

---

# Zero-shot protein stability prediction by inverse folding models: a free energy interpretation

---

**Jes Frellsen**<sup>\*†</sup>  
Technical University  
of Denmark

**Maher M. Kassem**<sup>\*</sup>  
Novonosis<sup>‡</sup>

**Tone Bengtsen**  
Novonosis<sup>‡</sup>

**Lars Olsen**  
Novonosis

**Kresten Lindorff-Larsen**  
University of  
Copenhagen

**Jesper Ferkinghoff-Borg**  
Novo Nordisk

**Wouter Boomsma**<sup>†</sup>  
University of  
Copenhagen

## Abstract

Inverse folding models have proven to be highly effective zero-shot predictors of protein stability. Despite this success, the link between the amino acid preferences of an inverse folding model and the free-energy considerations underlying thermodynamic stability remains incompletely understood. A better understanding would be of interest not only from a theoretical perspective, but also potentially provide the basis for stronger zero-shot stability prediction. In this paper, we take steps to clarify the free-energy foundations of inverse folding models. Our derivation reveals the standard practice of likelihood ratios as a simplistic approximation and suggests several paths towards better estimates of the relative stability. We empirically assess these approaches and demonstrate that considerable gains in zero-shot performance can be achieved with fairly simple means.

## 1 Introduction

Quantifying how amino acid substitutions affect protein structural stability is of fundamental importance for understanding human genetic diseases (Stein et al., 2019), and for our ability to design and optimise industrial enzymes and protein therapeutics (Listov et al., 2024). In recent years, multiplexed assays of variant effects (MAVE; see, e.g., Starita et al., 2017) experiments have greatly enhanced our ability to characterize variants experimentally, but still only scratch the surface compared to the astronomically large number of possible variants (we use the term variant to refer to a protein that differs from a reference, ‘wild-type’ protein by one or more amino acid substitutions). *In silico* prediction, therefore, remains an important tool, as a low-cost and fast alternative to experimental characterisation, but also to extrapolate meaningfully from experimental results to uncharacterized variants.

Despite the progress in high-throughput experimental characterisation, variant-effect data still falls in the low-data regime, and supervised learning in this domain has been observed to be prone to overfitting, exemplified by several large-scale studies on human variants (Livesey and Marsh, 2020). As a consequence, unsupervised or weakly supervised methods are often the most attractive approach to the problem. For protein stability prediction, inverse folding models, which provide a probability distribution over amino acid sequences given a fixed 3D structure, have emerged as particularly useful. In particular, variations of log-ratios based on the form

$$-\ln \frac{p(\text{variant sequence} \mid \text{wild-type structure})}{p(\text{wild-type sequence} \mid \text{wild-type structure})} \quad (1)$$

---

<sup>\*</sup>Equal contribution. <sup>‡</sup>Work performed while employed at Novonosis.

<sup>†</sup>Corresponding authors: jefr@dtu.dk & wb@di.ku.dk

have shown strong empirical correlation with experimental measurements of stability (Boomsma and Frellsen, 2017; Meier et al., 2021; Hsu et al., 2022; Dutton et al., 2024; Cagiada et al., 2025).

An inverse folding model describes amino acid preferences conditioned on their structural environment. While it is intuitively appealing to assume that likely amino acids correspond to stable protein structures, the formal connection between these probabilities and folding free energies remains incompletely understood. For instance, since thermodynamic stability is an ensemble property (a property of the entire structural distribution), it is not clear why it would be sufficient to condition on a single structure. Likewise, since thermostability is a balance between a folded and an unfolded state, one might reasonably wonder whether one should explicitly model the unfolded-state propensities, in addition to the current practice of generally considering the folded state alone. This raises the question: *Can we interpret these inverse folding models in terms of thermodynamic stability?*

In this paper, we derive a theoretical connection between inverse folding probabilities and thermodynamic stability and use this to elucidate current practices. We also suggest a number of improvements to the current protocols. Our contributions are:

- We derive a formal relationship between changes in thermodynamic stability ( $\Delta\Delta G$ ) and changes in inverse folding probabilities, and describe the approximations necessary to explain the current practice of simple probability ratios, cf. eq. (1).
- We show that the current practice corresponds to single-sample Monte Carlo estimates, and demonstrate empirical performance gains by using multiple samples from molecular dynamics (MD) simulations or the BioEmu generative model (Lewis et al., 2025).
- We show that the unfolded state can be disregarded, cf. eq. (1), but this introduces an additional factor in the probability ratio. We also propose and evaluate several strategies for explicitly modelling the unfolded state and show that these improve empirical performance.

## 2 Background

Assuming isothermal-isobaric conditions (i.e., an NPT ensemble), the stability of a protein is determined as the difference in Gibbs free energy,  $\Delta G$ , between its folded and unfolded states. These two states are *macro states* in the sense that they each correspond to a *set* of structural conformations. How this quantity changes as a result of changing one or several amino acid residues with another (substitutions) is referred to as a  $\Delta\Delta G$ . Denoting the original sequence as *wild type* (WT) with amino sequence  $\mathbf{a}$  and the new sequence as *variant* (MT) with sequence  $\mathbf{a}'$ , we have that

$$\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} = \Delta G_{\mathbf{a}'\rightarrow\text{F}}^{\text{U}} - \Delta G_{\mathbf{a}\rightarrow\text{F}}^{\text{U}} = (G_{\mathbf{a}'}^{\text{F}} - G_{\mathbf{a}'}^{\text{U}}) - (G_{\mathbf{a}}^{\text{F}} - G_{\mathbf{a}}^{\text{U}}). \quad (2)$$

### 2.1 Gibbs free energy

The Gibbs free energy is calculated from the Boltzmann distribution, expressing the probability of a protein microstate with structural degrees of freedom  $\chi \in \mathcal{X}$  and solvent  $\mathbf{w}$ , given by

$$p(\chi, \mathbf{w} | \mathbf{a}, \beta) = Z_{\mathbf{a}}^{-1} e^{-\beta H_{\mathbf{a}}(\chi, \mathbf{w})}, \quad Z_{\mathbf{a}} = \int e^{-\beta H_{\mathbf{a}}(\chi, \mathbf{w})} d\chi d\mathbf{w} \quad (3)$$

where  $H_{\mathbf{a}}(\chi, \mathbf{w}) = U_{\mathbf{a}}(\chi, \mathbf{w}) + PV(\chi, \mathbf{w})$  is the enthalpy of the microstate for amino acid sequence  $\mathbf{a}$ ,  $U_{\mathbf{a}}(\chi, \mathbf{w})$  is the internal energy (Hamiltonian),  $V(\chi, \mathbf{w})$  is volume of the microstate,  $P$  is the pressure, and  $\beta = \frac{1}{k_{\text{B}}T}$  is the inverse of the thermodynamic temperature. The Gibbs free energy associated with the amino acid sequence  $\mathbf{a}$  is defined as

$$G_{\mathbf{a}} = -\beta^{-1} \log Z_{\mathbf{a}}. \quad (4)$$

Since our focus is on the degrees of freedom of the protein, we integrate out the solvent degrees of freedom. Furthermore, we split the structure degrees of freedom  $\chi = (\mathbf{x}, \mathbf{s})$  into backbone degrees of freedom  $\mathbf{x}$  and side-chain degrees of freedom  $\mathbf{s}$ , and integrate out the side chains. The probability of a backbone configuration  $\mathbf{x}$  for a given amino acid sequence  $\mathbf{a}$  is then given by

$$p(\mathbf{x} | \mathbf{a}, \beta) = Z_{\mathbf{a}}^{-1} e^{-\beta H_{\mathbf{a}}(\mathbf{x})}, \quad Z_{\mathbf{a}} = \int e^{-\beta H_{\mathbf{a}}(\mathbf{x})} d\mathbf{x} \quad (5)$$

where  $H_{\mathbf{a}}(\mathbf{x})$  is the free energy associated with an implicit treatment of the solvent and side-chain freedom defined by  $e^{-\beta H_{\mathbf{a}}(\mathbf{x})} = \int e^{-\beta H_{\mathbf{a}}(\chi, \mathbf{w})} d\mathbf{s} d\mathbf{w}$ .

## 2.2 Thermodynamic stability

To calculate the stability of a protein, we usually partition<sup>3</sup> the space of structural degrees of freedom  $\mathfrak{X}$  into a folded subset  $\mathfrak{X}_a^F$  and an unfolded subset  $\mathfrak{X}_a^U$  (Brandts, 1969; Lindorff-Larsen and Teilmann, 2021). We will use  $S \in \{F, U\}$  to denote either of the two states. We will assume that the states can be fully characterized by the backbone degrees of freedom, such that the space of backbone degrees of freedom  $\mathbb{X}_a$  can be partitioned into a folded subset  $\mathbb{X}_a^F$  and an unfolded subset  $\mathbb{X}_a^U$ . For generality, we use a soft partitioning given by the probability  $p(S|\mathbf{x}, \mathbf{a})$ , where the hard partitioning above is a special case specified through the indicator function. We can then write the partition function (normalisation constant) for the state  $S$  as

$$Z_a^S = \int e^{-\beta H_a(\mathbf{x})} p(S|\mathbf{x}, \mathbf{a}) d\mathbf{x}. \quad (6)$$

Similarly to eq. (4), the free energy of state  $S$  is then given by

$$G_a^S = -\beta^{-1} \log Z_a^S, \quad (7)$$

and the folding stability of the protein with amino acid sequence  $\mathbf{a}$  is given by

$$\Delta G_a^{U \rightarrow F} = G_a^F - G_a^U \quad \text{or equivalently} \quad \beta \Delta G_a^{U \rightarrow F} = \ln \frac{Z_a^U}{Z_a^F}. \quad (8)$$

If we write the stability in terms of integrals over the Boltzmann distributions, we see that the stability can be expressed as a ratio of probabilities

$$\beta \Delta G_a^{U \rightarrow F} = \ln \frac{\int e^{-\beta H_a(\mathbf{x})} p(U|\mathbf{x}, \mathbf{a}) d\mathbf{x}}{\int e^{-\beta H_a(\mathbf{x})} p(F|\mathbf{x}, \mathbf{a}) d\mathbf{x}} = \ln \frac{\int p(\mathbf{x}|\mathbf{a}, \beta) p(U|\mathbf{x}, \mathbf{a}) d\mathbf{x}}{\int p(\mathbf{x}|\mathbf{a}, \beta) p(F|\mathbf{x}, \mathbf{a}) d\mathbf{x}} = \ln \frac{p(S=U|\mathbf{a}, \beta)}{p(S=F|\mathbf{a}, \beta)}, \quad (9)$$

where  $p(S|\mathbf{a}, \beta) = \int p(\mathbf{x}|\mathbf{a}, \beta) p(S|\mathbf{x}, \mathbf{a}) d\mathbf{x}$  denotes the probability of finding the protein with sequence  $\mathbf{a}$  in state  $S$ . Since  $p(F|\mathbf{a}, \beta) + p(U|\mathbf{a}, \beta) = 1$ , it follows from eq. (9) that

$$\beta \Delta G_a^{U \rightarrow F} = \ln \frac{1 - p(S=F|\mathbf{a}, \beta)}{p(S=F|\mathbf{a}, \beta)} = \ln \left( \frac{1}{p(S=F|\mathbf{a}, \beta)} - 1 \right), \quad (10)$$

which means that knowing  $p(S=F|\mathbf{a}, \beta)$  is sufficient for calculating the stability  $\beta \Delta G_a^{U \rightarrow F}$ .

