Predicting changes in protein thermostability upon point mutation with deep 3D convolutional neural networks

Bian Li[1,2,4], Yucheng T. Yang[1,2], John A. Capra[4*], Mark B. Gerstein[1,2,3*]

1 Program in Computational Biology and Bioinformatics, 2 Department of Molecular Biophysics and Biochemistry, 3 Department of Computer Science, Yale University, New Haven, CT 06520, USA; 4 Department of Biological Sciences and Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA

* To whom correspondence should be addressed: tony.capra@vanderbilt.edu (J.A.C.) and pi@gersteinlab.org (M.B.G.)

Abstract

Predicting mutation-induced changes in protein thermodynamic stability (ΔΔG) is of great interest in protein engineering, variant interpretation, and drug discovery. We introduce ThermoNet, a deep, 3D-convolutional neural network designed for structure-based prediction of the change in protein thermostability upon point mutation. To naturally leverage the image-processing power inherent in convolutional neural networks, we treat protein structures as if they were multi-channel 3D images. In particular, the inputs to ThermoNet are multi-channel voxel grids based on biophysical properties derived from raw atom coordinates; thus, the interpretability of its resulting predictions is improved compared to machine-learning approaches that take heterogeneous representations of predictive features as inputs. ThermoNet is trained with a data set balanced with direct and reverse mutations generated by symmetry-based data augmentation. It demonstrates improved performance compared to fifteen previously developed computational methods on a widely used blind test set. Unlike all other methods that exhibit a strong bias towards predicting destabilization, ThermoNet accurately predicts the effects of both stabilizing and destabilizing mutations. Finally, we demonstrate the practical utility of ThermoNet in predicting the ΔΔG landscape for two clinically relevant proteins, p53 and myoglobin, and all ClinVar missense variants. These predictions indicate that 85.1% of benign variants have effects on protein stability that are in the range expected to be neutral and that pathogenic variants are almost equally likely to be destabilizing as stabilizing (56.6% vs 43.4%, respectively). Overall, our results suggest that 3D convolutional neural networks can model the complex, non-linear interactions perturbed by mutations, directly from biophysical properties of atoms.

Key words: protein thermostability; single nucleotide variants; deep convolutional neural networks; symmetry-based data augmentation

Introduction

The thermodynamic stability of a protein, usually represented as the Gibbs free energy for the reaction of protein folding (ΔG), is a fundamental physical quantity. The magnitude of ΔG is collectively determined by the intramolecular interactions between amino acid residues within the protein and the interactions between the protein and the physiological environment around it (Li et al., 2018). When a mutation causes amino acid substitution in a protein, it is likely that the stability of the mutant protein will be affected compared to the wild-type protein. (Note that the term “wild-type” isn’t preferred because our genetic makeup is intrinsically diverse (Auton et al., 2015). A better term would be a “reference state”. However, as it will not cause confusion in this work and for consistence with many previous works, we use wild type throughout the text.) This change in protein thermodynamic stability (i.e. ΔΔG) incurred by mutation is of fundamental importance to medicine and biotechnology. Many disease-causing mutations are single-point

amino acid substitutions that lead to a substantial ΔΔG of the corresponding protein, and such single-point mutations are a key mechanism underlying a wide spectrum of molecular disorders (Stein et al., 2019; Wang and Moult, 2001; Yue et al., 2005). Given the huge number of variants discovered during large-scale population-level exome and genome sequencing studies and clinical genetic tests, there is a tremendous interest in predicting whether these variants are likely to exert any impact on protein function. In addition, in developing new biopharmaceuticals, one of the early goals is usually to design proteins with the intended thermostability. However, this task is often laborious, if not impossible, and usually involves experimentally screening an enormous number of mutant proteins (Huang et al., 2016). Thus, it is desirable to have an efficient and accurate computational tool to prioritize the set of mutant proteins to be experimentally tested.

Toward these goals, several programs have been developed for estimating ΔΔGs. These methods either rely on explicit biophysical modeling of amino acid interactions coupled with conformational sampling of protein structures (Bender et al., 2016; Gapsys et al., 2016; Kellogg et al., 2011; Worth et al., 2011; Yin et al., 2007) or apply machine/statistical learning to extract patterns from various types of amino acid sequence, evolutionary, and/or protein structural features (Cao et al., 2019; Capriotti et al., 2005; Dehouck et al., 2011; Fariselli et al., 2015; Guerois et al., 2002; Masso and Vaisman, 2008; Parthiban et al., 2006; Pires et al., 2014; Quan et al., 2016). While these methods have been useful in many applications (Buss et al., 2018; Roushar et al., 2019), they have substantial limitations. For example, physics-based methods are computationally demanding and low-throughput; these challenges have largely prevented them from being applied to large-scale protein engineering and variant interpretation tasks. In addition, several studies have highlighted significant bias in the predictions of machine learning-based methods; they tend to predict mutations as destabilizing more often than stabilizing (Fariselli et al., 2015; Pucci et al., 2018; Thiltgen and Goldstein, 2012; Usmanova et al., 2018). The main source of this bias likely comes from the facts that the training sets are dominated by experiment-derived destabilizing mutations and that machine learning methods are prone to overfitting to training sets (Fang, 2019; Pucci et al., 2018). Thus, there is a need for new methods that can make quantitative, unbiased prediction of ΔΔGs with high throughput.

Here, we describe ThermoNet, a computational framework based on deep 3D convolutional neural networks (3D CNNs) for predicting ΔΔGs upon single-point mutation. We address the bias towards predicting destabilization by training ThermoNet on a balanced data set generated through symmetry-based data augmentation, i.e. for each mutation, we consider both the direct and reverse versions. We model the structure of each mutation assuming that single-point mutations introduce negligible perturbation to the backbone of protein structure. We treat protein structures as if they were 3D images with voxels parameterized using atom biophysical properties (Jimenez et al., 2017; Jimenez et al., 2018). We leverage the power of the constrained architecture of CNNs in detecting spatially proximate features. These local biochemical interaction detectors are then hierarchically composed into more intricate features with the potential to describe the complex and nonlinear phenomenon of molecular interaction. We show that ThermoNet achieves state-of-the-art performance comparable to previously developed methods on a widely used blind test set with minimal prediction bias. We also demonstrate the applicability of ThermoNet by showing that ThermoNet accurately predicts the ΔΔGs of the missense mutations in two biologically important proteins, the p53 tumor suppressor protein and myoglobin

and that ThermoNet-predicted ΔΔGs of ClinVar variants fall within the experimentally observed range and are consistent with the expectations of a biophysical model of protein evolution.

Results

An overview of ThermoNet

To predict the ΔΔG of a point mutation, we take advantage of recent advances in deep learning for computer vision (LeCun et al., 2015) and the successes of deep convolutional neural networks in biophysical problems (Jimenez et al., 2017; Jimenez et al., 2018; Torng and Altman, 2017, 2019; Wallach et al., 2015). We treat protein structures as if they were 3D images with voxels parameterized using atom biophysical properties (Jimenez et al., 2017; Jimenez et al., 2018) (Fig. 1A,1B, and Table 1). For a given single-point mutation, ThermoNet requires that a 3D structure (either experimentally determined or modeled via homology modeling) of one of the alleles is available. As a first step, ThermoNet constructs a structural model for the mutant from the structure of the wild type using the Rosetta macromolecular modeling suite (Methods) (Bender et al., 2016; Leaver-Fay et al., 2011). ThermoNet assumes that the ΔΔG of a point mutation can be sufficiently captured by modeling the 3D physicochemical environment around the mutation site. It thus extracts predictive features by treating protein structures as if they were 3D images and voxelizing the space around the mutation site of both the wild-type structure and the corresponding mutant structural model (Fig. 1C). Each voxel is parameterized with seven predefined rules (Table 1) to characterize the physicochemical nature of its neighboring atoms. The feature maps are then stacked to create a tensor with size $[16,16,16,14]$ as input to the trained deep 3D-CNN, which generate a prediction of the ΔΔG the given mutation causes to the wild-type structure. The architecture of the 3D-CNN was tuned via five-fold cross-validation using a data set consisting of 3,214 distinct mutations in 148 proteins (Methods, Fig. S1). It consists of three 3D convolutional layers with 16, 24, and 32 neurons respectively and one densely connected layer of 24 neurons (Fig. 1D).

