

Predicting the Melting Point of Human C-Type Lysozyme Mutants

Deeptak Verma¹, Donald J. Jacobs^{2,*} and Dennis R. Livesay^{1,*}

¹Department of Bioinformatics and Genomics, ²Department of Physics and Optical Science, University of North Carolina at Charlotte, Charlotte, NC 28262 USA

Abstract: A complete understanding of the relationships between protein structure and stability remains an open problem. Much of our insight comes from laborious experimental analyses that perturb structure via directed mutation. The glycolytic enzyme lysozyme is among the most well characterized proteins under this paradigm, due to its abundance and ease of manipulation. To speed up such analyses, efficient computational models that can accurately predict mutation effects are needed. We employ a minimal Distance Constraint Model (mDCM) to predict the stability of a series of lysozyme mutants (specifically, human wild-type C-type lysozyme and 14 point mutations). With three phenomenological parameters that characterize microscopic interactions, the mDCM parameters are determined by obtaining the least squares error between predicted and experimental heat capacity curves. The mutants are chemically and structurally diverse, but have been experimentally characterized under nearly identical thermodynamic conditions (pH, ionic strength, etc.). The parameters found from best fits to heat capacity curves for one or more lysozyme structures are subsequently used to predict the heat capacity on the remaining. We simulate a typical experimental situation, where prediction of relative stabilities in an untested mutated structure is based on known results as they accumulate. From the statistical significance of these simulations, we establish that the mDCM is a viable predictor for relative stability of protein mutants. Remarkably, using parameters from any single fitting yields an average percent error of 4.3%. Across the dataset, the mDCM reproduces experimental trends sufficiently well ($R = 0.64$) to be of practical value to experimentalists when making decisions about which mutations to invest time and funds for characterization.

Keywords: Lysozyme, mutation, melting point, protein stability, heat capacity, distance constraint model.

INTRODUCTION

Due to the time and cost of molecular biology and biophysical experiments, accurate computational models to predict and explain the effects of point mutations on protein stability have long been desired. Despite some progress towards this goal, it remains largely an open computational biology problem. The majority of the successes thus far have been based on machine learning approaches (i.e., decision trees [1-3], support vector machines [4-6], and artificial neural networks [7-9]). While these methods can achieve impressive prediction accuracies, their interpretive utility spans a broad range (e.g., decision trees are somewhat interpretable, whereas artificial neural networks are not). And even under the best of circumstances, all of these empirical methods lack the descriptive power of first-principles calculations based on the underlying physics and chemistry. As such, we have been seeking to develop a biophysical calculation of protein stability.

Unfortunately, computational biophysics approaches have largely failed to achieve an expectable level of accuracy. The reason for this is that protein structures are extremely complicated, being dense networks of chemical interactions that lead to protein stabilities involving small differences between large free energy values. In fact, even

misplacement of a single hydrogen atom is sufficient to render a computational model wrong [10]. The primary exception to these failures is stability predictions of solvent exposed mutations. These mutations increase (or decrease) protein stability by optimizing (or destabilizing) long-range surface electrostatics. Because these mutations occur on the protein surface, which is less densely packed than the protein core, they are less susceptible to issues related to the intimate atomic details that frequently derail computational predictions within the core. As such, biophysical models that quantify surface electrostatics effects (i.e., Tanford-Kirkwood [11-13] and Poisson-Boltzmann theories [14-17]) have successfully reproduced experimental trends across several mutation sets and have provided explanations for a large number of confounding experimental results. When these methods fail, the origins of the mutant stability change are often generically explained as resulting from conformational changes, highlighting a fundamental limitation of nearly all biophysics-based stability prediction methods.

To address the problem of predicting mutant stability based on conformational considerations, we test herein the ability of our distance constraint model (DCM) to reproduce stability trends within human C-type lysozyme and 14 point mutations therein. The DCM is a phenomenological biophysics model that requires parameterization, usually done by fitting to experimental heat capacity, C_p , curves [18, 19]. After parameterization, the DCM quantifies both the enthalpic and entropic effects of all interactions within the protein, from which a wide variety of equilibrium thermodynamic quantities (i.e., C_p , free energy, T_m , etc.) are calculated. The

*Address correspondence to these authors at the University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223; USA; Tel: 704-687-8143; Fax: 704-687-8197; E-mail: djacobs1@uncc.edu; and Tel: 704-687-7995; Fax: 704-687-8667; E-mail: drlivesa@uncc.edu.

DCM approach is applied to 15 lysozyme structures each with measured heat capacity. The best-fit parameters derived from one such determination are applied to the remaining structures to assess predictive power. Over all possible permutations, this process results in an impressive average error of 4.3% (standard deviation = 3.6%) in the prediction of the experimental T_m 's, a commonly used surrogate for stability [20]. This translates to a Pearson correlation coefficient of 0.64 for predicted to experimental ΔT_m values, which is among the best values ever presented for prediction of point mutation stability focusing on conformational effects. In this approach, multiple parameter sets are found that fit the data well in addition to the above mentioned best-fit parameters associated with the lowest least squares error. In the attempt to boost statistical prediction accuracy by incorporating variability in model parameterization from additional near-optimal parameter sets and using additional parameter sets from multiple mutant structures, we find that, surprisingly, there is no statistically significant change in the predictive accuracy on average. This result is important, as it indicates that the predictions from any DCM parameterization resulting in a reasonably good fit to heat capacity is robust to transferability across mutant structures.

METHODS

Overview of the Distance Constraint Model

A complete description of protein stability should ideally account for a wide variety of chemical interactions, including: covalent bonding (i.e., bond stretching, angle bending and torsional effects), nonbonded interactions (i.e., long and short range ionic interactions, hydrogen bonds, dipole interactions and van der Waals contacts), solvation, etc. Most of which are affected by solvent pH, ionic strength and other co-solute concentrations. Standard simulation methods, such as molecular dynamics, attempt to describe most of the above terms, but rarely all. The “rules” constraining the simulation are based on energetic potentials; however, free energies are the primary protein stability metric of interest. Free energies are derived from a simulation *post priori* using the method of thermodynamic integration [21]. The primary advantage of simulation methods is that they are nearly chemically and physically complete, but simulation is extremely computationally intensive, making it prohibitive for large-scale analyses.

In response to the immense computational cost of molecular simulations, the DCM has been developed from conception to optimally balance computational efficiency with prediction accuracy by invoking both mechanical and thermodynamic viewpoints of macromolecular structure. The DCM is based on a free energy functional that decomposes the total free energy into constituent parts related to specific types of interactions. While total enthalpies can be calculated from the sum of the individual components, adding entropies over all components generally will overestimate conformational entropy [22, 23]. However, the utility of a free energy decomposition is restored using a DCM [24], where conformational entropy is additive over independent degrees of freedom (DOF). Herein, structure is recast as a topological network of distance constraints. Each constraint within the

topological framework (network) is associated with a component enthalpy and entropy value. The conformational part of the free energy of a given framework, $G(f)$, is reconstituted from the free energy decomposition that defines the types of interactions modeled as distance constraints. The free energy is calculated from the total enthalpy and entropy of a framework by:

$$G_{conf}(f) = \sum_t^{N_{int}} h_t N_t(f) - RT \sum_t^{N_{int}} \sigma_t I_t(f)$$

where N_{int} is the number of different types of modeled interactions, h_t is the enthalpy of interaction t , N_t is the number of times interaction type t occurs within framework f , σ_t is the pure entropy of a single distance constraint used to model interaction type t , R is the ideal gas constant, and I_t is the number of number of independent constraints of type t . To provide ease of interpretation and consistency, enthalpy parameters are depicted using Roman characters, whereas entropies are depicted by Greek symbols.

