Learning from physics-based features improves protein property prediction

Amy Wang
Department of Chemical Engineering[*]
Stanford University
Stanford, CA 94305
amywang1@stanford.edu

Ava P. Amini
Microsoft Research New England
Cambridge, MA
avasoleimany@microsoft.com

Alex X. Lu
Microsoft Research New England
Cambridge, MA
lualex@microsoft.com

Kevin K. Yang
Microsoft Research New England
Cambridge, MA
yang.kevin@microsoft.com

Abstract

Data-based and physics-based methods have long been considered as distinct approaches for protein property prediction. However, they share complementary strengths, such that integrating physics-based features with machine learning may improve model generalizability and accuracy. Here, we demonstrate that incorporating pre-computed energetic features in machine learning models improves performance in out-of-distribution and low training data regimes with two distinct protein engineering tasks. By training with sequence, structure, and pre-computed Rosetta energy features on graph neural nets, we achieve performance comparable to masked inverse folding pretraining with the same architecture.

1 Introduction

Proteins carry out a diverse range of complex functions essential to life and may possess valuable properties that can be optimized for industrial applications. Any small change to a protein's sequence can profoundly alter its conformation and function, which can be associated with disease [Li and Babu, 2018, Zheng et al., 2021, Cheng et al., 2021], or result in a new desirable variant in an engineering context [Romero and Arnold, 2009]. However, the effect of mutations on protein properties is non-additive, such that efforts to study mechanism or optimize function often require extensive experimental characterization, which is costly, time-consuming, and tedious. The ability to accurately predict the molecular properties of proteins *in silico* would therefore accelerate the rate of scientific discovery.

Machine learning is a promising approach for computationally predicting protein function [Fox et al., 2003, Yang et al., 2019, Wu et al., 2019, Hsu et al., 2022a]. While such approaches can effectively predict properties of unseen sequences and facilitate discovery of variants with desired characteristics, they require bespoke training datasets to effectively learn mappings between sequence and function. Collecting training data typically involves measuring libraries of protein variants using an experimental assay designed to capture some aspect of function. While some individual proteins have been systematically assayed for some functions via high-throughput experiments [Podgornaia and Laub, 2015, Wu et al., 2016, Sarkisyan et al., 2016], screening for many important other functional properties still requires painstaking characterization and assay development. Depending on

[*] Work performed during an internship at Microsoft Research

Preprint. Under review.

the efficiency of the data collection process, curating a training dataset for machine learning models may be more resource intensive than current pipelines that rely exclusively on experimental assays. Some strategies for predicting protein properties from scarce functional data include transfer learning from models pretrained on large datasets of protein sequences or structures [Bepler and Berger, 2021, Yang et al., 2022a, Wang et al., 2022]. However, recent work suggests that the relationship between fitness landscapes and evolutionary data is fundamentally limited [Weinstein et al., 2022].

On the other hand, because all molecular systems are bound by the laws of physics, every protein property stems from first principles. The physical characteristics of each amino acid in a protein determine energetically favorable atomic-level interactions, which fundamentally govern its dynamic processes and structure. Therefore, physics-based methods, which represent classical approaches in computational protein function prediction, should in theory be capable of directly predicting molecular properties from scarce data. In practice, atomistic modeling of protein systems with quantum-mechanical accuracy is computationally intractable, which has led to the development of force fields and sampling methods that approximate the underlying physics [Schymkowitz et al., 2005, Lopes et al., 2015, Alford et al., 2017]. These approaches have shown some success in protein property prediction but are computationally intensive and require domain expertise to develop [Glazer et al., 2009, Kirchner et al., 2011]. They also presently underperform relative to data-based methods [Li et al., 2020, Baek et al., 2021], possibly due to force field approximations, as energetic contributions of a few k_BT can significantly influence function [Lazaridis et al., 1995].

While data-based and physics-based methods historically represent distinct approaches for predicting protein function, their strengths and weaknesses complement each other. Machine learning models excel at adapting predictions to available data and identifying non-linear correlations but are limited by data scarcity. Physics-based models describe the functional properties of an individual physical system but are not data-aware and require mathematical approximations for computing non-linear relationships. Therefore, integrating biophysical knowledge with machine learning may help data-based methods achieve better generalization. Recent work has shown that incorporating physics-based $\Delta\Delta G$s as features improves some protein fitness predictions [Hsu et al., 2022a]. Additionally, Harmalkar et al. [2022] also demonstrated that Rosetta-computed features were useful for predicting antibody thermostability using machine learning, given that force fields directly model stability from protein structure.

