 500 µl using an Amicon 10 kDa centrifugal filter (MilliporeSigma, Burlington, MA) and loaded via a 1 ml loop for onto a Superdex 200 Increase 10/300 gl gel filtration column equilibrated in 100 mM NaCl, 10 mM DTT, 10 mM Tris-Cl, pH 7.5. Protein-containing fractions were concentrated to ~15 mg/ml based on *a priori* sequence-based extinction coefficients [74] and $OD_{280nm}$ values measured using a Nanodrop spectrophotometer (ThermoFisher, Waltham, MA), and the concentrated protein was immediately flash-frozen in aliquots in liquid nitrogen prior to storage at -80 ˚C pending use.

***Thermal stability assays using CD spectroscopy.*** An Applied Photophysics (Leatherhead, UK) Chirascan V100 spectropolarimeter with a Peltier-jacketed cell holder was used to collect circular dichroism (CD) spectra spanning 200-250 nm serially during a 3 ˚C/min thermal ramp nominally running from 10-84 ˚C. Data were measured from protein samples in a 0.5 mm quartz cuvette in 1 nm increments using a 0.25 integration time per point and a 1 nm bandwidth, corresponding to 35 sec per spectrum. Measurements of the actual cell temperature during the experiment, which were used for data display and analysis, indicated the actual range of the temperature ramp was from ~11-78 ˚C for all samples. Protein samples were diluted to 2 mg/ml using gel filtration buffer lacking DTT (*i.e.*, 100 mM NaCl, 10 mM Tris-Cl, pH 7.5) with the addition of 1 mM SAH for the MA_237 constructs. Global curve fitting of spectral data during the thermal ramp was performed from 215-230 nm using the program GLOBAL3 (Applied Photophysics) using double linear baseline correction (*i.e.*, before and after the observed transitions) to extract the thermodynamic parameters and melting temperatures (**Table 1** and **Fig. 3**). Each dataset was analyzed using the smallest number of transitions showing approximately random directions for the residuals for adjacent points in the CD *vs.* measured temperature plane.

***Solubility Assays.*** Protein stock solutions were diluted to working concentration in the same buffer used for gel-filtration chromatography containing different weight/volume concentrations of PEG3350. Following a 60 minute incubation at room temperature, the samples were spun for 10 min at 14,000 RPM in a microfuge to pellet particulates, and the concentration of protein in the supernatant was measured using the optical density at 280 nm measured in a Nanodrop spectrophotometer (ThermoFisher) based on the *a priori* extinction coefficient [74]. The centrifugation and measurement of concentration in the supernatant were repeated 24 hours later to ensure equilibrium had been reached.

***Protein crystallization.*** A subset of the crystal hits for hPDIa-9KR and MA_2137-D65R-11KR observed in the 1536-well screen conducted at the High-Throughput Crystallization Screening Center [53-58] at the Hauptman-Woodward Medical Research Institute (https://hwi.buffalo.edu/high-throughput-crystallization-center/) using 15 mg/ml protein stock concentrations were initially reproduced using the microbatch-under-oil method at 4 °C and 18 °C and subsequently optimized by seeding. The hPDIa-9KR stock was mixed at a 2:1 volume ration with a crystallization reagent comprising 24% (w/v) PEG 20k, 0.1 M potassium thiocyanate, 0.1 M MES, pH 6. The MA_2137-D65R-11KR stock was mixed at a 1:1 volume ration with a crystallization reagent comprising 30% (w/v) PEG 1k, 0.1 M HEPES, pH 7.5. All crystals were transferred into a similar crystallization solution supplemented with 20% (v/v) ethylene glycol prior to mounting and flash-freezing in liquid nitrogen. The WT and 5KR RNaseH constructs were screened in the same manner also using a protein stock concentration of 15 mg/ml but showed pervasive amorphous precipitation without yielding any crystallization hits, suggesting screening may have been conducted at too high a protein concentration.

***Crystal structure determination and refinement.*** X-ray diffraction data were collected from single crystals of PDIa-9M and MA_2137-D65R-11 using, respectively, the NE-CAT 24-ID-E and 24-ID-C beam lines at the Advanced Photon Source (**Table ED1**). The images were processed and scaled using XDS [75-78]. The structure of hPDIa-9KR was solved by molecular replacement using the program MOLREP [79] employing a search model comprising the first domain in the crystal structure of full-length hPDI (PDB id 4EKZ). The structure of MA_2137-D65R-11KR was solved using the same methods employing the structure of MA_2137-D65R (PDB id 6MRO) as the search model. Both structures were refined (**Table ED1**) using PHENIX [4] in conjunction with manual rebuilding in XtalView [80] and COOT [81].