The salient feature within the conformational free energy calculation of a framework is that total entropy is summed over a set of independent constraints, which are determined using efficient network rigidity graph algorithms [25, 26]. Starting from $3N_a$ DOF (N_a is the number of atoms), each constraint within flexible portions of the network removes one degree of freedom. However, when an interaction is added to an already rigid substructure of the network, no further reduction in entropy occurs because all available DOF within that region have already been consumed. Since assignment of which constraint is independent or redundant is not unique, the expression for conformational entropy is also not unique. This approach provides an upper bound estimate to conformational entropy regardless which set of independent constraints are considered. However, by adding constraints as dictated by an entropy spectrum [27] that preferentially orders them from smallest to largest entropy, a rigorous lowest upper bound is obtained. Note that a given chemical interaction can be modeled by more than one constraint. For example, covalent bonds and H-bonds are modeled as five constraints, and a torsion force is modeled as one [18].

If thermal fluctuations did not occur, the free energy of a given protein would simply be based on the above calculation using the native state structure, but this is of course not the case. While covalent bonding is appropriately described by a large set of quenched constraints that are present within each microstate of the ensemble, fluctuating constraints account for the forming and breaking of weak interactions that are critical to properly describe equilibrium behavior. Herein, we consider a ‘minimal’ set of fluctuating interaction types, specifically $N_{int} = 2$ types are considered: hydrogen bonds and torsion angle forces. Within this minimal DCM (mDCM), all possible hydrogen bonds (H-bonds) are defined by the native structure. H-bond enthalpies, h_{hb}^{pot} , are calculated from the native structure using the empirical potential from Dahiyat *et al.* [28]. Salt bridges are modeled as a special case of H-bonds. The entropic cost of forming an intramolecular H-bond is linearly related to h_{hb}^{pot} , whose slope

is defined by the parameter γ_{max} . Solvation terms are described through H-bonds to solvent; when an intramolecular H-bond breaks, there is a compensating reduction enthalpy given by the fitting parameter u_{sol} . As a consequence, the net effect of each intramolecular H-bond is given by $h_{hb}^{net} = h_{hb}^{pot} - u_{sol}$. While it is less immediately obvious, the reduction in entropy associated with torsional effects can also be modeled using distance constraints. Here, we introduce constraints across all i to $i+3$ atomic pairs, which includes side chain torsions. The torsions are segregated in an Ising-like manner where *native* torsions are associated with enthalpy and entropy values $\{v_{nat}, \delta_{nat}\}$. and *disordered* torsions are associated with $\{v_{dis}, \delta_{dis}\}$. Two important “minimal” aspects of the mDCM are that: (i.) other than h_{hb}^{pot} , all parameter values are treated phenomenologically; and (ii.) all parameters are treated universally regardless of residue type. The disordered dihedral angle enthalpy, v_{dis} , is our reference energy, which is defined as zero.

Hydrophobic considerations are the most severe omission from our current free energy decomposition scheme. The hydrophobic effect is a bulk colligative property related to an increase in the number of accessible DOF upon segregation of polar and nonpolar solvents. As such, they do not directly map to a set of distance constraints. For example, molecular dynamics does not directly model the hydrophobic effect because it is not explicitly included within the simulation “rules” defined by molecular mechanical force fields. Therein, the hydrophobic effect only emerges after thermodynamic integration of the trajectory phase space. Within the mDCM, hydrophobic interactions are indirectly included by two phenomenological terms that connect to order parameters describing the number of constraints within the system. This approach works well, and is tied to the observation that hydrophobic contacts track H-bond formation [29], meaning that our phenomenological H-bond parameters implicitly account for hydrophobic interactions. Thus, as we have discussed previously [30], the u_{sol} and v_{nat} parameters implicitly account for the hydrophobic effect.

Even with such a simple model, an exact calculation of the partition function for lysozyme is impossible due to an astronomical number ($\sim 2^{750}$) of possible frameworks. As such, a heterogeneous mean field approach has been developed to make the calculation tractable [18, 19]. Combining all the contributions described above, we arrive at the following free energy functional:

$$G(N_{hb}, N_{nat}) = U_{hb}(N_{hb}) - N_{hb}u_{sol} + N_{nat}v_{nat} - RTS_{conf}(N_{hb}, N_{nat} | \delta_{nat}, \delta_{dis}, \gamma_{max}) - RTS_{mix}(N_{hb}, N_{nat})$$

This functional has five adjustable parameters (depending on solvent conditions and protein fold). However, from our previous work, γ_{max} and δ_{dis} have been fixed and are treated as transferable parameters, leaving only $\{\delta_{nat}, v_{nat}, u_{sol}\}$ as free parameters within the mDCM (cf. Table 1). We have found that the three-free parameter mDCM provides a high degree of accuracy and robustness in predicting protein stability [19]. Typically the parameterization has been obtained by finding the appropriate parameter values to reproduce experimental C_p curves from differential scanning calorimetry (DSC) using simulated annealing. These mDCM parameters are physically meaningful with ranges that are

remarkably tight over a diverse set of proteins. This modeling approach has been found to provide accurate description of both thermodynamics and intrinsic flexibility in proteins in many applications [18, 19, 30-33].

Dataset Preparation and Simulated Annealing

Lysozyme, which is abundant in egg whites and secretions (i.e., tears, saliva, milk, etc.), is a general class of enzymes that degrade bacterial cell walls through hydrolysis of $\beta(1,4)$ glycosidic linkages. Members of the lysozyme superfamily share the same $\alpha+\beta$ structural motif within their active site region. Due to ease of production and characterization, human C-type lysozyme is a common model system for protein stability investigations. C_p curves for the human C-type lysozyme and 14 lysozyme mutant proteins have been obtained from published data [34-40]. Exact PDB codes are provided in Table 2. Since model parameters are dependent upon solvent conditions, only site directed mutants that are spatially distinct and characterized under the same experimental conditions are considered here. Specifically, all 15 proteins were characterized under near identical buffering conditions (pH = 2.7-2.8) and salt concentrations (0.05 M). The pH range is $2.67 \leq \text{pH} \leq 2.8$ with an average of 2.71, and with a 0.032 standard deviation. In addition, all of the considered C_p curves have been produced by the same group (Yutani *et al.*), which minimizes the risk that unforeseen factors (instrument, sample preparation, protein concentration estimates, etc.) are affecting the controls in experimental data. Note that, in practice, DSC is a notoriously difficult and finicky technique to perform [41], and in general it is difficult to find a large collection of systematic data. In order to look at the intrinsic variability in the DCM, and to justify our demand on the transferability of parameterization, it is important to have all the heat capacity measurements made under identical conditions. Fortunately, the methodical work by Yutani and co-workers presented us with the opportunity to consider a diverse collection of 14 point mutations that are structurally well distributed throughout the lysozyme structure, as shown in Fig. (1) which also indicates their solvent accessibility. Approximately half of the mutations are exposed to solvent, which as discussed above can be well described by long-range electrostatics models. However, we purposely omit an explicit long-range electrostatics component in the presented model to assess how well the mDCM does on its own. Incorporation of long-range electrostatics is expected to further improve model accuracy. To ensure proper ionization, the H++ server [42] is used to add hydrogen atoms to the structures as expected at pH 2.7 based on calculated pK_a values. The protonated structures are subsequently energy minimized prior to the simulated annealing fitting to determine $\{\delta_{nat}, v_{nat}, u_{sol}\}$ as done previously [18, 19]. Prior to fitting, all curves were shifted such that $C_{p,min} = 0$.

