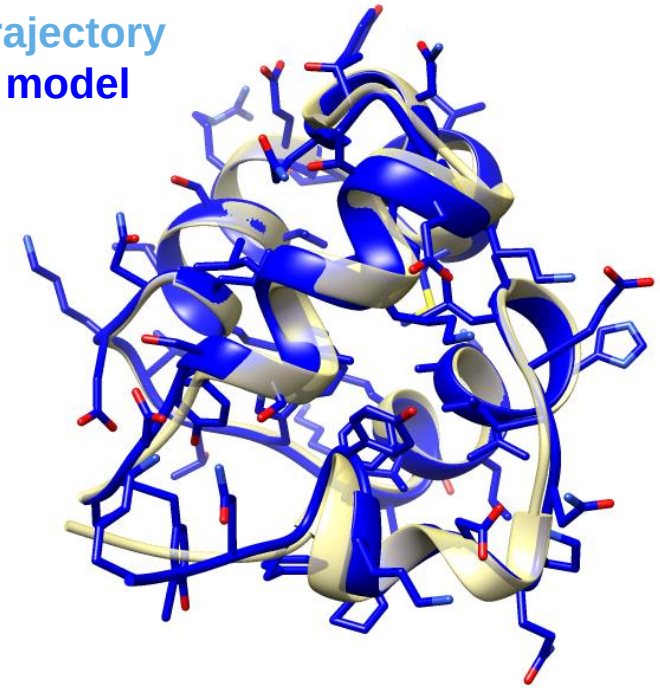
Protein Model Refinement via Molecular Dynamics Simulations with Improved Sampling Protocols

Lim Heo and Michael Feig

Feig Computational Biophysics Lab
MICHIGAN STATE UNIVERSITY

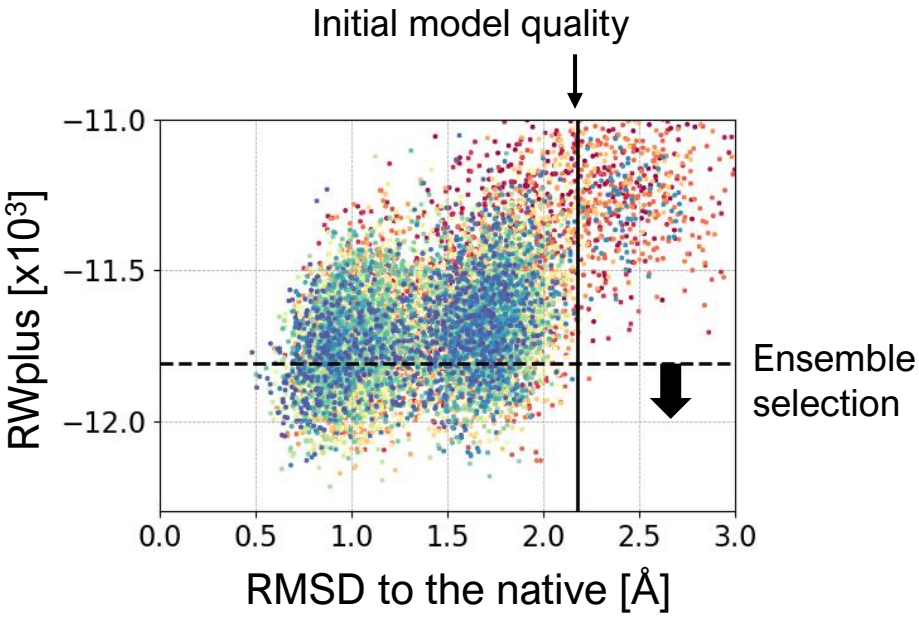
MD simulation and refinement

Native
Initial model
MD trajectory
Final model



Conformational sampling via MD simulations

Progress of sampling



Selecting ensemble of sample structures and averaging selected structures

CASP14 protocol (1)

- **FEIG-S: fully automated** refinement server

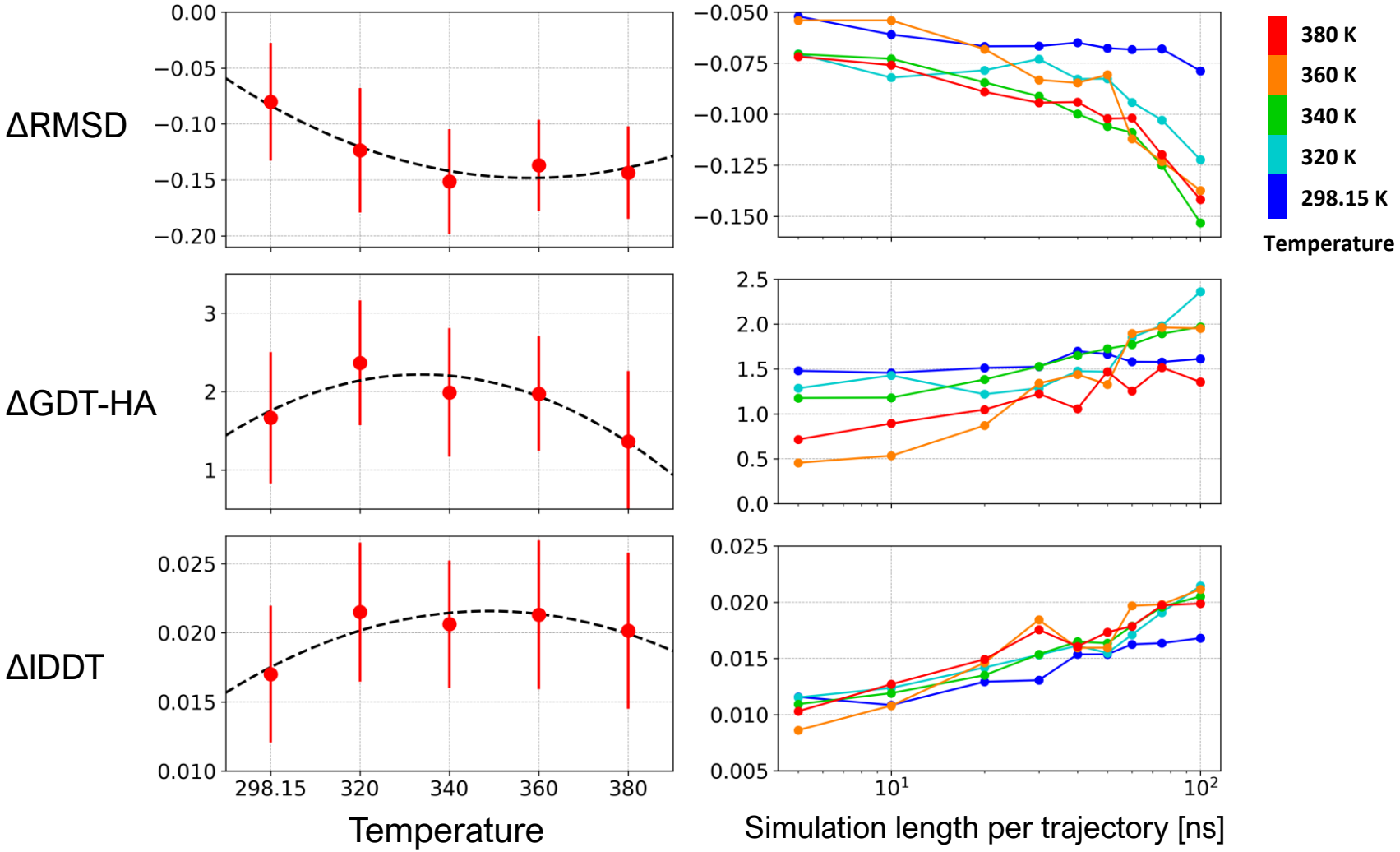
- Structure sampling via MD simulations
  - **Flat-bottom** harmonic restraints:
    Cartesian coordinates of Cα and **distance** between Cαs

$$E_{\text{Combined}}(\lambda) = \lambda \sum_{i} E_{\text{Cartesian}}\left(\mathbf{r}_i ; \mathbf{r}_i^{\text{o}}\right) + (1-\lambda) \sum_{j-i>3} E_{\text{Distance}}\left(d_{ij} ; d_{ij}^{\text{o}}\right)$$

