

A sequence-disjoint benchmark for T_m prediction reveals optimistic reporting

K. Lindqvist, E. Boateng

Department of Biochemistry, Northern Institute

Received 3 Jun 2024 · Accepted 21 Jan 2025

ABSTRACT

Published melting-temperature predictors are routinely evaluated on splits that allow the same protein sequence to appear in training and test under different buffer conditions. We construct BENCH-TM, a 318-protein benchmark with no sequence overlap between splits, and re-evaluate four published predictors. All four degrade substantially; the largest reported margin over a simple baseline does not survive.

1. Motivation

A protein measured in five buffers yields five records. Splitting at the record level places the same sequence on both sides of the split. The model can then memorise a sequence and interpolate the buffer effect, which inflates apparent accuracy.

2. BENCH-TM

BENCH-TM contains 318 proteins with at least one DSF measurement, clustered at 30% sequence identity, with whole clusters assigned to a single split. Buffer conditions are recorded as covariates rather than being folded into the label.

3. Re-evaluation

Table 3. Performance on BENCH-TM (sequence-disjoint, n = 318).

Method	Reported RMSE (°C)	BENCH-TM RMSE (°C)
SeqTm (Barros, 2021)	5.3	8.1
TSP-Struct (Nilsson, 2023)	4.6	7.6
ThermoNet-2 (Okafor, 2024)	4.9	7.8
Mean predictor (baseline)	—	8.4

4. Discussion

ThermoNet-2 reports 4.9 °C on its own record-level split and 7.8 °C here. The gap is consistent with sequence leakage rather than with any defect in the model. We stress that this is a reporting problem, not a fraud problem: the ThermoNet-2 authors describe their splitting procedure accurately in their Section 3. Readers simply tend not to reach Section 3. A mean-predictor baseline achieves 8.4 °C on BENCH-TM, so the practical margin for every method tested here is under one degree.

5. Recommendation

Report sequence-disjoint performance alongside any record-level number. Where a predictor is used to prioritise candidates, state the baseline it is being compared against.

