

# **Towards predictive biophysical and mechanistic models for disease-causing protein variants**

Amelie Stein<sup>1,\*</sup>, Douglas M. Fowler<sup>2</sup>, Rasmus Hartmann-Petersen<sup>1</sup> and Kresten Lindorff-Larsen<sup>1,\*</sup>

<sup>1</sup>Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Denmark

<sup>2</sup>Departments of Genome Sciences and Bioengineering, University of Washington, U.S.A.

\*Corresponding authors: Amelie Stein (amelie.stein@bio.ku.dk) and Kresten Lindorff-Larsen (lindorff@bio.ku.dk)

# Abstract

The rapid decrease in DNA sequencing cost is revolutionizing medicine and science. In medicine, genome sequencing has revealed millions of missense variants that change protein sequences, yet we only understand the molecular and phenotypic consequences of a small fraction. Within protein science, high-throughput deep mutational scanning experiments enable us to probe thousands of mutations in a single, multiplexed experiment. We review efforts that bring together these topics via experimental and computational approaches to determine the consequences of missense mutations in proteins. We focus on the role of changes in protein stability as a driver for disease, and how experiments, biophysical models and computation are together providing a framework for understanding and predicting how mutations affect cellular protein stability.

# Keywords

Protein stability; Deep mutational scanning; Protein quality control; Variant classification; Computational biophysics

# Highlights

- Human exome sequencing is revealing millions of missense variants that change protein sequences, but their phenotypic consequences are mostly unknown
- Deep mutational scanning and other high-throughput experiments provide simultaneous insights into the effects of thousands of variants
- Loss of protein stability is a common origin of inherited diseases, and computational predictions of protein stability are useful for assessing variant consequences
- Cellular protein quality control provides a mechanistic link between altered protein stability and cellular protein levels and degradation
- Computational biophysics, evolutionary sequence analyses and machine learning methods each provide information about variant consequences and may potentially be combined
- Mechanistic models for how mutations give rise to disease provide a starting point for therapeutic strategies

## 38 Introduction

39 Technological advances in DNA sequencing have made human genome sequencing on large scales not  
40 only feasible, but also affordable. The resulting data avalanche has highlighted the challenge of  
41 interpreting the phenotypic consequences of genetic variants [1,2]. Variant interpretation is particularly  
42 challenging since more than half of the distinct variants found in an analysis of >60,000 human exomes  
43 were only observed in a single individual [3] and since many diseases have a complex, polygenic origin  
44 [4]. Although the problem is difficult and complicated, the potential to improve the understanding,  
45 diagnosis and treatment of human diseases is enormous.

46 In this review, we focus on missense variants that result in a change from one amino acid to another  
47 (henceforth called *variants*). Further, we focus on recent efforts to understand and predict the effects  
48 these variants have on biophysical properties of proteins, and, consequently, their effect on function.  
49 While protein-coding regions only make up ~1.5% of the genome, around 5-10% of hits in genome-  
50 wide association studies fall into them, although linkage disequilibrium (joint inheritance of elements  
51 proximal on a chromosome) makes it challenging to identify precisely which of multiple nearby  
52 variants is causal [5]. Beyond diagnosis, we may use existing knowledge of proteins and their cellular  
53 pathways to help elucidate the disease-causing mechanisms. Because proteins can be targeted by small  
54 molecules or peptides, these insights can potentially open up therapeutic avenues.

## 55 Interpreting missense variation

56 Missense variants represent over 40% of the unique variants observed in the Exome Aggregation  
57 Consortium database [3], yet their phenotypic consequences are often difficult to predict. This is in  
58 contrast to nonsense or frameshift variants that cause large changes to the encoded protein and  
59 consequently are usually deleterious. As an example, systematic mutagenesis studies of the highly  
60 conserved protein ubiquitin have shown that many single missense mutations only have a minor impact  
61 on protein function in a cellular assay [6]. An analysis of similar high-throughput data across multiple  
62 proteins suggest that indeed about two thirds of single amino acid changes have only a minor effect on  
63 function [7]. Some variants are, however, severely detrimental and cause essentially complete loss of  
64 function. An interesting observation from further studies on ubiquitin is that, at least for this protein,  
65 there can be substantial variation of the effect of a mutation depending on the cellular status and  
66 conditions, so that most variants are detrimental under at least one condition [8].

67 In a clinical setting it would be useful to have robust methods and sufficient data for interpretation of  
68 genetic variants and accurate classifications of whether they are pathogenic or benign [9]. This is  
69 particularly important for diseases where such information can lead to clinical action [10]. To further  
70 our understanding of the origins of disease it would also be extremely valuable to have reliable  
71 predictors of the underlying mechanisms by which variants lead to disease.

72 There are several conceptual frameworks available to study, model and predict the phenotypic  
73 consequences and pathogenicity of mutations. For example, one may use cellular or biochemical assays  
74 to quantify the effects of the mutations on function and other properties, and recent developments are  
75 enabling such studies in high-throughput and with full coverage [11]. Another framework is to use  
76 bioinformatics and machine learning methods to integrate existing data, in particular information about  
77 sequence conservation, to interpret what sequence variation is compatible with function [12]. Finally,

one may use the accumulated knowledge about protein structure, function and folding to determine the likely effect of a variant [13]. These different approaches are not mutually exclusive and ongoing efforts indeed aim to combine them.

## Loss of protein stability as origin of disease

Protein stability is one of the most basic properties of a protein, and may be strongly affected by a missense mutation. As most proteins need to be folded to function, loss of stability may lead to loss of function. In the context of a biophysical or biochemical experiment stability generally refers to the thermodynamic or kinetic stability between a fully folded and globally unfolded state, but in a cellular and disease context many other factors and protein conformations play a role. These factors include interactions with the cellular protein quality control system, protein-protein interactions, cellular trafficking and post-translational modifications. Analyses linking the effect of a mutation on the thermodynamic stability of a protein with its cellular stability and pathogenicity suggest that loss of stability could be a main driver and origin of inherited diseases [14-18]. Thus, an improved understanding of the complex relationship between protein sequence, structure, folding and cellular stability could provide new possibilities for diagnosis and even treatment.

Experimental studies of protein folding and stability *in vitro* and *in vivo* may provide detailed, quantitative descriptions and mechanistic insights of the effects of mutations. Until recently, however, they were limited to studying the effects of a few mutations, generally limiting studies to retrospective analyses of variants already seen in patients. Recent developments in high-throughput experiments are, however, beginning to provide us with orders-of-magnitude more data to improve our models and understanding of protein stability, and to perform prospective studies of variants not yet seen in patients [19]. By leveraging the same advances in DNA sequencing that are enabling cheap sequencing of human genomes, deep mutational scanning (DMS) experiments are making it possible to study the effects of mutations on a scale not previously possible [20]. Combined with genetic selection systems, DNA sequencing methods can also be used to study the mechanisms and sequence specificity of cellular protein quality control [21].

