Coevolutionary landscape inference
and the context-dependence of mutations in beta-lactamase TEM-1

M. Figliuzzi,[1,2,3] H. Jacquier,[4,5] A. Schug,[6] O. Tenaillon,[4] M. Weigt[*,2,3]

[1]*Sorbonne Universités, UPMC, Institut de Calcul et de la Simulation, Paris, France*
[2]*Sorbonne Universités, UPMC, UMR 7238, Computational and Quantitative Biology, Paris, France*
[3]*CNRS, UMR 7238, Computational and Quantitative Biology, Paris, France*
[4]*INSERM, IAME, UMR 1137, Paris, France*
Université Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, Paris, France
[5]*Service de Bactériologie-Virologie, Groupe Hospitalier Lariboisiére-Fernand*
Widal, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France
[6]*Steinbuch Centre for Computing, Karlsruhe Institute for Technologie, Eggenstein-Leopoldshafen, Germany*
[*] *E-mail: martin.weigt@upmc.fr*

The quantitative characterization of mutational landscapes is a task of outstanding importance in evolutionary and medical biology: It is, e.g., of central importance for our understanding of the phenotypic effect of mutations related to disease and antibiotic drug resistance. Here we develop a novel inference scheme for mutational landscapes, which is based on the statistical analysis of large alignments of homologs of the protein of interest. Our method is able to capture epistatic couplings between residues, and therefore to assess the dependence of mutational effects on the sequence context where they appear. Compared to recent large-scale mutagenesis data of the beta-lactamase TEM-1, a protein providing resistance against beta-lactam antibiotics, our method leads to an increase of about 40% in explicative power as compared to approaches neglecting epistasis. We find that the informative sequence context extends to residues at native distances of about 20 from the mutated site, reaching thus far beyond residues in direct physical contact.

Keywords: Mutational landscape, genotype-phenotype mapping, epistasis, coevolution, statistical inference

INTRODUCTION

Protein mutational landscapes are genotype-to-phenotype mappings quantifying how mutations affect the biological functionality of a protein. They are closely related to fitness landscapes describing the replicative capacity of an organism as a function of its genotype [1]. Their comprehensive and accurate characterization is a task of outstanding importance in evolutionary and medical biology: It has a key role in our understanding of mutational pathways accessible in the course of evolution [2–4], it can lead to the identification of genetic determinants of complex diseases based on rare variants [5], and it can guide towards the understanding of the functional contribution of molecular alterations to oncogenesis [6]. In the context of antibiotic resistance, one of the most challenging problems in modern medicine, the understanding of the association between genetic variation and phenotypic effects can help to unveil patterns of adaptive mutations of the pathogens to gain drug resistance, and thereby hopefully guide toward the discovery of new therapeutic strategies [7].

One key issue in the description of a mutational landscape is to understand how much the effect of a mutation depends on the genetic background in which it appears [3, 8, 9]. For instance, in the field of human genetic diseases, is the presence of a mutation enough to predict a pathology or do we have to know the whole genotype to make that assertion? In a more formal way, this question is equivalent to quantifying how epistasis, i.e. the interaction between mutations through fitness, is shaping the mutational landscape. At the protein level, a destabilizing mutation might have a negligible pheno-typic effect in a very stable protein, but a large one in an unstable protein [10, 11]. If this destabilizing mutation increases, e.g., the enzymatic activity, it will be beneficial in a stable protein, and deleterious in an unstable one, cf. [12]. Hence the mutation is expected to be context dependent. Moreover, once a mutation has fixed, further mutations will build upon the specificity of that focal mutation, thereby creating a new genetic background with its specific interactions and interdependencies [13]. There are ample proofs of the existence of epistasis and condition dependent effects [12, 14–17]. Yet, it is not totally clear whether such interactions have a dominant or a minor effect in determining a mutation's phenotypic impact.

Recent technological advances have made it possible to simultaneously quantify the effects of thousands to hundreds of thousands of mutants through either growth competition [16, 18–21] or isolated allele experiments [11, 22, 23]. Experimental resolution can be good enough to detect even the effects of synonymous mutations [22]. Despite the development of such high-throughput methods, measured genotypes cover only a tiny fraction of sequence space: The number of possible mutants grows exponentially with the number of single mutations, such that checking the viability of all possible genotypes further than one or two mutations away from a reference sequence becomes infeasible, even for short polypeptides. More precisely, the number of distinct single-residue mutants for typical proteins is in the range of $10^3 - 10^4$. The number of all double mutants reaches the range of $10^6 - 10^8$. While this number is not yet experimentally accessible, it is needed to accurately assess the importance of epistasis. It has

been argued that existing mutagenesis data are not sufficient for accurate landscape regression [24]. Novel computational approaches exploring alternative data – in our case distant homologs – are thus urgently needed to gain a comprehensive picture of mutational landscapes. In this context, the growing amount of mutagenesis data offers the possibility to rigorously evaluate the performance of *in-silico* models of mutational landscapes.

Several computational methods for predicting mutational effects on protein function have been proposed over the years. A first class relies on *structural* information, more precisely on changes in the thermodynamic stability [25–30], which have been argued to play a key role in determining mutational effects [31–34]. A second class [35, 36] relies on *evolutionary* information extracted from independently evolving homologous proteins, showing variable amino-acid sequences but conserved structure and function. Evolution provides a multitude of informative 'experiments' on mutational landscapes. Critically important residues tend to be conserved, while unfavorable residues are observed less frequently.

None of these methods is able to model the effects of epistasis and sequence-context dependence of mutational effects. To overcome this limitation, we take inspiration from a recent development in structural biology. It has been recognized that coevolutionary information contained in large families of homologous proteins allows to extract accurate structural information from sequences alone [37]: Residues in contact in a protein's fold, even if distant along the primary sequence, tend to show correlated patterns of amino-acid occurrences. Inversely, correlated residues are not necessarily in contact, since correlations are inflated by indirect effects. Two residues, both being in contact to a third residue, will coevolve even if they are not in direct contact. The Direct-Coupling Analysis (DCA) [38, 39] has been proposed to disentangle such indirect effects from direct (i.e. epistatic) couplings, which in turn have been observed to accurately predict residue-residue contacts. DCA and closely related methods thereby guide tertiary [40–43] and quaternary [44–47] protein structure prediction; and shed light on specificity and crosstalk in bacterial signal transduction [48, 49].

In this paper we propose a variant of DCA which assigns to each mutant sequence a statistical score, which in a next step

In technical terms, a mutational landscape is given as a genotype-to-phenotype mapping. To each possible amino-acid sequence $(a_1, ..., a_L)$ consisting of L amino acids or gaps (L denotes the alignment width), a quantitative phenotype $\phi(a_1, ..., a_L)$ is assigned. The phenotypic effect of a mutation substituting the wild-type amino acid a_i at position i with amino acid b is measured by the difference score

$$\Delta\phi(a_i \to b) = \phi(a_1, ..., a_{i-1}, b, a_{i+1}, ..., a_L) - \\ - \phi(a_1, ..., a_{i-1}, a_i, a_{i+1}, ..., a_L) \quad (1)$$

between the mutant and the wild-type sequence. This function ϕ has, however, 20^L parameters, an astronomic number being far beyond any possibility of inference from data. Simplified parameterizations of ϕ reducing the number of parameters are needed. In general, a simple model can be inferred more robustly from limited data, but it risks to miss important effects.

is used for predicting the phenotype of the mutant sequence relative to the wild-type sequence. To evaluate the approach, we take the *Escherichia coli* beta-lactamase TEM-1, a model enzyme in biochemistry which provides resistance to beta-lactam antibiotics. Its mutational landscape has been quantitatively characterized measuring the minimum inhibitory concentration (MIC) of the antibiotic [11, 22, 50]. This abundance of mutagenesis data, the rich homology information and its well defined 3D structure make it a well-suited system for testing any computational model of protein mutational landscapes.