ThermoNet achieves state-of-the-art performance on blind test set

We systematically compared ThermoNet with fifteen ΔΔG predictors on a common, balanced data set S^{sym} to evaluate their performance and degree of bias with respect to the ΔΔG symmetry between direct and reverse mutations (Methods). The S^{sym} data set, which was constructed previously for assessing the biases of ΔΔG predictors (Pucci et al., 2018), consists of experimentally measured ΔΔG values for 342 direct mutations from fifteen protein chains for which the structures of both the wild-type and mutant proteins have been resolved by X-ray crystallography with a resolution of 2.5 Å or better. To increase the size of this data set and evaluate prediction bias of ThermoNet, we generated a reverse mutation for each of the direct mutations, thus totaling 684 mutations. This data set is by construction balanced with respect to stabilizing and destabilizing mutations. Thus, it is an independent test set that also enables evaluation of prediction bias.

Table 2 contains the values of root mean square error σ and the Pearson correlation coefficient r for all the methods surveyed in our study, reported separately for the direct and reverse mutations. Prediction bias is measured by two parameters, the Pearson correlation coefficient $r_{dir-rev}$ between the predictions for direct and those for reverse mutations and the δ value, defined as:

$\delta = \Delta\Delta G_{rev} + \Delta\Delta G_{dir}$. A perfectly unbiased predictor would give $r_{div-rev} = -1$ and $\langle\delta\rangle = 0\ kcal/mol$. ThermoNet achieves comparable prediction accuracy for both direct mutations ($r_{dir} = 0.58$ and $\sigma_{dir} = 1.42\ kcal/mol$, Fig. 2A) and the corresponding set of reverse mutations ($r_{rev} = 0.59$ and $\sigma_{rev} = 1.38\ kcal/mol$, Fig. 2C). The fractions of mutations for which the prediction error is within $0.5\ kcal/mol$ and $1.0\ kcal/mol$ are 38.9% and 62.6% for direct mutations and 41.8% and 64.3% for reverse mutations (Fig. 2B and 2D). PoPMuSiC^{sym} is the only other method with similar accuracy on direct and reverse mutations, but its accuracy is worse than ThermoNet (Table 2). ThermoNet's prediction accuracy on direct mutations is not ranked the best compared to other ΔΔG predictors; however, its performance on reverse mutations is the best by a large margin ($r_{rev} = 0.59$ vs. 0.48 for PoPMuSiC^{sym}). Compared to the other methods, ThermoNet also has the least prediction bias with the strongest correlation between direct and reverse predictions ($r_{dir-rev} = -0.95$, Fig. 2E, Table 2) and bias $\langle\delta\rangle$ (-0.05) comparable to PoPMuSiC^{sym}. We also report the distribution of δ, since $\langle\delta\rangle$ cannot distinguish large, but symmetric, bias from low bias (Methods). As shown in Fig. 2F, 36.3% and 91.2% of mutations have a prediction bias $<$ $0.1\ kcal/mol$ and $<$ $0.5\ kcal\ /\ mol$, respectively.

Structural models of reverse mutations are necessary for unbiased ΔΔG predictions

ThermoNet was trained on the spatial physicochemical environment around the mutation sites, which were represented using their raw atom coordinates. The primary training set consists of 3,214 missense mutations from 117 protein chains. This training set was obtained by excluding S^{sym} variants from the data set used to train STRUM (Quan et al., 2016). For each mutation from X to Y in the training set, which we designate as the "direct" mutation, we created a data point for the "reverse" mutation from Y to X with a ΔΔG value that has the same magnitude but opposite sign of the ΔΔG value of the direct mutation (Methods).

To evaluate whether the inclusion of the reverse mutations is necessary for the reduction in prediction bias, we trained an ensemble of ten networks using a data set consisting of only the 3,214 direct mutations and their associated experimental ΔΔGs. We applied this ensemble of networks to predict the ΔΔGs of the direct and reverse mutations of the S^{sym} blind test set. As shown in Fig. S2, ΔΔGs of direct mutations predicted by these models correlate reasonably well ($r = 0.59$ and $\sigma = 1.31\ kcal/mol$) with the experimental values and are comparable to the performance of the ensemble of networks trained using the balanced data set. In contrast, these models perform poorly ($r = -0.16$ and $\sigma = 2.79\ kcal/mol$) in predicting the ΔΔGs of the corresponding set of reverse mutations (Fig. S2). This suggests that the models were biased toward the training set which is dominated by destabilizing mutations. This is confirmed by the strongly positive correlation between the predictions for direct mutations and those for reverse mutations and the large prediction bias ($r_{dir-rev} = 0.57$ and $\langle\delta\rangle = 2.53\ kcal/mol$) (Fig. S2). Compared to the performance of the networks trained using the balanced data set, the results highlight the necessity of a balanced data set for correcting prediction bias.

Case studies: The p53 tumor suppressor protein and myoglobin.

We further tested ThermoNet by predicting the ΔΔGs of single-point mutations in the p53 tumor suppressor protein and myoglobin whose thermodynamic effects have previously been experimentally measured. The *TP53* tumor suppressor gene encodes the p53 transcription factor that is mutated in ~45% of all human cancers (Beroud and Soussi, 2003; Olivier et al. 2002).

Unlike most tumor suppressor proteins that are inactivated by deletion or truncation mutations, single amino acid substitutions in p53 often modify DNA binding or disrupt the conformation and stability of p53 (Olivier et al., 2002). Myoglobin is a cytoplasmic globular protein that regulates cellular oxygen concentration in cardiac myocytes and oxidative skeletal muscle fibers by reversible binding of oxygen through its heme prosthetic group (Ordway and Garry, 2004). The p53 data set consists of 42 mutations compiled in a previous study (Pires et al., 2014) within the DNA binding domain of p53. The myoglobin data set consists of 134 mutations scattered throughout the protein chain compiled in a previous study (Kepp, 2015). None of the mutations in these two data sets were present in the training set. We used published crystal structures of p53 (PDB ID: 2OCJ) and myoglobin (PDB ID: 1BZ6) to create one structural model for each of the mutations in these two data sets respectively using the *FastRelax* protocol in Rosetta (Tyka et al., 2011). These predictions were compared directly with the experimentally determined thermodynamic effects (Fig. 3). For p53, ΔΔGs of both direct and reverse mutations predicted by ThermoNet correlate with the experimental measurements ($r = 0.50$ and 0.65) and have little bias ($r_{dir-rev} = -0.91$ and $\langle \delta \rangle = -0.03\ kcal/mol$). However, the predicted ΔΔGs of myoglobin mutations correlate less well ($r = 0.37$ and 0.40 for direct and reverse mutations respectively) compared to those of p53, though the bias is also low. The poorer correlations for myoglobin are likely because the myoglobin data set consists of ΔΔG measurements obtained under various different experimental conditions which ThermoNet does not explicitly account for.

While the correlation between predicted and experimentally measured ΔΔGs is not perfect, the predictions are generally conservative – no mutations with low measured ΔΔG are predicted to have a high ΔΔG. These results demonstrate the utility of ThermoNet as rapid unbiased predictor of ΔΔG for mutations in clinically relevant proteins.

ΔΔG landscape of ClinVar missense variants

To further demonstrate the utility of ThermoNet in evaluating mutations with real-world relevance, we predicted the ΔΔG distributions of missense variants in ClinVar, a widely used resource of medically important variants (Landrum et al., 2018). For comparison, we also applied FoldX, a popular and freely available ΔΔG predictor (Buss et al., 2018; Guerois et al., 2002) to the ClinVar set. We first examined the overall ΔΔG distribution of ClinVar variants. The ΔΔGs of ClinVar variants predicted by ThermoNet range from -2.91 kcal/mol to +4.48 kca/mol (Fig. 4A). Experimentally measured ΔΔGs generally fall within -5 kcal/mol to +5 kcal/mol (DePristo et al., 2005; Guerois et al., 2002); thus, ThermoNet's predictions are consistent with the range of observed values. In contrast, the ΔΔGs of the same set of variants predicted by FoldX range from -6.64 kcal/mol to +57.2 kcal/mol (Fig. 4A), and 15.2% of ΔΔGs predicted by FoldX are outside the expected range of -5 kcal/mol to +5 kcal/mol.