Together, these developments are now being put to use to improve the predictions of clinical outcomes and to provide mechanistic models for diseases. Below we review recent developments in these areas, focusing on the role that loss of protein stability and resulting loss of function plays in human diseases. We begin with an overview of the cellular protein quality control system which recognizes unstable or misfolded proteins and target them for degradation, and thus is the mechanistic link between loss of stability and decreased cellular abundance of proteins. We proceed to describe how DMS experiments are transforming our ability to study functional and mechanistic consequences of mutations. We then describe recent developments in using computational methods to predict the consequences of mutations, and end by describing how insights into the mechanisms underlying loss of cellular protein stability may be used to develop new therapies.

## Cellular protein quality control

Since structurally destabilized or misfolded proteins may form various toxic inclusions or aggregates, all organisms have evolved a number of protective measures to guard against these potentially harmful

proteins. Collectively these mechanisms are known as protein quality control (PQC) systems, with the two main strategies being either refolding or degradation of the misfolded proteins [22,23].

During or after synthesis proteins may undergo transitions through various metastable folding intermediates towards the native state and be protected from aggregation by molecular chaperones; in a similar manner chaperones may also catalyse the refolding of proteins that become damaged after synthesis [22]. Degradative PQC, on the other hand, relies on proteases to irreversibly clear the intracellular environment of non-native proteins. Both of these PQC systems must be highly specific for incorrectly folded proteins, but also be broadly inclusive to ensure that many structurally diverse proteins can be targeted. Accordingly, defects in either of these systems can lead to accumulation of toxic protein species which in turn may trigger diseases, including several neurodegenerative disorders [24,25]. Conversely, an overaggressive destruction of structurally destabilized, but functional, proteins has been linked to various hereditary diseases, including cystic fibrosis [26,27] and Lynch syndrome [17,28,29]. It therefore becomes clear that substrate selection is a trade-off between specificity and recognition of a wide variety of substrates.

In eukaryotes, most protein degradation occurs in the cytosol and nucleus via the ubiquitin-proteasome system (UPS) or the autophagy-lysosomal pathway [30], with the latter system typically responsible for the degradation of highly misfolded and insoluble protein aggregates. Aggregation has also been linked to a number of diseases; however this is beyond the scope of this review and we refer the reader to a recent review [31]. The UPS generally targets soluble or partially soluble proteins through a process involving conjugation of a polyubiquitin chain to the substrate protein, thus targeting it to degradation by the 26S proteasome. Ubiquitin conjugation is catalysed by an enzymatic cascade that includes substrate specific E3 ubiquitin-protein ligases that add the ubiquitin chains to the target protein. The discriminating feature in a destabilized protein that elicits its recognition by E3s and degradation, the so-called degron, is despite tremendous recent efforts [21,32,33] not completely understood, but it is likely to involve hydrophobic regions that are buried in the native protein, but exposed in misfolded proteins (Fig. 1). We refer the reader to recent reviews of the role and components of the PQC that are important to the degradation of misfolded proteins [34,35].

In the context of disease-causing mutations, a key question is how much structural destabilization is tolerated before the PQC system kicks in? Recently, it was shown that the degree of protein destabilization correlates with the turnover rate in the Lynch-syndrome related protein MSH2 [17]. Surprisingly, however, as little as 3 kcal mol<sup>-1</sup> was sufficient to trigger degradation [17]. Although this figure is likely to vary from protein to protein, depending on how stable the wild type protein is, a 3 kcal mol<sup>-1</sup> destabilization is certainly not dramatic, compared to, for example, the average stability of 5 kcal mol<sup>-1</sup> for a series of small proteins [36]. It is, however, in agreement with genetic studies in yeast that have shown that the PQC system operates by following a better-safe-than-sorry principle and is thus highly diligent and prone to target proteins that are only slightly perturbed and still functional [29,37,38].

A key problem to tackle in the future is to understand better what structural features are actually recognized by the PQC system. For example, it is unclear whether cells generally recognize global or local unfolding events, and what the relationship is between such unfolding events and transient exposure of degron sequences (Fig. 1). In this context, a mutation causing a destabilization of a few

kcal mol<sup>-1</sup> could cause substantial increase in the population of locally unfolded structures, which in turn would lead to degradation and insufficient levels of the affected protein.

## Deep Mutational Scanning

Much of what we know about how proteins fold and are stabilized has been learned by studying individual amino acid changes. However, this one-at-a-time approach probes only a tiny fraction of the possible genetic variation we could observe in an individual, and hence limits our understanding and ability to predict phenotypic consequences. DMS experiments leverage cheap DNA sequencing to probe the effects of hundreds or thousands of variants in a single, multiplexed assay [20,39]. First, selection for a protein property of interest is applied to a large library of variants. Selections used so far include coupling protein activity to cell growth, coupling protein activity or stability to a fluorescent reporter, or selecting for ligand binding using phage or yeast display. Variants in the library change in frequency depending on how well they are able to perform under selective conditions. Finally, the frequency of each variant before and after the selection is read out using next-generation DNA sequencing and each variant's change in frequency is used to compute a functional score.

Most applications of DMS have employed selection for a biological function of the protein that can be probed in high throughput. For example, in a recent tour de force, the effect of variants of the *BRCA1* gene were assayed using saturation genome editing. Here, approximately 4,000 variants were introduced into 13 of *BRCA1*'s 24 exons using CRISPR/Cas9 editing of the genomic copy of *BRCA1* in a haploid cell line. The functional consequences of each variant on cell viability was measured using next-generation sequencing, and correlated strongly with existing expert-based assessment of pathogenicity. Variants that are common in the human population were more likely to be scored as functional in the assay. Importantly, this experiment also provided functional data for the several thousand variants that have not yet been seen in any patient. These unseen variants are of unknown pathogenicity, so the functional data will be of immediate use if any of them are seen in the future. An interesting observation was also that ~90% of all loss-of-function variants had no substantial changes in mRNA levels, suggesting that most missense variants—at least in *BRCA1*—affect function at the protein level. As observed from the results on ubiquitin discussed in the introduction, as well as a dual-assay DMS study of *BRCA1* [40], different assays and conditions might reveal different mutational sensitivities.