Prediction of T_m Values

After best-fit parameters have been determined for each mutant structure and C_p curve pair, we attempt to answer the following question: *How well does the mDCM reproduce T_m values on the remaining 14 structures using the best-fit parameters from the first?* After considering all possible per-

Table 1. Free Fitting Parameters Used by the mDCM

Interaction	Parameter	Treatment	Description
H-bonds	h_{hb}^{pot}	Empirical potential	Intramolecular H-bond enthalpy
	γ_{max}	Constant	H-bond pure entropy is linearly related to h_{hb}^{pot} whose slope is controlled by γ_{max}
	u_{sol}	Fitting	H-bond to solvent enthalpy ¹
Native torsion	δ_{nat}	Fitting	Native torsion angle pure entropy
	v_{nat}	Fitting	Native torsion angle enthalpy
Disordered torsion	δ_{dis}	Constant	Disordered torsion angle pure entropy
	v_{dis}	Constant	Disordered torsion angle enthalpy

¹ The net intramolecular H-bond enthalpy is calculated as $h_{hb}^{net} = h_{hb}^{pot} - u_{sol}$, where h_{hb}^{pot} is calculated from an empirical potential.

Table 2. Thermodynamic Characteristics and Best-Fit Parameters

Mutation	PDBID	T_m (K)	$C_{p,max}$ (kcal/[mol·K])	δ_{nat} (unitless)	v_{nat} (kcal/mol)	u_{sol} (kcal/mol)
Wild-type	1LZ1	338.7	17.5	1.16	-0.16	-1.84
K1A	1C45	336.7	13.1	1.36	-0.20	-1.71
V2A	1OUG	333.4	16.8	1.12	-0.28	-1.76
Y38F	1WQO	337.7	18.8	1.08	-0.23	-1.69
Y45F	1WQP	337.4	18.5	1.32	-0.26	-1.77
Y54F	1WQQ	337.3	17.3	1.36	-0.29	-1.90
I56T	1OUA	325.1	14.8	1.32	-0.29	-1.87
Q58G	1B7R	345.3	19.0	1.24	-0.29	-1.89
I59S	2MEG	326.2	14.4	1.32	-0.37	-1.94
Y63F	1WQR	337.6	18.5	1.24	-0.25	-1.87
P71G	1LHI	336.1	20.3	1.28	-0.32	-2.11
V74A	1OUH	337.3	18.8	1.28	-0.23	-1.77
V100A	1OUB	337.1	18.2	1.28	-0.35	-1.91
P103G	1LHJ	338.6	18.2	1.32	-0.19	-1.77
Y124F	1WQM	337.6	19.0	0.88	-0.39	-1.78
Average		336.1	17.5	1.24	-0.27	-1.84
Variation		1.5%	11.3%	10.5%	24.3%	5.8%

mutations, this thought experiment assesses how well, on average, the mDCM would do if only a single mutant structure and C_p curve pair were available before prediction on additional lysozymes. Meaning, we apply the best-fit parameters from mutant i to all of the remaining lysozyme structures. In each case, the predicted C_p curve is calculated, and its peak is used to identify the predicted T_m . We then collapse all $15 \times 14 = 210$ permutations into a single dataset and report average statistics.

As an extension, we also consider two scenarios where additional parameter sets are used in the prediction process. For example, we assess whether or not using the parameters from $n > 1$ ‘training’ lysozyme structures improves average prediction accuracy. Here, best-fit parameter sets from n different lysozymes are used to generate n different T_m predictions for a given target, which are simply averaged to give the final predicted value. However, complete enumeration of all possible permutations results in a combinatorial explosion

(>10⁸). Instead, for each target lysozyme, we have generated 100 random 14-character strings that simply list a unique identifier associated with each of the remaining lysozyme structures. For each value of $n \in \{1-14\}$, we include the first n lysozymes from the generated string within the ‘training’ set. The same 100 strings are used as we systematically consider all possible values of n . Over all 15 lysozyme structures, we determine the final predicted T_m value for each n by averaging over the 100 samples. This entire process was repeated ten times.

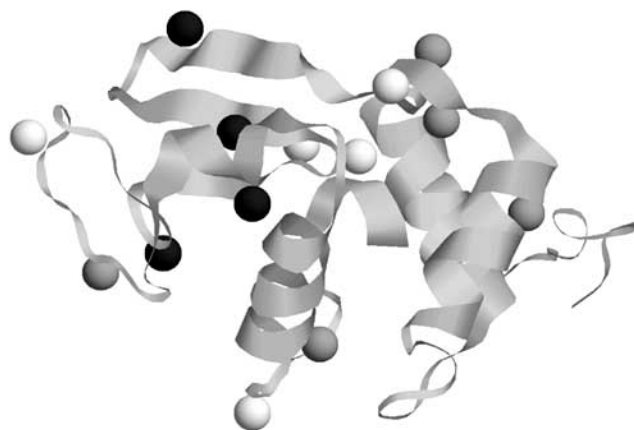


Fig. (1). The wild-type human C-type lysozyme structure. The CB atoms of the 14 mutated positions are highlighted and color-coded by highly (white), medium (light grey) and least (dark grey) solvent accessible.

While only best-fits have been discussed thus far, the simulated annealing procedure actually generates a large number of nearly optimal fits that are virtually indistinguishable by visual inspection. Fig. (2) plots kernel density functions generated using the R statistical package for the $m = 20$ best fit parameter sets for the wild-type structure. Across the parameters $\{\delta_{nat}, v_{nat}, u_{sol}\}$, the variation ranges from 2% to 24.2%. Similar results are observed for the lysozyme mutants. Related, Table 3 presents the percent variation in each parameter at three values of m ($m = 4, 8,$ and 16). Note that there is slight tendency for the variation to increase with m ; however, there are several examples where the opposite occurs, highlighting the stochastic nature of finding a good

parameter basin within the simulated annealing process. We assess whether or not using $m > 1$ of these near optimal fits improves average prediction accuracy. Meaning, for a given ‘training’ lysozyme mutant structure, we apply the m best parameters sets derived from it to each of the remaining structures, resulting in m different T_m predictions for each mutant trained on. As before, the final predicted T_m is simply the average over the m predictions for that structure. We consider each value of $m \in \{1-20\}$. Putting everything together, we consider m different parameter sets for each of the n proteins used to train on. From these two defined order parameters, we collapse the average statistics onto an $n \times m$ grid to assess if increasing the number of experimental mutants to fit to and/or the number of near-optimal parameter sets improves prediction accuracy.

RESULTS

Best-fit Parameters and Mutant Stability

Best-fits for each of the C_p curves are provided in Fig. (3) which highlights the quality of the model fits. Note that a simple solvent exposure model is actually sufficient to describe ΔC_p between low and high temperature [43]; however, such an approach cannot generate a peak within C_p , indicating that equilibrium fluctuations are not properly modeled. Within the mDCM, since we are not yet explicitly modeling solvation terms, the C_p baseline is subtracted. We analytically describe the baseline using a $\tanh(T)$ function. The rise of the $\tanh(T)$ function leads to a frequent slight overestimation of the C_p at high temperature, but this is mostly insignificant as the C_p peak is the primary region of interest. To the best of our knowledge, the mDCM remains the only free energy decomposition scheme capable of quantitatively reproducing experimental C_p peaks. The associated best-fit parameter values are provided in Table 2. All parameter values are consistent with ranges established in our prior works studying globular proteins [18, 19, 30-32], and are physically meaningful. The conservation within u_{sol} is especially noteworthy, which is due to the enforced criterion that all experimental DSC solvent conditions be the same.

