RoCO-ET: Structured Conditional Observation Learning for Enzyme Temperature Prediction

Anonymous authors
Paper under double-blind review

Abstract

Predicting temperature properties from enzyme sequences is important for selecting enzymes suited to target temperature conditions and guiding the design of experimental conditions. However, temperature observations are affected by multiple experimental conditions, and the relevant context is incompletely documented. Observations with large deviations can therefore exert excessive influence on model fitting. Predicting active temperature ranges additionally requires learning both range location and span. To address these challenges, we propose RoCO-ET (Robust Conditional Observation for Enzyme Temperature Prediction), which unifies the modeling of scalar temperatures and ordered active temperature ranges in a sequence-conditioned location–scale framework. The model uses conditional scales to measure observation deviations, modulates the influence of large standardized residuals on fitting through a Student-t objective, and treats the two range endpoints jointly as a complete observation during training. A readout parameterized by a center and a positive half-width produces ordered endpoints by construction, allowing record-level modulation to act on both range location and span. On three public tasks, compared with PatchET under the same protocol, RoCO-ET reduces root mean square error (RMSE) by 10.88% for optimum temperature prediction and by 6.46% for operational stability temperature prediction, and increases mean intersection over union (IoU) for active temperature ranges from 0.516 to 0.581. Matched ablations and training interventions further show that sequence-dependent scales and joint modulation of complete records together improve temperature prediction and range fitting.

1 Introduction

Temperature properties directly affect enzyme catalytic performance and the reaction conditions under which an enzyme can be used. Predictions based on amino acid sequences can support candidate enzyme screening and experimental design (Almeida & Marana, 2019; Qiu et al., 2025; Zhang et al., 2026). Practical applications require both estimates of characteristic temperatures, such as optimum temperature and operational stability temperature, and knowledge of the location and span of the active temperature range. Scalar predictions and complete range predictions jointly characterize how well candidate enzymes match target temperature conditions. Reliable learning of these properties also requires accounting for differences among observations that are not captured by the sequence.

Incomplete documentation of assay context makes observation heterogeneity a challenge for learning from sequence–temperature data. Previous work has shown that changes in assay duration and enzyme concentration can affect the observed optimum temperature of the same enzyme (Almeida & Marana, 2019). A review of 100 papers across 10 journals found that 45% did not report enzyme concentration (Halling et al., 2018). A study of data integration in BRENDA also documented missing experimental information, including pH and assay temperature (Figure 1a) (Schomburg et al., 2014). These reporting gaps across the field indicate that sequence inputs do not capture the full context of temperature assays. In addition to extracting sequence information, models therefore need to appropriately modulate the influence of observations with large deviations on fitting.

Existing methods continue to improve sequence representations and task adaptation, but there remains room to improve how observation deviations influence learning. Seq2Topt aggregates residue information through attention (Qiu et al., 2025). PatchET further learns hierarchical representations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

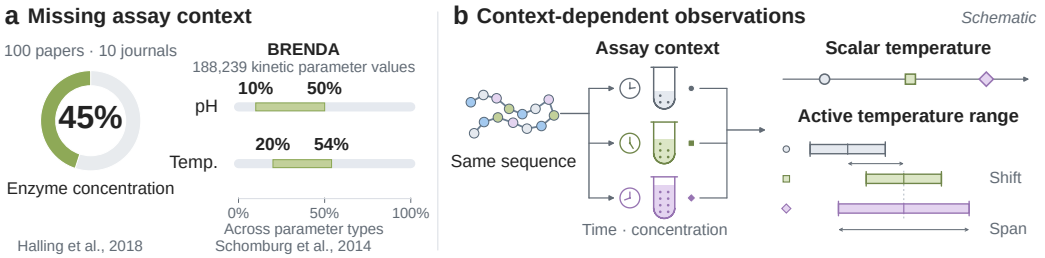

Figure 1: Assay context and temperature targets. (a) Missing assay information in 100 papers (Halling et al., 2018) and 188,239 BRENDA kinetic parameter records (Schomburg et al., 2014). (b) Schematic of context-dependent temperature observations and scalar versus range targets.

through interactions within and between segments, and adapts to different prediction targets using errors weighted by temperature group and endpoint ordering constraints (Zhang et al., 2026). These designs improve the use of sequence features, weighting across temperature groups, and output validity, but group weights and ordering constraints do not directly specify how to measure the deviation of an individual observation relative to a sequence-dependent scale. Thus, beyond improving representations, observation scales and the influence of residuals need to be incorporated into a unified learning process.

Active temperature range prediction further requires treating the complete range as one observation for robust learning. The lower and upper endpoints together determine the range's overall location and span, and a shift of the entire record or a change in its span affects the fitting of both endpoints (Figure 1b). Applying separate robust objectives to the endpoints modulates their individual large residuals but does not share information about the deviation of the complete record. Active temperature range prediction therefore requires combining ordered range generation with joint modulation at the record level, so that location and span are learned under the same observation principle.

To this end, we propose **RoCO-ET** (Robust Conditional Observation for Enzyme Temperature Prediction), whose sequence-conditioned location and scale branches produce temperature predictions and residual scales, respectively, while a Student-t objective modulates the influence of large standardized deviations. For active temperature ranges, a readout parameterized by a center and a positive half-width guarantees ordered endpoints. The residuals of both endpoints enter the same joint observation term to jointly modulate the learning of location and span. Across three public tasks, relative to PatchET under the same protocol, RoCO-ET reduces root mean square error (RMSE) for the two scalar prediction tasks by 10.88% and 6.46%, respectively, and increases mean intersection over union (IoU) for ranges from 0.516 to 0.581. Matched ablations and training interventions further distinguish the roles of conditional scales, joint modulation, and the structured readout.

Our main contributions are as follows:

- **Conditional observation learning.** We propose a unified framework for scalar temperatures and active temperature ranges that incorporates sequence-dependent scales into the measurement of observation deviations and robust fitting.
- **Structured record-level modulation.** We combine ordered range generation with a joint observation objective and characterize how complete observations jointly modulate the learning of range location and span.
- **Predictive and mechanistic validation.** We reduce scalar and endpoint errors on three public tasks and explain the predictive gains through matched ablations, controlled interventions, and range geometry analysis.

2 Related Work

Supervised modeling of enzyme temperature properties. DeepET transfers representations learned from organism growth temperatures to catalytic optimum and protein melting temperature prediction (Li et al., 2022). HotProtein combines growth-temperature annotations with structure-aware pretraining and factorized sparse tuning (Chen et al., 2023), while TemStaPro derives multithreshold

stability labels from growth temperatures (Pudžiuvelytė et al., 2024). Seq2Topt directly regresses catalytic optimum temperature using multi-head attention over residue representations (Qiu et al., 2025). More recently, ETopt combines multimodal features, resampling, and ensemble XGBoost regression for optimum-temperature prediction (Liu et al., 2026). We learn optimum temperature, operational stability temperature, and active temperature range separately.