ΔΔGs predicted by ThermoNet are also consistent with the expected ΔΔG distributions according to a biophysical model of protein evolution (DePristo et al., 2005). This model hypothesizes that fitness is a non-monotonic, concave function of protein stability, meaning that fitness decreases with increasing deviation from an optimal stability. The model also suggests that there is a "neutral zone" of 1.0 kcal/mol around the optimal stability in stability space, and mutations whose impact on stability fall within the neutral zone will have little effect on fitness (DePristo et al., 2005). We thus reasoned that the ΔΔGs of benign variants will fall within a narrow range from -0.5 kcal/mol

to +0.5 kcal/mol and that pathogenic variants are equally likely to be destabilizing or stabilizing (DePristo et al., 2005). To test this hypothesis, we examined the ΔΔG distributions of pathogenic and benign variants separately. Remarkably, the ΔΔGs of 85.1% of benign variants predicted by ThermoNet fall within the neutral zone, whereas FoldX only predicted 39.7% of benign variants to be in the neutral zone (Fig. 4B). Further, the ΔΔGs of pathogenic variants predicted by ThermoNet suggest pathogenic variants are nearly equally likely to be destabilizing (56.6%) as they are to be stabilizing (43.4%) (Fig. 4C). In contrast, FoldX predicted that 83.2% of pathogenic variants are destabilizing. As already demonstrated in previous studies, the bias is likely because FoldX was parameterized on an experimental ΔΔG data set dominated by destabilizing mutations (Pucci et al., 2018; Thiltgen and Goldstein, 2012; Usmanova et al., 2018).

ThermoNet's predictions are also consistent with the fact that variant pathogenicity can only be partially attributed to impacts on protein stability. ThermoNet predicts that 53.4% of pathogenic variants that have ΔΔGs within the neutral zone and 14.9% of benign variants have ΔΔGs outside the neutral zone (Fig. 4B and 4C). This is expected based on previous biochemical characterizations of pathogenic mutations. For example, Bromberg *et al.* collected 66 mutations with experimentally measured ΔΔG and functional annotations from the literature. The ΔΔGs of this set of mutations range from -4.3 to 4.96 kcal/mol and the authors found that 31% of mutations affecting function had ΔΔGs within the neutral zone while 19% functionally neutral mutations had ΔΔGs outside the neutral zone (Bromberg and Rost, 2009).

Discussion

Accurate modeling of protein thermostability is a complex and challenging task considering the delicate balance between the different thermodynamic state functions that contribute to protein stability (Li et al., 2018). The primary goal of this paper is to present a novel application of deep 3D convolutional neural networks to a challenge in structural bioinformatics: predicting changes in thermodynamic stability upon point mutation. We formulated the problem of ΔΔG prediction from a computer vision perspective and took full advantage of the power of the constrained architecture of convolutional neural networks in detecting spatially proximate features (LeCun et al., 2015). We developed ThermoNet, an ensemble of deep 3D convolutional neural networks, designed to predict ΔΔG upon point mutation. We showed that ΔΔG can be accurately predicted from protein structure using deep 3D convolutional neural networks without manual feature engineering. More importantly, we demonstrated that the pressing issue of bias in ΔΔG prediction can be remedied by modeling mutant structures and training ThermoNet with structures of both the wild-type and mutant proteins. While ThermoNet achieved comparable performance to previous methods on direct mutations, it performed better on reverse mutations than other methods by a large margin, and remarkably, reduced the magnitude of prediction bias.

ThermoNet treats protein structures as if they were 3D images, and it takes as input a 3D grid of voxels parameterized with seven chemical property channels. As such, this approach bypasses the tedious processes of manual feature engineering and feature selection that, if not done correctly, can often lead to over-optimistic estimation of model performance. The locally constrained deep convolutional architecture likely allows the system to model the complex, non-linear nature of molecular interactions. Our ThermoNet method demonstrates the potential of

deep convolutional neural networks for predicting biophysically meaningful information from protein structures and holds promise for protein engineering.

The fact that our approach relies on the availability of experimental structures or homology models and the 3D nature of the convolutional neural network create two limitations. First, while protein structures are being determined at an unprecedented pace, the faction of the human proteome with available experimental structure is estimated to be around 20% (Somody et al., 2017). Even when all the proteins whose structures can be modeled reliably are considered, only ~70% of the human proteome will have structural coverage (Somody et al., 2017). As the structures of many proteins can only be partially modeled, the space of the human proteome that one can apply ThermoNet to will be less than 70%. Second, compared to the 2D version with the same architecture, ThermoNet has four times more parameters: three convolutional layers of 16, 24, and 32 neurons respectively, and one dense layer with 24 neurons that takes an input tensor of the shape [16, 16, 16, 14] has 133,273 parameters. Training deep 3D CNNs is very demanding, requiring more GPU memory and more training data to avoid overfitting. While we demonstrated the potential of deep 3D CNNs in modeling ΔΔG of proteins, the relatively little training data available raises the question of whether deep 3D CNNs can model ΔΔGs and related thermodynamic properties at experimental accuracy. Nonetheless, the increasing adoption of deep mutational scanning techniques for systematic study of the molecular effects of mutations is generating an unprecedented amount of data. Furthermore, given rapid increases in GPU power and the number of structures of proteins and their complexes determined, we expect that deep 3D CNNs will be successfully applied to provide solutions to many biophysical problems such as modeling the impact of mutation on protein-protein, protein-DNA as well as protein-RNA interactions.

Methods

Protein thermostability

The thermostability ΔG of the folded form of a two-state protein, which is its Gibbs free energy of folding, is defined in relation to the concentration of folded $[folded]$ and the concentration of unfolded $[unfolded]$ forms:

$$\Delta \mathrm{G} = -RTln\frac{[folded]}{[unfolded]}$$

where T is the temperature, and R is the gas constant. More stable proteins, meaning that a higher fraction of the protein is in the folded form, have more negative values of ΔG. The impact of mutations on protein thermostability, ΔΔG, is defined in terms of the change in ΔG between the wild-type and mutant proteins:

$$\Delta\Delta \mathrm{G} = \Delta \mathrm{G}_{mutant} - \Delta \mathrm{G}_{wild-type} = -RTln\frac{[folded]_{mutant}/[unfolded]_{mutant}}{[folded]_{wild-type}/[unfolded]_{wild-type}}$$

such that a destabilizing mutation has a positive ΔΔG, whereas a stabilizing mutation has a negative ΔΔG. The values of ΔΔG resulting from single-point mutations usually range from -5 to 5 kcal/mol (Guerois et al., 2002). Assuming a constant concentration of the unfolded form of a

protein, the concentration of the folded form of a mutant that destabilizes the protein by 1 kcal/mol is only 18.5% of the folded form of the protein if the mutation had not occurred.

Thermodynamics of direct and reverse mutations

Consider a pair of proteins whose sequences differ only at a single position where the amino acid is X in one protein and Y in the other. Let the Gibbs free energies of folding of this pair of proteins be ΔG_X and ΔG_Y respectively. For such a pair of proteins, one can think of the protein Y as being generated by a "direct" mutation at the sequence location from amino acid X to Y and the change in the Gibbs free energy of folding caused by X to Y mutation is $\Delta\Delta G_{X\rightarrow Y} = \Delta G_Y - \Delta G_X$. One may also think of the protein X as being generated by a "reverse" mutation from amino acid Y to X and the change in the Gibbs free energy of folding caused by this reverse mutation is $\Delta\Delta G_{Y\rightarrow X} = \Delta G_X - \Delta G_Y = -\Delta\Delta G_{X\rightarrow Y}$. A well-performing, "self-consistent" method for predicting $\Delta\Delta G$s would not only give accurate $\Delta\Delta G$ predictions for the direct mutations, but also for the reverse mutations. The self-consistency requirement has been largely ignored by previously developed $\Delta\Delta G$ predictors (Pucci et al., 2018; Thiltgen and Goldstein, 2012; Usmanova et al., 2018).

