PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

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

1 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **2** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **3** Department of Biological Sciences and Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America, **4** Department of Computer Science, Yale University, New Haven, Connecticut, United States of America

* tony.capra@vanderbilt.edu (JAC); mark@gersteinlab.org (MBG)

Citation: Li B, Yang YT, Capra JA, Gerstein MB (2020) Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput Biol 16(11): e1008291. https://doi.org/10.1371/journal.pcbi.1008291

Editor: Piero Fariselli, Universita degli Studi di Torino, ITALY

Received: March 2, 2020

Accepted: August 26, 2020

Published: November 30, 2020

Data Availability Statement: All relevant data are within the manuscript, its Supporting information files and on https://github.com/gersteinlab/ThermoNet.

Funding: MBG and YTY were supported by National Science Foundation award (NSF DBI1660648), JAC was supported by National Institute of Health awards (R35 GM127087 and R01 GM126249), BL was supported by an American Heart Association Postdoctoral Fellowship (20POST35220002). The funders had

Abstract

Predicting mutation-induced changes in protein thermodynamic stability (ΔΔG) is of great interest in protein engineering, variant interpretation, and protein biophysics. We introduce ThermoNet, a deep, 3D-convolutional neural network (3D-CNN) designed for structure-based prediction of ΔΔGs upon point mutation. To leverage the image-processing power inherent in CNNs, we treat protein structures as if they were multi-channel 3D images. In particular, the inputs to ThermoNet are uniformly constructed as multi-channel voxel grids based on biophysical properties derived from raw atom coordinates. We train and evaluate ThermoNet with a curated data set that accounts for protein homology and is balanced with direct and reverse mutations; this provides a framework for addressing biases that have likely influenced many previous ΔΔG prediction methods. ThermoNet demonstrates performance comparable to the best available methods on the widely used S^{sym} test set. In addition, ThermoNet accurately predicts the effects of both stabilizing and destabilizing mutations, while most other methods exhibit a strong bias towards predicting destabilization. We further show that homology between S^{sym} and widely used training sets like S2648 and VariBench has likely led to overestimated performance in previous studies. Finally, we demonstrate the practical utility of ThermoNet in predicting the ΔΔGs for two clinically relevant proteins, p53 and myoglobin, and for pathogenic and benign missense variants from ClinVar. Overall, our results suggest that 3D-CNNs can model the complex, non-linear interactions perturbed by mutations, directly from biophysical properties of atoms.

Author summary

The thermodynamic stability of a protein, usually represented as the Gibbs free energy for the biophysical process of protein folding (ΔG), is a fundamental thermodynamic quantity. Predicting mutation-induced changes in protein thermodynamic stability (ΔΔG) is of

no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

great interest in protein engineering, variant interpretation, and protein biophysics. However, predicting ΔΔGs in an accurate and unbiased manner has been a long-standing challenge in the field of computational biology. In this work, we introduce ThermoNet, a deep, 3D-convolutional neural network (3D-CNNs) designed for structure-based ΔΔG prediction. To leverage the image-processing power inherent in CNNs, we treat protein structures as if they were multi-channel 3D images. ThermoNet demonstrates performance comparable to the best available methods. In addition, ThermoNet accurately predicts the effects of both stabilizing and destabilizing mutations, while most other methods exhibit a strong bias towards predicting destabilization. We also demonstrate that the presence of homologous proteins in commonly used training and testing sets for ΔΔG prediction methods has likely influenced previous performance estimates. Finally, we highlight the practical utility of ThermoNet by applying it to predicting the ΔΔGs for two clinically relevant proteins, p53 and myoglobin, and for pathogenic and benign missense variants from ClinVar.

This is a *PLOS Computational Biology* Methods paper.

Introduction

The thermodynamic stability of a protein, usually represented as the Gibbs free energy for the biophysical process of protein folding (ΔG), is a fundamental thermodynamic 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 biophysical environment surrounding it [1]. 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. (Note that the term "wild type" is not preferred because humans have substantial protein-coding genetic diversity [2]. A better term would be a "reference state". However, as it will not cause confusion in this work and for consistency with previous studies, we use "wild type" throughout the text.) This change in protein thermodynamic stability (i.e. ΔΔG) caused 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 [3–5]. Given the huge number of variants discovered by 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 thermodynamic stability. However, this task is often laborious, if not impossible, and usually involves experimentally screening an enormous number of mutant proteins [6]. 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 [7–11] or apply machine/statistical learning to extract patterns from various types of amino acid sequence, evolutionary, and/or protein structural features [12–20]. While these methods have been useful in many applications [21,22], 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. On the other hand, 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 [19,23–25]. The main source of this bias likely comes from the fact that the training sets are dominated by experiment-derived destabilizing mutations and that machine learning methods are prone to overfitting to training sets [24,26]. 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-CNNs for predicting ΔΔGs upon single-point mutation. We model the structure of each mutation assuming that single-point mutations introduce negligible perturbation to the overall architecture of protein structure. We treat protein structures as if they were 3D images with voxels parameterized using atom biophysical properties [27,28]. We leverage the power of the 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 address the bias in many previous methods towards predicting destabilization by training ThermoNet on a balanced data set generated through anti-symmetry-based data augmentation, i.e. for each mutation, we consider both the direct and reverse versions. We further demonstrate and address an unappreciated source of bias in previous performance estimates due to homology between training and evaluation sets. We show that ThermoNet achieves state-of-the-art performance comparable to previously developed methods on a widely used 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 missense 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 [29] and the successes of deep CNNs in biophysical problems [27,28,30–32]. We treat protein structures as if they were 3D images where voxels are parameterized using atom biophysical properties [27,28] (Fig 1A and 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) [9,33]. ThermoNet assumes that the ΔΔG of a point mutation can be sufficiently captured by modeling the 3D biophysical 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 biophysical 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 ensemble of ten deep 3D-CNNs, which generate a prediction of the ΔΔG the given mutation causes to the wild-type structure (Fig 1D and 1E, and Methods). Each of the component 3D-CNN models consists of three 3D convolutional layers with 16, 24, and 32 neurons respectively and one densely connected layer of 24 neurons (Fig 1E). These