Not surprisingly, the vast majority of all mutations destabilize (relative to wild-type) lysozyme. In fact, only the

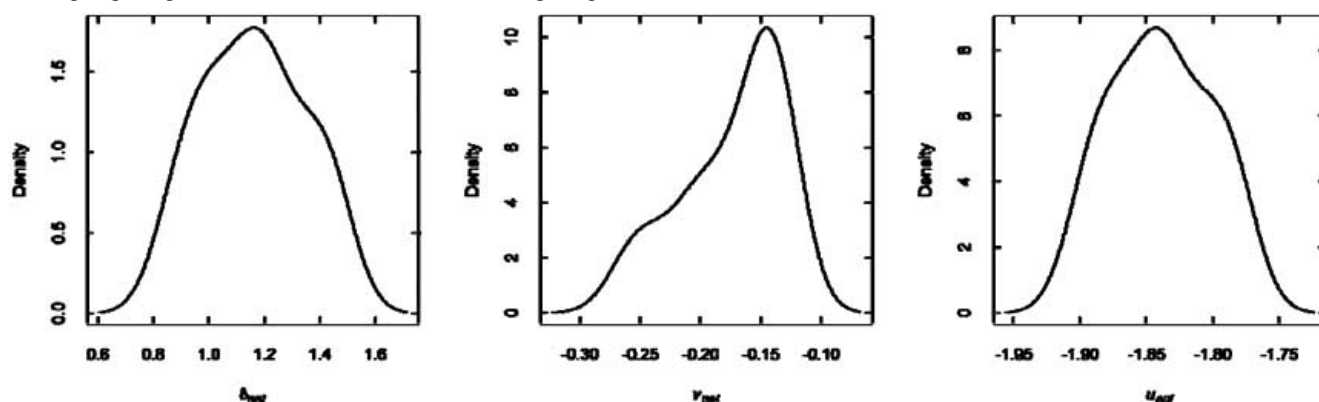


Fig. (2). Kernel density functions for each of the three model parameters generated from the $m = 20$ best wild-type lysozyme parameter sets. The percent variation for each parameter $\{\delta_{nat}, v_{nat}, u_{sol}\}$ is 15.9, 2.0 and 24.2 percent, respectively. The density functions and percent variation values for the lysozyme mutants are similar (cf, Table 3). The kernel density plots were generated using the R statistical package.

Table 3. Percent Variation within the *m* Best Parameter Sets

Mutation	PDBID	u_{sol} (kcal/mol)			v_{nat} (kcal/mol)			δ_{nat} (unitless)		
		<i>m</i> =4	<i>m</i> =8	<i>m</i> =16	<i>m</i> =4	<i>m</i> =8	<i>m</i> =16	<i>m</i> =4	<i>m</i> =8	<i>m</i> =16
Wild-type	1LZ1	1.9	1.3	1.8	20.9	17.3	23.2	12.7	10.4	14.1
Q58G	1B7R	0.5	0.6	1.4	5.3	5.5	9.9	5.6	6.3	10.4
K1A	1C45	2.1	1.5	4.1	11.3	9.2	12.0	1.7	1.5	10.9
P71G	1LHI	1.3	1.4	1.9	3.8	3.9	6.0	6.8	6.6	7.9
P103G	1LHJ	0.7	1.0	1.3	5.3	6.0	11.6	0.00	2.0	4.9
I56T	1OUA	2.1	2.8	5.0	9.7	8.4	10.4	12.9	13.8	15.3
V100A	1OUB	0.9	2.9	3.1	2.7	6.5	8.4	6.2	12.5	15.7
V2A	1OUG	0.3	2.2	2.7	2.8	3.7	9.2	2.1	11.9	13.7
V74A	1OUH	0.7	1.5	1.6	2.2	4.7	7.6	2.6	6.6	9.1
Y124F	1WQM	3.1	3.2	3.5	11.9	12.9	10.4	18.6	17.3	13.5
Y38F	1WQO	2.9	2.2	3.1	7.7	8.8	19.6	14.3	10.8	16.0
Y45F	1WQP	0.3	1.5	1.5	4.9	4.8	6.8	5.5	7.7	8.5
Y54F	1WQQ	1.6	1.3	1.3	4.6	3.8	8.5	3.1	5.7	11.5
Y63F	1WQR	0.8	0.8	1.2	4.0	4.8	5.2	1.6	2.4	3.7
I59S	2MEG	1.6	2.6	3.6	5.0	4.7	7.8	4.4	8.9	9.9

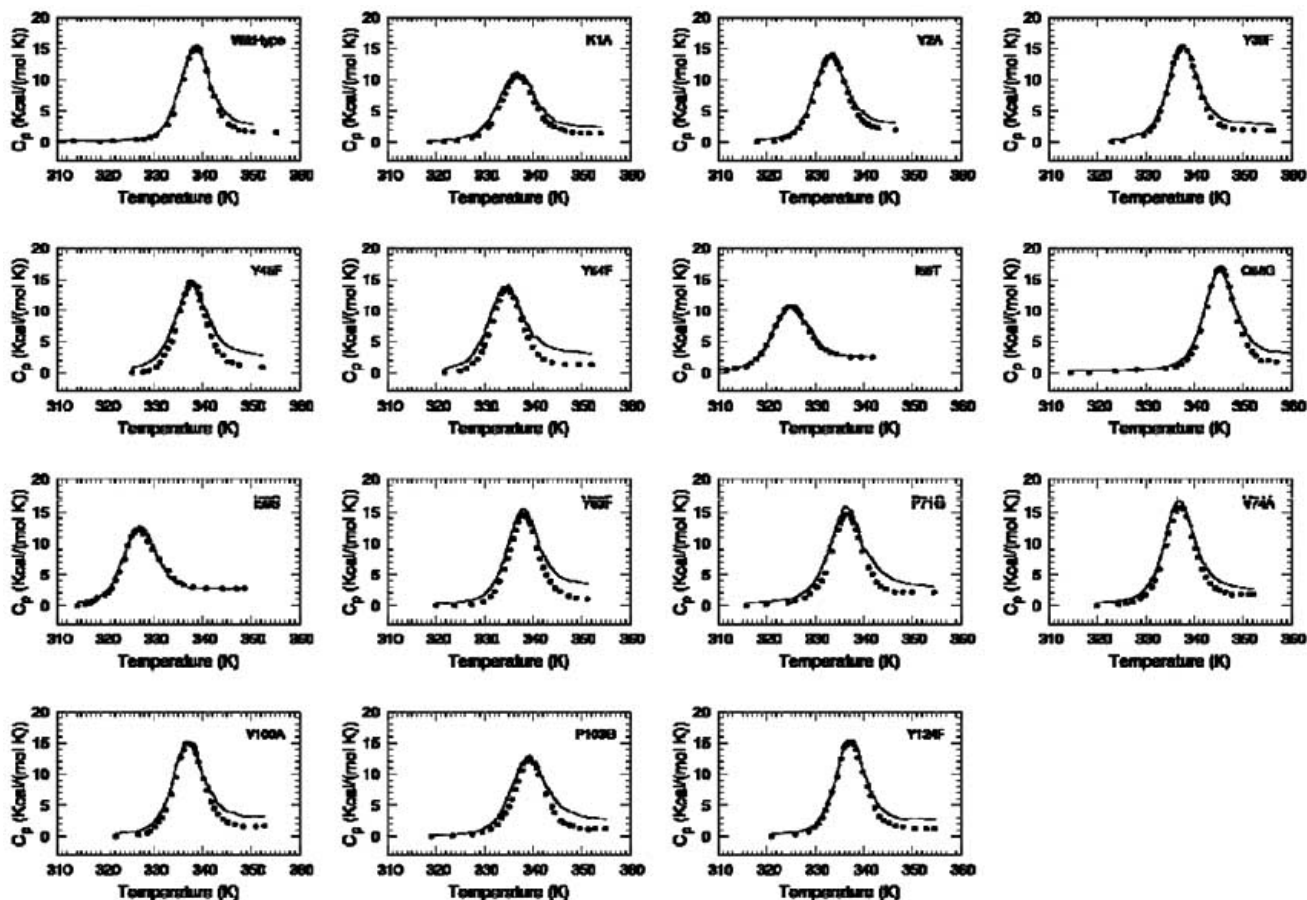


Fig. (3). The *m* = 1 absolute best-fits to the experimental heat capacity data for the human C-type lysozyme and the 14 different point mutations considered here. Experimental data points are shown by dots, whereas the mDCM predicted curves are shown in solid line. To facilitate comparisons, the coordinate ranges in all 15 examples are equal.