  - Simulations at a **higher temperature**, 360 K
  - **CHARMM36m** with modifications
    - Lower backbone dihedral angle energy barriers
    - Hydrogen-mass repartitioning (heavier hydrogens) → 4 fs/MD step
  - 500 ns/model (= 5 trajectories x 100 ns)
    (16 GPU hours on RTX2080Ti for R1056 (169 residues))
    (c.f. 2.25 μs/model for CASP13 protocol)

* Heo, L. *et al.*, Improved sampling strategies for protein model refinement based on molecular dynamics simulation, *ChemRxiv* (**2020**).

Refinement at higher temperature



*** CASP10 refinement targets**

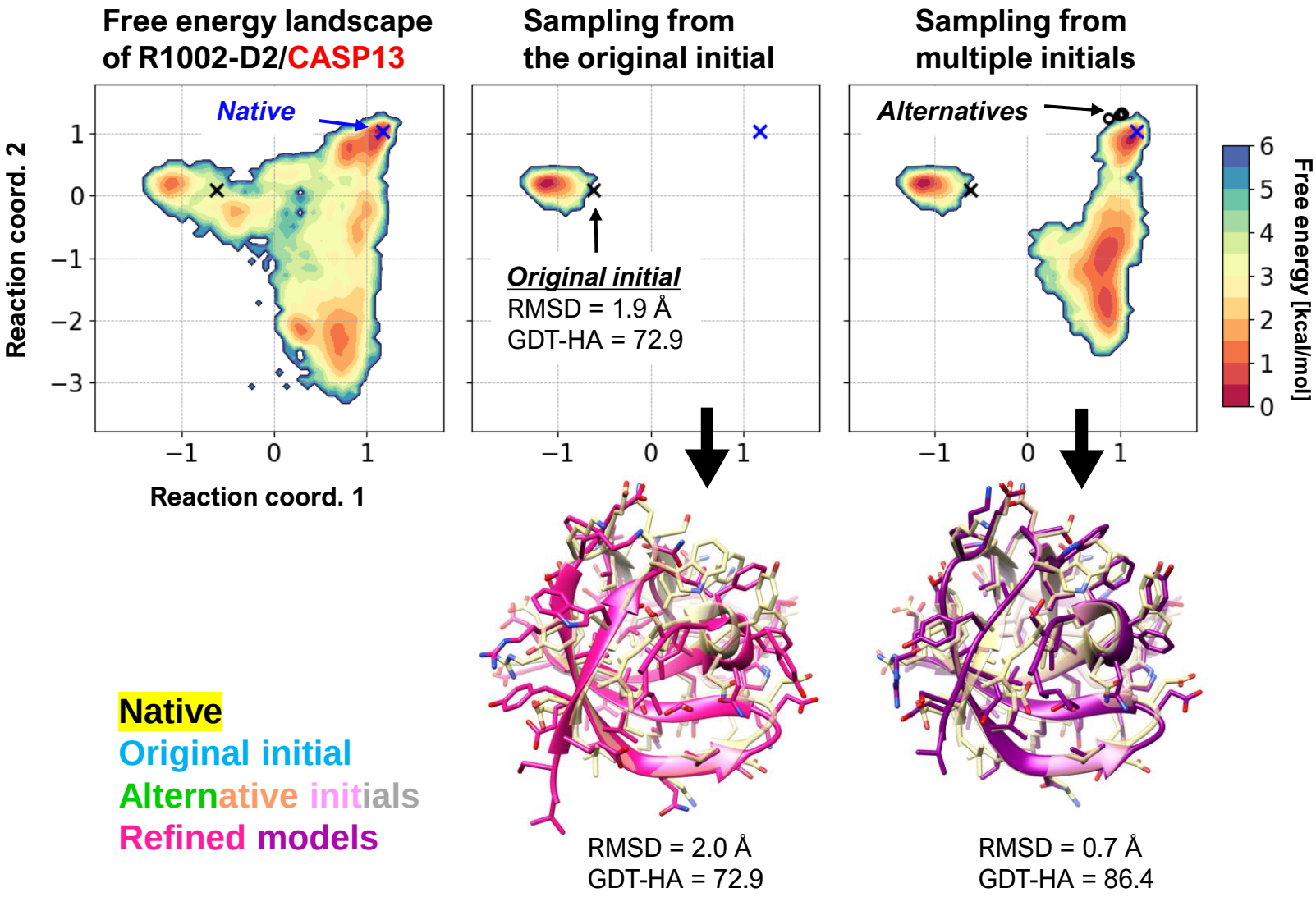
CASP14 protocol (2)

- Use of **multiple initial** models (6 targets)
  - Structure diversification prior to MD sampling to generate **kinetically favored** initial models
  - Simple **TBM** using sequence & structural similarity, further optimization by simplified Rosetta Iterative Hybridize

  - 5 initial models: **original initial** model + **4 alternatives**
  - 500 ns/model → 2.5 μs/target

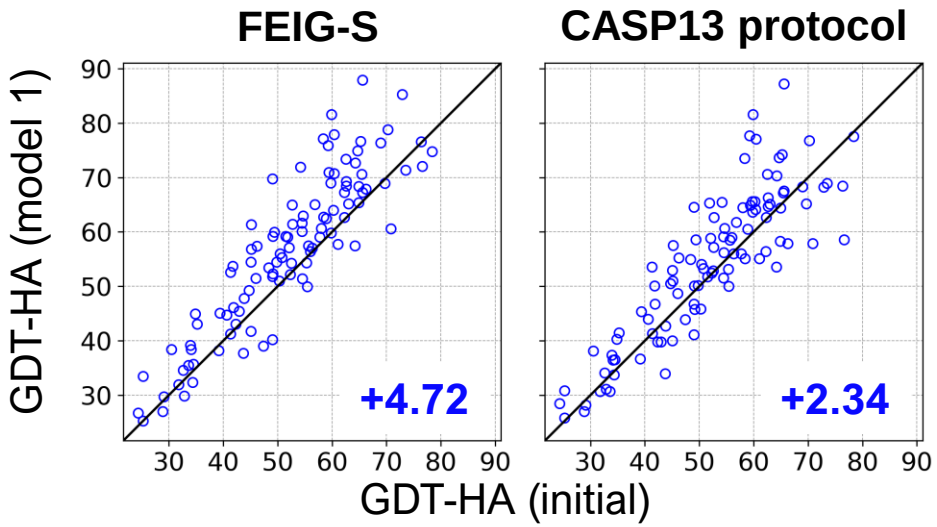
- **FEIG**
  - Hybridizing alternative submitted models from the same server group of the initial model
    - i.e., tFold-IDT model 1,4,5 were hybridized with the initial model of R1091-D2, tFold-IDT model 3