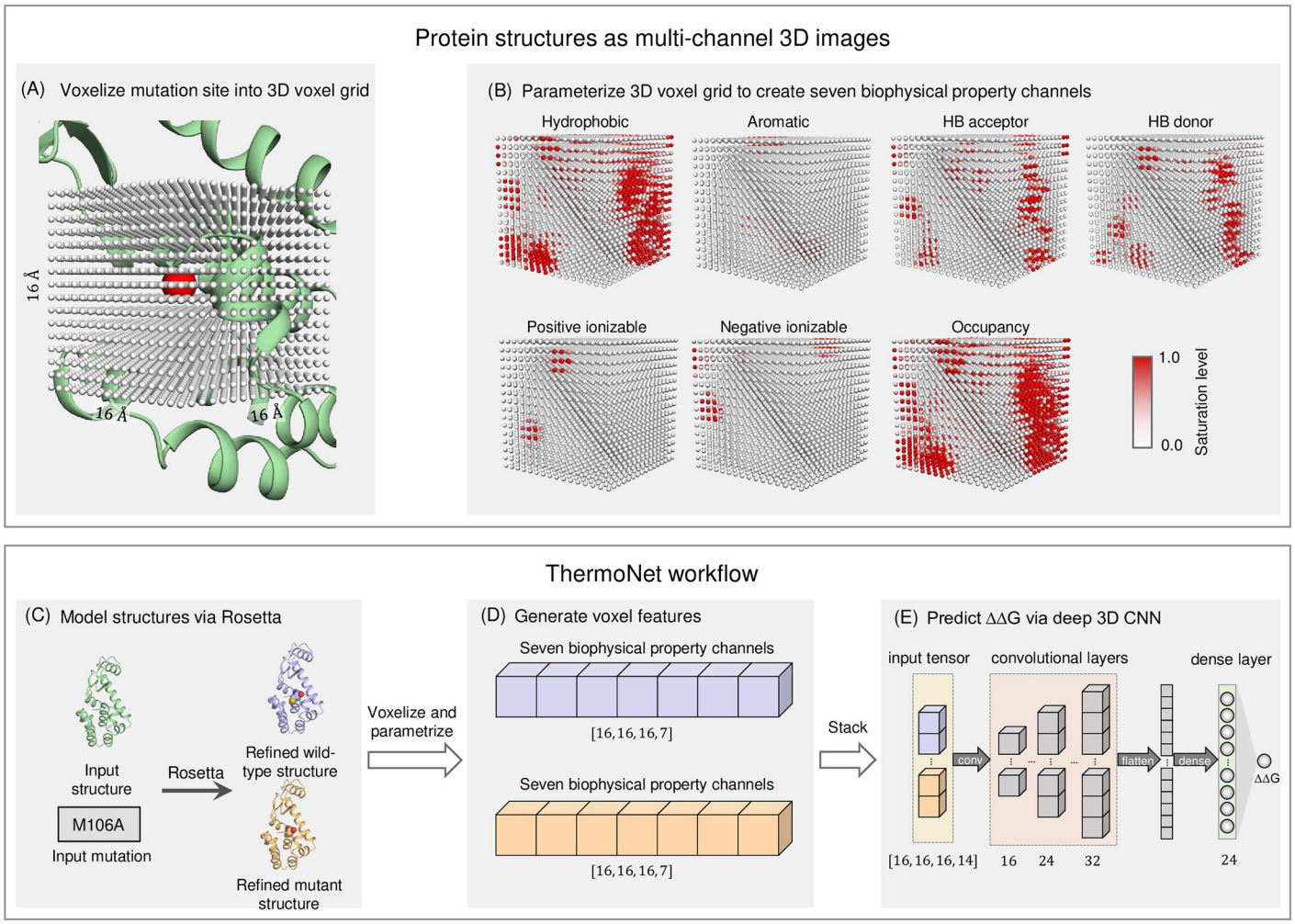

Fig 1. An overview of the ThermoNet computational framework. (A) Protein structures are treated as if they were 3D images. A 16 Å × 16 Å × 16 Å cubic neighborhood centered at the C_β atom (red sphere) of the mutated residue (or C_α atom in the case of a glycine) of an example protein (PDB ID: 1L63) is 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 biophysical 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 thermodynamic stability 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 ΔΔG the given mutation causes to the wild-type protein structure.

https://doi.org/10.1371/journal.pcbi.1008291.g001

Table 1. Biophysical property channels for protein structure voxels.

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

https://doi.org/10.1371/journal.pcbi.1008291.t001

architectural hyperparameters (i.e. number of neurons in convolutional and densely connected layers and sizes of the input voxel grid) were tuned via five-fold cross-validation (Methods, S1 Fig).

Creating data sets for robust training and testing of ThermoNet

The ability of a machine-learning model to generalize can be overestimated when there is data leakage between the training set and the test set. In structural bioinformatics problems such as ΔΔG prediction, such data leakage can result when the training set contains proteins that are homologous to proteins in the test set. This is because the effects of different mutations in the same protein or homologous proteins are not necessarily independent. In most previous methods for ΔΔG prediction, this data leakage issue was not fully appreciated, and homologous proteins were present between training and test sets. For example, a recent method used randomly selected subsets of mutants from the widely used S2648 data set for training and testing [34]. Not surprisingly, the training and test sets of shared 61 identical proteins (S1 Table).

The issue of having mutations from the same protein in both training and test sets was addressed in developing the mCSM and INPS methods, where the cross-validation procedure ensured that mutations of the same protein remained together in either the training or test set [16,19]. While this was a step in the right direction, grouping mutations at the protein level is not sufficient to remove the homology between training and test sets. For example, we found substantial homology between 132 proteins within S2648 (S2A Fig). Thus, splitting the S2648 data set for training and testing at the protein level is likely to end up with shared homology between the splits. In fact, using the PISCES server [35] to remove redundancy from the S2648 data set resulted in only 104 non-redundant proteins (out of 132) at the level of $< 25\%$ sequence identity (S2 Table). Homology is common among data sets used in training ΔΔG predictors; for example, many proteins in the VariBench [36] data set also share substantial homology (S2B Fig).

We further highlight that the widely used S2648 data set includes fourteen proteins from S^{sym} data set, which was previously used to evaluate a wide range of ΔΔG predictors [24,37], and another eight proteins that are putative homologs to proteins in S^{sym} (Fig 2A, S3 Table). In addition, a similar level of overlap also exists between the VariBench and Q3421 data sets and S^{sym} (Fig 2A, S4 and S5 Tables). Thus, performance estimates on S^{sym} of methods trained using S2648 or parameterized using VariBench are likely to be overly optimistic. For example, in a recent publication, a version of INPS trained using a data set obtained by removing all proteins with $> 25\%$ sequence identity to proteins in S^{sym} showed substantially reduced performance [38].

Thus, to train and evaluate ThermoNet, we implemented a rigorous procedure to reduce the sequence similarity between the training and test sets (S^{sym}, p53, myoglobin) by removing duplicate data points and pruning protein-level homology (Fig 2B). Our rigorous pruning of the starting Q3421 data set [15] resulted in a data set consisting of 1,744 distinct mutations. This data set was then augmented by creating a reverse mutation data point for each of the 1,744 direct mutations according to the anti-symmetry property of ΔΔG (Methods), thus giving to a total of 3,488 data points for the training of ThermoNet (Fig 2B). While the pruning nearly halved the size of available training data, as discussed in the following section, models trained using the resulting augmented Q3488 data set will be less likely to have an overestimated performance when evaluated on S^{sym}. Finally, to explore the influence of homology between training and test sets on estimates of model performance, we also augmented Q3214, the intermediate data set before the step of homology reduction, to train a different version of ThermoNet, called ThermoNet* (Fig 2B).