Sequence representations and output adaptation. Pretrained protein language models provide residue representations (Lin et al., 2023), and Light Attention combines attention weighting with max pooling (Stärk et al., 2021). Segment-Transformer uses multiscale segments and dual-group attention for stability prediction (Zhang et al., 2025). PatchET learns hierarchical segment interactions, uses temperature-group-weighted RMSE for scalar targets, and combines endpoint errors with an ordering penalty for ranges (Zhang et al., 2026). Recent work develops substructure-aware tuning through Magneton (Calef et al., 2026) and benchmarks fine-grained functional representations through VenusX (Tan et al., 2026). These advances refine protein representations; RoCO-ET focuses on how complete temperature observations influence fitting.

Conditional scales and robust observation learning. Conditional location–scale models describe input-dependent residuals and their influence on fitting. Hartmann and Vanhatalo combine Gaussian-process location and log-scale functions with Student-t observations (Hartmann & Vanhatalo, 2019). Seitzer et al. identify optimization pitfalls in Gaussian likelihood training and propose variance-weighted β-NLL (Seitzer et al., 2022). Faithful Heteroscedastic Regression adjusts Gaussian-model gradients to preserve squared-error-driven mean fitting (Stirn et al., 2023). RoCO-ET combines conditional scales, ordered outputs, and robust fitting around complete observations, using the same learning principle for scalar temperatures and active ranges.

3 Problem Formulation

We study the prediction of temperature properties from enzyme amino acid sequences. Given a sequence $x = (a_1, \ldots, a_L) \in \mathcal{X}$, where a_r denotes residue r and L is the sequence length, the goal is to predict a characteristic temperature or the active temperature range of the enzyme. Let $\mathcal{X} = \bigcup_{L \geq 1} \mathcal{A}^L$, where $\mathcal{A}$ is the set of amino acid symbols. Each temperature property is specified according to its own measurement definition, and all are expressed on a consistent temperature scale.

Temperature properties have two types of output structure. Scalar targets are described by a single temperature coordinate, such as the catalytic optimum temperature or a temperature defined by a given stability criterion. Range targets describe a single active range record through its lower and upper endpoints. The corresponding target spaces are

$$\mathcal{Y}_1 = \mathbb{R}, \qquad \mathcal{Y}_2 = \left\{ [y_{\mathrm{L}}, y_{\mathrm{U}}]^\top \in \mathbb{R}^2 : y_{\mathrm{L}} \leq y_{\mathrm{U}} \right\}. \tag{1}$$

The range endpoints describe the boundaries of the observed active range; its overall location and span together constitute the prediction target. For any temperature property τ, given an observation set $\mathcal{D}_\tau = \{(x_i, y_i)\}_{i=1}^{N_\tau}$, where $y_i \in \mathcal{Y}_{k_\tau}$ and $k_\tau \in \{1, 2\}$, we aim to learn

$$f_{\theta_\tau} : \mathcal{X} \longrightarrow \mathcal{Y}_{k_\tau}, \qquad \widehat{y} = f_{\theta_\tau}(x). \tag{2}$$

Supervision is provided at the level of sequence–temperature observation records, and the same sequence may correspond to multiple temperature observations. Differences in experimental context that are not supplied as inputs may remain in the prediction residuals. The learning objective is therefore to establish a predictive relationship from sequences to temperature properties using observation records while preserving the output structure. A separate model instance is trained for each property using the same modeling principle and that property's own observation set and measurement definition. We omit the task subscript below.

4 METHODOLOGY

4.1 FRAMEWORK AND SEQUENCE INTERFACE

RoCO-ET jointly learns **the target's temperature location and structure, and the influence of observation deviations on fitting**. A shared sequence representation feeds a location branch producing scalar temperatures or ordered ranges and a scale branch providing coordinatewise residual scales (Figure 2). A Student-t objective trains both branches on complete observations:

$$
\begin{aligned}
&z = z_\phi(x), \qquad m_\theta(x) = \mathcal{R}_k(h_{\psi_m}(z)) \,, \\
&D_\theta(x) = \mathrm{diag}(\mathcal{S}_{\boldsymbol{g}}(h_{\psi_s}(z))) \,, \\
&p_\theta(y \mid x) = t_\nu\big(y; m_\theta(x), D_\theta(x)^2\big) \,, \qquad \widehat{y}(x) = m_\theta(x).
\end{aligned}
\tag{3}
$$

Here, $\theta = (\phi, \psi_m, \psi_s)$ denotes the learnable parameters, $\mathcal{R}_k$ the structure-preserving location readout, and $\mathcal{S}_{\boldsymbol{g}}$ a bounded positive scale transformation. We use the location–scale Student-t density t_ν with $\nu = 4$, retaining ν in the derivations to expose its role. The framework combines conditional Student-t regression (Hartmann & Vanhatalo, 2019) with ordered outputs and record-level learning.

Shared sequence interface. A frozen ESM2 encoder (Lin et al., 2023) supplies residue representations, followed by a learnable projection and masked local residual processing. Mean and content-attention summaries are fused through a projection and residual MLP into z, shared by the location and scale branches. Pooling equations and implementation details are provided in Appendix B.1; parameters are not shared across temperature properties.

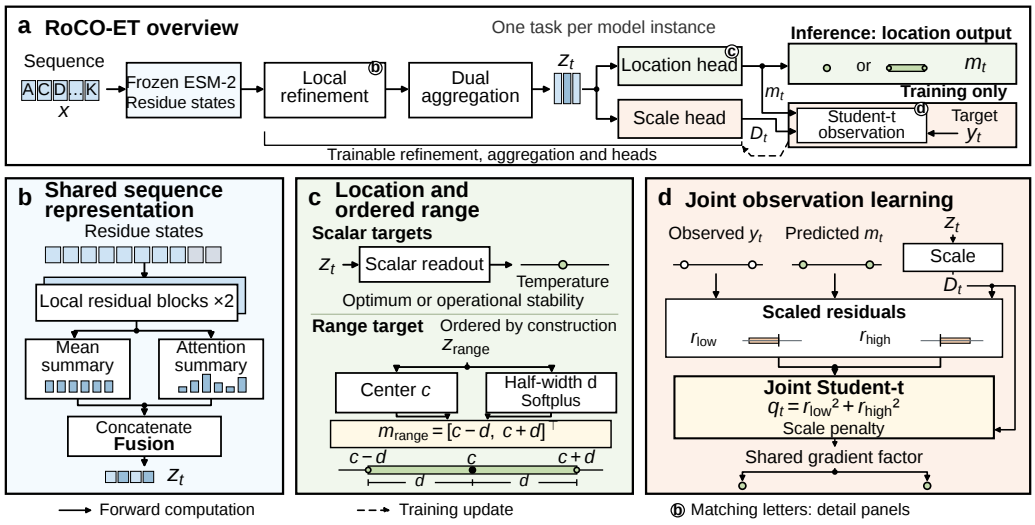

Figure 2: RoCO-ET's structured conditional observation framework.