We will show that coevolutionary models for mutational landscapes do not only provide quantitative predictions of mutational effects but, more importantly, they are able to capture the context dependence of these effects. In this way, the new approach manages to clearly outperform state-of-the-art approaches like SIFT [36] and PolyPhen-2 [35], which are based on independent-site models (even if, like in the case of PolyPhen-2, additional structural information is integrated into the prediction of mutational effects), which themselves outperform predictors based on structural stability. The approach is broadly applicable, as is illustrated in a small set of completely different systems: a RNA recognition motif [20], the glucosidase enzyme [23] and a PDZ domain [18]. In the last system positions most sensitive to mutation had been shown previously to fall into clusters of coevolving residues termed sectors [51]: Appling statistical inference we are able to get a more *quantitative* prediction of the impact of single point mutations in the domain. These findings illustrate the potential of coevolutionary landscape models in biomedical applications, via the *in-silico* prediction of mutational effects not only related to antibiotic drug resistance, but also to the role of mutations in rare diseases and cancer.

RESULTS

Evolutionary modeling of diverged beta-lactamase sequences to predict mutational effects of single-residue mutations in TEM-1

The pipeline of our approach is illustrated in Fig. 1. Even if these might be captured in more complex models, the latter risk to suffer from undersampling and thus overfitting effects. One of our aims is to find a good compromise between these two limitations.

The simplest non-trivial parametrization assumes position-specific but independent contributions of each residue,

$$\phi_{IND}(a_1, ..., a_L) = \sum_{i=1}^{L} \varphi_i(a_i) . \quad (2)$$

The contribution $\varphi_i(a_i)$ measuring the contribution of amino acid a_i in position i can be easily estimated from a multiple-sequence alignment (MSA) of homologous proteins using the framework of profile models (also called position-specific weight matrices), cf. *Methods* for details. Possibly existing epistatic effects are neglected. Within this modeling scheme, the score for a single amino-acid substitution simplifies from
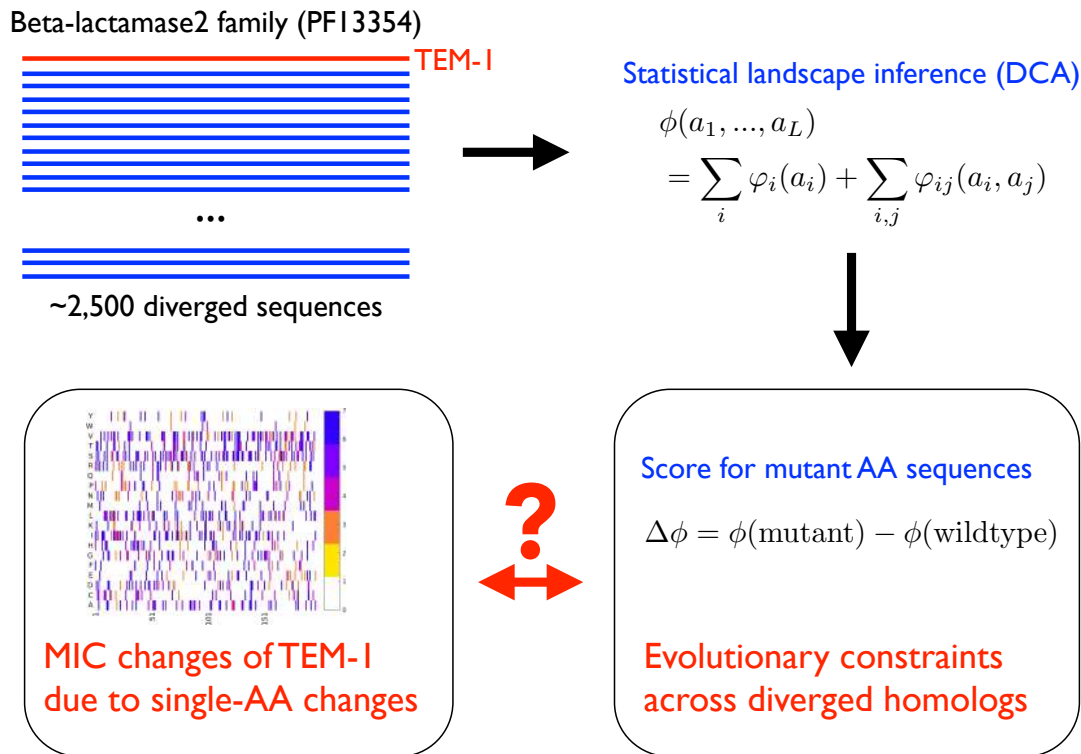
FIG. 1. **Pipeline of the mutational-landscape prediction**: The homologous Pfam family containing the protein of interest (the Beta-lactamase2 family PF13354 in the case of TEM-1) is used to construct a global statistical model using the Direct-Coupling Analysis (DCA). This model allows to score mutations by differences in the inferred genotype-to-phenotype mapping between the mutant and the wild-type amino-acid sequence. This score, which is expected to incorporate (co-)evolutionary constraints acting across the entire family, is used as a predictor of the phenotypic effects of single (or few) amino-acid substitutions in the protein of interest.

Eq. (1) to $\Delta\phi_{IND}(a_i \to b) = \varphi_i(b) - \varphi_i(a_i)$. It becomes immediately evident that the independent-residue model is unable to capture the context dependence of mutations, the substitution $a_i \to b$ is predicted to have identical effects if introduced into different sequence backgrounds. The score of a double mutation is simply given by the sum of the $\Delta\phi$-values of the two single-residue mutations.

The relation between statistically derived scores $\Delta\phi$ and the experimental MIC values may be nonlinear. The discrete nature of the latter introduces saturation effects, in particular for strongly deleterious mutations with MIC values below the lowest measured antibiotic concentration. To address these issues, we have designed a robust mapping of $\Delta\phi_{IND}(a_i \to b)$ to predicted MIC values $\hat{\mu}_{IND}(a_i \to b)$, cf. *Methods*, and compared them to the experimental MIC values $\mu_{exp}(a_i \to b)$ by linear correlation. A direct measurement of Spearman rank correlations between ϕ_{IND} and μ_{exp} leads to numerically very similar, but slightly less robust results.

The MIC predictions using model Eq. (2) show a Pearson correlation of $R = 0.63$ with the experimental MIC mea-

surements of single-residue substitutions in TEM-1. About $R^2 \simeq 39\%$ of the variability of the experimental results is thus explainable by an independent-site model built on the sequence variability between homologous sequences. Very similar correlations ($R^2 = 0.37$) are found when comparing experimental results and the probabilities of being tolerated as predicted by SIFT, which, like most state-of-the-art methods, is based on conservation profiles in sequence alignments. Higher accuracy is found for PolyPhen-2 ($R^2 = 0.48$): its improved performance results from the integration of a profile-based score with structural features and amino-acid properties.

However, all these predictions are based on the assumption that epistasis between mutations and context dependence can be neglected. The simplest model to challenge this assumption takes into account *pairwise epistatic interactions* between

different residue positions in the MSA,

$$\phi_{DCA}(a_1, ..., a_L) =$$
$$= \sum_{i,j=1}^{L} \varphi_{ij}(a_i, a_j) + \sum_{i=1}^{L} \varphi_i(a_i) \,, \qquad (3)$$

cf. *Methods*. The terms $\varphi_{ij}(a_i, a_j)$ parametrize the epistatic couplings between amino acids a_i and a_j in aligned positions i and j; if they would be set to zero the model would reduce to the independent-site model ϕ_{IND}. This model has been recently introduced within the *Direct-Coupling Analysis* (DCA) of residue coevolution with the aim to infer contacts between residues from sequence information alone, and to enable the prediction of tertiary and quaternary protein structures, cf. the references in the *Introduction* of this paper.

Estimating parameters from aligned sequences is a computationally hard task, but over the last years a number of accurate and computationally efficient approximate algorithms have been developed [38, 39, 52, 53]. Here we extend the mean-field scheme of Morcos et al. [39], cf. *Methods*. For TEM-1, standard DCA analysis accurately predicts tertiary contacts, cf. Fig. S1: More than 60 non-trivial residue-residue contacts (minimum separation of 5 residues along the sequence) are predicted without error, and more than 200 at a precision of 80%.

