Title: Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes

Authors: R. R. Cheng[1]*, O. Nordesjö[2], R. L. Hayes[3], H. Levine[1,4], S. C. Flores[2], J. N. Onuchic[1,5]*, F. Morcos[6]*

Affiliations:

[1]Center for Theoretical Biological Physics, Rice University, Houston, USA.

[2]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.

[3]Department of Biophysics, University of Michigan, Ann Arbor, USA.

[4]Department of Bioengineering, Rice University, Houston, USA.

[5]Departments of Physics & Astronomy, Chemistry, and Biosciences, Rice University, Houston, USA.

[6]Department of Biological Sciences, University of Texas at Dallas, Dallas, USA.

*To whom correspondence should be addressed: R. R. Cheng (ryan.r.cheng@gmail.com), J. N. Onuchic (jonuchic@rice.edu), F. Morcos (faruckm@utdallas.edu)

Abstract:
Two-component signaling (TCS) is the primary means by which bacteria sense and respond to the environment. TCS involves two partner proteins working in tandem, which interact to perform cellular functions while limiting interactions with non-partners (i.e., "cross-talk"). We construct a Potts model for TCS that can quantitatively predict how mutating amino acid identities affect the interaction between TCS partners and non-partners. The parameters of this model are inferred directly from protein sequence data. This approach drastically reduces the computational complexity of exploring the sequence-space of TCS proteins. As a stringent test, we compare its predictions to a recent comprehensive mutational study, which characterized the functionality of 20^4 mutational variants of the PhoQ kinase in *Escherichia coli*. We find that our best predictions accurately reproduce the amino acid combinations found in experiment, which enable functional signaling with its partner PhoP. These predictions demonstrate the evolutionary pressure to preserve the interaction between TCS partners as well as prevent unwanted "cross-talk". Further, we calculate the mutational change in the binding affinity between PhoQ and PhoP, providing an estimate to the amount of destabilization needed to disrupt TCS.

Introduction

Early theoretical work on protein folding postulated that proteins have evolved to be minimally frustrated [1, 2, 3], i.e., evolved to have favorable residue-residue interactions that facilitate folding into the native state while having minimal non-native energetic traps. The principle of minimal frustration provides intuition as to why protein sequences are not random strings of amino acids. The evolutionary constraint to fold into a particular, stable three-dimensional structure while minimizing the number of frustrated interactions greatly restricts the sequence-space of a protein [1, 3, 4]. Satisfaction of these constraints result in correlated amino acid identities within the sequences of a protein family. These correlated identities occur between different positions in a protein such as, for example, native contacts [5, 6, 7]. We refer to these quantifiable amino acid correlations as coevolution.

Of course, coevolution does not only arise from the constraint to fold. Proteins also fulfill cellular functions, which act as additional constraints on the sequences of proteins [8, 9, 10]. In the context of signal transduction, proteins have evolved to be able to preferentially bind to a signaling partner(s) as well as catalyze the chemical reactions associated with signal transfer. An important example is two-component signaling (TCS) [11, 12, 13, 14, 15, 16], which serves as the primary means for bacteria to sense the environment and carry out appropriate responses. TCS consists of two partner proteins working in tandem: a histidine kinase (HK) and a response regulator (RR). Upon the detection of stimulus by an extracellular sensory domain, the HK generates a signal via autophosphorylation. Its RR partner can then transiently bind to the HK and receive the signal (i.e., phosphoryl group), thereby activating its function as a transcription factor. The HK has also evolved to catalyze the reverse signal transfer reaction (i.e., phosphatase activity), acting as a sensitive switch to turn off signal transduction. To prevent signal transfer with the wrong partner (i.e., "cross-talk"), TCS partners have mutually evolved amino acids at their binding interface that confer interaction specificity [14, 15, 16]. Thus, the collection of protein sequences of TCS partners contains quantifiable coevolution between the HK and RR sequences.

Assuming that nature has sufficiently sampled the sequence-space of TCS proteins, the collection of protein sequences of TCS partners can be viewed as being selected under quasi-equilibrium from a Boltzmann distribution:

$$P(S_{\mathrm{TCS}}) = Z^{-1} \exp\left(-H(S_{\mathrm{TCS}}) / k_B T_{sel}\right) \qquad (1)$$

where S_{TCS} is the concatenated amino acid sequence of a HK and RR protein, P is the probability of selecting S_{TCS}, Z is the normalization (partition function), T_{Sel} is the evolutionary selection temperature [17], and H is an appropriate energy function in units of $k_B T_{Sel}$. Recently, maximum entropy-based approaches referred to as Direct Coupling Analysis (DCA) [18, 19, 20] have been successfully applied to infer the parameters of H (a Potts model) that governs the empirical amino acid sequence statistics. This has allowed for the direct quantification of the coevolution in protein sequence data (See Review: [21]). Early work using DCA to study TCS primarily focused on identifying the key coevolving residues between the HK and RR [20]. Highly coevolving residue pairs have been used as docking constraints in a molecular dynamics simulation to predict the HK/RR signaling complex [22], the autophosphorylation structure of a HK [23], and the homodimeric form (transcription factor) of the RR [24]. Recently, DCA has also been applied to quantify the determinants of interaction specificity between TCS proteins [25, 26], building on earlier coevolutionary approaches [27, 28]. In particular, DCA was used to predict the effect of point mutations on TCS phosphotransfer *in vitro* as well as demonstrate the reduced specificity between HK and RR domains in hybrid TCS proteins [26].

The experimental effort to determine the molecular origin of interaction specificity in TCS proteins (See Reviews: [13, 14, 15, 29]) precedes the recent computational efforts. Full knowledge of the binding interface between HK and RR was made possible through X-ray crystallography [30]. Scanning mutagenesis studies [31, 32, 33] provided insight on the subset of important interfacial residues that determine specificity. These key residues were mutated to enable a TCS protein to preferentially interact with a non-partner *in vitro* [31, 34]. However, the extent of possible amino acid identities that allow TCS partners to preferentially interact *in vivo* has remained elusive until recent comprehensive work by Podgornaia and Laub [35]. Their work focused on the PhoQ/PhoP TCS partners in *E. coli*, which control the response to low magnesium stress. PhoQ (HK) phosphorylates and dephosphorylates PhoP (RR) under low and high magnesium concentrations, respectively. Using exhaustive mutagenesis of 4 residues of PhoQ (20^4 = 160,000 mutational variants) at positions that form the binding interface with PhoP, Podgornaia and Laub [35] were able to characterize all mutants based on their functionality in *E. coli*. It was found that roughly 1% of all PhoQ mutants were functional, enabling *E. coli* to exhibit comparable responses to magnesium concentrations as the wild type PhoQ. This finding uncovered a broad degeneracy in the sequence-space of the HK protein that still maintained signal transfer efficiency as well as interaction specificity with its partner.

We ask whether amino acid coevolution inferred using DCA could capture the functional mutational variants observed in the comprehensive mutational study of PhoQ and if so, to what extent? Capturing this functionality requires that information gleaned from coevolution is sufficient to estimate the effect of mutations to PhoQ on its interactions with PhoP as well as on unwanted "cross-talk". Hence, our question is important to determine if coevolutionary methods can be extended from studying two interacting proteins to studying an interaction network (e.g., systems biology). Further, this question is of particular interest to those who want to engineer novel mutations in TCS proteins that can maintain or encode the interaction specificity of a TCS protein to its partner or a non-partner, respectively.