**Table 1.** ***Summary of expression, stability, and crystallization results for Bulk-KR mutant proteins.***[*]

| Target Protein | Construct | KR mutation sites | Expression | Solubility | $T_m$ (˚C) | $\Delta H_{vH}$ (kcal/mol) | # hits at 4 weeks | Crystal structure resolution (Å) |
|---|---|---|---|---|---|---|---|---|
| **hPDIa**<br>***120 amino acids*** | **WT** | – | +++ | +++ | **59.9, 67.9** | **204 ± 36, 131 ± 9** | **0** | ***n/a*** |
| | **2KR** | **42, 114** | +++ | +++ | **58.9, 68.6** | **201 ± 40, 117 ± 5** | ***n/a*** | ***n/a*** |
| | **5KR** | **+ 130, 69, 71** | +++ | +++ | **48.2, 65.7** | **47.7 ± 2, 125 ± 3** | ***n/a*** | ***n/a*** |
| | **7KR** | **+ 31, 131** | +++ | +++ | **49.5, 64.1** | **89.8 ± 5, 156 ± 5** | ***n/a*** | ***n/a*** |
| | **9KR** | **+ 57, 65** | +++ | +++ | **55.4, 60.4** | **163 ± 13, 161 ± 3** | **9** | **1.89 Å** |
| **MA_2137**<br>***194 amino acids*** | **WT** | – | +++ | +++ | **69.0** | **283 ± 13.5** | **75** | ***n/a*** |
| | **D65R** | – | +++ | +++ | **68.7** | **250 ± 4.60** | **126** | **1.60 Å** |
| | **D65R-3KR** | **126, 129, 194** | +++ | +++ | **67.1** | **194 ± 3.30** | ***n/a*** | ***n/a*** |
| | **D65R-5KR** | **+ 52 71** | +++ | +++ | **67.4** | **153 ± 7.10** | ***n/a*** | ***n/a*** |
| | **D65R-7KR** | **+ 172, 133** | +++ | +++ | **65.5** | **193 ± 3.90** | ***n/a*** | ***n/a*** |
| | **D65R-11KR** | **+ 8, 64, 142, 155** | +++ | +++ | **63.4** | **150 ± 2.60** | **238** | **1.91 Å** |
| **RNaseH**<br>***166 amino acids*** | **WT** | – | +++ | +++ | **45.2** | **135 ± 2.20** | **0** | ***n/a*** |
| | **2KR** | **31, 90** | +++ | +++ | ***n/a*** | ***n/a*** | ***n/a*** | ***n/a*** |
| | **5KR** | **+ 66, 89, 35** | +++ | +++ | **45.5** | **128 ± 2.60** | **0** | ***n/a*** |
| | **7KR** | **+ 107, 124** | +++ | + | ***n/a*** | ***n/a*** | ***n/a*** | ***n/a*** |
| | **11KR** | **+ 111, 62, 88, 101** | +++ | – | ***n/a*** | ***n/a*** | ***n/a*** | ***n/a*** |

[*] The constructs harboring increasing numbers of KR mutations also include all of those in the constructs of the same protein harboring fewer KR mutations. The code "*n/a*" stands for not applicable or not attempted.

