

Buffer-aware prediction closes most of the sequence-disjoint gap

I. Kowalski, N. Abara, K. Lindqvist

Department of Biochemistry, Northern Institute

Received 2 Dec 2025 · Accepted 14 Apr 2026

ABSTRACT

Sequence-only melting-temperature predictors lose most of their apparent accuracy under sequence-disjoint evaluation. We show that a large part of the residual error is not a failure of sequence modelling but a missing covariate: buffer composition. Supplying pH and ionic strength as explicit inputs reduces error on BENCH-TM from 7.8 to **6.1 °C**. We also show why this does not help in the setting most users care about.

1. Background

Lindqvist & Boateng (2025) showed that record-level splitting inflates reported accuracy, and that on a sequence-disjoint split no published predictor beats a mean predictor (8.4 °C) by more than one degree. That result has been read as a ceiling on sequence-only methods. We argue it is a ceiling on *sequence-only inputs*, which is a different claim.

2. Method

We retrain ThermoNet-2 unchanged in architecture, adding two scalar inputs: pH and ionic strength, both recorded in BENCH-TM as covariates and both discarded by every predictor evaluated there. No other change.

3. Results

Table 2. BENCH-TM, sequence-disjoint (n = 318).

Model	Inputs	RMSE (°C)
Mean predictor	—	8.4
ThermoNet-2 as published	sequence	7.8
ThermoNet-2 + pH	sequence, pH	6.9
ThermoNet-2 + pH + ionic strength	sequence, pH, I	6.1
Buffer only (no sequence)	pH, I	7.4

4. The part that matters

This does not help for prospective screening. At design time you are choosing between candidate sequences for a buffer you have not yet selected, and the buffer term is exactly what you do not know. Buffer-aware prediction is useful for interpolating across conditions for a protein you already have, and useless for ranking proteins you do not. Of the 318 proteins in BENCH-TM, 294 have measurements in only one buffer, so the model has no within-protein buffer variation to learn from for most of the set.

5. What we are not claiming

We are not claiming that sequence-only prediction has improved. The sequence-only number is unchanged at 7.8 °C. We are claiming that part of what was attributed to the difficulty of the sequence problem belongs instead to an omitted variable, and that reporting should separate the two.

