

Supplementary Information

ThermoNet-2: sequence-only prediction of protein melting temperature · Okafor, Deshpande, Vargas & Chen · J. Comput. Biophys. 41(3): 288–301 (2024) · doi:10.4821/jcb.2024.0288

Table S1. Records per protein sequence (top 10 of 212)

Protein	Records	Buffers	In train	In test
Lysozyme (hen)	11	7	9	2
Ribonuclease A	9	6	7	2
Myoglobin (horse)	8	5	6	2
Ubiquitin	8	5	7	1
Barnase	7	4	5	2
Carbonic anhydrase II	6	4	5	1
Chymotrypsin inhibitor 2	6	4	4	2
Staphylococcal nuclease	6	3	5	1
Cytochrome c	5	3	4	1
T4 lysozyme	5	3	4	1

Note S2. Sequence overlap between splits

148 of the 212 unique sequences (69.8%) appear in both the training and the test split under different buffer conditions. Of the 241 held-out records, **198 (82.2%) have at least one record of the same sequence in training.**

We report this for completeness. As described in the main text, splitting was performed at the record level, which is standard practice in this literature.

Table S3. Performance by overlap status

Held-out records	n	RMSE (°C)
Sequence also seen in training	198	4.1
Sequence not seen in training	43	8.0
All held-out records	241	4.9

The 43 records whose sequences do not appear in training are the closest available estimate of performance on a novel protein.