**Table 2.** ***Crystal-packing contacts in reference and Bulk KR protein structures.****

| Construct | WT | 9KR | | D65R | D65R-11KR | Totals |
|---|---|---|---|---|---|---|
| PDB ID (chain) | 4EKZ (A) | 8GDY (A) | 8GDY (B) | 6MRO (A) | 8GDU (A) | |
| Resolution (Å) | 2.51 | 1.93 | 1.93 | 1.6 | 1.95 | |
| Crystal Solvent Content | 44.20% | 37.40% | 37.40% | 35.20% | 51.20% | |
| Domain | hPDIa | hPDIa | hPDIa | MA_2137 | MA_2137 | |
| # residues in domain | 120 | 118 | 120 | 194 | 194 | 746 |
| # disordered K | 0 | 0 | 0 | 0 | 0 | 0 |
| # disordered R | 0 | 0 | 0 | 0 | 3 | 3 |
| # ordered surface residues | 75 | 74 | 74 | 118 | 118 | 459 |
| # ordered surface residues not K or R | 58 | 57 | 57 | 94 | 97 | 363 |
| # ordered K (all surface-exposed) | 11 | 2 | 2 | 13 | 2 | 30 |
| # ordered R (all surface-exposed) | 6 | 15 | 15 | 11 | 19 | 66 |
| # ordered R (native) | 6 | 6 | 6 | 11 | 11 | 40 |
| # ordered R (engineered) | n/a | 9 | 9 | n/a | 8 | 26 |
| # BB vdW contacts | 86 | 44 | 52 | 88 | 52 | 322 |
| # BB vdW contacts not K or R | 82 | 39 | 48 | 79 | 47 | 295 |
| # BB vdW contacts of K | 3 | 0 | 0 | 9 | 0 | 12 |
| # BB vdW contacts of R | 1 | 5 | 4 | 0 | 5 | 15 |
| # BB vdW contacts of native R | 1 | 0 | 0 | 0 | 1 | 2 |
| # BB vdW contacts of engineered R | n/a | 5 | 4 | n/a | 4 | 13 |
| BB vdW contacts per residue in domain | 0.72 | 0.37 | 0.43 | 0.45 | 0.27 | 0.43 |
| BB vdW per surface residue | 1.15 | 0.59 | 0.70 | 0.75 | 0.44 | 0.70 |
| BB vdW contacts per surface residue not K or R | 1.41 | 0.68 | 0.84 | 0.84 | 0.48 | 0.81 |
| BB vdW contacts per K | 0.27 | 0 | 0 | 0.69 | 0 | 0 |
| BB vdW contacts per R | 0.17 | 0.33 | 0.27 | 0 | 0.26 | 0.23 |
| BB vdW contacts per native R | 0.17 | 0 | 0 | 0 | 0.09 | 0.05 |
| BB vdW contacts per engineered R | n/a | 0.56 | 0.44 | n/a | 0.50 | 0.50 |
| # BB H-bonds | 9 | 0 | 4 | 8 | 7 | 28 |
| # BB H-bonds not K or R | 9 | 0 | 3 | 6 | 6 | 24 |
| # BB H-bonds K | 0 | 0 | 0 | 2 | 0 | 2 |
| # BB H-bonds R | 0 | 0 | 1 | 0 | 1 | 2 |
| # BB H-bonds native R | 0 | 0 | 0 | 0 | 0 | 0 |
| # BB H-bonds engineered R | n/a | 0 | 1 | n/a | 1 | 2 |
| BB H-bond per residue in domain | 0.08 | 0 | 0.03 | 0.04 | 0.04 | 0.04 |
| BB H-bond per surface residue | 0.12 | 0 | 0.05 | 0.07 | 0.06 | 0.06 |
| BB H-bonds per surface residue not K or R | 0.16 | 0 | 0.05 | 0.06 | 0.06 | 0.07 |
| BB H-bonds per K | 0 | 0 | 0 | 0 | 0 | 1 |
