Briefings in Bioinformatics, 2026, 27, bbag282
https://doi.org/10.1093/bib/bbag282
Published: 31 May 2026
Letter to Editor

OXFORD

Addressing flaws in the Seq2Topt dataset for the prediction of enzyme optimal temperature

Peng Zhou [1,2,*]

[1] State Key Laboratory of Submarine Geoscience, Second Institute of Oceanography, Ministry of Natural Resources, No. 36 North Bao Chu Road, Xihu District, 310012 Hangzhou, China
[2] Key Laboratory of Marine Ecosystem Dynamics, Ministry of Natural Resources and Second Institute of Oceanography, Ministry of Natural Resources, No. 36 North Bao Chu Road, Xihu District, 310012 Hangzhou, China

*Corresponding author. Second Institute of Oceanography, Ministry of Natural Resources, No. 36 North Bao Chu Road, Xihu District, 310012 Hangzhou, China. E-mail: zhoupeng@sio.org.cn

To the Editor,

I read the article from Qiu *et al.* with great interest, in which the authors developed a deep learning predictor, Seq2Topt, to estimate the optimal temperature (T_{opt}) of enzymes. Seq2Topt achieved a prediction performance with an RMSE of 12.26°C using only protein sequence information [1]. Temperature is one of the critical factors governing enzyme catalytic activity [2]. A comprehensive understanding of T_{opt} facilitates enzyme modification and industrial utilization through accurate T_{opt} prediction. Accordingly, the development of Seq2Topt provides a promising computational tool for enzyme mining and *in silico* enzyme design.

Nevertheless, although Seq2Topt yielded an RMSE of 12.26°C, the predictive performance still has considerable room for improvement. Improved model accuracy would enhance its practical value and applicability in real industrial and research scenarios. As the overall performance of deep learning models is fundamentally determined by the quality of their underlying datasets, I conducted a careful data inspection and verification. The relevant raw files were downloaded from the corresponding GitHub repository of Seq2Topt (https://github.com/SizheQiu/Seq2Topt). The training set accounted for 90% of the total dataset. I sorted the data in the file named "train.csv" in ascending order of annotated T_{opt} values and selected nine entries with low T_{opt} values (<15°C) for literature cross-validation and manual inspection. Surprisingly, rigorous data verification revealed that all nine selected entries were problematic. These nine anomalous records are summarized in Table 1, and the underlying causes of the inconsistent or missing experimental evidence are discussed in detail below.

F9FU71 is a linoleate diol synthase from *Fusarium oxysporum*. Its T_{opt} was annotated as 0°C in the Seq2Topt training set [1]. Nevertheless, available experimental study only indicated incubation on ice during enzyme assay [3], with no credible evidence supporting a T_{opt} of 0°C.

A9U908 and J7HET3 encode superoxide dismutase sodA and sodB in *Yersinia enterocolitica*, respectively. Both enzymes were assigned a T_{opt} of 4°C in the Seq2Topt training set. Although high enzymatic activity at 4°C had been documented in the literature [4], no activity measurements at lower temperatures were available. Accordingly, the reliability of the annotated T_{opt} of 4°C remains questionable.

Table 1 Nine anomalous entries found from Seq2Topt Dataset.

UniProt accession number	T_{opt} in Seq2Topt dataset (°C)	T_{opt} from literature (°C)
F9FU71	0	Not found
A9U908	4	Not found
J7HET3	4	Not found
P94368	5	Not found
Q5YXQ1	7.5	Not found
Q83V33	10	Not found
Q50539	13	Not found
Q50538	13	Not found
Q2WCS9	13	37

P94368 is an ADP-dependent NAD(P)H-hydrate dehydratase from *Bacillus subtilis*, with a labeled T_{opt} of 5°C in the Seq2Topt training set. After comprehensive literature retrieval [5–7], no experimental evidence or conclusive statement was found to confirm a T_{opt} of 5°C for this enzyme.

Q5YXQ1 is a maleate isomerase from *Nocardia farcinica*, recorded with a T_{opt} of 7.5°C in the Seq2Topt training set. However, no study has officially defined its T_{opt} as 7.5°C; only an optimal pH of 7.5 for this enzyme was reported [8]. Hence, the authenticity of this T_{opt} annotation requires further verification.

Q83V33 is an aminomuconate-semialdehyde dehydrogenase from *Pseudomonas fluorescens*. Seq2Topt annotated its T_{opt} as 10°C. In the relevant study, enzyme activity assays were merely conducted at 10°C [9], whereas this temperature was not determined as the optimum. No experimentally validated T_{opt} was found.

Q50538 and Q50539 represent two subunits of the corrinoid/iron–sulfur enzyme of the CO dehydrogenase complex from *Methanosarcina*

Received: May 3, 2026. **Accepted:** May 11, 2026

thermophila, with a unified annotated T_{opt} of 13°C in the Seq2Topt training set. Similar to Q83V33, the relevant literature only confirmed that multiple protein purification steps were carried out at 13°C for heterologously expressed proteins [10], rather than defining 13°C as their T_{opt}.