Having estimated ϕ_{DCA} from the MSA, we can follow the same strategy as in the independent-residue case. First, a mutational score is introduced as the difference of the ϕ-values of the mutated and the wild-type sequences, cf. Eq. (1). The inclusion of epistatic couplings leads to an *explicit context dependence* of the statistical score of a mutation $a_i \rightarrow b$ in position i on all other residues in the wild-type sequences,

$$\Delta\phi_{DCA}(a_i \rightarrow b \,|\, a_1, ..., a_{i-1}, a_{i+1}, ..a_L) =$$
$$= \varphi_i(b) - \varphi_i(a_i) + \sum_{j=1}^{L} \left[\varphi_{ij}(b, a_j) - \varphi_{ij}(a_i, a_j) \right] \,. \qquad (4)$$

In a second step, this difference score is mapped to predicted MIC values $\hat{\mu}_{DCA}(a_i \rightarrow b)$ and compared to the experimental values $\mu_{exp}(a_i \rightarrow b)$ by linear correlation.

Resulting predictions outperform the independent-residue modeling. DCA-predicted MIC values show a correlation of $R = 0.74$ with the experimental MIC measurements of single-residue substitutions in TEM-1, i.e. about $R^2 \simeq 55\%$ of the variability of the experimental results is explained by the DCA-inferred mutational landscape, see Fig. 3, as compared to the 39% reported before for the IND model. We find that DCA even outperforms the integrative modeling of PolyPhen-2 combining sequence profiles with structural and other prior biological knowledge, demonstrating the power of

It is interesting to observe that the IND model makes more predictions with very large deviations from the experimental data than the DCA model: There is an increased number of mutations, which are either predicted to be strongly deleterious even if they are close to neutral, or vice versa. Many of these strong errors are at least partially corrected by the

DCA in capturing epistatic effects in the TEM-1 mutational landscape.

Applying the same procedure to the data of Firnberg et al. [22], which are highly correlated with the data from Jacquier et al. [11] ($R = 0.94$), but slightly more precise than that, the correlation is slightly higher ($R = 0.76$, $R^2 = 0.58$). Excluding from the analysis those data which display large discrepancies between the two experiments (such discrepancies could be either due to experimental errors or due to antibiotic-specific effects) correlations between our computational score and both datasets rise above $R^2 = 0.65$, cf. Supplementary Fig. S2.

We conclude that sequence variability in the Pfam sequence alignments of distant homologs is highly informative about the local mutational landscape of TEM-1, despite the low typical sequence identity of only about 20% between the homologs and TEM-1. Moreover, accounting for context dependence has a crucial impact on the accuracy of an evolution-based approach, and that global inference methods like DCA can efficiently capture such dependencies.

Assessing the context dependence of mutational effects

To quantify more precisely the range of context dependence, we apply DCA to reduced MSA. These MSA contain the residue position carrying the mutation of interest, and all residues, which are, in a representative TEM-1 crystal structure (PDB: 1M40 [54]), within a distance d_{max} (we use the minimal distance between heavy atoms as the inter-residue distance). When using a very small $d_{max} \leq 1.2$, the mutated residue is considered on its own, when d_{max} is chosen to be larger than the maximum distance 46.9 existing within the PDB structure, we are back to the full DCA modeling of the previous section. Intermediate d_{max} interpolate between the two extreme cases. Doing so, we run DCA on sub-alignments of residues, which are not necessarily consecutive in the primary sequence but connected in the native fold, cf. the illustration of the procedure in Panel A of Fig. 2. Panel B shows the resulting correlations between MIC data and statistical predictions, in function of the cutoff distance d_{max}. We observe a rapid increase in predictive power when a structural neighborhood is taken into account, but the increase in correlation extends well beyond the directly contacting residues ($d_{max} \simeq 6$). The maximum correlation ($R^2 \simeq 0.57$) is reached around $d_{max} \simeq 20$, followed by a shallow decrease when including also more distant residues. This small decrease results probably from overfitting effects, since the number of model parameters grows quadratically in sequence length. The insert of Panel B shows the average fraction of residues included into the sub-MSA. At 20 it is slightly higher than 50%, i.e. the informative context of a mutation is given by more than half of the total number of residues in the protein.

DCA landscape model (cf. Supplementary Tables S1-S3). By the definition of the independent model in terms of frequency counts in individual MSA sequences, cf. *Methods*, a mutation with a low predicted IND score leads from a more frequent to a rare amino acid in the concerned MSA column. However, in the mutagenesis experiments some of these mutations
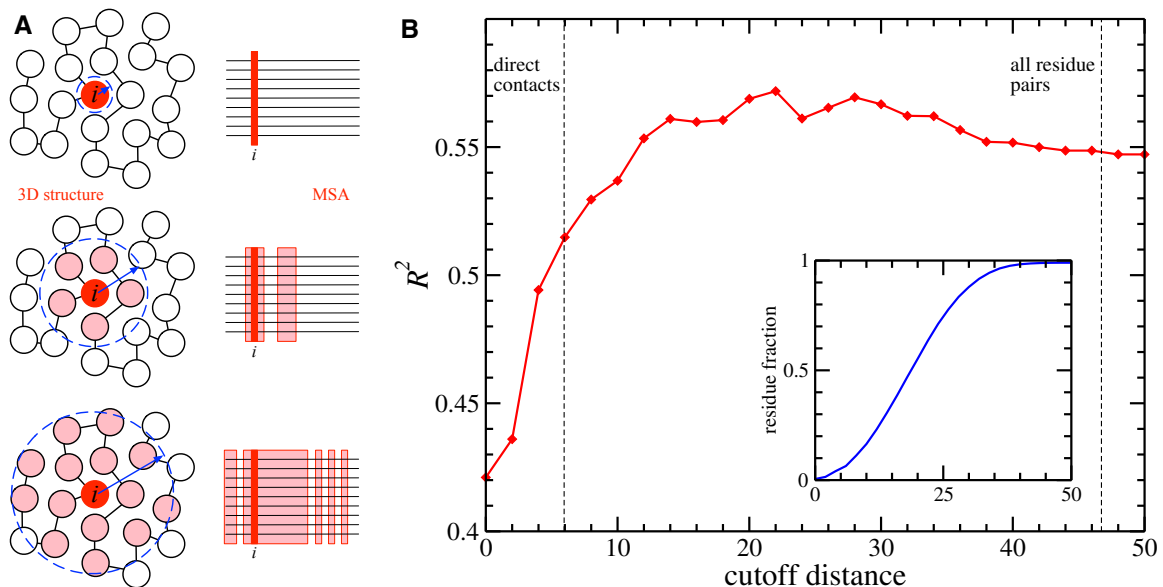
FIG. 2. **Context dependence of mutational effects:** Panel A illustrates the procedure of including all residues within a maximal native distance d_{max} into the prediction of the mutational effects of the residue of interest (labeled i in the figure). This leads to residue-specific sub-alignments, which consist of columns, which are not necessarily consecutive, but connected in 3D. The results are given in Panel B. The main figure shows the correlation R^2 between MIC data and our predictions, as a function of the cutoff distance d_{max}. The insert shows the average fraction of residues included into the reduced MSAs, again in dependence of d_{max}.

are found to be admissible in the specific sequence context of TEM-1, i.e. they are actually found to be close to neutral, examples being G52A, E61V, T112M, N152Y, A183V, T186P, D207V, D250Y (all target amino-acids are present in few tens of sequences in the MSA out of the about 2500 functional homologous sequences). For all of these cases, DCA is able to correct at least partially the statistical prediction. On the contrary, the independent-site model predicts that any mutation between two amino acids of similar frequency in the corresponding MSA column is close to neutral. Looking to the experimental MIC, substitutions D177N, A235D, I243N and G248E all predicted to be close to neutral, have strongly deleterious effects (MIC≤25). DCA corrects the mispredictions by at least two, on average by three MIC classes.

There is a small set of 9 mutations badly predicted by DCA. In none of these cases the independent modeling significantly ameliorates predictions. Interestingly, 6 out of these 9 mutations fall into the highly gapped part of the MSA: DCA display a significant loss of predictive power in the highly gapped positions of the MSA, and correlation between predicted and experimental MIC increases above $R^2 = 0.75$ when disregarding mutations in this region (see Supplementary Fig. S3).

Structural-stability predictions show lower correlations to MIC changes than sequence-based modeling

It has been proposed before that the role of most residues is to make the protein properly fold, and that mutations on these sites mainly alter protein stability and not its activity [31]: Hence an accurate estimation of the change in protein stability $\Delta\Delta G \equiv \Delta G^{mut} - \Delta G^{wt}$ should be able to account for a large fraction of mutational effects.