4.2 STRUCTURE-PRESERVING LOCATION READOUT

The location readout represents a scalar temperature by one coordinate and an active range by its center and positive half-width, preserving endpoint order by construction.

For a scalar target, the location head outputs the dimensionless quantity $u_\mu = h_{\psi_m}(z)$ and maps it back to a temperature coordinate:

$$
m(x) = b + g\, u_\mu(z), \qquad g > 0. \tag{4}
$$

b and g are the training-label median and robust scale anchor, computed and fixed before training.

For a range target, the location head outputs u_c, u_d to construct the center c and positive half-width d, respectively:

$$
\begin{aligned}
&c(x) = b_c + s_T u_c(z), \qquad d(x) = s_T \,\mathrm{softplus}(u_d(z)),\\
&m(x) = A \begin{bmatrix} c(x) \\ d(x) \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad s_T = \frac{g_{\mathrm{L}} + g_{\mathrm{U}}}{2} > 0.
\end{aligned} \tag{5}
$$

Here, b_c is the median training range midpoint, and $g_{\mathrm{L}}, g_{\mathrm{U}}$ are the endpoint scale anchors. The matrix A maps (c, d) to $[c - d, c + d]^\top$: c shifts both endpoints, while d expands or contracts the range. Positive softplus outputs guarantee $m_{\mathrm{L}} < m_{\mathrm{U}}$ and width $2d$, retaining location and span as two degrees of freedom.

4.3 Joint Observation Modeling with Conditional Scales

Equal absolute residuals can have different relative magnitudes across sequences or endpoints. RoCO-ET therefore learns coordinate scales from the sequence and measures standardized deviations for a scalar or complete endpoint pair.

The scale head outputs $a_j(z)$ and applies a bounded transformation to produce a scale for each coordinate:

$$
\begin{aligned}
&s_j(x) = g_j[\lambda_- + (\lambda_+ - \lambda_-)\sigma(a_j(z))], \qquad 0 < \lambda_- < \lambda_+ < \infty,\\
&D(x) = \mathrm{diag}(s_1(x), \ldots, s_k(x)).
\end{aligned} \tag{6}
$$

σ is the sigmoid function, and $g_j > 0$ is determined solely from the training labels; we set $\lambda_- = 0.05$ and $\lambda_+ = 3$. The positive lower bound prevents scale collapse to zero, while the finite upper bound prevents unbounded scale inflation. The residual scale s_j measures observation deviations, whereas the range half-width d determines the predicted span.

For an observation (x, y), define the residual, standardized residual, and deviation measure for the complete record as follows:

$$
e = y - m(x), \qquad r = D(x)^{-1} e, \qquad q(x, y) = \|r\|_2^2 = e^\top D(x)^{-2} e. \tag{7}
$$

The corresponding negative log-likelihood for a single observation is

$$
\ell_\nu(x, y) = C_{\nu,k} + \log\det D(x) + \frac{\nu + k}{2} \log\left(1 + \frac{q(x, y)}{\nu}\right), \tag{8}
$$

where $C_{\nu,k}$ depends only on ν and k. Scalar targets use $k = 1$; ranges use $k = 2$, combining $q = r_{\mathrm{L}}^2 + r_{\mathrm{U}}^2$ within one logarithm. Even with diagonal D, this is a multivariate Student-t term, not a sum of two univariate terms. The log-scale and standardized residual terms jointly determine scale updates.

The equivalent Gaussian–Gamma representation gives both endpoints a shared precision factor determined by the complete record (Appendix A.1). Training directly evaluates Equation (8), without outputting or sampling the auxiliary precision.

4.4 Record-Level Modulation

The joint observation objective directly modulates the influence of residuals on fitting through location gradients. Holding the positive scales fixed and taking partial derivatives of Equation (8) with respect to the location outputs gives

$$
\nabla_m \ell_\nu = -w_\nu(q) D^{-2} e, \qquad \frac{\partial \ell_\nu}{\partial m_j} = -w_\nu(q) \frac{e_j}{s_j^2}, \qquad w_\nu(q) = \frac{\nu + k}{\nu + q}. \tag{9}
$$

The coordinate-specific term e_j / s_j^2 retains each endpoint's deviation and scale, while $w_\nu(q)$ modulates all coordinates using the complete record. For ranges with $\nu = 4$, the shared factor is $6/(4 + r_{\mathrm{L}}^2 + r_{\mathrm{U}}^2)$. Holding one endpoint's nonzero residual and both positive scales fixed, increasing

the other endpoint's standardized deviation reduces the magnitude of the first endpoint's location gradient.

Redescending location response at large deviations. With the positive scales held fixed, the scale-standardized magnitude of the derivative with respect to the location output satisfies

$$\|D\nabla_m \ell_\nu\|_2 = \frac{(\nu+k)\sqrt{q}}{\nu+q} \le \frac{\nu+k}{2\sqrt{\nu}}, \qquad \lim_{q\to\infty} \|D\nabla_m \ell_\nu\|_2 = 0. \tag{10}$$

The magnitude peaks at $q = \nu$ and then decreases, so extreme standardized deviations exert progressively less influence on the location output.

Modulation of location and span. Let the observed center and half-width be $c_y = (y_{\mathrm{L}} + y_{\mathrm{U}})/2$ and $d_y = (y_{\mathrm{U}} - y_{\mathrm{L}})/2$. Then

$$\begin{aligned} &v = \begin{bmatrix} c \\ d \end{bmatrix}, \qquad \delta = \begin{bmatrix} c_y - c \\ d_y - d \end{bmatrix}, \qquad e = A\delta, \\ &q = \delta^\top G(x)\delta, \qquad G(x) = A^\top D(x)^{-2} A, \\ &\boxed{\nabla_v \ell_\nu = -w_\nu\big(\delta^\top G(x)\delta\big)\, G(x)\delta.} \end{aligned} \tag{11}$$

$G(x)$ is the residual metric induced by the endpoint scales and coordinate transformation. For equal endpoint scales s, $q = 2[(c_y - c)^2 + (d_y - d)^2]/s^2$. Thus, center and half-width deviations jointly determine the shared modulation factor, connecting ordered range generation to robust fitting of location and span. Appendix A.3 gives the derivation.