Q58G mutation has an increased T_m , which is increased by an astounding 6.6 K. As discussed in [40], the T_m of the P103G mutation is nearly identical to the wild-type. All of the other mutant T_m values range from 1 to ~14 K lower than that of the wild-type. While the wild-type structure has the lowest (most stabilizing) total H-bond enthalpy, there is actually only a weak correlation between total H-bond enthalpy and T_m ($R = -0.57$). In fact, the next most stabilizing total H-bond enthalpy is the P71G mutation, which has only the 11th largest T_m . Descriptions of the underlying H-bond networks are provided in Table 4. In all cases, the structures are of similar quality, as described by resolution and the observed R-value. Moreover, all of the mutant structures are from the same space group (P 2 2 2₁). Consequently, differences within the H-bond network can be reliably ascribed to conformational adjustments to relieve strain introduced by the mutation. As we have discussed previously [33], common descriptors (i.e., total H-bond enthalpy, average H-bond enthalpy, parameter values, etc.) are not good predictors of mDCM predictions across a set of closely related proteins. Descriptors based on global topological properties of the protein fold contain much less information than the mDCM, which is based on atomic level details affecting the network of distance constraints. Moreover, the way rigidity and flexibility propagate is non-trivial because network rigidity is a

long-range mechanical interaction that results in complex emergent behavior that cannot be captured solely from local or global network characteristics.

Average Prediction Accuracy Using a Single Parameter Set

To assess how well the mDCM describes the experimental T_m values, we apply the best-fit parameters from one of the above fits serving as a transferable set of parameters, to all remaining lysozyme structures. We repeat this same process for all 15 permutations. This is the simplest scenario presented in this report, corresponding to $m = 1$ and $n = 1$. Across all 210 T_m predictions, Fig. (4) highlights that more than 35% have errors less than 2%. The average error across all predictions is 4.3% (standard deviation = 3.6%). Clearly, the mDCM is doing a very good job at reproducing the experimental T_m values. Fig. (5) plots the average T_m for each structure using the other 14 parameter sets. In all but three cases, the experimental T_m is within the error range defined by \pm one standard deviation. Interestingly, the two starkest exceptions (wild-type and P71G) correspond to the two structures with the most stabilizing total H-bond enthalpy, suggesting that the mDCM free energy calculation might be slightly over-dependent upon very low total H-bond energies.

Table 4. Descriptions of the Wild-Type and Lysozyme Mutant Structures

Mutation	PDBID	Structure Resolution (Å)	Observed R-value	Total HB Enthalpy (kcal/mol)	Number of HBs	Average HB Enthalpy (kcal/mol)	T_m (K)
Wild-type	1LZ1	1.35	0.182	-613.9	240	-2.6	338.7
K1A	1C45	1.80	0.168	-567.8	245	-2.3	336.7
V2A	1OUG	1.80	0.173	-577.3	229	-2.5	333.4
Y38F	1WQO	1.80	0.170	-586.3	229	-2.6	337.7
Y45F	1WQP	1.80	0.174	-563.1	231	-2.4	337.4
Y54F	1WQQ	1.80	0.164	-565.9	229	-2.5	337.3
I56T	1OUA	1.80	0.148	-569.3	243	-2.3	325.1
Q58G	1B7R	1.80	0.160	-590.4	235	-2.5	345.3
I59S	2MEG	1.80	0.151	-554.3	239	-2.3	326.2
Y63F	1WQR	1.80	0.165	-589.1	239	-2.5	337.6
P71G	1LHI	1.80	0.156	-612.2	240	-2.6	336.1
V74A	1OUH	1.80	0.160	-586.1	235	-2.5	337.3
V100A	1OUB	1.80	0.160	-564.8	232	-2.4	337.1
P103G	1LHJ	1.80	0.152	-606.5	231	-2.6	338.6
Y124F	1WQM	1.80	0.164	-574.9	230	-2.5	337.6
Average		1.77	0.163	-581.5	235.1	-2.5	336.1
Variation		6.56%	5.740%	3.2%	2.3%	3.7%	1.5%

The Pearson correlation coefficient comparing the total H-bond enthalpy, number of H-bonds, and average H-bond enthalpy to the experimental T_m is, respectively, $R = -0.50$, -0.29 and -0.60 .

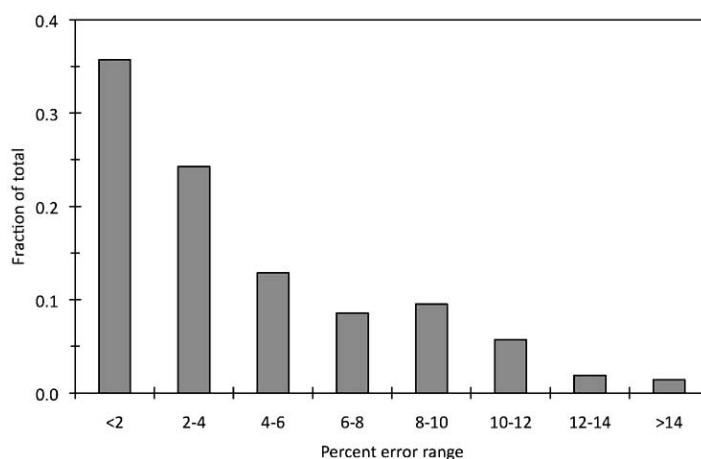


Fig. (4). Histogram plotting the accuracy of the mDCM T_m predictions when only a single parameter set is used.

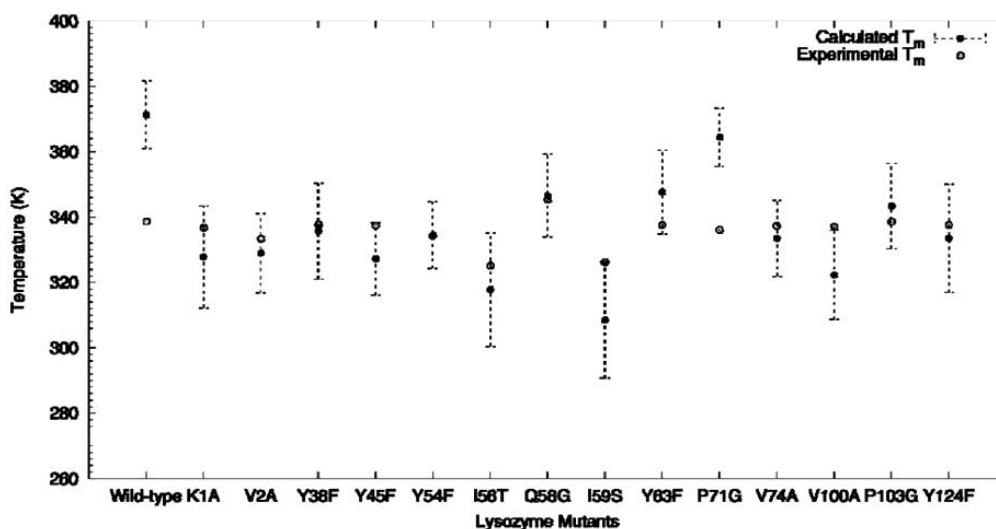


Fig. (5). The average T_m value for each structure using each of the other 14 parameter sets is plotted (error bars equal \pm one standard deviation). In all but three cases, the experimental T_m falls within the range defined by the error bars.