| BB H-bonds per R | 0 | 0 | 0.07 | 0 | 0.05 | 0.03 |
| BB H-bonds per native R | 0 | 0 | 0 | 0 | 0 | 0 |
| BB H-bonds per engineered R | n/a | 0 | 0.11 | n/a | 0.13 | 0.08 |
| # SC vdW contacts | 119 | 167 | 149 | 228 | 192 | 855 |
| # SC vdW contacts not K or R | 106 | 100 | 89 | 169 | 125 | 589 |
| # SC vdW contacts of K | 1 | 0 | 0 | 6 | 0 | 7 |
| # SC vdW contacts of R | 12 | 67 | 60 | 53 | 67 | 259 |
| # SC vdW contacts of native R | 12 | 20 | 20 | 53 | 34 | 139 |
| # SC vdW contacts of engineered R | n/a | 47 | 40 | n/a | 33 | 120 |
| SC vdW per residue in domain | 0.99 | 1.42 | 1.24 | 1.18 | 0.99 | 1.15 |
| SC vdW per surface residue | 1.59 | 2.26 | 2.01 | 1.93 | 1.63 | 1.86 |
| SC vdW contacts per surface residue not K or R | 1.83 | 1.75 | 1.56 | 1.80 | 1.29 | 1.62 |
| SC vdW contacts per K | 0.09 | 0 | 0 | 0.46 | 0 | 0 |
| SC vdW contacts per R | 2.00 | 4.47 | 4.00 | 4.82 | 3.53 | 3.92 |
| SC vdW contacts per native R | 2.00 | 3.33 | 3.33 | 4.82 | 3.09 | 3.48 |
| SC vdW contacts per engineered R | n/a | 5.22 | 4.44 | n/a | 4.13 | 4.62 |
| # SC H-bonds | 11 | 29 | 31 | 42 | 10 | 123 |
| # SC H-bonds not K or R | 7 | 16 | 18 | 32 | 0 | 73 |
| # SC H-bonds K | 1 | 0 | 0 | 2 | 0 | 3 |
| # SC H-bonds R | 3 | 13 | 13 | 8 | 10 | 47 |
| # SC H-bonds native R | 3 | 4 | 3 | 8 | 1 | 19 |
| # SC H-bonds engineered R | n/a | 9 | 10 | n/a | 9 | 28 |
| # K making 2, 1, 0 SC H-bonds | 0, 1, 10 | 0, 0, 2 | 0, 0, 2 | 1, 0, 12 | 0, 0, 2 | 1, 1, 28 |
| # R making 5, 4, 3, 2, 1, 0 SC H-bonds | 0, 0, 0, 1, 1, 4 | 1, 1, 0, 1, 2, 10 | 0, 1, 2, 1, 1, 10 | 1, 0, 0, 1, 1, 7 | 0, 0, 1, 2, 3, 13 | 2, 2, 3, 6, 9, 44 |
| # native R making 5, 4, 3, 2, 1, 0 SC H-bonds | 0, 0, 0, 1, 1, 4 | 0, 1, 0, 0, 0, 5 | 0, 0, 1, 0, 0, 5 | 1, 0, 0, 1, 1, 7 | 0, 0, 0, 0, 1, 10 | 1, 1, 1, 1, 5, 31 |
| # engineered R making 5, 4, 3, 2, 1, 0 SC H-bonds | n/a | 1, 0, 0, 1, 2, 5 | 0, 1, 1, 1, 1, 5 | n/a | 0, 0, 1, 2, 2, 3 | 1, 1, 2, 4, 5, 13 |
| SC H-bonds per residue | 0.09 | 0.25 | 0.26 | 0.22 | 0.05 | 0.16 |
| SC H-bond per surface residue | 0.15 | 0.39 | 0.42 | 0.36 | 0.08 | 0.27 |
| SC H-bonds per surface residue not K or R | 0.12 | 0.28 | 0.32 | 0.34 | 0 | 0.20 |
| SC H-bonds per K | 0.09 | 0 | 0 | 0.15 | 0 | 0.10 |
| SC H-bonds per R | 0.50 | 0.87 | 0.87 | 0.73 | 0.53 | 0.71 |
| SC H-bonds per native R | 0.50 | 0.67 | 0.50 | 0.73 | 0.09 | 0.48 |
| SC H-bonds per engineered R | n/a | 1.00 | 1.11 | n/a | 1.13 | 1.08 |

