










Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues

Bert Houben^{1,2} , Emiel Michiels^{1,2} , Meine Ramakers^{1,2}, Katerina Konstantoulea^{1,2} , Nikolaos Louros^{1,2} , Joffré Verniers^{1,2}, Rob van der Kant^{1,2} , Matthias De Vleeschouwer^{1,2}, Nuno Chicória^{1,2}, Thomas Vanpoucke^{1,2} , Rodrigo Gallardo^{1,2} , Joost Schymkowitz^{1,2,*}  & Frederic Rousseau^{1,2,**} 

Abstract

Many chaperones favour binding to hydrophobic sequences that are flanked by basic residues while disfavours acidic residues. However, the origin of this bias in protein quality control remains poorly understood. Here, we show that while acidic residues are the most efficient aggregation inhibitors, they are also less compatible with globular protein structure than basic amino acids. As a result, while acidic residues allow for chaperone-independent control of aggregation, their use is structurally limited. Conversely, we find that, while being more compatible with globular structure, basic residues are not sufficient to autonomously suppress protein aggregation. Using Hsp70, we show that chaperones with a bias towards basic residues are structurally adapted to prioritize aggregating sequences whose structural context forced the use of the less effective basic residues. The hypothesis that emerges from our analysis is that the bias of many chaperones for basic residues results from fundamental thermodynamic and kinetic constraints of globular structure. This also suggests the co-evolution of basic residues and chaperones allowed for an expansion of structural variety in the protein universe.

Keywords aggregation; gatekeepers; Hsp70; protein folding

Subject Category Translation & Protein Quality

DOI 10.15252/embj.2019102864 | Received 4 July 2019 | Revised 26 February 2020 | Accepted 27 February 2020 | Published online 1 April 2020

The EMBO Journal (2020) 39: e102864

See also **MB Koopman & SGD Rüdiger** (June 2020)

Introduction

Over the past few decades, many efforts have been made to determine general binding patterns for molecular chaperones. Overall,

chaperones tend to be promiscuous in their binding, which allows them to interact with and thereby assist the folding of a broad range of proteins (Bose & Chakrabarti, 2017). However, most chaperones do show a preference for hydrophobic segments, thereby shielding aggregation-prone regions that should normally be buried in the native state. Interestingly, in addition to a preference for hydrophobic amino acids, many elements of the proteostatic machinery also show a clear charge preference in that they bind basic residues and avoid acidic ones (Table 1). Indeed, the preference for basic residues over acidic ones is found in diverse chaperone systems and seems to be conserved across species and organelles. In the Hsp70 chaperone family, for example, this specific binding motif is shared by orthologs in bacteria (DnaK), yeast (Ssb) and mammalia (HSC70), and in the latter, it occurs both in the endoplasmic reticulum (BiP) and in the cytoplasm. Despite its pervasiveness however, a rationale for the preference of chaperones for basic/hydrophobic sequences has hitherto been lacking.

In parallel, our deepening understanding of the structural mechanisms driving protein aggregation allowed for the development of aggregation prediction algorithms. Using TANGO, we analysed the aggregation propensity of over 20 full proteomes of all kingdoms of life and found that about 20% of all residues in a typical globular domain reside within aggregation-prone regions (APRs). APRs are short (on average 7–8 residues), generally hydrophobic segments within the protein sequence that have a high propensity to aggregate by self-assembling via β -strand interactions (Rousseau *et al*, 2006; Fig 1). Importantly, this initial study also demonstrated that APRs of all proteomes are systematically N- and C-terminally flanked by the charged amino acids Arg, Lys, Asp and Glu, a pattern that was soon confirmed independently (Monsellier *et al*, 2008). We proposed that these counteract the aggregation propensity of APRs by charge repulsion and coined these charged residues in the flanks of APRs as “aggregation gatekeepers” (GKs) in reference to the proposition of Otzen *et al* (2000) that protein sequences undergo negative selection against aggregation. We chose the denomination “aggregation

¹ Switch Laboratory, VIB Center for Brain and Disease Research, Leuven, Belgium

² Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium

*Corresponding author. Tel: +32 16 37 25 70; E-mail: joost.schymkowitz@kuleuven.vib.be

**Corresponding author. Tel: +32 16 37 25 70; E-mail: Frederic.rousseau@kuleuven.vib.be

Table 1. Overview of protein quality control elements described to have binding preferences for hydrophobic and basic residues.

PQC element	Species	Subcellular location	Binding preferences reported	References
DnaK	<i>Escherichia coli</i>	Cytoplasm	R, K, hydrophobic	Rüdiger et al (1997); Van Durme et al (2009); de Crouy-Chanel et al (1996); Deuerling et al (2003)
DnaJ	<i>Escherichia coli</i>	Cytoplasm	R, hydrophobic	Rüdiger et al (2001)
TF	<i>Escherichia coli</i>	Cytoplasm	R, K, large aromatic	Patzelt et al (2001); Deuerling et al (2003)
SecB	<i>Escherichia coli</i>	Cytoplasm	R, K, aromatic	Knoblauch et al (1999)
ClpB	<i>Escherichia coli</i>	Cytoplasm	R, K, F, W, Y	Schlieker et al (2004)
Ssb	Yeast	Cytoplasm	R, K, V	Döring et al (2017)
HSP90	Mammalian	Cytoplasm	Moderate hydrophobicity and a positive net charge (in Tau)	Karagöz and Rüdiger (2015)
IRE1	Mammalian	Endoplasmic Reticulum	R, W, Y	Karagöz et al (2017)
BiP	Mammalian	Endoplasmic Reticulum	R, K, W, Y; aliphatic amino acids, R; large hydrophobic and aromatic, basic tolerated	Flynn et al (1991); Fourie et al (1994)
Hsc70	Mammalian	Cytoplasm	Large hydrophobic and aromatic, basic tolerated, acidic not tolerated	Fourie et al (1994)

gatekeepers” rather than “structural gatekeepers” to emphasize the specific position of these residues at the flanks of APRs. Several other studies confirmed the prevalence and importance of the GK pattern (Monsellier et al, 2008; Tartaglia & Vendruscolo, 2008; Buell et al, 2009; Wang et al, 2010b; Markiewicz et al, 2014; Sant’Anna et al, 2014; Estácio et al, 2015). We later confirmed that GKs constitute a *bona fide* and ubiquitous functional class specifically devoted to protein homeostasis. Indeed, GKs are evolutionarily conserved at a significant cost to the thermodynamic stability of the native structure (0.5 kcal/mol on average) (De Baets et al, 2014). Accordingly, we found that mutating GKs not only affects aggregation but also affects protein synthesis and degradation rates, heat stress response and overall cellular fitness (Beerten et al, 2012).

It remains unexplained however, why chaperones favour interaction with APRs capped by basic residues (Arg, Lys) at the expense of acidic residues (Asp, Glu) and whether there is a functional and/or evolutionary distinction between acidic and basic GKs.

Here, we show how the preference of the protein quality control machinery for basic residues is a consequence of fundamental structural and evolutionary constraints on globular structure. The evolution of globular structure resulted in the enrichment and fixation of amyloidogenic segments in natural protein sequences (Linding et al, 2004). In order to kinetically favour native folding over aggregation, these APRs are systematically flanked by charged residues that disfavour aggregation. Here, we demonstrate a fundamental difference in the ability of acidic and basic residues to achieve this feat. We find that while acidic residues (and especially Asp) are the most effective inhibitors of aggregation since they not only slow down aggregation kinetics, but also fundamentally alter the thermodynamic stability of the aggregates. On the contrary, they are not easily compatible with the diversity of structural contexts of APRs in native structures, because their short sidechains do not allow deep burial in the hydrophobic core. On the other hand, while being more

easily accommodated in the native structure, the basic residues Arg and Lys are much less efficient aggregation inhibitors that mostly act through slowing down the aggregation kinetics with minimal impact on the thermodynamic stability of the amyloid-like conformation. We show that Hsp70 is adapted to bind APRs that are capped by basic GKs, thereby compensating their poorer intrinsic ability to inhibit aggregation. The binding pocket of Hsp70 actually mimics the packing of hydrophobic residues in the core of globular proteins, while offering extra stabilization to the basic GKs by salt bridges, thereby compensating for their less-efficient inhibitory effect on aggregation kinetics.

Results

Acidic GKs flank the strongest APRs and are more conserved than positive GKs

To recapitulate the basic observation of aggregation GK enrichment at the flanks of APRs and to compare the different residue types, we analysed the cytoplasmic proteins of *Escherichia coli* using the TANGO algorithm (Fernandez-Escamilla et al, 2004). However, since TANGO calculations of the aggregation propensity incorporate the effects of GKs, we introduced a second step in the calculation, in which we recalculate the TANGO score of the isolated APRs, thus removing the GKs. We called the resulting aggregation propensity score of APRs without the mitigating effect of the flanking GKs the “PureTANGO” score, a detailed calculation of which can be found in the Materials and Methods section. As expected, the PureTANGO score is always higher than the TANGO score and we called the difference between both values the “GK effect”, as this reflects the impact of the GKs on the final TANGO score. A plot of the GK effect versus the positions close to the N-terminus and C-terminus of the

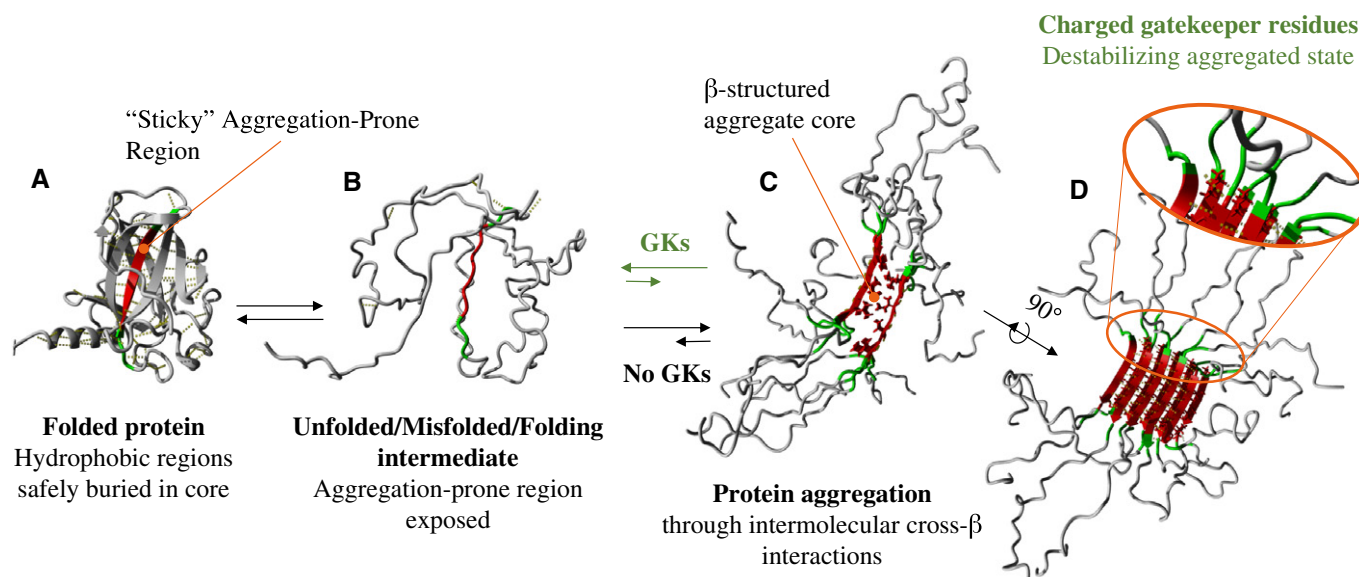


Figure 1. The APR-mediated aggregation process and its attenuation by aggregation gatekeepers (GKs).

A–D (A) In native protein structures, APRs (red) are usually buried within the hydrophobic core, rendering them harmless (structure adopted from PDB-entry 2AC0). Exposure of these APRs however (B) can lead to formation of highly stable aggregated species (C and D, structure adopted from PDB-entry 2M5N) that are detrimental to both protein function and cellular viability. Aggregation GKs (R, K, E and D, green) decrease APR-mediated aggregation tendency through charge repulsion. In doing so, GKs alleviate the threat posed by APRs and push the balance towards the native fold as indicated by the green arrows.

APRs shows that the GK effect increases closer to the APR (Fig 2A), closely mimicking the statistical enrichment profiles of GKs (Reumers *et al*, 2009). Furthermore, for each position there is a significantly higher GK effect for Asp and Glu than there is for Arg and Lys, revealing that Asp and Glu are predicted by the TANGO algorithm to be stronger aggregation breakers, with Asp generally being the strongest. The weaker GK effect of positively charged amino acids in TANGO derives from several factors, most notably the more favourable burial energies of the longer positively charged side chains, which is not entirely offset by their slightly higher entropic costs (Fernandez-Escamilla *et al*, 2004). To further analyse the impact of the intrinsic aggregation propensity, we compared the GK enrichment in the 30% strongest versus the 30% weakest APRs in terms of PureTANGO score in the same *E. coli* protein set. This revealed a strong enrichment of negatively charged GKs at the flanks of the 30% strongest APRs in the set, suggesting these residues might have some selective advantage in this group (Fig 2B). To corroborate this, we compared the conservation of GK residues, both in *E. coli* (Fig 2C) and in the human proteome (Fig 2D), and found that in both cases, there is a stronger conservation of the negative than the positive GKs, again suggesting there may be a selective advantage to the negative GKs.