To answer this question, we first infer a Potts model energy function, H (see Eq. 1), which forms the basis for quantifying how mutations affect the interaction between a HK and RR protein. Focusing on the parameters of H that are related to interprotein coevolution, we constructing an energy landscape to quantify TCS interactions, H_{TCS}, for a given sequence of an HK and RR protein. H_{TCS} serves as a proxy for signal transfer efficiency, allowing us to quantify the interaction between any HK and RR protein. Further, we can assess how mutations affect the HK/RR interaction by computing the mutational change in H_{TCS} between the mutant sequence, $S_{\mathrm{TCS}}^{\mathrm{mutant}}$, and the wild type sequence, $S_{\mathrm{TCS}}^{\mathrm{WT}}$:

$$\Delta H_{\mathrm{TCS}} = H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) - H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{WT}}). \qquad (2)$$

Considering the concatenated sequence of PhoQ and PhoP, we compute Eq. 2 for the 20^4 PhoQ mutational variants. We find that mutants with the most favorable ΔH_{TCS} (e.g., most negative) were classified as functional HKs by Podgornaia and Laub [35]—i.e., true positive predictions. Next, we focus on mutations predicted to be favorable by Eq. 2 that were classified as non-functional in experiment. Expanding our analysis of the PhoQ mutants beyond its interaction with PhoP, we consider how mutations affect the signal transfer efficiency, H_{TCS}, between PhoQ and all of the RR proteins in *E. coli*. We find that many of these non-functional mutants exhibit "cross-talk" reactions according to our model, accounting for their non-functionality. If we exclude these promiscuous variants, we can better isolate the true positive predictions that are functional from false positives that are non-functional. We next constructed a coevolutionary

energy landscape for the HK alone, H_{HK}. We use H_{HK} to assess the extent to which a model that only considers the HK can capture the experimentally observed mutational phenotypes. While such a model is not better at identifying functional mutations than H_{TCS}, we find the most favorable mutations that preserve autophosphorylation are also the most favorable for preserving the HK/RR interaction. This demonstrates the evolutionary pressure to simultaneously preserve both monomeric function and complex formation. Finally, we estimate the mutational change in binding affinity in the PhoQ/PhoP bound complex using the Zone Equilibration of Mutants (ZEMU) method [36], a combined physics- and knowledge-based approach for free energy calculations. Consistent with what we would expect, we find that mutations that destabilize the HK/RR interaction tend to be non-functional with very high statistical significance. Non-functional mutants are on average destabilized by ~2 kcal/mol with respect to functional mutants. Further, we provide inconclusive support for the view that the mutations that overly strengthen the binding affinity may also be deleterious towards TCS.

The work described herein demonstrates that a coevolutionary model built from sequence data can directly connect molecular details at the residue-level to mutational phenotypes in bacteria. This has broad applications in systems biology, but also in synthetic biology since our computational framework can be used to select mutations that enhance or suppress interactions between TCS proteins. A more detailed description of our computational approaches can be found in the Materials and Methods section.

Results

Mutational change in coevolutionary energy landscape, ΔH_{TCS}, for PhoQ/PhoP interaction

Focusing on the Dimerization and Histidine phosphotransfer domain (DHp) and the Receiver (REC) domain (Fig. 1A) which form the HK/RR binding interface (Fig. 1C), we construct an energy landscape, H_{TCS} (Eq. 5), as a proxy for signal transfer efficiency. For each of the 1,659 functional and 158,341 non-functional PhoQ-mutational variants identified by Podgornaia and Laub [35], we compute the mutational change in energy, ΔH_{TCS}, between PhoQ and PhoP. As an initial step, we only consider the PhoQ/PhoP sequence, i.e., we do not yet consider other RR proteins than PhoP. A histogram of ΔH_{TCS} is generated for all mutational variants (Fig. 2A). The distribution of the functional mutants tends more towards favorable ΔH_{TCS} than the distribution of non-functional mutants, but more interestingly, the most favorable predictions of our model contain mostly functional mutations. This is made clear by a plot of the Positive Predictive Value (PPV) for the top N mutational variants ranked by ΔH_{TCS} (Fig. 2B) from most favorable to most deleterious. The top 25 mutational variants ranked by ΔH_{TCS} contain 20 functional mutants and 5 non-functional mutants (i.e., PPV=0.8).

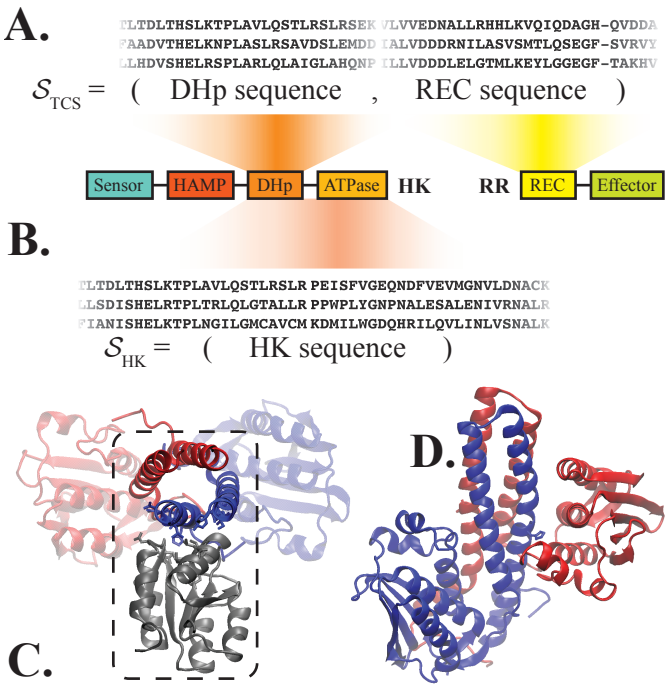

Fig. 1. **TCS domain interactions of interest.** We focus only on HK proteins that have the following domain architecture from N to C terminus: sensor, HAMP, DHp, and ATPase. Likewise, we consider RR proteins that consist of a REC domain followed by an effector domain. (A) The interaction between the DHp and REC domains of the HK and RR proteins, respectively, form the TCS complex. Sequences of TCS partners are collected and stored as the concatenated sequence of the DHp and REC domains, S_{TCS} (See Materials and Methods). (B) We also consider a model of the HK sequences, S_{HK}, consisting of the DHp and ATPase domains. (C) A representative structure of the HK/RR TCS complex previously predicted for the KinA/Spo0F complex in *B. subtilis* [26]. The HK homodimer is shown in red and blue while the receiver domain of the RR is shown in gray. The dashed box highlights the DHp and REC interface. (D) The crystal structure of a representative autophosphorylation state is shown for the HAMP-containing HK CpxA in *E. coli* [37]. The ATPase domain of one HK protein (red) is bound to the DHp domain of the other HK protein (blue).

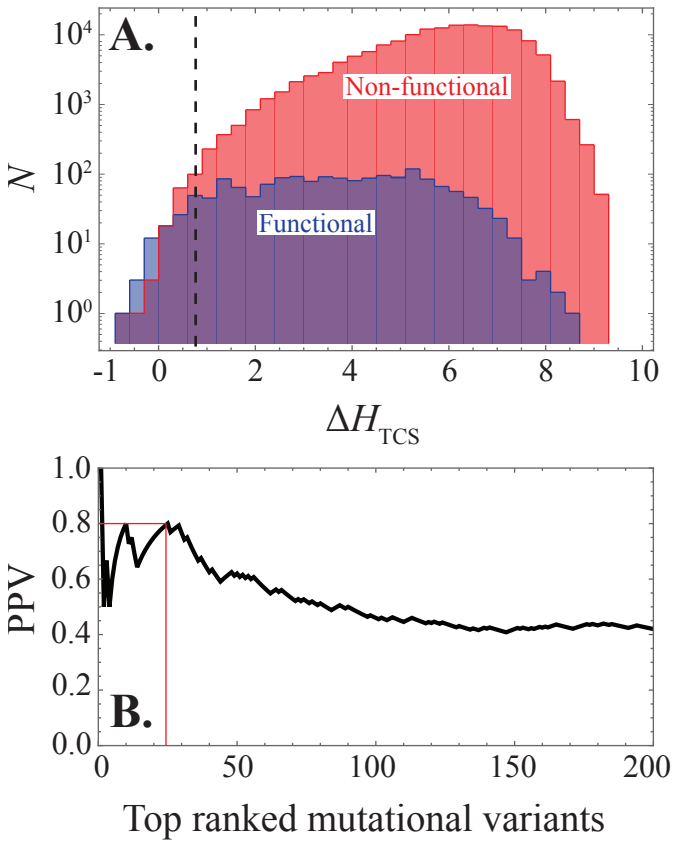

Fig. 2. **Effect of mutations on the PhoQ/PhoP interaction.** (A) Considering the concatenated sequence of PhoQ/PhoP, a histogram of ΔH_{TCS} (Eq. 5) is plotted for the functional (blue) and non-functional (red) mutational variants reported by Podgornaia and Laub [29]. The color purple shows parts of the plot where the blue and red histograms overlap. The dashed line roughly partitions the 200 most favorable mutational variants given by ΔH_{TCS}, which contains more functional than non-functional mutants. By definition, $\Delta H_{TCS} = 0$ corresponds to the wild type PhoQ/PhoP and $\Delta H_{TCS} < 0$ corresponds to mutations that we predict to be more favorable to PhoQ/PhoP signaling than the wild type. (B) We plot the positive predictive value (PPV) as a function of the *N* mutational variants ranked by ΔH_{TCS} from the most to least favorable for the first 200 mutants. PPV=TP/(TP+FP), where true positives (TP) and false positives (FP) refer to the fraction of mutants that are functional or non-functional, respectively, in the top *N* ranked variants. The thin red lines denote that the top 25 ranked mutational variants have a PPV of 0.8.