Many bioinformatic programs have been developed for estimating protein stability change upon mutation: among them MUpro [25] and I-Mutant2.0 [26], which take the sole sequence as input, PoPMuSiC [28] and I-Mutant2.0(sequence+structure)[26], which consider both sequence and structure. Since these methods show incoherent predictions in between each other, cf. Supplementary Fig. S4, we complement them by extensive force-field molecular simulations at all-atom resolution to estimate protein stability changes $\Delta\Delta G$ induced by single point mutations; cf. *Methods* for details. A score can be assigned to any substitution of amino acid a_i in position i by amino acid b,

$$\Delta\phi_{stab}(a_i \rightarrow b) \equiv -\Delta\Delta G(a_i \rightarrow b) , \qquad (5)$$

and then mapped to predicted MIC values $\hat{\mu}_{stab}(a_i \rightarrow b)$ using the before-mentioned scheme. Pearson correlations between predicted and experimental MIC are calculated: We find that, while those methods which consider not only sequence but also structural information ($R^2 = 0.13$ for PoPMuSiC and $R^2 = 0.14$ I-Mutant2.0(sequence+structure)) largely outperform those who do not ($R^2 \sim 0.02$ for MUpro and I-Mutant), one gets only a modest further improvement letting the mutated polypeptide relax via molecular simulations ($R^2 = 0.17$ for molecular simulations, see Fig. 3).

It is well known that residues buried in the protein core are important determinants of protein stability. Mutation affecting these sites tend to be highly destabilizing [55–58]. Therefore, we test also to what extent solvent accessibility explains the

experimental mutation effects. Upon defining

$$\Delta\phi_{RSA}(a_i \rightarrow b) = \alpha_i , \qquad (6)$$

where α_i is the relative solvent accessible surface area (RSA) of residue a_i in position i. We use Michel Sanner's Molecular Surface (MSMS) algorithm [59] applied to the PDB structure 1M40 to estimate surface accessible surface areas (SAS), normalized by the maximum accessibilities given in [60]. We find that $R^2 = 0.20$ of the variability of the experimental fitness is explainable via RSA. In general, we find that different accessibility estimates provide very similar results, including the absolute SAS, cf. the *Supplement*. Indeed, a simple binary classifier roughly distinguishing buried from exposed residues is almost as informative as RSA and SAS values (Fig. S5). Note that the score $\Delta\phi_{RSA}$ does not depend on the target amino acid b, but only on the wild-type structure. Note also that this R^2-value, while been greater than those achieved through molecular simulations, is substantially smaller than all statistical sequence scores derived from homologs.
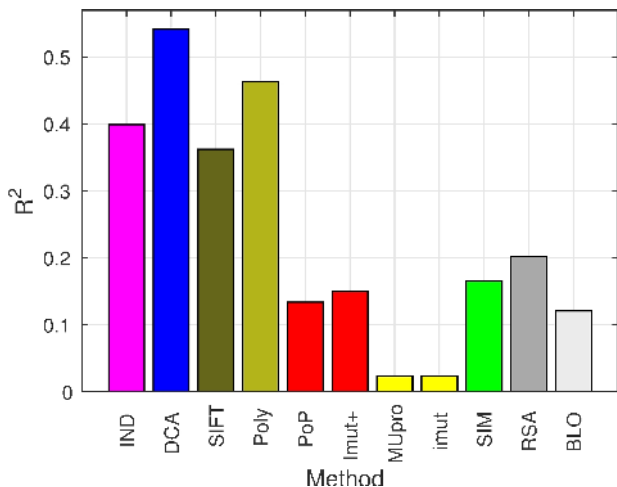

FIG. 3. R^2 **between experimental fitness and predicted fitness** for the following features: Independent-residue model (IND), Direct-Coupling Analysis (DCA), SIFT (SIFT), Polyphen-2 (Poly), PoPMuSiC (PoP), I-Mutant2.0(sequence+structure) (Imut+), MUpro (MUpro), I-Mutant2.0 (Imut), molecular simulations (SIM), relative solvent accessibility (RSA) and Blosum62 substitution matrix (BLO).

The failure of stability-based predictions of mutational effects may result from strong-effect mutations in or close to the active site, whose phenotypic effect is unrelated to protein stability. To assess this effect, we have repeated our analysis including only 111 mutations falling into the extended active site, cf. the Supplementary Fig. S6 for details. The R^2-values for both statistical models (IND and DCA) go up strongly ($R^2_{IND} = 0.52, R^2_{DCA} = 0.67$), while the structure-based predictors show little or no gain at all. This demonstrates, that evolutionary information accurately predicts the effects of mutation falling into the active site, and structural information does not.

Being grounded on complementary sources of information, predictions by evolution- and structure-based methods are not strongly correlated, as shown in Supplementary Fig. S4. A linear combination of DCA with structural predictors, however, yields only little increase in correlation: the explained variance of experimental data gets to $0.60 \sim 0.61$ when performing a bivariate linear regression between DCA scores and either solvent accessibility or Polyphen-2 predictions, as displayed in Supplmentary Fig. S7.

DCA landscape modeling spots stabilizing mutations and captures protein-specific substitution scores

The TEM-1 beta-lactamase has been the subject of intense studies with regard to protein structure, function, and evolution, and a number of structurally stabilizing substitutions have been identified [19, 61–63]: P62S, V80I, G92D, R120G, E147G, H153R, M182T (strongly stabilizing), L201P, I208M, A184V, A224V, I247V, T265M, R275L/Q, and N276D (positions are indicated using standard Ambler numbering [64]). Some of them were found to influence the resistance phenotype [65]. Notably, the five highest DCA scores $\Delta\phi_{DCA}$ out of all considered mutants belong to this set: M182T, H153R, E147G, L201P and G92D (with a large gap separating the likelihood of the strongly stabilizing M182T from the scores of the other four, cf. Fig. 4). More quantitatively, we found that the Gibbs Free Energy change relative to wild type $\Delta\Delta G$ of a different, small set of mutations (most of which not affecting Amoxicillin resistance) characterized by four independent studies [19, 61–63] are highly correlated with DCA scores ($R_{DCA} = 0.81$) but less correlated when using independent model ($R_{IND} = 0.62$).

We further investigate whether the statistical analysis of homologous sequences is able to capture *protein-specific amino-acid substitution effects*, i.e. if the effect of a specific amino-acid substitution (averaged over all sequence positions where this mutation appears) is better described by our statistical model than it would be by Blosum matrices, which are estimated from many distinct aligned protein sequences. To this aim, a matrix of average substitution scores is built from the set of experimental MIC values, cf. Fig. 5. We also construct an analogous matrix for the DCA-predicted MIC values of the same set of mutations, and quantify correlations between predicted and experimental average effects computing a Pearson correlation weighting each term with the square root of the number of measured mutations falling in the related class. We find a very large correlation ($R^2 = 0.72$) between average experimental and predicted substitution matrices. This value has to be compared with the substantially lower correlation found when comparing the mutational effects in TEM-1 with the Blosum62 matrix ($R^2 = 0.34$), which provides amino-acid substitution scores averaged over many proteins. All other inference methods show substitution scores with correlations to MIC, which are comparable to or lower than the correlations between MIC and Blosum62.
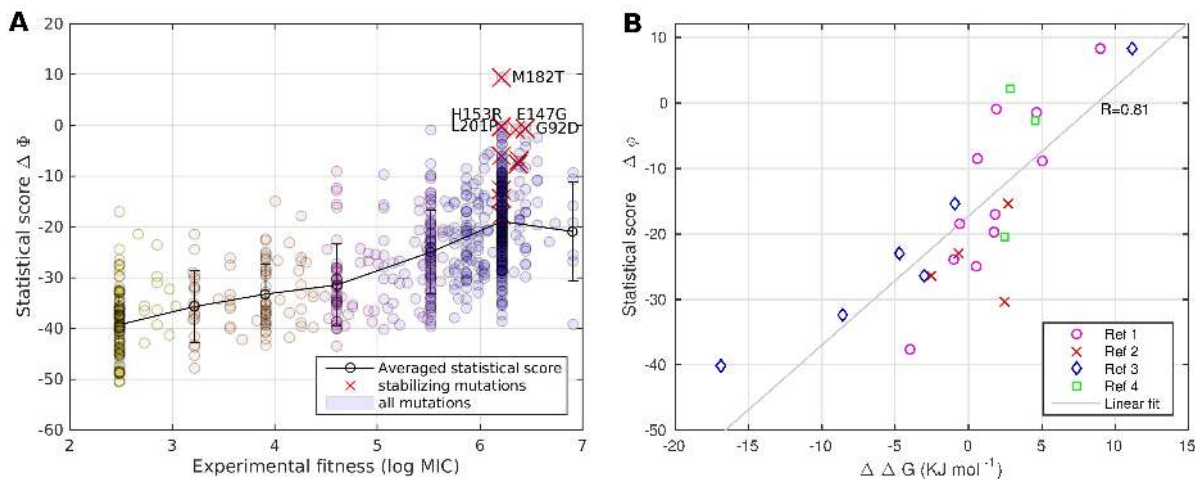
FIG. 4. **Statistical scores and thermodynamic stabilities** Panel A: Scatter plot of the log odd ratio $\Delta\phi_{DCA}$ vs the experimental fitness μ_{exp}, stabilizing mutations mentioned in the text are highlighted in red. The highest scoring mutations are M182T, G92D, H153R, L201P and E147G, all reported as stabilizing. Panel B: $\Delta\phi_{DCA}$ for a smaller set of single mutations is now plotted vs. the change in Gibbs Free Energy relative to wild type $\Delta\Delta G \equiv \Delta G_{mut} - \Delta G_{wt}$, as measured by four independent studies (Ref1, Ref2, Ref3 and Ref4 are [61],[62],[63] and [19] respectively).