Acidic GKs are more efficient aggregation inhibitors than basic GKs

To experimentally compare the efficiency of different charged GK residues to inhibit the aggregation of their cognate APRs, we measured the aggregation kinetics and solubility of a set of 66 peptides consisting of six variants of 11 *E. coli* APRs. The set of APRs assessed in this study is shown in Table 2. GK variants of

each APR flanked at both ends by one of the four charged residues, i.e. Asp, Glu, Arg or Lys, were produced (Fig 3A). Control peptides were added with either Ala in the GK position (AA) or no gatekeeping residues at all (noGK). All of the peptides in this set were dissolved to a concentration of 100 μ M in PBS, and aggregation events studied through time-resolved dynamic light scattering (DLS), amyloid-specific dye-binding kinetics—both the widely used Thioflavin-T (ThT) and the more recently developed polythiophene dye pFTAA (Hammarström *et al*, 2010)—imaging using transmission electron microscopy (TEM) and endpoint critical concentration measurements. Figure 3B–F shows the results of these analyses for the peptide set derived from the aggregating core of the β -galactosidase (β -Gal) enzyme of *E. coli*. Time-resolved DLS of the initial stages of this process showed that AA-gatekept and non-gatekept APR peptides rapidly form particles with a hydrodynamic radius of over a micron (Fig 3B and Appendix Fig S11). Introducing charged GKs slows down the aggregation process, with basic GKs doing so to a lesser extent than acidic GKs. ThT- and pFTAA-binding experiments show that the aggregated particles formed by the positively charged and non-gatekept peptides consist of β -structured species (Fig 3C and D). Endpoint critical concentration measurements of this peptide set revealed that after an incubation period of 7 days at a starting concentration of 100 μ M, peptides with basic GKs have precipitated more than the negative ones (Fig 3E). Moreover, GKs in general favour APR solubility as compared to a non-gatekept APR. Finally, TEM imaging of the aggregated peptides confirmed that the acidic GK peptides underwent little aggregation, and that the species that were formed are mostly amorphous structures, although the Glu-gatekept peptide does show limited fibre formation (Fig 3F). The amorphous-like species are also found in some buffer control grids, suggesting some of them are in fact

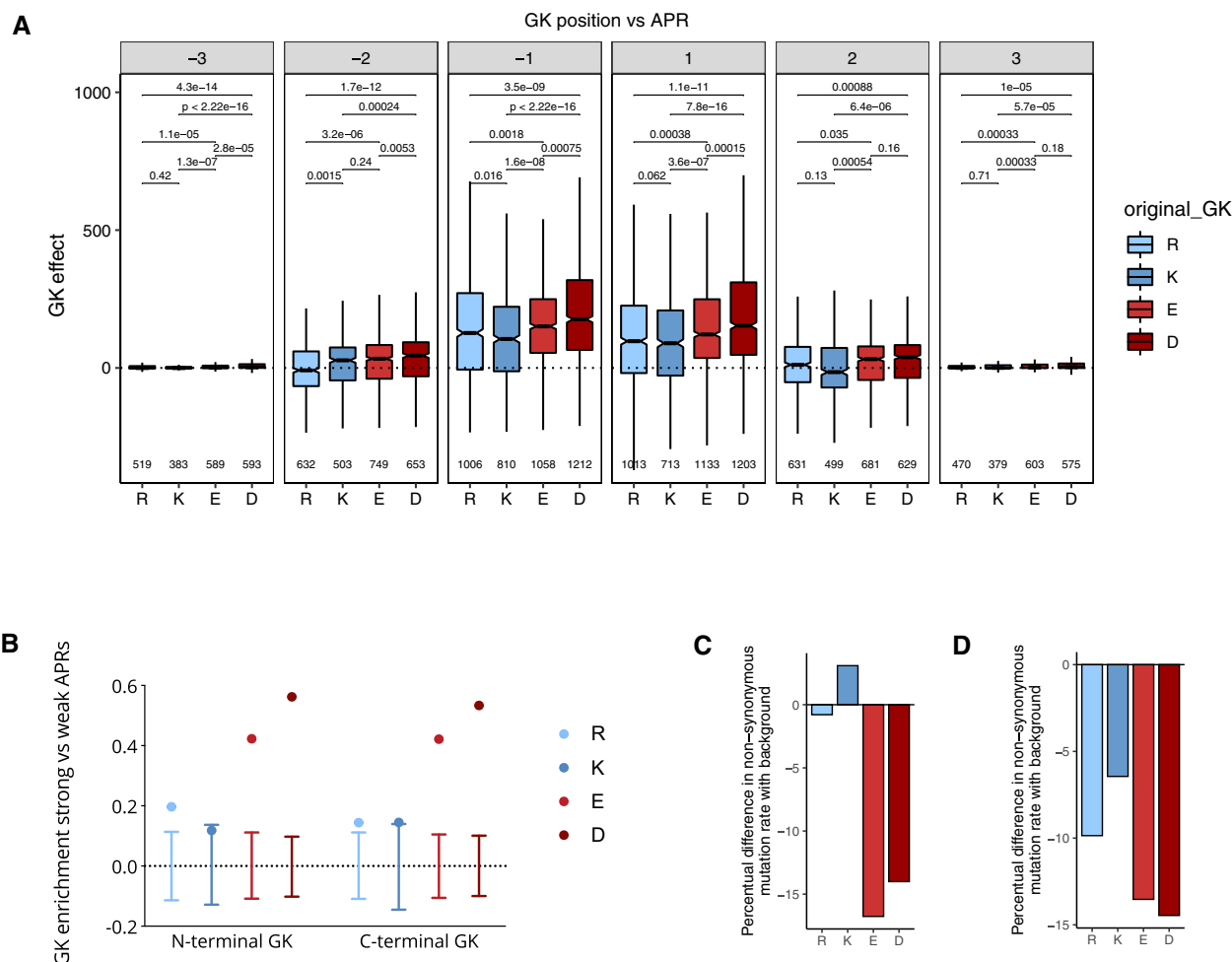


Figure 2. Acidic GKs are stronger aggregation breakers, are more conserved and more enriched in the flanks of strong APRs than basic GKs.

A Boxplots showing GK effect—the difference between TANGO score and PureTANGO score—per charged GK, and per position with reference to the APR. Negative numbers indicate N-terminal positions, and positive numbers refer to C-terminal positions. The upper whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least 1.5 * IQR from the lower hinge. Notches extend 1.58 * IQR/sqrt(n) from either side of the median and represent a 95% confidence interval for the median value. Significance was determined through Kruskal–Wallis test with a *post hoc* Wilcoxon analysis for pairwise testing and Bonferroni correction for multiple testing. Adjusted *P*-values are indicated for each comparison. The number of observations per group is indicated beneath each box.

B Log(odds enrichment) scores for charged GKs in the flanks of the 30% strongest versus the 30% weakest APRs in the *Escherichia coli* cytoplasmic proteome. Whiskers indicate 95% confidence intervals for no enrichment as determined through bootstrapping analysis.

C Percentual difference in non-synonymous mutation rate in *E. coli* for GK versus background, i.e. versus the corresponding residue in non-gatekeeper positions.

D Percentual difference in non-synonymous mutation rate in mammalian species for GK versus background, i.e. versus the corresponding residue in non-gatekeeper positions.

staining artefacts. The positively, Ala- and non-gatekept peptides all showed widespread aggregation. Interestingly, the non-gatekept peptides form amorphous aggregates, while Arg-, Lys- and Ala-protected peptides show formation of amyloid fibrillar structures. These data suggest that although the aggregates formed by non-gatekept peptides are stained by both amyloid-specific dyes, they do not seem to mature into fibres discernable through TEM imaging. Indeed, aggregating peptides can in fact be cross- β structured in nature without long-range structure, leading them to bind amyloid-specific dyes while appearing amorphous under an electron microscope (Wang *et al*, 2010a). Together, these biophysical data indicate that acidic GKs more efficiently slow down the aggregation process of the β -Gal aggregating core than basic GKs. The combined

results of all the 11 aggregating cores in our peptide set largely confirm the generality of these findings (Appendix Figs S1–S11). A comprehensive overview of these results is provided in Fig 3G—where peptides were classified on whether or not they show a kinetic in the dye-binding and DLS analyses, and whether they form amyloid fibres detectable in TEM imaging—and Fig 3H—showing the distribution of endpoint solubilities of the entire peptide set, grouped per GK. Time-resolved DLS analyses confirm that basic peptides are less likely to prevent aggregation over the first 15 h than acidic GKs, with Asp performing the best. In fact, basic GKs only marginally outperform non-gatekept variants in this assay. ThT binding confirms the acidic GK superiority, although it does seem to suggest a better performance of basic GKs than indicated by the DLS

Table 2. Aggregating peptide cores used in this study.

Peptide ID	UniProt accession	Gene	APR position	APR core
1	POA993	fbp	105	YVVL
3	P06715	gor	283	GYIV
4	POA7A9	ppa	16	IYVVI
5	P76141	lsrR	22	IAWFYH
6	POA763	ndk	129	IAYFFG
7	POCOR7	rlmE	200	VYIVATG
8	POAGB6	rpoE	68	SAFYTWLY
9	P33012	sbmC	47	WVAVYY
10	P00959	metG	250	YFYVWL
11	POAGJ7	trmL	2	LNIVLY
β -Gal	P00722	lacZ	454	SVIIWSLGN

analyses. We wondered whether this is an artefact caused by the positive charge of the ThT dye that might disturb binding to positively charged peptides, and therefore employed the negatively charged pFTAA as a control. This obviously imposes the reverse charge bias, but does correlate more strongly with the DLS data, and shows that a negative ThT signal is not indicative of the absence of amyloid species in the positively gatekept peptides (peptide 5 is a strong example of this pattern). Moreover, TEM imaging, which is not biased by charge, again confirms the finding from the DLS analyses, showing that positively gatekept peptides are more likely to form discernable amyloid fibres. The fibrillar nature of basic GK peptide aggregates demonstrates that basic GKs are thermodynamically still compatible with amyloid-like structure. These combined data also generalize the finding that positively flanked APRs are even more likely to form highly ordered fibres than their non-gatekept counterparts. Finally, endpoint concentration measurements show that GKs thermodynamically disfavour aggregation, and that acidic GKs tend to perform better at this task (Fig 3H). Taken together, these biophysical measurements demonstrate that acidic GKs are intrinsically much better at inhibiting the aggregation of APRs.

To assess whether these differences in intrinsic GK activity can be extrapolated to APRs on a proteome-wide scale, we determined whether gatekeeping affects protein solubility in the *in vitro* translation screen performed by Niwa *et al* (2009). This dataset contains solubility data upon *in vitro* translation for a large amount of *E. coli* open reading frames in the absence of chaperones. In light of the results of our biophysical peptide analyses outlined above, we stipulated that those APRs that are not flanked by any negative charges have the highest risk of aggregating. We then analysed whether proteins that contain a strong APR (PureTANGO score > 70) that is not protected by at least one negative GK have lower solubility than those that have all their APRs protected by at least one negative charge in the three flanking positions on either side. We found that indeed, proteins with an “unprotected” APR are significantly less soluble (Fig 3I). As a control, we assessed whether having a strong APR that is not protected by at least one positive charge also undermines solubility, which it does not. This again shows in an independent experimental dataset that positive residues are less efficient GKs on the protein scale.

Dissection of the energetic contributions to GK efficiency

To dissect the energetics of the GK effects of the different amino acid types, we gathered a set of known cross- β zipper structures (Table 3) from the PDB and assessed the thermodynamic effects of the introduction of GKs at flanking positions in these structures using the FoldX force field (Guerois *et al*, 2002; Schymkowitz *et al*, 2005) (Fig 4A, left panel). FoldX calculates the separate energetic contributions of different structural parameters including electrostatics, H-bonding, desolvation and van der Waals energies and main-chain and sidechain entropy. In accordance with the experimental observations above, FoldX correctly predicts acidic residues to be more destabilizing than basic residues to cross- β amyloid zippers (Fig 4B). Moreover, FoldX modelling shows that this difference mainly stems from stronger electrostatic repulsion and lower stabilization through van der Waals packing and desolvation of hydrophobic atoms for acidic versus basic residues. These differences to a large extent stem from the fact that acidic sidechains are much shorter and less hydrophobic than basic residues. Indeed, their longer aliphatic sidechains also allow for the energetically more favourable burial and packing of hydrophobic atoms into the aggregates and the increased flexibility of Arg and Lys allows for a larger distance between repulsive charges. Given that Coulomb's law of charge–charge interactions stipulates a quadratic dependence on the distance between the charges, even relatively small increases in the average charge distance have a major impact on the effective repulsion energy. In order to confirm that their increased flexibility allows basic residues to move their charged atoms further apart when engaged in a cross- β structure, we measured the distances between the charged moieties of the most centrally situated GKs in our zipper models (at the N-terminus and C-terminus of chains E and F, as shown in Fig 4A), and their closest adjacent charge (Fig 4C). Clearly, across all the zipper structures analysed, Arg and Lys show the greatest capacity to separate their charged moieties. Asp, the shortest and most rigid GK, is severely limited in its ability to distance its charged atoms from those of neighbouring polypeptides, leading to a net increase in charge repulsion. Intriguingly, although Arg clearly has the capacity to move charged moieties apart, the guanidinium groups are regularly found in closer vicinity to each other than even the carboxylate groups of Asp, suggesting a stabilizing effect rather than a disrupting one (Fig 4A, right-hand side). In fact, the planar stacking of Arg residues has been described in protein structures at similar distances to the 4 Å we observe here (Magalhaes *et al*, 1994; Neves *et al*, 2012). The occurrence of these unusually close charged residues is attributed to strong stabilizing water networks that bridge the stacked guanidinium moieties and compensate for the charge repulsion. This analysis again confirms that desolvation of the negatively charged carboxylate groups is energetically more unfavourable than that of the positively charged groups, in particular guanidinium, pointing to an additional force permitting Arg-gatekept APRs to form amyloid fibres.