[*] The abbreviation n/a stands for not applicable. Ratios per engineered arginine reflect exclusively ordered residues and exclude the three disordered arginine residues in the MA_2137-D65R-11KR structure. Candidate H-bonds were initially identified based on the participating heteroatoms having an internuclear separation ≤ 3.5 Å, but they were included in the count only if visually confirmed to have reasonable interaction geometry. Among the five crystal structures analyzed here, only two potential H-bonding interactions in crystal-packing interfaces fulfilled the distance criterion but failed the geometric evaluation. Individual atoms fulfilling the basic distance and geometric criteria with two different potential H-bonding partner atoms were counted as contributing two H-bonds [38], consistent with Coulomb's law being additive. Atom pairs with internuclear separation ≤ 4.0 Å not meeting the H-bonding criteria were counted as van der Waals contacts.

# Figure 1

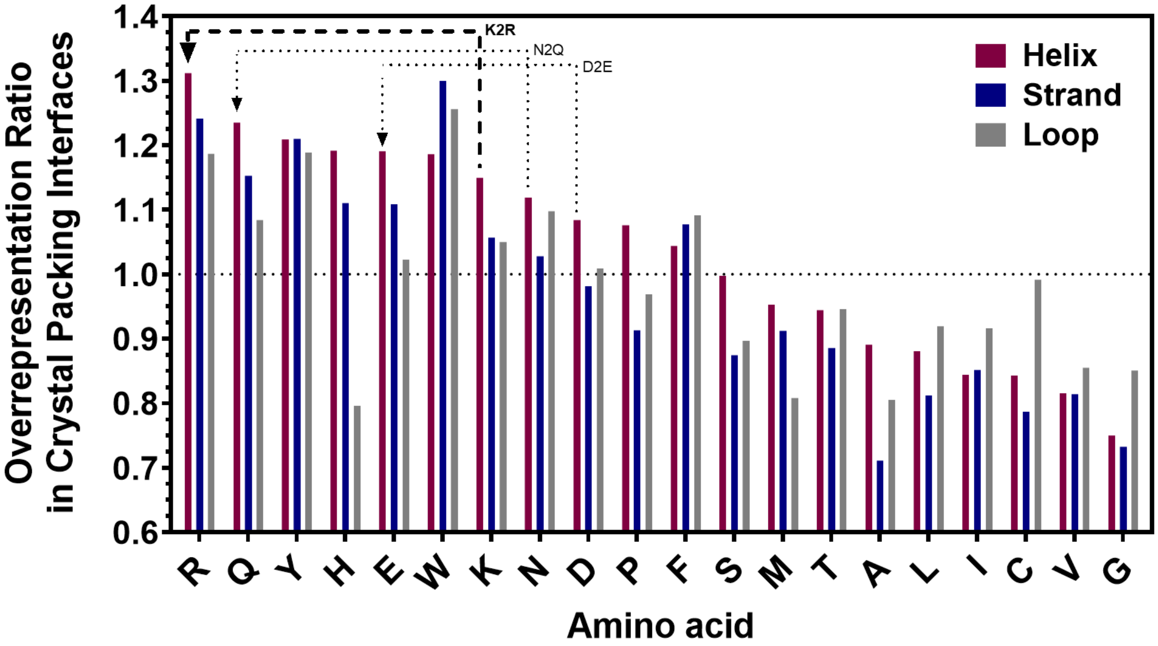

**Figure 1.** ***Overrepresentation ratios*** [28] ***of amino-acids in crystal-packing interfaces compared to overall surface composition in 87,684 crystal structures deposited in the Protein Data Bank (PDB).*** The red circles/arrows schematize the gain in packing probability produced by K-to-R mutations. The overrepresentation ratios are segregated by protein 2° structure as assessed by DSSP [82], and the amino acids are ordered in decreasing order of overrepresentation ratio in α-helical secondary structure.