Data sets and symmetry-based data augmentation

The data set used to train ThermoNet was derived from the Q3421 data set compiled in a previous study (Quan et al., 2016). The Q3421 data set contains 3,421 distinct single-point mutations in 150 proteins collected from the ProTherm database (Kumar et al., 2006). The impact of these mutations on the thermostability of the protein structure have been measured experimentally and expressed quantitatively as $\Delta\Delta G$ values. Mutations were excluded from the Q3421 data if they were also in the test set (see following). The final data set used in the training of ThermoNet consisted of 3,214 distinct mutations in 148 proteins. This data set was augmented by creating a reverse mutation data point for each of the 3,214 direct mutations, thus giving to a total of 6,428 data points. The data set was randomly divided into ten equally sized, mutually exclusive subsets each consisting of 10% of the direct mutations and the corresponding 10% of reverse mutations. In the training of each component model of ThermoNet, nine subsets were combined to form a training set and the remaining one subset was used as a validation set. The data set used to test ThermoNet and to compare it with fifteen previously developed $\Delta\Delta$G predictors was a common, balanced data set called S^{sym} consisting of 342 pairs of proteins with known crystal structures (Pucci et al., 2018). The members forming each pair differ at only a single position in the protein sequence. The $\Delta\Delta$G values of the 342 direct mutations have been experimentally measured and the $\Delta\Delta$G values of the corresponding 342 reverse mutations were assigned using anti-symmetry.

Modeling mutant structures

We treat each mutation as a pair of proteins whose sequences differ only at a single sequence position. For each pair of proteins, we designate the one whose structure has been experimentally resolved as protein X and the other as protein Y. Structures of the X proteins were collected from the Protein Data Bank (Rose et al., 2017) and were relaxed in the Rosetta all-atom energy function `ref2015` (Alford et al., 2017) using the Rosetta *FastRelax* protocol (Tyka et al., 2011). To prevent large-scale conformational shift from the input PDB structure, atoms were constrained to their starting locations with a harmonic penalty potential during relaxation. The same Rosetta *FastRelax* protocol was also employed to create structural models for each of the Y proteins from

the corresponding relaxed structure of the X protein by supplying a Rosetta `resfile` specifying the mutation $X \rightarrow Y$ to make. The structures of both proteins of each protein pair in the test set were collected from the Protein Data Bank (Rose et al., 2017) and were also relaxed using the same Rosetta *FastRelax* protocol.

Voxelization of the neighborhood of mutation site

We treated each protein structure as a collection of volume elements (voxels) in 3D space. Just as a pixel element in an image has color channels, we parameterized a voxel in a protein structure by a set of k chemical property channels: $[v_1, v_2, \ldots, v_k]$ where the value v_i of each property channel indicates the level of saturation of property i at this voxel (Fig. 1A and 1B). For each mutation (a pair of proteins), we superimposed the mutant structure onto the wild-type structure such that the root-mean-squared distance between them is minimized and collected a grid of $16 \times 16 \times 16$ voxels from both structures. We parameterized each voxel with seven property channels each of a distinct chemical nature according to AutoDock4 atom types (Morris et al., 2009) as in the work of Jimenez et al. (Jimenez et al., 2018) (Table 1). This resulted in a tensor of the shape $[16, 16, 16, 7]$ for a single structure and a tensor of the shape $[16, 16, 16, 14]$ when the two tensors from both structures are concatenated to represent the mutation. The grid was centered at the C_β atom of the mutation site amino acid (or C_α atom if it's a glycine) where each voxel is a unit cube whose sides are 1 Å long. The level of saturation $f(d)$ of each property channel at each voxel is determined by the van der Waals radius r_{vdw} of the atom designated to have that property and its distance d to the center of the voxel through the following formula:

$$f(d) = 1 - \exp\left[-\left(\frac{r_{vdw}}{d}\right)^{12}\right]$$

The computation of tensors from protein structures was performed using routines implemented in the HTMD Python library (version 1.17) for molecular simulations (Doerr et al., 2016) and a Python program for creating a data set from a list of mutations is provided in Supplementary Code.

Model architecture

Convolutional neural networks are a type of deep-learning model commonly used in computer vision applications. They have recently proven to perform well in residue contact prediction (Kandathil et al., 2019a; Schaarschmidt et al., 2018; Shrestha et al., 2019; Wang et al., 2018) and protein tertiary structure prediction (Greener et al., 2019; Kandathil et al., 2019b; Xu, 2019; Xu and Wang, 2019). We selected CNNs for protein structure-based $\Delta\Delta G$ prediction because we formulated this problem as a computer vision problem by treating protein structures as if they were 3D images. Each of the component models of ThermoNet features a sequential organization of three 3D convolutional layers (Conv3D), one 3D max pooling layer (MaxPool3D), followed by one fully connected layer (Dense) (Fig. S1). Convolutions operate over 4D tensors, called feature maps, with three spatial axes (length, width, and height) as well as a depth axis. The convolution operation extracts 3D patches of shape $[3, 3, 3]$ with stride 1 from its input feature map and applies the same transformation to all patches, producing an output feature map. This output feature map is still a 4D tensor: it has a length, height, and a width whose values are determined by the shape of convolution patches and size of the stride, but its depth, which is also called number of filters, is a hyperparameter of the layer (see the section on hyperparameter search below). The number of filters in the three Conv3D layers were 16, 24, and 32 respectively. All convolution operation

outputs from each Conv3D layer are transformed by the rectified linear activation function (ReLU). The transformed outputs from the last Conv3D layer are pooled by taking the maximum activation of each $2 \times 2 \times 2$ grid. The max-pooled activations are then flattened into a 1D vector of features which are fully connected with a dense layer of 64 ReLU units. This model architecture was implemented using Keras (Chollet, 2015) with TensorFlow (Abadi et al., 2016) as the backend.

Training ThermoNet

ThermoNet is an ensemble of ten deep 3D convolutional neural networks, each trained to perform the best on a validation set. Each of the component models was trained on nine subsets (collectively known as the training set), and its generalization performance was monitored on the remaining one subset (validation set). This process was iterated ten times each using a different one of the ten subsets as the validation set and the remaining nine subsets as the training set. All component models of ThermoNet were trained using the Adam optimizer (Kingma and Ba, 2014) for 200 epochs with default hyperparameters (maximum learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Kernel weights of the model were initialized using the Glorot uniform initializer and updated after processing each batch of eight training examples. The mean squared error (MSE) of the predicted $\Delta\Delta$G values from the experimental measurements was used as the loss function during training. When training a deep neural network, one often cannot predict how many epochs will be needed to get to an optimal validation loss. We monitored the MSE of the predictions on a separate validation set consisting of 10% variants randomly selected from the training set during training. Training was stopped when the MSE on the validation set stopped decreasing for ten consecutive epochs. To regularize the model, the dense layer was placed between two dropout layers with a dropout rate of 0.5 in each layer (Fig. S1). Each of the final component models was the one that produced the lowest MSE on the validation set. The final predicted $\Delta\Delta$G value is the average of predictions from the ten models.

Hyperparameter search

The design of deep CNNs entails many architectural choices to account for number of hidden convolutional layers and fully connected layers, number of filters, filter size, strides, padding, dropout rate among many other hyperparameters. We initially created a voxel grid with size $16 \times 16 \times 16$ at a resolution of 1 Å for each chemical property channel following the procedure described in (Jimenez et al., 2017) to cross-validate our network architectures. Considering the limited training data set available in our study, we tried some smaller architectures with cross-validation to decide the optimal one, rather than simply adapting the widely used, much larger, network architectures in computer vision applications. We restricted our deep CNNs to have three convolutional layers and one fully connected while considering several hidden layer sizes. Our results from five-fold cross-validation suggest that the architecture of the $16 \times 24 \times 32$ convolutional configuration combined with a fully connected layer of size 24 achieved the best performance (Fig. S1). An additional consideration for our deep 3D CNN architecture is the dimension of the local box. The size of the local box specifies the structural information accessible by the network and therefore is a hyperparameter of our method. We cross-validated the voxelization scheme with grid of sizes $8 \times 8 \times 8$, $12 \times 12 \times 12$, $16 \times 16 \times 16$, and $20 \times 20 \times 20$ at a resolution of 1 Å. All voxelization schemes draw a cubic box around the mutation site with lateral

length of l Å where l equals 8, 12, 16, or 20 Å. Our results from five-fold cross-validation indicate that the voxel grid with size $16 \times 16 \times 16$ gives the best prediction performance (Fig. S1).