## 2.3 Databases of structure-sequence pairs

Consider a dataset of structure-sequence pairs  $D = \{(\chi_i, \mathbf{a}_i)\}_i$  that is sampled from the marginal of some data-generating distribution  $p_D(\chi, \mathbf{a}, \beta)$ . The dataset could, for instance, be the Protein Data Bank (PDB; Berman et al., 2000) or some subset of it. We make the fairly strong **assumption** that all structures in the database are sampled approximately from their respective Boltzmann distribution with potentially different and unknown (latent)  $\beta$ . That is, for any  $\mathbf{a}$  and  $\beta$ , we assume

$$p_D(\chi|\mathbf{a}, \beta) \approx p(\chi|\mathbf{a}, \beta). \quad (11)$$

Note that thermodynamics do not, per se, make any statements about the marginal distribution over the amino sequences. Furthermore, the data marginal over sequences,  $p_D(\mathbf{a})$ , may not reflect the true biological sequence prevalence due to data collection biases, as documented for the PDB (Gerstein, 1998; Orlando et al., 2016).

## 2.4 Inverse folding models

We assume a joint model  $p_\theta(\mathbf{a}, \mathbf{x}, \beta)$  over the sequence  $\mathbf{a}$ , backbone degrees of freedom  $\mathbf{x}$ , and inverse temperature  $\beta$ . Since available datasets typically do not include the inverse temperature, this limits what can be learned in practice. From a dataset of structure-sequence pairs, we learn the conditional distribution  $p_\theta(\mathbf{a}|\mathbf{x})$ , referred to as the *inverse folding model*, as well as the marginal distribution  $p_\theta(\mathbf{a})$  over sequences. For simplicity, we let  $\theta$  denote the combined parameters of all components, although in practice the parameter sets may be disjoint and estimated separately.

Building on the assumption in eq. (11), we further assume that the model provides a good estimate of the conditional distribution over structures given a sequence and inverse temperature, such that

$$p_\theta(\mathbf{x}|\mathbf{a}, \beta) \approx p_D(\mathbf{x}|\mathbf{a}, \beta) \approx p(\mathbf{x}|\mathbf{a}, \beta). \quad (12)$$

This reflects the idea that the learned posterior approximates the Boltzmann distribution well.

<sup>3</sup>This means that  $\mathfrak{X}_a = \mathfrak{X}_a^F \cup \mathfrak{X}_a^U$  and  $\mathfrak{X}_a^F \cap \mathfrak{X}_a^U = \emptyset$ .

### 3 Methods

Consider a wild-type sequence  $\mathbf{a}$  and a variant sequence  $\mathbf{a}'$  that differ by one or more amino acid substitutions. The question is now, can we utilise inverse folding models to calculate the change in stability between the two proteins? The definition of the change in stability is given by the difference in folding free energy between the mutant and wild-type. We can reformulate this as

$$\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} = \underbrace{\ln \frac{p(S=U|\mathbf{a}',\beta)}{p(S=F|\mathbf{a}',\beta)}}_{\beta\Delta G_{\mathbf{a}'\rightarrow\mathbf{F}}^U} - \underbrace{\ln \frac{p(S=U|\mathbf{a},\beta)}{p(S=F|\mathbf{a},\beta)}}_{\beta\Delta G_{\mathbf{a}\rightarrow\mathbf{F}}^U} = \underbrace{\ln \frac{p(S=U|\mathbf{a}',\beta)}{p(S=U|\mathbf{a},\beta)}}_{\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^U} - \underbrace{\ln \frac{p(S=F|\mathbf{a}',\beta)}{p(S=F|\mathbf{a},\beta)}}_{\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^F}, \quad (13)$$

where the second form expresses the mutation effects on the folded and unfolded states in terms of the two *pseudo* change in free energy terms  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^S$ , which we define for analysis but are not physical quantities. In the following, we show how these terms can be estimated using importance sampling.

#### 3.1 Change in stability using inverse folding models

To calculate the pseudo change in free energy  $\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^S$ , we begin by expressing  $p(S|\mathbf{a}',\beta)$  using importance sampling, similarly to free energy perturbation (Zwanzig, 1954). That is,

$$p(S|\mathbf{a}',\beta) = \int p(\mathbf{x}|\mathbf{a}',\beta)p(S|\mathbf{x},\mathbf{a}')d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|S,\mathbf{a},\beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}',\beta)p(S|\mathbf{x},\mathbf{a}')}{p(\mathbf{x}|S,\mathbf{a},\beta)} \right] \quad (14)$$

Using Bayes' theorem, we can express the structure posterior as  $p(\mathbf{x}|S,\mathbf{a},\beta) = p(\mathbf{x}|\mathbf{a},\beta)p(S|\mathbf{x},\mathbf{a})/p(S|\mathbf{a},\beta)$ , which allows us to rewrite the importance sampling expression from eq. (14) as

$$\frac{p(S|\mathbf{a}',\beta)}{p(S|\mathbf{a},\beta)} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|S,\mathbf{a},\beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}',\beta)p(S|\mathbf{x},\mathbf{a}')}{p(\mathbf{x}|\mathbf{a},\beta)p(S|\mathbf{x},\mathbf{a})} \right]. \quad (15)$$

To evaluate the ratio  $\frac{p(\mathbf{x}|\mathbf{a}',\beta)}{p(\mathbf{x}|\mathbf{a},\beta)}$ , we use the approximation assumption from eq. (12) and apply Bayes' theorem in both numerator and denominator, i.e.,

$$\frac{p(\mathbf{x}|\mathbf{a}',\beta)}{p(\mathbf{x}|\mathbf{a},\beta)} \approx \frac{p_\theta(\mathbf{x}|\mathbf{a}',\beta)}{p_\theta(\mathbf{x}|\mathbf{a},\beta)} = \frac{p_\theta(\mathbf{a}'|\mathbf{x},\beta)p_\theta(\mathbf{x}|\beta)}{p_\theta(\mathbf{a}|\mathbf{x},\beta)p_\theta(\mathbf{x}|\beta)} / \frac{p_\theta(\mathbf{a}'|\beta)}{p_\theta(\mathbf{a}|\beta)} = \frac{p_\theta(\mathbf{a}'|\mathbf{x},\beta)}{p_\theta(\mathbf{a}|\mathbf{x},\beta)} \frac{p_\theta(\mathbf{a}|\beta)}{p_\theta(\mathbf{a}'|\beta)}. \quad (16)$$

Substituting this into eq. (15), we arrive at

$$\frac{p(S|\mathbf{a}',\beta)}{p(S|\mathbf{a},\beta)} \approx \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x}|S,\mathbf{a},\beta)} \left[ \frac{p_\theta(\mathbf{a}'|\mathbf{x},\beta)}{p_\theta(\mathbf{a}|\mathbf{x},\beta)} \frac{p(S|\mathbf{x},\mathbf{a}')}{p(S|\mathbf{x},\mathbf{a})} \right] \frac{p_\theta(\mathbf{a}|\beta)}{p_\theta(\mathbf{a}'|\beta)}. \quad (17)$$

**Assumptions** Recall that the inverse folding model introduced in section 2.4 does not depend explicitly on  $\beta$ . To evaluate eq. (17), we assume that the considered  $\beta$  is equally representative of both sequences,  $\mathbf{a}$  and  $\mathbf{a}'$ , and of the structure under our model, i.e.,  $p_\theta(\beta|\mathbf{a}) \approx p_\theta(\beta|\mathbf{a}')$  and  $p_\theta(\beta|\mathbf{x},\mathbf{a}) \approx p_\theta(\beta|\mathbf{x},\mathbf{a}')$ . This assumption may be reasonable when the two sequences differ by only a few substitutions. It then follows that

$$\frac{p_\theta(\mathbf{a}'|\mathbf{x},\beta)}{p_\theta(\mathbf{a}|\mathbf{x},\beta)} \approx \frac{p_\theta(\mathbf{a}'|\mathbf{x})}{p_\theta(\mathbf{a}|\mathbf{x})} \quad \text{and} \quad \frac{p_\theta(\mathbf{a}'|\beta)}{p_\theta(\mathbf{a}|\beta)} \approx \frac{p_\theta(\mathbf{a}')}{p_\theta(\mathbf{a})}. \quad (18)$$

Furthermore, we assume that a small number of sequence substitutions does not alter the backbone-to-state mapping, that is,  $p(S|\mathbf{x},\mathbf{a}) \approx p(S|\mathbf{x},\mathbf{a}')$ , and we can thus assume that the ratio of these terms is 1 in eq. (17). Together, these assumptions yield

$$\frac{p(S|\mathbf{a}',\beta)}{p(S|\mathbf{a},\beta)} \approx \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x}|S,\mathbf{a},\beta)} \left[ \frac{p_\theta(\mathbf{a}'|\mathbf{x})}{p_\theta(\mathbf{a}|\mathbf{x})} \right] \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}')}, \quad (19)$$

providing a means to approximate  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^S$  using an inverse folding model.