4.5 Training and Inference

Location and scale gradients jointly update the shared representation. The log-scale derivative and the two gradient paths are derived in Appendix A.4. Identical sequences yield identical conditional locations and scales, while different labels affect fitting through their respective q values.

For each temperature property, we independently minimize the mean negative log-likelihood over the training records:

$$\mathcal{L}(\theta) = \frac{1}{N_{\mathrm{tr}}} \sum_{(x_i, y_i) \in \mathcal{D}_{\mathrm{tr}}} \ell_4(x_i, y_i; \theta). \tag{12}$$

ESM2 remains frozen; the sequence aggregation module and both heads are jointly trained from random initialization using Equation (12) as the sole prediction loss. At inference, only the sequence is required: $\widehat{y}(x) = m_\theta(x)$ outputs a scalar or $[c - d, c + d]^\top$. Scales modulate training residuals; the location branch determines the final prediction. Algorithm A1 summarizes the procedure.

5 Experiments

5.1 Experimental Setup

Data and tasks. We use curated BRENDA-derived data for optimum temperature and active temperature range prediction, and the existing benchmark's label definition and partition for operational stability (Chang et al., 2021; Zhang et al., 2026; 2025). Optimum temperature, operational stability, and active range prediction use 10,370, 3,454, and 1,818 sequence–temperature records, respectively, with partitions in Table A1 and preprocessing in Appendix B.1.

Comparison methods and statistics. Table 1 presents representative published benchmarks; Table A2 contains the complete published comparisons. All published scores come from Zhang et al. (2026). We separately compare local PatchET, matched Huber, and RoCO-ET under the same protocol, reporting means and sample standard deviations over three fixed seeds, with ablations and interventions paired by seed. Per-record range analyses use seed 1 and all 182 test records.

Table 1: Performance on three enzyme-temperature tasks.

Method	Scalar temperature				Active temperature range			
	Opt. RMSE ↓	Opt. MAE ↓	Stab. RMSE ↓	Stab. MAE ↓	JRMSE ↓	IoU ↑	MOR ↑	W-MAE ↓
DeepET†	14.960	11.790	30.200	24.660	–	–	–	–
Seq2Topt†	15.670	11.590	25.660	19.440	–	–	–	–
Segment-Transformer†	–	–	24.030	18.090	–	–	–	–
PatchET†	**13.880**	**9.690**	**22.660**	**17.310**	15.997	–	0.707	–
PatchET (local)	13.913 ±0.075	10.429 ±0.233	22.440 ±0.147	17.211 ±0.571	15.590 ±0.105	0.516 ±0.009	0.656 ±0.019	12.712 ±0.217
Huber	13.243 ±0.085	9.862 ±0.172	22.173 ±0.131	16.620 ±0.237	15.320 ±0.082	0.531 ±0.018	0.680 ±0.023	12.654 ±0.236
RoCO-ET	**12.400** ±0.056	**9.296** ±0.041	**20.990** ±0.110	**15.835** ±0.130	**14.307** ±0.067	**0.581** ±0.009	**0.774** ±0.005	**10.697** ±0.103

†Published reference (Zhang et al., 2026); full results in Table A2. Local cells: mean above ± sample SD (three seeds). Errors: °C. Bold/underline: best/second-best within each block. –: not shown.

Training and evaluation. We freeze ESM2-150M and train each task independently with AdamW, selecting models on validation data before test evaluation. Scalar metrics are RMSE and mean absolute error (MAE), with Pearson and Spearman correlations additionally reported for optimum temperature. Range metrics are joint endpoint RMSE (JRMSE), mean interval IoU, ground-truth range coverage (MOR), and center/width MAE. MOR measures ground-truth coverage; IoU also accounts for predicted length beyond the true boundaries, distinguishing overlap from mere widening. Temperature errors are in degrees Celsius. Appendix B provides implementation, optimization, selection rules, and metric definitions.

5.2 Overall Prediction Performance

RoCO-ET reduces scalar temperature errors and improves active temperature range quality. Compared with PatchET under the same protocol, it reduces RMSE by 10.88% and 6.46% on the two scalar tasks, respectively, while also lowering MAE (Table 1). For range prediction, JRMSE decreases from 15.590 to 14.307 °C, IoU increases from 0.516 to 0.581, and width MAE decreases from 12.712 to 10.697 °C. These gains span endpoint fitting, overall overlap, and span quality. The primary error metrics on all three tasks are also lower than the published reference results in the table. The following experiments distinguish the contributions of the joint objective, conditional scales, and output structure.

5.3 Effect of Record-Level Modulation

Joint modulation based on the complete record helps address deviations in the overall location or span of a temperature range. We use Independent-t, which retains the same conditional scales and ordered readout, as a direct control and apply paired interventions to 20% of the training range records. The analytical response shows that increasing the deviation at one endpoint also weakens the joint objective's location response at the other endpoint (Figure 3a). Under the strongest whole-range shift, the increase in JRMSE relative to each method's own reference without additional perturbation is 0.907 °C for RoCO-ET, compared with 1.192 °C for Independent-t (Figure 3b).

Under the strongest shift and span interventions, the joint objective achieves IoU gains of 0.040 and 0.046, respectively, with gains in the same direction across all three paired seeds (Figure 3c). Independent-t performs better under single-endpoint interventions, indicating that independent coordinatewise modulation is better suited to localized distortions, whereas the main gains from the joint objective arise for changes affecting the observation as a whole. Appendix C.1 reports absolute errors under strong interventions and analyses of the unperturbed endpoint.

5.4 Observation and Readout Ablations

Matched ablations show that a joint objective for complete records provides gains beyond independent heavy-tailed fitting. Although Independent-t retains conditional scales and ordered

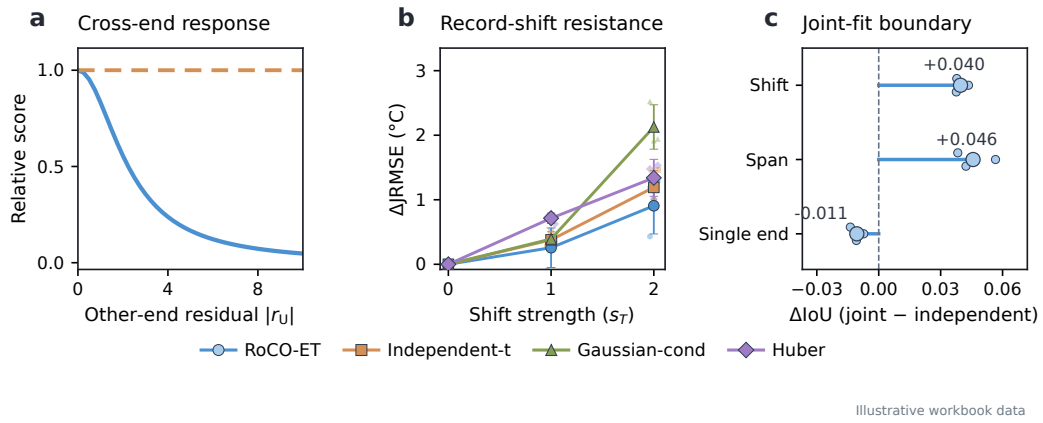

Figure 3: Analytical responses and training interventions. (a) Lower-endpoint response normalized at $r_U = 0$, with $r_L = 1$, $\nu = 4$, and fixed positive scales. (b) Test JRMSE changes under whole-range shifts relative to each method's unperturbed reference. (c) Paired IoU differences (joint minus independent) at the strongest interventions. Small/large markers: three seed results/means; error bars: sample SD.