In practice, the primary goal of a computational model of protein stability is to assess relative stability of a mutant to the wild-type. To that end, Fig. (6A) provides a scatter plot of each predicted ΔT_m value ($T_{m,mut} - T_{m,wt}$) versus the experimental ΔT_m . The Pearson correlation coefficient is $R = 0.64$, which is among the best values ever reported for biophysical models focusing solely on conformationally derived properties. Because the mDCM over predicts the wild-type T_m so drastically, all predicted ΔT_m values are negative. However, Fig. (6B) plots the average ΔT_m ' values (defined as: $T_{m,mut}^{pred} - T_{ref}$, here $T_{ref} = T_{m,wt}^{exp}$) versus the experimental ΔT_m values, which demonstrates that once an appropriate reference point has been established, the mDCM does a very good job of predicting stabilizing mutations to be stabilizing (quadrant 1) and destabilizing mutations to be destabilizing (quadrant 3). Only three predictions are located in an incorrect quadrant. Note that this arbitrariness in defining a reference point is generally unnecessary to resurrect a satisfactory quadrant clustering using any of the other structures as a

reference point. The sole other exception is P71G, whose T_m is also over predicted by the mDCM.

Can Accuracy be Improved by Additional Parameterization?

Naively, it is expected that training on additional parameter sets from $n > 1$ lysozymes should improve average prediction accuracy. Similarly, it is expected that increasing parameter diversity by using m near optimal fits (up to some point before fit quality degrades) should also improve prediction accuracy. However, this is not the case here. Fig. (7) plots multiple cross-sections from the $n \times m$ landscape. Specifically, the entire series of n is plotted for five different values of m over four different lysozyme examples. In each case, the y-axis plots the $\langle \Delta T_m \rangle = \langle T_m^{pred} - T_m^{exp} \rangle$. Error bars indicate \pm one standard deviation. To make sure that our sampling procedure is not statistically biased, we have reported average values over ten different simulations of 100 samples each. The average behavior is shown to be largely consistent across all ten simulations, and in each case the

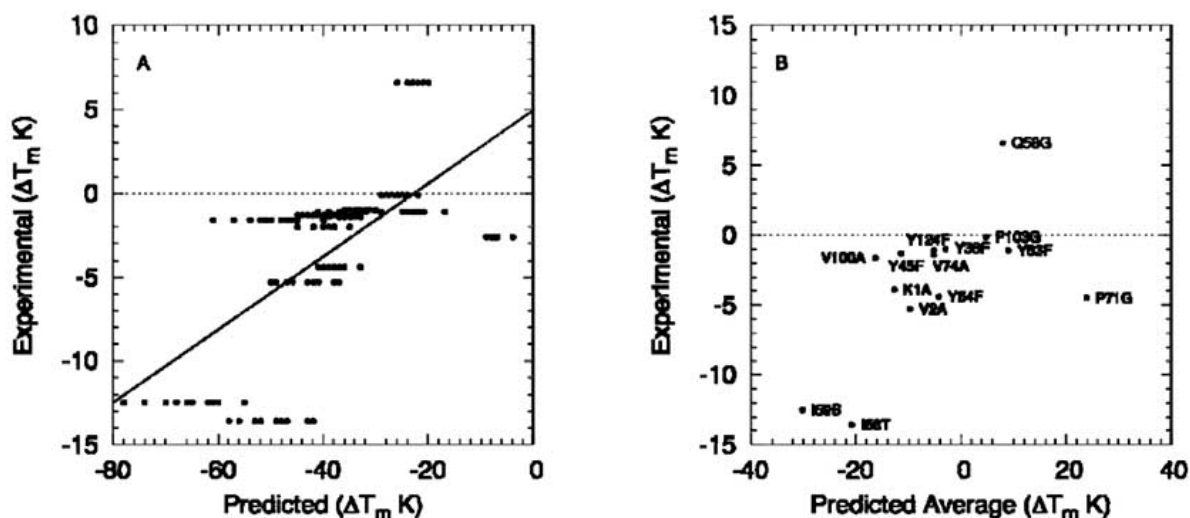


Fig. (6). (A) The ΔT_m values ($T_{m,mut} - T_{m,wt}$) for each of the $14 \times 13 = 182$ cases is plotted against the experimental equivalent. The Pearson correlation coefficient is $R = 0.64$. (B) Average ΔT_m values ($T_{m,mut(pred)} - T_{m,wt(exp)}$) versus the experimental ΔT_m values. The Pearson correlation coefficient is $R = 0.60$.