Refinement with multiple initials

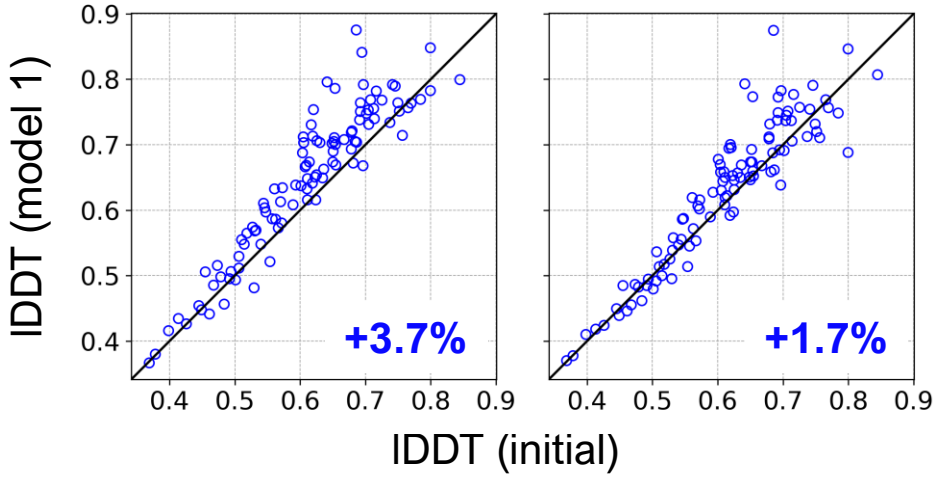


Progress? benchmark results

*** CASP11-13 refinement targets**



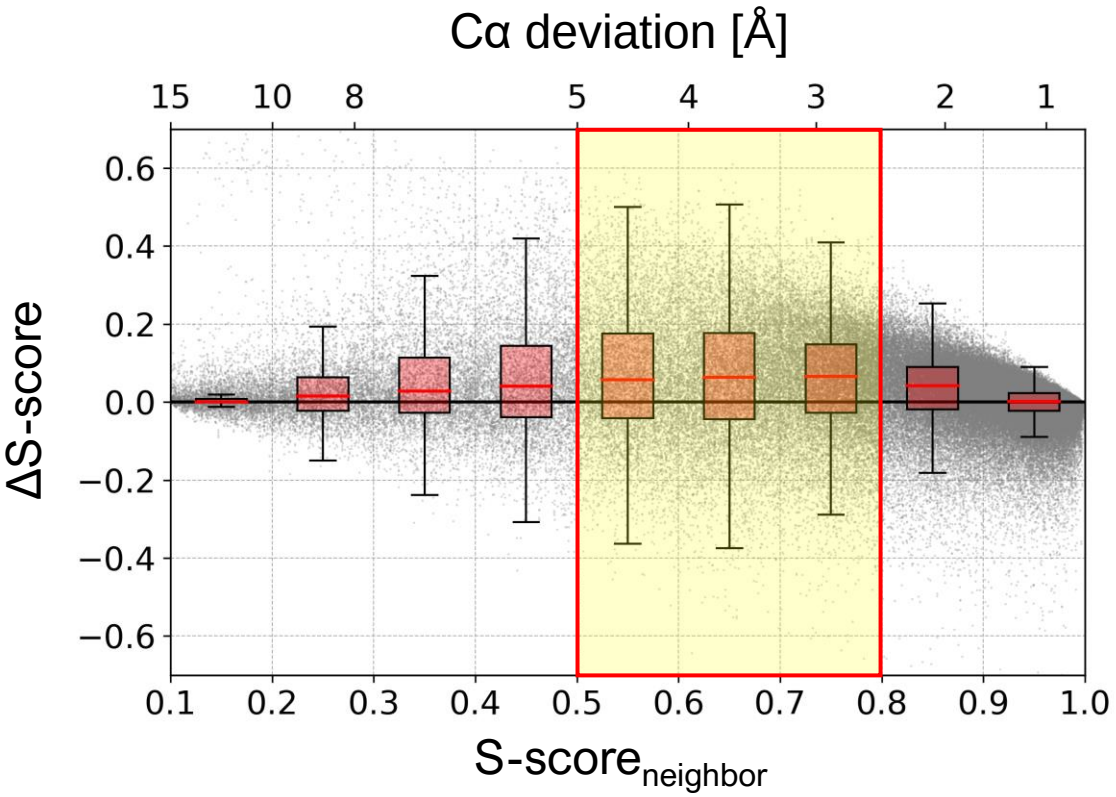
CASP12 protocol: +2.34



CASP12 protocol: +1.5%

Where is refinement possible?

*** CASP11-13 refinement targets**

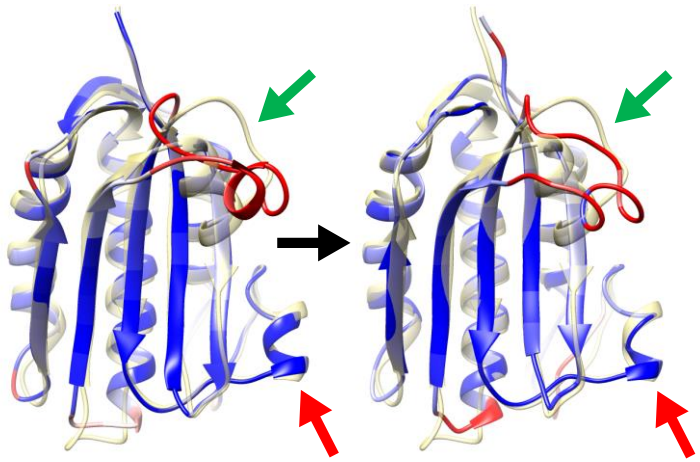


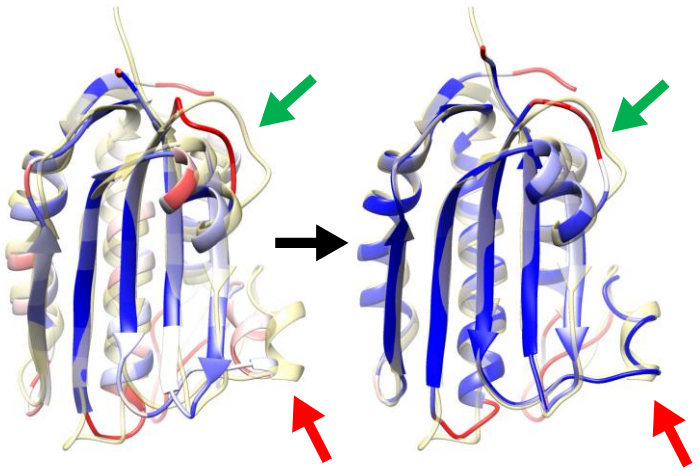
$$\text{S-score} = \frac{1}{\left(1+(d/5\,\text{Å})^2\right)}$$