DISCUSSION

The central aim of this paper is the accurate computational inference of protein mutational landscapes to predict the phenotypic effect of mutations. This is exemplified in the case of the TEM-1 protein of *E. coli*, a beta-lactamase providing antibiotic drug resistance against beta-lactams, like penicillin, amoxicillin or ampicillin.

To reach this aim, we have extracted information about a protein and its potential mutants, which is hidden in the sequence variability of *diverged but functional* homologs of this protein. The central ingredient of our analysis is a careful modeling of residue coevolution by Direct-Coupling Analysis, i.e. the modeling includes pairwise epistasis between residues. This approach, initially developed in the context of structural biology in order to predict residue-residue contacts from sequences, has been used to define a score for each mutation, which was found to explain 55% resp. 58% of the phenotypic variability in the two corresponding experimental TEM-1 data sets [11, 22]. This value is substantially higher than what can be obtained by a more standard modeling approach based on sequence profiles (39% of variability explained), which does not include epistasis, or on changes in structural stability. Furthermore, our coevolutionary approach clearly outperforms state-of-the-art approaches like SIFT and PolyPhen-2, which are based on non-epistatic models.

However, epistatic effects are not equally important for all residues, which may explain that some authors disagree on the contribution of the sequence context to mutational effect [13, 66, 67]. The relevant context determining the effect of a mutation of a residue is not only given by its direct physical neighbors, but extends to a distance of about 20. The informative context thus includes, on average, roughly half of all residues in the aligned TEM-1 sequence. This result agrees

with the finding that interactions from second shell and beyond might be important for protein function [68]. Having a look to the physico-chemical properties of the wild-type and the mutant amino-acids, we observe, e.g., that mutations substituting a hydrophobic residue with a hydrophilic one are almost equally well described by the DCA and by the independent model ($R^2_{DCA} - R^2_{IND} \simeq 5\%$), due to the structurally highly disruptive effect of a hydrophilic residue in a buried site, and thus the absence of hydrophilic residues in the corresponding column of the sequence alignment. On the contrary, the more moderate effect of replacing a small by a large amino acid depends strongly whether the context is able to accomodate this change or not, and thus the independent model performs much worse than the DCA model ($R^2_{DCA} - R^2_{IND} \simeq 26\%$). Concentrating on mutations from amino acids of given physico-chemical characteristics (hydrophobicity, charge, volume) toward a target amino acid of either different (e.g. hydrophobic to hydrophilic) or conserved characteristics (e.g. hydrophobic to hydrophobic) we find that the DCA predictions are stable, with R^2-values between 49 and 64%, while the ones of the IND model vary much more strongly (25-55%). In none of the considered cases, the independent model was able to outperform the coevolutionary one.

Our findings demonstrate that the *local mutational landscape* dictating the mutational effects in TEM-1 is closely related to the (co-)evolutionary pressures acting globally across the entire homologous protein family. This result is quite remarkable: Despite a low typical sequence identity of about 20% between homologous beta-lactamases and TEM-1, their sequence statistics provides quantitative information about the effect of single-residue substitutions in TEM-1. We are thus able to infer landscapes and predict quantitatively mutational effects even in cases, where mutaganesis data are not sufficiently numerous, cf. [24]. This complements recent findings,
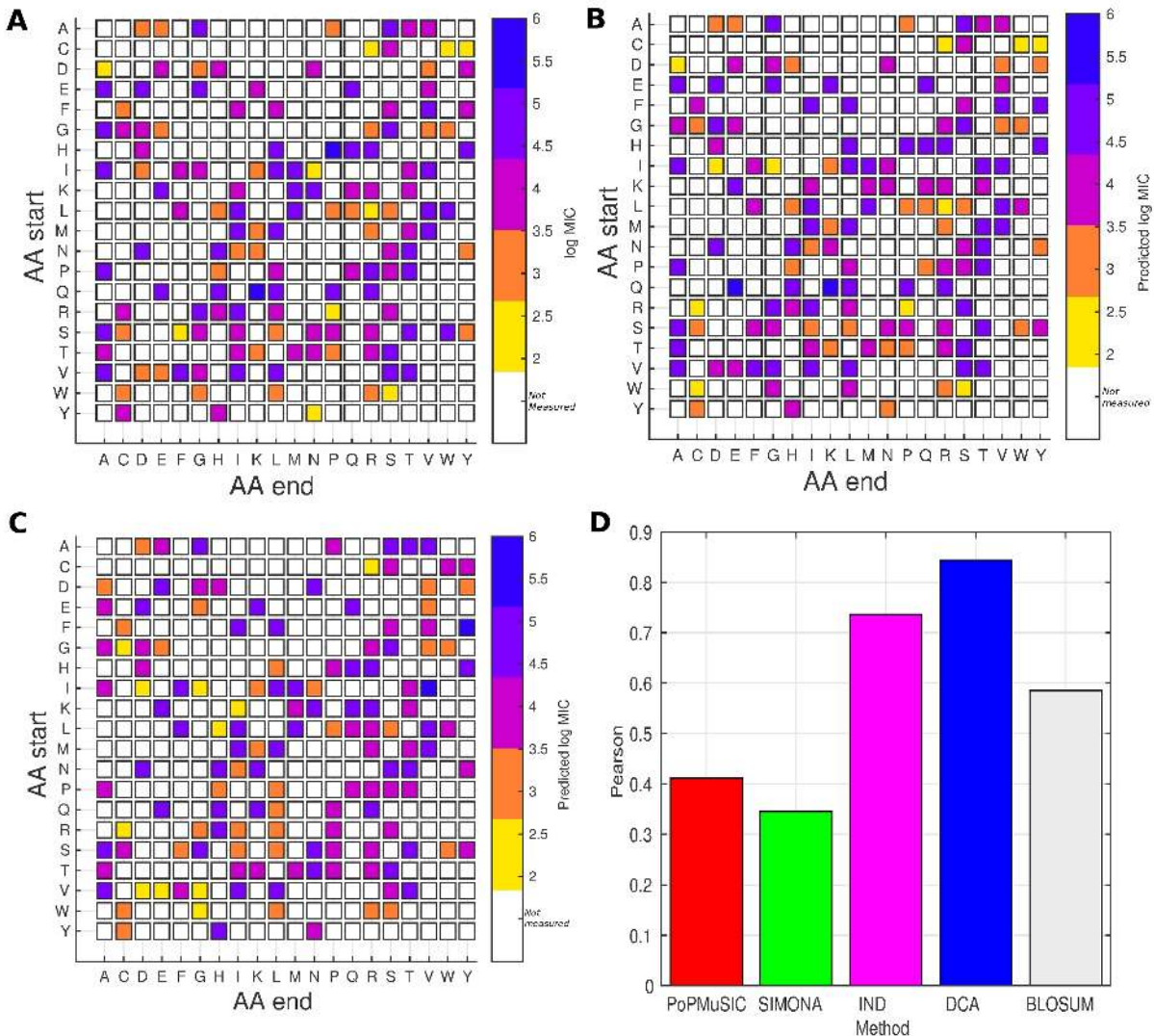
FIG. 5. **Protein-specific amino-acid substitution effects in TEM-1:** Amino acid substitution effects, averaged over experimental measurements (Panel A), DCA predictions (Panel B), and extracted from BLOSUM62 (Panel C). Blues squares correspond to nearly neutral mutations (log MIC > 5.3), while yellow squares correspond to highly deleterious mutations (log MIC < 2.6). White squares are used for unobserved substitutions. The histogram in Panel D shows R^2 between averaged computational and experimental amino-acid substitution effects.