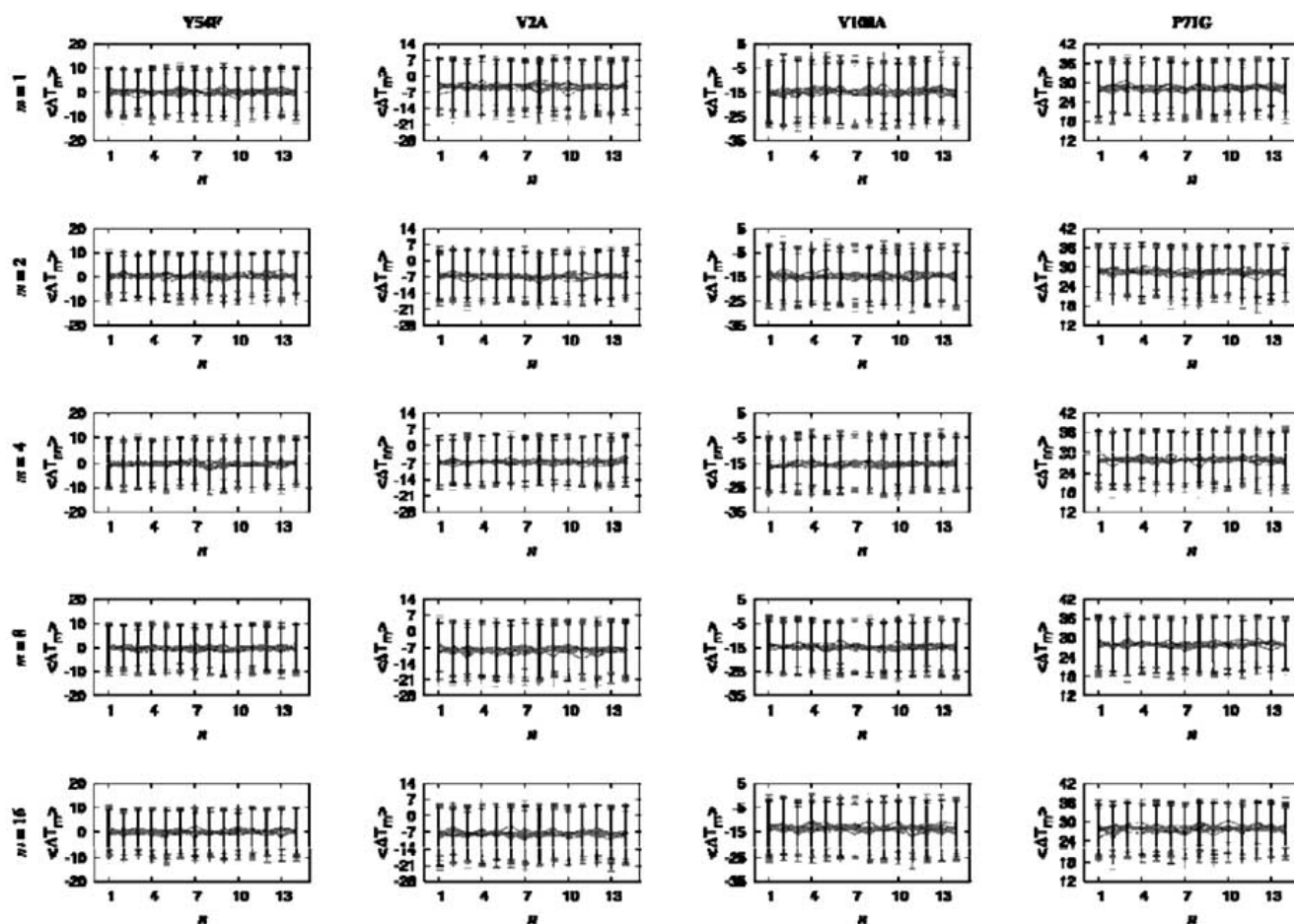


Fig. (7). Cross-sections of the $n \times m$ landscape for four different lysozyme mutant examples (columns). In each case, average prediction accuracy is reported over all values of n for a given value of m . Five different values of m (rows) are shown. In all cases, our results surprisingly demonstrate that increasing the amount of parameter diversity does not improve the average difference between the experimental and predicted T_m values. The results from ten different simulations are shown superimposed on each other to show that our results are robust. Across the ten simulations, the average behavior is largely conserved, and in each case the average ΔT_m values are within the error bars of the other nine.

average ΔT_m values are well within the error bars of the other nine simulations. Unexpectedly, no accuracy trends with increasing n or m are observed, meaning increased parameter diversity does not improve average prediction accuracy. Rather, for a given target, any particular parameter set gives a similar accuracy to any other set, indicating that the quality of the prediction is almost entirely dependent on the target structure itself. For example, Y45F is among the best-predicted structures, resulting in $\langle \Delta T_m^{Y45F} \rangle \approx 0 \pm 10\text{K}$. Conversely, as discussed above, the P71G mutation is particularly problematic, resulting in $\langle \Delta T_m^{P71G} \rangle \approx 28 \pm 10\text{K}$. Note that while a difference of 28K might seem large at first glance, it is in fact only an 8.3% percent error. The mutations V2A and V100A are shown as intermediate examples.

DISCUSSION

The initial objective of this work was to improve the predictive value of the mDCM in a scenario that employs contextual learning. The idea was that a set of best-fit parameters for the mDCM, based on an experimentally determined structure and heat capacity measurement, could be used to predict the relative stability of protein mutants. As more experiments are performed, additional best-fit parameterizations could be determined based on the new systems, thereby boosting statistics, and, as such, better accuracy would occur as more experimental data is obtained. While it was initially surprising that increasing the amount of parameter diversity does not improve prediction accuracy in a statistically significant way, this result can be viewed in two ways. First, there may exist additional features of a protein that we can incorporate to filter out better parameter sets for a given structure. Second, these results reveal the saturation of accuracy inherent within the mDCM.

With the former view, perhaps the original objective of a context learning approach can be recovered by using a more sophisticated statistical analysis. In this work, the considered model parameters provide a range of predicted T_m values, but all parameters sets are treated with equal weighting. Meaning, the collective statistics from the 'good' and 'bad' sets for a given structure cancel out in the average statistics, leading to \pm standard deviation $\sim 20\text{K}$. If we could, somehow, only apply the parameter sets best suited to a particular structure, then the average prediction accuracy will naturally improve as n increases. To that end, it would be necessary to develop a classifier that identifies a good parameter set for a given structure to allow for a knowledge-based weighted average. For example, along these lines we considered estimating a target value for the u_{sol} parameter describing the average enthalpy for H-bonding to solvent as a function of global intramolecular properties of H-bonds. However, virtually no correlation between descriptors of global network properties and model parameters (in addition to model predictions as mentioned above) are found. Therefore, boosting the predictive accuracy using a classification scheme is likely to produce only marginal gains on prediction accuracy of a model that is intrinsically oversimplified. The expected marginal gain leads us to view these results in terms of a physical interpretation.

Based on the simplicity of the mDCM, where the H-bond network is featured so prominently in its free energy functional [18, 19], it should be surprising that the mDCM predictions yield such a high degree of accuracy and transferability of parameters. However, the mDCM has consistently proven to be a very robust and reliable model [18, 19, 30-32], presumably because of the encoded information that lies within the H-bond network. Of course, the importance of H-bonds has been part of a long-standing paradigm in protein biophysics [44]. This work benchmarks the best accuracy level that can be hoped for using the mDCM since we are working with essentially an ideal system for its application. Yet, we are not suggesting the mDCM is the end of the story. On the contrary, the minimal DCM does not explicitly model other essential mechanisms such as hydrophobic and long-range electrostatic interactions. As demonstrated by Guerois *et al.* [45], inclusion of essential mechanisms will improve model accuracy. Similarly, we expect a considerable gain in accuracy will come forth when we employ a more complete free energy decomposition scheme, which we are currently in the process of developing.

CONCLUSION

We establish that the mDCM is a viable approach to predict the relative stability of protein mutants. Even using a single parameter set from some previously fit example, the average error of the method when applied to an unknown example is very good (average percent error = 4.3%), and it does a reasonably good job of reproducing experimental trends ($R = 0.64$), which is definitely good enough to be of practical value to experimentalists when making decisions about which mutations to invest time and funds for characterization. The results also point to the intrinsic limits of such a simplified model, and points to the need to develop a more complete free energy decomposition scheme.

ACKNOWLEDGEMENTS

This work was supported by National Institute of Health grant R01 GM073082 to DJJ and DRL. The graph-rigidity algorithm is claimed in U.S. Patent Number 6,014,449, which has been assigned to the Board of Trustees Michigan State University. Used with permission.

LIST OF ABBREVIATIONS

DCM	=	Distance constraint model
DOF	=	Degrees of freedom
mDCM	=	Minimal distance constraint model
DSC	=	Differential scanning calorimetry
PDB	=	Protein data bank

REFERENCES

- [1] Huang, L.T.; Saraboji, K.; Ho, S.Y.; Hwang, S.F.; Ponnuswamy, M.N.; Gromiha, M.M. Prediction of protein mutant stability using classification and regression tool. *Biophys. Chem.*, **2007**, *125*(2-3), 462-470.
- [2] Huang, L.T.; Gromiha, M.M.; Ho, S.Y. Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. *J. Mol. Model.*, **2007**, *13*(8), 879-890.