Note that it would be mathematically tempting to write the importance sampler in eqs. (14) and (15) using the unconditional Boltzmann distribution  $p(\mathbf{x}|\mathbf{a},\beta)$  as the proposal distribution. However, we note that it is much more difficult to sample from  $p(\mathbf{x}|\mathbf{a},\beta)$  than from the conditional  $p(\mathbf{x}|S,\mathbf{a},\beta)$  as it requires sampling both the folded and unfolded states. Unfolding events are typically rare and difficult to sample in, e.g., MD simulations, and we would need to sample multiple folding-unfolding events to obtain a low-variance estimate. If no unfolded structures are sampled, it would (erroneously) imply that  $p(F|\mathbf{a},\beta) = 1$ , see appendix A.1 for further discussion.

### 3.2 Change in stability from folded and unfolded ensembles

By combining eqs. (13) and (19), we can express the change in stability as

$$\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} \approx \ln \mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{U},\mathbf{a},\beta)} \left[ \frac{p_{\theta}(\mathbf{a}'|\mathbf{x})}{p_{\theta}(\mathbf{a}|\mathbf{x})} \right] - \ln \mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{F},\mathbf{a},\beta)} \left[ \frac{p_{\theta}(\mathbf{a}'|\mathbf{x})}{p_{\theta}(\mathbf{a}|\mathbf{x})} \right], \quad (20)$$

which follows from the important observation that the marginal sequence probabilities  $p_{\theta}(\mathbf{a})$  and  $p_{\theta}(\mathbf{a}')$  cancel between the folded and unfolded terms.

Equation (20) represents a key result: it shows that the change in thermodynamic stability can be estimated using an inverse folding model. The terms inside the expectations can be computed using an inverse folding model, and the full expression becomes tractable through Monte Carlo estimation, provided we can sample structures from the conditional structure distributions  $p_{\theta}(\mathbf{x}|S, \mathbf{a}, \beta)$  for both the unfolded and folded ensembles.

For the folded state,  $p_{\theta}(\mathbf{x}|\mathbf{F}, \mathbf{a}, \beta)$ , we can approximate this distribution using structural data available in the dataset. Usually, the dataset will only contain very few structures for each sequence, and if only a single structure  $\mathbf{x}_{\mathbf{a}}$  is available for sequence  $\mathbf{a}$ , a one-sample approximation yields

$$\mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{F},\mathbf{a},\beta)} \left[ \frac{p_{\theta}(\mathbf{a}'|\mathbf{x})}{p_{\theta}(\mathbf{a}|\mathbf{x})} \right] \approx \frac{p_{\theta}(\mathbf{a}'|\mathbf{x}_{\mathbf{a}})}{p_{\theta}(\mathbf{a}|\mathbf{x}_{\mathbf{a}})}. \quad (21)$$

Presumably, a more accurate estimate could be obtained by sampling local structural variations around  $\mathbf{x}_{\mathbf{a}}$  through molecular simulation. Similarly, simulations could be used to approximate the expectation for the unfolded state. We explore these strategies empirically in section 4.

### 3.3 Change of stability from a folded ensemble or structure

Recall from eq. (10) that knowing  $p(\mathbf{F}|\mathbf{a}, \beta)$  is sufficient for determining the stability of a protein. In this section, we investigate the extent to which we can estimate the change in stability from a folded ensemble alone. We consider two approaches leading to the same expression: in the first case, we assume that  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{U}} \approx 0$ , and in the second case we only consider *ranking* mutations.

#### 3.3.1 Simplified change of stability estimation via unfolded-state invariance

If we consider a stable wild-type sequence, i.e.,  $p(\mathbf{F}|\mathbf{a}, \beta)$  is close to 1, and a variant with a similar or higher folding probability  $p(\mathbf{F}|\mathbf{a}', \beta)$ , then the folded term  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{F}}$  is close to zero. In this case, the dominant contribution to the stability change  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'}$  arises from the unfolded term  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{U}}$ , cf. eq. (20). Thus, for relatively stable variants, the unfolded pseudo-free-energy change,  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{U}}$ , is the key term to evaluate. Conversely, for strongly destabilising mutations, where  $p(\mathbf{F}|\mathbf{a}', \beta) < p(\mathbf{U}|\mathbf{a}, \beta)$ , the dominant term in  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'}$  is the folded term  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{F}}$ . These two cases are illustrated in fig. 3. Consequently, if our main interest lies in quantifying strongly destabilising mutations, we may neglect the unfolded state and approximate the stability change as

$$\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} \approx -\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{F}} \approx -\ln \mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{F},\mathbf{a},\beta)} \left[ \frac{p_{\theta}(\mathbf{a}'|\mathbf{x})}{p_{\theta}(\mathbf{a}|\mathbf{x})} \right] - \ln \frac{p_{\theta}(\mathbf{a})}{p_{\theta}(\mathbf{a}')}, \quad (22)$$

where the second term accounts for the conditional sequence probabilities under the model. If only a single structure  $\mathbf{x}_{\mathbf{a}} \sim p_{\theta}(\mathbf{x}|\mathbf{F}, \mathbf{a})$  is available, we can approximate the expectation with a one-sample estimator

$$\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} \approx -\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\mathbf{F}} \approx -\ln \frac{p_{\theta}(\mathbf{a}'|\mathbf{x}_{\mathbf{a}})}{p_{\theta}(\mathbf{a}|\mathbf{x}_{\mathbf{a}})} - \ln \frac{p_{\theta}(\mathbf{a})}{p_{\theta}(\mathbf{a}')}. \quad (23)$$

The expression in eq. (23) closely resembles standard practice in the field, cf. eq. (1), and thus provides an explanation for zero-shot prediction of inverse-folding models. However, we note that the expression includes an additional correction term that accounts for the frequency of the substituted amino acid under the model (or in the underlying dataset). The fact that the raw log-odds scores work well in practice suggests that this is not a dominating term, but we would expect performance to improve when including it. We investigate this empirically in section 4.

### 3.3.2 Ranking changes of stability

When comparing the stability change of multiple variants  $\mathbf{a}'^{(1)}$  to  $\mathbf{a}'^{(n)}$ , one would ideally compute and compare their respective values  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'^{(i)}}$ . However, note that the term  $\beta\Delta G_{\mathbf{a}}^{\text{U}\rightarrow\text{F}}$  is constant across all variants and thus cancels out when comparing values. As a result, ranking variants by their  $\beta\Delta G_{\mathbf{a}'^{(i)}}^{\text{U}\rightarrow\text{F}}$  values preserves the same ordering. Moreover, since  $p(\text{F}|\mathbf{a},\beta)$  is constant,  $\Delta G_{\mathbf{a}'}^{\text{U}\rightarrow\text{F}}$  is a monotonic function of  $p(\text{F}|\mathbf{a}',\beta)$  and thus also a monotonic function of  $-\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\text{F}}$ . See fig. 3 for an illustration and appendix A.2 for a detailed derivation. Therefore, ranking a set of variants  $\mathbf{a}'^{(1)}, \dots, \mathbf{a}'^{(n)}$  by  $-\beta\Delta\tilde{G}_{\mathbf{a}'^{(i)}\rightarrow\mathbf{a}}^{\text{F}}$  yields the same ordering as ranking them by their full stability changes  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'^{(i)}}$ . This implies that if we are only interested in ranking variants, rather than computing exact stability changes, we can ignore the unfolded ensemble and instead use  $-\beta\Delta\tilde{G}_{\mathbf{a}'^{(i)}\rightarrow\mathbf{a}}^{\text{F}}$ , as given by eqs. (22) and (23).

Importantly, this ranking argument does not rely on the approximation  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^{\text{U}} \approx 0$ , but still leads to the same practical expression. This helps explain why strong Spearman correlations have been observed between the simple log-ratio expression  $-\ln(p_{\theta}(\mathbf{a}'|\mathbf{x}_{\mathbf{a}})/p_{\theta}(\mathbf{a}|\mathbf{x}_{\mathbf{a}}))$  and experimentally measured values of stability changes (Meier et al., 2021), as the Spearman coefficient is a purely rank-based metric.

### 3.4 Change in stability with sequence models

In the previous sections, we derived expressions for estimating the change in thermodynamic stability using inverse folding models. These derivations relied on the ability to sample structural ensembles for both folded and unfolded states. However, in a practical setting, it may be preferable or more convenient to estimate free energy changes using only sequence-based models, without requiring structure-conditioned models.

We assume a joint probabilistic model over amino acid sequences and structural states of the form

$$p_{\gamma}(\mathbf{a}, S) = p_{\gamma}(\mathbf{a} | S)p_{\gamma}(S), \quad (24)$$

where  $\gamma$  denotes the model parameters. Using Bayes' theorem, the pseudo free energy change for a given structural state  $S$  can be expressed as

$$\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^S = \ln \frac{p(S | \mathbf{a}', \beta)}{p(S | \mathbf{a}, \beta)} \approx \ln \frac{p_{\gamma}(S | \mathbf{a}')}{p_{\gamma}(S | \mathbf{a})} = \ln \frac{p_{\gamma}(\mathbf{a}' | S)}{p_{\gamma}(\mathbf{a} | S)} + \ln \frac{p_{\gamma}(\mathbf{a})}{p_{\gamma}(\mathbf{a}')}, \quad (25)$$

where we assume access to both the marginal and conditional probabilities under the model, and that  $p(S | \mathbf{a}, \beta) \approx p_{\gamma}(S | \mathbf{a})$  for all  $\mathbf{a}$ . This means that the change in thermodynamic stability can be estimated using only a state-conditional sequence model; see appendix A.3 for further details and section 4 and table 2 for empirical results.

This framework also enables us to combine estimates from different sources by using a sequence-based model for one state and a structure-based model for the other. A practical special case arises when we have a good characterisation of  $p_{\gamma}(\mathbf{a} | \text{U})$  from data on, e.g., intrinsically disordered proteins or regions, which predominantly represent the unfolded ensemble. By combining this with an inverse folding model for the folded state, and assuming that the marginal sequence probabilities agree across models, i.e.,  $p_{\gamma}(\mathbf{a}) = p_{\theta}(\mathbf{a})$ , the change in stability can be expressed as

$$\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'} \approx \ln \frac{p_{\gamma}(\mathbf{a}' | \text{U})}{p_{\gamma}(\mathbf{a} | \text{U})} - \ln \mathbb{E}_{\mathbf{x} \sim p_{\theta}(\mathbf{x} | \text{F}, \mathbf{a}, \beta)} \left[ \frac{p_{\theta}(\mathbf{a}' | \mathbf{x})}{p_{\theta}(\mathbf{a} | \mathbf{x})} \right]. \quad (26)$$

This hybrid approach is particularly useful when only folded structures are available, as a state-conditional sequence model can provide an estimate of the unfolded pseudo-change in stability. We evaluate this hybrid approach empirically in section 4 and include ablations across different combinations of sequence- and structure-based models in table 2.