Because thermodynamic principles underlie all functional characteristics of proteins, we reasoned that incorporating physics-derived interaction energies with machine learning should broadly improve protein property prediction beyond stability predictions. To test this hypothesis, we assessed whether energetic features improve the out-of-distribution and small training set performance of graph neural networks for two protein engineering datasets. We find that physics-based features enable equal or better performance compared to pretrained models using the same architecture. Ultimately, our results represent a proof of concept that learning biophysical principles can be broadly useful for protein property prediction.

2 Methods

We compare model performance across two protein engineering datasets, *Rma* NOD and GB1 (See Section A.1 for detailed descriptions). Protein fitness is defined as a metric for both stability and binding partner affinity.

Our baseline model takes in graph representations of sequence and structure from physics-based *in silico* mutagenesis for each variant. We concatenate graph representations of pre-computed energetic features and compare performance on protein property prediction (Figure 1). To compare our approach against a purely data-driven method, we also finetune a pretrained masked inverse folding language model.

2.1 Predicting protein properties with Geometric Vector Perceptrons

Previous work demonstrated that graph representations effectively capture spatial relationships in protein structures. Geometric vector perceptrons (GVP) have been used to predict protein sequences from structure, assess model quality within a pool of candidate structures, and make zero-shot predictions of mutation fitness effects [Jing et al., 2020, Hsu et al., 2022b].

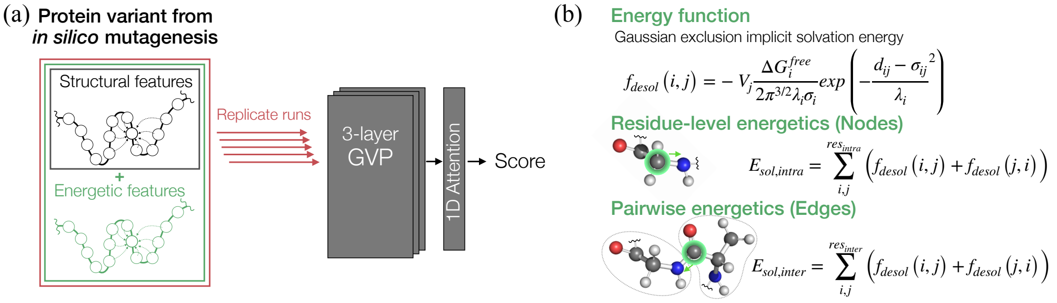

(a) Protein variant from *in silico* mutagenesis
Structural features
+
Energetic features
Replicate runs
3-layer GVP
1D Attention
Score
(b) Energy function
Gaussian exclusion implicit solvation energy
$f_{desol}(i,j) = -V_j \frac{\Delta G_i^{free}}{2\pi^{3/2}\lambda_i\sigma_i} exp\left(-\frac{d_{ij}-\sigma_{ij}}{\lambda_i}^2\right)$
Residue-level energetics (Nodes)
$E_{sol,intra} = \sum_{i,j}^{res_{intra}} \left(f_{desol}(i,j) + f_{desol}(j,i)\right)$
Pairwise energetics (Edges)
$E_{sol,inter} = \sum_{i,j}^{res_{inter}} \left(f_{desol}(i,j) + f_{desol}(j,i)\right)$

Figure 1: Leveraging energetic features to improve protein property prediction. (a) Structural features are represented as a graph, and pre-computed energetics from Rosetta are concatenated as additional features. Replicate runs of each variant from *in silico* mutagenesis are treated as distinct examples. (b) Example of a Rosetta-derived energetic parameter. Node features describe the sum of energetic contributions between atoms within a residue, whereas edge features similarly describe interaction energies between two residues. See Section A.3 for an overview of other Rosetta energy terms.

Nodes and edges describe the one-hot encoded amino acid representation and spatial features of the protein backbone (see Section A.2). We use three graph propagation steps in which messages from 30 neighboring nodes and edges update each node embedding, and a node-wise attention layer with a subsequent dense feed-forward network then reduces node embeddings to a scalar property (Figure 1). Table A1 provides additional details on model hyperparameters and training.