# Figure 2

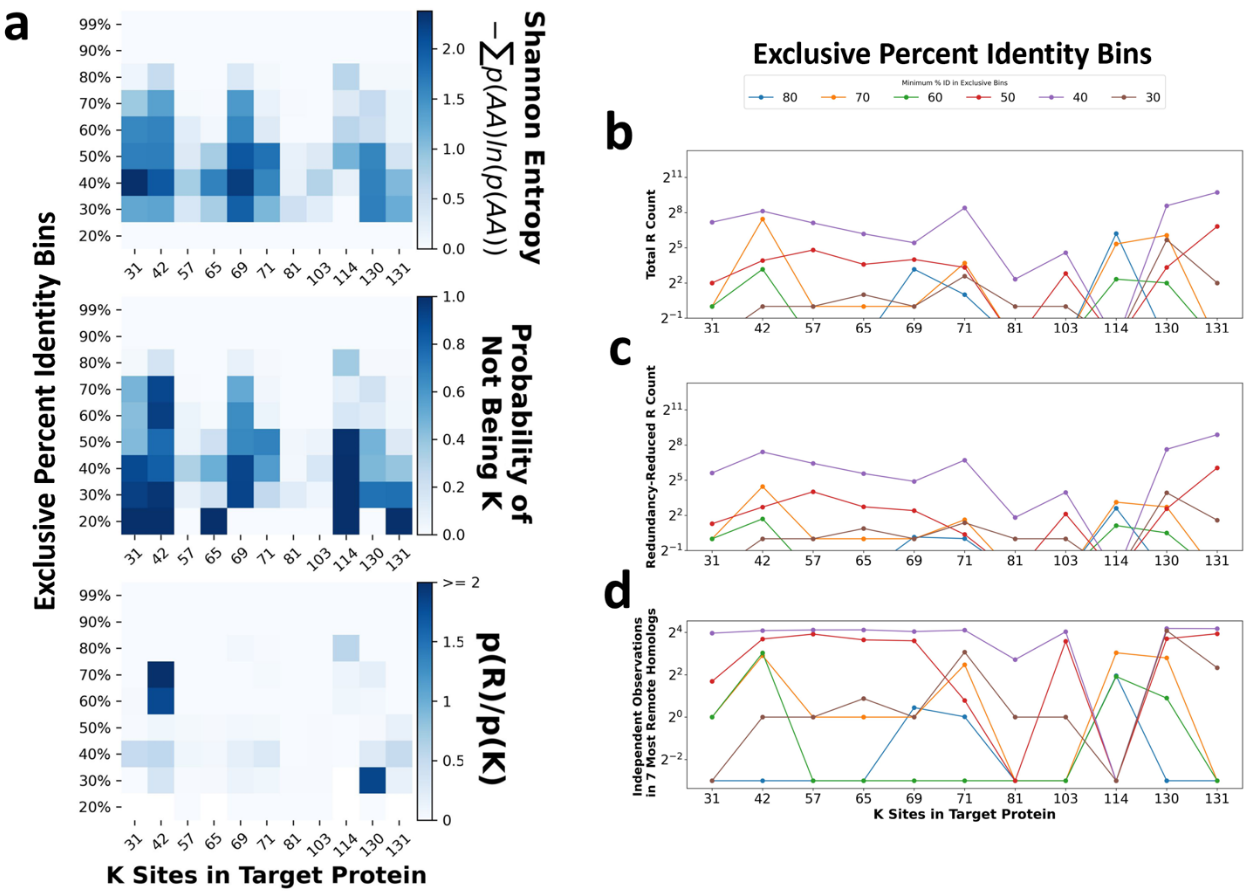

**Figure 2.** ***Representative output from the Bulk-R webserver analyzing lysine-to-arginine substitution patterns in homologous proteins.*** All sequence analyses are conducted on sets of proteins spanning 10% ranges (bins) of sequence identity relative to the target protein, with the exception of the highest identity bin which spans 90-99% identity to avoid mutant protein sequences. **(a)** The top graph shows the Shannon entropy, the middle graph shows the fraction of residues other than lysine, and the bottom graph shows the ratio of arginine/lysine residues at each lysine site in the native sequence in the indicated sequence identity bin. **(b,c)** The total (panel **b**) and heuristically redundancy-reduced (panel **c**) counts of arginine residues observed at each lysine site in the native sequence in the indicated sequence identity bin. **(d)** The expectation value for the number of independent observations of arginine in up to the seven most diverged sequence pairs in the bin that have arginine at the indicated site. The graph in panel **c** displays the results of the first redundancy-correction calculation described in the text, while the graph in panel **d** displays the results of the second.