that patterns of polymorphism and covariation in patient derived (and thus highly similar) HIV sequences are informative about their replicative capacities [69, 70], thanks to high mutation rates in the HIV virus. Further more, coevolutionary patterns in protein families were recently found to be closely related to protein energetics and folding landscapes [71, 72].

We expect that the modeling approach via DCA can be improved along several lines. First, prediction accuracy depends critically on the quality and size of the training multiple-sequence alignment. As we have shown, the prediction for gapped (and typically less well-aligned) positions is substantially worse than the one for ungapped (thus better alignable) ones (R^2-values raging from 30% to 78% from the most to the least gapped positions). We therefore excluded gapped

sequences from the training alignment, but this procedure reduces the sequence number and thus the statistics for the un-gapped positions.

Second, the current DCA approach is purely statistical and based on evolutionary information. It does not take into account any complementary knowledge about the protein under study. We have, however, observed that the integration of structural knowledge helps to increase the prediction accuracy. Fitting the model only for residues within about 20 from the mutated residue, the R^2-value raises slightly by about 2%. The effect of integrating the DCA-score and the solvent-accessible surface area is even larger, leading to a gain in R^2 of more than 6%. A very similar increase (7%) is obtained when combining DCA with PolyPhen-2, the latter being built upon

a profile model and structural information. These increases are based on a simple linear regression scheme with threefold crossvalidation: It will be interesting to explore more sophisticated approaches, e.g. integrating prior structural knowledge via a Bayesian inference scheme directly into the statistical-inference procedure.

Even if the integration of complementary information may substantially improve our prediction accuracy, the most important contribution is, however, coming from the careful inclusion of epistatic effects into our modeling approach to mutational landscapes, as shown by a partial-correlation analysis in Fig. S8.

From a computational point of view, the approach is widely applicable beyond the specific case of TEM-1 and antibiotic drug resistance. To check this practically, we have analyzed further systems in the *Supplement*: a PDZ domain [18], a RNA recognition motif [20] and the glucosidase enzyme [23], cf. Supplementary Text S1 and Figs. S9-S11. DCA predictions systematically outperform independent-site models neglecting epistasis and all other tested methods. Only PolyPhen-2 reaches, in two cases out of four, comparable performance. Despite this encouraging finding, correlations between experiment and computation are numerically smaller than those observed for TEM-1. We expect this reduction to result from discrepancies between the measured phenotypes (e.g. protein stability, binding affinity) and those under evolutionary selection (fitness); MIC is without doubt a better proxy for fitness than most molecular phenotypes. However, to systematically support this idea, large-scale experiments assessing the impact of mutations on multiple phenotypic traits in the same protein would be necessary. In summary, despite not representing a comprehensive survey, currently available data suggest a large potential for coevolutionary models in biomedical applications, via the *in silico* prediction of the role of mutations in rare diseases and cancer.

METHODS

Data

Mutational data

The original dataset [11] was used directly at the translated amino-acid level. It contains 8621 (4094 distinct) measurements of amoxicillin MIC. Among these 8112 do not include stop codons, 2440 are repeated measures of the wild-type sequence, 3129 ($N_{multiple}$ = 2051 distinct) have all mutations inside the part of the sequence covered by the Pfam domain (i.e. subject to the presented statistical analysis). Finally, among the latter set, there are N_{single} = 742 distinct single mutation. Each measurement z_i falls in 9 discrete classes: $12.5, 25, 50, 125, 250, 500, 1000, 2000, 4000$ (mg/L) (no single point mutation has $z > 1000$). For a given phenotype where amino acid a_i in position i is replaced with amino acid b we have defined a unique experimental fitness $\mu_{exp}(a_i \to b)$ taking the logarithmic average on all measurements (whenever multiple measurements were available):

$$\mu_{exp}(a_i \to b) = \frac{1}{N(a_i \to b)} \sum_{i=1}^{N(a_i \to b)} log(z_i) \qquad (7)$$

where $N(a_i \to b)$ is the number of measurements of mutation $a_i \to b$.

Homologous sequences and preprocessing of the training set

The genomic model was learned from a multiple sequence alignment (MSA) of sequences belonging to the Pfam Beta-lactamase2 family (PF13354) [73]. We have used HMMer [74] to search against the Uniprot protein sequence database (version updated to March 2015). The resulting MSA is L = 197 sites long, and contains 5119 distinct sequences. After removing all sequences with more than 5 gaps, 2462 sequences are retained and used for the statistical analysis. They have an average sequence identity $\sim$ 20% with the TEM-1 wild-type sequence.

Statistical sequence modeling

Independent model – sequence profile

The basic assumption of the independent model Eq. (2) is the additivity of the mutational effects of different positions in the amino-acid sequence. In terms of statistical sequence models, this corresponds to a *sequence profile model*, which assigns to each sequence the factorized probability

$$P_{IND}(a_1, ..., a_L) = \prod_{i=1}^{L} f_i(a_i) \qquad (8)$$

with $f_i(a)$ being the frequency of aminoacid a in column i of the MSA, see below for a precise definition of this frequency. The factorized form of this expression suggests to use log-probabilities as a computational predictor of the genotype-to-phenotype mapping,

$$\phi_{IND}(a_1, ..., a_L) = \log P_{IND}(a_1, ..., a_L) . \qquad (9)$$

This leads to an explicit expression of the phenotypic contribution of amino acid a in site i: $\varphi_i(a) = \log f_i(a)$.

Epistatic model – Direct-Coupling Analysis

Following last paragraph's idea to identify the computational predictor of the genotype-to-phenotype mapping with the log-probability of a statistical model inferred from an MSA of TEM-1 homologs, the latter takes the form

$$P_{DCA}(a_1, ..., a_L) = \frac{1}{Z} \exp\{\phi_{DCA}(a_1, .., a_L)\} \qquad (10)$$

where

$$\phi_{DCA}(a_1, ..., a_L) = \sum_{i=1}^{L} \varphi_i(a_i) + \sum_{i,j>j=1}^{L} \varphi_{ij}(a_i, a_j) \qquad (11)$$

is given in Eq. (3), and the so-called partition function $Z = \sum_{a_1,...,a_L} \exp\{\phi_{DCA}(a_1, .., a_L)\}$ is a normalization factor. The statistical model P_{DCA} thus takes the form of a generalized Potts model or, equivalently, a pairwise Markov random field. The same model was introduced in the Direct-Coupling Analysis of residue coevolution [38, 39]. Inferring model parameters φ from the MSA is a computationally hard task, we therefore follow the mean-field approximation introduced in [39]. In this context, the epistatic couplings can be determined by inversion of the empirical covariance matrix $C_{ij}(a, b)$ for the co-occurrence of amino-acids a and b in positions i and j of the same protein sequence. Once the model parameters are determined, the context-dependent mutational effects can be estimated using Eq. 4.