In order to experimentally verify the general conclusion from these modelling calculations, we synthesized a set of peptides containing APRs flanked by non-natural basic and acidic amino acids where we altered the length of the aliphatic sidechains that separate the charged moiety from the residue backbone. 2,3-diaminopropionic acid (DAP) was used as short version of Lys, with only one methyl group separating the charged moiety from the backbone,

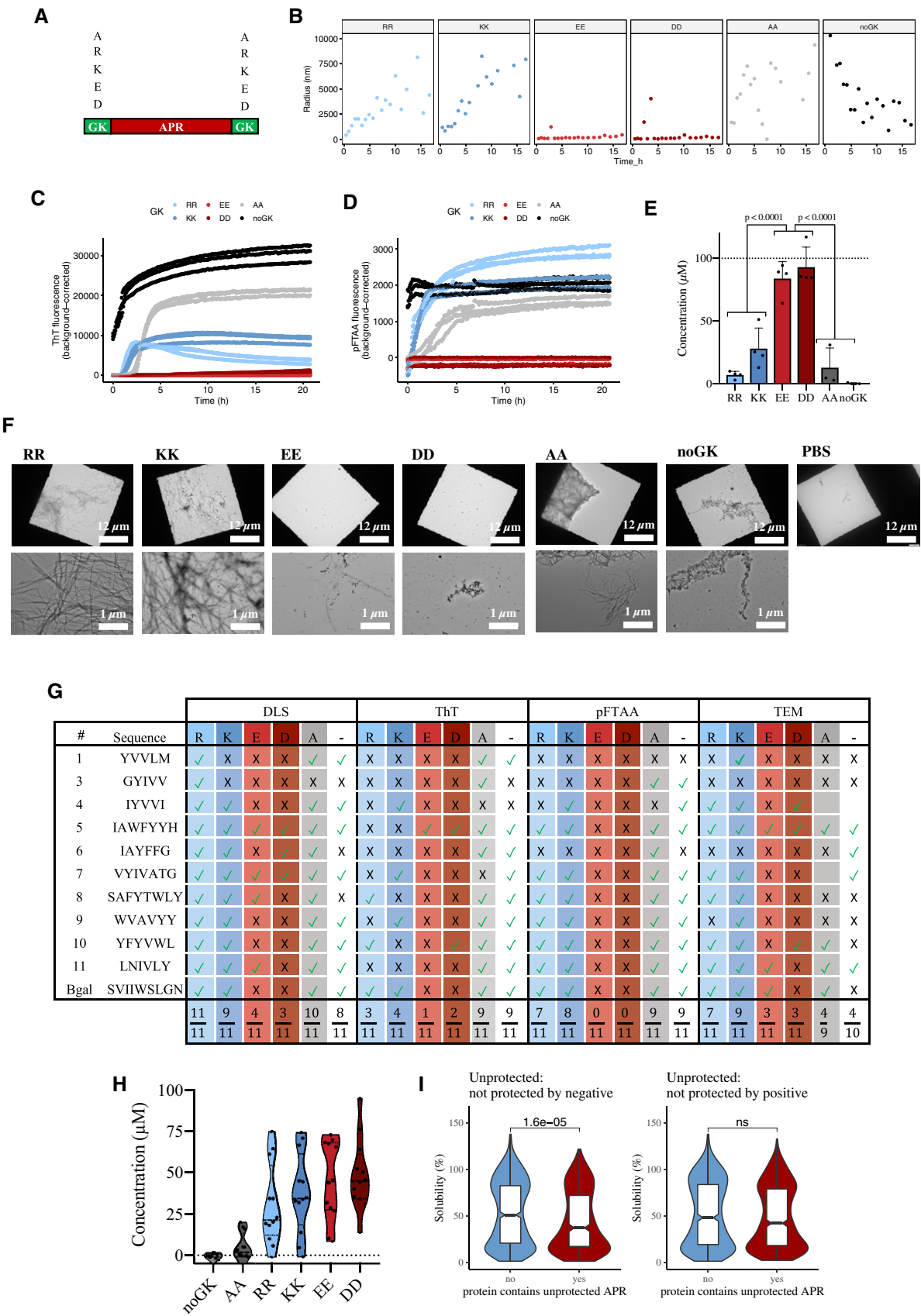


Figure 3.

Figure 3. Aggregating peptides are more potently solubilized by acidic than by basic GKs.

- A Schematic representation of the peptide design used in this study. An aggregating peptide core (or APR, red) is flanked on either side by one of four charged residues (Arg, Lys, Asp or Glu), an Ala on either side, or no flanking residues (noGK).
- B Time-resolved dynamic light scattering (DLS) analysis indicating average particle size over time for each of the peptides in the β -Gal peptide set (aggregating core sequence SVIIWSLGN). One representative result is shown, and two independent repeats can be found in Appendix Fig S11.
- C ThT binding kinetics of the β -Gal peptide set. Background-corrected fluorescence over time is shown for 3 independent repeats.
- D pFTAA binding kinetics of the β -Gal peptide set. Background-corrected fluorescence over time is shown for 3 independent repeats.
- E Endpoint solubility analysis quantifying the concentration of peptide remaining in the supernatant following ultracentrifugation for the β -Gal peptide set. Bars depict mean concentration, and whiskers indicate standard deviation ($n = 4$). Statistical significance was determined through two-way ANOVA with Bonferroni's correction for multiple comparisons. P -values for significant differences are indicated.
- F TEM images of the β -Gal peptide set after 3 days of incubation. Top panels give an overview image of one square in a TEM grid, and lower panel shows enlarged image (magnification 12,000 \times). Scale bars indicate 12 μ m (top row) or 1 μ m (bottom row).
- G Combined results for the aggregation analyses for the entire peptide set. Peptides were classified on whether they show a kinetic in the dye-binding and DLS analyses, and whether they form amyloid fibres detectable in TEM imaging.
- H Violin plots showing the distribution of endpoint solubilities across 11 aggregating cores, grouped by GK. The dotted line indicates a concentration of 0 μ M, corresponding to completely insoluble peptides.
- I Solubility analysis of the *in vitro* solubility screen performed by Niwa *et al* (2009). Left panel shows combined boxplots and violin plots depicting the solubility distributions of proteins containing at least one strong APR (TANGO score > 70) not protected by any negative GKs (red) versus proteins where all strong APRs are flanked by at least one negative charge (blue). The right panel shows a control analysis, in which proteins were classified by the occurrence of at least one strong APR (TANGO score > 70) not flanked by any positive GKs (red) versus proteins where all strong APRs are flanked by at least one positive charge (blue). The upper boxplot whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least 1.5 \times IQR from the lower hinge. Notches extend 1.58 \times IQR/sqrt(n) from either side of the median and represent a 95% confidence interval for the median value. Statistical significance was determined through Kruskal–Wallis test with a *post hoc* Wilcoxon analysis for pairwise testing. P -values are indicated.

instead of four. On the other hand, 2-aminoheptanedioic acid (AHD) was used as a longer substitute for Asp and Glu, with four methyl groups separating the charged moiety from the backbone, instead of 1 and 2, respectively. To assess the effect of sidechain length on electrostatic repulsion, we measured the effects on aggregation of these non-natural GKs in a Tris buffer at pH 8 containing a range of salt concentrations using time-resolved DLS (Fig 4D and Appendix Fig S12). We found that with NaCl concentrations up to 75 μ M, DAP slows down aggregation more strongly than Lys, indicating that sidechain length and the effect on charge repulsion associated with it indeed help determine the lower efficiency of positive GKs. However, increasing the sidechain length of negatively charged GKs does not equivalently alter their effect on the aggregation kinetics, suggesting again that the chemical nature of the carboxyl group itself is intrinsically more efficient at inhibiting aggregation. This inherent superiority may in part be explained by the larger desolvation penalty of the carboxylate group as compared to the amine and guanidinium moieties (Collins, 1997; Mason *et al*, 2003; Trevino *et al*, 2007). In addition to this, Asp and Glu have a higher tendency to form sidechain–mainchain hydrogen bonds, which compete with the backbone–backbone hydrogen bonds necessary for cross- β formation (Eswar & Ramakrishnan, 2000). This parameter is not directly picked up in our modelling analyses since our zipper set is already fully mainchain–mainchain hydrogen bonded, and FoldX does not sample altered backbone conformations that would be necessary to detect sidechain–mainchain hydrogen-bond formation in individual strands.

Structural constraints restrict the usage of acidic GKs

To investigate the compatibility of acidic and basic GK residues with the globular structure context of APRs, we analysed a set of 230 *E. coli* protein structures consisting of monomeric, cytoplasmic proteins with a resolution of < 3 Å and a mutual sequence similarity less than 70%, representing a diversity of protein structural topologies. Using the FoldX force field, we assessed how different GKs affect thermodynamic stability of the folded structure. In

accordance with previous analyses (De Baets *et al*, 2014), both acidic and basic GKs on average destabilize native protein structures to a similar extent (0.5 kcal/mol on average, Fig 5A). However, analysing mainchain burial of the charged residues we find that basic residues are generally more buried in the structure than acidic ones (Fig 5B). Additionally, comparing GKs with non-GK charged residues—i.e. charged residues not flanking APRs—we find that there is no difference in burial between acidic residues in GK and non-GK positions, while basic residues occupy more deeply buried positions when they take the role of GK. This demonstrates that basic residues have a higher adaptability to be positioned at the flanks of APRs where they can be buried deeply into the hydrophobic core. In contrast, due to their short sidechains and higher desolvation penalties, acidic residues cannot be buried more deeply at APR flanks, which restricts their use as GKs. This is further illustrated in Fig 5C and D, showing that Asp is already fully buried at a depth of 5 Å, while at the other extreme, Arg can be accommodated to depths over 6.5 Å. Dissecting the energetic contribution of burial, we indeed find that Asp and Glu pay higher penalties for the burial of polar atoms than Arg and Lys do (Fig 5E). Furthermore, Asp shows a very low contribution of the burial of hydrophobic atoms to stability, that plateaus at lower depths, since it has only one methyl group in its short sidechain (Fig 5F). Glu and Arg show a similar pattern of hydrophobic stabilization, as they respectively have two and three methyl groups that they bury at similar depths. Lysine, with the longest sidechain, has the most hydrophobic atoms (and hence length of sidechain) to bury and can keep burying until a large depth (about 6.5 Å). Similar patterns are observed for van der Waals packing energies, where the shorter the sidechain, the less atoms can be packed into the structure (Fig 5G). Conversely, the entropic penalties for sidechain burial scale roughly linearly with sidechain length (Pickett & Sternberg, 1993), but the penalty this implies for basic residues does not fully off-set the enthalpic contributions (Fig 5H). In short, these analyses indicate that the more potent acidic GKs are less readily buried in protein cores, in part as a consequence of their shorter sidechain lengths, and this restricts their use. For APRs that require incorporation of

Table 3. Amyloid zipper structures used in this study.

PDB ID	Sequence	Protein
2OLX	NNQQ	Sup35
2ONX	NNQQ	Sup35
4QXX	GNLVS	Eosinophil major basic protein (EMBP)
1YJO	NNQQNY	Sup35
2ON9	VQIVYK	Tau
2ONW	SSTSAA	Bovine pancreatic RNase A
3DG1	SSTNVG	IAPP
3FTR	SSTNVG	IAPP
3FVA	NNQNTF	Prion Elk
3PPD	GGVLVN	Prostatic acid phosphatase
3Q2X	NKGAI	Abeta
4R0P	IFQINS	Lysozyme
4RP7	TIITLE	p53
2OL9	SNQNNF	Prion
2Y3J	AIIGLM	Abeta
4XFO	TAVVTN	Transthyretin
5E5X	ANFLVH	IAPP
2ONV	GGVVIA	Abeta
3FPO	HSSNNF	IAPP
3SGS	GDVIEV	AlphaB-crystallin
3LOZ	LSFSKD	Beta-2-microglobulin
3NVE	MMHFGN	Prion syrian hamster
3PZZ	GAIGL	Abeta
2OMP	LYQLEN	Insulin
2OMQ	VEALYL	Insulin
2Y29	KLVFFA	Abeta
2Y2A	KLVFFA	Abeta
3FOD	AILSST	IAPP
3FR1	NFLVHS	IAPP
4TUT	GGYMLG	Prion
4UBY	GGYVLG	Prion
5E5Z	L VHSSN	IAPP
2OKZ	MVGGVW	Abeta
2ONA	MVGGVW	Abeta
3NHC	GYMLGS	Prion M129 mutant
3NHD	GYVLGS	Prion V129 mutant
4WBU	GYMLGS	Prion
4WBV	GYVLGS	Prion
3OW9	KLVFFA	Abeta
4UBZ	GGYLLG	Prion
4XFN	AEWFT	Transthyretin
1YJP	GNNQQNY	Sup35
2OMM	GNNQQNY	Sup35
3DGJ	NNFGAIL	IAPP
3FTK	NVGSNTY	IAPP

Table 3 (continued)

PDB ID	Sequence	Protein
3FTL	NVGSNTY	IAPP
3HYD	LVEALYL	Insulin
4NIO	CVTGIAQ	SOD (I149T)
4NIP	GVIGIAQ	SOD
4R0U	TGTAVA	Alpha-synuclein
4R0W	VTGVTA	Alpha-synuclein
4NIN	DSVISLS	SOD
4RP6	LTIITLE	p53
3FTH	NFLVHSS	IAPP
4W5Y	GYMLGSA	Prion
4W67	GYVLGSA	Prion
4W71	GYLLGSA	Prion
5E5V	NFGAILS	IAPP
4W5L	GGYLLGS	Prion
4W5M	GGYMLGS	Prion
4W5P	GGYVLGS	Prion
2Y3K	MVGGVVIA	Abeta
2Y3L	MVGGVVIA	Abeta
4ZNN	GVHGVTTVA	Alpha-synuclein
4RIL	GAVVTGVTAVA	Alpha-synuclein
2M5N	YTIAALLSPYS	Transthyretin

GKs at very deeply buried positions, the less effective positively charged GKs are the only option that is structurally compatible with native state stability. For experimental validation of the differing burial capacity of the charged residues, we turned to a study conducted by Varadarajan and colleagues (Bajaj *et al.*, 2005). Briefly, they mutated every residue in the *E. coli* toxin Ccdb to each of the charged residues and assessed the effects of this mutation on protein function, as a corollary for folding and native state stability. Using these data, we assessed the depth at which the insertion of a charged residue becomes predominantly detrimental to protein function (Fig 5I). Clearly, Asp and Glu destabilize the structure at lower depths than Arg and Lys. In fact, from a depth of 5 Å, more than half of all Asp mutations abrogate protein function. Glu reaches this threshold at about 6.5 Å, while Arg and Lys can be buried up to 7 Å before they become predominantly detrimental to protein structure. These data correlate well with our assessments in Fig 5C and D.