System-level analysis using ΔH_{TCS}: functional mutants limit "cross-talk"

Mutations that may enhance signal transfer efficiency between PhoQ and PhoP *in vitro* may still result in a non-functional PhoQ/PhoP system *in vivo*. This would occur if the mutations to PhoQ sufficiently encoded it to preferentially interact with another RR in *E. coli*. For this reason, we

focused our computational analysis on the subset of mutational variants that preserve PhoQ/PhoP specificity by limiting "cross-talk" according to our coevolutionary model.

We first calculate the signal transfer efficiency, H_{TCS}, between the wild type PhoQ sequence and all of the non-hybrid RR proteins in *E. coli* (Fig. 3A). We find that for wild type PhoQ, the most favorable H_{TCS} (most negative) is with its known signaling partner, PhoP. As a consistency check, we also plot H_{TCS} for different combinations of the cognate partners TCS proteins in *E. coli* (Fig. S1). This result is consistent with previous computational predictions that used information-based quantities [25, 26] to quantify interaction specificity.

Extending upon Fig. 3A, we assess "cross-talk" in our model by calculating H_{TCS} between each PhoQ mutant and all of the non-hybrid RR in *E. coli*. We exclude all mutant-PhoQ variants that have a more favorable H_{TCS} with a non-partner RR. These excluded mutants are excellent candidates for engineering specificity in *E. coli*. Applying our exclusion criterion, we find that only 181 functional and 1,532 non-functional variants remain, i.e., 89% and 99% of the functional and non-functional variants, respectively, were removed. A histogram of the remaining (cross-talk excluded) mutants as a function of ΔH_{TCS} (Fig. 3B) shows that a filter based on interaction specificity is better able to isolate the true positive (functional) variants. Notably, the first 17 ranked variants are all functional variants. Once again, ranking the filtered variants by ΔH_{TCS} from the most favorable to the least favorable, we can plot the PPV (Fig. 3C) for the top N ranked variants. We find that the cross-talk excluded PPV tends to lie above the original PPV from Fig. 2B. Further, these functional predictions tend to be 3- and 4-point mutations (Fig. S2), highlighting their non-trivial nature.

Construction of an energy landscape to assess mutational effect on autophosphorylation

We further extend our analysis by examining whether a model based solely on HK (intraprotein) coevolution can identify the functional and non-functional mutational variants of PhoQ. We construct an energy landscape of HK intraprotein interactions, H_{HK} (Eq. 6), focusing on the DHp and ATPase domains (Fig. 1B). These domains form a binding interface during autophosphorylation (Fig. 1C). To test the fidelity of our inferred model, we plot the top coevolving residue pairs in the HK using the Direct Information (DI) [18, 20] metric (Fig. S3). We find that the top coevolving pairs are contacts in the experimentally determined autophosphorylation state of the HK [37].

We generate a 2D histogram of ΔH_{HK} and ΔH_{TCS} for the 20^4 PhoQ mutational variants (Fig. 4A). Note that H_{HK} (Eq. 5) and H_{TCS} (Eq. 6) are models of the intraprotein (HK only) and interprotein (HK/RR) coevolution, respectively. Interestingly, the mutational variants as a function of ΔH_{HK} and ΔH_{TCS} are highly correlated with Pearson correlations of 0.72 and 0.75 for the functional and non-functional mutants, respectively (Fig. 4A). This reflects the evolutionary constraint to satisfy both HK/RR phosphotransfer and HK autophosphorylation simultaneously. This constraint restricts the sequence-space of the HK to the same subset of amino acid residues on the binding region of the DHp domain. However, we find that ΔH_{HK} is not better at identifying functional mutations than ΔH_{TCS}, which is quantified by the plot of PPV versus the N top mutational variants ranked by ΔH_{HK} (Fig. S4).

Next, we examine the subset mutational variants that limit "cross-talk" (detailed in

previous subsection) in a 2D histogram of ΔH_{HK} and ΔH_{TCS} (Fig. 4B). We find that while the remaining functional mutants are clustered around more favorable values of ΔH_{TCS}, they cover a wide range of deleterious mutational changes in ΔH_{HK}. This would suggest that the HK is perhaps more tolerant of mutations that may reduce autophosphorylation but more sensitive to mutations that reduce interaction specificity for phosphotransfer or phosphatase activity. Interestingly, it was reported that the response regulator PhoP, is capable of receiving a phosphoryl group from acetyl-phosphate [35] and thus, mutations that minimally reduce PhoQ autophosphorylation activity but preserving interaction specificity with PhoP may still be identified as functional in experiment.

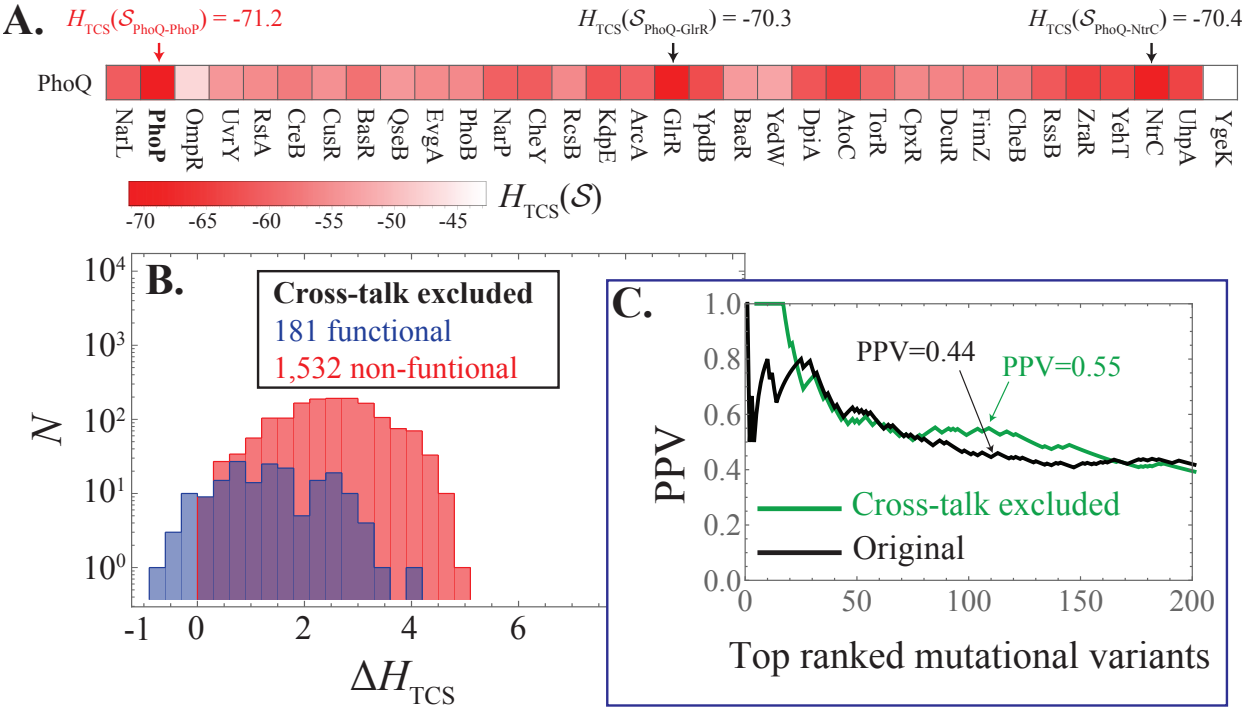

Fig. 3. Excluding mutational variants that are inferred to "cross-talk". (A) A grid plot showing H_{TCS} (Eq. 5) computed for the wild type PhoQ sequence with all of the non-hybrid RR protein sequences in *E. coli*, respectively. The most favorable energy (most negative) is between PhoQ and its partner PhoP. (B) We plot the Cross-talk excluded subset (181 functional 1,532 non-functional) in a histogram as a function of the ΔH_{TCS} similar to Fig. 2A. (C) We plot the PPV as a function of the N top mutational variants ranked by ΔH_{TCS} for the first 200 mutants. The PPV for the cross-talk excluded mutational variants from Fig. 3B are plotted in green while the original PPV (Fig. 2B) is shown in black.