Details of statistical inference

To take into account phylogenetic correlations and sampling biases in the training set, each sequence $(a_1^m, ..., a_L^m)$, $m = 1, ..., M$, of the MSA appears in the statistics with the following weight,

$$w_m = \left(1 + \sum_{m \neq m'} \theta\left(d_{m,m'} - \vartheta L\right)\right)^{-1}, \qquad (12)$$

with $d_{mm'}$ being the Manhattan distance (number of mismatches) between sequences m and m' and θ being the Heaviside step function whose value is zero for negative argument and one for positive argument. The reweighting threshold is set to $\vartheta = 0.8$ as usually done in DCA [39]).

Due to finite sampling, the statistics of the MSA has to be regularized introducing pseudocounts:

$$f_{ij}^\lambda(a, b) = \frac{\lambda}{q^2} + \frac{1-\lambda}{M_{eff}} \sum_{m=1}^{M} w_m \delta_{a_i^m,a} \delta_{a_j^m,b} \qquad (13)$$

$$f_i^\lambda(a) = \frac{\lambda}{q} + \frac{1-\lambda}{M_{eff}} \sum_{m=1}^{M} w_m \delta_{a_i^m,a} \qquad (14)$$

with $M_{eff} = \sum_{m=1}^{M} w_m$ and δ the Kronecker's delta whose value is one if the variables are equal, and zero otherwise. We have included pseudocounts at two levels: First, for the inference of epistatic couplings we have used large pseudocounts ($\Lambda_2 = 0.5$), needed to correct for systematic biases introduced by the MF approximation [75],

$$\varphi_{ij}(a, b) = [C^{-1}]_{ij}(a, b) \qquad (15)$$

$$C_{ij}(a, b) = f_{ij}^{\Lambda_2}(a, b) - f_i^{\Lambda_2}(a) f_j^{\Lambda_2}(b) \qquad (16)$$

for all amino acids a and b. Following [76], also diagonal terms $\phi_{ii}(a, b) = [C^{-1}]_{ij}(a, b)$ are included. Couplings with gaps are set to zero, $\varphi_{ij}(a, -) = \varphi_{ij}(-, a) = 0$, cf. [39].

Smaller pseudocounts of Bayesian size ($\Lambda_1 = \frac{1}{M_{eff}}$) have been used in the regularization of single site frequencies to infer the fields:

$$\varphi_i(a) = \log\left(\frac{f_i^{\Lambda_1}(a)}{f_i^{\Lambda_1}(-)}\right) - \sum_{j,b} \varphi_{ij}(a, b) f_j^{\Lambda_1}(b) \qquad (17)$$

The same small regularization $\Lambda_1 = \frac{1}{M_{eff}}$ has been adopted in the independent-site model.

Mapping scores to MIC values

To compare computational predictions with experimental MIC values, we map computational scores $\Delta\phi(a_i \rightarrow b)$ into predicted MIC $\hat{\mu}(a_i \rightarrow b)$, by first sorting them and then associating to the n_{th} highest score $\Delta\phi_{n_{th}}$ the n_{th} highest experimental MIC value $\mu_{exp}(n_{th})$,

$$\hat{\mu}(\Delta\phi_{n_{th}}) = \mu_{exp}(n_{th}) . \qquad (18)$$

We subsequently compute linear correlations between the predicted MIC $\hat{\mu}$ and the experimental one μ_{exp}, resulting in nonlinear rank correlations between experimental fitnesses and raw computational scores $\Delta\phi$.

This procedure has proved to be more robust than the standard Spearman rank correlations, because of the peculiar distribution of experimental data (bimodal with many repeated measures), and helpful to reduce the statistical weight of outliers (such as strongly destabilizing mutations in the distribution of $\Delta\Delta G$ predicted by molecular simulations). However, numerical values of Spearman correlations are in general not very different from those obtained by our procedure.

Structural stability predictions

Bioinformatic predictors

A list of predicted $\Delta\Delta G$ of *E. coli* TEM-1 protein point mutations for the web-based programs mentioned in the article have been downloaded from the SPROUTS database [27].

Force-field based molecular simulations

Computation of protein thermodynamic stability is computationally very demanding: A direct calculation of thermodynamic stability by molecular dynamics simulations implies the sampling of complete folding and unfolding events. This is presently infeasible for proteins of the size of TEM-1 (286 amino acids). An alternative, less expensive approach to estimate mutational effects on pritein stability is to look for locally stable configuration performing small structural relaxations from a reference structure, with the wild type amino acid replaced by the mutant amino acid. Assuming that the protein can be described by a two-state system (folded vs. unfolded), and that both the entropy of the folded and the free

energy of the unfolded are not sensibly affected by the mutation, we can approximate

$$\Delta\Delta G \equiv \Delta G_{wt} - \Delta G_{mut} \simeq E_{wt}^{fold} - E_{mut}^{fold} \equiv \Delta E \ . \qquad (19)$$

Moreover, as thermodynamic stability is an equilibrium property, one can replace expensive molecular-dynamics simulations with more efficient Monte-Carlo sampling.

Molecular simulations were performed using SIMONA [77], a Monte-Carlo based simulation software for efficient molecular simulations which have proved useful to obtain reproducible folding in a series of test cases [78, 79]. As reference structure for molecular relaxations we have taken a highly resolved (0.8) structure (PDB: 1M40 [54]). Further details of the simulations are reported in next section.

Details and calibration of the molecular simulations

To estimate the thermodynamic stability of TEM-1 mutants we have executed the following steps:

1. Starting from a sufficiently close reference state (in our case the SIMONA-relaxed structure of the wild type molecule), the wild-type amino acid is replaced by the mutant one.

2. Monte-Carlo simulations are performed under SIMONA, to locally minimize the energy function.

3. The resulting energy change $\Delta E = E_{mut} - E_{wt}$ is determined.

In the simulation, we have included the complete forcefield PFF03v4-all parallel OpenMP (scale 1.0), which makes use the amber99sb-star-ildn dihedral potential with an implicit solvent model. It contains the following contributions:

$$V(\{\vec{r_i}\}) = \sum_{ij} V_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^6 \right] +$$
$$+ \sum_{ij} \frac{q_i q_j}{\epsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{hBonds} V_{hb}. \qquad (20)$$

where r_{ij} represents the distance between atoms i and j, and $g(i)$ the type of amino-acid i, V_{ij} and R_{ij} are Lennard-Jones parameters, q_i and $\epsilon_{g(i)g(j)}$ are the partial charges and group-specific dielectric constants for non trivial electrostatic interactions, σ_i and A_i are the free energy per unit area and the area of atom i in contact with fictitious solvent respectively, and finally V_{hb} is a short range interaction term for backbone-backbone hydrogen bonding [78].

I. SUPPLEMENTARY MATERIAL

Supplementary Tables S1-S3, Figures S1-S13, Texts S1 and a Matlab implementation of DCA modeling and sequence scoring are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

[1] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. Proceedings of the 6th International Congress of Genetics: 356-366, 1932.

[2] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, 1987.

[3] Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, 2006.

[4] Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.

[5] Elizabeth T Cirulli and David B Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010.

[6] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, page 407, 2011.

[7] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D Walker, and Arup K Chakraborty. Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[8] Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332 (6034):1193–1196, 2011.

[9] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034):1190–1192, 2011.

[10] Jesse D Bloom, Jonathan J Silberg, Claus O Wilke, D Allan Drummond, Christoph Adami, and Frances H Arnold. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):606–611, 2005.

[11] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Em-

manuelle Petit, Julie Poulain, Guilène Barnaud, et al. Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences*, 110(32):13067–13072, 2013.

[12] Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14(8):559–571, 2013.

[13] David D Pollock, Grant Thiltgen, and Richard A Goldstein. Amino acid coevolution induces an evolutionary stokes shift. *Proceedings of the National Academy of Sciences*, 109(21): E1352–E1359, 2012.

[14] Michael S Breen, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor A Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, 2012.

[15] J Arjan GM de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, 2014.

[16] Anna I. Podgornaia and Michael T. Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, 2015. doi:10.1126/science.1257360. URL http://www.sciencemag.org/content/347/6222/673.abstract.

[17] Martijn F Schenk, Ivan G Szendro, Merijn LM Salverda, Joachim Krug, and J Arjan GM de Visser. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Molecular Biology and Evolution*, 30(8):1779–1787, 2013.

[18] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422): 138–142, 2012.

[19] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, and Timothy Palzkill. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *Journal of Molecular Biology*, 424(3):150–167, 2012.