The results of growth-based saturation genome editing experiments like those described for *BRCA1* above depend on the combined effects that a mutation may have on numerous properties including RNA splicing, expression levels, protein function, protein-protein interaction, post-translational modifications and protein folding and stability. Because the cellular growth rate may capture many of the biologically-relevant effects of variants it can be extremely accurate and useful for assessing the pathogenicity. On the other hand, the results may be less informative for disentangling the mechanism by which each variant exerts an effect, and the knowledge obtained is not easily transferable to studying the effects of variants in other proteins.

To enable more widespread analysis of variant consequences without needing to establish protein-specific assays, and to learn more general rules regarding the relationship between protein stability and cellular abundance, we have recently developed Variant Abundance by Massively Parallel sequencing (VAMP-seq, Fig. 2). VAMP-seq measures the impact of variants on the steady-state cellular abundance

of a protein [41]. Here, a library of variants of the protein of interest is fused to GFP (Fig. 2a). Then, the library is expressed in cultured mammalian cells such that each cell expresses one and only one variant (Fig. 2b). The stability of the variant dictates the stability of the GFP fusion, so each cell's GFP fluorescence reports on the abundance of the protein variant. Cells are sorted into bins based on their fluorescence, next-generation sequencing is used to determine the frequency of every variant in each bin, and variant frequencies are used to compute abundance scores (Fig. 2c). Thus, a single VAMP-seq experiment provides quantitative abundance data for thousands of variants simultaneously and enables one to separate mutations with modest effects on stability from those that are substantially destabilizing (Fig. 2d).

In the context of enabling computational prediction methods, it is worth highlighting that a single VAMP-seq experiment provides information about a number of variants comparable in size to the entire database used to train current state-of-the-art models for predicting protein stability [42,43] (Fig. 2e). Another advantage is that DMS experiments generally target most or all of the 19 possible amino acid substitutions at each position. This comprehensive data is useful in the clinic because it can be used to aid the interpretation of any variant. Moreover, unlike the majority of available biophysical data that is highly biased [44] and mostly consist of side chain truncations to alanine or glycine (Fig. 2e), comprehensive functional and stability data can both be used to provide insight into a specific protein and can also be used to guide the development of improved pathogenicity prediction methods. DMS is already a widely-applied method, and will become even more useful as methods for generating and sequencing variant libraries improve and decrease in cost. We also note that DMS and related high-throughput experiments may provide very useful information for understanding and improving protein function and stability for example in protein engineering and design [45,46].

## Predicting the consequences of missense variation

While experimental testing of variants is expanding in scope and scale, computational predictions of variant consequences will continue to be the only widely applicable method to assess pathogenicity for the foreseeable future. A number of predictors have been trained specifically for this purpose, often using known benign and pathogenic variants [47]. Here, we instead focus on three distinct approaches developed to address more general questions concerning how changes in the protein sequence affect, for example, protein stability or general functional properties. These methods have not been specifically trained on pathogenic variants; instead, they were created to capture thermostability of folding, evolutionary tolerance, and patterns observed in DMS experiments, respectively. To illustrate the outcome and performance of these three classes of prediction methods, we show the results of stability calculations (Fig. 3a), a sequence likelihood model (Fig. 3b) and the DMS-based prediction method (Fig. 3c) on the protein MSH2, and discuss them in more detail below.

Modelling amino acid substitution(s) directly in a protein's 3-dimensional structure should, in principle, enable an accurate assessment of the resulting change in folding energy. Two tools that take this approach are FoldX [43] and Rosetta [48], which each predict mutational effects on stability with an accuracy of about 1 kcal mol<sup>-1</sup> and a correlation coefficient of ~0.7 (depending on test set [42]). In addition to predicting stability effects, these and related methods have been shown to successfully identify pathogenic variants in several proteins [14,17]. In selected cases, experimental validation yielded a correlation between the predicted loss of stability and cellular protein levels [17,41,49]. In

addition to classifying unstable variants as pathogenic, stability predictions have the additional advantage of indicating the likely underlying mechanism; this information is useful when developing therapeutic strategies (see below).

Prediction methods that focus on a specific mechanism such as loss of stability will, of course, not capture variants that give rise to disease via different mechanisms. Thus, stability predictions are most useful when combined with other predictors [47] [50-52]. Analysis of the conservation patterns in a multiple sequence alignment of a protein family is a powerful and general approach to identify substitutions that are pathogenic by their paucity in, or absence from, the alignment, and indeed is used in most prediction methods [47]. A recent development is the construction of higher-order statistical models that examine both conservation at individual sites and also between multiple sites [53-56]. While these latter approaches generally provide greater accuracy than methods that analyse each site independently [57], they require a larger number of homologous sequences. This restriction arises because the methods involve building global sequence models rather than examining each site independently. Analyses that consider both site conservation and pairwise co-varying positions have successfully been applied to predict variant pathogenicity [57,58], and more recently, more general models have been introduced [12].

Because evolutionary conservation across a protein family is likely to capture residues required for the protein's core function, these approaches can identify variants that affect many protein properties including stability, enzymatic activity, post-translational modifications or protein-protein interactions. Thus, a conserved variant may be neutral from the perspective of thermodynamic folding energy but have strong functional consequences. On the other hand, evolutionary sequence analysis may miss pathogenic changes where the residue in question is critical only for human biology, or in a small branch of the protein family's phylogenetic tree. In this context, recent analyses focusing on mutational tolerance in non-human primates are particularly interesting [59].

As an alternative to analyses of conservation through deep multiple sequence alignments, one may use other sources of data to learn what kind of amino acid changes typically lead to perturbed function. Here, DMS experiments now provide us with a large collection the functional effects of tens of thousands of substitutions across a diverse set of proteins [7]. Annotation of this functional data with biochemical and coarse-grained structural features was combined with machine learning to create Envision, a tool for quantitative prediction of the effect of missense variants [60]. In contrast to the biophysical modelling and sequence conservation analysis approaches discussed above, Envision does not require specific data on the protein in question beyond its sequence, and is thus more widely applicable than stability calculations and statistical sequence analysis, yet it successfully identified many pathogenic variants in a recent benchmark [60].

As an example of the power of using these three prediction paradigms, we show their application to the protein MSH2, where mutations may lead to cancer predisposition (Lynch syndrome) (Fig. 3). Specifically, as previously described [17], we used FOLDX [43] to calculate changes in protein stability from the structure of MSH2 and Gremlin [54] to analyse a multiple sequence alignment of MSH2. Finally, we used a Envision [60], the abovementioned machine learning method trained on DMS data, structure and sequence features, to predict the consequences of mutations. In contrast to our previous work that focus on a smaller set of mutations, we here used ClinVar [61] to select 21

pathogenic and 66 benign variants, and also analysed the 587 missense variants of MSH2 found in gnomAD [3].