2.2 Incorporating energetic features from PyRosetta

To perform *in silico* mutagenesis and compute energetic features for each mutant in the dataset, we used a semi-empirical physics-based macromolecular modeling package called PyRosetta [Chaudhury et al., 2010]. PDB IDs 6WK3 and 2GI9 were used as reference structures for *Rma* NOD and GB1, respectively [Wittmann et al., 2020, Franks et al., 2006]. For every mutation, side chains of amino acids within 10Å of the mutated residue were repacked, and a FastRelax simulation was performed with the Rosetta Energy Function 2015 (REF15) to find a protein conformation corresponding to a local energetic minimum.

REF15 describes the decomposable all-atom energetic parameters of a protein structure. For computational efficiency, the atomic energetic contributions were summed across each residue. Thus, the energetic features of a protein structure are defined by intra-residue energies and inter-residue energies for each residue pair, as described in Table A3. Because prior work suggests that the empirical weights of the energetic terms likely contribute to the inaccuracy of REF15, we used the unweighted energies [Rubenstein et al., 2018].

Proteins are highly dynamic, and their conformational distribution contributes towards protein function. However, the PDB structure is a static snapshot and therefore inadequately represents the physics, which describe the forces and motions within a molecular environment. Therefore, we consider several possible conformations a protein variant may adopt by performing five pyRosetta relaxation simulations. We augment the training data by considering each of the five pyRosetta runs (‘samples’) for a protein variant as a distinct example. Likewise, we perform test-time augmentation by averaging predictions from the five samples.

Given that residue-level energy features map directly to a protein structure, we concatenated intra-residue energetics to node features and inter-residue energetics to edge features in all models that include energetic features (Figure 1B). To normalize and rescale inputs, we took the difference between energetic features for each variant with those computed from the reference PDB structure corresponding to the wild-type sequence.

3 Results

We assess the ability of energetic features to improve performance on out-of-distribution splits and low training sample size in-distribution splits. Mean square error (MSE) was used as the primary

Table 1: MSE for out-of-domain prediction of *Rma* NOD enantioselectivity. Uncertainties are standard deviations over 3 random seeds.

	No energetics		With energetics	
	Baseline	Pretrain	5 samples	1 sample
3-vs-many	0.08 ± 0.01	0.10 ± 0.01	**0.07 ± 0.00**	**0.07 ± 0.00**
4-vs-many	0.08 ± 0.02	0.08 ± 0.02	0.08 ± 0.01	**0.07 ± 0.01**

Table 2: MSE for out-of-domain prediction of GB1 fitness from FLIP. Uncertainties are standard deviations over 3 random seeds.

	No energetics		With energetics	
	Baseline	Pretrain	5 samples	1 sample
2-vs-many	1.41 ± 0.14	**1.16 ± 0.06**	**1.16 ± 0.05**	1.32 ± 0.13

metric to assess model performance, as rank correlation metrics do not properly account for the fitness score magnitudes.

3.1 Models with energetic features match pretrained performance on out-of-distribution predictions

The ability to accurately predict the functional effect of several mutations from variants with fewer mutations can reduce tedious and costly experimental characterization. Given that physics-based methods generalize well, we hypothesized that incorporating energetic features may improve these out-of-distribution predictions. Table 1 shows model performance on predicting the enantioselectivity of *Rma* NOD variants with up to seven mutations after training only on 74 variants with mutations at three positions (‘3-vs-many’), or 208 variants with mutations at four positions (‘4-vs-many’) (Table A2). Validation and test were randomly split. We find that pretraining does not improve predictions, consistent with previous work with a different masked inverse folding pretrained model [Yang et al., 2022a]. On the other hand, training with energetic features produces modest improvements in out-of-domain prediction.