## 4 Experiments

To evaluate the effects of the different assumptions and approximations discussed in the previous section, we conclude the paper with a series of experiments in which the individual terms are estimated

	Single sample for folded state	Multiple samples (MD) for folded state
Ignoring unfolded state	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)}$	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F$
Ignoring unfolded state. p(a) correction	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}'')}$	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}'')}$
Unfolded state using Monte Carlo	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$
Unfolded state using minimal-context inverse-folding	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a}_i \mathbf{x}_{a_{i-1:i+1}})}{p_\theta(\mathbf{a}'_i \mathbf{x}_{a_{i-1:i+1}})}$	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a}_i \mathbf{x}_{a_{i-1:i+1}})}{p_\theta(\mathbf{a}'_i \mathbf{x}_{a_{i-1:i+1}})}$
Unfolded state using IDP statistics	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$

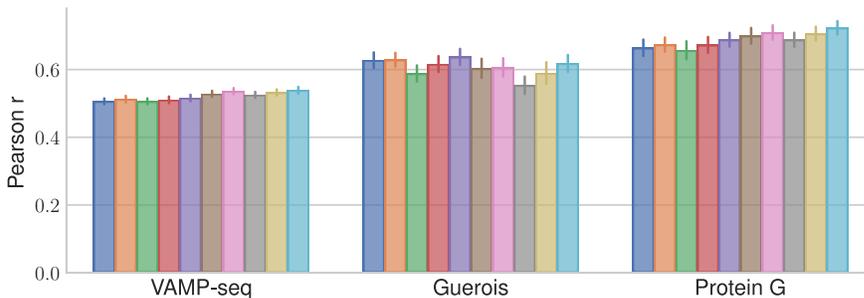


Figure 1: Correlation coefficients obtained using the different expressions discussed in the paper, all involving the inverse-folding model ESM-IF. The top-left expression is the approach typically employed as a zero-shot predictor of protein stability. The left column contains methods that consider only a single folded structure, while the right column considers a structural ensemble from an MD simulation. The different rows represent increasingly accurate approximations to the  $\beta\Delta\Delta G$  (see text for details). The error bars represent the standard error of the mean calculated using 100 bootstrap samples. For simplicity, we used the bracket notation  $\langle \cdot \rangle_S$  to denote the expectation  $\mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}|S, \alpha, \beta)}[\cdot]$ .

using available computational methods on a representative selection of protein datasets. In the following sections, we discuss these choices in turn. We will initially conduct our experiments using the pre-trained ESM inverse folding (ESM-IF) model (Hsu et al., 2022), as it has been shown to perform well in a zero-shot setting (Notin et al., 2023). An ablation with ProteinMPNN is discussed subsequently.

#### 4.1 Data

Our primary analysis considers three different data sets. The first is a high-quality data set measuring the thermodynamic stability of nearly all variants of a single 56-residue protein, the B1 domain of Protein G (hereafter called Protein G; Nisthal et al., 2019). The second is an older benchmark set, compiled for the FoldX prediction method by Guerois et al. (2002), mostly consisting of entries from the ProTherm database (Gromiha et al., 1999). This set also directly provides experimental changes in thermodynamic stability, but is heterogeneous in terms of experimental conditions and is known to be biased towards substitutions in which a large amino acid residue is replaced by a smaller one and, in particular, mutations to Alanine (Stein et al., 2019). Finally, we include data generated using variant abundance by massively parallel sequencing (VAMP-seq) experiments that probe stability only indirectly, by quantifying the variant abundance in cultured cells using a combination of fluorescent tags and sequencing (Matreyek et al., 2018). The data generated by VAMP-seq have previously been shown to correlate with both biophysical measurements (Matreyek et al., 2018) and computational predictions (Cagiada et al., 2021) for protein stability, despite the assay reflecting cellular factors beyond thermodynamic stability. We will refer to these sets as Protein G, Guerois, and VAMP-seq, respectively.

We will conclude the experiments section with a brief discussion of scaling. For this purpose, we employ a subset of the mega-scale stability dataset (Tsuboyama et al., 2023) as included in ProteinGym (Notin et al., 2023). Both VAMP-seq and the mega-scale set represent recent developments in high-throughput assays based on deep sequencing, which are becoming increasingly common for variant characterisation. Combined, the chosen data sets are selected to reflect the different quality/noise regimes in experimental stability data (see appendix B.1 for details).

## 4.2 The folded ensemble

We approximate the expectation over  $p_\theta(\mathbf{x}|\mathbb{F}, \mathbf{a}, \beta)$  using unbiased molecular dynamics (MD) simulations. Using the OpenMM framework (Eastman et al., 2017), 20 ns simulations were conducted at 300 K using 2 femtosecond time steps with the Langevin integrator, combined with the Amber 14 force field with a TIP3P water model, adding counter ions to assure overall neutrality. See appendix B.2 for details on the choice of simulation ensemble.

When considering only the folded state, a sequence correction factor arises in eq. (23). For simplicity, we will in our experiments use a position independent sequence model estimated from  $p_D$ , meaning that for a single mutation amino acid  $i$  the factor becomes  $\ln(p_\theta(\mathbf{a})/p_\theta(\mathbf{a}')) = \ln(p_\theta(\mathbf{a}_i)/p_\theta(\mathbf{a}'_i))$ . We also did the analysis using the ESM2 language model (Lin et al., 2022), but found this choice to be detrimental (see section 4.5 and table 2).

## 4.3 The unfolded ensemble

For the unfolded ensemble, it is less clear what the best approach is, and we therefore try different strategies. In the first approach, we conduct Metropolis-Hastings simulations in the Phaistos framework (Boomsma et al., 2013), using the TorusDBN (Boomsma et al., 2008, 2014) and Basilisk (Harder et al., 2010) statistical models to obtain reasonable backbone and side-chain conformations, but otherwise keep the chain in an unfolded state. We simulated segments with five flanking amino acids on each side of the position of interest, running for 10,000 iterations, where each 100th structure was saved. In the second approach, we again evaluate the ESM-IF model on segments, but unlike the first approach, we now extract a single fixed segment from the crystal structure (i.e., the folded state). The fragment length is kept short (1 flanking amino acid to each side) to approximate an unfolded state, and no structural averaging is done. This approach is similar to that introduced by Dutton et al. (2024), but using segments of length 3 instead of 1. The third approach differs from the first two by not considering the unfolded structural ensemble at all, and instead using a sequence model as detailed in eq. (26). Specifically, we consider protein disorder as a proxy for the unfolded state, and estimate  $p_\gamma(\mathbf{a}|\mathbb{U}_{\text{IPD}})$  using amino acid frequencies obtained from disordered regions according to the ‘curated-disorder-uniprot’ annotation in the MobiDB (Piovesan et al., 2021) database (extracted Jan 21, 2021).

## 4.4 Ablations and scaling

Ablations were performed on the Protein G dataset. First, we repeated all analyses using the ProteinMPNN model (Dauparas et al., 2022). Second, we probed whether the expensive MD simulations could be replaced by predicted ensembles using the BioEmu model (Lewis et al., 2025). After observing promising results, the BioEmu approach was repeated on the larger mega-scale stability dataset (Tsuboyama et al., 2023).

## 4.5 Results

Results for folded/unfolded strategies are reported in fig. 1 and appendix B.1 with additional results on sequence models in table 2. The ProteinMPNN and BioEmu results are found in table 3, while the mega-scale results are displayed in fig. 2.

**Folded state: single-sample vs multi-sample approximation** The left column in the legend in fig. 1 corresponds to methods that use only a single native structure to approximate the expectation, while the right column approximates the ensemble average using multiple sampled structures. For the VAMP-seq and Protein-G dataset, this choice consistently improved performance. On the Guerois dataset, no general trends can be observed, and we observe extensive fluctuations among the 40 different structures in the dataset, reflecting the heterogeneous nature of this older dataset, and perhaps indicating issues with our MD simulations for some of these systems (see also fig. 4).

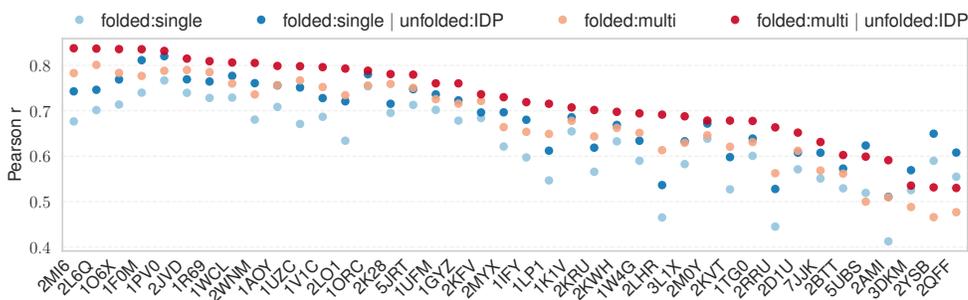


Figure 2: Scaling using BioEmu (Lewis et al., 2025). When replacing MD simulations of the folded state with structural ensembles generated using BioEmu (20 samples; filtered to include only folded states), we still observe consistent improvements over the single-structure performance on this subset of the mega-scale data set (Tsuboyama et al., 2023). This suggests that learned generators of molecular ensembles offer a promising route to scaling our ensemble-based approach in practice.

**Considering only the folded state: the  $p_\theta(\mathbf{a})$  correction** For the folded-only case, the experiments generally show an improvement of including the  $\ln(p_\theta(\mathbf{a})/p_\theta(\mathbf{a}'))$  correction term, but the effect is fairly minor as anticipated. We note that in light of eq. (26), the correction term in eq. (23) can be interpreted as a specific choice of model of the unfolded state, namely the one where  $p_\gamma(\mathbf{a} | \mathbf{U})$  is assumed to factorize over positions and follow the general amino acid propensities. This perspective provides another reason why this simplistic model might not provide very accurate results.