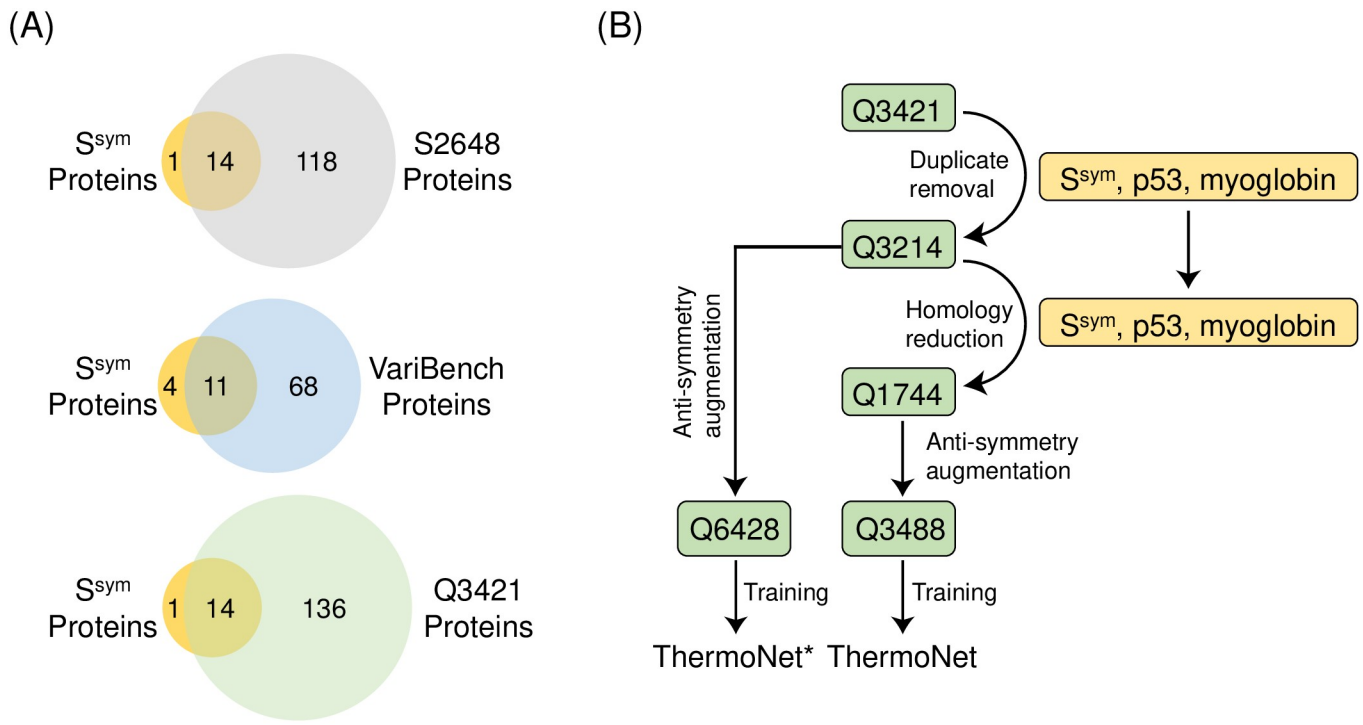

Fig 2. Data set curation and identification of shared homology. (A) Venn diagrams showing the amount of overlap at the protein level between three widely used training sets S2648, VariBench, and Q3421 for ΔΔG predictors and the S^{sym} test set. Numbers in these diagrams indicate protein counts. Upper panel and lower panel indicate that both S2648 and Q3421 share 14 identical proteins with S^{sym}; middle panel indicates that VariBench and S^{sym} share 11 identical proteins. All three data sets share additional homology with S^{sym}, which is presented in S3, S4, and S5 Tables, respectively. (B) Creating data sets for robust training and testing of ThermoNet. We started with the Q3421 set of 3421 mutations from 150 proteins. (Numbers in data set names indicate the number of unique mutations the data set contains.) After homology reduction and anti-symmetry data augmentation (Methods), this data curation workflow gives a training set of 3488 mutations with an equal representation of stabilizing and destabilizing changes and reduced homology to the S^{sym} test set. A separate data set called Q6428 was also created by augmenting the Q3214 data set before homology reduction to train ThermoNet*.

https://doi.org/10.1371/journal.pcbi.1008291.g002

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

We systematically compared ThermoNet and ThermoNet* with seventeen ΔΔG predictors on the S^{sym} balanced data set to evaluate their performance and degree of bias with respect to the ΔΔG anti-symmetry between direct and reverse mutations (Methods). A brief summary of the characteristics of these ΔΔG predictors and their references are given in S6 Table. In short, the predictors are based on diverse features and strategies with some, like ThermoNet based only on structural information, while others like DDGun3D [37] and STRUM [15], integrate structural information with sequence and evolutionary features. The S^{sym} data set, which was constructed previously for assessing the biases of ΔΔG predictors [24], consists of experimentally measured ΔΔG values for 342 direct and the corresponding reverse mutations (a total of 684 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. This data set is by construction balanced with respect to stabilizing and destabilizing mutations, thus enabling the evaluation of prediction bias. However, as noted in the previous section, many proteins in S^{sym} overlap or are homologous to proteins in commonly used training sets (Fig 2A).

To evaluate performance, we computed the root-mean-square error σ and the Pearson correlation coefficient r separately for direct and reverse mutations. We measured prediction bias by two statistics, the Pearson correlation coefficient $r_{dir-rev}$ between the predictions for

Table 2. Comparative analysis using the balanced 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
DDGun3D	1.42	0.56	1.46	0.53	-0.99	-0.02
DDGun	1.47	0.48	1.50	0.48	-0.99	-0.01
ThermoNet	1.56	0.47	1.55	0.47	-0.96	-0.01
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 ΔΔG values for the direct mutations in S^{sym}. Many of these mutations belong to the training set of the machine-learning-based methods tested [24], so their performances are likely to be overestimated. σ_{rev} and r_{rev} are the root mean square deviation and the Pearson correlation coefficient between the predicted and experimental ΔΔG values for the reverse mutations in S^{sym}. These reverse 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 DDGun and DDGun3D were obtained from reference [37] and those for all other methods were obtained from [24]. The methods are ranked according to their performance, σ_{rev}, on reverse mutations. Both ThermoNet* and ThermoNet were trained using a data set balanced with direct and reverse mutations, but the data set for training ThermoNet* was not homology-reduced with respect to S^{sym} (Fig 2B).

https://doi.org/10.1371/journal.pcbi.1008291.t002

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 $r_{dir-rev} = -1$ and $\langle\delta\rangle = 0$ *kcal/mol*, where $\langle\delta\rangle$ is the average of δ.

ThermoNet achieves strong prediction accuracy that is comparable for direct mutations ($r_{dir} = 0.47$ and $\sigma_{dir} = 1.56$ *kcal/mol*; Table 2, Fig 3A) and the corresponding set of reverse mutations ($r_{rev} = 0.47$ and $\sigma_{rev} = 1.55$ *kcal/mol*, Fig 3C). This suggests that ThermoNet did not overfit to direct mutations. The fractions of mutations for which the prediction error is within 0.5 *kcal/mol* and 1.0 *kcal/mol* are 36.3% and 58.8% for direct mutations and 36.5% and 58.5% for reverse mutations (Fig 3B and 3D). ThermoNet successfully reduces prediction bias with a near-perfect $r_{dir-rev}$ (-0.96) and a negligible $\langle\delta\rangle$ (−0.01 *kcal/mol*) (Fig 3E). 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 3F, 40.9% and 96.2% of mutations have a prediction bias < 0.1 *kcal/mol* and < 0.5 *kcal/mol*, respectively.

We also report the performance of ThermoNet*, a different version of ThermoNet trained using the Q6428 data set augmented from the intermediate data set Q3214 before the step of homology reduction in our data set curation procedure (Fig 2B and Methods). ThermoNet* was trained in the exact same way as ThermoNet except that the homology of its training set Q3214 to S^{sym} was retained. Thus, the parameterization of ThermoNet* is comparable to