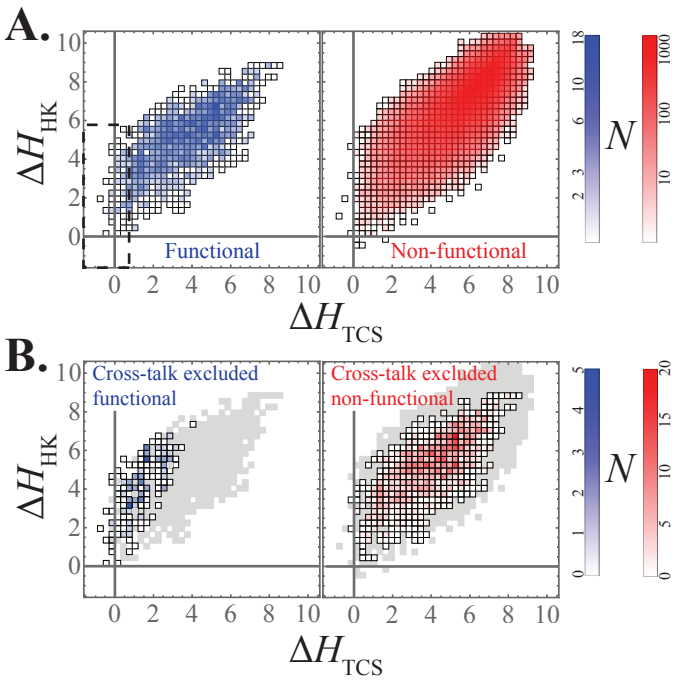

Fig. 4. Effect of mutation on PhoQ autophosphorylation. (A) Considering a 2D histogram of ΔH_{TCS} and ΔH_{HK} for the functional variants (blue) and non-functional variants (red). The two solid lines corresponding to $\Delta H_{TCS} = 0$ and $\Delta H_{HK} = 0$ represent the mutational change with respect to the wild type PhoQ. Additionally, $\Delta H_{TCS} < 0$ and $\Delta H_{HK} < 0$ correspond to mutations that are more favorable than the wild type. The number of variants in each bin is shown on the logarithmic scale by the shade of blue or red, respectively. We observe an enrichment of functional mutants for lower values of both ΔH_{TCS} and ΔH_{HK}. (B) The 2D histogram is plotted for the subset of mutational variants that limit "cross-talk" by preserving the specificity of the PhoQ/PhoP interaction (labeled Cross-talk excluded). The gray squares denote the parts of the original 2D histogram (panel A) that have been removed.

Mutational change in the binding affinity using a combined physics- and knowledge-based approach

We used ZEMu (See Materials and Methods) to compute the mutation-induced change in the binding affinity, $\Delta\Delta G_{TCS}^{ZEMu}$, between PhoQ and PhoP. The calculation converged for 42,985 mutants (702 functional and 42,283 non-functional) from a randomly selected subset of the 20^4 variants. A histogram of $\Delta\Delta G_{TCS}^{ZEMu}$ is plotted for these mutants in Fig. 5A. A histogram of ΔH_{TCS} for the same subset of mutants is shown in Fig. S5A. On the population level, functional mutations exhibit a mean $\Delta\Delta G_{TCS}^{ZEMu}$ of 1.76 ± 0.06 kcal/mol lower than that of the non-functional mutants, with a Wilcoxon rank-sum test p-value $< 2.2 \times 10^{-16}$. This indicates that destabilizing mutations of ~2 kcal/mol are sufficient for disrupting TCS. Furthermore, destabilizing mutations that are more than 2 standard deviations greater than the mean $\Delta\Delta G_{TCS}^{ZEMu}$ for functional variants are significantly less likely to be functional, with a p-value $< 10^{-6}$ computed from a cumulative

binomial distribution (based on the 6157 mutants above this threshold, 19 of which are functional).

We next examine the potentially deleterious effect of mutations that overly stabilize the binding affinity between PhoQ and PhoP. Although we find that all 56 mutants with $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}} < -5$ kcal/mol are non-functional (Fig. 5A), this has no statistical significance (p-value ~ 0.4). However, when we further examine the subset of mutants explored using ZEMu while excluding the mutants that exhibited "cross-talk" based on our coevolutionary energy, ΔH_{TCS}, in the previous subsection, we find that the remaining subset of mutants falls within roughly ±5 kcal/mol from the wild type PhoQ/PhoP binding affinity (Fig. 5B). Selecting this subset of 363 mutational variants (92 functional and 272 non-functional) from the population of 42,985 variants is statistically significant with a p-value < 10^{-81} from a cumulative hypergeometric distribution. A histogram of ΔH_{TCS} for this subset of mutational variants is shown for consistency in Fig. S5B. Fig. 5B shows that a combination of our ZEMu calculation and coevolutionary energy, ΔH_{TCS}, can provide support (albeit inconclusive) for the existence of an upper limit to the binding affinity for TCS proteins that would disrupt the transient nature of TCS signaling, since both the strong and weak binding mutants exhibited diminished signal transfer efficiency (i.e., ΔH_{TCS}) with PhoP respective of the other RR proteins in *E. coli*. However, further statistical analysis would be necessary to firmly establish such an upper limit for TCS.

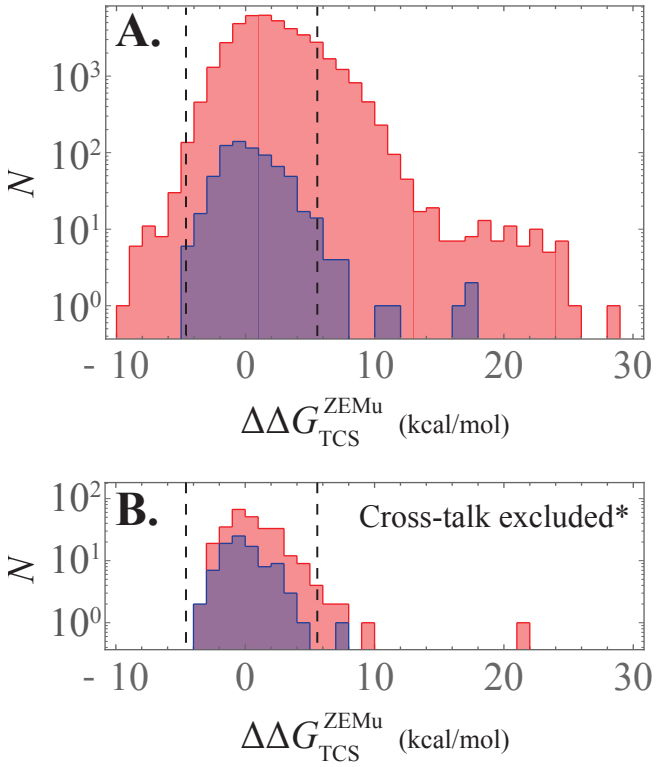

Figure 5. Mutational change in binding affinity for PhoQ/PhoP interaction. (A) A histogram of $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ (See Materials and Methods), is plotted for the 702 functional (blue) and 42,283 non-functional (red) mutational variants analyzed in our study. The dashed lines denote ±2 standard deviations from the mean of the functional (blue) distribution. (B) A histogram of $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ is plotted for the subset of mutants in panel A

that overlap with the subset of mutants that limited "cross-talk" based on the coevolutionary energy, ΔH_{TCS} (i.e., Cross-talk excluded subset in Fig. 3B). The resultant subset consists of 92 functional and 271 non-functional mutational variants, respectively, and is denoted with an asterisk in the label.

Discussion

Treating a large collection of amino acid sequence data for TCS partner proteins as independent samples from a Boltzmann equilibrium distribution, we infer a coevolutionary energy landscape, H_{TCS}. This energy function captures coevolving amino acid combinations that give rise to interaction specificity in TCS systems. In the past, we were able to predict how a point mutation to a TCS protein affects its ability to transfer signal to its partner *in vitro* [26]. Our present work shifts the paradigm of coevolution-based analysis towards systems biology by extending our analysis to include how those mutations affect "cross-talk" in a bacterial organism. We demonstrate that our most favorable predictions for multiple site mutations can accurately capture *in vivo* TCS functionality, consistent with the comprehensive mutational study of Podgornaia and Laub [35]. This is not a trivial computational task, since inferring coevolutionary information from sequence data is highly underdetermined and estimates ~10^7 parameters (for Eq. 3) from ~10^3 sequences of TCS partners. Adding to the problem complexity, it is plausible that the full functional sequence space has not yet being explored by evolutionary process [16, 38]. Despite this, the coevolutionary landscape is predictive and identifies mutational variants that are not found in nature, e.g., none of the mutational sequences are included as input data in our model. We have demonstrated the feasibility of generating predictions using coevolution and the predictive power of such an approach will only systematically improve as more sequences of TCS partners are collected.

