Sim2Real Transfer Learning for Nanobody Thermal Stability Prediction

◯Kentarou Sasaki[1], Taihei Murakami[1,2] , Yasuhiro Matsunaga[1,3]
[1]Saitama University, [2]Epsilon Molecular Engineering, [3]RIKEN R-CCS

ABSTRACT

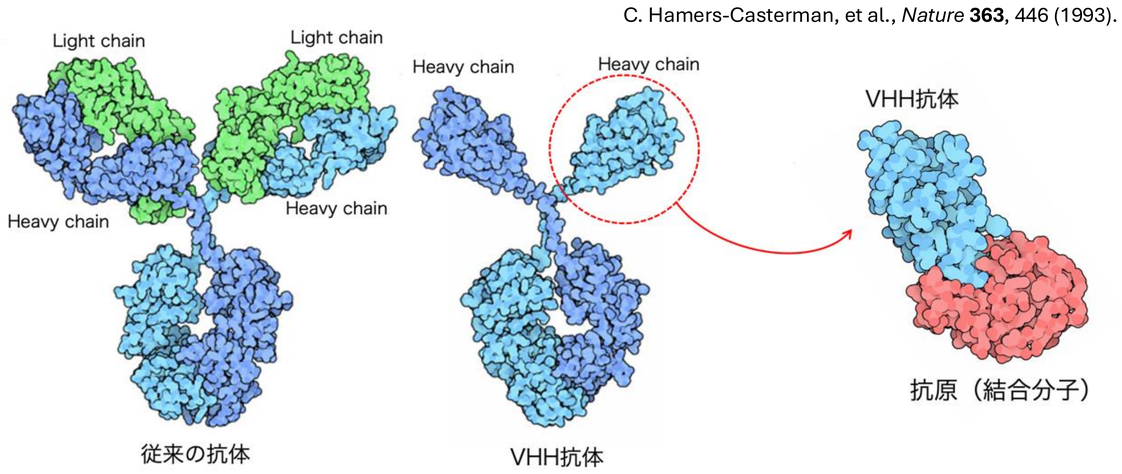
VHH antibodies (nanobodies) are small and easy to engineer, making them promising therapeutic candidates. However, their thermal stability is sequence-dependent, and experimental measurement of the melting temperature (T_m) is costly. While molecular dynamics simulations can compute stability-related quantities such as ΔΔG, these calculations are computationally demanding and require significant HPC resources. Sim2Real transfer learning, which leverages large-scale simulation data to enhance predictions on limited experimental data, has proven effective in materials science. In this work, we apply Sim2Real transfer learning to nanobody thermal stability prediction. Since experimental T_m values and simulation-derived ΔΔG values cannot be directly combined, we propose a multitask learning approach with shared representations to bridge the gap between simulation and real-world data.

BACKGROUND

VHH antibodies and Current Challenges

- VHH antibodies are small and easy to engineer, making them attractive candidates for therapeutic applications.
- Their thermal stability is highly dependent on amino acid sequence, and experimental evaluation of stability, typically via measurement of the melting temperature (T_m), is costly and time-consuming.

C. Hamers-Casterman, et al., *Nature* **363**, 446 (1993).



Machine Learning and Sim2Real

S. Minami, et al., *Npj Comput Mater* **11**, 146 (2025).

- Machine learning–based methods have been developed to predict protein stability; however, their predictive performance is often limited by the scarcity of experimental data.
- In this study, we aim to improve prediction accuracy by leveraging simulation-derived data through Sim2Real transfer learning.
- In addition, we quantitatively evaluate how simulation-derived data contribute to predictive performance relative to experimental data, assessing the extent to which simulation samples can compensate for limited experimental measurements.

METHOD

ΔG

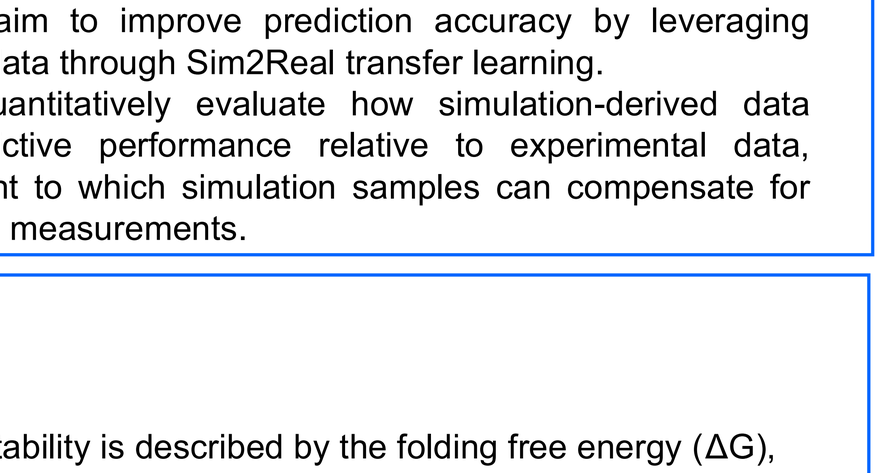
- Protein structural stability is described by the folding free energy (ΔG), defined as the free energy difference between the folded and unfolded states; however, direct computation of ΔG is generally difficult because it requires extensive sampling of a vast conformational space.

ΔΔG

- ΔΔG corresponds to the difference between the folding free energy (ΔG) of a mutant protein and that of the wild-type sequence, representing a relative change in structural stability induced by mutation.
  Because ΔΔG can be computationally evaluated, it is adopted as the target variable for training data in this study.