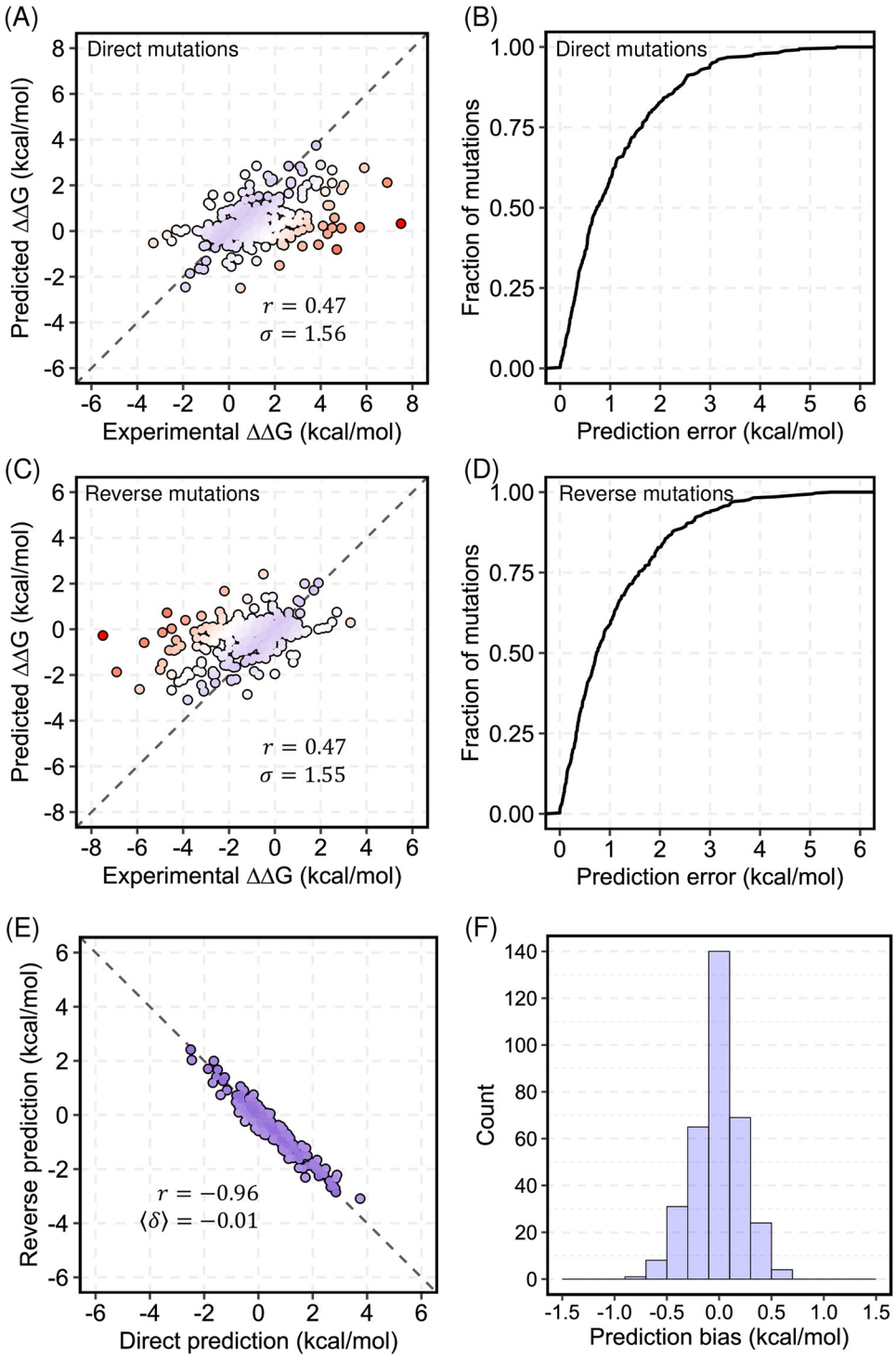

Fig 3. Performance of ThermoNet on the blind test set. (A) Performance of ThermoNet on predicting ΔΔG for direct mutations; The Pearson correlation coefficient (r) between predicted values and experimentally determined values is 0.47, and the root-mean-square error (σ) of predicted values from experimentally determined values is 1.56 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 ΔΔG for the reverse mutations (r = 0.47, σ = 1.55 kcal/mol). (D) Cumulative distribution of ThermoNet prediction error on reverse mutations. (E) Direct versus reverse ΔΔG 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 successfully reduces prediction bias with $r = -0.96$ and $\langle\delta\rangle = -0.01$ *kcal/mol*. (F) Distribution of ThermoNet prediction bias.

https://doi.org/10.1371/journal.pcbi.1008291.g003

previous methods that did not consider homology. As expected, evaluation of ThermoNet* on S^{sym} shows even better performance in σ_{rev} (1.38 vs. 1.55 kcal/mol) and r_{rev} (0.59 vs. 0.48) than ThermoNet, which was trained using the data set obtained after homology reduction (Table 2, Fig 2B, and Methods). Thus, the performance of many previously developed methods is likely to be substantially lower if they had been trained using a data set that shared no homology with S^{sym}. In contrast, ThermoNet's strong performance, even after removing homology reduction, suggests robust generalization in real-life applications.

Compared to other ΔΔG predictors, ThermoNet* achieves the best performance on reverse mutations, and the methods that outperform it on direct mutations all have substantial bias against reverse mutations ($\sigma_{rev} > 2.09$ kcal/mol and $\langle\delta\rangle <$ -0.58). The seemingly good performance of many machine learning-based methods on direct mutations, but poor performance on reverse mutations suggests potential overfitting due to unbalanced training sets [24–26,39]. ThermoNet also performs well, but as a result of the reduction in performance due to removing homology between training and test sets, the DDGun and DDGun3D methods outperform it on direct and reverse mutations. Unfortunately, it is not possible to retrain and evaluate all the other methods on the homology pruned training set, so we cannot directly compare the other methods to ThermoNet. Nonetheless, the fact that it still outperforms most suggests its utility and robustness.

Structural models of reverse mutations are necessary for unbiased ΔΔG predictions

To evaluate whether the inclusion of the reverse mutations is necessary for the reduction in prediction bias, we trained a predictor following the same procedure for training ThermoNet but using a data set consisting of only the 1,744 direct mutations and their associated experimental ΔΔGs (i.e. the Q1744 data set in Fig 2B). We applied this predictor to predict the ΔΔGs of the direct and reverse mutations of the S^{sym} test set. As shown in S3 Fig, ΔΔGs of direct mutations predicted by these models correlate reasonably well ($r = 0.47$ and $\sigma = 1.38$ *kcal/mol*) with the experimental values and are comparable to the performance of the ensemble of networks trained using the balanced data set Q3488. In contrast, these models perform poorly ($r = -0.06$ and $\sigma = 2.40$ *kcal/mol*) in predicting the ΔΔGs of the corresponding set of reverse mutations (S3 Fig). 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 ($r_{dir-rev} = 0.35$ and $\langle\delta\rangle = 1.63$ *kcal/mol*) (S3 Fig). Compared to the performance of ThermoNet, which was trained using the balanced data set Q3488, 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. 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 [40]. Myoglobin