Q2WCS9 is a glutamate dehydrogenase (GDH) from *Lactiplantibacillus plantarum*, annotated with a T_{opt} of 13°C in the Seq2Topt training set. In contrast, a relevant study showed that the GDH activity of the cytoplasm fraction of *Lactiplantibacillus plantarum* strains tested reached its optimum at 37°C [11], revealing a remarkable data discrepancy.

In summary, all nine entries inspected were identified as anomalous through literature cross-checking. The Seq2Topt study stated that its dataset ($n = 2917$) was retrieved from the GitHub repository of TOMER [12], which was originally obtained from the BRENDA database [13]. Data quality fundamentally affects the training, optimization, and performance evaluation of machine learning models. To mitigate the above data-related issues, the existing T_{opt} annotations should be thoroughly inspected and corrected. Upon careful verification of the T_{opt} annotations, a more reliable dataset is expected.

I hope these remarks contribute to further refinement of enzyme T_{opt} prediction models and promote their more reliable practical applications.

Conflicts of interest

The author declares that there is no conflict of interest.

Funding

None declared.

Data availability

No new data were generated in this study. The letter refers to the data provided by Seq2Topt study [1].

References

1. Qiu S, Hu B, Zhao J *et al.* Seq2Topt: a sequence-based deep learning predictor of enzyme optimal temperature. *Brief Bioinform* 2025;**26**:bbaf114. https://doi.org/10.1093/bib/bbaf114
2. Nelson DL, Cox MM, Hoskins AA. *Lehninger Principles of Biochemistry*8th edn. Austin: Macmillan Learning, 2021.
3. Sooman L, Oliw EH. Discovery of a novel linoleate dioxygenase of *Fusarium oxysporum* and linoleate diol synthase of *Colletotrichum graminicola*. *Lipids* 2015;**50**:1243–52. https://doi.org/10.1007/s11745-015-4078-9
4. Dhar MS, Gupta V, Virdi JS. Detection, distribution and characterization of novel superoxide dismutases from *Yersinia enterocolitica* Biovar 1A. Skurnik M, editor. *PLoS One* 2013;**8**:e63919. https://doi.org/10.1371/journal.pone.0063919
5. Shumilin IA, Cymborowski M, Chertihin O *et al.* Identification of unknown protein function using metabolite cocktail screening. *Structure* 2012;**20**:1715–25. https://doi.org/10.1016/j.str.2012.07.016
6. Petrovova M, Tkadlec J, Dvoracek L *et al.* NAD(P)H-hydrate dehydratase—a metabolic repair enzyme and its role in *Bacillus subtilis* stress adaptation. Appanna VD, editor. *PLoS One* 2014;**9**:e112590. https://doi.org/10.1371/journal.pone.0112590
7. Zhang R, Grembecka J, Vinokour E *et al.* Structure of *Bacillus subtilis* YXKO—a member of the UPF0031 family and a putative kinase. *J Struct Biol* 2002;**139**:161–70. https://doi.org/10.1016/S1047-8477(02)00532-4
8. Fisch F, Fleites CM, Delenne M *et al.* A covalent succinylcysteine-like intermediate in the enzyme-catalyzed transformation of maleate to fumarate by maleate isomerase. *J Am Chem Soc* 2010;**132**:11455–7. https://doi.org/10.1021/ja1053576
9. Huo L, Davis I, Liu F *et al.* Crystallographic and spectroscopic snapshots reveal a dehydrogenase in action. *Nat Commun* 2015;**6**:5935. https://doi.org/10.1038/ncomms6935
10. Maupin-Furlow J, Ferry JG. Characterization of the cdhD and cdhE genes encoding subunits of the corrinoid/iron-sulfur enzyme of the CO dehydrogenase complex from *Methanosarcina thermophila*. *J Bacteriol* 1996;**178**:340–6. https://doi.org/10.1128/jb.178.2.340-346.1996
11. De Angelis M, Calasso M, Di Cagno R *et al.* NADP-glutamate dehydrogenase activity in nonstarter lactic acid bacteria: effects of temperature, pH and NaCl on enzyme activity and expression: GDH activity in nonstarter lactic acid bacteria. *J Appl Microbiol* 2010;**109**:1763–74. https://doi.org/10.1111/j.1365-2672.2010.04804.x
12. Gado JE, Beckham GT, Payne CM. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *J Chem Inf Model* 2020;**60**:4098–107. https://doi.org/10.1021/acs.jcim.0c00489
13. Schomburg I, Jeske L, Ulbrich M *et al.* The BRENDA enzyme information system—from a database to an expert system. *J Biotechnol* 2017;**261**:194–206. https://doi.org/10.1016/j.jbiotec.2017.04.020

Briefings in Bioinformatics, 2026, **27**, bbag282
https://doi.org/10.1093/bib/bbag282
Letter to Editor