Performance evaluation

The following measures were adopted to evaluate the performance of ThermoNet and to facilitate comparison with previously developed methods. The primary measures for evaluating prediction accuracy were the Pearson correlation coefficient (r) between experimental and predicted $\Delta\Delta G$s and the root-mean-squared error (σ) of predictions. For a set of n data points (x_i, y_i), the formula for calculating r and σ are defined as follows:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

where the (x_i, y_i) tuple denotes the experimental and predicted $\Delta\Delta G$ values of mutation i, respectively, and n denotes the number of mutations in the data set. The measures for evaluating prediction bias were the Pearson correlation coefficient between the predictions for direct mutations and those for reverse mutations and the parameter δ which is defined as: $\delta = \Delta\Delta G_{inv} + \Delta\Delta G_{dir}$ and was previously used to quantify prediction bias (Thiltgen and Goldstein, 2012). An unbiased predictor should have $\delta = 0$ for every mutation. The average of δ, $\langle\delta\rangle$, taken over all mutations in the S^{sym} data set was used in two previously studies to quantify prediction bias (Montanucci et al., 2019; Pucci et al., 2018). While we report $\langle\delta\rangle$ in this work, we note that $\langle\delta\rangle$ is flawed because biases toward opposite directions will be washed out when summed. To give a more transparent presentation of prediction bias, we also plot the distribution of δ and report the average of absolute bias, i.e. $\langle|\delta|\rangle$.

ClinVar variants

We retrieved the ClinVar database (Landrum et al., 2018) in VCF format on August 15, 2019 and ran the VCF file through the Variant Effect Predictor (version 97) (McLaren et al., 2016) to annotate the consequences of all ClinVar variants. We created a set of missense variants that can be mapped to protein structures to demonstrate the applicability of ThermoNet to clinically relevant variants. Our evaluation set consists of solely ClinVar missense variants that are labeled as "pathogenic" or "likely pathogenic" for true positive (pathogenic) variants and "benign" or "likely benign" for true negative (benign) variants. All variants are required to have a review status of at least one star and no conflicting interpretation. Due to the dependency of ThermoNet on 3D structures, we also require variants in the evaluation set to be mappable to available protein structures. Any ClinVar variant designated as “no assertion criteria provided”, “no assertion provided”, “no interpretation for the single variant”, or not covered by either a structure or homology model was excluded from the evaluation set. Collectively, these restrictions resulted in 3,305 pathogenic variants and 985 benign variants that can be mapped to experimental structures deposited in the Protein Data Bank (Rose et al., 2017).

Author contribution

B.L. and M.B.G. conceived the study. B.L. implemented ThermoNet. B.L. and Y.T.Y performed the analysis. B.L. performed the work on predicting the ΔΔG distributions of ClinVar variants using ThermoNet at Vanderbilt University with the help from J.A.C. M.B.G. and J.A.C. supervised the project. B.L. wrote the manuscript with help from all authors.

Code and data availability

ThermoNet source code and raw data supporting the analysis of this work is available at https://github.com/gersteinlab/ThermoNet.

Acknowledgement

The authors would like to thank the Yale Center for Research Computing for supporting high-performance computing for this work. This work was also conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M.*, et al.* (2016). TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (Savannah, GA, USA: USENIX Association), pp. 265-283.
Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K.*, et al.* (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput *13*, 3031-3048.
Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.
Bender, B.J., Cisneros, A., Duran, A.M., Finn, J.A., Fu, D., Lokits, A.D., Mueller, B.K., Sangha, A.K., Sauer, M.F., Sevy, A.M.*, et al.* (2016). Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. Biochemistry *55*, 4748-4763.
Bromberg, Y., and Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. BMC bioinformatics *10 Suppl 8*, S8.
Buss, O., Rudat, J., and Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J *16*, 25-33.
Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J.Z. (2019). DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. Journal of Chemical Information and Modeling.
Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research *33*, W306-W310.
Chollet, F. (2015). keras (\url{https://github.com/fchollet/keras}).
Dehouck, Y., Kwasigroch, J.M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC bioinformatics *12*.
DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet *6*, 678-687.
Doerr, S., Harvey, M.J., Noe, F., and De Fabritiis, G. (2016). HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. Journal of Chemical Theory and Computation *12*, 1845-1852.
Fang, J. (2019). A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief Bioinform.

Fariselli, P., Martelli, P.L., Savojardo, C., and Casadio, R. (2015). INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics *31*, 2816-2821.
Gapsys, V., Michielssens, S., Seeliger, D., and de Groot, B.L. (2016). Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. Angew Chem Int Edit *55*, 7364-7368.
Greener, J.G., Kandathil, S.M., and Jones, D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun *10*, 3977.
Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology *320*, 369-387.
Huang, P.S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320-327.
Jimenez, J., Doerr, S., Martinez-Rosell, G., Rose, A.S., and De Fabritiis, G. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics *33*, 3036-3042.
Jimenez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. (2018). K-DEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of Chemical Information and Modeling *58*, 287-296.
Kandathil, S.M., Greener, J.G., and Jones, D.T. (2019a). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins.
Kandathil, S.M., Greener, J.G., and Jones, D.T. (2019b). Recent developments in deep learning applied to protein structure prediction. Proteins.
Kellogg, E.H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins-Structure Function and Bioinformatics *79*, 830-838.
Kepp, K.P. (2015). Towards a "Golden Standard" for computing globin stability: Stability and structure sensitivity of myoglobin mutants. Biochimica et biophysica acta *1854*, 1239-1248.
Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization.
Kumar, M.D.S., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Research *34*, D204-D206.
Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W.*, et al.* (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062-D1067.
Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W.*, et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol *487*, 545-574.
LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436-444.
Li, B., Fooksa, M., Heinze, S., and Meiler, J. (2018). Finding the needle in the haystack: towards solving the protein-folding problem computationally. Crit Rev Biochem Mol Biol *53*, 1-28.
Masso, M., and Vaisman, I.I. (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics *24*, 2002-2009.
McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122.
Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., and Fariselli, P. (2019). DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC bioinformatics *20*.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. Journal of Computational Chemistry *30*, 2785-2791.
Ordway, G.A., and Garry, D.J. (2004). Myoglobin: an essential hemoprotein in striated muscle. J Exp Biol *207*, 3441-3446.
Parthiban, V., Gromiha, M.M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Research *34*, W239-W242.
Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics *30*, 335-342.
Pucci, F., Bernaerts, K.V., Kwasigroch, J.M., and Rooman, M. (2018). Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics *34*, 3659-3665.
Quan, L.J., Lv, Q., and Zhang, Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics *32*, 2936-2946.
Rose, P.W., Prlic, A., Altunkaya, A., Bi, C.X., Bradley, A.R., Christie, C.H., Di Costanzo, L., Duarte, J.M., Dutta, S., Feng, Z.K.*, et al.* (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research *45*, D271-D281.
Roushar, F.J., Gruenhagen, T.C., Penn, W.D., Li, B., Meiler, J., Jastrzebska, B., and Schlebach, J.P. (2019). Contribution of Cotranslational Folding Defects to Membrane Protein Homeostasis. J Am Chem Soc *141*, 204-215.
Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., and Bonvin, A. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins *86 Suppl 1*, 51-66.
Shrestha, R., Fajardo, E., Gil, N., Fidelis, K., Kryshtafovych, A., Monastyrskyy, B., and Fiser, A. (2019). Assessing the accuracy of contact predictions in CASP13. Proteins.
Somody, J.C., MacKinnon, S.S., and Windemuth, A. (2017). Structural coverage of the proteome for pharmaceutical applications. Drug Discov Today *22*, 1792-1799.
Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem Sci *44*, 575-588.
Thiltgen, G., and Goldstein, R.A. (2012). Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. Plos One *7*.
Torng, W., and Altman, R.B. (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC bioinformatics *18*, 302.
Torng, W., and Altman, R.B. (2019). High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics *35*, 1503-1512.
Tyka, M.D., Keedy, D.A., Andre, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol *405*, 607-618.
Usmanova, D.R., Bogatyreva, N.S., Bernad, J.A., Eremina, A.A., Gorshkova, A.A., Kanevskiy, G.M., Lonishin, L.R., Meister, A.V., Yakupova, A.G., Kondrashov, F.A.*, et al.* (2018). Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics *34*, 3653-3658.
Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery (arXiv).
Wang, S., Sun, S.Q., and Xu, J.B. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins-Structure Function and Bioinformatics *86*, 67-77.
Wang, Z., and Moult, J. (2001). SNPs, protein structure, and disease. Hum Mutat *17*, 263-270.
Worth, C.L., Preissner, R., and Blundell, T.L. (2011). SDM-a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research *39*, W215-W222.
Xu, J. (2019). Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A.