The results show clearly that, although these methods have not been trained on population genetics data or disease mutations, they are able to separate known disease-causing variants from benign variants with relatively high accuracy. For example, benign variants generally have modest effects on stability, whereas many pathogenic variants are highly destabilizing. It is also worth noting that only three of the XX pathogenic variants seen in ClinVar have actually been observed in the ~150,000 genome and exome sequences available in gnomAD. Thus, there is a clear trend that more common population variants are predicted to have milder effects, whereas many uncommon variants and pathogenic variants are predicted to have more dramatic effects (Fig. 3A). These observations imply that there is a clear difference in the distribution of predicted scores between benign and pathogenic variants (Fig. 3B) which in turn can be transformed into relatively accurate predictions (Fig. 3C). Nonetheless, the analyses also show that these predictions of functional effects are not yet alone sufficient to fully separate benign from pathogenic variation.

## Therapeutic possibilities

In addition to the prospect for improved diagnosis via prediction of pathogenicity, the experimental and computational studies discussed above provide new opportunities for treatment of diseases. For mutations that gives rise to disease via loss of stability, intracellular degradation and thereby loss of function, it might be possible to rescue function via restabilization. In particular, because the PQC is overzealous in targeting potentially functional, but mildly destabilized proteins, many disease-causing variants might be sufficiently functional that pathogenicity could potentially be averted if the proteins were stabilized [29] (Fig. 4).

The most dramatic approach is perhaps to inhibit the proteasome, and proteasome inhibitors are indeed already approved drugs [62]. In many cases, a more direct and elegant approach might be to target the components in the PQC that are relevant for degrading a specific disease-causing variant. To enable this approach, we need to map in much greater detail the E3 enzymes and chaperones involved in recognizing specific substrates and targeting them for degradation. As an example, in yeast, certain mutant variants of MSH2 linked to Lynch syndrome can be rescued by deleting the E3 ligase that targets the MSH2 variants for degradation, thus restoring cellular MSH2 protein levels and MSH2 function [28]. Thus, targeting the equivalent, but still unknown [63], human E3 ligase may provide treatment options for individuals with certain MSH2 variants. Since a number of the PQC E3s display overlapping substrate specificity [64], this will likely be complicated. Other strategies involve increasing or decreasing the levels of chaperones that either aid in refolding or degradation [65,66].

Some protein variants might be so unstable that even inhibiting their degradation would not be sufficient to restore cellular stability and function. These variants might, however, be rescued via small molecules that bind directly to the destabilized variant protein [67]. This chemical chaperones or corrector approach has already been shown to rescue function for example in mutant p53 [68] and CFTR [69].

## 319 Outlook

320 Widespread access to cheap DNA sequencing is transforming medicine and science. Within precision  
321 medicine, genome or exome sequencing provides possibilities for finding causal variants and for  
322 improved diagnosis and possible treatment. Within protein science, DMS experiments are enabling the  
323 study of the effects of thousands of mutations in a single experiment. Recent efforts are bringing these  
324 fields together by using DMS to help classify variants as benign or pathogenic, and by providing data  
325 to benchmark or train prediction methods for variant classification. These approaches may be  
326 particularly important for so-called rare genetic disease that are difficult to diagnose from population-  
327 based studies [70].

328 So far, these approaches have mostly been applied to simple, monogenic Mendelian disorders. In the  
329 future it will be interesting to investigate whether they can improve polygenic risk scores that aggregate  
330 information across variants in multiple genes. Here it is worth noting how stability predictions for  
331 protein-protein complexes provide a direct mechanism for finding apparently non-additive effects. For  
332 example, two variants that individually only cause a mild change in the stability of the complex may,  
333 when combined, have a dramatic effect because of the non-linear relationship between energy and  
334 population of the complex.

335 One of the problems in assessing the importance of loss of stability for disease is that we do not fully  
336 understand when and why the current prediction methods fail. This is in part due to the fact that they  
337 were trained and benchmarked on a biased dataset that mostly focuses on mutations where a large  
338 amino acid is mutated to a smaller one, often alanine or glycine. We expect that unbiased functional  
339 data from DMS experiments will be extremely useful in assessing and parameterizing prediction  
340 methods for a much wider set of amino acid changes. An important problem to tackle in the future is to  
341 map genetic variants on to accurate structural models for the entire human proteome [71], and to  
342 develop prediction methods that are robust towards structural noise in homology models. Finally, an  
343 important open question is how the different prediction methods are best combined, and how they can  
344 both provide accurate predictions of pathogenicity and aid in developing mechanistic hypotheses for  
345 the origin of disease.

## 346 Outstanding Questions

- 347 • What are the structural features of the unfolded and misfolded states, and how are they  
348 recognized by the PQC system?
- 349 • Are there generic PQC components including chaperones and E3s that target a wide range of  
350 human missense variants?
- 351 • When current predictors fail, why is that? Can we develop confidence scores to identify less  
352 reliable predictions?
- 353 • Can biophysics, statistical sequence analysis and machine learning on DMS data improve  
354 polygenic risk scores?
- 355 • How are predictors best combined, both to improve accuracy and to develop mechanistic  
356 hypotheses for the origin of genetic diseases?
- 357 • Can we develop therapeutic strategies to target many different variants in a single protein, or  
358 variants in different proteins that are degraded by similar pathways?

## 359 **Acknowledgements**

360 We thank Drs. Sofie V. Nielsen and Caspar E. Christensen for helpful comments on the manuscript and  
361 assistance with preparing the figures.

## 362 **Funding**

363 Our work in this area has been supported by grants from the Lundbeck Foundation (A.S., R.H.-P. and  
364 K.L.-L.), The Danish Cancer Society (R.H.-P.), the Novo Nordisk Foundation (R.H.-P. and K.L.-L.),  
365 the Danish Council for Independent Research (Natural Sciences) (to R.H.-P), the National Institute of  
366 General Medical Sciences (1R01GM109110 to D.M.F.). D.M.F. is a CIFAR Azrieli Global Scholar.