outputs, RoCO-ET further reduces range JRMSE by 0.623 °C while improving both IoU and MOR (Table 2). This comparison changes only how the two endpoint residuals enter the objective. Together with the record-level intervention results, it supports the predictive value of joint modulation.

The primary errors on all three tasks decrease progressively from T-fixed to T-global to RoCO-ET, indicating that sequence-dependent scales outperform fixed scales or global constant scales. Gaussian-cond remains weaker than the full model despite using the same conditional scale parameterization, supporting the role of the Student-t observation objective. Replacing the center–positive-half-width readout with a direct endpoint readout and an ordering penalty increases JRMSE to 14.747 °C and reduces IoU to 0.552, indicating that the center–positive-half-width parameterization helps fit the complete range. Variant definitions and comparisons across output structures are provided in Appendices B.2 and C.2.

Table 2: Ablations of the observation objective, scales, and readout.

Variant	Opt. RMSE ↓	Stab. RMSE ↓	Range JRMSE ↓	IoU ↑	MOR ↑
RoCO-ET	**12.400** ± 0.056	**20.990** ± 0.110	**14.307** ± 0.067	**0.581** ± 0.009	**0.774** ± 0.005
Independent-t	–	–	14.930 ± 0.075	0.545 ± 0.004	0.711 ± 0.014
Gaussian-cond	13.107 ± 0.070	21.950 ± 0.130	15.187 ± 0.085	0.539 ± 0.008	0.673 ± 0.001
T-global	<u>12.723</u> ± 0.091	<u>21.497</u> ± 0.122	14.767 ± 0.061	<u>0.560</u> ± 0.003	<u>0.739</u> ± 0.013
T-fixed	12.967 ± 0.067	21.817 ± 0.120	14.983 ± 0.080	0.542 ± 0.012	0.707 ± 0.015
Direct-joint	–	–	<u>14.747</u> ± 0.106	0.552 ± 0.019	0.720 ± 0.037

Mean ± sample SD over three seeds; errors in °C. RoCO-ET runs match Table 1. Variant definitions: Appendix B.2.

5.5 Range Location and Span Quality

Range prediction gains involve both improved overlap and reduced span errors. Among the 182 paired records evaluated with the fixed seed 1, 98 have higher IoU, 110 have lower width error, and 69 improve on both measures (Figure 4a). These per-record results connect aggregate metric gains to changes in individual ranges; representative patterns are shown in Appendix F.

The global widening control further distinguishes increased coverage from range fitting. Multiplying the PatchET half-width by 1.25 raises MOR to 0.759, close to RoCO-ET's 0.768, but IoU reaches only 0.525, below RoCO-ET's 0.573, and width MAE increases to 17.247 °C (Figure 4b). RoCO-ET achieves a better combination of coverage and overlap, consistent with jointly learning location

and span. Its predictive advantage is also retained under sequence-group holdout and a second representation interface. Appendices D–E report groupwise performance, parameter analyses, and model size.

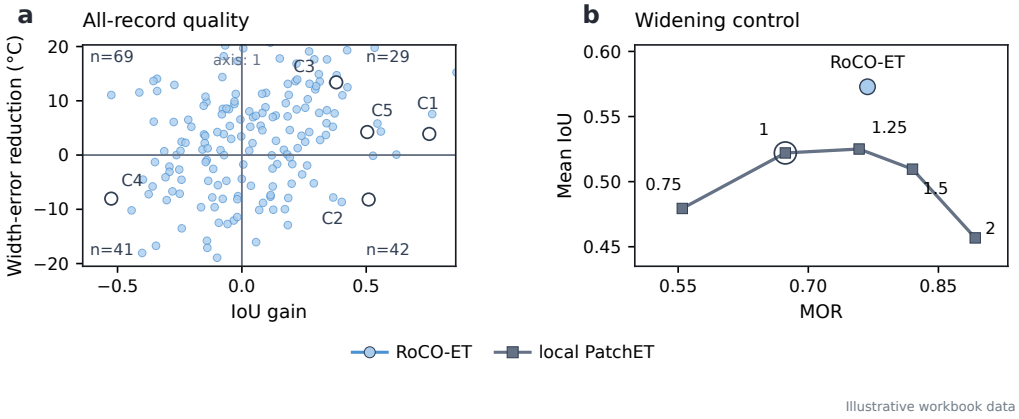

Figure 4: Range predictions for 182 test records (seed 1). (a) IoU differences (RoCO-ET minus PatchET) versus width-error differences (PatchET minus RoCO-ET). (b) MOR and IoU with fixed PatchET centers and scaled half-widths. Numbers denote multipliers; the ring marks the validation-selected $\alpha = 1.00$.

6 Conclusion

We propose RoCO-ET, which combines sequence-conditioned scales with record-level robust fitting to support both scalar temperature and ordered active temperature range prediction. For ranges, the center–positive-half-width readout and joint Student-t objective allow the complete observation to jointly modulate learning of location and span. Comparisons across three tasks, matched ablations, and training interventions show that this design reduces prediction errors and improves range overlap and span quality. These results highlight that, for heterogeneous temperature observations, explicitly modeling how observation deviations affect learning provides an effective modeling approach alongside stronger sequence representations.

Reproducibility Statement

Section 4 describes the model readout, observation objective, and training and inference procedures, with mathematical derivations and pseudocode in Appendix A. Appendix B provides data partitions, implementation details, variant definitions, and evaluation metrics. Appendices C–E cover interventions, supplementary evaluation settings, parameter analyses, and model size; Appendix F presents range examples and the calculations for the widening control.

Statement on AI Use

Generative AI tools were used to improve the clarity and readability of the manuscript and to assist with limited code editing and debugging for the experiments. All AI-assisted text and code were reviewed and verified by the authors.

References

Vitor M. Almeida and Sandro R. Marana. Optimum temperature may be a misleading parameter in enzyme characterization and application. *PLOS ONE*, 14(2):e0212977, 2019. doi: 10.1371/journal.pone.0212977.