Similar predictions to those discussed herein can readily be used to engineer novel protein-protein interactions in TCS systems. Such a strategy would potentially complement already existing strategies to match novel inputs with outputs via modular engineering [39, 40, 41, 42, 43]. The strength of our coevolutionary approach is that it makes possible an efficient search of sequence-space for mutations at arbitrary positions in either the HK or RR that desirably enhance or suppress its interaction with a RR or HK, respectively. It can also readily be applied to study the *in vivo*, system-level effect of mutating a TCS protein on insulating its interaction with a desired partner or enabling "cross-talk" with non-partners. Our study highlights an intuitive but key principle for selecting mutations to a TCS protein that encodes specificity *in vivo*: mutations must be selected to enhance protein-protein interactions with a desired partner while limiting protein-protein interactions with undesired partners. While also intuitive, we demonstrate that mutations that significantly destabilize the binding affinity result in the loss of signaling. Further, we estimate that destabilization of ~2 kcal/mol in the binding affinity between TCS partners is sufficient to disrupt TCS.

It is also important to note that coevolutionary methods described here for identifying mutational phenotypes (e.g., response to magnesium stress) is that they are not particular to TCS systems. This framework is transferable to other systems where molecular interactions coevolve to preserve function, opening the window to a large set of open problems in molecular and systems biology. Our results further extend the idea that a combination of coevolutionary based methods, molecular modeling and experiment can be used to identify the proper amino acids

sites and identities that can be used to identify mutational phenotypes. Our study highlights the important role of coevolution in maintaining protein-protein interactions, as in the case of bacteria signal transduction. Statistical methods that probe coevolution not only allow us to connect molecular, residue-level details to mutational phenotypes, but also to explore the evolutionary selection mechanisms that are employed by nature to maintain interaction specificity, e.g., negative selection [44]. Further investigations of other systems that are evolutionarily constrained to maintain protein-protein interactions could elucidate the extent at which our methods can be used in alternative systems. One potential example is the toxin-antitoxin protein pairs in bacteria, which was the focus of recent experimental work [45] elucidating the determinants of interaction specificity.

Materials and Methods

Sequence database for HK and RR inter-protein interactions: DHp and REC

We obtain multiple sequence alignments (MSA) from Pfam [46] (Version 28), focusing on the DHp (PF00512) and REC (PF00072) domains of the HK and RR, respectively (Fig. 1A). The first 4 positions (columns) of PF00512 were removed due to poor alignment of the PhoQ sequence at those positions. The remaining DHp MSA has a length of $L_{\mathrm{DHp}} = 60$. Each REC MSA had a default length of $L_{\mathrm{REC}} = 112$. Here, we considered HK proteins that have the same domain architecture as the PhoQ kinase from *E. coli*, i.e., DHp domain sandwiched between an N-terminal HAMP domain (PF00672) and a C-terminal ATPase domain (PF02518). The remaining HK (DHp) sequences were paired with a TCS partner RR (REC) by taking advantage of the observation that TCS partners are typically encoded adjacent to one another under the same operon [47, 48], i.e., ordered locus numbers differ by 1. Further, we exclude all TCS pairs that are encoded adjacent to multiple HKs or RRs. Each DHp and REC sequence that was paired in this fashion was concatenated into a sequence (Fig. 1A), $S_{\mathrm{TCS}} = (A_1, A_2, ..., A_{L-1}, A_L)$ of total length L where A_i is the amino acid at position i which is indexed from 1 to $q = 21$ for the 20 amino acids and MSA gap. The DHp sequence is indexed from positions 1 to L_{DHp} and REC sequence from positions $L_{\mathrm{DHp}} + 1$ to the total length of $L = L_{\mathrm{DHp}} + L_{\mathrm{REC}} = 172$. Our remaining dataset (External Databases S1) consisted of 6,519 non-redundant concatenated sequences.

Sequence database for HK intra-protein interactions: DHp and ATPase

We searched the Representative Proteomes (RP55) [49] database for HK proteins using the Jackhmmer algorithm on the HMMER web server [50] with default parameters. Once again, we restricted our curated set of HK proteins to ones having the PhoQ domain architecture. We restricted our MSA to a length (number of columns) of $L_{\mathrm{HK}} = 222$ by only including the DHp and ATPase domains. Our remaining dataset (External Databases S2) consisted of 4,483 non-redundant HK sequences of the form: $S_{\mathrm{HK}} = (A_1, A_2, ..., A_{L-1}, A_L)$ with a total length $L = L_{\mathrm{HK}} = 222$.

Inference of parameters of coevolutionary model

The collection of protein sequences for a protein family or coevolved families can be viewed as being selected from a Boltzmann equilibrium distribution, i.e., $P(s) = Z^{-1}\exp(-\beta H(s))$, where $s = (A_1, A_2, ..., A_L)$ is a sequence, $\beta = (k_B T_{sel})^{-1}$ is the inverse of the evolutionary selection temperature [17], and $H(s)$ is an appropriate energy function in units of $k_B T_{sel}$. However, more appropriate for our interests is the inverse problem of inferring an appropriate $H(s)$ when provided with an abundant number of protein sequences. Typical approaches to this problem have applied the principle of maximum entropy to infer a least biased model that is consistent with the input sequence data [18, 20], e.g., the empirical single-site and pairwise amino acid probabilities, $P_i(A_i)$ and $P_{ij}(A_i, A_j)$, respectively. The solution of which is the Potts model:

$$H(\mathrm{s}) = -\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} J_{ij}(A_i, A_j) - \sum_{i=1}^{L} h_i(A_i) \tag{3}$$

where A_i is the amino acid at position *i* for a sequence in the MSA, $J_{ij}(A_i, A_j)$ is the pairwise statistical couplings between positions *i* and *j* in the MSA with amino acids A_i and A_j, respectively, and $h_i(A_i)$ is the local field for position *i*. We estimate the parameters of the Potts model, $\{\mathbf{J}, \mathbf{h}\}$, using the pseudo-likelihood maximization Direct Coupling Analysis (plmDCA) (See Ref: [19] for full computational details).

Inference problems of this nature exhibit a known gauge freedom associated with being able to add energy to the inferred fields that can be subtracted from the couplings to maintain a constant *H*. We fix the gauge by adopting the Ising condition, which can be obtained for a Potts model in any particular gauge choice, $\{\hat{\mathbf{J}}, \hat{\mathbf{h}}\}$, using the transformation:

$$\begin{aligned} J_{ij}(a,b) &= \hat{J}_{ij}(a,b) - \hat{J}_{ij}(:,b) - \hat{J}_{ij}(a,:) + \hat{J}_{ij}(:,:) \\ h_i(a) &= \hat{h}_i(a) - \hat{h}_i(:) + \sum_{\substack{j=1 \\ j\neq i}}^{L} \left(\hat{J}_{ij}(a,:) - \hat{J}_{ij}(:,:) \right) \end{aligned} \tag{4}$$

where the colon symbol (:) denotes an average over all amino acids identities at its respective position, i.e., $\hat{J}_{ij}(:,b)$ is the average statistical coupling between positions *i* and *j* where position *j* has amino acid *b* and the average is taken over all possible amino acid combinations at position *i*. The Ising condition has the following property: $\sum_{a=1}^{q} J_{ij}(a,b) = \sum_{b=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} h_i(a) = 0$, where $q = 21$ represents the 20 amino acids and MSA gap. This gauge condition ensures that the ensemble of random sequences has a mean energy of 0.

Previous studies have used DCA to identify highly coevolving pairs of residues to predict the native state conformation of a protein [51, 52, 53], including repeat proteins [54], as well as identify additional functionally relevant conformational states [52, 55, 56] and multi-meric states [18, 20, 22, 24, 56]. Structural and coevolutionary information share complementary roles in the molecular simulations of proteins (See review: [57]). The Potts model (Eq. 3) obtained from DCA has been related to the theory of evolutionary sequence selection [17] as well as mutational changes in protein stability [17, 58, 59]. Additional work has applied DCA to protein folding to predict the effect of point mutations on the folding rate [60] as well as construct a statistical potential for native

contacts in a structure-based model of a protein [61] to better capture the transition state ensemble. Finally, DCA has been used to identify relevant protein-protein interactions in biological interaction networks [25, 62] as well as identify high fitness variants for a number of proteins by relating mutational changes in stability to organism fitness [63, 64] or viral fitness [65].