**Three models for the unfolded state** Estimating the contribution from the unfolded state using a Monte Carlo simulation worked less well than expected, generally performing worse than the simple log-odds baseline. One explanation could be that ESM-IF has been trained on structures generated by AlphaFold, and thus has learnt specific geometric features that may not be present in the structures generated by our Monte Carlo simulations. Another explanation could be that the sequence- and local structure signal in ESM-IF dominates when no structural environment is present. Since our Monte Carlo sampler uses a proposal distribution that guarantees native-like local structure, ESM-IF apparently displays folded-like preferences when evaluated on unfolded fragments with native local structure. Replacing the Monte Carlo simulation with a single short structural fragment extracted from the folded state yields some improvements, in line with previous reporting (Dutton et al., 2024). Remarkably, the best approach was to approximate the unfolded state using amino acid frequency statistics from disordered regions. Despite its simplicity and ease of implementation, it generally outperforms the other models of the unfolded state on the considered datasets.

**Ablation and scaling** The ablation results in table 3 confirm earlier reports that the likelihoods from ProteinMPNN are less informative for zero-shot stability prediction than those from ESM-IF (Notin et al., 2023). It has been shown that the autoregressive order of evaluation can improve likelihoods (Dutton et al., 2024), but this was not implemented here. Despite the different baseline performance levels, both models improve similarly across the different strategies. The BioEmu ablation shows only a very minor drop in performance relative to the MD ensembles, suggesting that BioEmu can serve as an efficient proxy when averaging structural ensembles. When applied to a range of proteins from the mega-scale dataset, the BioEmu protocol indeed yields consistent improvements over the baseline at a fraction of the computational cost of a full MD analysis (fig. 2).

## 5 Related work

The idea of connecting free energies to the statistical properties of experimentally determined protein structures has a long history (Tanaka and Scheraga, 1976; Lifson and Levitt, 1979; Sippl, 1990; Miyazawa and Jernigan, 1996; Simons et al., 1999). Before the advent of deep models, these approaches have essentially been based on deriving approximate free-energy stability energies from single and pairwise amino-acid statistics (Borg et al., 2012). Commonly used physico-chemical based modelling tools for calculating  $\Delta\Delta G$  include the FoldX (Schymkowitz et al., 2005) and Rosetta

force-fields (Alford et al., 2017). While providing reasonable estimates for conservative mutations, these force-fields are known to be sensitive to the specific choice of the backbone template in a given application and, as such, do not yield consistent  $\Delta\Delta G$  estimates when applied to the full native ensemble of backbone structures. The connection between  $\Delta\Delta G$  and inverse-folding likelihoods was initially explored by Boomsma and Frellsen (2017) in the context of a 3D convolutional model. This study introduced a correction term compensating for the base frequencies of amino acids similar to eq. (23), but argued for it in terms of the unfolded state, while we show here that it follows more naturally as a consequence of assuming zero contribution from the unfolded state. More recently, a study demonstrated performance gains by including a correction term by evaluating an inverse-model only on the coordinates of the amino acid in question, motivating it as a representation of the unfolded state (Dutton et al., 2024). Our paper provides the theoretical basis for this argument, and we include a very similar strategy in our experiments (using fragments of length 3). Finally, contemporaneously with our work, a recent study on binding affinity prediction reported substantial performance gains by explicitly incorporating the unfolded state (Jiao et al., 2024), using Bayes’ theorem in a similar way as we do in eq. (15), but without considering the full structural ensembles as we do here. Deng et al. (2025) subsequently extended this approach by fine-tuning the inverse folding model on folding stability data.

## 6 Discussion

Log-odds scores from protein inverse-folding models correlate remarkably well with changes in protein stability, but the underlying reasons for this correspondence have remained incompletely understood. In this paper, we take steps to establish a formal connection between the two. We demonstrate that the standard log-odds practice arises as a consequence of a specific set of assumptions, and explore how these assumptions can be relaxed to improve zero-shot prediction further.

Based on our experiments, two choices appear to have the most significant impact over the simple log-odds baseline: 1) including a contribution from the unfolded ensemble, and 2) approximating the structural ensemble of the folded state with more than a single sample. From a practical perspective, both are potentially inconvenient in that they involve molecular simulation. Fortunately, our experiments indicate that the unfolded state can be approximated by a simple static distribution extracted from disordered regions, and we show that computationally-convenient proxies can also be found for the folded ensemble, based on recent generative models of molecular ensembles (Lewis et al., 2025). We therefore anticipate that these improvements can be readily implemented on top of any existing pre-trained free energy model. Finally, we note that while our study has focused on protein stability, it should extend directly to the analysis of binding affinity, generalizing the approach derived by Jiao et al. (2024).

**Limitations** While our derivations are general, the experiments section necessitates choices regarding practical implementations of the individual terms. We believe we have made reasonable choices, but have not exhaustively explored the space of possible models. We consider our experimental section as a proof-of-concept, exemplifying that a better theoretical treatment *can* lead to gains in performance. The relative size of these performance gains will depend on the protein and the models used to approximate the terms, and cannot be conclusively established from our limited set of experiments. Another outstanding issue is that our analysis does not explain the recent observation that inverse-folding likelihoods also correlate surprisingly well with *absolute* stabilities (Cagiada et al., 2025).

**Broader impact** As machine learning models play an increasing role in science, it is important to understand how such models work and interact with existing interpretable models. By establishing a link between pre-trained protein models and the free-energy considerations that drive our physical understanding of protein stability, we hope to make these models more broadly applicable to the scientific community. Although we acknowledge that inverse-folding models can be considered dual-use technologies, we believe any such risks are mitigated by the fact that our work focuses on a theoretical understanding of an existing model class, rather than the development of new predictive capabilities.

**Availability** Code for reproducing the experiments is accessible at: [https://github.com/MachineLearningLifeScience/inverse\\_folding\\_free\\_energies](https://github.com/MachineLearningLifeScience/inverse_folding_free_energies)

## Acknowledgments and Disclosure of Funding

JF and WB are supported by the Center for Basic Machine Learning Research in Life Science (MLLS) through the Novo Nordisk Foundation (NNF20OC0062606), and the Pioneer Centre for AI (DNRF grant P1). KL-L is supported by the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (NNF18OC0033950).

## References

- R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. 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.
- H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.
- W. Boomsma and J. Frellsen. Spherical convolutions and their application in molecular modelling. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3433–3443. Curran Associates, Inc., 2017.
- W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.
- W. Boomsma, J. Frellsen, T. Harder, S. Bottaro, K. E. Johansson, P. Tian, K. Stovgaard, C. Andretta, S. Olsson, J. B. Valentin, et al. Phaistos: A framework for markov chain monte carlo simulation and inference of protein structure. *Journal of computational chemistry*, 34(19):1697–1705, 2013.
- W. Boomsma, P. Tian, J. Frellsen, J. Ferkinghoff-Borg, T. Hamelryck, K. Lindorff-Larsen, and M. Vendruscolo. Equilibrium simulations of proteins using molecular fragment replacement and nmr chemical shifts. *Proceedings of the National Academy of Sciences*, 111(38):13852–13857, 2014.
- M. Borg, T. Hamelryck, and J. Ferkinghoff-Borg. On the physical relevance and statistical interpretation of knowledge-based potentials. In T. Hamelryck, K. Madia, and J. Ferkinghoff-Borg, editors, *Bayesian Methods in Structural Bioinformatics*, volume 1, chapter 3, pages 97–121. Springer-Verlag, Berlin, 2012.
- J. F. Brandts. Conformational transitions of proteins in water and in aqueous mixtures. In S. Timasheff and G. Fasman, editors, *Structure and Stability of Biological Macromolecules*, volume 2, chapter 3, pages 213–290. Marcel Dekker, New York, 1969.
- M. Cagiada, K. E. Johansson, A. Valanciute, S. V. Nielsen, R. Hartmann-Petersen, J. J. Yang, D. M. Fowler, A. Stein, and K. Lindorff-Larsen. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Molecular biology and evolution*, 38(8):3235–3246, 2021.
- M. Cagiada, S. Ovchinnikov, and K. Lindorff-Larsen. Predicting absolute protein folding stability using generative models. *Protein Science*, 34(1):e5233, 2025.
- J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel, et al. Robust deep learning-based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.
- A. Deng, K. D. Householder, F. Wu, K. C. Garcia, and B. L. Trippe. Predicting mutational effects on protein binding from folding energy. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13129–13151. PMLR, 13–19 Jul 2025.

- O. Dutton, S. Bottaro, M. Invernizzi, I. Redl, A. Chung, F. Hoffmann, L. Henderson, S. Ruschetta, F. Airoldi, B. M. Owens, et al. Improving inverse folding models at protein stability prediction without additional training or data. *bioRxiv*, pages 2024–06, 2024.
- P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.
- M. Gerstein. How representative are the known structures of the proteins in a complete genome? a comprehensive structural census. *Folding and Design*, 3(6):497–512, 1998.
- M. M. Gromiha, J. An, H. Kono, M. Oobatake, H. Uedaira, and A. Sarai. Protherm: thermodynamic database for proteins and mutants. *Nucleic acids research*, 27(1):286–288, 1999.
- R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.
- T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. E. Johansson, and T. Hamelryck. Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC bioinformatics*, 11(1): 1–13, 2010.
- C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. Learning inverse folding from millions of predicted structures. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 17–23 Jul 2022.
- X. Jiao, W. Mao, W. Jin, P. Yang, H. Chen, and C. Shen. Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. *arXiv preprint arXiv:2410.09543*, 2024.
- S. Lewis, T. Hempel, J. Jiménez-Luna, M. Gastegger, Y. Xie, A. Y. K. Foong, V. G. Satorras, O. Abdin, B. S. Veeling, I. Zaporozhets, Y. Chen, S. Yang, A. E. Foster, A. Schneuing, J. Nigam, F. Barbero, V. Stimper, A. Campbell, J. Yim, M. Lienen, Y. Shi, S. Zheng, H. Schulz, U. Munir, R. Sordillo, R. Tomioka, C. Clementi, and F. Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025.
- S. Lifson and M. Levitt. On obtaining energy parameters from crystal structure data. *Comput Chem.*, 3:49–50, 1979.
- Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.
- K. Lindorff-Larsen and K. Teilum. Linking thermodynamics and measurements of protein stability. *Protein Engineering, Design and Selection*, 34:gzab002, 03 2021.
- D. Listov, C. A. Goverde, B. E. Correia, and S. J. Fleishman. Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology*, 25(8):639–653, 2024.
- B. J. Livesey and J. A. Marsh. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular systems biology*, 16(7):e9380, 2020.
- K. A. Matreyek, L. M. Starita, J. J. Stephany, B. Martin, M. A. Chiasson, V. E. Gray, M. Kircher, A. Khechaduri, J. N. Dines, R. J. Hause, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics*, 50(6):874–882, 2018.
- J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021.