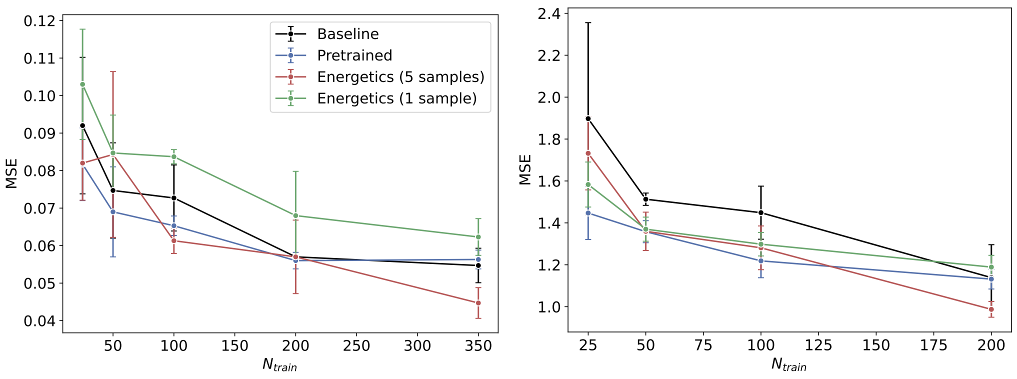
Additionally, Table 2 compares performance for the GB1 2-vs-many split from FLIP [Dallago et al., 2021], in which models predict fitness on variants with three or four mutations after being trained on variants with one or two mutations (Table A2). Performance on the 2-vs-many split between other pretrained models and baselines varied significantly, whereas performance between models on other GB1 splits showed little separation [Dallago et al., 2021, Yang et al., 2022b]. We find that the performance of models trained with augmented energetic features is comparable to that of the pretrained model, where pretraining and energetics both confer substantial gains in performance relative to the baseline.

3.2 Energetic features improve performance in low data regimes

Models trained on small training datasets may achieve better performance with energetic features, which can help accelerate scientific discovery. Thus, we compared model performance across several different training sample sizes (N_{train}) from randomly sampled splits, with results shown in Figure 2.

For the *Rma* NOD dataset, we tested on a held out set of 171 variants and randomly shuffled the remaining 385 examples for 3 different seeds for train and validation splits. Subsampled variants were partitioned in a 90% train and 10% validation split (Table A2). Training with augmented energetic features improves model accuracy relative to baseline, except for $N_{\text{train}} = 50$, as shown in Figure 2a.

Additionally, we used the GB1 sampled split from FLIP [Dallago et al., 2021] to test on a held out set of 1745 variants. We subsample within the train and validation splits after randomly shuffling for 3 different seeds such that validation sample sizes were 10% that of N_{train} (Table A2). Here, we find that training with augmented energetic features improves model accuracy relative to baseline across all analyzed training sample sizes, as shown in Figure 2b.



(a) *Rma* NOD

(b) GB1

Figure 2: Comparison of MSE for small training set sizes.

Interestingly, these preliminary results suggest that energetic features may improve in-distribution predictive performance relative to pretrained models for intermediate values of N_{train} but not for low N_{train}. The difference in performance between pretrained models and baseline narrows with increasing training sample sizes. Pretrained and baseline models both achieve a mean MSE of roughly 0.055 and 1.13 for *Rma* NOD ($N_{\text{train}} = 350$) and GB1 ($N_{\text{train}} = 200$), respectively. Models trained with augmented energetic features achieve considerably lower mean MSE with 0.045 ± 0.004 and 0.98 ± 0.04 for *Rma* NOD and GB1, respectively. However, pretrained models outperform those trained with energetic features at low N_{train} and benefit from having smaller variance across seeds due to the consistent weight initialization.

3.3 Sampling energetic features from multiple conformations generally improves model performance

The outputs from a single Rosetta run represent a static representation of a protein, which does not adequately represent the conformational dynamics that may influence function. To determine whether a broader sample of conformational space would improve performance, we compare performance after performing several runs Rosetta runs on each variant. Including energetic features from just one pyRosetta run results in worse performance compared to a pretrained model for out-of-distribution GB1, and compared to the baseline for in-distribution *Rma* NOD across all N_{train}. However, using five pyRosetta runs almost always improves performance (Figure 2, Table 1, Table 2), matching or exceeding performance compared to the pretrained condition. Ultimately, we find that data augmentation via several rounds of pyRosetta runs is required to improve model performance relative to the baseline and pretrained conditions. These observations support current biophysical principles describing protein mechanism, that the molecular details of a protein's dynamic conformation and chemical environment significantly influence its behavior.

4 Discussion

In this work, we demonstrate that energetic features computed from a physics-based method improve the performance of graph neural networks on two protein engineering tasks. Strikingly, models trained with just sequence, structure, and energetic features can achieve competitive performance, relative to pretrained models with a similar number of parameters, on out-of-distribution predictions and in low-data regimes. However, more experiments on a wider range of tasks, such as thermostability, mechanostability, fluorescence, and catalytic efficiency, are required to truly assess the extent to which biophysical principles improve the generalizability and robustness of neural network predictions.