Mutational changes in coevolutionary energy landscapes

For the concatenated sequences of HK (DHp) and RR (REC) (Fig. 1A), we infer a Potts model (Equation 3). We focus on a subset of parameters in our model consisting of the interprotein couplings, J_{ij}, between positions in the DHp and REC domains that are in close proximity in a representative structure of the TCS complex (Fig. 1C). All local fields terms, h_i, are included to partially capture the effect of mutations on domain stability. These considerations allow us to construct our energy landscape for TCS:

$$H_{\mathrm{TCS}}(S_{\mathrm{TCS}}) = -\sum_{i=1}^{L_{\mathrm{DHp}}} \sum_{j=L_{\mathrm{DHp}}+1}^{L_{\mathrm{DHp}}+L_{\mathrm{REC}}} J_{ij}(A_i, A_j) \times \Theta(c - r_{ij}) - \sum_{i=1}^{L_{\mathrm{DHp}}+L_{\mathrm{REC}}} h_i(A_i) \qquad (5)$$

where S_{TCS} is the concatenated sequence of the DHp and REC domains, the double summation is taken over all interprotein statistical couplings between the DHp and REC domains, Θ is a Heaviside step function, c is the a cutoff distance of 16Å which was determined in a previous study [17], and r_{ij} is the minimum distance between residues i and j in the representative structure. Mutational changes in the energy are then computed using Equation 2, i.e., $\Delta H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) = H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) - H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{WT}})$.

Likewise, for the HK model (Fig. 1B) we infer an energy function with the form of Equation 3 and focus on the statistical couplings between positions in the HK that are in close proximity in a representative structure of the HK autophosphorylation state (Fig. 1D):

$$H_{\mathrm{HK}}(S_{\mathrm{HK}}) = -\sum_{i=1}^{L_{\mathrm{HK}}-1} \sum_{j=i+1}^{L_{\mathrm{HK}}} J_{ij}(A_i, A_j) \times \Theta(c - r_{ij}) - \sum_{i=1}^{L_{\mathrm{HK}}} h_i(A_i) \qquad (6)$$

where Θ is a Heaviside step function, c is the a cutoff distance of 16Å, and r_{ij} is the minimum distance between residues i and j in the representative structure. The mutational changes in energy are then computed using $\Delta H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{mutant}}) = H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{mutant}}) - H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{WT}})$.

Zone Equilibration of Mutants (ZEMu) calculation

ZEMu consists of a multiscale minimization by dynamics, restricted to a flexibility zone of five residues about each substitution site [36], which is followed by a mutational change in stability using FoldX [66]. ZEMu has been used to explain the mechanism of Parkinson's disease associated mutations in Parkin [67, 68]. The minimization is done in MacroMoleculeBuilder (MMB), a general-purpose internal coordinate mechanics code also known for RNA folding [69], homology modeling [70], morphing [71], and fitting to density maps [72].

We use the Zone Equilibration of Mutants (ZEMu) [36] method to predict the mutational change in binding energy between PhoQ and PhoP. ZEMu first treats mutations as small perturbations on the structure by using molecular dynamics simulations (See Ref. [36] for full

computational details) to equilibrate the local region around mutational sites. ZEMu can then estimate the binding affinity between the mutant-PhoQ/PhoP, $\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{mutant})$, and the wild type-PhoQ/PhoP, $\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{WT})$, using the knowledge-based FoldX [66] potential. This allows for the calculation of the mutational change in binding affinity as:
$\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}} = \Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{mutant}) - \Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{WT})$.

ZEMu calculation was performed according to Ref: [36], with the following two differences. First, due to the large number of mutations we capped the computer time permitted to 3 core-hours per mutant, whereas in [36] the limit was 48 hours. This meant that of 122802 mutants attempted, 42923 completed within the time limit, whereas in [36], almost all mutants converged. The major reason for non-convergence in the current work involved mutation to bulky or constrained residues. Steric clashes produced by such residues force the error-controlled integrator [73] to take small time steps and hence use more computer time. Exemplifying this, the amino acids F, W and Y are the most common residues for non-converging mutations at positions 285 and 288 in PhoQ. The second difference was that we permitted flexibility in the neighborhood of all four possible mutation sites, even when not all of them were mutated, whereas in [36] only the mutated positions were treated as flexible. This allowed us to compare all of the mutational energies to a single wild type simulation, also performed with flexibility at all four sites.

Acknowledgments: We would like to thank Drs. Michele Di Pierro and Lena Simine for helpful comments. **Funding**: This research was supported by the NSF INSPIRE award (MCB-1241332) and the NSF-funded Center for Theoretical Biological Physics (PHY-1427654). SF and ON acknowledge funds from eSSENCE (http://essenceofescience.se/), Uppsala University, and the Swedish Foundation for International Cooperation in Research and Higher Education (STINT). We also acknowledge a generous allocation of supercomputer time from the Swedish National Infrastructure for Computing (SNIC) at Uppmax, and applications assistance from Drs. Rudberg, Karlsson, and Freyhult.

Author contributions: R.R.C., F.M., and J.N.O designed the research with the assistance of H.L., S.C.F., and O.N.; R.R.C and O.N. performed the research; R.R.C. and R.L.H. curated the protein databases that were used in our study; R.R.C., F.M., S.F. and O.N. wrote the paper.

Competing interests: The authors declare no competing interests.

Supplementary Materials

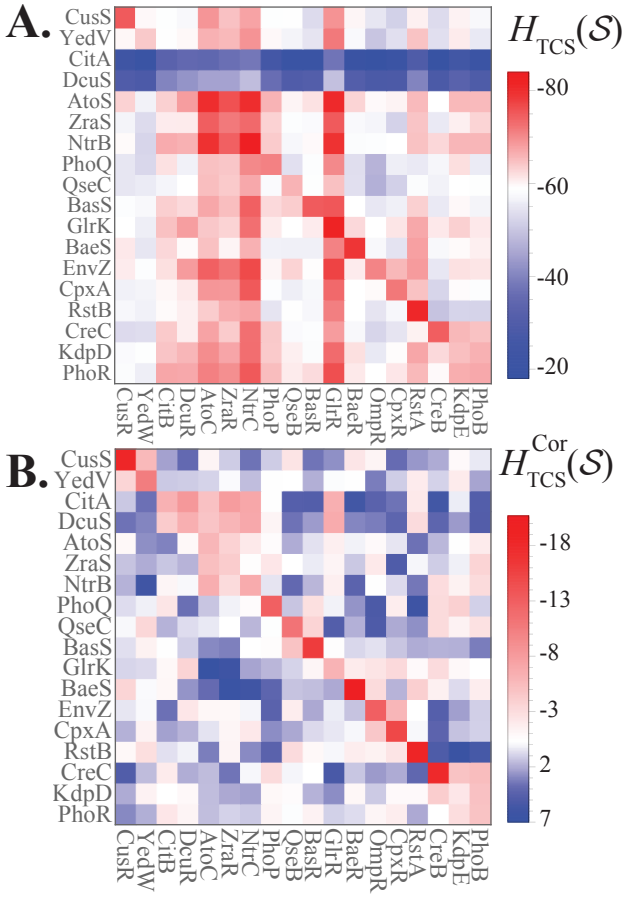

Fig. S1. H_{TCS} **captures partners in *E. coli*.** (A) A grid plot of H_{TCS} where all of the cognate HK (y-axis) is plotted against the cognate RR (x-axis). TCS partners encoded adjacent to one another on an operon are referred to as cognate partners. The plot is arranged in such a way that the diagonal reflects the interaction between known partners [47]. The strength of H_{TCS} is shown in the legend, where more negative suggests better signal transfer efficiency based on our model (i.e., more favorable interaction). Rows 3 and 4 are consistently unfavorable due to the poor alignment of CitA and DcuS, which contain mostly gaps in the MSA. (B) We can apply an average product correction to the square matrix in Fig. S1A as $M_{ij} \rightarrow M_{ij} - M_{:j}M_{i:}/M_{::}$. Here, ":" denotes an average over the row or column, respectively. We refer to this corrected matrix as $H^{\mathrm{Cor}}_{\mathrm{TCS}}$, which satisfies the property that the sum of all rows and columns equals zero, respectively. This allows for the partner interactions to be clearly visualized along the diagonal. However, there is no physical basis for why the mean row or column should be set to the same reference energy. Hence, we do not use any correction to our coevolutionary energy, H_{TCS}, in our study.