# References

- 1 Shendure, J. and Akey, J.M. (2015) The origins, determinants, and consequences of human mutations. *Science* 349, 1478–1483
- 2 Manolio, T.A. *et al.* (2017) Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell* 169, 6–12
- 3 Lek, M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291
- 4 Martin, H.C. *et al.* (2018) Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 42, eaar6731
- 5 Gusev, A. *et al.* (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95, 535–552
- 6 Roscoe, B.P. *et al.* (2013) Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *J Mol Biol* 425, 1363–1377
- 7 Gray, V.E. *et al.* (2017) Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* 207, 53–61
- 8 Mavor, D. *et al.* (2018) Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biology Open* 7, bio036103–8
- 9 Richards, S. *et al.* (2015), Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology., presented at the Genetics in medicine conference, 17, pp. 405–424
- 10 MacArthur, D.G. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–476
- 11 Findlay, G.M. *et al.* (2018) Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 372, 2235
- 12 Riesselman, A.J. *et al.* (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* DOI: 10.1038/s41592-018-0138-4
- 13 Kroncke, B.M. *et al.* (2016) Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. *Biochemistry* 55, 5002–5009
- 14 Pey, A.L. *et al.* (2007) Predicted Effects of Missense Mutations on Native-State Stability Account for Phenotypic Outcome in Phenylketonuria, a Paradigm of Misfolding Diseases. *The American Journal of Human Genetics* 81, 1006–1024
- 15 Casadio, R. *et al.* (2011) Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Human Mutation* 32, 1161–1170
- 16 Pal, L.R. and Moul, J. (2015) Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. *J Mol Biol* 427, 2271–2289
- 17 Nielsen, S.V. *et al.* (2017) Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* 13, e1006739
- 18 Stein, A. *et al.* (2018) Loss of Protein Stability is a Strong Indicator of Pathogenicity. *in prep*
- 19 Starita, L.M. *et al.* (2017) Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* 101, 315–325

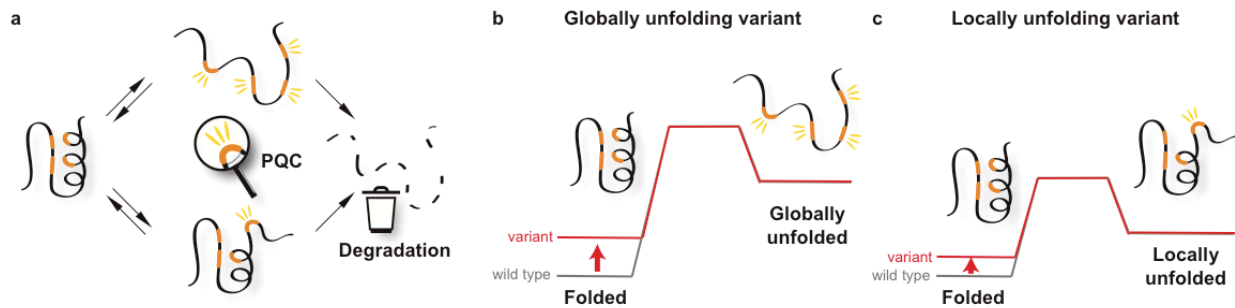
- 409 20 Fowler, D.M. and Fields, S. (2014) Deep mutational scanning: a new style of protein science. *Nat*  
410 *Methods* 11, 801–807
- 411 21 Geffen, Y. *et al.* (2016) Mapping the Landscape of a Eukaryotic Degronome. *Mol Cell* 63, 1055–  
412 1065
- 413 22 Hartl, F.U. *et al.* (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475,  
414 324–332
- 415 23 Kettern, N. *et al.* (2010) Chaperone-assisted degradation: multiple paths to destruction. *Biol Chem*  
416 391, 481–489
- 417 24 Ciechanover, A. and Kwon, Y.T. (2017) Protein Quality Control by Molecular Chaperones in  
418 Neurodegeneration. *Front Neurosci* 11, 185
- 419 25 Rubinsztein, D.C. (2006) The roles of intracellular protein-degradation pathways in  
420 neurodegeneration. *Nature* 443, 780–786
- 421 26 Ahner, A. *et al.* (2007) Small heat-shock proteins select deltaF508-CFTR for endoplasmic  
422 reticulum-associated degradation. *Mol Biol Cell* 18, 806–814
- 423 27 Meacham, G.C. *et al.* (2001) The Hsc70 co-chaperone CHIP targets immature CFTR for  
424 proteasomal degradation. *Nat Cell Biol* 3, 100–105
- 425 28 Arlow, T. *et al.* (2013) Proteasome inhibition rescues clinically significant unstable variants of the  
426 mismatch repair protein Msh2. *Proc Natl Acad Sci U S A* 110, 246–251
- 427 29 Kampmeyer, C. *et al.* (2017) Blocking protein quality control to counter hereditary cancers. *Genes*  
428 *Chromosomes Cancer* 56, 823–831
- 429 30 Kwon, Y.T. and Ciechanover, A. (2017) The Ubiquitin Code in the Ubiquitin-Proteasome System  
430 and Autophagy. *Trends in Biochemical Sciences* 42, 873–886
- 431 31 Chiti, F. and Dobson, C.M. (2017) Protein Misfolding, Amyloid Formation, and Human Disease:  
432 A Summary of Progress Over the Last Decade. *Annu Rev Biochem* 86, 27–68
- 433 32 Maurer, M.J. *et al.* (2016) Degradation Signals for Ubiquitin-Proteasome Dependent Cytosolic  
434 Protein Quality Control (CytoQC) in Yeast. *G3 (Bethesda)* 6, 1853–1866
- 435 33 Rosenbaum, J.C. *et al.* (2011) Disorder targets disorder in nuclear quality control degradation: a  
436 disordered ubiquitin ligase directly recognizes its misfolded substrates. *Mol Cell* 41, 93–106
- 437 34 Clausen, R. *et al.* (2015) Mapping the Conformation Space of Wildtype and Mutant H-Ras with a  
438 Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Comput Biol* 11, e1004470–26
- 439 35 Enam, C. *et al.* (2018) Protein Quality Control Degradation in the Nucleus. *Annu Rev Biochem* 87,  
440 725–749
- 441 36 Maxwell, K.L. *et al.* (2005) Protein folding: defining a “standard” set of experimental conditions  
442 and a preliminary kinetic data set of two-state proteins. *Protein Sci* 14, 602–616
- 443 37 Gardner, R.G. *et al.* (2005) Degradation-mediated protein quality control in the nucleus. *Cell* 120,  
444 803–815
- 445 38 Kriegenburg, F. *et al.* (2014) A chaperone-assisted degradation pathway targets kinetochore  
446 proteins to ensure genome stability. *PLoS Genet* 10, e1004140
- 447 39 Fowler, D.M. *et al.* (2014) Measuring the activity of protein variants on a large scale using deep  
448 mutational scanning. *Nature Protocols* 9, 2267–2284
- 449 40 Starita, L.M. *et al.* (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain  
450 Variants. *Genetics* 200, 413–422
- 451 41 Matreyek, K.A. *et al.* (2018) Multiplex assessment of protein variant abundance by massively  
452 parallel sequencing. *Nature Genetics* 50, 874–882