Ultimately, integrating data-based models with physics may enable development of methods that are data-aware, principled, and more accurate. Given the complementarity of data-based and physics-based methods, such strategies may unlock a new paradigm in protein modeling and design that could lead to new scientific understanding, disease treatments, and enhanced protein variants.

References

Xiao-Han Li and M. Madan Babu. Human Diseases from Gain-of-Function Mutations in Disordered Protein Regions. *Cell*, 175(1):40–42, September 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.08.059. URL https://www.sciencedirect.com/science/article/pii/S0092867418311206.

Fan Zheng, Marcus R. Kelly, Dana J. Ramms, Marissa L. Heintschel, Kai Tao, Beril Tutuncuoglu, John J. Lee, Keiichiro Ono, Helene Foussard, Michael Chen, Kari A. Herrington, Erica Silva, Sophie N. Liu, Jing Chen, Christopher Churas, Nicholas Wilson, Anton Kratz, Rudolf T. Pillich, Devin N. Patel, Jisoo Park, Brent Kuenzi, Michael K. Yu, Katherine Licon, Dexter Pratt, Jason F. Kreisberg, Minkyu Kim, Danielle L. Swaney, Xiaolin Nan, Stephanie I. Fraley, J. Silvio Gutkind, Nevan J. Krogan, and Trey Ideker. Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, 374(6563):eabf3067, October 2021. doi: 10.1126/science.abf3067. URL https://www.science.org/doi/10.1126/science.abf3067. Publisher: American Association for the Advancement of Science.

Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R. Martin, Ruisheng Wang, Jin Huang, Tong Hao, Hong Yue, Jing Ma, Yuan Hou, Jessica A. Castrillon, Jiansong Fang, Justin D. Lathia, Ruth A. Keri, Felice C. Lightstone, Elliott Marshall Antman, Raul Rabadan, David E. Hill, Charis Eng, Marc Vidal, and Joseph Loscalzo. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature Genetics*, 53(3):342–353, March 2021. ISSN 1546-1718. doi: 10.1038/s41588-020-00774-y. URL https://doi.org/10.1038/s41588-020-00774-y.

Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009. ISSN 1471-0080. doi: 10.1038/nrm2805. URL https://www.nature.com/articles/nrm2805. Number: 12 Publisher: Nature Publishing Group.

Richard J Fox, Ajoy Roy, Sridhar Govindarajan, Jeremy Minshull, Claes Gustafsson, Jennifer T Jones, and Robin Emig. Optimizing the search algorithm for protein engineering by directed evolution. *Protein Engineering*, 16(8):589–597, 2003.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, 2019. doi: 10.1038/s41592-019-0496-6.

Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences USA*, 2019.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 2022a.

Anna I Podgornaia and Michael T Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, 2015. doi: 10.1126/science.1257360.

Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016. doi: 10.7554/eLife.16965.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, Natalya S Bogatyreva, Peter K Vlasov, Evgeny S Egorov, Maria D Logacheva, Alexey S Kondrashov, Dmitry M Chudakov, Ekaterina V Putintseva, Ilgar Z Mamedov, Dan S Tawfik, Konstantin A Lukyanov, and Fyodor A Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397, 2016. doi: 10.1038/nature17995.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, June 2021. ISSN 2405-4720. doi: 10.1016/j.cels.2021.05.017.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *bioRxiv*, 2022a.

Zichen Wang, Steven A. Combs, Ryan Brand, Miguel Calvo Rebollar, Panpan Xu, George Price, Nataliya Golovach, Emmanuel Oluwatobi Salawu, Colby Wise, Sri Priya Ponnapalli, and Peter M. Clark. LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, 12, 2022.

Eli N Weinstein, Alan N Amin, Jonathan Frazer, and Debora S Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *bioRxiv*, 2022.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue):W382–W388, July 2005. ISSN 0305-1048. doi: 10.1093/nar/gki387. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160148/`.

Pedro EM Lopes, Olgun Guvench, and Alexander D MacKerell Jr. Current status of protein force fields for molecular dynamics. *Methods in Molecular Biology*, 2015.

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Dariya S. Glazer, Randall J. Radmer, and Russ B. Altman. Improving Structure-Based Function Prediction Using Molecular Dynamics. *Structure*, 17(7):919–929, July 2009. ISSN 0969-2126. doi: 10.1016/j.str.2009.05.010. URL `https://www.sciencedirect.com/science/article/pii/S0969212609002226`.

