Introduction to stability data

Challenge:

Predicting multi-mutant mini-protein stability
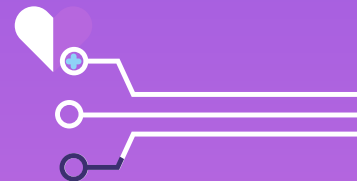
CBioVikings   biolib   REBBLS
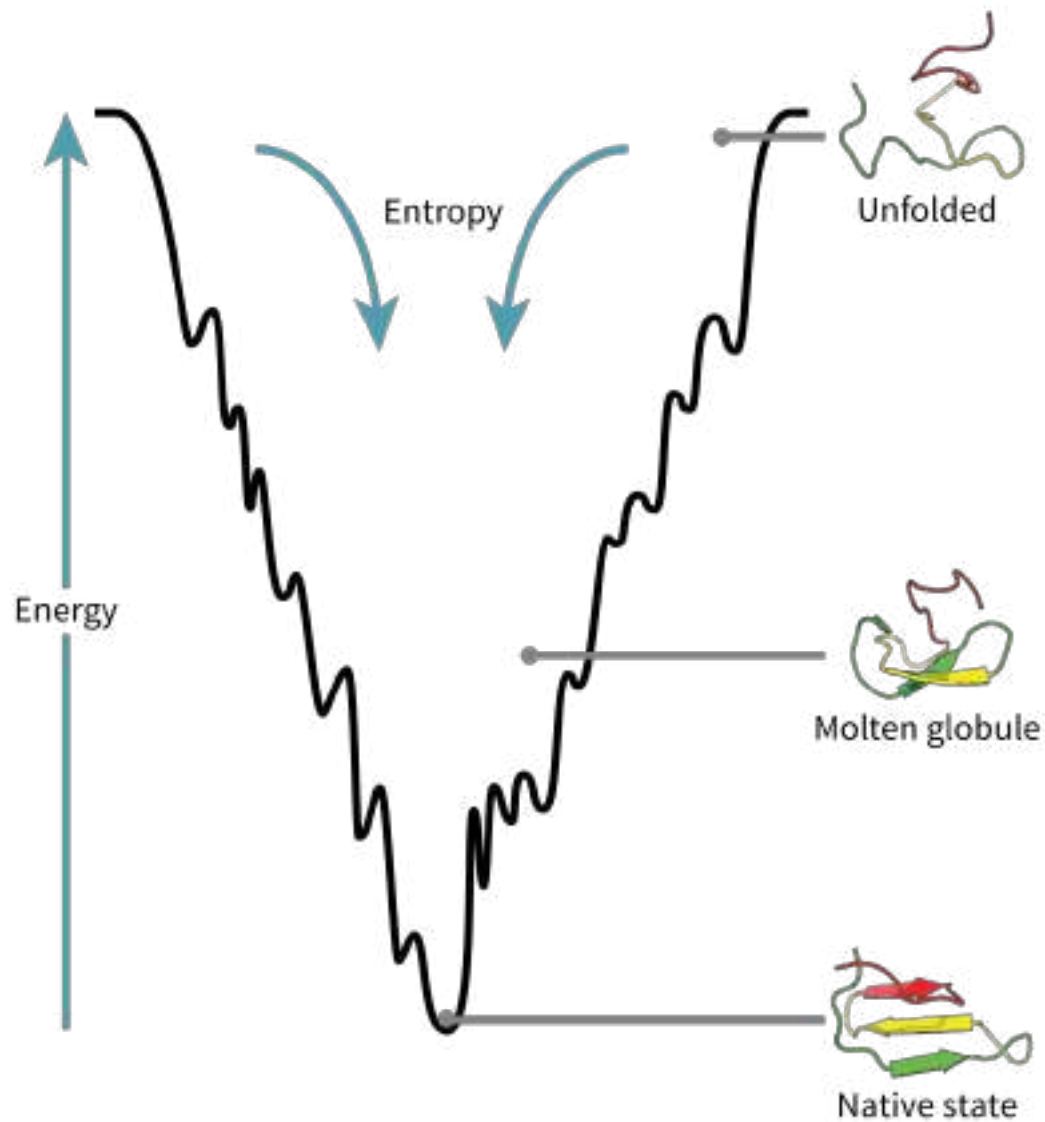
Protein stability