Robert Calef, Arthur Liang, Manolis Kellis, and Marinka Zitnik. Greater than the Sum of Its Parts: Building Substructure into Protein Encoding Models. In *International Conference on Learning Representations*, pp. 64192–64228, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6901d8b7969d4377f543f6ea7a92eb8d-Paper-Conference.pdf.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 49(D1):D498–D508, 2021. doi: 10.1093/nar/gkaa1025.

Tianlong Chen, Chengyue Gong, Daniel Jesus Diaz, Xuxi Chen, Jordan Tyler Wells, Qiang Liu, Zhangyang Wang, Andrew Ellington, Alex Dimakis, and Adam Klivans. HotProtein: A Novel Framework for Protein Thermostability Prediction and Editing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YDJRFWBMNby.

Peter Halling, Paul F. Fitzpatrick, Frank M. Raushel, Johann Rohwer, Santiago Schnell, Ulrike Wittig, Roland Wohlgemuth, and Carsten Kettner. An empirical analysis of enzyme function reporting for experimental reproducibility: Missing/incomplete information in published papers. *Biophysical Chemistry*, 242:22–27, 2018. doi: 10.1016/j.bpc.2018.08.004.

Marcelo Hartmann and Jarno Vanhatalo. Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic Student-t model. *Statistics and Computing*, 29(4):753–773, 2019. doi: 10.1007/s11222-018-9836-0.

Gang Li, Filip Buric, Jan Zrimec, Sandra Viknander, Jens Nielsen, Aleksej Zelezniak, and Martin K. M. Engqvist. Learning deep representations of enzyme thermal adaptation. *Protein Science*, 31 (12):e4480, 2022. doi: 10.1002/pro.4480.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Chunhui Liu, Bolun Li, Hong Jiang, Jianan Sun, Jiayi Ren, Xiujuan Huang, Peng Xu, Mengxiao Li, Yu Xin, Suxue Chen, and Xiangzhao Mao. Accelerated discovery of thermophilic biocatalysts for marine polysaccharide bioprocessing via ETopt. *International Journal of Biological Macromolecules*, 351:151068, 2026. doi: 10.1016/j.ijbiomac.2026.151068.

Ieva Pudžiuvelytė, Kliment Olechnovič, Egle Godliauskaite, Kristupas Sermokas, Tomas Urbaitis, Giedrius Gasiunas, and Darius Kazlauskas. TemStaPro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics*, 40(4):btae157, 2024. doi: 10.1093/bioinformatics/btae157.

Sizhe Qiu, Bozhen Hu, Jing Zhao, Weiren Xu, and Aidong Yang. Seq2Topt: a sequence-based deep learning predictor of enzyme optimal temperature. *Briefings in Bioinformatics*, 26(2):bbaf114, 2025. doi: 10.1093/bib/bbaf114.

Ida Schomburg, Antje Chang, and Dietmar Schomburg. Standardization in enzymology—Data integration in the world's enzyme information system BRENDA. *Perspectives in Science*, 1(1–6): 15–23, 2014. doi: 10.1016/j.pisc.2014.02.002.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021. doi: 10.1093/bioadv/vbab035.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful Heteroscedastic Regression with Neural Networks. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5593–5613. PMLR, 2023. URL https://proceedings.mlr.press/v206/stirn23a.html.

Yang Tan, Wenrui Gou, Bozitao Zhong, Huiqun Yu, Liang Hong, and Bingxin Zhou. VenusX: Unlocking Fine-Grained Functional Understanding of Proteins. In *International Conference on Learning Representations*, pp. 44475–44507, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/49c1879ae366644ce2c17fb39ddea982-Paper-Conference.pdf.

Ziqi Zhang, Shiheng Chen, Runze Yang, Zhisheng Wei, Wei Zhang, Lei Wang, Zhanzhi Liu, Fengshan Zhang, Jing Wu, Xiaoyong Pan, Hongbin Shen, Longbing Cao, and Zhaohong Deng. Modeling Enzyme Temperature Stability from Sequence Segment Perspective. *Journal of Chemical Information and Modeling*, 65(20):10932–10944, 2025. doi: 10.1021/acs.jcim.5c01674.

Ziqi Zhang, Runze Yang, Longbing Cao, and Zhaohong Deng. PatchET: Learning Enzyme Temperature Properties Through Patch-Based Neural Architectures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28671–28679, 2026. doi: 10.1609/aaai.v40i34.40099.

A Mathematical Derivations and Training Procedure

A.1 Shared-Precision Representation of the Joint Student-t Model

Let $D = \mathrm{diag}(s_1, \ldots, s_k)$ with $s_j > 0$. The Student-t observation density is

$$p(y \mid x) = \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2}\det D}\left(1+\frac{q}{\nu}\right)^{-(\nu+k)/2}, \qquad q = (y-m)^\top D^{-2}(y-m). \tag{A1}$$

Its negative logarithm gives Equation (8) in the main text, where

$$C_{\nu,k} = \frac{k}{2}\log(\nu\pi) + \log\Gamma(\nu/2) - \log\Gamma((\nu+k)/2). \tag{A2}$$

Using the shape–rate parameterization, introduce a shared precision variable:

$$\omega \sim \mathrm{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \qquad Y \mid x, \omega \sim \mathcal{N}\left(m(x), \frac{D(x)^2}{\omega}\right). \tag{A3}$$

The terms involving ω in the product of the Gaussian conditional density and the Gamma density are

$$p(y, \omega \mid x) \;\propto\; \omega^{(\nu+k)/2-1}\exp\left[-\frac{\nu+q}{2}\omega\right]. \tag{A4}$$

Integrating over ω yields Equation (A1). The posterior distribution of the shared precision is

$$\omega \mid x, y \sim \mathrm{Gamma}\left(\frac{\nu+k}{2}, \frac{\nu+q}{2}\right), \qquad \mathbb{E}[\omega \mid x, y] = \frac{\nu+k}{\nu+q} = w_\nu(q). \tag{A5}$$

The two range endpoints share the same precision variable, so the marginalized observation term depends jointly on both endpoint residuals even when D is diagonal. This representation explains the origin of the record-level factor; training evaluates Equation (8) directly.

A.2 Boundedness and Redescending Behavior of the Location Response

Holding D fixed and using $\nabla_m q = -2D^{-2}e$, we obtain

$$\nabla_m \ell_\nu = \frac{\nu + k}{2(\nu + q)} \nabla_m q = -\frac{\nu + k}{\nu + q} D^{-2} e. \tag{A6}$$

Let $u = \|D^{-1}e\|_2 = \sqrt{q} \geq 0$. Then

$$\|D\nabla_m \ell_\nu\|_2 = h(u), \qquad h(u) = \frac{(\nu + k)u}{\nu + u^2}, \qquad h'(u) = \frac{(\nu + k)(\nu - u^2)}{(\nu + u^2)^2}. \tag{A7}$$

Thus, h attains its maximum at $u = \sqrt{\nu}$, with value $(\nu + k)/(2\sqrt{\nu})$, and tends to zero as $u \to \infty$, giving Equation (10) in the main text.