Xu, J., and Wang, S. (2019). Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins.
Yin, S.Y., Ding, F., and Dokholyan, N.V. (2007). Eris: an automated estimator of protein stability. Nature Methods *4*, 466-467.
Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol *353*, 459-473.

Table 1. Chemical property channels (AutoDock4 atom types) of a protein structure voxel

Property	**Rule**
Hydrophobic	Aliphatic or aromatic carbon atoms
Aromatic	Aromatic carbon atoms
Hydrogen bond donor	Nitrogen, oxygen, sulfur atoms with lone-pair electrons
Hydrogen bond acceptor	Polar hydrogen atoms
Positive ionizable	Atoms with positive charge
Negative ionizable	Atoms with negative charge
Occupancy	All atom types

Table 2. Comparative analysis using the balanced blind test set S^{sym}.

Method	σ_{dir}	r_{dir}	σ_{rev}	r_{rev}	$r_{dir-rev}$	$\langle\delta\rangle$
ThermoNet	1.42	0.58	**1.38**	**0.59**	**-0.95**	-0.05
$PoPMuSiC^{sym}$	1.58	0.48	1.62	0.48	-0.77	**0.03**
MAESTRO	1.36	0.52	2.09	0.32	-0.34	-0.58
FoldX	1.56	0.63	2.13	0.39	-0.38	-0.47
PoPMuSiC 2.1	1.21	0.63	2.18	0.25	-0.29	-0.71
SDM	1.74	0.51	2.28	0.32	-0.75	-0.32
iSTABLE	1.10	0.72	2.28	-0.08	-0.05	-0.60
I-Mutant 3.0	1.23	0.62	2.32	-0.04	0.02	-0.68
NeEMO	1.08	0.72	2.35	0.02	0.09	-0.60
DUET	1.20	0.63	2.38	0.13	-0.21	-0.84
mCSM	1.23	0.61	2.43	0.14	-0.26	-0.91
MUPRO	**0.94**	**0.79**	2.51	0.07	-0.02	-0.97
STRUM	1.05	0.75	2.51	-0.15	0.34	-0.87
Rosetta	2.31	0.69	2.61	0.43	-0.41	-0.69
AUTOMUTE	1.07	0.73	2.61	-0.01	-0.06	-0.99
CUPSAT	1.71	0.39	2.88	0.05	-0.54	-0.72

σ_{dir} and r_{dir} are the root mean square deviation and the Pearson correlation coefficient between the predicted and experimental $\Delta\Delta G$ values for the direct mutations in S^{sym}, from reference state to mutant. These mutations belong to the training data set of the machine-learning-based methods tested (Pucci et al., 2018), so their predictions are likely to be overfitted and σ_{dir} and r_{dir} to be underestimated and overestimated, respectively.

σ_{inv} and r_{inv} are the root mean square deviation and the Pearson correlation coefficient between the predicted and experimental $\Delta\Delta G_0$ values for the reverse mutations in S^{sym}, from mutant to reference state. These mutations do not belong to the training data sets and thus constitute an independent test set.

The parameter δ quantifies the prediction bias and is defined as: $\delta = \Delta\Delta G_{rev} + \Delta\Delta G_{dir}$. A perfectly non-biased tool should have $\delta = 0$ for every mutation. We used here its average value $\langle\delta\rangle$ taken over all mutations that belong to S^{sym}.

Numbers for methods other than ThermoNet were obtained from (Pucci et al., 2018).