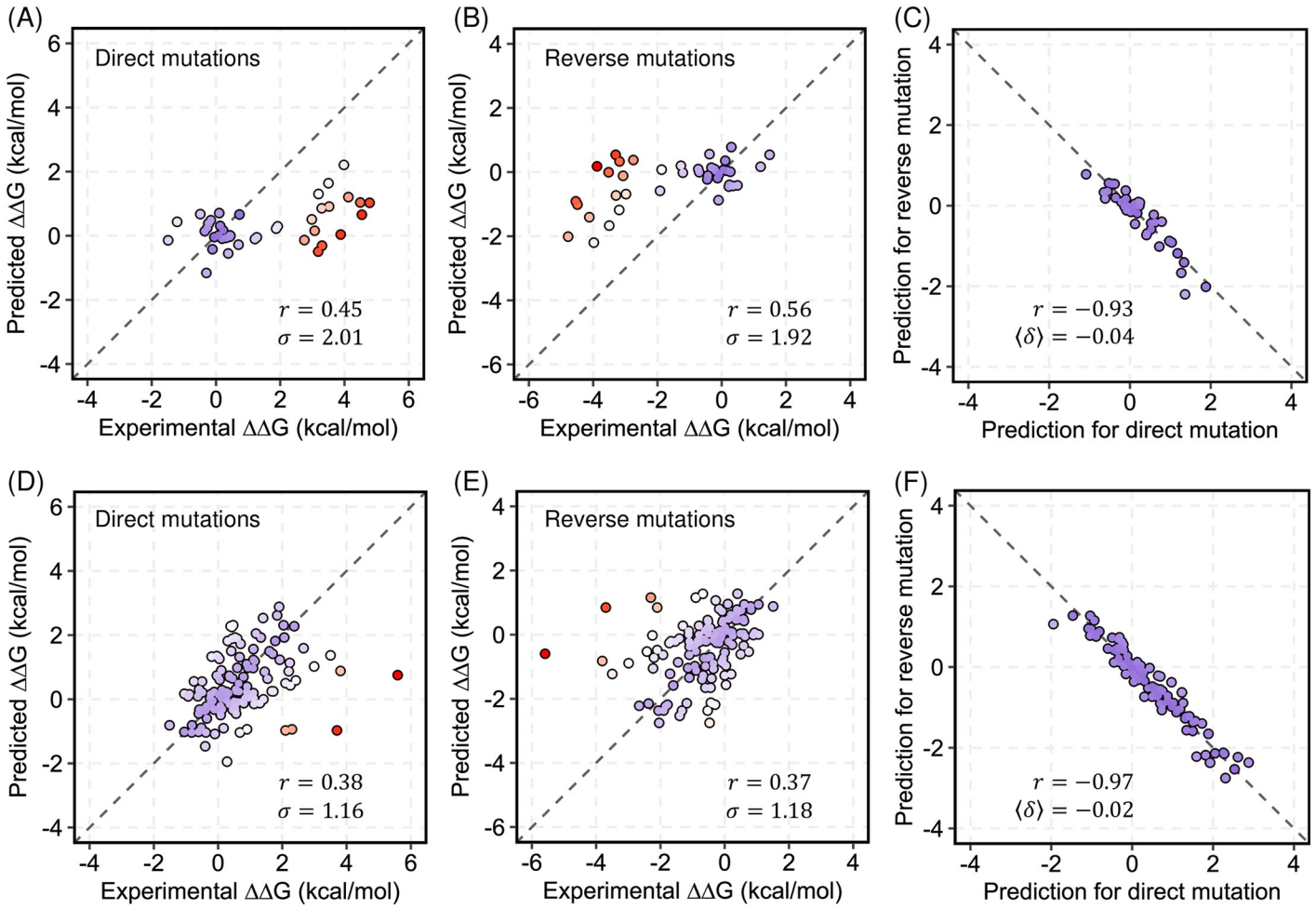

Fig 4. ThermoNet predicted well the ΔΔGs of mutations in the p53 tumor suppressor protein and myoglobin. (A) Performance of ThermoNet on predicting ΔΔG for the direct mutations in p53 ($r = 0.45$, $\sigma = 2.01$ kcal/mol). (B) Performance of ThermoNet on predicting ΔΔG for the reverse mutations in p53 ($r = 0.56$, $\sigma = 1.92$ kcal/mol). (C) Direct versus reverse ΔΔG values of all p53 mutations predicted by ThermoNet ($r_{dir-rev} = -0.93$ and $\langle\delta\rangle = -0.04$ *kcal/mol*). (D) Performance of ThermoNet on predicting ΔΔG for the direct mutations in myoglobin ($r = 0.38$, $\sigma = 1.16$ kcal/mol). (E) Performance of ThermoNet on predicting ΔΔG for the reverse mutations in myoglobin ($r = 0.37$, $\sigma = 1.18$ kcal/mol). (F) Direct versus reverse ΔΔG values of all myoglobin mutations predicted by ThermoNet, with a Pearson correlation of $r_{dir-rev} = -0.97$ and $\langle\delta\rangle = -0.02$ *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.

https://doi.org/10.1371/journal.pcbi.1008291.g004

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 [41]. The p53 data set consists of 42 mutations within the DNA binding domain of p53 [16]. The myoglobin data set consists of 134 mutations scattered throughout the protein chain [42]. We note that none of the mutations in these two data sets were present in the training set and that proteins that are likely to be homologous to p53 and myoglobin were also removed from the training set of ThermoNet (Fig 2B, Methods). 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 [43]. These predictions were compared directly with the experimentally determined ΔΔGs (Fig 4).

For p53, ΔΔGs of both direct and reverse mutations predicted by ThermoNet correlate well with the experimental measurements ($r = 0.45$ and 0.56) and have little bias ($r_{dir-rev} = -0.93$

and $\langle\delta\rangle = -0.04\ kcal/mol$). However, the predicted ΔΔGs of myoglobin mutations correlate less well ($r = 0.38$ and 0.37 for direct and reverse mutations respectively) compared to those of p53, though the bias is also low ($r_{dir-rev} = -0.97$ and $\langle\delta\rangle = -0.02\ kcal/mol$). The poorer correlations for myoglobin are likely because the myoglobin data set consists of ΔΔG measurements obtained under various experimental conditions which ThermoNet does not explicitly account for. In fact, after excluding four data points (L29N, A130L, and two data points corresponding to A130K) with the biggest prediction error (> 3 kcal / mol), the Pearson correlations increase to 0.52 and 0.51 for direct and reverse mutations respectively. 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.

We also compared ThermoNet with four other biophysics-based methods: FoldX, Rosetta, SDM, and CUPSAT (see S6 Table for a summary and references of these methods). We were not able to include all methods because both p53 and myoglobin are already in the S2648 set that was used for training most of the other machine learning-based ΔΔG predictors and they are also in the VariBench [36] data set used to derive parameters of the DDGun model [37]. We recognize that there is also a possibility that mutation ΔΔG values from p53 and myoglobin were used in the derivation of parameters of the biophysics-based methods tested here. Our comparison indicates that both FoldX and Rosetta predictions have a better correlation than ThermoNet while also reasonably anti-symmetric (S7 and S8 Tables). However, as shown in Table 2 and demonstrated in previous studies [24,25], both FoldX and Rosetta are likely to show bias toward predicting destabilization when tested on larger data sets.

ΔΔG landscape of ClinVar missense variants

Previous work has shown that variant deleteriousness can only be partially attributed to ΔΔG [44] and that both stabilization and destabilization can cause disease [45]. We sought to use ThermoNet to obtain a less biased picture of the impact of benign and pathogenic variants on protein stability. We applied ThermoNet to predict the ΔΔG distributions of pathogenic and benign missense variants in ClinVar, a widely used resource of medically important variants [46]. For comparison, we also applied FoldX, a popular and freely available ΔΔG predictor [13,22] to the ClinVar set. We first examined the overall predicted ΔΔG distribution of ClinVar variants. The ΔΔGs of ClinVar variants predicted by ThermoNet range from -2.75 kcal/mol to +3.75 kca/mol (Fig 5A). Experimentally measured ΔΔGs generally fall within -5 kcal/mol to +5 kcal/mol [13,47]; 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 5A), 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 [47]. 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 [47]. We thus reasoned that the ΔΔGs of benign variants should fall within a narrow range from -0.5 kcal/mol to +0.5 kcal/mol and that pathogenic variants should be equally likely to be destabilizing or stabilizing [47]. To test this hypothesis, we examined the ΔΔG distributions of pathogenic and benign variants separately. The ΔΔGs of 80.2% of benign variants predicted by

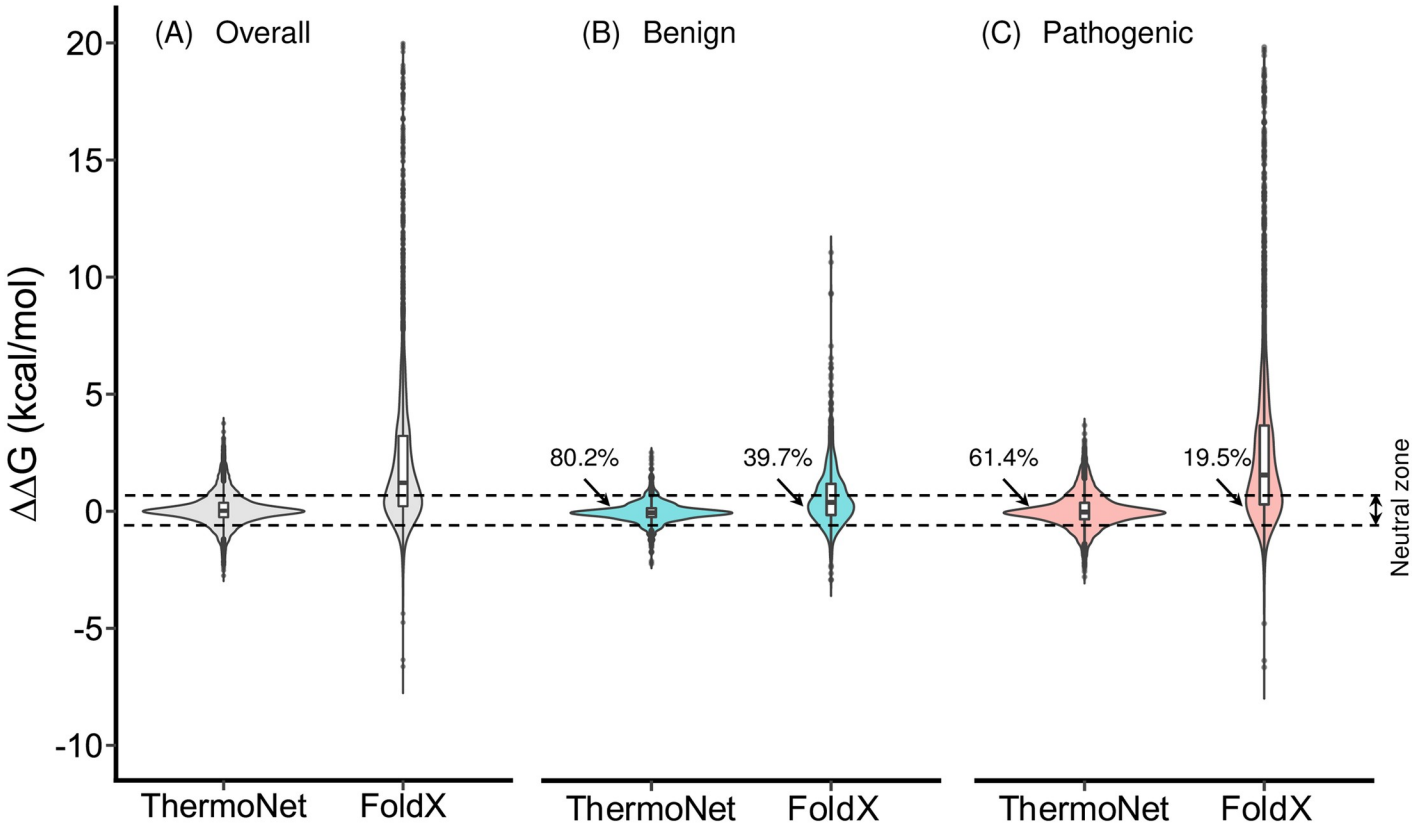

Fig 5. Predicted ΔΔG distributions of ClinVar missense variants. (A) The overall ΔΔG distributions of ClinVar variants predicted by ThermoNet and FoldX. ThermoNet's predictions are consistent with the expected range based on experimentally determined ΔΔG values (-5 kcal/mol to +5 kcal/mol). In contrast, 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 80.2% of benign variants predicted by ThermoNet fall within the neutral zone (-0.5 to +0.5 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 (47.3%) as destabilizing (52.7%). 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.