- 453 42 Ó Conchúir, S. *et al.* (2015) A Web Resource for Standardized Benchmark Datasets, Metrics, and  
454 Rosetta Protocols for Macromolecular Modeling and Design. *PLoS ONE* 10, e0130433–18
- 455 43 Guerois, R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a  
456 study of more than 1000 mutations. *J Mol Biol* 320, 369–387
- 457 44 Yang, Y. *et al.* (2018) PON-tstab: Protein Variant Stability Predictor. Importance of Training Data  
458 Quality. *IJMS* 19, 1009
- 459 45 Wrenbeck, E.E. *et al.* (2016) Deep sequencing methods for protein engineering and design. *Curr*  
460 *Opin Struct Biol* 45, 36–44
- 461 46 Gupta, K. and Varadarajan, R. (2018) Insights into protein structure, stability and function from  
462 saturation mutagenesis. *Curr Opin Struct Biol* 50, 117–125
- 463 47 Niroula, A. and Vihinen, M. (2016) Variation Interpretation Predictors: Principles, Types,  
464 Performance, and Choice. *Human Mutation* 37, 579–597
- 465 48 Park, H. *et al.* (2016) Simultaneous Optimization of Biomolecular Energy Functions on Features  
466 from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* DOI:  
467 10.1021/acs.jctc.6b00819
- 468 49 Bershtein, S. *et al.* (2013) Protein Quality Control Acts on Folding Intermediates to Shape the  
469 Effects of Mutations on Organismal Fitness. *Mol Cell* 49, 133–144
- 470 50 De Baets, G. *et al.* (2012) SNPeff 4.0: on-line prediction of molecular and structural effects of  
471 protein-coding variants. *Nucleic Acids Res* 40, D935–9
- 472 51 Raimondi, D. *et al.* (2016) Multilevel biological characterization of exomic variants at the protein  
473 level significantly improves the identification of their deleterious effects. *Bioinformatics* 32,  
474 1797–1804
- 475 52 Wagih, O. *et al.* (2018) Comprehensive variant effect predictions of single nucleotide variants in  
476 model organisms. DOI: 10.1101/313031
- 477 53 Weigt, M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by  
478 message passing. *Proc Natl Acad Sci U S A* 106, 67–72
- 479 54 Balakrishnan, S. *et al.* (2011) Learning generative models for protein fold families. *Proteins* 79,  
480 1061–1078
- 481 55 Lapedes, A. *et al.* Using Sequence Alignments to Predict Protein Structure and Stability With  
482 High Accuracy. *arXiv*. 12-Jul-(2012), 1–29. URL: <https://arxiv.org/pdf/1207.2484v1.pdf>
- 483 56 Marks, D.S. *et al.* (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation.  
484 *PLoS ONE* 6, e28766–20
- 485 57 Feinauer, C. and Weigt, M. (2017) Context-Aware Prediction of Pathogenicity of Missense  
486 Mutations Involved in Human Disease. *bioRxiv* DOI: 10.1101/103051
- 487 58 Hopf, T.A. *et al.* (2017) Mutation effects predicted from sequence co-variation. *Nature Publishing*  
488 *Group* 35, 128–135
- 489 59 Sundaram, L. *et al.* (2018) Predicting the clinical impact of human mutation with deep neural  
490 networks. *Nature Genetics* 50, 469
- 491 60 Gray, V.E. *et al.* (2017) Quantitative Missense Variant Effect Prediction Using Large-Scale  
492 Mutagenesis Data. *Cell Systems* DOI: 10.1016/j.cels.2017.11.003
- 493 61 Landrum, M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting  
494 evidence. *Nucleic Acids Res* 46, D1062–D1067
- 495 62 Beck, P. *et al.* (2012) Covalent and non-covalent reversible proteasome inhibition. *Biol Chem* 393,  
496 1101–1120

- 497 63 Boomsma, W. *et al.* (2016) Bioinformatics analysis identifies several intrinsically disordered  
498 human E3 ubiquitin-protein ligases. *PeerJ* 4, e1725–18
- 499 64 Samant, R.S. *et al.* (2018) Distinct proteostasis circuits cooperate in nuclear and cytoplasmic  
500 protein quality control. *Nature* 86, 27
- 501 65 Kirkegaard, T. *et al.* (2010) Hsp70 stabilizes lysosomes and reverts Niemann-Pick disease-  
502 associated lysosomal pathology. *Nature* 463, 549–553
- 503 66 Kirkegaard, T. *et al.* (2016) Heat shock protein-based therapy as a potential candidate for treating  
504 the sphingolipidoses. *Sci Transl Med* 8, 355ra118–355ra118
- 505 67 Pereira, D.M. *et al.* (2018) Tuning protein folding in lysosomal storage diseases: the chemistry  
506 behind pharmacological chaperones. *Chem Sci* 9, 1740–1752
- 507 68 Joerger, A.C. and Fersht, A.R. (2016) The p53 Pathway: Origins, Inactivation in Cancer, and  
508 Emerging Therapeutic Approaches. *Annu Rev Biochem* 85, 375–404
- 509 69 Van Goor, F. *et al.* (2011) Correction of the F508del-CFTR protein processing defect in vitro by  
510 the investigational drug VX-809. *Proc Natl Acad Sci U S A* 108, 18843–18848
- 511 70 Wright, C.F. *et al.* (2018) Assessing the pathogenicity, penetrance and expressivity of putative  
512 disease-causing variants in a population setting. *bioRxiv* DOI: 10.1101/407981
- 513 71 Glusman, G. *et al.* (2017) Mapping genetic variations to three-dimensional protein structures to  
514 enhance variant interpretation: a proposed framework. *Genome Med* 9, 113
- 515 72 Fersht, A.R. *et al.* (1992) The folding of an enzyme. I. Theory of protein engineering analysis of  
516 stability and pathway of protein folding. *J Mol Biol* 224, 771–782
- 517 73 Allen, M. *et al.* (2018) Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ*  
518 *Preprints* DOI: 10.7287/peerj.preprints.27137v1  
519