MD-FEP

- Physics-based approach using molecular dynamics simulations and free energy perturbation
- High accuracy but computationally expensive



Rosetta

- Structure-based approach using Monte Carlo sampling with an empirical energy function
- Computationally efficient but less accurate than physics-based methods

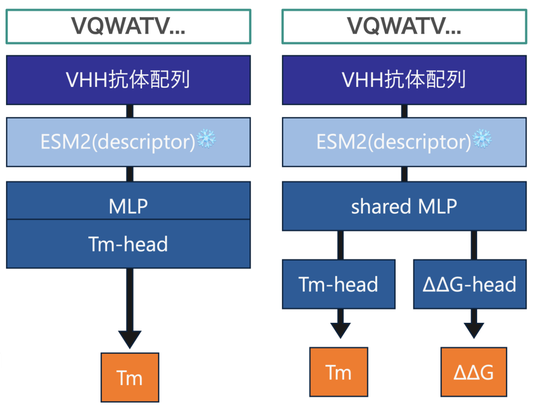
Multitask learning framework

S. Minami, et al., *Npj Comput Mater* **11**, 146 (2025).

To enable Sim2Real transfer, the model was trained within a multitask learning framework. The architecture consists of a shared MLP backbone that learns a common latent representation across tasks, along with task-specific heads for predicting experimental melting temperatures (T_m) and simulation-derived ΔΔG values.

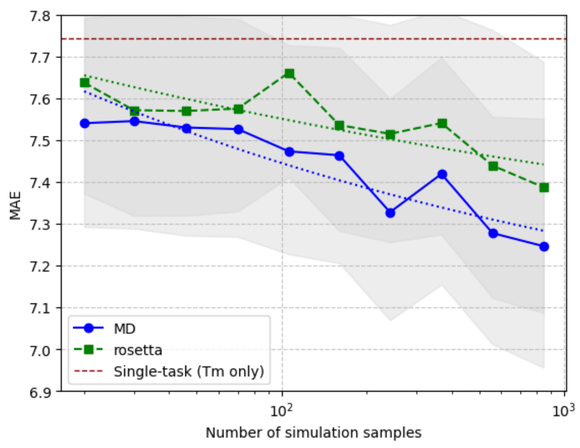
Computational details

- To evaluate the effect of data scale, we varied the number of simulation samples from 10 to 846 while fixing the number of experimental T_m samples at 56.
- Conversely, we varied the number of experimental T_m samples from 10 to 567 while fixing the number of simulation samples at 846, and conducted validation under these settings.



RESULTS

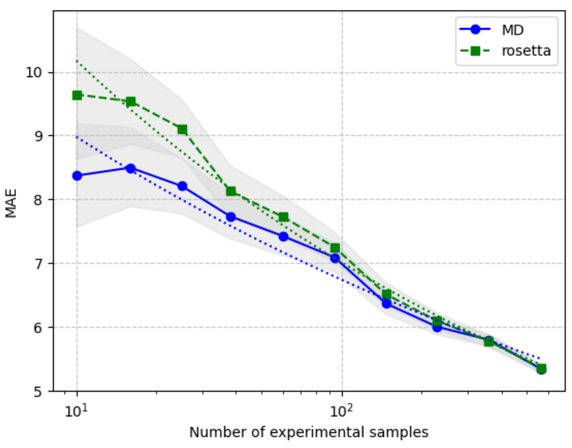
Scaling for simulation data



- Multitask learning consistently outperformed the single-task model trained only on experimental melting temperature (T_m) data.
- Furthermore, a comparison between ΔΔG values computed by MD-FEP and Rosetta showed that using MD-FEP–derived ΔΔG led to improved prediction accuracy. The corresponding approximation is shown below.

$$y = 0.326662\ n^{-0.183735} + 6.946006$$

Scaling for experimental data



- When scaling analysis was performed based on experimental data, a stronger scaling behavior was observed compared to that based on simulation data.
- Furthermore, a comparison between ΔΔG values computed by MD-FEP and Rosetta showed that using MD-FEP–derived ΔΔG was more effective in improving prediction accuracy.

$$y = 3.819658\ n^{-0.150731} + 1.331720$$

The marginal rate of substitution is estimated based on the scaling results obtained from simulation and experimental data.

$$\left.\frac{dy}{dn}\right|_{n=846} \approx -2.034 \times 10^{-5} \qquad \left.\frac{dy}{dm}\right|_{m=56} \approx -5.630 \times 10^{-3}$$

$$\left.\frac{dn}{dm}\right|_{(n,m)=(846,56)} \approx 276.9$$

The marginal rate of substitution was estimated to be 276.9.

CONCLUSION

- In this study, we demonstrated Sim2Real transfer learning for predicting the thermal stability of nanobodies by integrating simulation-derived ΔΔG data and experimental melting temperature (T_m) data through multitask learning.
- The estimated marginal rate of substitution (277:1) provides a quantitative guideline on the extent to which simulation data should be leveraged to compensate for limited experimental data, thereby contributing to the efficient allocation of HPC resources in biomolecular simulations.

ON-GOING STUDIES

- Performing multitask learning using folding free energies (ΔG) instead of ΔΔG to incorporate simulation data more directly related to the melting temperature (T_m)

ACKNOWLEDGEMENTS

- This work was supported by RIKEN TRIP AGIS