https://doi.org/10.1371/journal.pcbi.1008291.g005

ThermoNet fall within the neutral zone, whereas FoldX only predicted 39.7% of benign variants to be in the neutral zone (Fig 5B). Further, the ΔΔGs of pathogenic variants predicted by ThermoNet suggest pathogenic variants are nearly equally likely to be destabilizing (52.7%) as they are to be stabilizing (47.3%) (Fig 5C). 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 [23–25].

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 61.4% of pathogenic variants that have ΔΔGs within the neutral zone and 19.8% of benign variants have ΔΔGs outside the neutral zone (Fig 5B and 5C). 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 [44].

Discussion

Accurate modeling of protein thermodynamic stability is a complex task due to the delicate balance between the different thermodynamic state functions that contribute to protein stability [1]. The primary goal of this paper is to present a novel application of deep 3D-CNNs to a fundamental 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 CNNs in detecting spatially proximate features [29]. We developed ThermoNet, a method based on deep 3D-CNNs, to predict ΔΔG upon point mutation. We showed that ΔΔG can be predicted from protein structure with reasonable accuracy using deep 3D-CNNs without manual feature engineering. While ThermoNet achieved comparable performance to previous methods on direct mutations, it performed better on reverse mutations than most methods by a large margin, and remarkably, reduced the magnitude of prediction bias.

In addition to introducing ThermoNet, we also address two methodological challenges in the development and evaluation of computational methods for ΔΔG prediction: lack of anti-symmetry and data leaks due to homology between proteins in the training and test sets. Previously, it was shown that the lack of anti-symmetry in ΔΔG prediction can be effectively addressed either by using input features that are anti-symmetric by construction [19,37,38,48,49] or by training the predictor using both direct and reverse mutations [19,24]. In addition, when the statistical model is parametric, one may identify the terms that are responsible for breaking the symmetry and make correction accordingly [49]. However, when the predictor is nonparametric, meaning that the terms of the statistical learning model are not established *a priori*, the only way is to train the model with data set balanced with direct and reverse mutations so that it learns the anti-symmetry [24]. For structure-based ΔΔG predictors, this approach requires knowing the 3D structures of both the wild type and mutant protein. While mutant structures determined via experimental techniques are scarce, in this work, we demonstrated that mutant protein structures obtained through molecular modeling-based data augmentation can also be effectively used as substitutes for experimental structures to remedy the lack of anti-symmetry in ΔΔG prediction.

Recently, the potential for data leak between training and testing due to the inclusion of mutations from the same proteins has been appreciated [16,19]; however, the effects of including mutations from homologous proteins in training and test sets are less appreciated and understood. For example, while the S^{sym} data set was reasonably constructed and has been used practically to evaluate ΔΔG predictors, comparison of predictor performance is complicated by the presence of homology. Our results suggest that that such homology can influence performance estimates. The ThermoNet* model, which was trained before homology reduction, achieved stronger performance than the ThermoNet model trained after homology reduction (Fig 2B). In real-world applications, there will be homology between proteins used to train prediction models and the proteins to which they are applied. However, given the relatively small number of proteins included in commonly used training sets and the fact that they are not representative of the full diversity of protein folds and functions, we believe that the inclusion of mutations from proteins with shared evolutionary histories is likely to bias performance estimates. In the future, it will be valuable to explore this issue further and construct training sets that reflect the evolutionary relationships expected in various applications.

ThermoNet treats protein structures as if they were 3D images, and it takes as input a 3D grid of voxels parameterized with seven biophysical 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. Recently, a spherical convolutional architecture in which concentric voxel grids parameterized by atom masses and charges were used as input to predict the ΔΔGs of direct mutations with good accuracy [50]. Thus, together with the current work, these results demonstrate the potential of deep CNNs for predicting biophysically meaningful information from protein structures and hold 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 CNN create two limitations. First, while protein structures are being determined at an unprecedented pace, the fraction of the human proteome with available experimental structure is estimated to be around 20% [51]. Even when all the proteins whose structures can be modeled reliably are considered, only ~70% of the human proteome will have structural coverage [51]. 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 [52]. 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 thermodynamic stability

The thermodynamic stability 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:

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

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 stability, ΔΔG, is defined in terms of the change in ΔG between the wild-type and mutant proteins:

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

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 [13]. A mutation that destabilizes a protein with a typical stability (ΔG = -5 kcal/mol) by 1 kcal/mol will reduce the equilibrium constant for the folding reaction of this protein by a factor of 5.1 at physiological temperature (310 K).

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. This self-consistency requirement has been largely ignored by previously developed $\Delta\Delta G$ predictors [23–25].

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 [15]. The Q3421 data set contains 3,421 distinct single-point mutations in 150 proteins collected from the ProTherm database [53]. The effects of these mutations on the stability of protein structure have been measured experimentally and expressed quantitatively as ΔΔG values. We first excluded those mutations from the Q3421 data set that were also in the S^{sym} test set (see following). To reduce the sequence similarity between proteins in the training set of ThermoNet and the proteins it was tested on, we also removed all proteins that are likely homologous (BLAST e-value < 0.001) to p53, myoglobin, and proteins in S^{sym} from Q3421. Estimation of homology was accomplished by running the blastp program [54] using protein sequences in the S^{sym} data set and the sequences of p53 and myoglobin as queries against protein sequences in Q3421. Our rigorous pruning of the Q3421 data set resulted in a final data set consisting of 1,744 distinct mutations in 127 proteins. This data set was augmented by creating a reverse mutation data point for each of the 1,744 direct mutations, thus giving to a total of 3,488 data points for training ThermoNet. 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 ΔΔG predictors was a common, balanced data set called S^{sym} consisting of 342 pairs of proteins with known crystal structures [24]. The members forming each pair differ at only a single position in the protein sequence. The ΔΔG values of the 342 direct mutations have been experimentally measured and the ΔΔG 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 (PDB) [55] and were relaxed in the Rosetta all-atom energy function ref2015 [56] using the Rosetta *FastRelax* protocol [43]. 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 S^{sym} test set were collected from the PDB [55] 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 pixels element in an image have color channels, we parameterized voxels in a protein structure by a set of k biophysical 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 [57] as in the work of Jimenez et al. [28] (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 [58]. A Python program for creating input tensors from a list of mutations is provided in GitHub repository at https://github.com/gersteinlab/ThermoNet. The orientations of protein structures were taken directly from the retrieved PDB files without further adjustment, although our Python program does provide an option for rotating input structures about all three Cartesian axes.

Model architecture

CNNs are a type of deep-learning model commonly used in computer vision applications. They have recently proven to perform well in residue contact prediction [59–62] and protein tertiary structure prediction [63–66]. We selected CNNs for protein structure-based ΔΔG 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) (S1 Fig). 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 24 ReLU units. This model architecture was implemented using Keras [67] with TensorFlow [68] as the backend.