For the cross-endpoint response of a range, hold $r_{\mathrm{L}} = 1$, $\nu = 4$, and the positive scales fixed. The magnitude of the lower-endpoint derivative of the joint objective is proportional to $6/(5 + r_{\mathrm{U}}^2)$. Normalizing by its value at $r_{\mathrm{U}} = 0$ gives

$$\frac{|\partial \ell_{\mathrm{joint}}/\partial m_{\mathrm{L}}|}{|\partial \ell_{\mathrm{joint}}/\partial m_{\mathrm{L}}|_{r_{\mathrm{U}}=0}} = \frac{5}{5 + r_{\mathrm{U}}^2}. \tag{A8}$$

For the independent-endpoint objective, the lower-endpoint derivative depends only on r_{L}, so its response under the same normalization is always 1, as shown in Figure 3a.

A.3 Joint Learning in Center–Half-Width Coordinates

Let the observed center and half-width be $c_y = (y_{\mathrm{L}} + y_{\mathrm{U}})/2$ and $d_y = (y_{\mathrm{U}} - y_{\mathrm{L}})/2$, respectively, and define

$$v = \begin{bmatrix} c \\ d \end{bmatrix}, \qquad \delta = \begin{bmatrix} c_y - c \\ d_y - d \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}. \tag{A9}$$

Since $e = A\delta$, the joint squared standardized residual of a complete record and the corresponding gradient with respect to the location coordinates are

$$\begin{aligned} &q = \delta^\top G(x)\delta, \qquad G(x) = A^\top D(x)^{-2} A, \\ &\nabla_v \ell_\nu = -w_\nu(\delta^\top G\delta) G\delta. \end{aligned} \tag{A10}$$

The endpoint scales and the fixed coordinate transformation jointly determine G. Writing $a = s_{\mathrm{L}}^{-2}$ and $b = s_{\mathrm{U}}^{-2}$, we can expand it as

$$G = \begin{bmatrix} a + b & b - a \\ b - a & a + b \end{bmatrix}. \tag{A11}$$

When the endpoint scales are equal, $G = 2s^{-2}I$, and the sum of squared center and half-width residuals determines w_ν. When the scales differ, G retains the coordinate relationships induced by the endpoint scales, and location and span still share the same record-level factor.

A.4 Conditional Scales and Gradients of the Shared Representation

Write $\eta_j = \log s_j$. Using $\log\det D = \sum_j \eta_j$ and $\partial q/\partial \eta_j = -2r_j^2$, we obtain

$$\frac{\partial \ell_\nu}{\partial \eta_j} = 1 - w_\nu(q) r_j^2. \tag{A12}$$

The log-scale term and standardized residuals jointly drive scale updates. Letting $J_m = \partial m/\partial z$ and $J_\eta = \partial \eta/\partial z$, we further obtain

$$\nabla_z \ell_\nu = -J_m^\top w_\nu(q) D^{-2} e + J_\eta^\top \left[\mathbf{1} - w_\nu(q)(r \odot r)\right]. \tag{A13}$$

The first term updates the representation through the location branch, and the second through the scale branch. Identical sequences produce identical conditional locations and scales, whereas different observed labels affect fitting in a record-dependent manner through their respective values of q. The gradient-path experiment in Appendix C.2 directly compares training through both paths with blocking the second path's gradient to the representation.

A.5 Training and Inference from the Location Output

Algorithm A1 Independent Training and Inference for Each RoCO-ET Task

Require: Training set $\mathcal{D}_{\mathrm{tr}}$, validation set $\mathcal{D}_{\mathrm{val}}$, frozen residue encoder, optimization configuration; output dimension k.

Ensure: Location prediction model selected using the validation set.

1: Compute the location reference b or b_c and positive scale anchors g_j from the training labels.
2: Initialize the trainable projection, local residual processing, dual aggregation module, location head, and scale head.
3: For each training epoch and each batch within that epoch:
   a. Extract frozen residue representations and obtain z through the trainable sequence interface.
   b. Generate the scalar location or ordered temperature range $m(x)$ using Equation (4) or Equation (5).
   c. Generate conditional scales $D(x)$ using Equation (6).
   d. Compute the residuals and q using Equation (7), and average the negative log-likelihood in Equation (8) over the batch.
   e. Update the task network parameters while retaining both the location and scale gradient paths.
4: Select the model using the validation metric for the corresponding task and apply early stopping (see Appendix B.1).
5: At inference, input a sequence and directly return the location output $m(x)$.

B Experimental Settings, Evaluation, and Complete Comparisons

B.1 Data and Implementation

Each of the three tasks uses its own fixed data partition, and counts refer to sequence–temperature observation records. The optimum temperature and active temperature range records underwent filtering for valid annotations, accession splitting, and deduplication; the operational stability task retains the label definition of the existing benchmark. The local optimum temperature training set contains 8,346 records, one fewer than the published reference. Table A1 lists the actual data partitions and training settings.

Sequence interface and task network. We use frozen ESM2-150M (`facebook/esm2_t30_150M_UR50D`) and extract the 640-dimensional residue states from layer 30. Sequences are converted to uppercase and stripped of whitespace, retaining at most the first 998 N-terminal residues; special tokens and padding positions are excluded from pooling. The residue states are projected to 384 dimensions and passed through two masked local residual blocks. Both blocks use a convolution kernel size of 9, with dilation rates of 1 and 3, respectively, and a $384 \to 768 \to 384$ channel transformation. A mask for valid residues is applied both before and after local processing.

Let $H_r \in \mathbb{R}^h$ denote the locally processed representation and V the set of valid residues. Mean and content-attention summaries are computed as

$$
\begin{aligned}
p_{\mathrm{mean}} &= \frac{1}{|V|}\sum_{r\in V} H_r, \qquad \alpha_r = \frac{\exp(\psi(H_r))}{\sum_{v\in V}\exp(\psi(H_v))}, \\
p_{\mathrm{att}} &= \sum_{r\in V} \alpha_r H_r, \qquad z = \mathrm{Fusion}([p_{\mathrm{mean}}; p_{\mathrm{att}}])\,.
\end{aligned}
\tag{B1}
$$

The scalar attention function ψ uses a 96-dimensional hidden layer and a $\tanh$ activation. Mean pooling preserves the overall residue state, while attention pooling weights residues by sequence content. The concatenated 768-dimensional summary is projected back to 384 dimensions by Fusion and then fused by a residual MLP. The location and scale heads are two-layer MLPs with dimensions $384 \to 192 \to k$ and $384 \to 128 \to k$, respectively, where $k = 1$ for scalar tasks and $k = 2$ for the range task. The local modules and prediction heads use GELU and a dropout rate of 0.1, Fusion uses a dropout rate of 0.15, and the modules incorporate LayerNorm. Models for the three tasks are initialized and trained separately; the location and scale branches share a representation only within the same task.