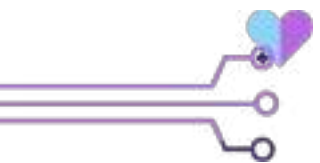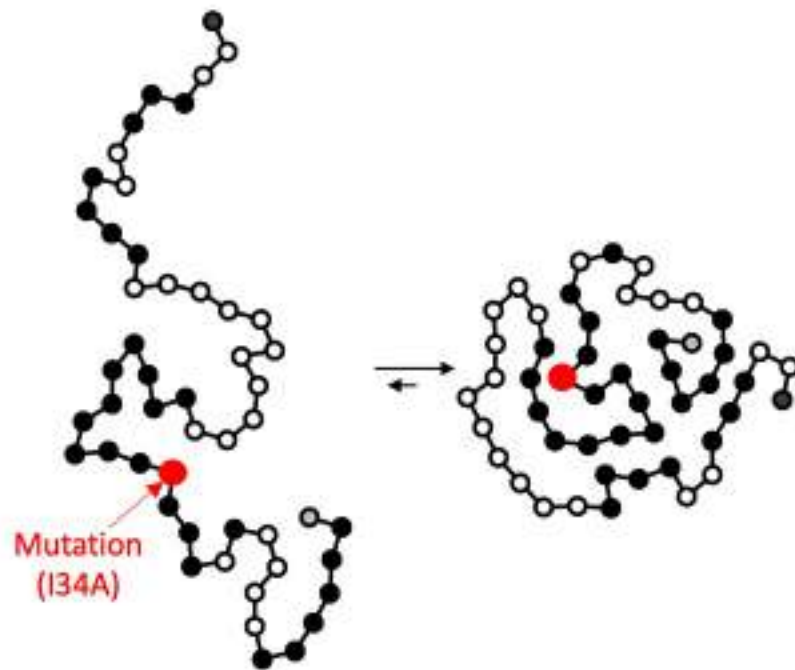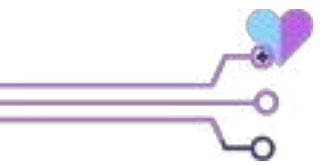Energy diagram showing Entropy, Unfolded, Molten globule, and Native state.

Protein stability is the extent to which the native state is more favorable than all the competing states.

Mutagenesis

**Protein sequence determines fold**

Multi-mutant miniprotein data set

single_muts_train

	name	sequence	secondary_structure	stabilityscore
0	EHEE_rd2_0005.pdb_V25L	GSSTTRYRFTDEEEARRAAKEWARRGYQLHVTQNGTYWEVEVR	TTTTEEEEEETTHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	1.72
1	HHH_rd2_0134.pdb_R22D	SKDEAQREAERAIRSGNKEEADRILEEAGYSPEQAERIRKLG	THHHHHHHHHHHHHTTTHHHHHHHHHHHTTTHHHHHHHHHHHHHT	1.22
2	HHH_rd2_0134.pdb_Q6K	SKDEAKREAERAIRSGNKEEARRILEEAGYSPEQAERIRKLG	THHHHHHHHHHHHHTTTHHHHHHHHHHHHTTTHHHHHHHHHHHHHT	1.2
3	EHEE_0882.pdb_E2I	GSSQITIEVEDEEEARRVAKELRKKGYEVKIERRGNKWHVHRT	TTTTEEEEEETTHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	1.59
4	EEHEE_rd3_1716.pdb_A26R	TEVHLGDIKLKYPNPEQAKKAAEKLRQKYNLTWTVIGDYVKIE	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	0.88
5	EEHEE_rd3_1716.pdb_Y12Q	TEVHLGDIKLKQPNPEQAKKAAEKLAQKYNLTWTVIGDYVKIE	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	0.41
6	EEHEE_rd3_1498.pdb_H4V	GTLVLNGVTVKVPSLEKAIKAAKKFAKKYNLEVQVHGNTVHVH	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	1.06
7	EEHEE_rd3_1716.pdb_E43Y	TEVHLGDIKLKYPNPEQAKKAAEKLAQKYNLTWTVIGDYVKIY	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	1.19
8	EEHEE_rd3_1702.pdb_K17W	TTIHVGDLTLKYDNPKWAYEIAKKLAKKYNLTVTIKNGKITVT	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	1.03
9	HHH_rd3_0138.pdb_K10F	ERRKIEEIAFKLYQSGNPEAARRFLRKAGISEEEEIERILQKAG	THHHHHHHHHHHHTTTHHHHHHHHHHHHTTTHHHHHHHHHHHHHT	1.88
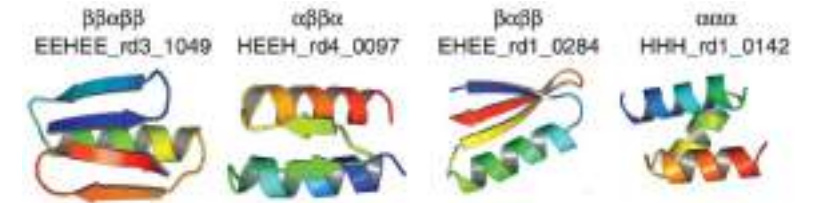
multiple_muts_train