# Figure 3

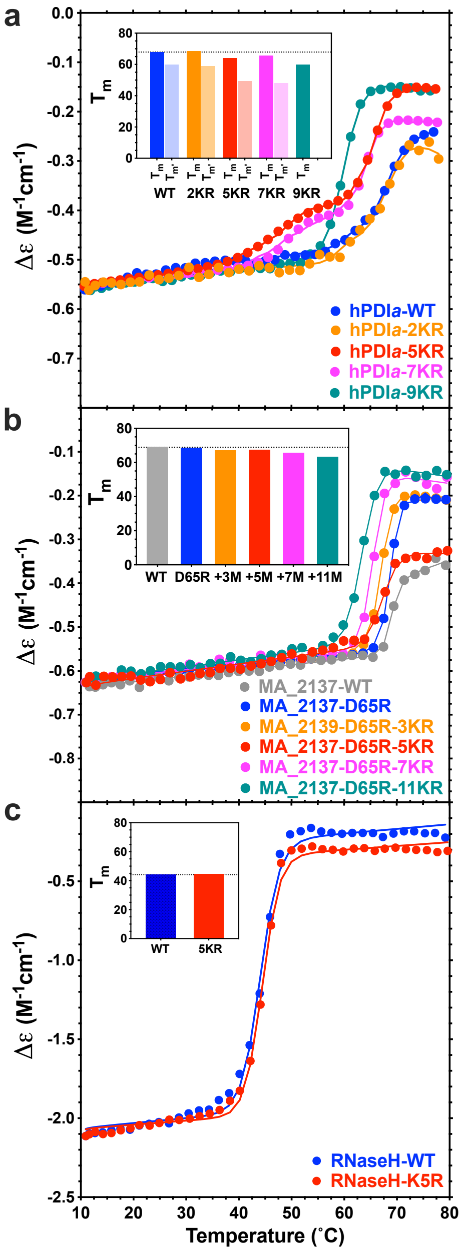

**Figure 3.** ***Thermodynamic stability of bulk-KR mutants characterized using thermal denaturation experiments monitored by circular dichroism (CD) spectroscopy.*** Experiments were conducted using protein samples at 2 mg/ml scanned at a rate of 3 ˚/min in a buffer containing 100 mM NaCl, 10 mM Tris-Cl, pH 7.5, with the addition of 1 mM SAH for the MA_2317 constructs. The suppression of the low-temperature transition in the hPDIa-9KR construct could be attributable to an intra-helical salt bridge between the sidechains of residues E62 and R65. The latter residue is one of the KR mutations in this construct not shared by the hPDIa-7KR construct, and the sidechain of residue K65 does not make any H-bonds at all in the crystal structure of the multidomain construct (PDB id 4EKZ). Therefore, the sidechain salt-bridge produced by the K65R mutation could potentially stabilize local structure in the hPDIa domain.