Training ThermoNet

ThermoNet is an ensemble of ten deep 3D-CNNs, 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. We employed this procedure to obtain an ensemble of ten models because model ensembling has been suggested to produce better predictions [69]. Evaluation of this ensemble was performed on a separate test set (see below). All component models of ThermoNet were trained using the Adam optimizer [70] 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 (S1 Fig). 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 [27] 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 (S1 Fig). 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 × 16 × 16 gives the best prediction performance (S1 Fig).

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:

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

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

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:$\delta = \Delta\Delta G_{rev} + \Delta\Delta G_{dir}$ and was previously used to quantify prediction bias [23]. 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 [24,37]. 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.

ClinVar variants

We retrieved the ClinVar database [47] in VCF format on August 15, 2019 and ran the VCF file through the Variant Effect Predictor (version 97) [71] 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 the pathogenic group and "benign" or "likely benign" for the benign group. All variants were required to have a review status of at least one star and no conflicting interpretation. All ClinVar variants designated as “no assertion criteria provided”, “no assertion provided”, “no interpretation for the single variant”, or not covered by protein structure were excluded from the evaluation set. Due to the dependency of ThermoNet on 3D structures, we also require variants in the evaluation set to be mappable to available protein structures. The residue-level mapping of ClinVar variants onto protein structures was based on the SIFTS resource that provides residue-level mapping between UniProt and PDB entries [72]. Collectively, these restrictions resulted in 3,510 pathogenic variants and 950 benign variants that can be mapped to experimental structures deposited in the PDB [55]. The mapped variants along with PDB IDs can be found in our GitHub repository at https://github.com/gersteinlab/ThermoNet.

Supporting information

S1 Fig. Architecture of the deep 3D convolutional neural network model and hyperparameter search. (A) The overall organization of the model begins with the input tensor and ends with a final layer that outputs ΔΔG prediction. The numbers in parentheses before the Flatten layer represent the dimensionality of the output from each layer in the format (width, height, depth, property channels). The number in parentheses starting from the Flatten layer represent the number of output features from each of the densely connected layers. This optimized architecture was determined through cross-validation. (B) Results from cross validating the sizes of the convolutional layers while keeping the size of the densely connected layer at 32 neurons. (C) Results from cross validating the size of the densely connected layer while keeping the sizes of the convolutional layers at (16, 24, 32). (D) Results from cross validating the dimensions of the input grid.
(PDF)

S2 Fig. Pairwise percent sequence identity of the proteins in the S2648 and VariBench data sets. (A) A heatmap representation of the pairwise percent sequence identity matrix of the proteins in the S2648 data set. (B) A heatmap representation of the pairwise percent sequence identity matrix of the proteins in the VariBench data set. It is obvious from these two heatmaps that there is substantial pairwise homology (percent identity $> 25\%$) in both S2648 and VariBench. The pairwise identity matrices were obtained using the Clustal Omega multiple sequence alignment program [73].
(PDF)

S3 Fig. CNNs trained using only direct mutations show a large prediction bias. (A) Performance of an ensemble of ten networks trained using only the set of 1,744 direct mutations on predicting the ΔΔGs of the direct mutations in the blind test set; The Pearson correlation coefficient (r) between predicted values and experimentally determined values is 0.47, and the root-mean-square deviation (σ) of predicted values from experimentally determined values is 1.38 kcal/mol. (B) Performance of the same ensemble of ten networks on predicting the ΔΔGs of the reverse mutations in the blind test set; The Pearson correlation coefficient (r) between predicted values and experimentally determined values is -0.06, and the root-mean-square deviation (σ) of predicted values from experimentally determined values is 2.40 kcal/mol. (C) Direct versus reverse ΔΔG values of all the mutations in the blind test set predicted by the same ensemble of networks. (B) and (C) highlight that the models trained with only direct mutations have a large bias and, when compared to the models trained using the balanced data set, the necessity of adding reverse mutations to correct the bias. The dots are colored in gradient from blue to red such that blue represents the most accurate prediction and red represents the least accurate.
(PDF)

S1 Table. PDB IDs of Identical proteins between PoPMuSiC-2.0 training and test sets.
(DOCX)

S2 Table. PDB IDs of non-redundant proteins returned by submitting proteins in the S2648 data set to the PISCES server.
(DOCX)

S3 Table. Proteins in the S2648 data set (Subject) that are either identical or likely to be homologous to proteins in the S^{sym} data set (Query).
(DOCX)

S4 Table. Proteins in the VariBench data set (Subject) that are either identical or likely to be homologous to proteins in the S^{sym} data set (Query).
(DOCX)

S5 Table. Proteins in the Q3421 data set (Subject) that are either identical or likely to be homologous to proteins in the S^{sym} data set (Query).
(DOCX)

S6 Table. A brief summary of the characteristics of methods presented in Table 2.
(DOCX)

S7 Table. Comparison of ThermoNet with four other methods on p53.
(DOCX)

S8 Table. Comparison of ThermoNet with four other methods on myoglobin.
(DOCX)

Acknowledgments

The authors would like to thank the Center for Research Computing at Yale University and the Advanced Computing Center for Research and Education at Vanderbilt University for supporting high-performance computing.

Author Contributions

Conceptualization: Bian Li, Mark B. Gerstein.

Data curation: Bian Li, Yucheng T. Yang.

Formal analysis: Bian Li, Yucheng T. Yang.

Funding acquisition: Bian Li, John A. Capra, Mark B. Gerstein.

Methodology: Bian Li, John A. Capra, Mark B. Gerstein.

Software: Bian Li.

Supervision: John A. Capra, Mark B. Gerstein.

Visualization: Bian Li, John A. Capra, Mark B. Gerstein.

Writing – original draft: Bian Li.

Writing – review & editing: Bian Li, Yucheng T. Yang, John A. Capra, Mark B. Gerstein.

References

1. Li B, Fooksa M, Heinze S, Meiler J. Finding the needle in the haystack: towards solving the protein-folding problem computationally. Crit Rev Biochem Mol Biol. 2018; 53(1):1–28. https://doi.org/10.1080/10409238.2017.1380596 PMID: 28976219
2. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245
3. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001; 17(4):263–70. https://doi.org/10.1002/humu.22 PMID: 11295823
4. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005; 353(2):459–73. https://doi.org/10.1016/j.jmb.2005.08.020 PMID: 16169011
5. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem Sci. 2019; 44(7):575–88. https://doi.org/10.1016/j.tibs.2019.01.003 PMID: 30712981

6. Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016; 537 (7620):320–7. https://doi.org/10.1038/nature19946 PMID: 27629638

7. Gapsys V, Michielssens S, Seeliger D, de Groot BL. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. Angew Chem Int Edit. 2016; 55(26):7364–8.

8. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins-Structure Function and Bioinformatics. 2011; 79 (3):830–8. https://doi.org/10.1002/prot.22921 PMID: 21287615

9. Bender BJ, Cisneros A, Duran AM, Finn JA, Fu D, Lokits AD, et al. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. Biochemistry. 2016; 55(34):4748–63. https://doi.org/10.1021/acs.biochem.6b00444 PMID: 27490953

10. Yin SY, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. Nature Methods. 2007; 4(6):466–7. https://doi.org/10.1038/nmeth0607-466 PMID: 17538626

11. Worth CL, Preissner R, Blundell TL. SDM-a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research. 2011; 39:W215–W22. https://doi.org/10.1093/nar/gkr363 PMID: 21593128

12. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC bioinformatics. 2011; 12.

13. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology. 2002; 320(2):369–87. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393

14. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Research. 2006; 34:W239–W42. https://doi.org/10.1093/nar/gkl190 PMID: 16845001

15. Quan LJ, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics. 2016; 32(19):2936–46. https://doi.org/10.1093/bioinformatics/btw361 PMID: 27318206

16. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014; 30(3):335–42. https://doi.org/10.1093/bioinformatics/btt691 PMID: 24281696

17. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research. 2005; 33:W306–W10. https://doi.org/10.1093/nar/gki375 PMID: 15980478

18. Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics. 2008; 24(18):2002–9. https://doi.org/10.1093/bioinformatics/btn353 PMID: 18632749

19. Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics. 2015; 31(17):2816–21. https://doi.org/10.1093/bioinformatics/btv291 PMID: 25957347

20. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. Journal of Chemical Information and Modeling. 2019. https://doi.org/10.1021/acs.jcim.8b00697 PMID: 30759982

21. Roushar FJ, Gruenhagen TC, Penn WD, Li B, Meiler J, Jastrzebska B, et al. Contribution of Cotranslational Folding Defects to Membrane Protein Homeostasis. J Am Chem Soc. 2019; 141(1):204–15. https://doi.org/10.1021/jacs.8b08243 PMID: 30537820

22. Buss O, Rudat J, Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J. 2018; 16:25–33. https://doi.org/10.1016/j.csbj.2018.01.002 PMID: 30275935

23. Thiltgen G, Goldstein RA. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. Plos One. 2012; 7(10). https://doi.org/10.1371/journal.pone.0046084 PMID: 23144695

24. Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics. 2018; 34(21):3659–65. https://doi.org/10.1093/bioinformatics/bty348 PMID: 29718106

25. Usmanova DR, Bogatyreva NS, Bernad JA, Eremina AA, Gorshkova AA, Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics. 2018; 34(21):3653–8. https://doi.org/10.1093/bioinformatics/bty340 PMID: 29722803

26. Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief Bioinform. 2019.

27. Jimenez J, Doerr S, Martinez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017; 33(19):3036–42. https://doi.org/10.1093/bioinformatics/btx350 PMID: 28575181

28. Jimenez J, Skalic M, Martinez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model. 2018; 58(2):287–96. https://doi.org/10.1021/acs.jcim.7b00650 PMID: 29309725

29. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–44. https://doi.org/10.1038/nature14539 PMID: 26017442

30. Torng W, Altman RB. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC bioinformatics. 2017; 18(1):302. https://doi.org/10.1186/s12859-017-1702-0 PMID: 28615003

31. Torng W, Altman RB. High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics. 2019; 35(9):1503–12. https://doi.org/10.1093/bioinformatics/bty813 PMID: 31051039

32. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv. 2015; arXiv:1510.02855

33. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–74. https://doi.org/10.1016/B978-0-12-381270-4.00019-6 PMID: 21187238

34. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics. 2009; 25(19):2537–43. https://doi.org/10.1093/bioinformatics/btp445 PMID: 19654118

35. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19 (12):1589–91. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846

36. Yang Y, Urolagin S, Niroula A, Ding X, Shen B, Vihinen M. PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. Int J Mol Sci. 2018; 19(4). https://doi.org/10.3390/ijms19041009 PMID: 29597263

37. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC bioinformatics. 2019; 20 (Suppl 14):335. https://doi.org/10.1186/s12859-019-2923-1 PMID: 31266447

38. Montanucci L, Savojardo C, Martelli PL, Casadio R, Fariselli P. On the biases in predictions of protein stability changes upon variations: the INPS test case. Bioinformatics. 2019; 35(14):2525–7. https://doi.org/10.1093/bioinformatics/bty979 PMID: 30496382

39. Pucci F, Bourgeas R, Rooman M. High-quality Thermodynamic Data on the Stability Changes of Proteins Upon Single-site Mutations. Journal of Physical and Chemical Reference Data. 2016; 45(2).

40. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat. 2002; 19(6):607–14. https://doi.org/10.1002/humu.10081 PMID: 12007217

41. Ordway GA, Garry DJ. Myoglobin: an essential hemoprotein in striated muscle. J Exp Biol. 2004; 207 (Pt 20):3441–6. https://doi.org/10.1242/jeb.01172 PMID: 15339940

42. Kepp KP. Towards a "Golden Standard" for computing globin stability: Stability and structure sensitivity of myoglobin mutants. Biochimica et biophysica acta. 2015; 1854(10 Pt A):1239–48. https://doi.org/10.1016/j.bbapap.2015.06.002 PMID: 26054434

43. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, et al. Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol. 2011; 405(2):607–18. https://doi.org/10.1016/j.jmb.2010.11.008 PMID: 21073878

44. Bromberg Y, Rost B. Correlating protein function and stability through the analysis of single amino acid substitutions. BMC bioinformatics. 2009; 10 Suppl 8:S8. https://doi.org/10.1186/1471-2105-10-S8-S8 PMID: 19758472

45. Ancien F, Pucci F, Godfroid M, Rooman M. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep. 2018; 8(1):4480. https://doi.org/10.1038/s41598-018-22531-2 PMID: 29540703

46. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018; 46(D1):D1062–D7. https://doi.org/10.1093/nar/gkx1153 PMID: 29165669

47. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 2005; 6(9):678–87. https://doi.org/10.1038/nrg1672 PMID: 16074985

48. Savojardo C, Martelli PL, Casadio R, Fariselli P. On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief Bioinform. 2019; 21(5):1856–1858. https://doi.org/10.1093/bib/bbz168 PMID: 31885042

49. Pucci F, Bernaerts K, Teheuxa F, Gilisa D, Roomana M. Symmetry Principles in Optimization Problems: an application to Protein Stability Prediction. IFAC-PapersOnLine. 2015; 48(1):458–63.

50. Boomsma W, Frellsen J, editors. Spherical convolutions and their application in molecular modelling. Advances in Neural Information Processing Systems; 2017; p3433–3443

51. Somody JC, MacKinnon SS, Windemuth A. Structural coverage of the proteome for pharmaceutical applications. Drug Discov Today. 2017; 22(12):1792–9. https://doi.org/10.1016/j.drudis.2017.08.004 PMID: 28843631

52. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nature Methods. 2014; 11(8):801–7. https://doi.org/10.1038/nmeth.3027 PMID: 25075907

53. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and Pro-NIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Research. 2006; 34:D204–D6. https://doi.org/10.1093/nar/gkj103 PMID: 16381846

54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

55. Rose PW, Prlic A, Altunkaya A, Bi CX, Bradley AR, Christie CH, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research. 2017; 45(D1): D271–D81. https://doi.org/10.1093/nar/gkw1000 PMID: 27794042

56. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017; 13(6):3031–48. https://doi.org/10.1021/acs.jctc.7b00125 PMID: 28430426

57. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. Journal of Computational Chemistry. 2009; 30(16):2785–91. https://doi.org/10.1002/jcc.21256 PMID: 19399780

58. Doerr S, Harvey MJ, Noe F, De Fabritiis G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. Journal of Chemical Theory and Computation. 2016; 12(4):1845–52. https://doi.org/10.1021/acs.jctc.6b00049 PMID: 26949976

59. Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins. 2019; 87(12):1092–1099. https://doi.org/10.1002/prot.25779 PMID: 31298436

60. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins. 2018; 86 Suppl 1:51–66. https://doi.org/10.1002/prot.25407 PMID: 29071738

61. Wang S, Sun SQ, Xu JB. Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins-Structure Function and Bioinformatics. 2018; 86:67–77. https://doi.org/10.1002/prot.25377 PMID: 28845538

62. Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, et al. Assessing the accuracy of contact predictions in CASP13. Proteins. 2019; 87(12):1058–1068. https://doi.org/10.1002/prot.25819 PMID: 31587357

63. Xu J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A. 2019 2019; 116(34):16856–16865. https://doi.org/10.1073/pnas.1821309116 PMID: 31399549

64. Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins. 2019; 87(12):1069–1081. https://doi.org/10.1002/prot.25810 PMID: 31471916

65. Kandathil SM, Greener JG, Jones DT. Recent developments in deep learning applied to protein structure prediction. Proteins. 2019: 87(12):1179–1189. https://doi.org/10.1002/prot.25824 PMID: 31589782

66. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun. 2019; 10(1):3977. https://doi.org/10.1038/s41467-019-11994-0 PMID: 31484923

67. Chollet F. keras. \url{https://github.com/fchollet/keras}; 2015.

68. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation; Savannah, GA, USA. 3026899: USENIX Association; 2016. p. 265–83.

69. Chollet F. Deep Learning with Python. Shelter Island, NY: Manning Pulications; 2018.

70. Kingma DP, Ba JL. Adam: a method for stochastic optimization. 2015;arXiv:1412.6980

71. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):122. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

72. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. 2019; 47(D1):D482–D9. https://doi.org/10.1093/nar/gky1114 PMID: 30445541

73. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7:539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835