- S. Miyazawa and R. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.
- A. Nisthal, C. Y. Wang, M. L. Ary, and S. L. Mayo. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proceedings of the National Academy of Sciences*, 116(33):16367–16377, 2019.
- P. Notin, A. Kollasch, D. Ritter, L. van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, Y. Gal, and D. Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64331–64379, 2023.
- G. Orlando, D. Raimondi, and W. F. Vranken. Observation selection bias in contact prediction and its implications for structural bioinformatics. *Scientific Reports*, 6(1):36679, 2016.
- D. Piovesan, M. Necci, N. Escobedo, A. M. Monzon, A. Hatos, I. Mičetić, F. Quaglia, L. Paladin, P. Ramasamy, Z. Dosztányi, et al. Mobidb: intrinsically disordered proteins in 2021. *Nucleic acids research*, 49(D1):D361–D367, 2021.
- J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The foldx web server: an online force field. *Nucleic acids research.*, 33:W382–W388, 2005.
- K. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34:82–95, 1999.
- M. Sippl. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.
- L. M. Starita, N. Ahituv, M. J. Dunham, J. O. Kitzman, F. P. Roth, G. Seelig, J. Shendure, and D. M. Fowler. Variant interpretation: functional assays to the rescue. *The American Journal of Human Genetics*, 101(3):315–325, 2017.
- A. Stein, D. M. Fowler, R. Hartmann-Petersen, and K. Lindorff-Larsen. Biophysical and mechanistic models for disease-causing protein variants. *Trends in biochemical sciences*, 44(7):575–588, 2019.
- S. Tanaka and H. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9:945–950, 1976.
- K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, and G. J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.
- R. W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 08 1954.

## A Details on evaluating the change in stability using inverse folding models

### A.1 Using the unconditional Boltzmann distribution as the proposal

In eq. (15), we use the conditional Boltzmann distribution  $p(\mathbf{x}|S, \mathbf{a}, \beta)$  as the proposal, since it is easier to sample from than  $p(\mathbf{x}|\mathbf{a}, \beta)$ . Here, we will investigate the unconditional proposal distribution.

We can write the importance sampler from eq. (14) using the unconditional Boltzmann distribution as

$$p(S|\mathbf{a}', \beta) = \int p(\mathbf{x}|\mathbf{a}')p(S|\mathbf{x}, \mathbf{a}', \beta) d\mathbf{x} = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{a}, \beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}', \beta)p(S|\mathbf{x}, \mathbf{a}')}{p(\mathbf{x}|\mathbf{a}, \beta)} \right]. \quad (27)$$

In the extreme case, when we try to sample from  $p(\mathbf{x}|\mathbf{a}, \beta)$ , we efficiently only sample the folded state, i.e., we use  $p(\mathbf{x}|\mathbf{F}, \mathbf{a}, \beta)$  as the proposal distribution. This corresponds to assuming that  $p(S|\mathbf{a}', \beta) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{F}, \mathbf{a}, \beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}', \beta)p(S|\mathbf{x}, \mathbf{a}')}{p(\mathbf{x}|\mathbf{a}, \beta)} \right]$ . Using Bayes' rule on this assumption gives us that

$$p(S|\mathbf{a}', \beta) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{a}, \beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}', \beta)p(S|\mathbf{x}, \mathbf{a}')}{p(\mathbf{x}|\mathbf{a}, \beta)} \frac{p(\mathbf{F}|\mathbf{x}, \mathbf{a})}{p(\mathbf{F}|\mathbf{a}, \beta)} \right] \quad (28)$$

$$= \frac{1}{p(\mathbf{F}|\mathbf{a}, \beta)} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{a}, \beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}', \beta)p(S|\mathbf{x}, \mathbf{a}')}{p(\mathbf{x}|\mathbf{a}, \beta)} p(\mathbf{F}|\mathbf{x}, \mathbf{a}) \right] \quad (29)$$

$$\leq \frac{1}{p(\mathbf{F}|\mathbf{a}, \beta)} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{a}, \beta)} \left[ \frac{p(\mathbf{x}|\mathbf{a}', \beta)p(S|\mathbf{x}, \mathbf{a}')}{p(\mathbf{x}|\mathbf{a}, \beta)} \right] = \frac{p(S|\mathbf{a}')}{p(\mathbf{F}|\mathbf{a})}, \quad (30)$$

which is only true for  $p(\mathbf{F}|\mathbf{a}, \beta) = 1$ . So the assumption implies that  $p(\mathbf{F}|\mathbf{a}, \beta) = 1$ .

### A.2 Monotonicity of stability

We aim to show that  $\Delta G_{\mathbf{a}' \rightarrow \mathbf{F}}^{\mathbf{U} \rightarrow \mathbf{F}}$  is a monotone increasing function of  $-\beta \Delta \tilde{G}_{\mathbf{a}' \rightarrow \mathbf{a}}^{\mathbf{F}}$ . In the following, all distributions are conditioned on  $\beta$ , but we omit it to simplify notation. Starting from eq. (10), we can write

$$\beta \Delta G_{\mathbf{a}' \rightarrow \mathbf{F}}^{\mathbf{U} \rightarrow \mathbf{F}} = \ln \left( \frac{1}{p(S=\mathbf{F}|\mathbf{a}')} - 1 \right) \quad (31)$$

$$= \ln \left( \frac{1}{\frac{p(S=\mathbf{F}|\mathbf{a})}{p(S=\mathbf{F}|\mathbf{a}')} p(S=\mathbf{F}|\mathbf{a}')} - 1 \right) \quad (32)$$

$$= \ln \left( \frac{1}{p(S=\mathbf{F}|\mathbf{a}) \exp \left( \ln \frac{p(S=\mathbf{F}|\mathbf{a}')}{p(S=\mathbf{F}|\mathbf{a})} \right)} - 1 \right) \quad (33)$$

$$= \ln \left( \frac{\exp \left( -\ln \frac{p(S=\mathbf{F}|\mathbf{a}')}{p(S=\mathbf{F}|\mathbf{a})} \right)}{p(S=\mathbf{F}|\mathbf{a})} - 1 \right) \quad (34)$$

$$= \ln \left( \frac{\exp(y)}{p(S=\mathbf{F}|\mathbf{a})} - 1 \right) =: f(y) \quad (35)$$

where  $y = -\beta \Delta \tilde{G}_{\mathbf{a}' \rightarrow \mathbf{a}}^{\mathbf{F}} = -\ln \frac{p(S=\mathbf{F}|\mathbf{a}')}{p(S=\mathbf{F}|\mathbf{a})}$ . We note that  $y \mapsto \exp(y)$  is monotonically increasing and  $z \mapsto \ln(z/c - 1)$  is monotonically increasing for  $c > 0$ . Since  $p(S=\mathbf{F}|\mathbf{a})$  is a positive constant and function composition preserves monotonicity, we have that  $f(y)$  is a monotonically increasing function of  $y$ .

### A.3 Details on using sequence models for change in stability

By combining eq. (13) with eq. (25), the change in thermodynamic stability can be approximated as

$$\beta \Delta \Delta G_{\mathbf{a} \rightarrow \mathbf{a}'} \approx \ln \frac{p_\gamma(\mathbf{a}'|\mathbf{U})}{p_\gamma(\mathbf{a}|\mathbf{U})} - \ln \frac{p_\gamma(\mathbf{a}'|\mathbf{F})}{p_\gamma(\mathbf{a}|\mathbf{F})}, \quad (36)$$

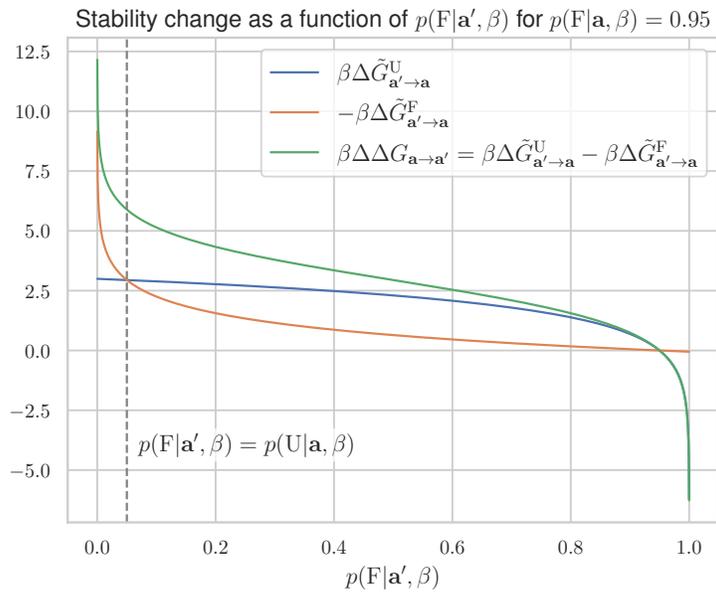


Figure 3: Stability change calculated from the two pseudo-free-energy terms,  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^S$ , plotted as a function of  $p(\mathbf{F}|\mathbf{a}',\beta)$  for  $p(\mathbf{F}|\mathbf{a},\beta) = 0.95$ . As discussed in section 3.3.1, we observe that when  $p(\mathbf{F}|\mathbf{a}',\beta)$  is around 0.95, the unfolded-state term,  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^U$ , provides the dominant contribution to the stability change,  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'}$ , whereas for  $p(\mathbf{F}|\mathbf{a}',\beta) < p(\mathbf{U}|\mathbf{a},\beta)$ , the folded-state term,  $\beta\Delta\tilde{G}_{\mathbf{a}'\rightarrow\mathbf{a}}^F$ , dominates. We also observe that the stability change,  $\beta\Delta\Delta G_{\mathbf{a}\rightarrow\mathbf{a}'}$ , is a monotonic function of the variant’s folding probability,  $p(\mathbf{F}|\mathbf{a}',\beta)$ , as discussed in section 3.3.2 and appendix A.2.

where the marginal sequence probabilities  $p_\gamma(\mathbf{a})$  and  $p_\gamma(\mathbf{a}')$  cancel between the folded and unfolded terms, analogous to the cancellation in eq. (20). This expression shows that we can estimate the change in stability using only a sequence model that provides conditional likelihoods given the structural state, i.e., a model capable of computing  $p_\gamma(\mathbf{a} | S)$  for  $S \in \{\mathbf{F}, \mathbf{U}\}$ .