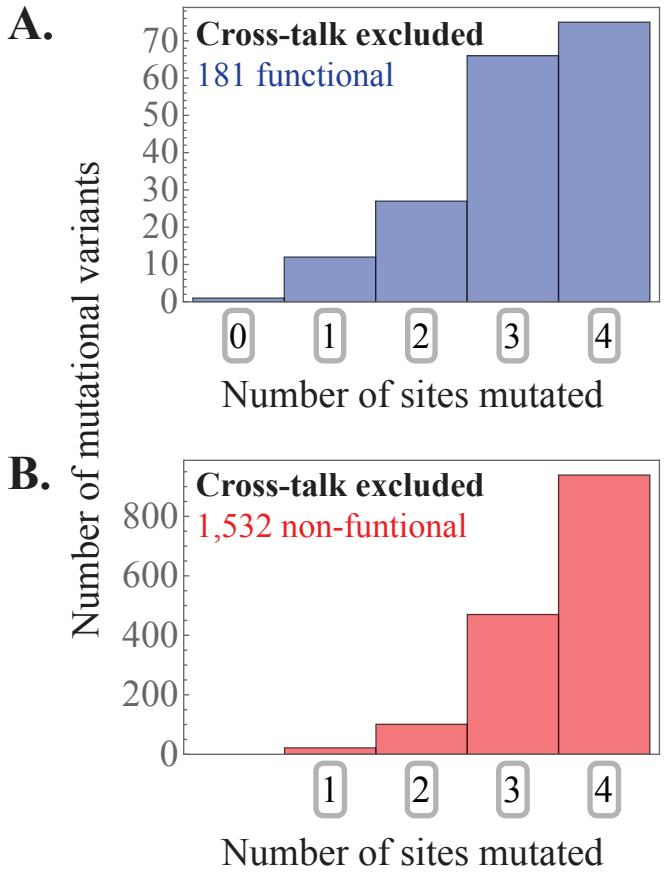

Fig. S2. Number of sites mutated in cross-talk excluded subset. A histogram of the 0, 1, 2, 3, and 4-site mutants is plotted for the (A) 181 functional and (B) 1,532 non-functional mutational variants in our cross-talk excluded subset.

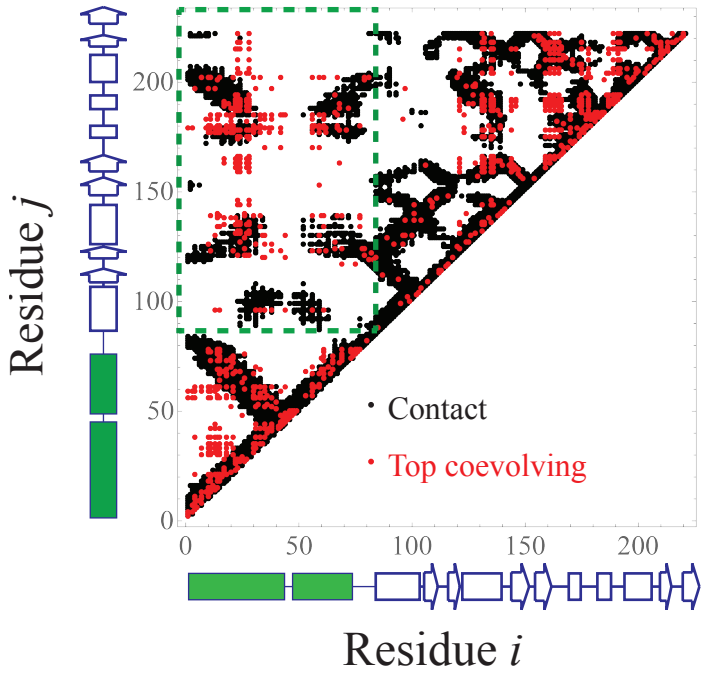

Figure S3. High coevolving pairs of residues in HK correspond to contacts in autophosphorylation state. A contact map is plotted for the crystal structure of the CpxA homodimer (HK) in *E. coli* (PDB: 4BIW[37]) in black where each point denotes two residues in the structure that are within 12Å of one another. The secondary structure of the CpxA protein is shown on the x- and y-axes where the solid green color denotes the DHp domain while the ATPase domain is shown in a solid white. The dashed green box in the contact map contains the DHp/ATPase amino acid residue pairs. The 4BIW structure captures the autophosphorylation state of the HK. Since 4BIW is a homodimer, the inter- and intraprotein contacts are projected onto the contact map of a single monomer. Plotted alongside the contact map are the top 1000 coevolving pairs of residues predicted by DCA using the coevolutionary model constructed for the HK alone in red (See Materials and Methods). The Direct Information (DI) metric is used as a proxy for the amount of coevolution between sites (See Refs: [18, 20] for more details). The highly coevolved pairs of residues correspond well with the contacts in the autophosphorylation state of the HK.

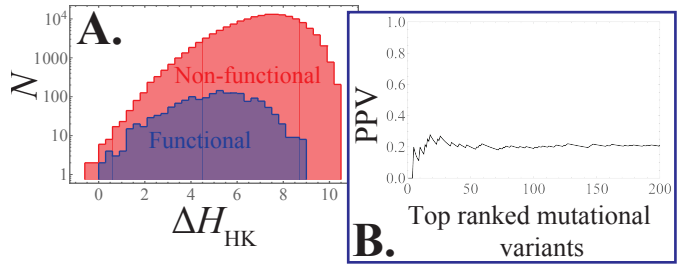

Figure S4. Effect of mutation on the PhoQ autophosphorylation: 1D histogram.

(A) A histogram of the mutational change in our coevolutionary energy, ΔH_{HK} (Eq. 6), is plotted for the functional (blue) and non-functional (red) mutational variants. The color purple shows parts of the plot where the blue and red histograms overlap. By definition, $\Delta H_{\mathrm{HK}} = 0$ corresponds to the mutational change in energy with respect to the wild type PhoQ. (B) We plot the positive predictive value, PPV=TP/(TP+FP), as a function of the *N* mutational variants ranked by ΔH_{HK} from the most to least favorable for the first 200 mutants. Once again, true positives (TP) and false positives (FP) refer to the fraction of mutants that are functional or non-functional, respectively, in the top *N* ranked variants.

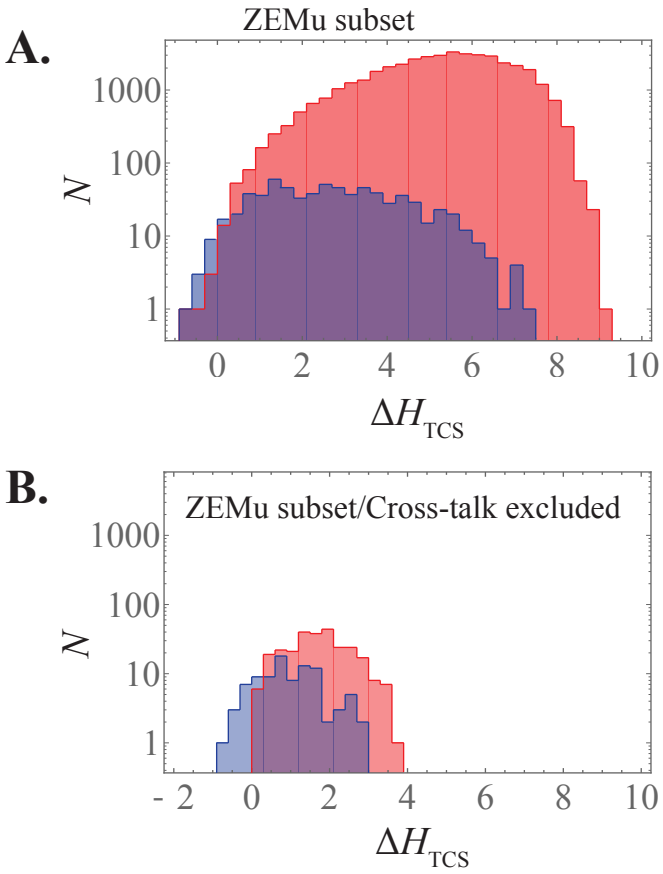

Figure S5. Histogram of mutational change in coevolutionary energy, ΔH_{TCS}, for subset of mutational variants explored by ZEMu calculation. (A) A histogram of the mutational change in our coevolutionary energy, ΔH_{HK} (Eq. 6), is plotted for the subset of 42,985 mutational variants explored by ZEMu, 702 functional (blue) and 42,283 non-functional (red). The analogous plot for the full dataset of 20^4 mutational variants is in Fig. 2A. (B) The subset of mutational variants that were explored by ZEMu and overlap with the Cross-talk excluded subset of Fig. 3B is plotted in a histogram of ΔH_{HK}.

External Database S1. Collection of partnered sequences of DHp and REC.

External Database S2. Collection of HAMP-containing HK sequences.