## Figures

### Figure 1



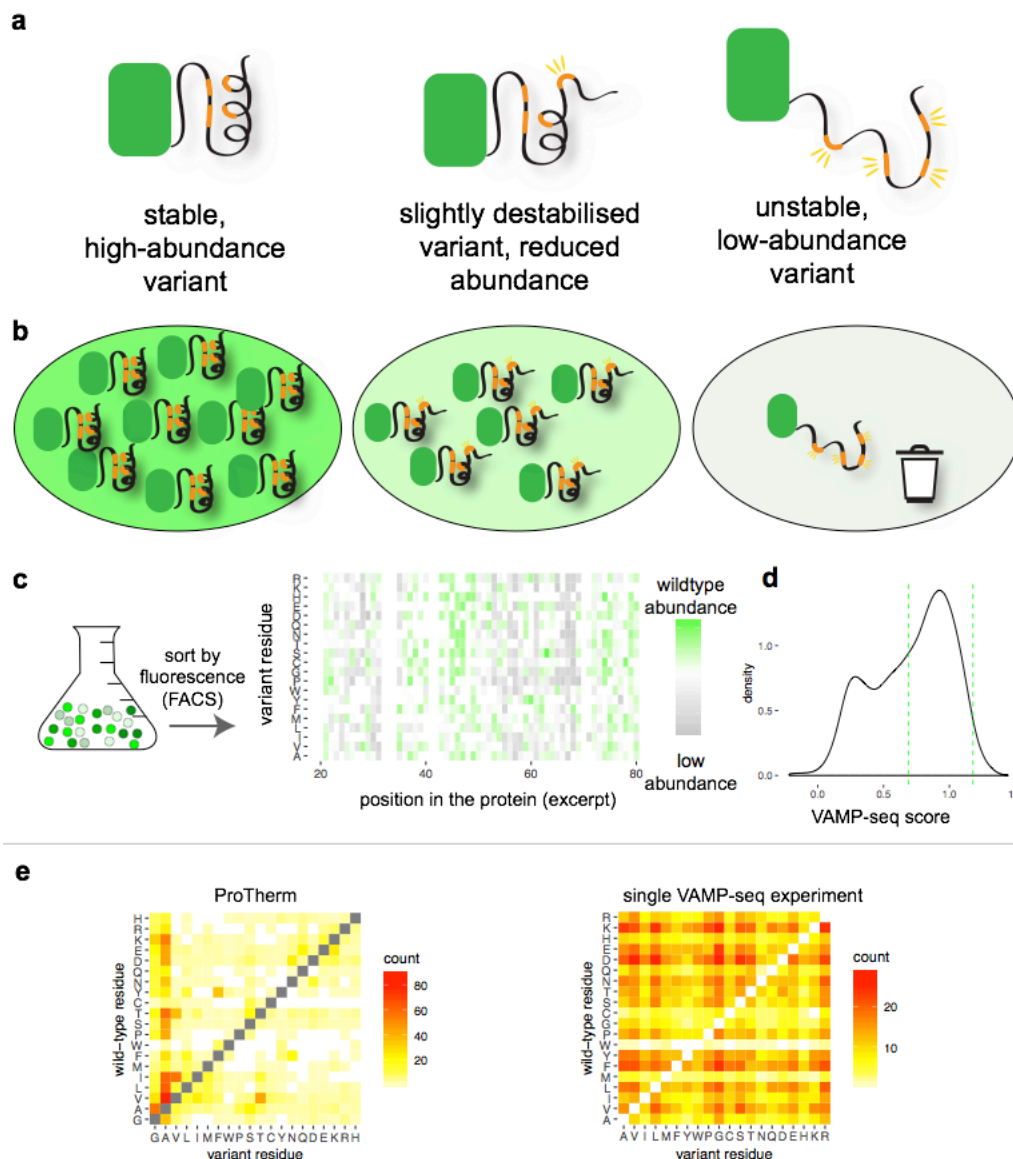
523

524

**Fig 1.** Mechanisms for cellular protein quality control and degradation, and effects of mutations on the folding energy landscape. (a) In a folded protein (left), the degradation signals (degrons, orange) are generally buried inside the protein. Upon local and partial unfolding (bottom route) or full unfolding (top route) one or more degrons may become exposed. The cellular protein quality control (PQC) components (magnifying glass), such as molecular chaperones and E3 ubiquitin-protein ligases, scan the cell for such degradation signals and target the substrates for degradation (right). Mutations may affect all of these steps including increasing the populations of unfolded or partially unfolded states, or creating or removing degron sequences. (b) A globally destabilising variant brings the free energy of the folded conformation closer to that of the fully unfolded state, increasing the population of this state and making the protein more easily targeted for degradation. (c) Because local unfolding involves smaller free energy differences, amino acid changes with more modestly destabilizing effects may still cause substantial increase in locally unfolded states, and possible exposure of degrons. In this way such variants can have a stronger effect in the cell than one would expect from the predicted thermodynamic change of global stability.