DnaK compensates for the inferior gatekeeping potential of basic residues

As shown in Table 1, several groups have reported on a binding preference of DnaK for basic residues in a hydrophobic context. Moreover, in a cellulose-based peptide–protein-binding assay, Rüdiger *et al.* (1997) observed specific binding of DnaK to a hydrophobic stretch of 4–5 residues (an APR) flanked by positively charged residues (GKs). Here, we verify this specific binding pattern and show that DnaK binding to the β-Gal APR is

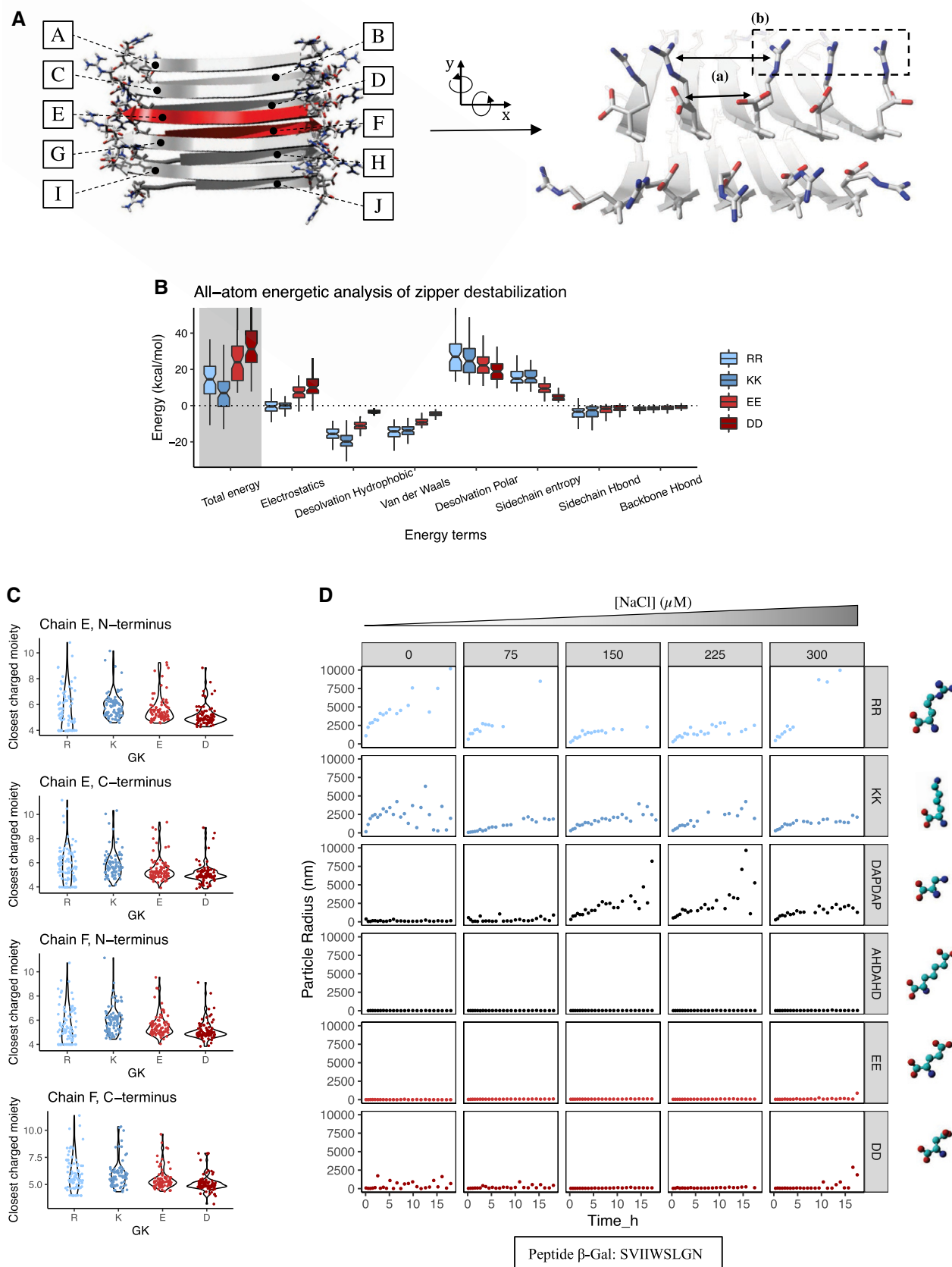


Figure 4.

Figure 4. Dissection of the energetic contributions to GK efficiency.

- A Left panel: overlay of the models resulting from the introduction of Arg and Asp into the 2M5N zipper structure, with the identifier indicated for each strand. Strands E and F are highlighted in red. Right panel: the same structure turned 90° about both the x- and y-axes. Arrows in (a) indicate the capacities of Arg and Asp to distance their charged moieties. (b) shows an example of planar Arg stacking.
- B Boxplots showing all-atom analysis of the energetic effects of the introduction of GKs into a set of amyloid steric zipper structures. The upper whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least 1.5 * IQR from the lower hinge. Notches extend $1.58 * \text{IQR}/\sqrt{n}$ from either side of the median and represent a 95% confidence interval for the median value. Grey shaded area indicates total energy associated with the introduction of GKs into zipper structures, the remainder of the plot shows the breakdown of this total energy into its constituent energy terms. Each boxplot shows the distribution of 72 observations.
- C Violin plots showing the distribution of distances to the closest charged moiety in a neighbouring GK residue in the central-most positions (strands E and F) of the steric zipper models. Each violin shows the distribution of 83 observations.
- D Time-resolved dynamic light scattering (DLS) analysis showing average particle radius over time for the β -Gal peptides flanked by both natural and non-natural GKs, at increasing NaCl concentrations (indicated in μM). Structures of the amino acids flanking the aggregating core are indicated on the right-hand side, and the sequence of the aggregating core is indicated below.

stronger when this APR is flanked by basic GKs, versus when it is flanked with acidic ones, indicating that basic GKs indeed favour DnaK binding (Fig 6A). As a positive control, we included the aggregating core VLYLQ from the *E. coli* $\sigma 32$ transcription factor, which has been shown to be a DnaK binding site (Rodriguez *et al*, 2008). Similarly to the β -Gal APR, placing basic GKs in the flanks of the $\sigma 32$ APR strongly increases DnaK binding. To assess whether this binding has a direct impact on APR aggregation kinetics, we constructed a set of folding-incompetent model proteins consisting of the same peptides studied in isolation above, fused to the aminoterminal of GFP (Fig 6B). By separating the APR from GFP by a rigid linker, we hope to minimize confounding effects arising from direct interaction between the APR and the GFP moieties. The main confirmation that our APRs do not majorly interfere with GFP folding is derived from the fluorescence of these constructs, which require correct three-dimensional assembly of the chromophore and hence the entire fold. Since the APR in our construct is isolated from its hydrophobic core, it cannot be removed from the solvent by native protein folding or even hydrophobic collapse, thereby ensuring continued exposure of the APR and its flanking GKs to the cellular environment, including the chaperones. In this way, we believe our model allows to study the effect of GKs and their interactions in the denatured state, which would otherwise be too short-lived to study reliably. To assess intrinsic solubility, our model constructs were first expressed in an *in vitro* (cell-free) translation setup, following the protocol established by Niwa *et al* (2009). This system is completely devoid of proteostatic machinery components, allowing for a clean interpretation of solubility readouts. These analyses show GFP fused to APRs flanked with positive GKs to be very poorly soluble, while, in line with the peptide results discussed above, the negative GKs strongly diminish the aggregation tendency instilled by the APR, rendering the constructs as soluble as GFP without an APR fused to it (Fig 6C and D). Interestingly, flanking the APR with Ala (AA construct) severely diminishes expression levels, confounding direct comparison since aggregation is strongly concentration-dependent. Supplementing the reaction with the DnaK chaperone system—consisting of DnaK, DnaJ and GrpE—strongly mitigates the observed solubility differences between the charged GKs, showing that chaperones can compensate for the intrinsic weakness of the positive GKs. The non-gatekept AA construct is unresponsive to DnaK supplementation in this assay, both in terms of solubility and expression levels.

We subsequently cloned the folding-incompetent constructs into a suitable vector for expression in *E. coli* K12, to assess solubility *in cellulo*, thus in the presence of the full proteostatic machinery. These experiments showed that the differences in solubility between positively and negatively gatekept constructs in cells are less outspoken than in an *in vitro* context, suggesting a buffering effect coming from the proteostatic machinery. Still, a trend is retained that GFP constructs with positive GKs yield less soluble protein than the negatively charged constructs (Fig 6E and F). To further address this, we expressed the model proteins in a DnaK knockout *E. coli* line, as well as in the wild-type clone with overexpression of the DnaK-DnaJ-GrpE system. In this setup, the non-gatekept AA constructs are responsive to DnaK levels, decreasing in solubility upon DnaK deletion and increasing when DnaK is overexpressed. Strikingly, Arg-gatekept constructs are even more responsive to DnaK levels, decreasing in soluble expression levels upon DnaK depletion and increasing in soluble expression to levels similar to those of GFP under DnaK overexpression conditions. Lys-gatekept constructs also strongly increase in solubility upon DnaK overexpression, although reduction in soluble expression is not detected in the DnaK knockout condition. Glu-gatekept constructs benefit from DnaK overexpression, while Asp constructs show no response to either DnaK knockout or overexpression. Super-resolution fluorescence microscopy further revealed that non-gatekept, positively gatekept and Glu-gatekept constructs form fluorescent inclusion bodies, indicating that aggregation events are driven by the APRs, with minimal effects on GFP folding (Fig 6G). For the Asp-gatekept construct, no such inclusions were observed, consistent with the *in vitro* observation that these GKs are very effective in preventing aggregation. Addition of DnaK partly solubilized the inclusion bodies, as shown by increased diffuse GFP fluorescence, confirming the solubilization observed in Fig 6E and F.

In order to determine the molecular basis for the stronger DnaK binding of basic residues, we performed a stability analysis on a set of structures of the DnaK substrate-binding domain (SBD) bound to a peptide substrate, generated by Zahn *et al* (2013). Through the FoldX force field, we introduced GK mutations into the bound substrate peptides and assessed the effects on the stability of the complex (Fig 7A). Intriguingly, DnaK binding and aggregate structure destabilization show a remarkably similar energetic profile (Figs 7A and 4B). DnaK binding to basic residues is energetically favourable, while acidic residues destabilize the interaction. These differences mainly stem from increased

stabilization through the burial and van der Waals packing of hydrophobic atoms in basic residues, as well as electrostatic interactions and sidechain hydrogen bonding. Figure EV1 offers a more detailed breakdown of these energetic contributions for each amino acid position in the substrate (defined relative to the DnaK inner binding pocket as indicated in Fig 7B). Strikingly, the interaction energy for acidic residues is worse than that of basic residues in each position, aside from the binding pocket (position 0). In this position, only Lys seems to be tolerated, largely due to the burial of all its hydrophobic atoms, as well as a strong hydrogen-

bond formation (as depicted in Fig 7D). The substrate-wide stabilization of the interaction with basic residues is mainly governed by favourable burial and packing of hydrophobic atoms (reflected in hydrophobic desolvation and van der Waals packing). The reason for this is evident from Fig 7B, which shows an example of the conformation of each of the four charged GKs in each position along a DnaK-bound substrate in the 1dkz PDB-structure. In its closed conformation, DnaK forms a hydrophobic binding tunnel, in which the substrate peptide is completely buried. As discussed before, basic residues are more amenable to burial in a

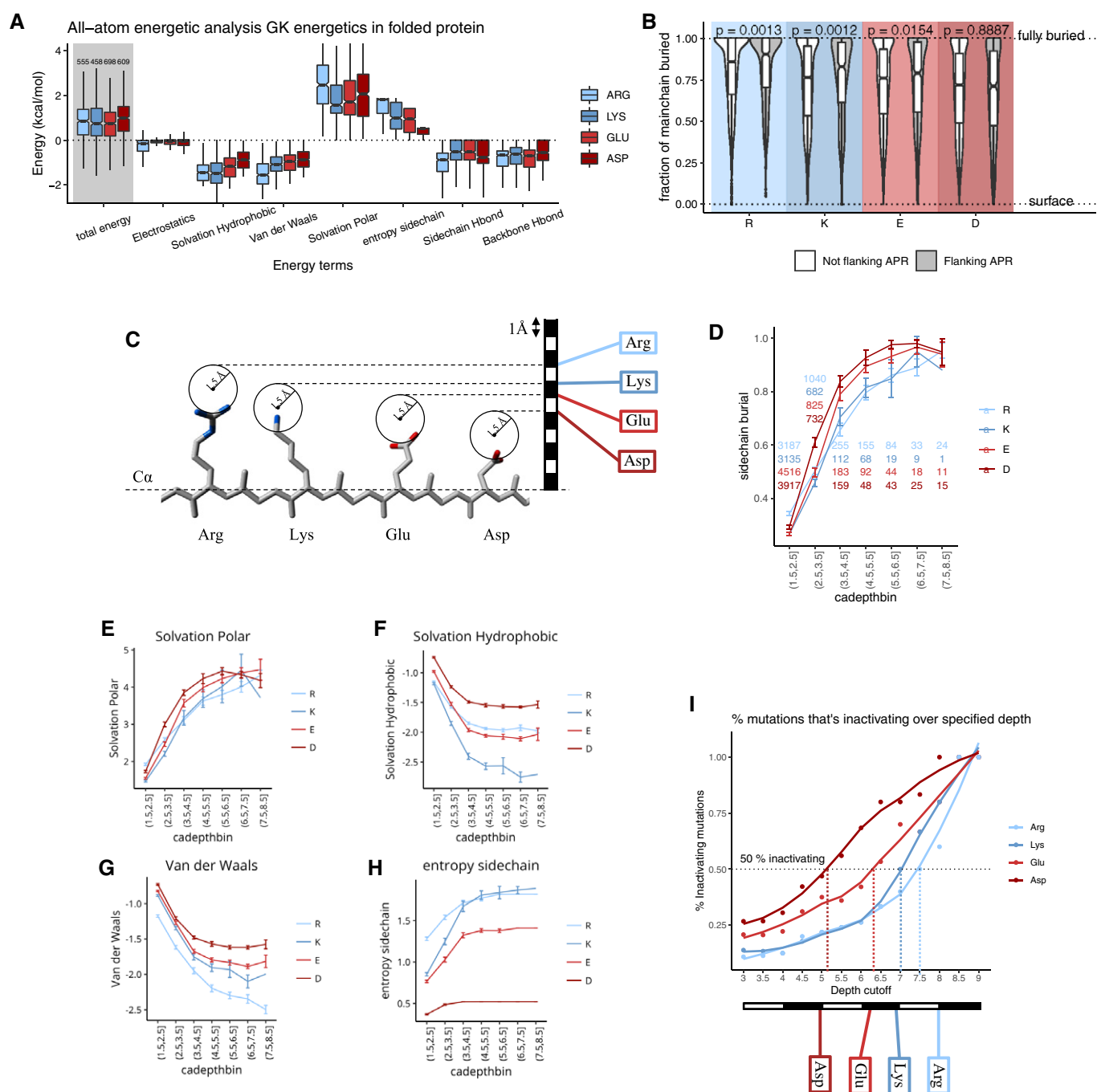


Figure 5.

Figure 5. Structural constraints on the usage of GKs in the native fold.