## B Experimental details

### B.1 Dataset preprocessing details

The Guerois data set (Guerois et al., 2002) contains 988 entries that we filtered to contain only single amino acid substitutions (i.e. no double and triple substitutions). 911 entries remained after filtering and are associated to 40 PDB structures.

The Protein G data set (Nisthal et al., 2019) contains 907 entries associated with a single PDB (PDBID:1PGA). We used the values labelled as “ddG(mAvg)\_mean” which are associated with the lowest median uncertainty reported to be 0.1 kcal/mol. For 107 of the very destabilising entries, only a lower bound of 4.0 kcal/mol were reported. Note that these values are the  $\Delta\Delta G$  of unfolding, therefore, we inverted the sign to obtain  $\Delta\Delta G$  values for folding.

The VAMP-seq data (Matreyek et al., 2018) contains 8096 entries for two proteins, TPMT and PTEN (associated with two structures with PDBID: 2H11 and 1D5R, respectively). After filtering to include only amino acid residues that are resolved in the protein structures, 6909 entries remain. We use the values labelled as “score” with a negative sign.

For our last experiment, we considered a subset of the mega-scale thermodynamic folding stability dataset (Tsuboyama et al., 2023), with a total of 42253 variants across 40 proteins (PDB-IDs can be found in Figure 2). The experimental values were retrieved through the ProteinGym (Notin et al., 2023) interface without further processing. The structural ensembles for the wild-type variants of these proteins were retrieved from the supporting material of the BioEmu paper (Lewis et al., 2025).

## B.2 Choice of simulation ensemble

For the molecular dynamics simulations used in our study, we initially conducted simulations in an NVT ensemble. Over the course of the study, we refined this protocol and switched to an NPT simulation setup, which we used for the simulations for the Protein G, TPMT and PTEN in the VAMP-seq and Protein G datasets. Since the change in protocol turned out to have very minor effect on the prediction accuracy we decided not to rerun the simulations for the 40 proteins in the Guerois data set.

## B.3 BioEmu samples

The BioEmu model produces both folded and unfolded conformations. Since we only use BioEmu as a proxy for the folded state in our analysis, we filter out conformations with a fraction of native contacts (FNC) below 0.5. Since BioEmu aims to sample states proportionally to their Boltzmann weights, and the considered proteins are stable, less than 1% of samples were discarded by this procedure.

## B.4 Resources

The experiments in this paper comprise running Monte Carlo and molecular dynamics simulations for 40 proteins, in addition to model evaluation of pretrained models on all samples. Since no training was involved, no large scale GPU-resources were necessary for this study.

## C Licenses and references for used assets

Below we list the external software and dataset assets used in this study, along with their licenses and access information:

**ESM-IF** The ESM-IF inverse folding model (Hsu et al., 2022) is released under the MIT license. Source code and pretrained models are available at: <https://github.com/facebookresearch/esm>

**Phaistos Framework** The Phaistos framework (Boomsma et al., 2013), used for Monte Carlo simulations of unfolded structures, is available under the LGPLv2 or GPLv3 license. The source code is available at: <https://sourceforge.net/projects/phaistos>

**OpenMM** The OpenMM molecular dynamics engine (Eastman et al., 2017) was used for MD simulations. It is released under a mix of licenses, including MIT, LGPLv3, CC BY 3.0 and several other licenses for specific parts (see details at <https://github.com/openmm/openmm/blob/master/docs-source/licenses/Licenses.txt>) and available at: <https://github.com/openmm/openmm>

**Protein G Dataset** The thermodynamic stability measurements for the B1 domain of Protein G by Nisthal et al. (2019) is available with no license from: [https://www.protabank.org/study\\_analysis/3xESLyS9/](https://www.protabank.org/study_analysis/3xESLyS9/)

**Guerois Benchmark Set** This benchmark set of experimental  $\Delta\Delta G$  values was compiled by Guerois et al. (2002), with data sourced from the ProTherm database (Gromiha et al., 1999). The dataset is available through the ProTherm database at: <https://web.iitm.ac.in/bioinfo2/prothermdb>

**VAMP-seq Dataset** Stability-related data for the TPMT and PTEN proteins were taken from Matreyek et al. (2018) and are available for non-profit, non-commercial use at: <https://github.com/FowlerLab/VAMPseq>

**MobiDB Disorder Statistics** Amino acid frequencies for disordered regions were extracted from the ‘curated-disorder-uniprot’ entries in the MobiDB database (Piovesan et al., 2021). The database is available under the CC BY 4.0 licence at: <https://mobidb.bio.unipd.it>

**Mega-scale thermodynamic folding stability dataset** The mega-scale thermodynamic folding stability dataset (Tsuboyama et al., 2023) was accessed through the ProteinGym interface (Notin et al., 2023). It is licensed under the MIT licence.

Table 1: Tabular overview of the Pearson correlation coefficients presented in the fig. 1, where the numbers in parentheses are the standard error of the mean.

Strategy		Guerois	Protein G	VAMP-seq
<i>folded-only, single-sample</i>	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)}$	0.63 (0.02)	0.66 (0.02)	0.51 (0.01)
<i>folded-only, <math>p_\theta(\mathbf{a})</math> correction</i>	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}')}$	0.63 (0.02)	0.67 (0.02)	0.51 (0.01)
<i>folded single-sample, unfolded multi-sample (MC)</i>	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$	0.59 (0.02)	0.66 (0.02)	0.51 (0.01)
<i>folded single-sample, inv-fold single-aa as unfolded</i>	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a}_1 \mathbf{x}_{\mathbf{a}_1-1:i+1})}{p_\theta(\mathbf{a}'_1 \mathbf{x}_{\mathbf{a}'_1-1:i+1})}$	0.62 (0.02)	0.67 (0.02)	0.51 (0.01)
<i>folded single-sample, IDP aa-stats as unfolded</i>	$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$	0.64 (0.02)	0.69 (0.02)	0.52 (0.01)
<i>folded-only, multi-sample</i>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F$	0.6 (0.03)	0.7 (0.03)	0.53 (0.01)
<i>folded only, multi-sample (MD), <math>p_\theta(\mathbf{a})</math> correction</i>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}')}$	0.61 (0.03)	0.71 (0.02)	0.54 (0.01)
<i>folded multi-sample (MD), unfolded multi-sample (MC)</i>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$	0.55 (0.03)	0.69 (0.02)	0.53 (0.01)
<i>folded multi-sample, inv-fold single-aa as unfolded</i>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a}_1 \mathbf{x}_{\mathbf{a}_1-1:i+1})}{p_\theta(\mathbf{a}'_1 \mathbf{x}_{\mathbf{a}'_1-1:i+1})}$	0.59 (0.02)	0.71 (0.02)	0.53 (0.01)
<i>folded multi-sample (MD), IDP aa-stats as unfolded</i>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$	0.62 (0.03)	0.72 (0.02)	0.54 (0.01)

$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)}$ <p><i>folded-only, single-sample</i></p>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F$ <p><i>folded-only, multi-sample</i></p>
$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}')}$ <p><i>folded-only, <math>p_\theta(\mathbf{a})</math> correction</i></p>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a})}{p_\theta(\mathbf{a}')}$ <p><i>folded only, multi-sample (MD), <math>p_\theta(\mathbf{a})</math> correction</i></p>
$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$ <p><i>folded single-sample, unfolded multi-sample (MC)</i></p>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \left\langle \frac{p_\theta(\mathbf{a} \mathbf{x}_a)}{p_\theta(\mathbf{a}' \mathbf{x}_a)} \right\rangle_U$ <p><i>folded multi-sample (MD), unfolded multi-sample (MC)</i></p>
$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a}_i \mathbf{x}_{a_{i-1:i+1}})}{p_\theta(\mathbf{a}'_i \mathbf{x}_{a_{i-1:i+1}})}$ <p><i>folded single-sample, inv-fold single-aa as unfolded</i></p>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a}_i \mathbf{x}_{a_{i-1:i+1}})}{p_\theta(\mathbf{a}'_i \mathbf{x}_{a_{i-1:i+1}})}$ <p><i>folded multi-sample (MD), inv-fold single-aa as unfolded</i></p>
$-\ln \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$ <p><i>folded single-sample, IDP aa-stats as unfolded</i></p>	$-\ln \left\langle \frac{p_\theta(\mathbf{a}' \mathbf{x}_a)}{p_\theta(\mathbf{a} \mathbf{x}_a)} \right\rangle_F - \ln \frac{p_\theta(\mathbf{a} \mathbf{U}_{IDP})}{p_\theta(\mathbf{a}' \mathbf{U}_{IDP})}$ <p><i>folded multi-sample (MD), IDP aa-stats as unfolded</i></p>

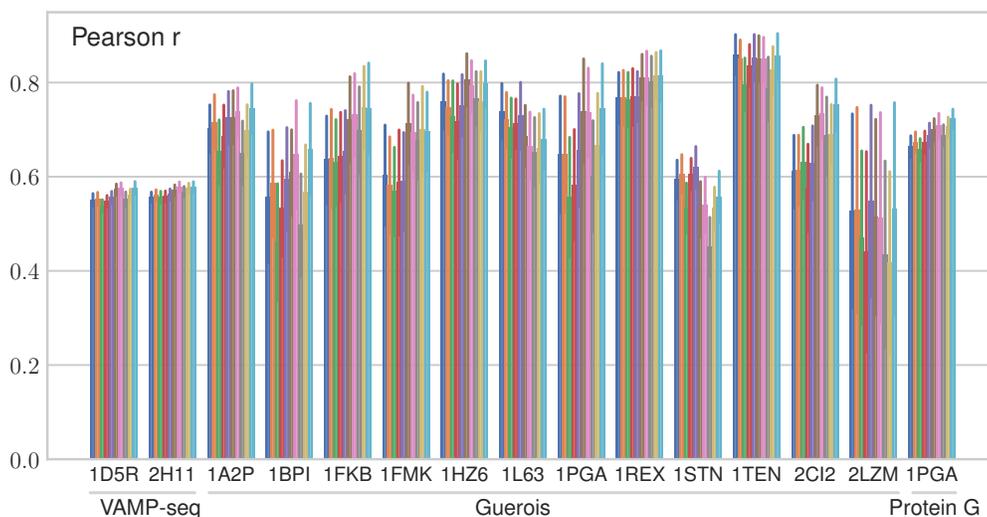


Figure 4: A breakdown of the performance for the individual proteins within the three original datasets. Since correlations are computed, only proteins with at least 20 variant observations are included. The top-left variant is the approach typically employed as zero-shot predictor for protein stability prediction. The left column are methods based that consider only a single folded structure, while the right column considers a structural ensemble from an MD simulation. Note the considerable variation among the proteins in the Guerois set.