[20] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *RNA*, 19(11):1537–1551, 2013.

[21] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of molecular biology*, 425(8):1363–1377, 2013.

[22] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution*, 31(6):1581–1592, 2014.

[23] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, page 201422285, 2015.

[24] Jakub Otwinowski and Joshua B. Plotkin. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences*, 111(22): E2301–E2309, 2014.

[25] Jianlin Cheng, Arlo Randall, and Pierre Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1125–1132, 2006.

[26] Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, 33(suppl 2):

W306–W310, 2005.

[27] Mathieu Lonquety, Zoé Lacroix, Nikolaos Papandreou, and Jacques Chomilier. Sprouts: a database for the evaluation of protein stability upon point mutation. *Nucleic Acids Research*, 37(suppl 1):D374–D379, 2009.

[28] Yves Dehouck, Jean M Kwasigroch, Dimitri Gilis, and Marianne Rooman. Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12(1):151, 2011.

[29] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7:61–80, 2006.

[30] John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23 (15):1875–1882, 2007.

[31] C Scott Wylie and Eugene I Shakhnovich. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences*, 108(24):9916–9921, 2011.

[32] Adrian WR Serohijos and Eugene I Shakhnovich. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Current Opinion in Structural Biology*, 26:84–91, 2014.

[33] Jesse D Bloom and Matthew J Glassman. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Computational Biology*, 5:e1000349, 2009.

[34] Julian Echave, Eleisha L Jackson, and Claus O Wilke. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Physical biology*, 12(2): 025002, 2015.

[35] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4): 248–249, 2010.

[36] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

[37] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[38] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[39] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[40] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, 2012.

[41] Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, and José N Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109 (26):10340–10345, 2012.

[42] Thomas A Hopf, Lucy J Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.

[43] Timothy Nugent and David T Jones. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences*, 109(24):E1540–E1547, 2012.

[44] Alexander Schug, Martin Weigt, José N Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106 (52):22124–22129, 2009.

[45] Angel E Dago, Alexander Schug, Andrea Procaccini, James A Hoch, Martin Weigt, and Hendrik Szurmant. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences*, 109(26):E1733–E1742, 2012.

[46] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, 3, 2014.

[47] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *eLife*, 3, 2014.

[48] Andrea Procaccini, Bryan Lunt, Hendrik Szurmant, Terence Hwa, and Martin Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PLoS One*, 6(5):e19729, 2011.

[49] Ryan R Cheng, Faruck Morcos, Herbert Levine, and José N Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, 111 (5):E563–E571, 2014.

[50] Helen C Davison, Mark EJ Woolhouse, and J Chris Low. What is antibiotic resistance and how can we measure it? *Trends in Microbiology*, 8(12):554–559, 2000.

[51] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[52] Carlo Baldassi, Marco Zamparo, Christoph Feinauer, Andrea Procaccini, Riccardo Zecchina, Martin Weigt, and Andrea Pagnani. Fast and accurate multivariate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PLoS One*, 9(3):e92721, 2014.

[53] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.

[54] George Minasov, Xiaojun Wang, and Brian K Shoichet. An ultrahigh resolution structure of tem-1 β-lactamase suggests a role for glu166 as the general base in acylation. *Journal of the American Chemical Society*, 124(19):5333–5340, 2002.

[55] Jay W Ponder and Frederic M Richards. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193(4):775–791, 1987.

[56] Carlos D Bustamante, Jeffrey P Townsend, and Daniel L Hartl. Solvent accessibility and purifying selection within proteins of escherichia coli and salmonella enterica. *Molecular Biology and Evolution*, 17(2):301–308, 2000.

[57] Eric A Franzosa and Yu Xia. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular biology and evolution*, 26(10):2387–2395, 2009.

[58] Luciano A Abriata, Timothy Palzkill, and Matteo Dal Peraro. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PloS one*, 10(2), 2015.

[59] Michel F Sanner, Arthur J Olson, and Jean-Claude Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[60] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE*, 8(11):e80635, 11 2013. doi: 10.1371/journal.pone.0080635. URL http://dx.doi.org/10.1371%2Fjournal.pone.0080635.

[61] Insa Kather, Roman P Jakob, Holger Dobbek, and Franz X Schmid. Increased folding stability of tem-1 beta-lactamase by in vitro selection. *Journal of molecular biology*, 383(1):238–251, 2008.

[62] X Raquet, M Vanhove, J Lamotte-Brasseur, S Goussard, P Courvalin, and J-M Frère. Stability of tem β-lactamase mutants hydrolyzing third generation cephalosporins. *Proteins: Structure, Function, and Bioinformatics*, 23(1):63–72, 1995.

[63] Xiaojun Wang, George Minasov, and Brian K Shoichet. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of Molecular Biology*, 320(1): 85–95, 2002.

[64] RP Ambler, AF Coulson, Jean-Marie Frère, Jean-Marie Ghuysen, Bernard Joris, M Forsman, RC Levesque, G Tiraby, and SG Waley. A standard numbering scheme for the class a beta-lactamases. *Biochemical Journal*, 276(Pt 1):269, 1991.

[65] Merijn LM Salverda, JAGM De Visser, and Miriam Barlow. Natural evolution of tem-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiology Reviews*, 34(6):1015–1036, 2010.

[66] Orr Ashenberg, L Ian Gong, and Jesse D Bloom. Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences*, 110 (52):21071–21076, 2013.

[67] Zhengting Zou and Jianzhi Zhang. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Molecular Biology and Evolution*, page msv091, 2015.

[68] Sarah M Drawz, Christopher R Bethel, Kristine M Hujer, Kelly N Hurless, Anne M Distler, Emilia Caselli, Fabio Prati, and Robert A Bonomo. The role of a second-shell residue in modifying substrate and inhibitor interactions in the shv β-lactamase: a study of ambler position asn276. *Biochemistry*, 48 (21):4557–4566, 2009.

[69] Karthik Shekhar, Claire F Ruberman, Andrew L Ferguson, John P Barton, Mehran Kardar, and Arup K Chakraborty. Spin models inferred from patient-derived viral sequence data faithfully describe hiv fitness landscapes. *Physical Review E*, 88(6): 062705, 2013.

[70] Jaclyn K. Mann, John P. Barton, Andrew L. Ferguson, Saleha Omarjee, Bruce D. Walker, Arup Chakraborty, and Thumbi Ndung'u. The fitness landscape of hiv-1 gag: Advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput Biol*, 10(8):e1003776, 08 2014.

[71] Sara Lui and Guido Tiana. The network of stabilizing contacts in proteins studied by coevolutionary data. *The Journal of chemical physics*, 139(15):155103, 2013.

[72] Faruck Morcos, Nicholas P Schafer, Ryan R Cheng, José N Onuchic, and Peter G Wolynes. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences*,

111(34):12408–12413, 2014.

[73] Robert D Finn, Alex Bateman, Jody Clements, Penelope Cog-gill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the pro-tein families database. *Nucleic Acids Research*, page gkt1223, 2013.

[74] Jaina Mistry, Robert D Finn, Sean R Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic Acids Re-search*, 41(12):e121–e121, 2013.

[75] JP Barton, S Cocco, E De Leonardis, and R Monasson. Large pseudocounts and l 2-norm penalties are necessary for the mean-field inference of ising and potts models. *Physical Re-*

view E, 90(1):012132, 2014.

[76] Toshiyuki Tanaka. Mean-field theory of boltzmann machine learning. *Physical Review E*, 58(2):2302, 1998.

[77] Timo Strunk, Moritz Wolf, Martin Brieg, K Klenin, A Biewer, Frank Tristram, M Ernst, PJ Kleine, N Heilmann, Ivan Kon-dov, et al. Simona 1.0: an efficient and versatile framework for stochastic simulations of molecular and nanoscale systems. *Journal of Computational Chemistry*, 33(32):2602–2613, 2012.

[78] A Schug, T Herges, and W Wenzel. Reproducible protein fold-ing with the stochastic tunneling method. *Physical Review Let-ters*, 91(15):158102, 2003.

[79] A Verma, A Schug, KH Lee, and W Wenzel. Basin hopping simulations for all-atom protein folding. *The Journal of Chem-ical Physics*, 124(4):044515, 2006.