- A Boxplots showing all-atom analysis of GK energetic contributions to a set of native folds from the *Escherichia coli* cytoplasmic proteome. The upper whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least $1.5 * \text{IQR}$ from the lower hinge. Notches extend $1.58 * \text{IQR}/\sqrt{n}$ from either side of the median and represent a 95% confidence interval for the median value. Grey shaded area indicates total energy associated with the placement of GKs in the native fold, and the remainder of the plot shows the breakdown of this total energy into its constituent energy terms. The number of observations per GK is indicated above the boxes depicting the distribution of the total energy for each of the GKs.
- B Burial analysis on a set of *E. coli* structures of cytoplasmic, monomeric proteins. Violin plots and boxplots indicate distribution of the degree of mainchain burial for each of the charged amino acids, grouped by their usage as GK. The upper whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least $1.5 * \text{IQR}$ from the lower hinge. Notches extend $1.58 * \text{IQR}/\sqrt{n}$ from either side of the median and represent a 95% confidence interval for the median value. Statistical differences between charged residues flanking APRs versus charged residues elsewhere in the protein were determined through Wilcoxon test. *P*-values are indicated.
- C Schematic representation of the approximate sidechain lengths of the four charged residues. Circles indicate a probe radius of 1.5 Å used here for the determination of distance from the solvent-accessible surface.
- D Fraction of sidechain buried versus α depth bin for all charged residues in the dataset. Mean values are plotted, and whiskers indicate edges of a 95% confidence interval for the population mean based on a t-distribution. The number of observations per group is indicated.
- E–H Dissection of energetic components with increasing mainchain depths for each charged residue. Mean values are plotted, and whiskers indicate edges of a 95% confidence interval for the population mean based on a t-distribution. The number of replicates per group is indicated. The numbers of observations per group are the same as those indicated in panel (D).
- I Analysis of an exhaustive mutational fitness screen conducted by Varadarajan and colleagues (Bajaj *et al.*, 2005). The percentage of mutations found to be inactivating is plotted against the depth of burial in the protein core for each of the charged residues. The dotted line indicates where charge introduction becomes predominantly destabilizing. Sidechain lengths derived from panel (C) are indicated on the x-axis.

hydrophobic core than acidic residues due to their long, hydrophobic sidechains. In fact, our modelling clearly shows that basic residues, while burying their hydrophobic atoms, are more capable of placing their polar atoms outside of the binding tunnel, thereby stably binding DnaK. Acidic residues on the other hand do not have as many hydrophobic atoms to bury and are less capable of reaching the solvent-accessible surface with their polar atoms, yielding a very unstable interaction with DnaK. On top of these burial elements, electrostatics for basic residues are also generally favourable, albeit more so on the C-terminal side of the substrate. On this side of the DnaK SBD, several acidic residues are available for salt bridge formation with basic residues, as shown in Fig 7C.

Taken together, these data show that DnaK binding is driven by the same energetics that confer poor aggregation-breaking characteristics to basic residues, i.e. favourable burial of their longer sidechains. Furthermore, the interaction is stabilized through specific electrostatic interactions as well as hydrogen bonding. This suggests that DnaK evolved to recognize poorly aggregation-breaking GKs by their longer sidechain lengths as well as their charged moiety. The modelling data also explain why Glu,

although it is not a strong binder, does benefit from DnaK overexpression more than Asp, as it has a longer sidechain and therefore more beneficial burial characteristics.

DnaK substrates are enriched in poorly protected APRs

All of the above suggests that the charge specificity of DnaK, and by extension other chaperones, serves as a compensation mechanism for the ineffective gatekeeping capacity of basic residues. In other words, chaperones recognize vulnerable APRs by their lack of effective gatekeeping. The question then arises whether chaperone substrates are enriched in APRs that are ineffectively gate-kept. To analyse this, we used a chaperone dependence database devised by Niwa *et al.* (2012) in a follow-up study on the *in vitro* solubility database we used for the analysis in Fig 3I. In short, the authors performed *in vitro* translation in the presence of molecular chaperones on a set of proteins determined to be insoluble in the initial database. We analysed whether proteins that were found to be DnaK-dependent in this assay are enriched in APRs that are not protected by any negative charges. We indeed found a significant enrichment for such APRs in DnaK substrates

Figure 6. DnaK compensates for the inferior gatekeeping potential of basic residues.

- A DnaK binding to cellulose-bound peptide array. Sequences of the aggregating cores are indicated in the top of the panel, and the GKs flanking them in each instance are indicated above the blots.
- B Schematic representation of folding-incompetent model protein. The β -Gal APR is flanked on either side by one of the charged GKs, and separated from the GFP moiety by a rigid linker. Additionally, the construct is C-terminally His-tagged.
- C Representative Western blot of *in vitro* translation analysis of folding-incompetent model constructs. Bands show total (T) and soluble (S) fractions after centrifugation at 21,000 *g* for 30 min. “+DnaK” indicates addition of DnaK mix—containing DnaK, DnaJ and GrpE.
- D Bar plots showing mean solubility as determined through quantification of Western blots in (C), whiskers indicate standard deviation ($n = 4$). Statistically significant differences between the –DnaK and +DnaK conditions were determined through two-way ANOVA with Bonferroni’s correction for multiple comparisons. *P*-values for significant differences are indicated.
- E Representative Western blot of the solubility analysis of GK variants upon overexpression in *Escherichia coli* K12. Bands show soluble (Sol) and pellet fractions after centrifugation at 17,100 *g* for 15 min. “–DnaK” indicates expression in *E. coli* K12 DnaK knockout line, and “+DnaK” indicates co-expression of pKJE7, encoding DnaK, DnaJ and GrpE.
- F Barplots showing mean solubility as determined through quantification of Western blot bands in (E), whiskers indicate standard deviation ($n = 7$, except for the AA construct, where $n = 4$). Statistical significance was determined through two-way ANOVA followed by Tukey’s multiple comparisons test. *P*-values are indicated.
- G Structured illumination microscopy (SIM) images of *E. coli* after 3 h of expression of one of the GK variants. GFP fluorescence is shown in green. “ΔDnaK” indicates expression in the *E. coli* K12 DnaK knockout line, “+DnaK” indicates co-expression of pKJE7, encoding DnaK, DnaJ and GrpE.

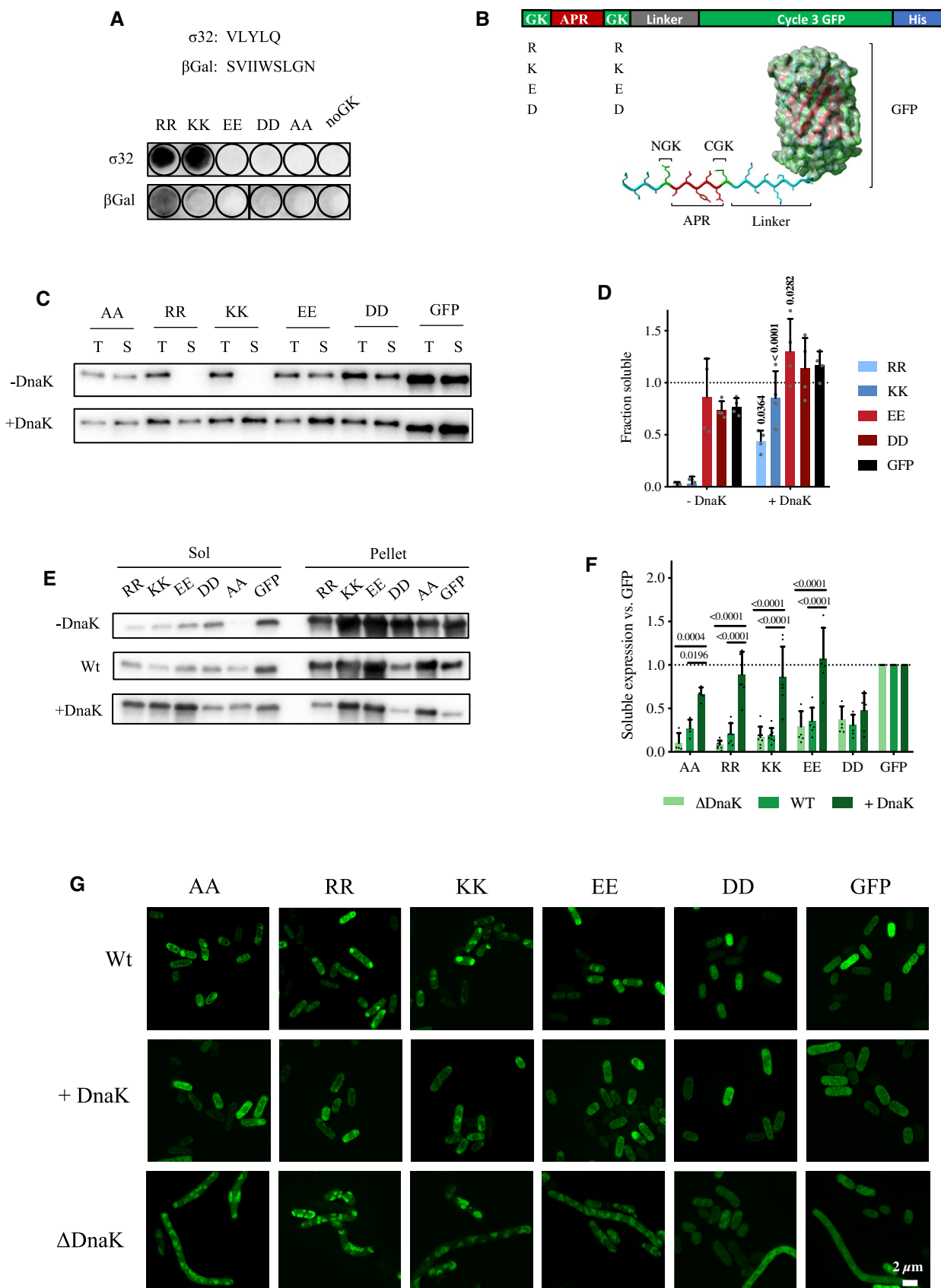


Figure 6.

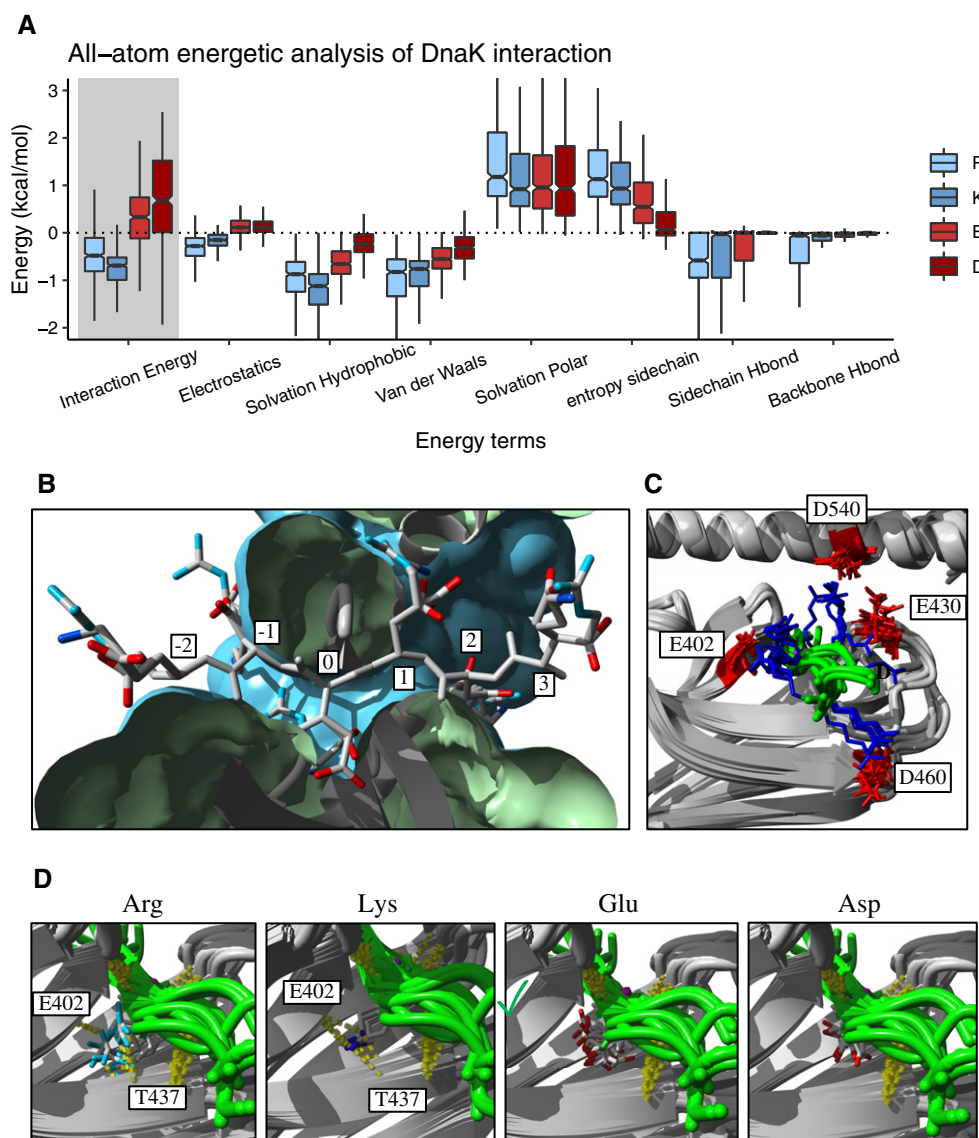


Figure 7. Energetic dissection of GK–DnaK interaction.

- A** Boxplots showing all-atom energetic analysis of the introduction of charged GKs into the DnaK–substrate interaction structures. The upper whiskers indicate the largest value no further than 1.5 times the inter-quartile range (IQR) from the upper hinge, and the lower whiskers show the smallest value at least $1.5 \times \text{IQR}$ from the lower hinge. Notches extend $1.58 \times \text{IQR}/\sqrt{n}$ from either side of the median and represent a 95% confidence interval for the median value. Grey shaded area indicates total energy associated with the placement of GKs into substrate peptides, and the remainder of the plot shows the breakdown of this total energy into its constituent energy terms. The number of replicates per box is 561.
- B** Example of the modelled conformation of each of the 4 charged residues in each position along the DnaK substrate. Variants derived from PDB-structure 1dkz are shown.
- C** Overlay of salt bridges between basic residues (blue) in DnaK substrates (green) and negatively charged residues (red) on the DnaK surface (grey). Positions of the negatively charged residues are indicated.
- D** Hydrogen-bond (yellow dashed lines) formation by Arg, Lys, Glu and Asp in the DnaK-binding pocket (position 0) with the DnaK backbone (grey). Hydrogen-bond acceptor residues are indicated.

(Fig 8A). Moreover, this enrichment significantly scales with increased PureTANGO score of the APRs (P -value = 0.0084), indicating that DnaK dependence is correlated with poorly protected aggregation potential. As a control, the same analysis was performed, this time regarding APRs that are not protected by any positive charges to be unprotected. However, these were not significantly enriched in DnaK substrates.