References

1. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. *Proteins-Structure Function and Genetics* **21**, 167-195 (1995).

2. Bryngelson JD, Wolynes PG. Spin-Glasses and the Statistical-Mechanics of Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 7524-7528 (1987).

3. Onuchic JN, LutheySchulten Z, Wolynes PG. Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem* **48**, 545-600 (1997).

4. Leopold PE, Montal M, Onuchic JN. Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8721-8725 (1992).

5. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317 (1994).

6. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* **7**, 349-358 (1994).

7. Neher E. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America* **91**, 98-102 (1994).

8. Ferreiro DU, Komives EA, Wolynes PG. Frustration in biomolecules. *Q Rev Biophys* **47**, 285-363 (2014).

9. Wolynes PG. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218-230 (2015).

10. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface* **11**, 20140419 (2014).

11. Stock AM, Robinson VL, Goudreau PN. Two-component signal transduction. *Annual Review of Biochemistry* **69**, 183-215 (2000).

12. Hoch JA. Two-component and phosphorelay signal transduction. *Current Opinion in Microbiology* **3**, 165-170 (2000).

13. Casino P, Rubio V, Marina A. The mechanism of signal transduction by two-component systems. *Current Opinion in Structural Biology* **20**, 763-771 (2010).

14. Szurmant H, Hoch JA. Interaction fidelity in two-component signaling. *Current Opinion in Microbiology* **13**, 190-197 (2010).

15. Laub MT, Goulian M. Specificity in Two-Component Signal Transduction Pathways. *Annual Review of Genetics* **41**, 121-145 (2007).

16. Capra EJ, Laub MT. Evolution of two-component signal transduction systems. *Annu Rev Microbiol* **66**, 325-347 (2012).

17. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12408-12413 (2014).

18. Morcos F*, et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293-E1301 (2011).

19. Ekeberg M, Lovkvist C, Lan YH, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* **87**, (2013).

20. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 67-72 (2009).

21. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nature reviews Genetics* **14**, 249-261 (2013).

22. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences* **106**, 22124-22129 (2009).

23. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences* **109**, E1733-E1742 (2012).

24. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep-Uk* **5**, (2015).

25. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. *PloS one* **6**, e19729 (2011).

26. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, (2014).

27. Li L, Shakhnovich EI, Mirny LA. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proceedings of the National Academy of Sciences* **100**, 4463-4468 (2003).

28. Burger L, van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4**, (2008).

29. Podgornaia AI, Laub MT. Determinants of specificity in two-component signal transduction. *Current Opinion in Microbiology* **16**, 156-162 (2013).

30. Casino P, Rubio V, Marina A. Structural Insight into Partner Specificity and Phosphoryl Transfer in Two-Component Signal Transduction. *Cell* **139**, 325-336 (2009).

31. Capra EJ, Perchuk BS, Lubin EA, Ashenberg O, Skerker JM, Laub MT. Systematic Dissection and Trajectory-Scanning Mutagenesis of the Molecular Interface That Ensures Specificity of Two-Component Signaling Pathways. *PLoS Genetics* **6**, (2010).

32. Tzeng Y-L, Hoch JA. Molecular recognition in signal transduction: the interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis. *Journal of Molecular Biology* **272**, 200-212 (1997).

33. Qin L, Cai S, Zhu Y, Inouye M. Cysteine-Scanning Analysis of the Dimerization Domain of EnvZ, an Osmosensing Histidine Kinase. *Journal of Bacteriology* **185**, 3429-3435 (2003).

34. Skerker JM*, et al.* Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell* **133**, 1043-1054 (2008).

35. Podgornaia AI, Laub MT. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673-677 (2015).

36. Dourado DFAR, Flores SC. A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins-Structure Function and Bioinformatics* **82**, 2681-2690 (2014).

37. Mechaly AE, Sassoon N, Betton JM, Alzari PM. Segmental Helical Motions and Dynamical Asymmetry Modulate Histidine Kinase Autophosphorylation. *Plos Biol* **12**, (2014).

38. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. *Nature reviews Genetics* **17**, 109-121 (2016).

39. Schmid SR, Sheth RU, Wu A, Tabor JJ. Refactoring and Optimization of Light-Switchable Escherichia coli Two-Component Systems. *Acs Synth Biol* **3**, 820-831 (2014).

40. Whitaker WR, Davis SA, Arkin AP, Dueber JE. Engineering robust control of two-component system phosphotransfer using modular scaffolds. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 18090-18095 (2012).

41. Ganesh I, Ravikumar S, Lee SH, Park SJ, Hong SH. Engineered fumarate sensing Escherichia coli based on novel chimeric two-component system. *J Biotechnol* **168**, 560-566 (2013).

42. Tabor JJ, Levskaya A, Voigt CA. Multichromatic control of gene expression in Escherichia coli. *J Mol Biol* **405**, 315-324 (2011).

43. Hansen J, Benenson Y. Synthetic biology of cell signaling. *Nat Comput* **15**, 5-13 (2016).

44. Zarrinpar A, Park S-H, Lim WA. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676-680 (2003).

45. Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **163**, 594-606 (2015).

46. Finn RD*, et al.* Pfam: the protein families database. *Nucleic Acids Research* **42**, D222-D230 (2014).

47. Yamamoto K, Hirao K, Oshima T, Aiba H, Utsumi R, Ishihama A. Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *Journal of Biological Chemistry* **280**, 1448-1456 (2005).

48. Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: A system-level analysis. *Plos Biol* **3**, 1770-1788 (2005).

49. Chen CM*, et al.* Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. *PloS one* **6**, (2011).

50. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29-W37 (2011).

51. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences* **109**, 10340-10345 (2012).

52. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 13567-13572 (2015).

53. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nature biotechnology* **30**, 1072-1080 (2012).

54. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Capturing coevolutionary signals inrepeat proteins. *BMC Bioinformatics* **16**, 207 (2015).

55. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, (2013).

56. Malinverni D, Marsili S, Barducci A, De Los Rios P. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS computational biology* **11**, (2015).

57. Noel JK, Morcos F, Onuchic JN. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. *F1000Res* **5**, (2016).

58. Lui S, Tiana G. The network of stabilizing contacts in proteins studied by coevolutionary data. *J Chem Phys* **139**, 155103 (2013).

59. Contini A, Tiana G. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J Chem Phys* **143**, 025103 (2015).

60. Mallik S, Das S, Kundu S. Predicting protein folding rate change upon point mutation using residue-level coevolutionary information. *Proteins-Structure Function and Bioinformatics* **84**, 3-8 (2016).

61. Cheng RR, Raghunathan M, Noel JK, Onuchic JN. Constructing sequence-dependent protein models using coevolutionary information. *Protein Sci* **25**, 111-122 (2016).

62. Feinauer C, Szurmant H, Weigt M, Pagnani A. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *PloS one* **11**, e0149166 (2016).

63. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution* **33**, 268-280 (2016).

64. Hopf TAI, John B.; Poelwijk, Frank J.; Springer, Michael; Sander, Chris; Marks, Debora S. Quantification of the effect of mutations using a global probability model of natural sequence variation. *arXiv*, (2015).

65. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38**, 606-617 (2013).

66. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology* **320**, 369-387 (2002).

67. Caulfield TR, Fiesel FC, Moussaud-Lamodiere EL, Dourado DFAR, Flores SC, Springer W. Phosphorylation by PINK1 Releases the UBL Domain and Initializes the Conformational Opening of the E3 Ubiquitin Ligase Parkin. *PLoS computational biology* **10**, (2014).

68. Fiesel FC*, et al.* Structural and Functional Impact of Parkinson Disease-Associated Mutations in the E3 Ubiquitin Ligase Parkin. *Hum Mutat* **36**, 774-786 (2015).

69. Flores SC, Altman RB. Turning limited experimental information into 3D models of RNA. *Rna* **16**, 1769-1778 (2010).

70. Flores SC, Wan Y, Russell R, Altman RB. Predicting RNA structure by multiple template homology modeling. *Pac Symp Biocomput*, 216-227 (2010).

71. Tek A, Korostelev AA, Flores SC. MMB-GUI: a fast morphing method demonstrates a possible ribosomal tRNA translocation trajectory. *Nucleic Acids Res* **44**, 95-105 (2016).

72. Flores SC. Fast fitting to low resolution density maps: elucidating large-scale motions of the ribosome. *Nucleic Acids Res* **42**, e9 (2014).

73. Flores SC, Sherman MA, Bruns CM, Eastman P, Altman RB. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. *Ieee Acm T Comput Bi* **8**, 1247-1257 (2011).