- [3] Huang, L.T.; Gromiha, M.M.; Ho, S.Y. iPTREE-STAB, interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics*, **2007**, *23*(10), 1292-1293.
- [4] Cheng, J.; Randall, A.; Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **2006**, *62*(4), 1125-1132.
- [5] Capriotti, E.; Fariselli, P.; Calabrese, R.; Casadio, R. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **2005**, *21* Suppl 2, ii54-58.
- [6] Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0, predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **2005**, *33*(Web Server issue), W306-310.
- [7] Rost, B.; Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **1994**, *19*(1), 55-72.
- [8] Caballero, J.; Fernandez, L.; Abreu, J.I.; Fernandez, M. Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *J. Chem. Inf. Model*, **2006**, *46*(3), 1255-1268.
- [9] Capriotti, E.; Fariselli, P.; Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **2004**, *20*(Suppl 1), i63-68.
- [10] Alber, T. Mutational effects on protein stability. *Annu. Rev. Biochem.*, **1989**, *58*, 765-798.
- [11] Strickler, S.S.; Gribenko, A.V.; Keiffer, T.R.; Tomlinson, J.; Reihle, T.; Loladze, V.V.; Makhatazde, G.I. Protein stability and surface electrostatics, a charged relationship. *Biochemistry*, **2006**, *45*(9), 2761-2766.
- [12] Makhatazde, G.I.; Loladze, V.V.; Ermolenko, D.N.; Chen, X.; Thomas, S.T. Contribution of surface salt bridges to protein stability, guidelines for protein engineering. *J. Mol. Biol.*, **2003**, *327*(5), 1135-1148.
- [13] Makhatazde, G.I.; Loladze, V.V.; Gribenko, A.V.; Lopez, M.M. Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions. *J. Mol. Biol.*, **2004**, *336*(4), 929-942.
- [14] Torrez, M.; Schultehrich, M.; Livesay, D.R. Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces. *Biophys. J.*, **2003**, *85*(5), 2845-2853.
- [15] Dong, F.; Vijayakumar, M.; Zhou, H.X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys. J.*, **2003**, *85*(1), 49-60.
- [16] Zhou, H.X.; Dong, F. Electrostatic contributions to the stability of a thermophilic cold shock protein. *Biophys. J.*, **2003**, *84*(4), 2216-2222.
- [17] Dong, F.; Zhou, H.X. Electrostatic contributions to T4 lysozyme stability, solvent-exposed charges versus semi-buried salt bridges. *Biophys. J.*, **2002**, *83*(3), 1341-1347.
- [18] Jacobs, D.J.; Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.*, **2005**, *88*(2), 903-915.
- [19] Livesay, D.R.; Dallakyan, S.; Wood, G.G.; Jacobs, D.J. A flexible approach for understanding protein stability. *FEBS Lett.*, **2004**, *576*(3), 468-476.
- [20] Razvi, A.; Scholtz, J.M. Lessons in stability from thermophilic proteins. *Protein Sci.*, **2006**, *15*(7), 1569-1578.
- [21] Meirovitch, H. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Curr. Opin. Struct. Biol.*, **2007**, *17*(2), 181-186.
- [22] Dill, K.A. Additivity principles in biochemistry. *J. Biol. Chem.*, **1997**, *272*(2), 701-704.
- [23] Mark, A.E.; van Gunsteren, W.F. Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J. Mol. Biol.*, **1994**, *240*(2), 167-176.
- [24] Jacobs, D.J.; Dallakyan, S.; Wood, G.G.; Heckathorne, A. Network rigidity at finite temperature, relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, **2003**, *68*(6 Pt 1), 061109.
- [25] Jacobs, D.J.; Rader, A.J.; Kuhn, L.A.; Thorpe, M.F. Protein flexibility predictions using graph theory. *Proteins*, **2001**, *44*(2), 150-165.
- [26] Jacobs, D.J.; Thorpe, M.F. Generic rigidity percolation, The pebble game. *Phys. Rev. Lett.*, **1995**, *75*(22), 4051-4054.
- [27] Vorov, O.K.; Livesay, D.R.; Jacobs, D.J. Helix/coil nucleation, a local response to global demands. *Biophys. J.*, **2009**, *97*(11), 3000-3009.
- [28] Dahiyat, B.I.; Gordon, D.B.; Mayo, S.L. Automated design of the surface positions of protein helices. *Protein Sci.*, **1997**, *6*(6), 1333-1337.
- [29] Fernandez, A.; Kardos, J.; Goto, Y. Protein folding, could hydrophobic collapse be coupled with hydrogen-bond formation? *FEBS Lett.*, **2003**, *536*(1-3), 187-192.
- [30] Livesay, D.R.; Jacobs, D.J. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins*, **2006**, *62*(1), 130-143.
- [31] Jacobs, D.J.; Livesay, D.R.; Hules, J.; Tasayco, M.L. Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model. *J. Mol. Biol.*, **2006**, *358*(3), 882-904.
- [32] Livesay, D.R.; Huynh, D.H.; Dallakyan, S.; Jacobs, D.J. Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem. Cent. J.*, **2008**, *2*, 17.
- [33] Mottonen, J.M.; Xu, M.; Jacobs, D.J.; Livesay, D.R. Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *Proteins*, **2009**, *75*(3), 610-627.
- [34] Yamagata, Y.; Kubota, M.; Sumikawa, Y.; Funahashi, J.; Takano, K.; Fujii, S.; Yutani, K. Contribution of hydrogen bonds to the conformational stability of human lysozyme, calorimetry and X-ray analysis of six tyrosine --> phenylalanine mutants. *Biochemistry*, **1998**, *37*(26), 9355-9362.
- [35] Takano, K.; Yamagata, Y.; Fujii, S.; Yutani, K. Contribution of the hydrophobic effect to the stability of human lysozyme, calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants. *Biochemistry*, **1997**, *36*(4), 688-698.
- [36] Takano, K.; Ota, M.; Ogasahara, K.; Yamagata, Y.; Nishikawa, K.; Yutani, K. Experimental verification of the 'stability profile of mutant protein' (SPMP) data using mutant human lysozymes. *Protein Eng.*, **1999**, *12*(8), 663-672.
- [37] Takano, K.; Tsuchimori, K.; Yamagata, Y.; Yutani, K. Effect of foreign N-terminal residues on the conformational stability of human lysozyme. *Eur. J. Biochem.*, **1999**, *266*(2), 675-682.
- [38] Funahashi, J.; Takano, K.; Yamagata, Y.; Yutani, K. Contribution of amino acid substitutions at two different interior positions to the conformational stability of human lysozyme. *Protein Eng.*, **1999**, *12*(10), 841-850.
- [39] Funahashi, J.; Takano, K.; Ogasahara, K.; Yamagata, Y.; Yutani, K. The structure, stability, and folding process of amyloidogenic mutant human lysozyme. *J. Biochem.*, **1996**, *120*(6), 1216-1223.
- [40] Herning, T.; Yutani, K.; Inaka, K.; Kuroki, R.; Matsushima, M.; Kikuchi, M. Role of proline residues in human lysozyme stability, a scanning calorimetric study combined with X-ray structure analysis of proline mutants. *Biochemistry*, **1992**, *31*(31), 7077-7085.
- [41] Robertson, A.D.; Murphy, K.P. Protein structure and the energetics of protein stability. *Chem. Rev.*, **1997**, *97*(5), 1251-1268.
- [42] Gordon, J.C.; Myers, J.B.; Folta, T.; Shoja, V.; Heath, L.S.; Onufriev, A. H++, a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.*, **2005**, *33*(Web Server issue), W368-371.
- [43] Gomez, J.; Hilsner, V.J.; Xie, D.; Freire, E. The heat capacity of proteins. *Proteins*, **1995**, *22*(4), 404-412.
- [44] Pace, C.N.; Shirley, B.A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J.*, **1996**, *10*(1), 75-83.
- [45] Guerois, R.; Nielsen, J.E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes, a study of more than 1000 mutations. *J. Mol. Biol.*, **2002**, *320*(2), 369-387.