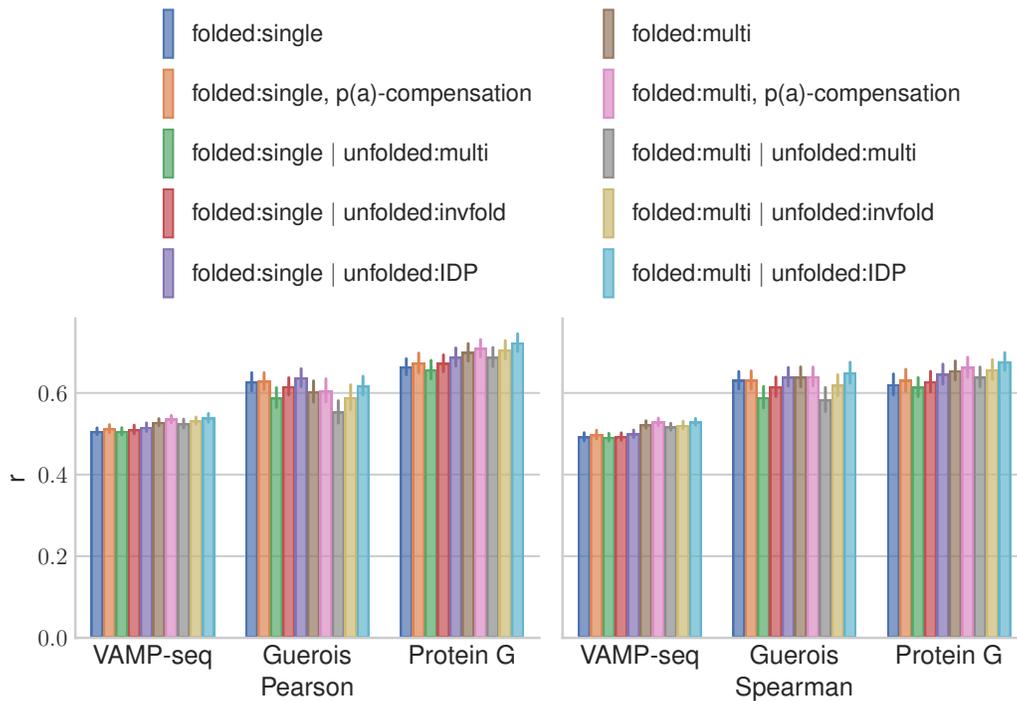


Figure 5: Pearson and Spearman correlations behave similarly. As expected from our derivations, the relationship between zero-shot scores and stability is linear, and employing a rank-based procedure like Spearman rho is therefore not necessary.

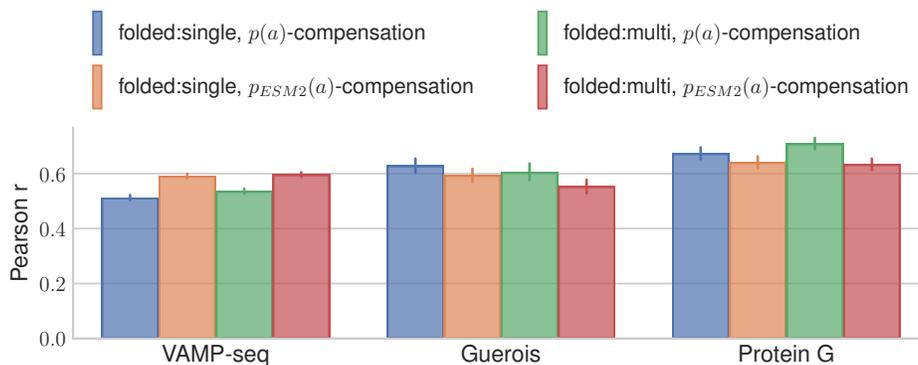


Figure 6: Compensating with a richer model for  $p(a)$ . A comparison between using the simple site-independent  $p(a)$  and a full language model, ESM2 (Lin et al., 2022). The language model does not provide a benefit over the simpler model. One potential explanation could be that the ESM2 model itself captures a considerable structural signal, as illustrated by its use for structure prediction in ESMFold (Lin et al., 2022).

Table 2: Pearson correlation coefficients for different strategies, focusing on the sequence-only strategies (as compensation terms or as a model for the folded state). The first 10 rows in this table are also present in appendix B.1, but are repeated here for ease of comparison. Numbers in parentheses represent standard errors of the mean.

Strategy		Guerois	Protein G	VAMP-seq
folded:single		0.63 (0.02)	0.66 (0.02)	0.51 (0.01)
	unfolded:multi	0.59 (0.02)	0.66 (0.02)	0.51 (0.01)
	unfolded:invfold	0.62 (0.03)	0.67 (0.02)	0.51 (0.01)
	unfolded:IDP	0.64 (0.02)	0.69 (0.02)	0.52 (0.01)
folded:multi		0.60 (0.02)	0.70 (0.02)	0.53 (0.01)
	unfolded:multi	0.55 (0.03)	0.69 (0.02)	0.53 (0.01)
	unfolded:invfold	0.59 (0.03)	0.71 (0.02)	0.53 (0.01)
	unfolded:IDP	0.62 (0.03)	0.72 (0.02)	0.54 (0.01)
folded:single, $p(a)$ -compensation		0.63 (0.02)	0.67 (0.02)	0.51 (0.01)
folded:multi, $p(a)$ -compensation		0.61 (0.03)	0.71 (0.02)	0.54 (0.01)
folded:single, $p_{ESM2}(a)$ -compensation		0.59 (0.03)	0.64 (0.02)	0.59 (0.01)
folded:multi, $p_{ESM2}(a)$ -compensation		0.55 (0.03)	0.63 (0.02)	0.60 (0.01)
folded: $p(a)$		0.05 (0.04)	0.12 (0.04)	0.09 (0.01)
	unfolded:multi	-0.04 (0.03)	-0.04 (0.04)	0.05 (0.01)
	unfolded:invfold	-0.08 (0.04)	-0.01 (0.04)	0.01 (0.01)
	unfolded:IDP	0.14 (0.04)	0.21 (0.03)	0.12 (0.01)
folded: $p_{ESM2}(a)$		0.38 (0.04)	0.41 (0.03)	0.49 (0.01)
	unfolded:multi	0.34 (0.03)	0.37 (0.03)	0.48 (0.01)
	unfolded:invfold	0.35 (0.03)	0.39 (0.03)	0.46 (0.01)
	unfolded:IDP	0.4 (0.03)	0.45 (0.03)	0.50 (0.01)

Table 3: Ablation on the Protein G dataset of 1) the inverse-folding model employed and 2) the procedure for generating structural ensembles. As has been observed previously (Notin et al., 2023), ProteinMPNN generally produces lower correlation scores than ESM-IF, but displays a relatively larger performance boost with the more advanced strategies. Remarkably, the MD and BioEmu results are almost identical on these datasets, suggesting that BioEmu could provide a reasonable approximation to the expensive MD structural ensembles. Values are Pearson correlation scores and standard errors of the mean are provided in parenthesis.

Strategy	Ensemble type	ESM-IF	MPNN
folded:single		0.66 (0.02)	0.4 (0.03)
folded:single, $p(a)$ -compensation		0.67 (0.02)	0.43 (0.03)
folded:single   unfolded:multi		0.66 (0.02)	0.39 (0.03)
folded:single   unfolded:invfold		0.67 (0.02)	0.35 (0.03)
folded:single   unfolded:IDP		0.69 (0.02)	0.48 (0.02)
folded:multi	MD	0.70 (0.02)	0.50 (0.02)
	BioEmu	0.69 (0.02)	
folded:multi, $p(a)$ -compensation	MD	0.71 (0.02)	0.53 (0.02)
	BioEmu	0.70 (0.02)	
folded:multi   unfolded:multi	MD	0.69 (0.02)	0.49 (0.02)
	BioEmu	0.68 (0.02)	
folded:multi   unfolded:invfold	MD	0.71 (0.02)	0.43 (0.03)
	BioEmu	0.70 (0.02)	
folded:multi   unfolded:IDP	MD	0.72 (0.02)	0.58 (0.02)
	BioEmu	0.72 (0.02)	

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly state the paper's aims: providing a formal connection between inverse folding model probabilities and thermodynamic stability, and demonstrating how to improve prediction accuracy through refined estimators. The claims are well supported by the theory and experiments presented.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper includes a dedicated limitations paragraph in the Discussion section (6), covering both modelling assumptions (e.g., dataset approximates Boltzmann distributions) and practical limitations (e.g., choices in implementation, scope of experiments).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper does not include formal theorems but provides complete derivations with clearly stated assumptions (e.g., dataset approximates Boltzmann distribution) throughout the main text and appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient information to reproduce the experiments, including dataset sources, simulation setups (e.g., MD protocols), the version of the pretrained ESM-IF model, a description of model evaluation, and detailed formulas for the calculations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The authors commit to releasing the code and all simulated data with open access upon acceptance. All other datasets and models used are publicly available and properly cited.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper and appendix specifies dataset filtering, simulation parameters (e.g., temperature, time steps, force fields), and model evaluation methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports correlations with standard errors of the mean, and provides a detailed breakdown across proteins in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The Discussion section (6) includes a paragraph on resources used, including the absence of large scale GPU training.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper adheres to the NeurIPS Code of Ethics. There are no ethical concerns raised by the datasets or methods used.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Discussion section (6) includes a “Broader impact” paragraph, addressing both positive (scientific interpretability) and potential negative (dual-use) consequences of inverse folding models.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release models or data posing a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and tools (e.g., PDB, OpenMM, ESM-IF) are properly credited and referenced. The licenses of used assets are listed in appendix C.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper will release code and datasets used for reproducibility, and the accompanying code is well documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.