# Figure 4

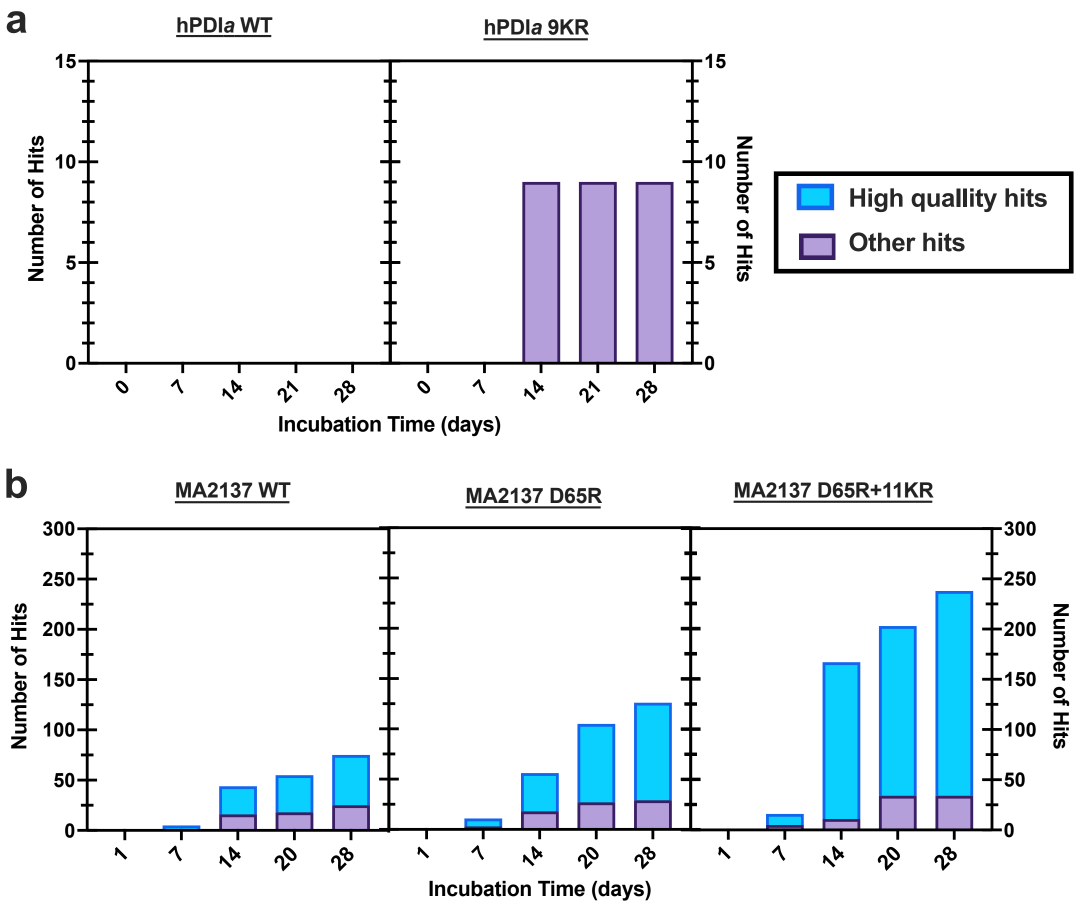

**Figure 4.** ***Crystallization hit counts from the 1536-condition high-throughput automated microbatch-under-oil crystallization screen at the National Crystallization Center at the Hauptman-Woodward Institute (HWI).*** These screens were conducted during the summer of 2021 using generation 19 of the HWI crystallization cocktail collection. The proteins were at ~15 ml/ml in the stock solutions used for screening, which contained 100 NaCl, 10 mM DTT, 10 mM Tris-Cl, pH 7.5. The screening plates are maintained at 4 ˚C for the first week and then moved to 25 ˚C for the remainder of the screening period [55-57, 83].

**Figure 5**

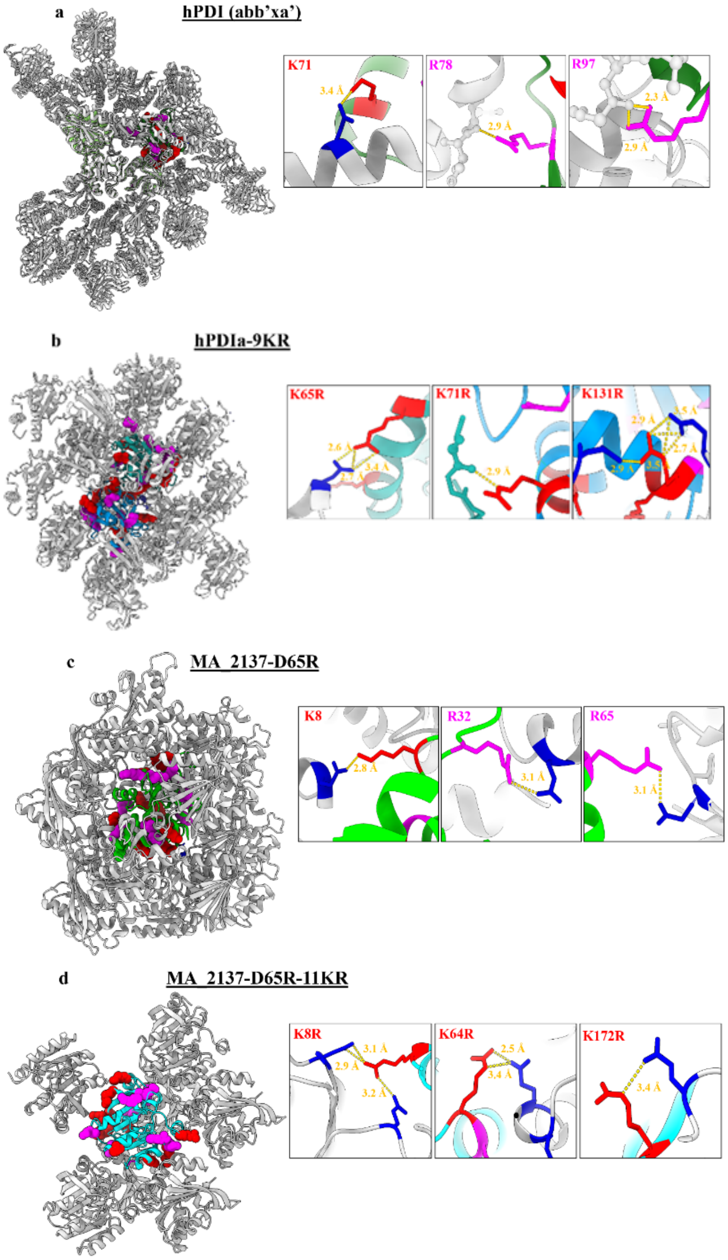

**Figure 5.** ***Crystal-packing in structures from proteins containing bulk-KR mutations.*** The protein backbone is shown in ribbon representation, colored gray for symmetry mates and shades of green, blue, cyan, and teal for the subunits modeled in the asymmetric unit of each crystal structure. The sidechains of the native arginines (magenta) and residues mutated from lysine-to-arginine (red) are shown in space-filling representation on the left and in stick representation in the zoomed-in views of local packing interactions in the boxes to the right. Asp, asn, glu, and gln residues making H-bonds to the illustrated lysine and arginine residues in the boxes are shown in blue stick representation, while the backbone atoms making H-bonds to those residues are shown in ball-and-stick representation.