The evolutionary link between basic GKs and molecular chaperones

Our results suggest co-evolution between aggregation GKs and molecular chaperones and suggest the possibility that the evolutionary advantage resulting from the functionalities of complex protein folds and larger proteins was facilitated by the introduction of GKs that are

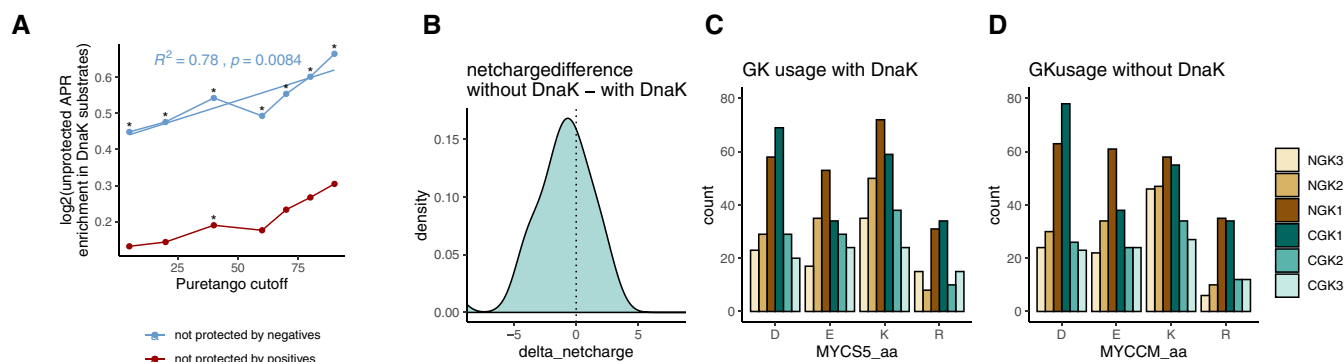


Figure 8. How DnaK dependence and GK usage are linked.

- A Enrichment of APRs not protected by a specific charge in DnaK substrates, as determined by Niwa *et al* (2012), analysed versus PureTANGO score of the APR considered. Statistical significance was determined through hypergeometric testing, with Bonferroni correction for multiple testing. **P*-value < 0.05. *P*-values and Pearson correlation coefficient (R^2) for linear regression are indicated.
- B Distribution of the difference in net charge between orthologs from a *Mycoplasma* species devoid of DnaK (*Mycoplasma crocodyli*, indicated as MYCCM), and its closest DnaK-expressing relative (*Mycoplasma synoviae*, indicated as MYCS5).
- C GK usage in DnaK-expressing *M. synoviae* orthologs.
- D GK usage in DnaK-depleted *M. crocodyli* orthologs.

more compatible with burial, but therefore also less potent aggregation breakers. This in turn may have sparked the specialization of molecular chaperones to pick out basic residues in a hydrophobic context, thereby homing in on the most vulnerable parts of folding proteins. Although definitively proving this evolutionary causation is impossible, we here provide some evidence to support this hypothesis. First off, in a phage-display study conducted by Jespers *et al* (2004), domain antibodies were selected for their aggregation resistance. The resulting aggregation-resistant antibodies had accumulated acidic residues in their unstructured regions, suggesting that, in the absence of molecular chaperones and when structurally feasible, evolution favours acidic GKs. We further wondered whether the evolutionary loss of DnaK would affect GK usage. To assess this, we analysed a species of *Mycoplasma* that has been shown to have lost DnaK expression, namely *Mycoplasma crocodyli* (Vishnyakov & Borchsenius, 2013). We identified a set of orthologous proteins between this species and a close DnaK-expressing relative, *Mycoplasma synoviae*. Interestingly, orthologs from the species lacking DnaK have indeed shifted their net charge towards the negative side (Fig 8B). Furthermore, they show a reduced usage of positive GKs, especially Lysine, suggesting some of the evolutionary advantage of basic GKs is lost upon DnaK depletion (Fig 8C and D).

Discussion

Here, we show how the preference of the protein quality control machinery for basic residues is most likely a consequence of fundamental structural and evolutionary constraints of globular structure. Classical globular structure formation in an aqueous environment requires the tertiary packing of hydrophobic primary sequence segments into a hydrophobic core, accompanied by the formation of secondary structure elements such as β -sheets. Inevitably, these prerequisites lead to the existence of structurally frustrated sequence segments with an intrinsic capacity to not only contribute to the native fold but also engage in β -aggregation, here referred to as

APRs. In order to kinetically favour folding to the globular native state over amyloid-like assembly, APRs in globular proteins are systematically N- and C-terminally capped by charged residues. These aggregation GKs favour native folding by kinetically disfavouring amyloid-like assembly through charge repulsion. However, GK placement comes at a cost as, apart from kinetically disfavouring the aggregated state, they also thermodynamically destabilize the native fold (De Baets *et al*, 2014) and slow the folding reaction. Indeed, it has been shown that folding without GKs occurs faster and yields more thermodynamically stable protein, albeit with a higher risk of aggregate formation (Kurnik *et al*, 2012). In the work presented here, we uncover fundamental differences between charged residues in terms of their destabilization of both the folded structure and the aggregated states (Fig 9A). Although all charged residues kinetically disfavour aggregation, we show acidic GKs to be the better aggregation breakers intrinsically, destabilizing the aggregated state both kinetically and thermodynamically by virtue of their shorter sidechains. Basic residues, on the other hand, kinetically hinder the aggregation process but barely destabilize the aggregated state, thereby still allowing amyloid formation to occur and *ad infinitum* leading to largely insoluble species (Fig 3B–I and Appendix Figs S1–S10). Conversely, the same characteristics that make acidic GKs more potent also render them largely incompatible with burial in the native structure, while basic residues are more amenable to native fold incorporation. By specifically homing in on positively charged residues, many proteostatic components offer a solution to this apparent paradox between aggregation and folding. The Hsp70 SBD, for example, mimics burial in the folded state by enveloping its substrate peptides and therefore recognizes basic residues by the exact characteristics that make them inferior aggregation gatekeepers (leading to the strikingly similar energetic signatures in Figs 4B and 7A). Well-placed negative charges on the SBD further promote identification of APRs poorly protected by their positive flanks. In this way, molecular chaperones cause an apparent shift in unfolded-to-aggregated state transition energies, effectively compensating for the differences in aggregation-breaking

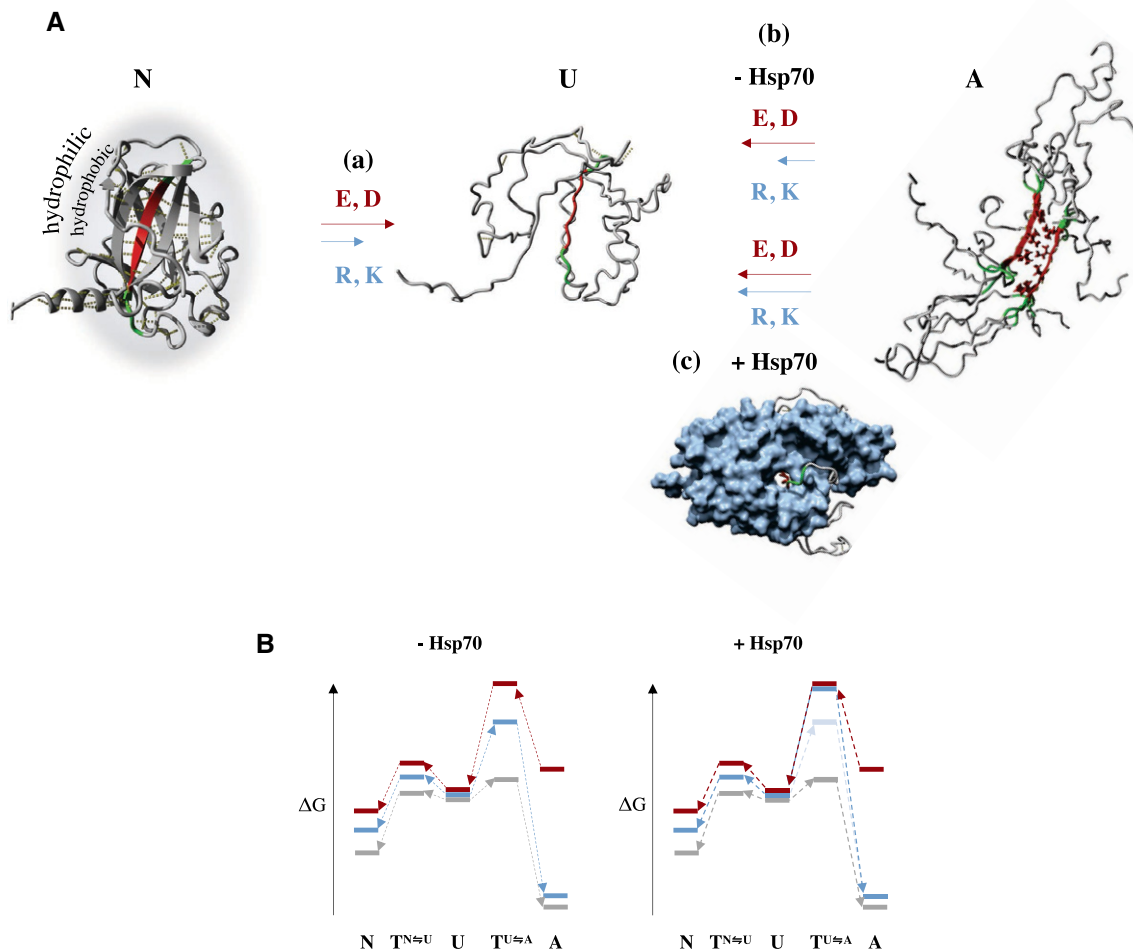


Figure 9. Energetic model of the combined effects of aggregation GKs and molecular chaperones.

N = Native fold, U = Unfolded state, A = Aggregated state, T = Transition state.

A Effects of GK placement on the energetics of the native, unfolded and aggregated states. GKs generally come with a cost to the native fold stability, with negative GKs being more destabilizing on average owing to their shorter sidechains and resulting increased burial costs (a). On the other hand, acidic residues more effectively destabilize the aggregated state than basic residues (b). Consequently, structural constraints in the native state may lead to APRs that are inadequately protected by their basic GKs. Hsp70 specifically recognizes these APRs through their inadequate GKs, and effectively compensates for the difference in destabilization of the aggregated state between basic and acidic residues (c).

B Energetic landscape of the model in (A). Acidic GKs are indicated in red, basic GKs in blue and the non-gatekept energy landscape in grey. GKs thermodynamically destabilize and kinetically slow down the folding reaction, resulting in higher energies for N and $T^{N \rightarrow U}$. Acidic residues have a similar effect on the aggregated state, decreasing aggregate formation kinetics and thermodynamic stability. Basic residues, on the other hand, only marginally affect aggregate stability and mostly work through deceleration of the aggregation process through an increase in $T^{U \rightarrow A}$ energy. Through specifically binding positive residues, Hsp70 slows aggregate formation kinetics for positively gatekept APRs. Hence, Hsp70 increases GK potential of basic residues, preventing aggregation and allowing for a stable native fold.

potency between basic and acidic residues (Fig 9B), as exemplified by the almost complete restoration of the balance in the amount of folded protein shown in Fig 6D and F. Our findings on the potency of acidic GKs rationalize work from other groups, including that of Jespers *et al*, who found through a phage-display setup that domain antibodies selected for their aggregation resistance are strongly enriched in acidic residues (Jespers *et al*, 2004).

To ensure interpretability of our findings, most of the work regarding GK potency and DnaK binding was performed on peptides flanked on both sides by the same GK, allowing us to draw conclusions on the four charged residues independently. Obviously, this does not fully cover the scope of GK patterns found in nature, and

the effects of different GK combinations on protein solubility and chaperone dependence remain to be explored further.

Interestingly, it has been proposed that early in evolution, protein sequence space was limited to combinations of a reduced amino acid repertoire, which gradually evolved into the now-canonical 20-amino acid system. Acidic residues have been proposed to belong to the low-complexity original alphabet, both because of their relative chemical simplicity and because of their role as precursors in the biosynthesis of more complex amino acids, including the basic residues Arg and Lys (Crick, 1968; Wong, 1975). Random protein design efforts from such a reduced alphabet, consisting of only Val, Ala, Gly, Asp and Glu, showed that, at similar levels of hydrophobicity,

the reduced-alphabet system produced more soluble protein than the 20-amino acid system (Doi *et al*, 2005). This again indicates that the negatively charged amino acids more effectively solubilize proteins, for which we here provide a mechanistic explanation. Given all of the above, we believe our findings hint at significant events in protein evolution, as the introduction of chemically complex basic residues gives rise to GKs that are more compatible with extended hydrophobic cores while offering rudimentary protection against aggregation, allowing for an exploration of expanded structure space. Finally, the co-evolution of molecular chaperones to specifically recognize positively charged GKs consolidates them as truly effective aggregation breakers, expanding their use in complex folds and highly abundant proteins (Ramakrishnan *et al*, 2019).

Apart from its evolutionary implications, our work exposes universal, fundamental rules governing protein architecture and folding. This information is key in understanding proteostasis and by extension proteostatic disorders, as well as in establishing rules for the rational design of protein sequences and expression conditions for bioproduction.

Materials and Methods

APR prediction and PureTANGO calculation

All APR and GK predictions were performed using TANGO, a statistical mechanics algorithm for the prediction of aggregation-prone stretches based on primary sequence (Fernandez-Escamilla *et al*, 2004). Using this method, the entire *E. coli* proteome (UniProt *E. coli* K12 reference proteome UP000000625; The UniProt Consortium, 2018) was scanned for APRs and the GKs flanking them. TANGO returns a score for each APR, which quantifies its aggregation propensity. As this score is influenced by the residues flanking the APR, it does not reflect the intrinsic strength of the APR core in the absence of the protecting GKs. For this reason, all APR cores were artificially flanked with Ala, which has negligible effects on aggregation propensity, and reanalysed with TANGO. The resulting APR strengths are direct measures of the aggregation tendency of the APR core, without any interference from GKs. This new APR strength is referred to as the PureTANGO score.

GK effect analysis

In order to determine the effect of GKs on APR strength, cytoplasmic APRs were identified by cross-referencing the APR list determined above with the StepDB subcellular location database (Orfanoudaki & Economou, 2014). Each GK in the original proteome sequences was then mutated to Ala, after which TANGO scores for each protein were recalculated. Finally, the differences between the original sequence and the protein sequence with a single GK mutated to Ala were calculated, and the resulting value is referred to here as the GK effect.