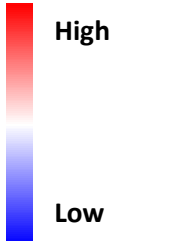
$$\text{S-score}_{\text{neighbor}}(r) = \frac{\sum_s w(s)\ \text{S-score}(r+s)}{\sum_s w(s)}$$

$$w(s) = e^{-(s/s_0)^2/2}$$

Example of more refinable model

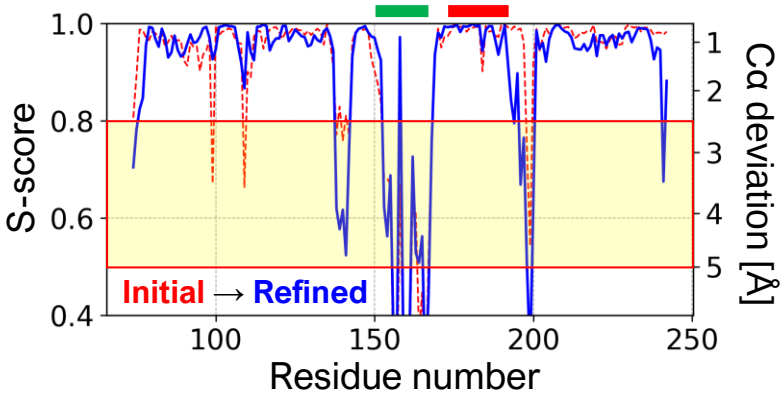


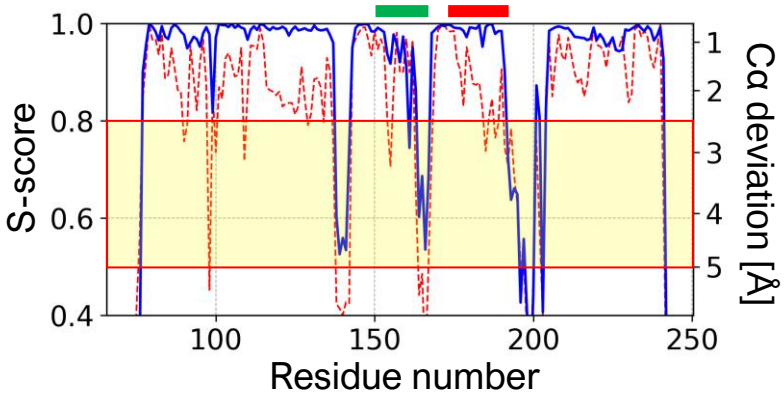




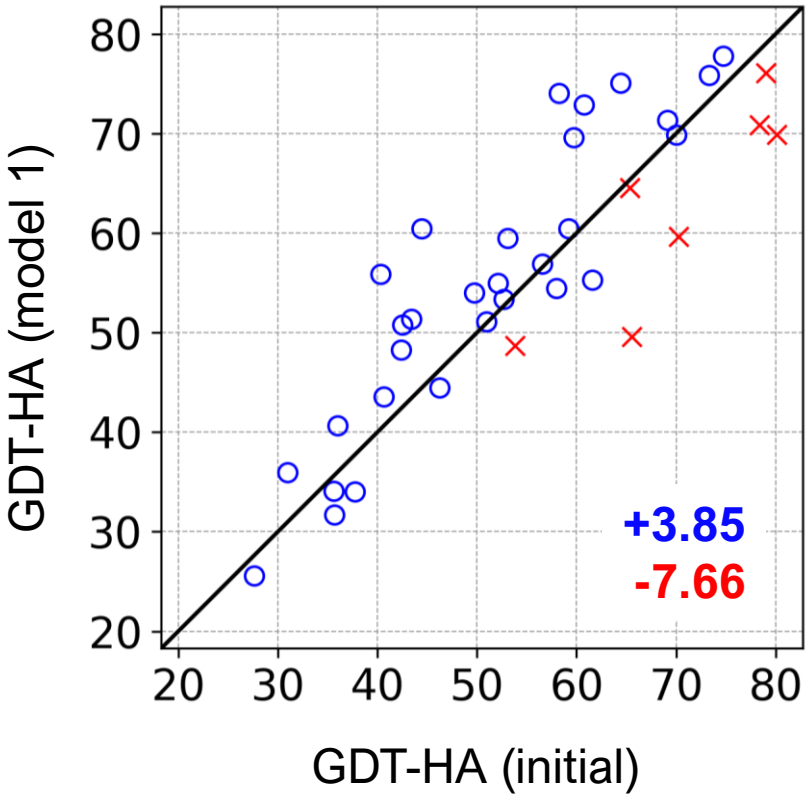
TR893 (original)
GDT-HA = 69.7 → 66.3 (-3.4)
lDDT = 0.737 → 0.719 (-0.018)

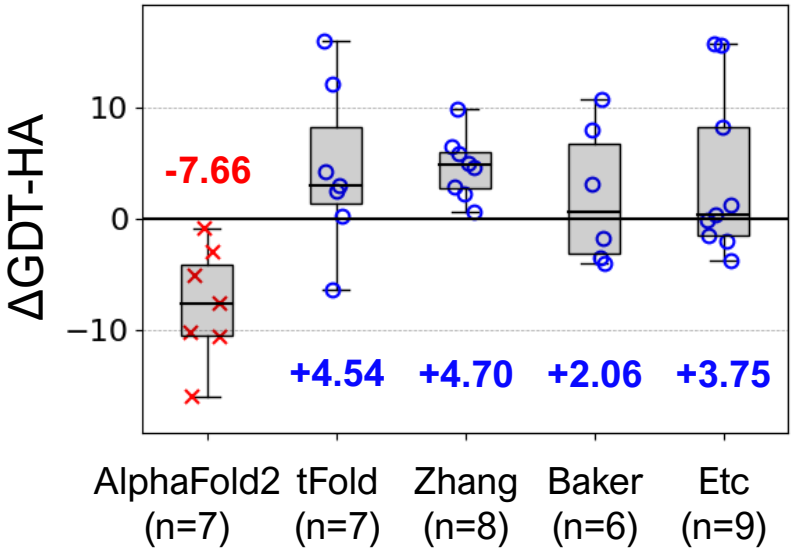
TR893 (alternative)
GDT-HA = 49.3 → **73.7** (+24.4 / **+4.0**)
lDDT = 0.681 → **0.777** (+0.096 / **+0.040**)