	name	sequence	secondary_structure	stabilityscore
0	EHEE_rd1_0392.pdb	GSSTETYEVHTTDEAERVRKELEKKGYKVEVRSDGTEYEIRSE	TTTTEEEEEETTHHHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	-0.32
1	EHEE_rd1_0925.pdb_random	GSSEKEDVHRFEIATWERDKQLEERTGREGRTRDIANREDERE	TTTTEEEEEETTHHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	-0.5
2	HEEH_rd1_0699.pdb_hp	EARKSDGPSGVKRAKKNEDKETDRKTALYNKIQIVAKEITENE	THHHHHHHHHHHHHTTEEEEETTEEEEETTTHHHHHHHHHHHHHHHT	-0.33
3	EEHEE_rd1_0108.pdb_hp	GSAQFKNGNEESEVPYKGRRVKQEPKKDANEWERQGTEGRSDG	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	-0.18
4	EEHEE_rd1_0483.pdb	GSGTVELNGNRYTSNSEEDIREWIRRQGATTSETSNGKYQVHR	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	-0.12
5	EHEE_rd1_0417.pdb	GSSTTTYKVTDENERKRVTKEIKKRGGEVRTQEENGTFKVEER	TTTTEEEEEETTHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	-0.94
6	EHEE_rd1_0275.pdb_hp	GSSDRYDTGEYQKQIRTGKKDVKNTAEKGTDDRERLERNLTNV	TTTTEEEEEETTHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	-0.94
7	EEHEE_rd1_0592.pdb_hp	GSKEGYWKVQEVKFKNNTTAKNKADQFAVRNERTEEKEAKATG	TEEEEETTEEEEETTTHHHHHHHHHHHHHHHHHTTTEEEEETTEEEET	-0.02
8	HEEH_rd1_0384.pdb_hp	KEQQEKRGDTVCRDKARLKLPRAEVEKVKKKGNDKIVYYTEDE	THHHHHHHHHHHHHTTEEEEETTEEEEETTTHHHHHHHHHHHHHHHT	-0.43
9	EHEE_rd1_0585.pdb_random	GSSKVTYRNARQYNDDGVTEQSEEFKKETKKKTATVLEVESKG	TTTTEEEEEETTHHHHHHHHHHHHHHHHHTTTEEEEEEETTEEEEEEET	-0.37
10	HEEH_rd1_0441.pdb_random	VLREQSKKSKGVAAKERGYNRDKKEVPHKRKKIETTTKRKEIE	THHHHHHHHHHHHHTTEEEEETTEEEEETTTHHHHHHHHHHHHHHHT	-0.4
11	HEEH_rd1_0399.pdb_random	NQEVADYQADQSEIREKGAKHLRKTRKAKEVQKENKKENRRA	THHHHHHHHHHHHHTTEEEEETTEEEEETTTHHHHHHHHHHHHHHHT	-0.32

Secondary structure codes:
- H -> α-helix
- E -> β-sheet
- T -> loop

βββαββ
EEHEE_rd3_1049

αββα
HEEH_rd4_0097

βαββ
EHEE_rd1_0284

ααα
HHH_rd1_0142

High positive stability score -> stable protein

Performance of existing stability predictors

Single mutations

Multiple mutations

Experimental multi-mutation ΔΔG vs. Predicted multi-mutation ΔΔG