Escherichia coli conservation analyses

Non-synonymous coding rates were obtained from Martincorena *et al* (2012). For each of the charged amino acids, the deviation in

non-synonymous mutation rate for residues in GK positions versus the genome average for that amino acid was calculated using the following formula:

$$\% \text{ deviation from average} = \frac{\text{rate in GK positions} - \text{average rate}}{\text{average rate}}$$

Mammalian conservation analyses

Mammalian orthologous protein clusters were obtained from the OMA browser (Altenhoff *et al*, 2017). Multiple alignments were made using the MAFFT algorithm (using the FFT-NS-2 algorithm with options `–retree 2—reorder`), and orthologs with at least 50% sequence identity were retained for further analysis. Membrane proteins were filtered out through UniProt subcellular location annotation. Consensus sequences and sequence conservation were calculated using the *conserv()* and *consensus()* functions from the R package bio3D (Grant *et al*, 2006). Next, in line with the work of Martincorena *et al* (2012), residues with 100% sequence identity were considered conserved, while residues with less than 100% identity were considered to undergo non-synonymous mutations. The deviation in non-synonymous mutation rate for residues in GK positions versus the genome average for that amino acid was then calculated using the following formula:

$$\% \text{ deviation from average} = \frac{\text{rate in GK positions} - \text{average rate}}{\text{average rate}}$$

Enrichment scores for GKs in strong versus weak APRs

From the set of APRs from *E. coli* cytoplasmic proteins described above, the 30% strongest and 30% weakest APRs according to their PureTANGO score were selected. For each of these groups, the occurrence of each of the charged GKs was determined as the proportion of all APRs flanked with that specific GK. The ratio of proportions in strong versus weak APRs was then calculated, and a logarithm of this ratio yielded a log-odds score. This value is positive in case of an enrichment and negative in case of a depletion of a certain GK in the flanks of strong APRs. To determine statistical significance, a bootstrapping approach was used to calculate 95% confidence intervals for the null-hypothesis that the proportion of GKs in strong versus weak APRs are equal. Any value outside such a confidence interval is considered statistically significant.

Peptide set design

To construct a set of aggregating peptides for solubility analyses, APRs were selected from the set of cytoplasmic *E. coli* proteins described above. To ensure adequate aggregation propensity, only APRs with a TANGO score over 60 were selected, with a maximum length of eight amino acids. To facilitate accurate concentration determination of peptides through absorbance measurements at 280 nm, only APRs containing Trp and/or Tyr were considered. GK variants of each APR flanked at both ends by one of the four charged residues, i.e. Asp, Glu, Arg and Lys, were produced as well as control peptides with either Ala in the GK positions (AA), or no gatekeeping residues at all (NoGK).

Peptide aggregation analyses

All peptides were ordered from Genscript in an N-terminally acetylated and C-terminally amidated form. Peptides were dissolved in DMSO, after which these stocks were sonicated for 20 min and filtered through 0.2- μ m PVDF filters. Stock concentrations were determined using a NanoDrop 2000 (Thermo Fisher Scientific). Stocks were then diluted to the appropriate peptide concentration (100 μ M unless stated otherwise) in PBS (unless stated otherwise) with a final concentration of 5% DMSO. For ThT- and pFTAA-binding kinetics, 10 μ M ThT or 1 μ M pFTAA was added to the peptide samples. Dye binding was measured over time through excitation at 440 nm and emission at 480 and 520 nm, for ThT and pFTAA, respectively. Measurements were performed in a Fluostar OMEGA and Fluostar fluorescence plate reader (BMG LABTECH), respectively, at 25°C with 5 s of shaking before each readout. Dynamic light scattering analyses for time-resolved particle size measurements were performed at 25°C using a Wyatt DynaPro Plate Reader I. For endpoint concentrations, peptide preparations were left at room temperature for 7 days and subsequently subjected to ultracentrifugation at 76,000 g for 1 h at 4°C, after which supernatant concentrations were measured using a NanoDrop 2000 (Thermo Fisher Scientific).

TEM imaging

TEM images were taken after 3 days' incubation of the peptide samples prepared as described above. Formvar film-coated 400-mesh copper grids (Agar Scientific Ltd., UK) were glow-discharged to improve adsorption efficiency. Next, 10 μ l of each sample was adsorbed for 10 min after which the grids were washed by contact with one drop of ultrapure water. Negative staining was performed by contact with one drop of uranyl acetate (1% w/v) for 1 min. This was followed by three additional wash steps by contact with three drops of ultrapure water, after which grids were air-dried and stored in a vacuum desiccator. Grids were examined using a JEM-1400 transmission electron microscope (Jeol, Japan) at 80 keV.

Proteome-wide *in vitro* protein solubility and DnaK dependence dataset analysis

The proteome-wide protein solubility and DnaK dependence datasets were obtained from studies conducted by Niwa *et al* (2009, 2012). The set was restricted to cytoplasmic proteins through cross-referencing with the StepDB database (Orfanoudaki & Economou, 2014). APRs, flanking GKs and PureTANGO scores were determined for all proteins in the dataset as described above. Proteins were then classified on whether they contain at least one strong APR (PURETango > 70) that has no acidic in its three N-terminal or C-terminal flanking positions (referred to as “unprotected” APR). Solubility differences between groups that have such an APR versus background were then assessed. As a control, the same analysis was performed, only now considering APRs that are not protected by at least one positive GK to be “unprotected”.

Next, proteins were classified as DnaK-dependent if their observed increase in solubility upon the addition of DnaK to the cell-free translation reaction exceeds 50%. The enrichment of APRs that are not flanked by any negative GKs (“unprotected” APRs) in DnaK substrates versus DnaK-independent proteins was assessed.

Enrichment scores were determined using the following equation:

$$\text{Enrichment} = \log_2 \left(\frac{\% \text{ proteins with unprotected APR in DnaK – dependent group}}{\% \text{ proteins with unprotected APR in DnaK – independent group}} \right)$$

These enrichments were calculated at increasing PureTANGO cutoffs (i.e. PureTANGO scores above which APRs are considered in the analysis). Statistical significance was determined through hypergeometric testing with Bonferroni correction for multiple comparisons. As a control, the same analysis was performed, this time regarding APRs that are not flanked by at least one positive GK to be “unprotected”.

Zipper structure modelling

In order to assess destabilizing effects of GKs on steric zipper structures, the FoldX force field was used (Guerois *et al*, 2002; Schymkowitz *et al*, 2005). A set of 68 known steric zipper structures was compiled to test GK effects on zipper stability. If a PDB file contained several zipper modes, these were deconvoluted into separate structure files. Each of these structures was then adjusted to contain 10 molecules (two interdigitating β -sheets made up of five identical strands) using YASARA (Krieger & Vriend, 2014). YASARA was also used for all other structure visualization in this study. Next, the structure energies were minimized using the FoldX “RepairPDB” command, after which both flanks of each molecule were mutated first to Ala, and then to one of the charged GKs using the FoldX “BuildModel” command. Stability effects of the latter mutations were then analysed.

To assess the distances between the charged moieties of the GKs introduced into the zipper structures, distances between the central-most atoms (“CZ”, “NZ”, “CD” and “CG” for Arg, Lys, Glu and Asp, respectively) were assessed using the BioPDB package from Biopython. To remove any edge effects (strand at the edges of a zipper would have more freedom of movement), only the innermost strands (E and F, as indicated in Fig 4A) were considered.

Native fold gatekeeper effect calculations

We queried the PDB for a set of monomeric cytoplasmic *E. coli* structures with a sequence identity below 70% and a resolution below 3 Å. This yielded approximately 400 PDB structures (Table 4). APRs and GK positions were determined using TANGO, energies of the structures were minimized using the FoldX “RepairPDB” command, and stability and burial calculations for each amino acid were performed using the “SequenceDetail” command. C_{α} and average residue distance from the solvent-accessible surface were determined using the *ResidueDepth* function from Biopython, which utilizes Michael Sanner's MSMS program.

Charged residue introduction analysis in the *Escherichia coli* CcdB toxin

Empirical data on the effects of the introduction of charged residues on the function of the *E. coli* CcdB protein were obtained from Bajaj *et al* (2005), as well as information on the backbone depth of the mutated residue. For each of the charged residues, the fraction of

Table 4. List of identifiers for the *Escherichia coli* PDB structures used in the native protein structure analyses.

PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID
5dmm	3cb4	4as1	3nbx	1aj2	3v97	1q2l	5cay
2h5e	2iif	1rqj	2mbr	1vht	1li5	1flz	4p32
3lfu	3c8f	1eso	1aOp	1ahn	2pjd	4rul	2z70
3avu	1mwi	2rn2	1tjl	1f76	2zpa	5mfr	1ako
1k4n	1x6j	2pth	1cke	1a04	1onr	3i49	3eps
1u0b	2vec	1ef9	4ari	4pe8	1qfj	2ggc	3i3m
1kon	1mla	1k0g	1gsq	2fdf	1vlo	3ss7	1o0b
2reb	1ujc	1nzj	1gg4	3zjt	1o9b	1poh	2fdi
1nij	3dh3	2fdh	1w78	1ks9	1jsx	5isv	4arc
3a7l	4jht	1jcl	1ksk	2ix0	1xwy	1ecr	2ewj
1p7t	1ew4	2x5o	4xpu	1r3f	3awi	3zgz	3agp
3cuz	5e1l	2x6t	1nyl	3zju	1e4y	4xr1	1g1b
1po9	1ni9	3kr6	3l8e	4aq7	2w70	4jxz	3h9c
1euy	1il2	1o0c	2olw	1c0a	1rc6	2iv2	3zjv
4s3r	2bon	4blu	1jyh	2fdg	4oby	4q43	4ir1
3avt	1gmX	1cib	1xEO	2olr	1lqa	4xr0	1szw
1e8c	2iie	1cde	2frx	1hzt	1g7v	1mjc	3bzn
1s03	2btd	4ir9	2zcu	2owo	1uqu	2fd8	1mpg
3uf7	2dxa	3r0d	2xva	2oug	1bs1	3rfa	5c5j
1z9t	4kaz	4dt4	1zjw	1sv6	4z8d	5hr3	1vi3
4r8u	1mwj	4h4d	1wxh	1v9f	1e5k	1i7h	2wu1
1qf6	1zyl	1y79	3qou	1x8m	5t3d	1uj8	1exd
1mug	2ar0	4jxx	4qnx	1ff3	1mzr	1k7j	1iov
1o65	5hr7	4xr3	1yt3	1k7k	4g9s	1zmr	4myd
4kb0	1grj	4kjj	5xm5	1pui	4i8o	1vly	1k8w
4zci	4dcm	3kgd	3ieu	2i06	1xm5	1vb3	1gts
3avw	4cqn	2fkb	1fmt	3a7r	3eye	1mul	1wq5
1gtk	2r4t	3i2o	1jbe	1w26	1diz	3cmw	3avx
1kag	4jyz	2azo	2obl	3ffv	3bf7	2i05	1fiy
2qcu	2fdk	1dm9	3lbf	1yix	1dfu	2pqx	1sqg
1sdi	2cxa	1qyr	3efp	2r1r	3avy		
1t8k	1qtq	2gmw	1w8g	3vnn	4auk		
4kb1	5hr6	1l5j	1nmn	1urh	1gtr		

instances of the introduction of that residue that were observed to be disruptive to protein function was calculated. This was then repeated systematically with increasing depth cutoffs, i.e. distance of the residue backbone to the solvent above which residues were considered.

DnaK purification

DnaK purification was performed using a slightly adapted version of the method described by Chang *et al* (2008). Briefly, DnaK was expressed in *E. coli* at 25°C overnight. Cell pellets were produced through centrifugation and resuspended in 25 mM Tris, 10 mM KCl, 5 mM MgCl₂, pH 7.5 (buffer A). Cells were disrupted in a French

press, and nucleic acids were degraded through sonication. Cleared cell extracts were then applied to a Q-Sepharose fast flow column (GE Healthcare), and protein was eluted with a gradient 10–500 mM KCl. Fractions containing DnaK were applied to ATP-agarose (Sigma) and eluted with 3 mM ATP after wash steps with buffer A and buffer A containing 1 M KCl. Finally, DnaK-containing fractions were run over a Superdex 200 gel filtration column (GE Healthcare). Purified protein was aliquoted, flash-frozen and stored at –80°C.

DnaK binding analysis to peptide array

Peptide arrays were produced through SPOT synthesis on acid-stable cellulose membranes with PEG spacer (Aims-Scientific) using

the Intavis Multiprep RSi synthesis robot. Peptides were synthesized from C-terminus to N-terminus, starting with a GGS linker. Membranes were incubated in 50% methanol for 10 min, followed by three short washes in TBS-T (TBS, 0.1% Tween-20). Next, membranes were blocked in 4% BSA in TBS-T for 1 h, followed by incubation with 100 nM DnaK in 25 mM Tris, 10 mM KCl, 5 mM MgCl₂, pH 7.5, 0.1% Tween-20 (buffer B) for 1 h at room temperature. Membranes were then washed in buffer B for three times 5 min, followed by incubation with HRP-linked His-tag antibody (BioLegend antibody 652504) for 1 h. Finally, membranes were washed in buffer B for three times 5 min, and developed through chemiluminescence using a ChemiDoc XRS (Bio-Rad).

In vitro translation and solubility analysis

In vitro translation assays were performed using the New England Biolabs Inc. PURExpress® *In Vitro* Protein Synthesis Kit. Where mentioned, reactions were supplemented with DnaK mix, obtained from Cosmo Bio as part of their PUREfrex® system. Both the PURExpress® and PUREfrex® systems are based on the PUREsystem™ devised by Shimizu *et al* (2001). Linear template DNA with a T7 promotor for expression in the PURExpress® system was produced through PCR according to the manufacturer's instructions. Protein solubility upon cell-free translation was determined as previously described by Taguchi and colleagues (Niwa *et al*, 2009). Briefly, cell-free translation was performed for 1 h at 37°C, with or without the addition of DnaK mix following the manufacturer's instructions, after which soluble and insoluble fractions were separated through centrifugation at 21,000 g for 30 min at 4°C. Total and soluble fractions were diluted 1:10 in 8 M urea to completely unfold and dissolve any protein produced, and protein levels in each fraction were determined through SDS-PAGE followed by Western blotting. Blots were developed using chemiluminescence after incubation with primary anti-GFP antibody (Cell Signaling Technologies antibody 2555S) and secondary HRP-conjugated antibody (Promega antibody W4011). Blots were quantified using Bio-Rad's Image Lab™ Software. Soluble expression was determined by calculating the ratio of soluble over total protein.

In cellulo expression and solubility analysis

Folding-incompetent model protein expression constructs were subcloned into the Thermo Fisher Scientific pBAD/Myc-His A vector under an arabinose-inducible promotor. The resulting vectors were transformed into *E. coli* K12 MG1655 (wild-type) or the *E. coli* DnaK knockout strain from the *E. coli* Keio Knockout Collection (Dharmacon). Where indicated, co-transformation with vector pKJE7 obtained from the Takara Bio was performed. This vector overexpresses DnaK, DnaJ and GrpE upon arabinose induction. For protein expression and solubility analyses, bacterial strains were grown overnight in Lysogeny broth (LB) supplemented with ampicillin for GFP expression and both ampicillin and chloramphenicol for co-expression of the GFP constructs with pKJE7. For expression of the DnaK knockout line, LB was supplemented with ampicillin and kanamycin. The overnight cultures were diluted 1:100 in fresh LB supplemented with the appropriate antibiotics and grown to an OD of about 0.6, after which expression was induced with 0.2% arabinose. Expression was allowed to

proceed for 3 h after which cells were lysed in B-PER™ reagent supplemented with 0.1 mg/ml lysozyme (Sigma-Aldrich), cOmplete™ Protease Inhibitor Cocktail (Sigma-Aldrich) and Pierce™ universal nuclease for cell lysis (ThermoFisher). Cells were lysed on ice for 30 min, after which soluble and insoluble fractions were separated through centrifugation at 17,100 g for 30 min at 4°C. Supernatant was removed and the insoluble fraction dissolved in an equal volume of 8 M urea. GFP in soluble and insoluble fractions was then quantified through SDS-PAGE followed by Western blotting. Blots were developed using chemiluminescence after incubation with primary anti-GFP antibody (Cell Signaling Technologies antibody 2555S) or anti-DnaK antibody (USBio D8076) and secondary HRP-conjugated antibody. Blots were quantified using Bio-Rad's Image Lab™ Software.