FEIG-S: CASP14 results

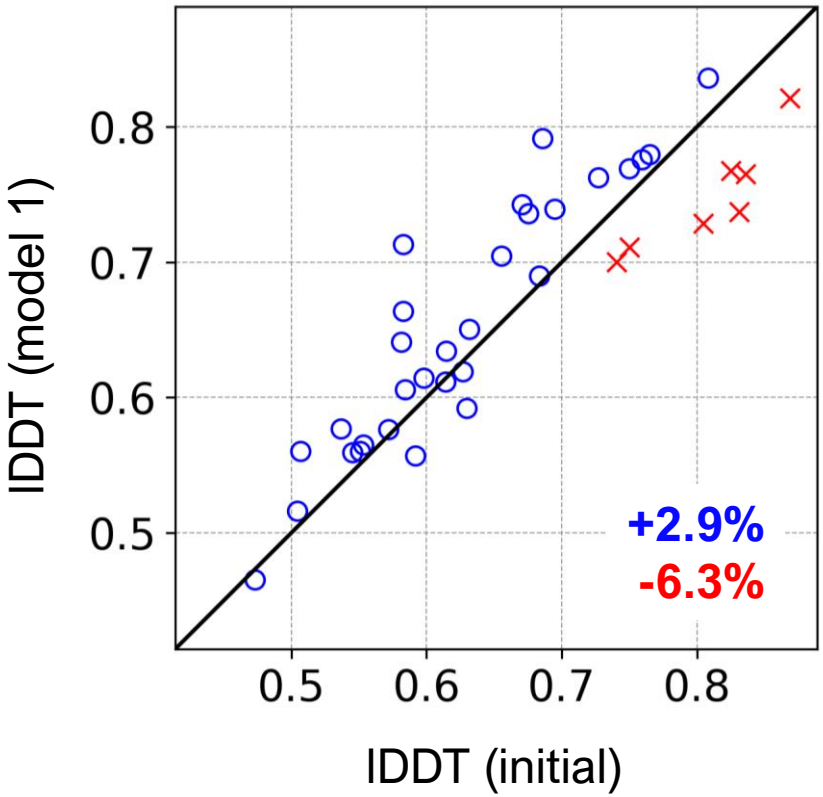


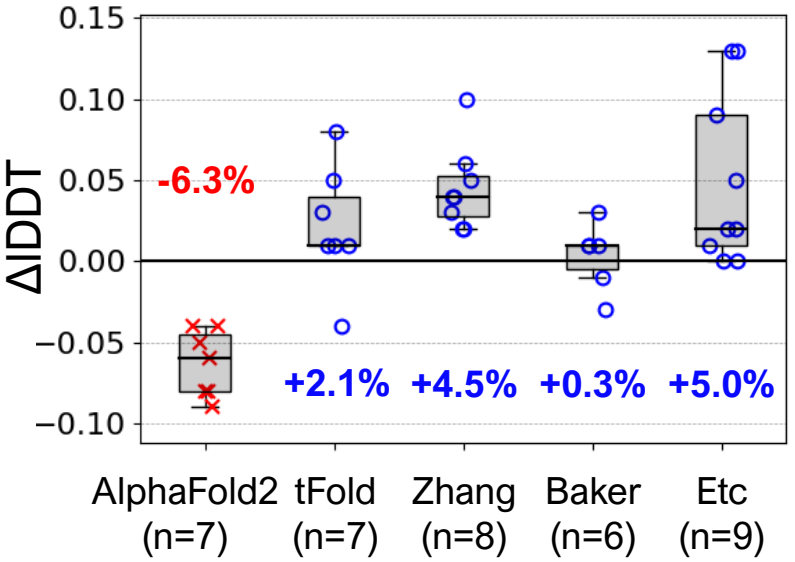


CASP13 protocol: +1.70 / -5.89
CASP12 protocol: +1.39 / -0.53

Targets
originated from AlphaFold2 (7 domains)
originated from other groups (30 domains)

FEIG-S: CASP14 results

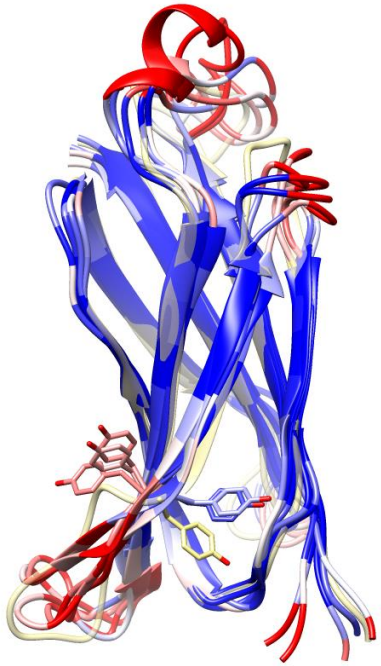


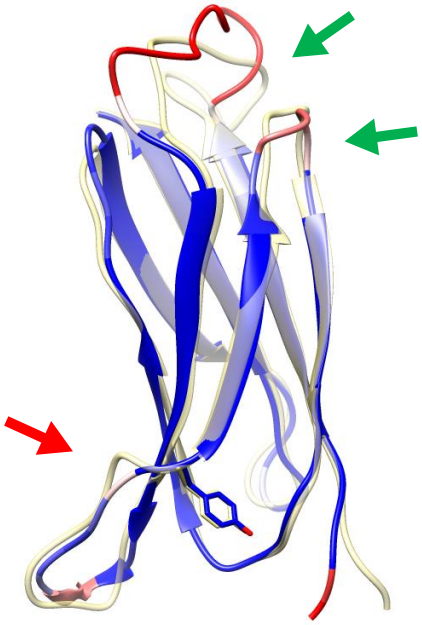


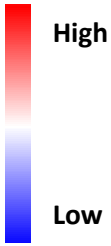
CASP13 protocol: +1.1% / -5.5%
CASP12 protocol: +0.6% / -3.3%

Targets
originated from AlphaFold2 (7 domains)
originated from other groups (30 domains)

What was improved?



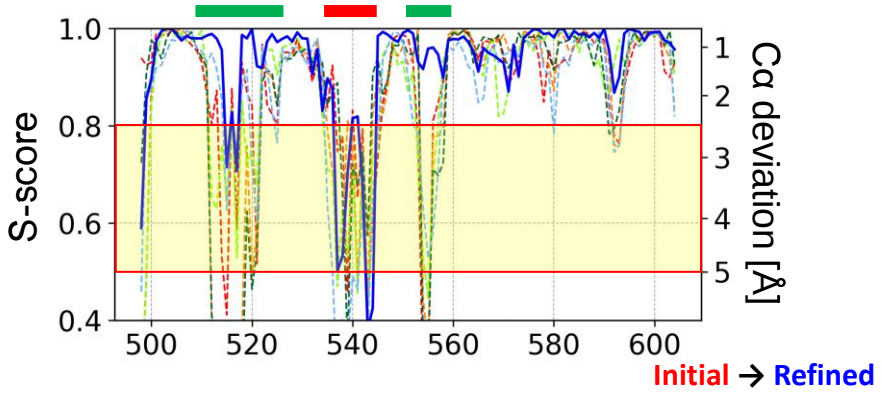




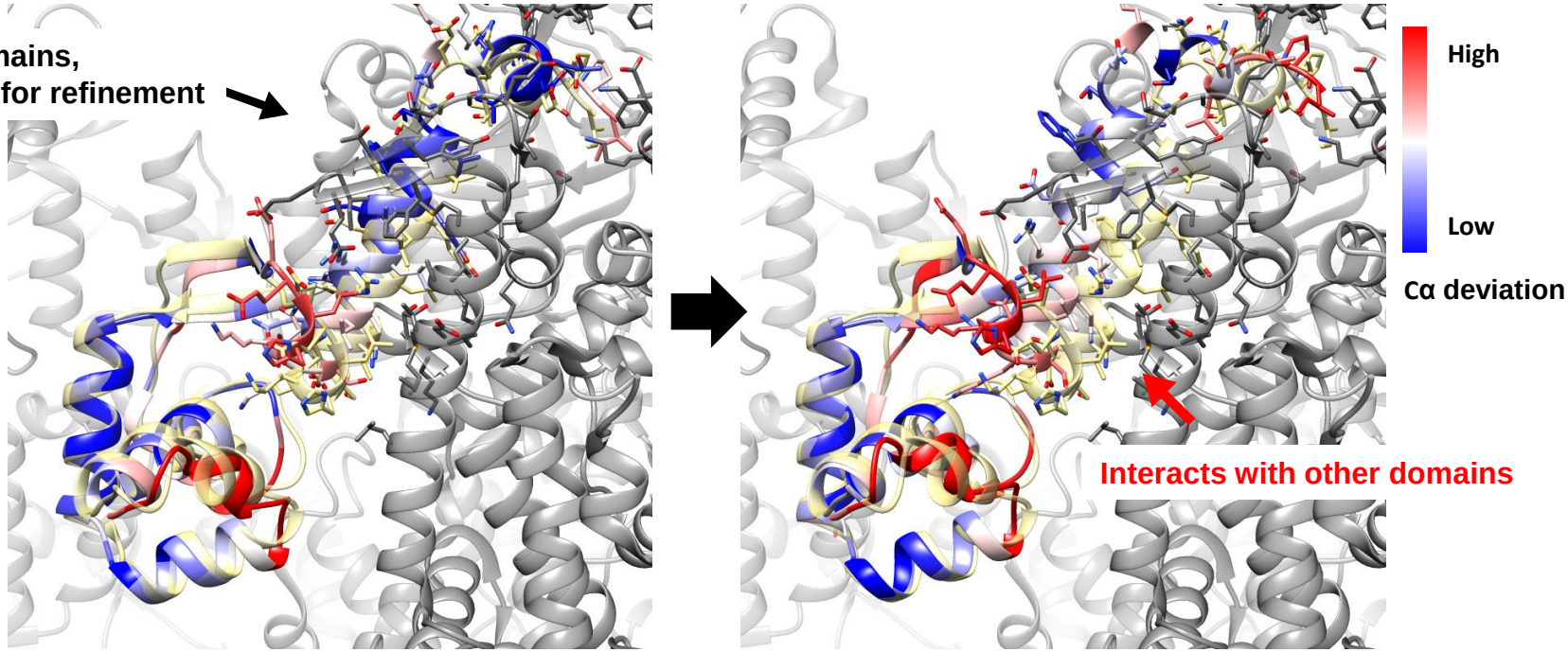
R1091-D2 (tFold-IDT)

GDT-HA = 60.8 → 76.9 (+16.1)

lDDT = 0.695 → 0.760 (+0.065)



What went wrong?