538

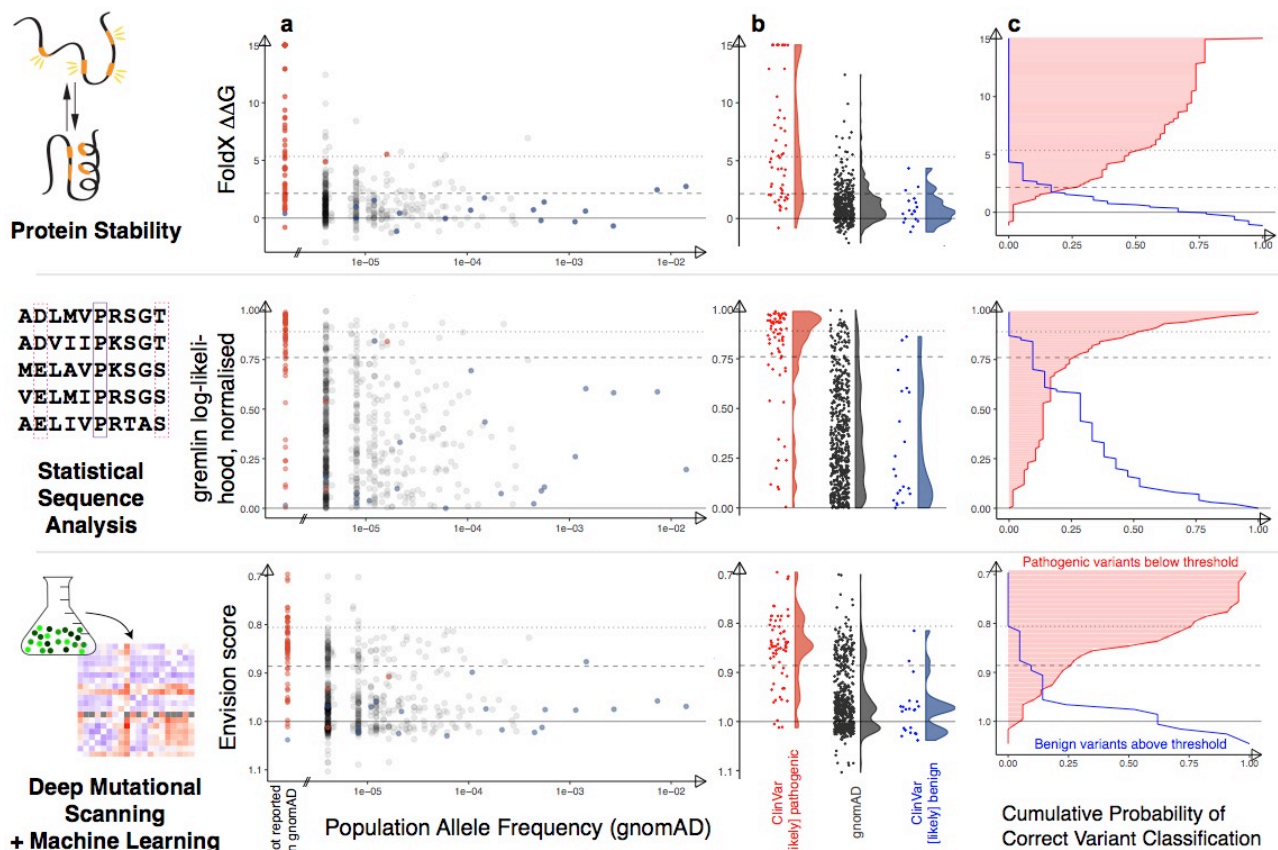
**Figure 2**



**Fig. 2.** Deep mutational scanning for protein stability and variant abundance. Panels A–C outline the VAMP-seq method [41]: (a) generation of a large library of variants, typically all possible 19 variants at each site, and fusion to GFP; (b) abundance of the respective variant fusion construct determines each cell's fluorescence; (c) fluorescence activated cell sorting, followed by sequencing and data analysis allows for the quantification of the abundance of each variant. (d) Distribution of VAMP-seq scores for missense variants in the protein PTEN, normalized such that unity corresponds to the wild type protein sequence and zero to the average of the 1% lowest scoring variants [41]. Green lines

550 indicate the 5<sup>th</sup> and 95<sup>th</sup> percentile for synonymous variants; 56% of the missense variants fall within  
551 this range. (e) Accurate biophysical measurements of the change in protein stability upon amino acid  
552 changes have been collected over many years [42], but are dominated by mutations to alanine, and a  
553 few other chemically, structurally, biophysically-motivated substitutions [72] (left). In contrast, a  
554 single VAMP-seq experiment provides data for a comparable number of variants, but is less bias  
555 chemically (right).

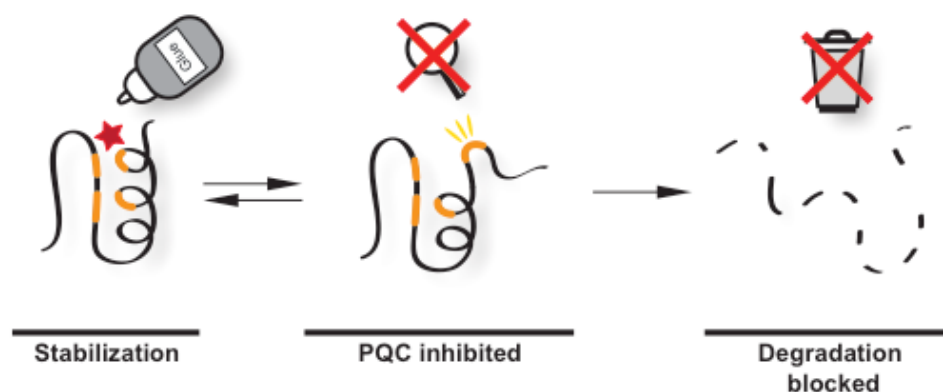
**Figure 3**



**Fig. 3.** Three paradigms for predicting the consequences of amino acid changes. We illustrate the utility of (top) stability predictions, (middle) evolutionary analyses and (bottom) a regression model trained on deep mutational scanning data to predict the consequences for pathogenic and benign MSH2 variants from the ClinVar database [61]. (a) The allele frequencies in the gnomAD database of genome sequences (gnomad.broadinstitute.org) are plotted against the predicted score of the variant. The variant scores are ordered so that detrimental variants are shown at the top, and stability prediction scores were truncated at 15 kcal mol<sup>-1</sup>. Red and blue points are those reported as (likely) pathogenic and benign, respectively, in ClinVar. The left-most “column” of points (labelled “not reported in gnomAD”) contains variants reported in ClinVar, but not observed in gnomAD; they mostly correspond to known pathogenic variants expected to be found at very low allele frequencies. (b) Raincloud plots [73] illustrating the predicted score distributions of pathogenic (red), population (grey) and benign (blue) variants. For all three prediction methods there is a clear, yet also non-perfect, separation between pathogenic and benign variants. (c) Cumulative distribution functions showing which fraction of variants are above/below any given score threshold. The red curve shows the fraction of pathogenic variants below the value (false negatives) and the blue curve the fraction of benign variants above the threshold (false positives). The horizontal dashed lines indicate the respective

576 threshold for 25% false negative predictions, and the dotted lines are the thresholds for no false  
577 positives. Solid lines indicate the respective predictor's value for the wild type. Overall the plots  
578 illustrate that all three predictors correctly identify many of the pathogenic variants as detrimental, and  
579 most of the benign variants as tolerated. The "area under the curve" (AUC) in a receiver operating  
580 characteristic (ROC) analysis is 0.91, 0.90, and 0.91 for the three methods, respectively. To address the  
581 imbalance between the sizes in the pathogenic and benign datasets, the pathogenic dataset was split in  
582 three; these AUCs are averages over these three ROC analyses.

**Figure 4**



**Fig. 4.** Rescuing protein stability as a strategy for therapy. The cellular levels of a destabilized protein variant may be increased by blocking the PQC system (magnifying glass; middle) or the degradation machinery (trashcan; right). Alternatively, a small molecule (star) that associates with the native form of the protein may act as a “glue” to stabilize the protein.