SIM imaging of fixed *Escherichia coli* cells

After 3 h of overexpression, cells were fixed by adding 2.5% PFA and 0.04% glutaraldehyde (final concentrations) to culture media, followed by incubation at room temperature for 15 and 30 min on ice. Cells were then washed in PBS, resuspended in GTE buffer and finally put on a glass slide for microscopy. SIM imaging was performed using the Zeiss Elyra S.1, controlled through the Zeiss ZEN 2012 software (black edition). GFP was excited using a 488 nm laser, and emission was captured through band-pass 495–550-nm filter. Images were taken at three grating angles, and the composite super-resolution image produced through the Structured Illumination image processing functionality in the Zeiss ZEN software.

DnaK interaction modelling

To analyse the effect of GK introduction on DnaK-substrate interaction, a set of PDB structures for DnaK bound to substrate peptide were taken from Zahn *et al* (2013). Using this set of 19 different structures offers variability in the substrate backbone. Structures were obtained from the PDB, and each DnaK-substrate pair was parsed into a separate structure file, yielding 31 different structures. The energy of the structures was minimized using the FoldX “RepairPDB” command, after which substrates were mutated to PolyAla using the FoldX “BuildModel” command, in order to minimize effects of the original sequence on the observed differences in stability. Next, individual amino acids in the substrate were mutated to each of the four charged GKs using the “BuildModel” command, and interaction energies between DnaK and substrate calculated through the “AnalyzeComplex” command. For each mutation, three repeats were performed, increasing the chances of sampling all possible conformations. Differences in interaction energy between the mutated substrate and the polyAla energy were then calculated and analysed through R statistical software.

Mycoplasma GK usage analysis

Protein orthologs between *M. synoviae* and *M. crocodyli* were obtained from the OMA orthology database (identifiers “MYCS5” and “MYCCM”, respectively) (Altenhoff *et al*, 2017). Orthologs were aligned using the MAFFT algorithm (Misawa *et al*, 2002) (with

“-globalpair -maxiterate 1000 -leavegapregion” options) and sequence identity calculated from these alignments. Orthologous proteins were analysed using TANGO to determine the position of APRs and flanking GKs. Orthologous groups were then filtered for those that had a sequence identity over 50%, leaving 170 orthologous groups. Next, GK usage was assessed for those APRs with a TANGO score above 30 in *M. synoviae* and was still identified as APRs in *M. crocodyli*.

Protein structure visualization

All protein structure visualizations were performed using YASARA Structure (Krieger & Vriend, 2014).

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Expanded View for this article is available online.

Acknowledgements

The Switch Laboratory was supported by grants from the European Research Council under the European Union's Horizon 2020 Framework Programme ERC Grant agreement 647458 (MANGO) to JS, the Flanders Institute for Biotechnology (VIB, grant no. C0401), the Hercules Foundation (AKUL/15/34—G0H1716N), Funds for Scientific Research Flanders (FWO, G051817N), the Flanders Agency for innovation by Science and Technology (IWT, SBO grant 60839) and the Belgian Science Policy Office (BELSPO, IAP contract P7/16). BH was supported by PhD Fellowship from the IWT (file nr. 141546). SIM microscopy was performed at the VIB Bioimaging Core at KU Leuven (LiMoNe), and TEM was performed at the Electron Microscopy Facility of VIB-KU Leuven. NL was funded by Fund for Scientific Research Flanders Post-doctoral Fellowship (FWO) [12P0919N to NL].

Author contributions

FR, JS and BH devised the study. BH performed most of the experimental work *in vitro* and *in cellulo*, as well as *in silico* analyses. FR, JS and BH wrote the manuscript. EM, MR and JV performed peptide aggregation analyses. MR produced cellulose-based peptide arrays. KK performed TEM analyses. NL compiled and curated the dataset for zipper structure stability assessments and assisted in DnaK-binding experiments. RK contributed a curated dataset from SCOPe for initial assessments of native fold stability. NC and TV performed the mammalian conservation analyses. RG synthesized peptides for preliminary assessments of non-natural GKs. MDV performed peptide synthesis and labelling.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Altenhoff AM, Gonnet GH, Train C-M, Dylus D, Glover NM, de Fariás TM, Dessimoz C, Warwick Vesztrocy A, Zile K, Stevenson C et al (2017) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 46: D477–D485

- Bajaj K, Chakrabarti P, Varadarajan R (2005) Mutagenesis-based definitions and probes of residue burial in proteins. *Proc Natl Acad Sci USA* 102: 16221–16226
- Beerten J, Jonckheere W, Rudyak S, Xu J, Wilkinson H, De Smet F, Schymkowitz J, Rousseau F (2012) Aggregation gatekeepers modulate protein homeostasis of aggregating sequences and affect bacterial fitness. *Protein Eng Des Sel* 25: 357–366
- Bose D, Chakrabarti A (2017) Substrate specificity in the context of molecular chaperones. *IUBMB Life* 69: 647–659
- Buell AK, Tartaglia GG, Birkett NR, Waudby CA, Vendruscolo M, Salvatella X, Welland ME, Dobson CM, Knowles TPJ (2009) Position-dependent electrostatic protection against protein aggregation. *ChemBioChem* 10: 1309–1312
- Chang L, Bertelsen EB, Wisén S, Larsen EM, Zuiderweg ERP, Gestwicki JE (2008) High-throughput screen for small molecules that modulate the ATPase activity of the molecular chaperone DnaK. *Anal Biochem* 372: 167–176
- Collins KD (1997) Charge density-dependent strength of hydration and biological structure. *Biophys J* 72: 65–76
- Crick FHC (1968) The origin of the genetic code. *J Mol Biol* 38: 367–379
- de Crouy-Chanel A, Kohiyama M, Richarme G (1996) Specificity of DnaK for Arginine/Lysine and effect of DnaJ on the amino acid specificity of DnaK. *J Biol Chem* 271: 15486–15490
- De Baets G, Van Durme J, Rousseau F, Schymkowitz J (2014) A genome-wide sequence-structure analysis suggests aggregation gatekeepers constitute an evolutionary constrained functional class. *J Mol Biol* 426: 2405–2412
- Deuring E, Patzelt H, Vorderwülbecke S, Rauch T, Kramer G, Schaffitzel E, Mogk A, Schulze-Specking A, Langen H, Bukau B (2003) Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Mol Microbiol* 47: 1317–1328
- Doi N, Kakukawa K, Oishi Y, Yanagawa H (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng Des Sel* 18: 279–284
- Döring K, Ahmed N, Riemer T, Suresh HG, Vainshtein Y, Habich M, Riemer J, Mayer MP, O'Brien EP, Kramer G et al (2017) Profiling Ssb-nascent chain interactions reveals principles of Hsp70-assisted folding. *Cell* 170: 298–311.e20
- Estácio SG, Leal SS, Cristóvão JS, Faísca PFN, Gomes CM (2015) Calcium binding to gatekeeper residues flanking aggregation-prone segments underlies non-fibrillar amyloid traits in superoxide dismutase 1 (SOD1). *Biochim Biophys Acta* 1854: 118–126
- Eswar N, Ramakrishnan C (2000) Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Eng Des Sel* 13: 227–238
- Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22: 1302–1306
- Flynn GC, Pohl J, Flocco MT, Rothman JE (1991) Peptide-binding specificity of the molecular chaperone BiP. *Nature* 353: 726–730
- Fourie AM, Sambrook JF, Gething MJH (1994) Common and divergent peptide binding specificities of hsp70 molecular chaperones. *J Biol Chem* 269: 30470–30478
- Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22: 2695–2696
- Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320: 369–387

- Hammarström P, Simon R, Nyström S, Konradsson P, Åslund A, Nilsson KPR (2010) A fluorescent pentameric thiophene derivative detects *in vitro*-formed prefibrillar protein aggregates. *Biochemistry* 49: 6838–6845
- Jaspers L, Schon O, Famm K, Winter G (2004) Aggregation-resistant domain antibodies selected on phage by heat denaturation. *Nat Biotechnol* 22: 1161
- Karagöz GE, Rüdiger SGD (2015) Hsp90 interaction with clients. *Trends Biochem Sci* 40: 117–125
- Karagöz GE, Acosta-Alvear D, Nguyen HT, Lee CP, Chu F, Walter P (2017) An unfolded protein-induced conformational switch activates mammalian IRE1. *Elife* 6: e30700
- Knoblauch NTM, Rüdiger S, Schönfeld H-J, Driessen AJM, Schneider-Mergener J, Bukau B (1999) Substrate specificity of the SecB chaperone. *J Biol Chem* 274: 34219–34225
- Krieger E, Vriend G (2014) YASARA view—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics* 30: 2981–2982
- Kurnik M, Hedberg L, Danielsson J, Oliveberg M (2012) Folding without charges. *Proc Natl Acad Sci USA* 109: 5705–5710
- Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L (2004) A comparative study of the relationship between protein structure and β -aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 342: 345–353
- Magalhaes A, Maigret B, Hoflack J, Gomes JNF, Scheraga HA (1994) Contribution of unusual Arginine-Arginine short-range interactions to stabilization and recognition in proteins. *J Protein Chem* 13: 195–215
- Markiewicz BN, Oyola R, Du D, Gai F (2014) Aggregation gatekeeper and controlled assembly of Trpzip β -hairpins. *Biochemistry* 53: 1146–1154
- Martincorena I, Seshasayee ASN, Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485: 95–98
- Mason PE, Neilson GW, Dempsey CE, Barnes AC, Cruickshank JM (2003) The hydration structure of guanidinium and thiocyanate ions: implications for protein stability in aqueous solution. *Proc Natl Acad Sci USA* 100: 4557–4561
- Misawa K, Katoh K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059–3066
- Monsellier E, Ramazzotti M, Taddei N, Chiti F (2008) Aggregation propensity of the human proteome. *PLoS Comput Biol* 4: e1000199
- Neves MAC, Yeager M, Abagyan R (2012) Unusual Arginine formations in protein function and assembly: rings, strings, and stacks. *J Phys Chem B* 116: 7006–7013
- Niwa T, Ying B-W, Saito K, Jin W, Takada S, Ueda T, Taguchi H (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc Natl Acad Sci USA* 106: 4201–4206
- Niwa T, Kanamori T, Ueda T, Taguchi H (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc Natl Acad Sci USA* 109: 8937–8942
- Orfanoudaki G, Economou A (2014) Proteome-wide subcellular topologies of *E. coli* polypeptides database (STEPdb). *Mol Cell Proteomics* 13: 3674–3687
- Otzen DE, Kristensen O, Oliveberg M (2000) Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci USA* 97: 9907–9912
- Patzelt H, Rüdiger S, Brehmer D, Kramer G, Vorderwülbecke S, Schaffitzel E, Waitz A, Hesterkamp T, Dong L, Schneider-Mergener J et al (2001) Binding specificity of *Escherichia coli* trigger factor. *Proc Natl Acad Sci USA* 98: 14244–14249
- Pickett SD, Sternberg MJE (1993) Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol* 231: 825–839
- Ramakrishnan R, Houben B, Rousseau F, Schymkowitz J (2019) Differential proteostatic regulation of insoluble and abundant proteins. *Bioinformatics* 35: 4098–4107
- Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2009) Protein sequences encode safeguards against aggregation. *Hum Mutat* 30: 431–437
- Rodríguez F, Arsène-Ploetze F, Rist W, Rüdiger S, Schneider-Mergener J, Mayer MP, Bukau B (2008) Molecular basis for regulation of the heat shock transcription factor σ 32 by the DnaK and DnaJ chaperones. *Mol Cell* 32: 347–358
- Rousseau F, Serrano L, Schymkowitz JWH (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* 355: 1037–1047
- Rüdiger S, Germeroth L, Schneider-Mergener J, Bukau B (1997) Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J* 16: 1501–1507
- Rüdiger S, Schneider-Mergener J, Bukau B (2001) Its substrate specificity characterizes the DnaJ co-chaperone as a scanning factor for the DnaK chaperone. *EMBO J* 20: 1042–1050
- Sant'Anna R, Braga C, Varejao N, Pimenta KM, Grana-Montes R, Alves A, Cortines J, Cordeiro Y, Ventura S, Foguel D (2014) The importance of a gatekeeper residue on the aggregation of transthyretin: implications to transthyretin-related amyloidoses. *J Biol Chem* 289: 28324–28337
- Schlieker C, Weibezahn J, Patzelt H, Tessarz P, Strub C, Zeth K, Erbse A, Schneider-Mergener J, Chin JW, Schultz PG et al (2004) Substrate recognition by the AAA+ chaperone ClpB. *Nat Struct Mol Biol* 11: 607–615
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* 33: W382–W388
- Shimizu Y, Inoue A, Tomari Y, Suzuki T, Yokogawa T, Nishikawa K, Ueda T (2001) Cell-free translation reconstituted with purified components. *Nat Biotechnol* 19: 751
- Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev* 37: 1395–1401
- The UniProt Consortium (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46: 2699
- Trevino SR, Scholtz JM, Pace CN (2007) Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J Mol Biol* 366: 449–460
- Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS Comput Biol* 5: e1000475
- Vishnyakov IE, Borchsenius SN (2013) Mycoplasma heat shock proteins and their genes. *Microbiology* 82: 653–667
- Wang L, Schubert D, Sawaya MR, Eisenberg D, Riek R (2010a) Multidimensional structure-activity relationship of a protein in its aggregated states. *Angew Chemie Int Ed* 49: 3904–3908
- Wang X, Zhou Y, Ren J-J, Hammer ND, Chapman MR (2010b) Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. *Proc Natl Acad Sci USA* 107: 163–168
- Wong JT-F (1975) A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA* 72: 1909–1912
- Zahn M, Berthold N, Kieslich B, Knappe D, Hoffmann R, Sträter N (2013) Structural studies on the forward and reverse binding modes of peptides to the chaperone DnaK. *J Mol Biol* 425: 2463–2479