R1040v1 (AlphaFold2)

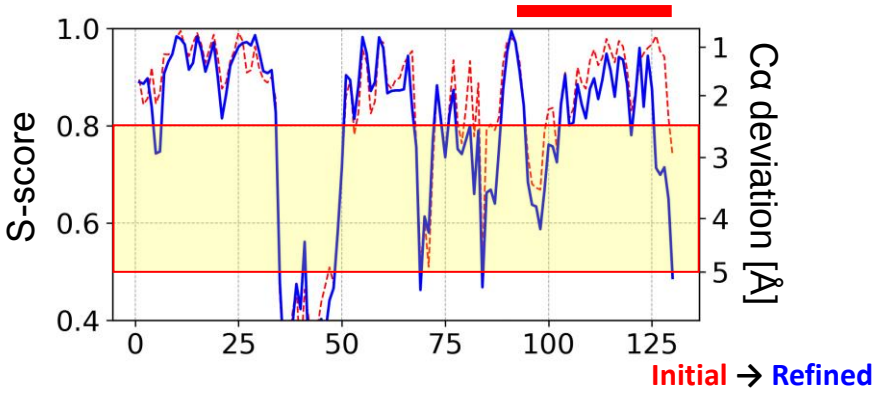
GDT-HA = 53.9 → 48.7 (-5.2)

lDDT = 0.741 → 0.700 (-0.041)

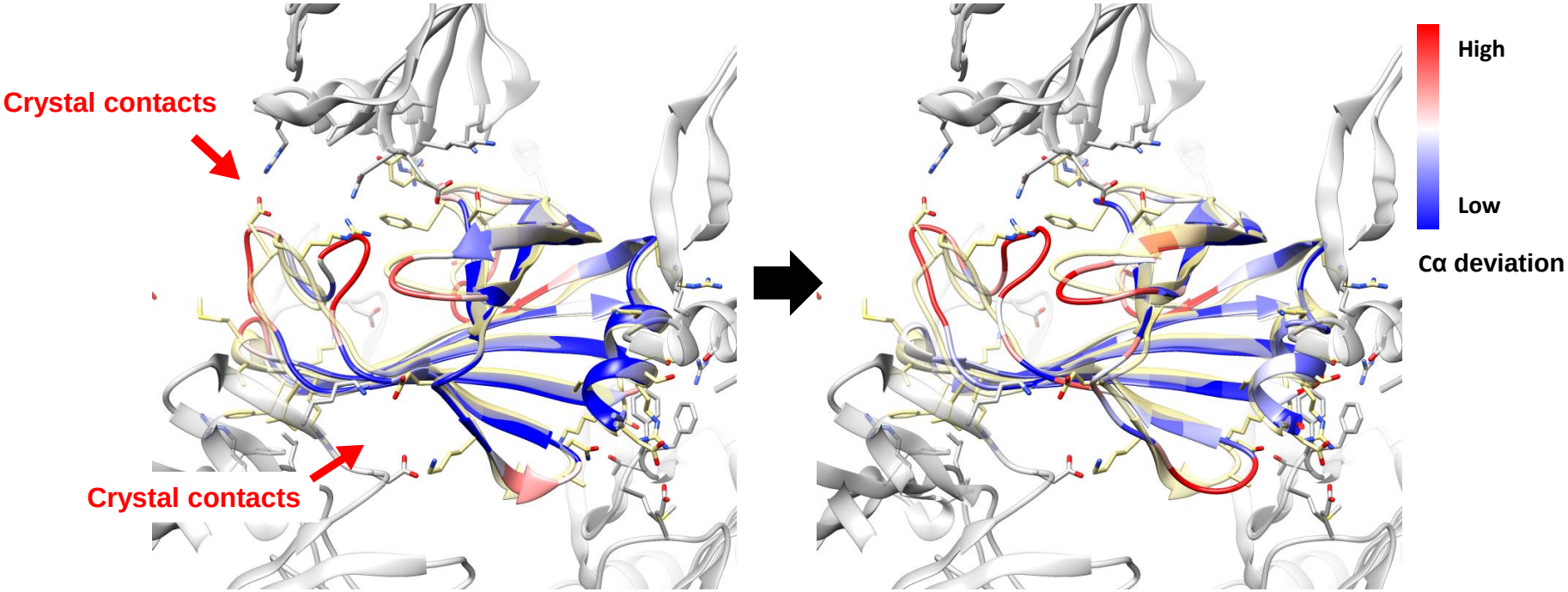
For non-interfacial residues

GDT-HA = 67.0 → 68.3 (+1.3)

lDDT = 0.713 → 0.711 (-0.002)



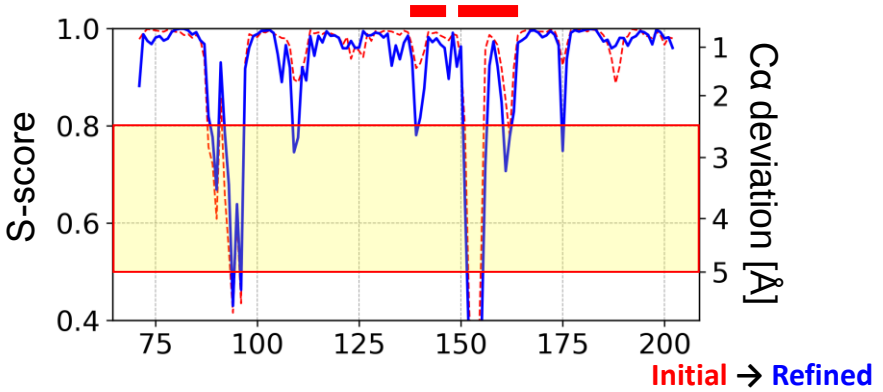
What went wrong?