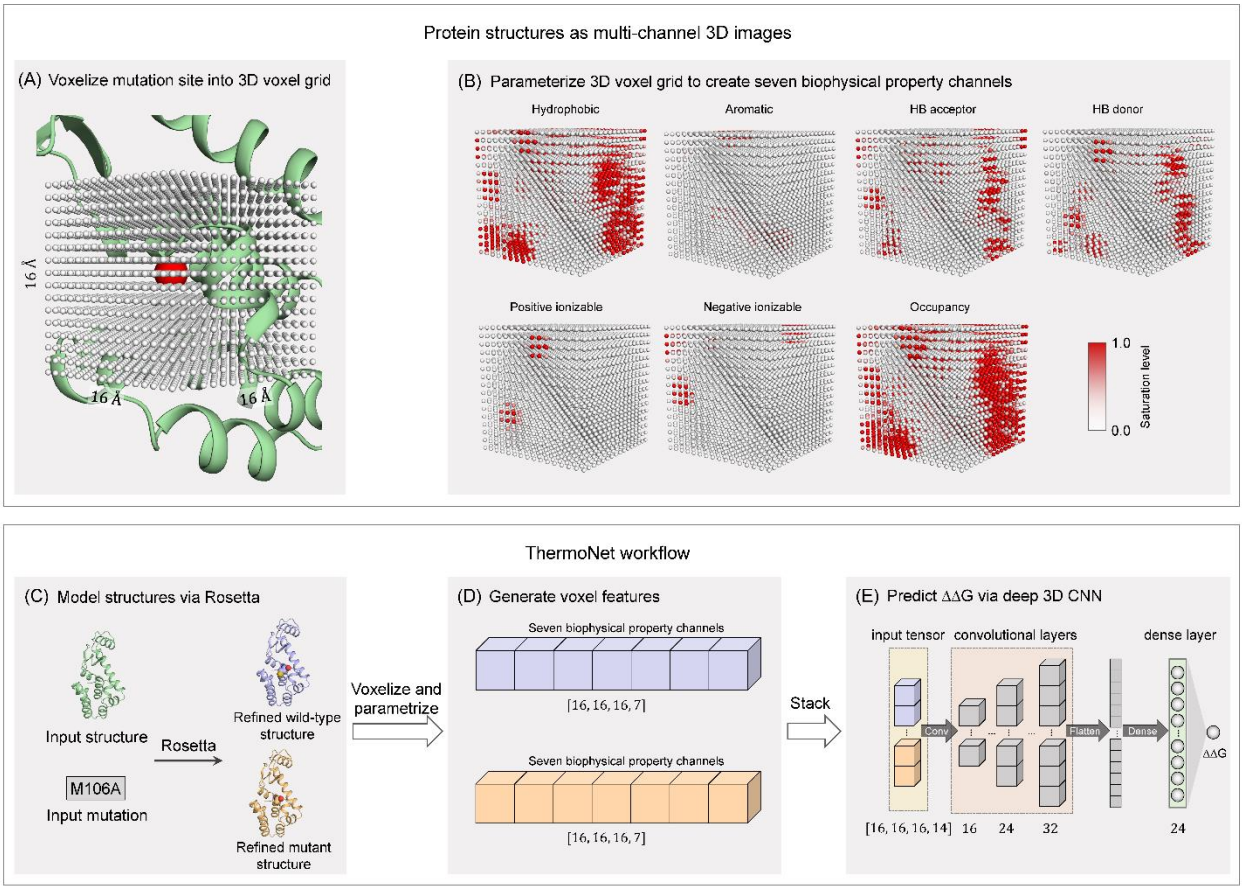

Figure 1. An overview of the ThermoNet computational framework. (A) Protein structures are treated as if they were 3D images. A $16\ \text{Å} \times 16\ \text{Å} \times 16\ \text{Å}$ cubic neighborhood centered at the C_β atom (red sphere) of the mutated residue (or C_α atom in the case of a glycine) of an example protein (PDB ID: 1L63) is discretized into a 3D voxel grid at a resolution of 1 Å. Each voxel is represented by a gray dot. (B) Just as an RGB image has three color channels, the 3D voxel grid is parameterized with seven chemical property channels: hydrophobic, aromatic, hydrogen bonding donor, hydrogen bond acceptor, positive ionizable, negative ionizable, and occupancy. The saturation level of each voxel ranges from 0.0 to 1.0 and is colored accordingly (Methods). (C) To predict the change in thermostability caused by a given single-point mutation, ThermoNet calls Rosetta to refine the wild-type structure and to create a structural model of the mutant protein. (D) ThermoNet voxelizes the space around the mutation site of both the Rosetta-refined wild-type structure and the corresponding mutant structural model. Both the 3D voxel grid of the wild-type structure and that of the mutant model are parameterized accordingly to create two $[16, 16, 16, 7]$ feature maps. (E) The feature maps are then stacked to create a $[16, 16, 16, 14]$ tensor as an input to the trained deep 3D convolutional neural network. The final output of the network is the predicted $\Delta\Delta G$ the given mutation causes to the wild-type protein structure.

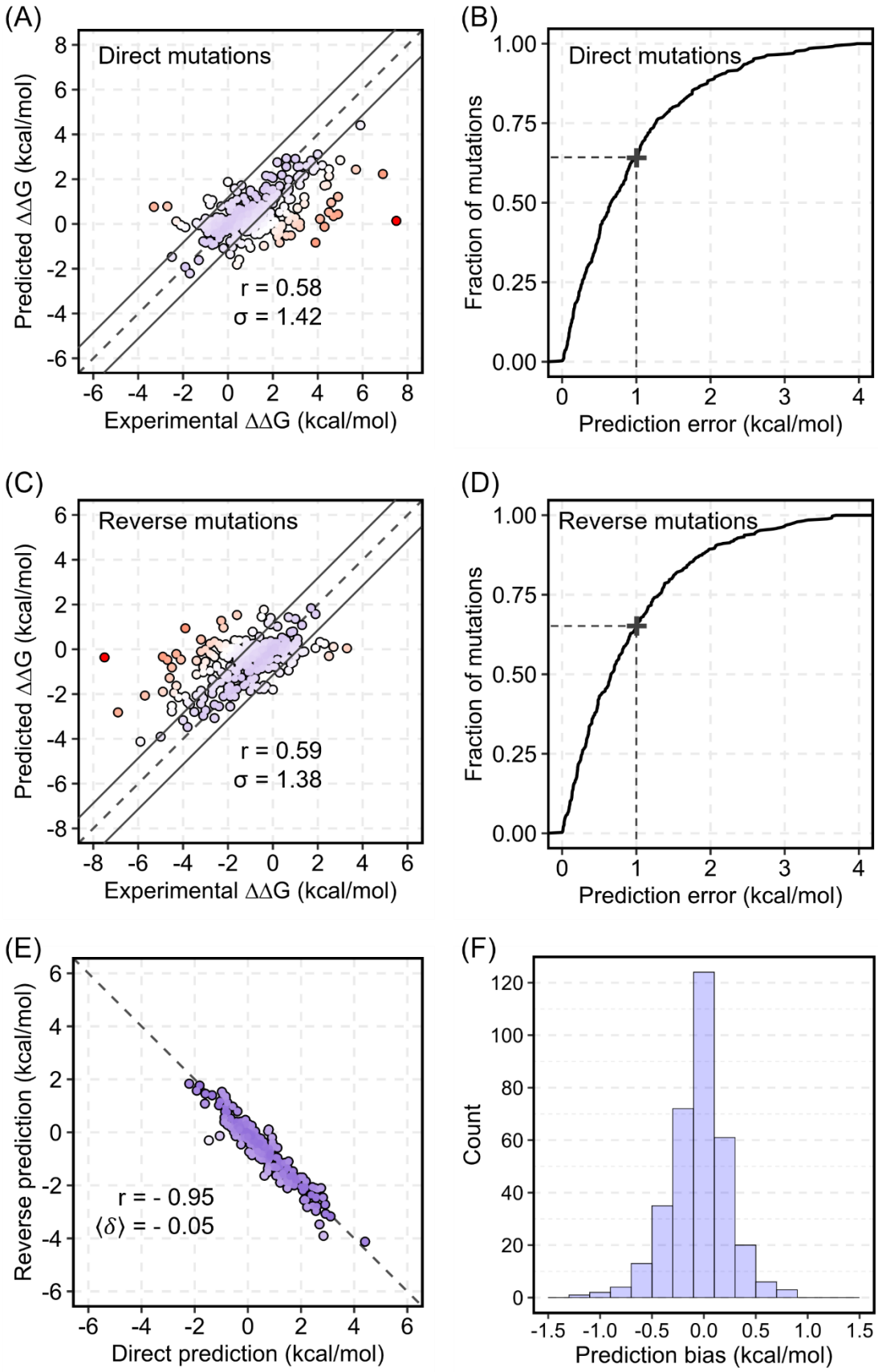
(A)
Direct mutations
r = 0.58
σ = 1.42
Predicted ΔΔG (kcal/mol)
Experimental ΔΔG (kcal/mol)
(B)
Direct mutations
Fraction of mutations
Prediction error (kcal/mol)
(C)
Reverse mutations
r = 0.59
σ = 1.38
Predicted ΔΔG (kcal/mol)
Experimental ΔΔG (kcal/mol)
(D)
Reverse mutations
Fraction of mutations
Prediction error (kcal/mol)
(E)
r = - 0.95
⟨δ⟩ = - 0.05
Reverse prediction (kcal/mol)
Direct prediction (kcal/mol)
(F)
Count
Prediction bias (kcal/mol)

Figure 2. Performance of ThermoNet on the blind test set. (A) Performance of ThermoNet on predicting $\Delta\Delta$G for direct mutations; The Pearson correlation coefficient (r) between predicted values and experimentally determined values is 0.58, and the root-mean-square deviation (σ) of predicted values from experimentally determined values is 1.42 kcal/mol. The dots are colored in gradient from blue to red such that blue represents the most accurate prediction and red indicates the least accurate prediction. (B) Cumulative distribution of ThermoNet prediction error on direct mutations. (C) Performance of ThermoNet on predicting $\Delta\Delta$G for the reverse mutations (r = 0.59, $\sigma =$ 1.38 kcal/mol). (D) Cumulative distribution of ThermoNet prediction error on reverse mutations. The crosses on the cumulative distribution curves in (C) and (D) correspond to the fraction of mutations that fall within the region bounded by the two solid diagonal lines in (A) and (C) respectively, i.e. the fractions (62.6% and 64.3%) of mutations for which prediction error is less than 1.0 kcal/mol. (E) Direct versus reverse $\Delta\Delta$G values of all the mutations in the blind test set predicted by ThermoNet. A perfectly unbiased predictor would give $r = -1$ and $\langle\delta\rangle = 0\ kcal/mol$. ThermoNet is much less biased than existing methods with $r = -0.95$ and $\langle\delta\rangle = -0.05\ kcal/mol$. (F) Distribution of ThermoNet prediction bias.