Barbara Kirchner, Philipp J di Dio, and Juerg Hutter. Real-world predictions from ab initio molecular dynamics simulations. *Multiscale Molecular Methods in Applied Chemistry*, pages 109–153, 2011.

Bian Li, Yucheng T. Yang, John A. Capra, and Mark B. Gerstein. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Computational Biology*, 16(11):e1008291, November 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008291. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008291`. Publisher: Public Library of Science.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. doi: 10.1126/science.abj8754. URL `https://www.science.org/doi/10.1126/science.abj8754`. Publisher: American Association for the Advancement of Science.

Themis Lazaridis, Georgios Archontis, and Martin Karplus. Enthalpic contribution to protein stability: Insights from atom-based calculations and statistical mechanics. *Advances in Protein Chemistry*, 1995.

Ameya Harmalkar, Roshan Rao, Jonas Honer, Wibke Deisting, Jonas Anlahr, Anja Hoenig, Eva Czwilkla, Julia Sienz-Widmann, Doris Rau, Austin Rice, Timothy P Riley, Danqing Li, Hannah B Catterall, Christine E Tinberg, Jeffrey J Gray, and Kathy Y Wei. Towards generalizable prediction of antibody thermostability using machine learning on sequence and structure features. *bioRxiv*, 2022.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022b.

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, 2010.

Bruce J Wittmann, Anders M Knight, Julie L Hofstra, Sarah E Reisman, SB Jennifer Kan, and Frances H Arnold. Diversity-oriented enzymatic synthesis of cyclopropane building blocks. *ACS catalysis*, 10(13):7112–7116, 2020.

W Trent Franks, Benjamin J Wylie, Sara A Stellfox, and Chad M Rienstra. Backbone conformational constraints in a microcrystalline U-15N-labeled protein by 3D dipolar-shift solid-state NMR spectroscopy. *Journal of the American Chemical Society*, 128(10):3154–3155, 2006.

Aliza B. Rubenstein, Kristin Blacklock, Hai Nguyen, David A. Case, and Sagar D. Khare. Systematic Comparison of Amber and Rosetta Energy Functions for Protein Structure Evaluation. *Journal of Chemical Theory and Computation*, 14(11):6015–6025, November 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b00303. URL `https://doi.org/10.1021/acs.jctc.8b00303`. Publisher: American Chemical Society.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Kevin K Yang, Alex X Lu, and Nicolo K Fusi. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 2022b.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, pages 15794–15805, 2019.

A Appendix

A.1 Description of datasets

We chose protein engineering datasets that have an existing PDB structure and that have previously been evaluated on state-of-the-art pretrained models [Yang et al., 2022a]. Because these libraries differ in the breadth and depth in mutation sequence coverage, we are able to assess the effects of energetic features more holistically.

1. *Rma* NOD: Wu et al. [2019] constructed a small library of 556 *Rhodothermus marinus* (*Rma*) nitric oxide dioxygenase (NOD) variants with mutations at seven positions that affect the relative selective yield of carbon-silicon bond enantiomers ('enantioselectivity'). Targeted amino acid positions for mutation were rationally selected based on putative mechanistic hypotheses. This suggests that incorporating biophysical principles should be helpful, although the connection between protein energetics and enantioselectivity is less direct.

2. GB1:Wu et al. [2016] characterized the fitness of 149,361 variants out of 160,000 possible combinations of mutations at four positions of GB1, the binding domain of an immunoglobulin binding protein G from Streptococcal bacteria. Due to the depth of combinatorial sequences tested, the GB1 dataset is considered a gold standard for studying interactions between mutations. Protein fitness incorporates both stability and binding affinity metrics. We use the downsampled splits from FLIP [Dallago et al., 2021]. Although stability is directly associated with protein energetics, the amino acids targeted for mutation were optimized for a rugged fitness landscape such that their properties are difficult to rationally predict. Additionally, while binding affinity is also directly associated with the underlying energetics between the interacting substrates, here we consider GB1 only in its isolated state, and not bound to protein G, when computing energetic features.

A.2 GVP training details

As in Jing et al. [2020], our baseline models consider the one-hot sequence encoding and strutural features. Scalar node features include the one-hot representation of each amino acid identity and sine and cosine transformations of the dihedral angles. Vector node features include forward and reverse unit vectors from Cα_{i-1} – Cα_i and Cα_{i+1} – Cα_i, and a unit vector from Cβ_i – Cα_i. Edges for each node are defined with respect to its 30 closest neighbors, where scalar features describe the distance between nodes and a sine transformation of the distance along the backbone, and vector features include the unit vector along Cα_j – Cα_i.