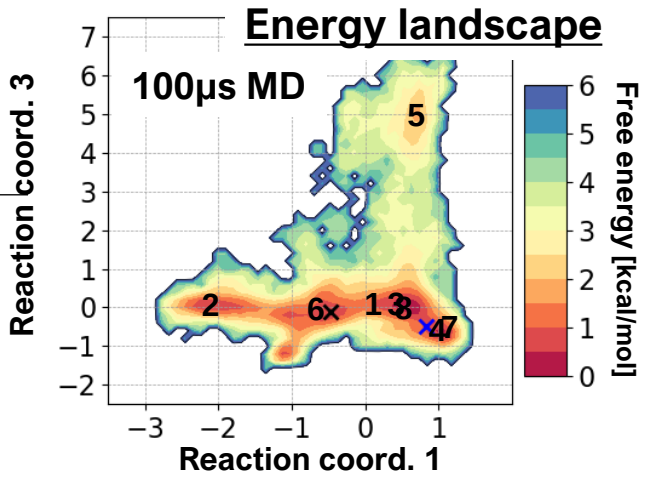
R1074v1 (AlphaFold2)

GDT-HA = 78.4 → 70.8 (-7.6)

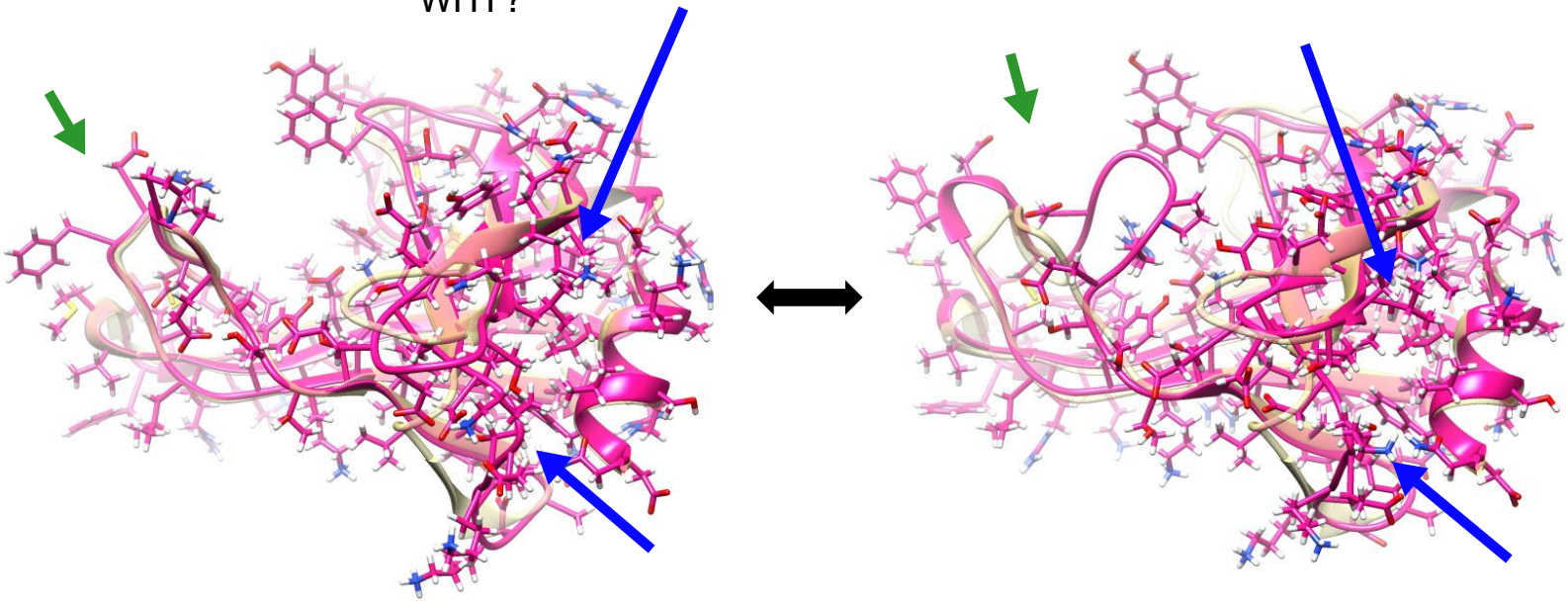
lDDT = 0.836 → 0.765 (-0.071)



What went wrong?



Many exposed **hydrophobic** residues.
WHY?



Crystal structure-like state

FE = 0.5 kcal/mol
GDT-HA = 82
RMSD = 1.4 Å

AlphaFold model-like state

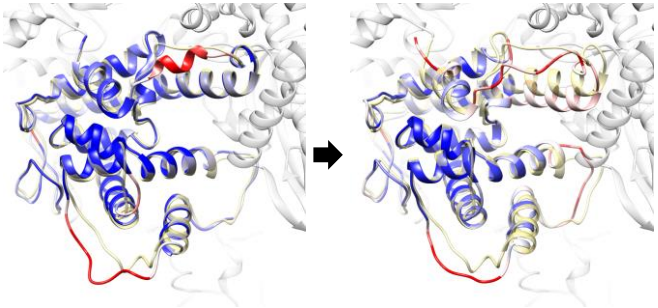
FE = 0.0 kcal/mol
GDT-HA = 74
RMSD = 2.5 Å

What went wrong?

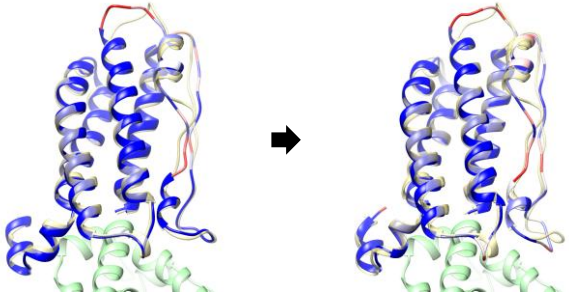
Domain-domain interactions

R1042v2 (AlphaFold2): 65.6 → 49.6 (**-16.0**)

R1041v1 (AlphaFold2): 70.3 → 59.6 (**-10.7**)



R1053v2 (AlphaFold2): 80.1 → 69.9 (**-10.2**)



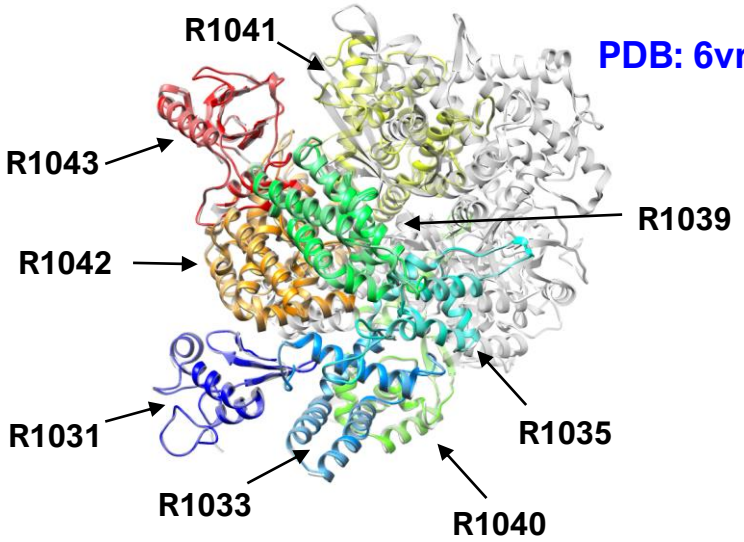
R1045s2 (tFold): 61.6 → 55.3 (**-6.3**)

R1040v1 (AlphaFold2): 53.9 → 48.7 (**-5.2**)

R1042v1 (BAKER-exp): 35.7 → 31.7 (**-4.0**)

R1033 (E2E): 37.8 → 34.0 (**-3.8**)

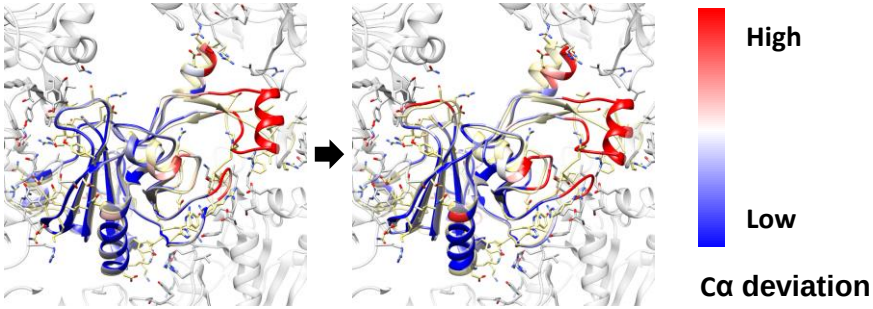
*** GDT-HA changes via refinement (FEIG-S) are shown.**