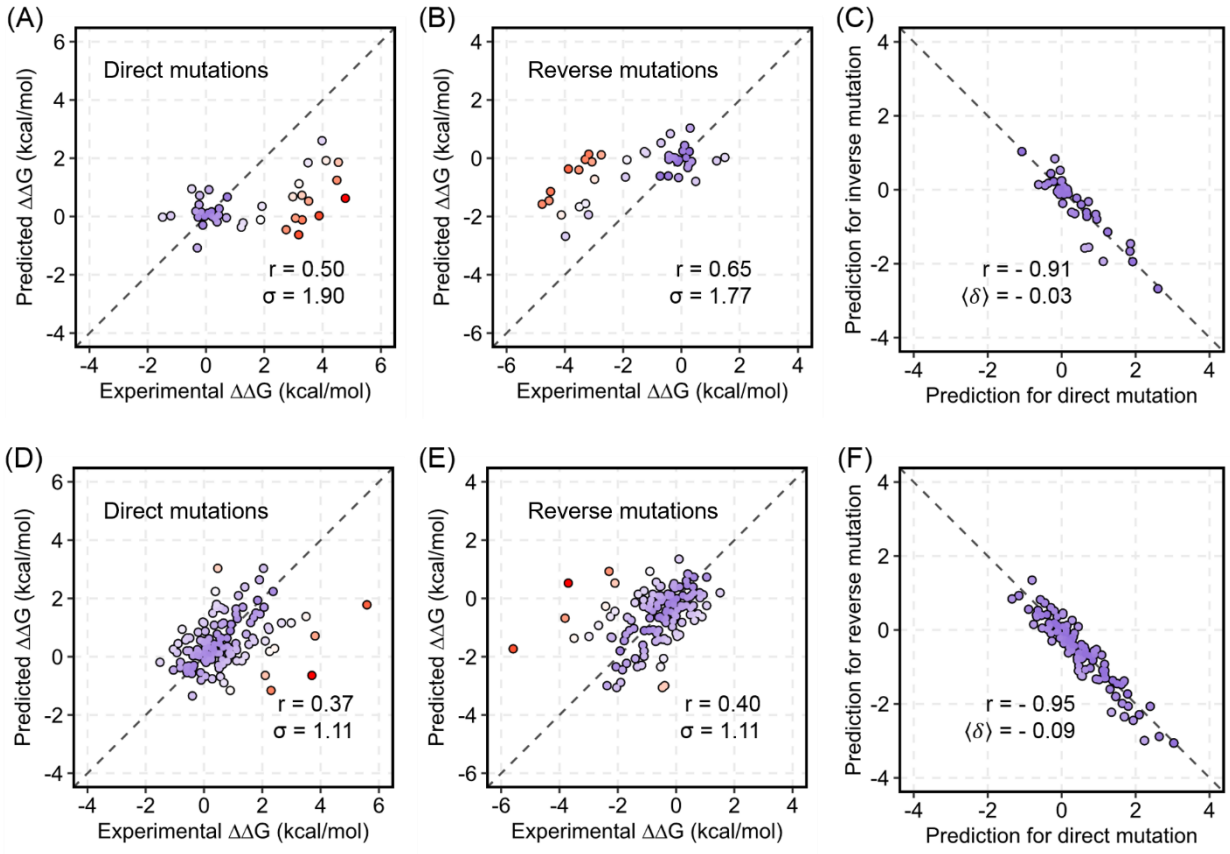

Figure 3. Predicted ThermoNet ΔΔG landscapes on p53 tumor suppressor protein and myoglobin. (A) Performance of ThermoNet on predicting ΔΔG for the direct mutations in p53 (r = 0.50, σ = 1.90 kcal/mol). (B) Performance of ThermoNet on predicting ΔΔG for the reverse mutations in p53 (r = 0.65, σ = 1.77 kcal/mol). (C) Direct versus reverse ΔΔG values of all p53 mutations predicted by ThermoNet ($r = -0.91$ and $\langle\delta\rangle = -0.03\ kcal/mol$). (D) Performance of ThermoNet on predicting ΔΔG for the direct mutations in myoglobin (r = 0.37, σ = 1.11 kcal/mol). (E) Performance of ThermoNet on predicting ΔΔG for the reverse mutations in myoglobin (r = 0.40, σ = 1.11 kcal/mol). (F) Direct versus reverse ΔΔG values of all myoglobin mutations predicted by ThermoNet, with a Pearson correlation of $r = -0.95$ and $\langle\delta\rangle = -0.09\ kcal/mol$. The dots are colored in gradient from blue to red such that blue represents the most accurate prediction and red indicates the least accurate prediction.

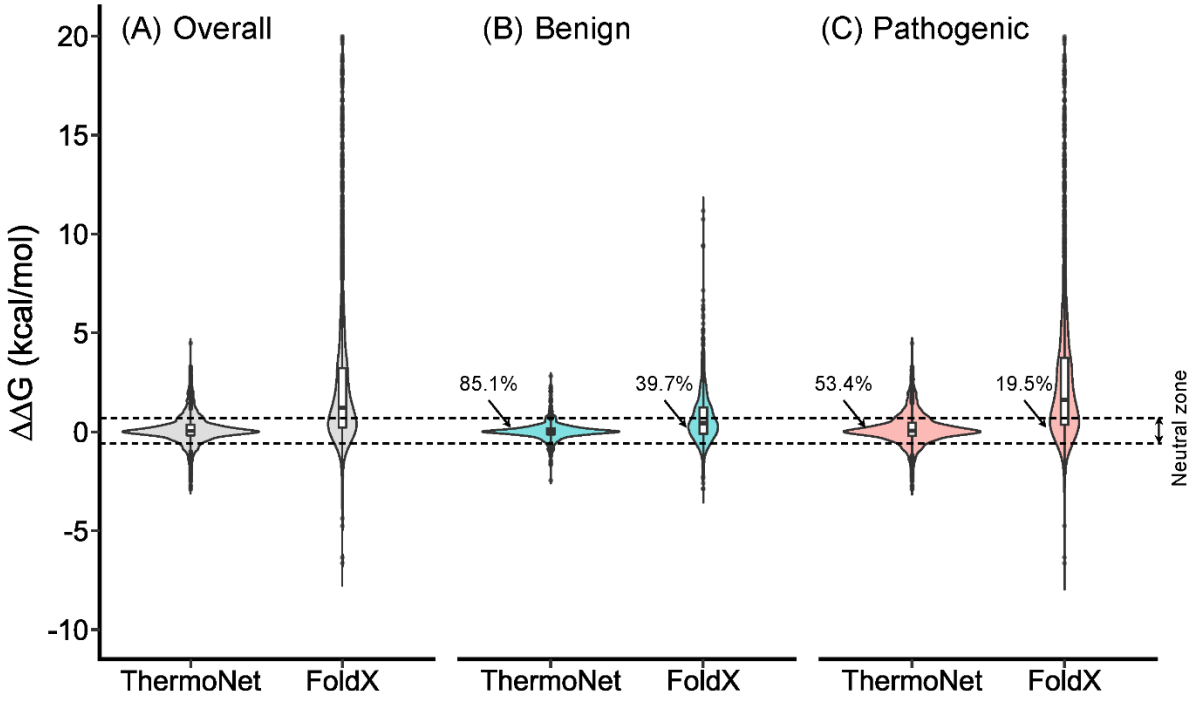

Figure 4. Predicted ΔΔG distributions of ClinVar missense variants. (A) The overall ΔΔG distributions of ClinVar variants predicted by ThermoNet and FoldX. ThermoNet's predictions are consistent with the expected range based on experimentally determined ΔΔG values (-5 kcal/mol to +5 kcal/mol). In contrast, more than 15% of ΔΔGs predicted by FoldX are outside the expected range. (B) The ΔΔG distributions for ClinVar benign variants predicted by ThermoNet and FoldX. (C) The ΔΔG distributions of ClinVar pathogenic variants predicted by ThermoNet and FoldX. The ΔΔGs of 85.1% of benign variants predicted by ThermoNet fall within the neutral zone (-0.05 to +0.05 kcal/mol, region between dashed lines), in which variants are not expected to influence fitness. FoldX only predicted 39.7% of benign variants to be in the neutral zone. Further, the ΔΔGs of pathogenic variants predicted by ThermoNet suggest pathogenic variants are nearly equally likely to be stabilizing (43.4%) as destabilizing (56.6%). In contrast, FoldX predicted that 83.2% of pathogenic variants are destabilizing. Variants for which FoldX ΔΔG is > 20 kcal/mol are omitted for clarity. Percentages represent the fractions of variants whose ΔΔGs are predicted to be in the neutral zone.