Previous work [Yang et al., 2022a] trained a masked inverse folding protein language model parameterized as a structured graph neural network. Here, we apply similar methods to 3-layer GVP model trained on several thousands of structures CATH 4.2 using training, validation and testing splits from Ingraham et al. [2019] to serve as basis of comparison for a purely data-driven model. Briefly, the pretraining task reconstructs a corrupted protein sequence inspired by BERT, conditioned on its backbone structure. We then use the pretrained weights with the best validation loss as a starting point for training on a downstream task. All baseline and pretrained GVP models considered in this study had 608,587 parameters, whereas models with energetic features had 610,151 parameters after concatenating 13 node features and 7 edge features pre-computed from Pyrosetta (Section A.3).

Hyperparameters for each task were consistent across all models and determined based on screening for optimal pretrained performance out-of-distribution performance (Table A1). All models had a hidden scalar node dimension of 100, vector node dimension of 16, scalar edge dimension of 32, and hidden vector edge dimension of 1 and were trained with an Adam optimizer. All test MSEs are reported for the epoch corresponding to the best validation loss.

Values for the number of unique sequences in each training, validation, and test split are reported in Table A2.

Table A2: Number of unique sequences in each split across tasks.

	Train	Validation	Test
***Rma* NOD**			
3-vs-many	74	241	241
4-vs-many	208	174	174
random sample	N_{train}	0.1 $\times N_{train}$	171
GB1			
2-vs-many	381	43	8309
random sample	N_{train}	0.1 $\times N_{train}$	1745

Table A1: Model hyperparameters.

	Pretraining	Protein engineering tasks	
	CATH	*Rma* NOD	GB1
Epochs	200	500	500
Warm-up updates	1000	0	0
Learning rate	0.001 (max)	0.0001	0.0005
Batch size (Maximum number of nodes)	6000	1000	1000
Dropout rate	0.5	0	0

A.3 Description of Rosetta energy terms

We consider the residue-level and pairwise residue energetics from the PyRosetta REF15 force field as node and edge features in the graph neural net, respectively. A description for the decomposed energy terms is provided in Table A3, as described in Alford et al. [2017].

Figure 1 extended description: The Gaussian exclusion implicit solvation model describes the energy required to remove contacting water when an atom *i* is approached by neighboring atom *j* is parameterized by $\Delta G^{free}, \lambda, \sigma, V$ for vapor-water transfer free energy, correlation length, atomic radius, and desolvating atomic volume, respectively.

Table A3: Descriptions of decomposed Rosetta energy terms.

Per-residue energy terms (node features)	
fa_intra_sol_xover4	Gaussian exclusion implicit solvation energy between protein atoms in the same residue
omega	backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°
yhh_planarity	sinusoidal penalty for nonplanar tyrosine χ_3 dihedral angle
per_residue_hbond_sr_bb	energy of short-range hydrogen bonds
per_residue_hbond_lr_bb	energy of long-range hydrogen bonds
per_residue_hbond_bb_sc	energy of backbone–side-chain hydrogen bonds
per_residue_hbond_sc	energy of side-chain–side-chain hydrogen bonds
pro_close	penalty for an open proline ring and proline ω bonding energy
ref	reference energies for amino acid types
dslf_fa13	energy of disulfide bridges
Empirical energetic terms	-
fa_dun	probability that a chosen rotamer is native-like given backbone ϕ, ψ angles
rama_prepro	probability of backbone ϕ, ψ angles given the amino acid type
p_aa_pp	probability of amino acid identity given backboneϕ, ψ angles
Pairwise energy terms (edge features)	
fa_sol	Gaussian exclusion implicit solvation energy between protein atoms in different residues
fa_atr	attractive energy between two atoms on different residues separated by a distance d
fa_rep	repulsive energy between two atoms on different residues separated by a distance d
fa_elec	energy of interaction between two nonbonded charged atoms separated by a distance d
pairwise_hbond_unweighted	summed energy of all hydrogen bond contributions between atoms on different residues
pairwise_hbond_weights	weights for hydrogen bond contributions dependent on the environment and surface exposure
lk_ball_wtd	orientation-dependent solvation of polar atoms assuming ideal water geometry