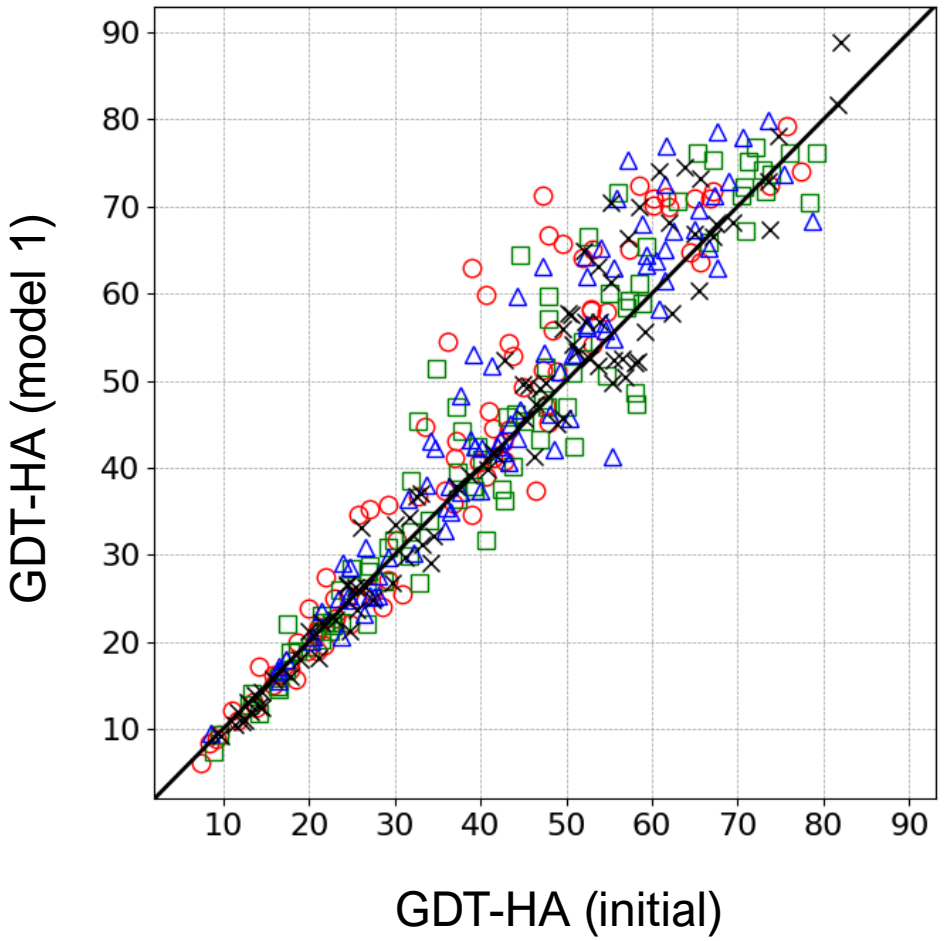
Crystal contacts

R1074v1 (AlphaFold2): 78.4 → 70.8 (**-7.6**)

R1067v2 (AlphaFold2): 79.1 → 76.0 (**-3.1**)



Refinement in the era of ML



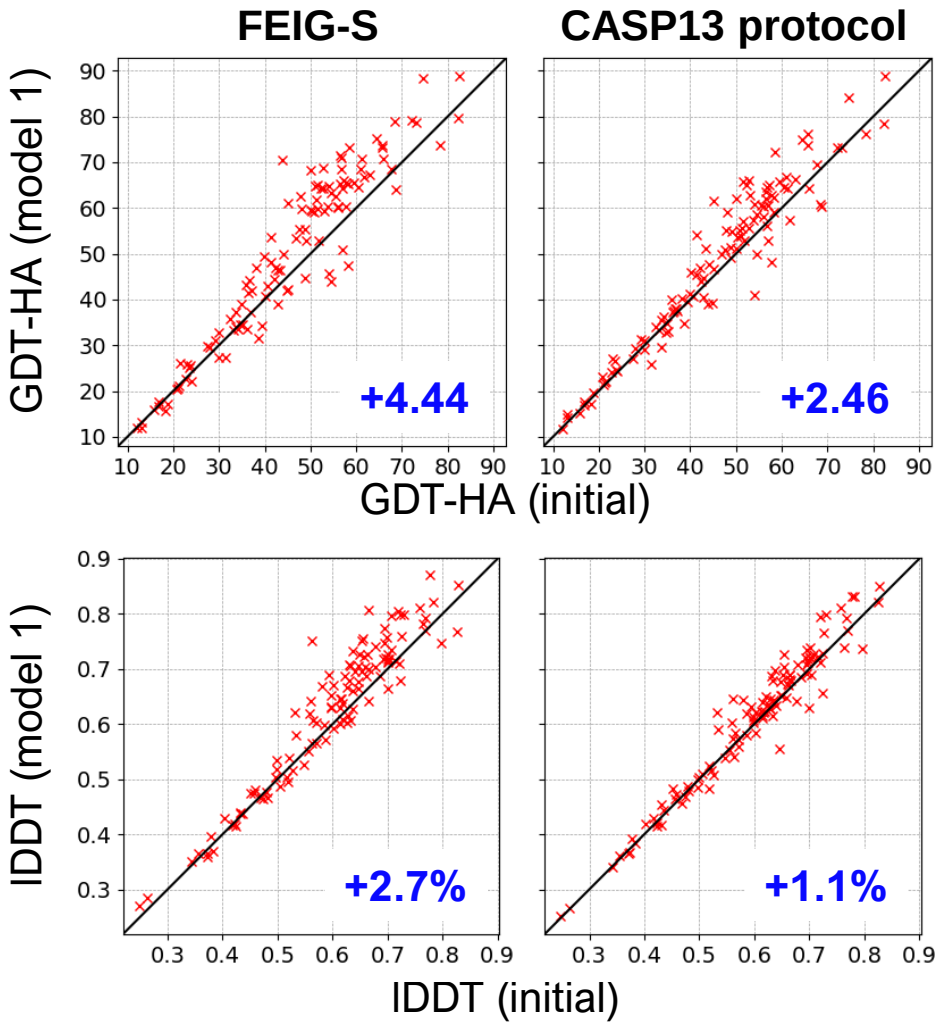
FEIG-S (trRosetta+TBM):
37.2 → **40.6 (+3.4)**

FEIG-R1 (RaptorX model 1):
41.9 → **43.1 (+1.2)**

FEIG-R2 (Zhang-s model 1):
44.5 → **47.2 (+2.8)**

FEIG-R3 (Baker-s model 1):
42.2 → **43.3 (+1.1)**

Refinement of CASP13 AlphaFold models



CASP12 protocol: +1.38

CASP12 protocol: +0.5%

* Heo, L. and Feig, M., High-accuracy protein structures by combining machine-learning with physics-based refinement, *Proteins* (**2020**).

Summary

- We improved sampling procedure for refinement
  - Optimized MD simulation procedure
  - Used multiple initial models

- Successes
  - Roughly correct residues could be improved
- Limitations
  - Errors with significant energy barrier could not be fixed
- Failures
  - Occasionally missing inter-domain/protein contacts made models worse

Acknowledgements

Feig Computational Biophysics Lab
MICHIGAN STATE UNIVERSITY

- CASP participants, assessors, target providers, and organizers

- NIH funding (R35 GM126948)
- XSEDE computing resources (TG-MCB09003)