Training-label references. The location references and scale anchors are computed from the training labels and fixed before training. For each coordinate, let b_j denote its median. We use the normal-consistent median absolute deviation for all coordinates:

$$
b_j = \mathrm{median}_{1\le i\le N_{\mathrm{tr}}}\, y_{ij}, \qquad g_j = \frac{\mathrm{median}_{1\le i\le N_{\mathrm{tr}}}\,|y_{ij} - b_j|}{\Phi^{-1}(0.75)},
\tag{B2}
$$

where Φ is the standard normal cumulative distribution function. A lower bound of 0.1 °C is imposed on the scale anchors. The values of g_j for optimum temperature, stability, and the lower and upper range endpoints are 10.378, 19.274, 14.826, and 22.239 °C, respectively, and the range location scale is $s_T = 18.533$ °C; all calculations use unrounded values. Conditional scales use the bounds $(0.05g_j, 3g_j)$ in Equation (6), with g_j as the initialization reference. The range center and half-width use initialization references of 45 °C and 15 °C, respectively, obtained from the training statistics.

Optimization and model selection. The main comparisons retain the task-specific learning rates listed in Table A1. All matched supplementary experiments use AdamW with a learning rate of 3×10^{-4}, training and evaluation batch sizes of 128 and 256, respectively, a maximum of 36 epochs, and an early-stopping patience of 8. Weight decay is 0.01, the gradient norm is clipped at 1.0, and linear warmup and linear decay are used. Optimization settings are held constant within each paired comparison. For supplementary experiments, scalar models are selected by the lowest validation RMSE, with ties broken by lower MAE; range models are selected by the lowest validation JRMSE, with ties broken by higher MOR. Any remaining ties are resolved in favor of the earlier epoch. Parameters are updated using only the training set, and the test set is used for final evaluation after loading the selected model.

Local experiments use seeds 1, 2, and 3. Test metrics are computed for each run and then summarized by the mean and sample standard deviation. Intervention and parameter analyses use differences paired by seed; experiments on gradient paths, data partitions, and representation interfaces each use their own paired reference. Per-record range analyses consistently use seed 1.

B.2 Comparison Methods and Matched Variants

Published comparisons use the values reported in the benchmark paper. Comparisons under the same protocol separately evaluate local PatchET, matched Huber, and RoCO-ET. Matched variants retain the same cached residue representations, local processing, pooling interface, and paired training settings while varying the observation objective, scale formulation, range readout, or gradient path.

Joint and independent observations. The full model uses the joint Student-t objective in Equation (8). Independent-t retains the conditional scales and the center–positive-half-width readout but replaces the range objective with the sum of two univariate Student-t terms:

$$
\ell_{\mathrm{ind}} = \sum_{j\in\{\mathrm{L,U}\}} \left[C_{\nu,1} + \log s_j + \frac{\nu+1}{2}\log\left(1 + \frac{r_j^2}{\nu}\right)\right], \qquad \nu = 4.
\tag{B3}
$$

Table A1: Data partitions and implementation settings.

Task	Train	Validation	Test
opt	8346	987	1037
stability	2798	311	345
range	1464	172	182

Setting	Value
Residue encoder	Frozen ESM2-150M; layer 30, 640 dimensions
Visible N-terminal prefix	998 residues
Training / evaluation batch	128 / 256
Maximum epochs / patience	36 / 8
Main comparison LR: optimum	1.5×10^{-4}
Main comparison LR: stability / range	3.0×10^{-4}
Matched supplementary LR	3.0×10^{-4}
Optimizer / weight decay	AdamW / 0.010
Fusion dropout	0.150
Degrees of freedom	$\nu = 4$
Conditional scale bounds	$(0.050g_j, 3.000g_j)$
Model seeds	1, 2, 3
Supplementary selection: scalar / range	Validation RMSE / JRMSE

Counts denote sequence–temperature observation records.

Each endpoint is modulated independently by its own residual, in direct contrast to the full model's shared record-level factor.

Fixed, global, and conditional scales. T-fixed fixes $s_j = g_j$. T-global learns one input-independent global scale for each output coordinate, using the same sigmoid bounds as the full model and g_j as the initialization reference; the original input-dependent scale head remains frozen. The full model instead generates $s_j(x)$ from the sequence representation. The Huber control fixes the scales at g_j and applies a Huber objective with threshold 1 to the standardized residuals $\widetilde{r}_j = (y_j - m_j)/g_j$. The objective is summed over coordinates and averaged over records, without a log-scale term; the threshold corresponds to one training scale anchor for each coordinate.

Gaussian observation controls. Gaussian-cond and Gaussian-fixed use conditional scales and fixed scales $s_j = g_j$, respectively. Both use covariance $2D^2$, with the observation objective

$$\ell_{\mathrm{G}} = C_{\mathrm{G},k} + \log\det D + \frac{1}{4}\sum_{j=1}^{k} r_j^2, \qquad r = D^{-1}(y - m). \tag{B4}$$

For the same D, this Gaussian parameterization has the same covariance as the Student-t model with $\nu = 4$, preserving a consistent second-order scale convention in the comparison.

Direct endpoint readouts. Direct-joint and Direct-independent generate the lower and upper endpoints directly as $m_j = b_c + s_T u_j(z)$, using a joint bivariate Student-t objective and a sum of two univariate Student-t terms, respectively. Both include a normalized ordering penalty:

$$\ell_{\mathrm{ord}} = \lambda_{\mathrm{ord}} \left[\max\left(0, \frac{m_{\mathrm{L}} - m_{\mathrm{U}}}{s_T}\right)\right]^2, \qquad \lambda_{\mathrm{ord}} = 1. \tag{B5}$$

Both the direct endpoint readouts and the center–positive-half-width readout use 30 °C and 60 °C as endpoint initialization references. The two direct readout variants retain the same conditional scales, task network, and training configuration to compare the effects of the output parameterization and the observation unit.

Scale gradient path. The Stop scale path variant applies $\mathrm{stopgrad}(z)$ at the input to the scale head. The scale head is still updated by the observation loss, and backpropagation through the location path proceeds as usual. Gradients from the scale branch no longer propagate to Fusion, the local residual

modules, or the residue projection. This change preserves the form of the forward computation and specifically tests the contribution of the scale path to learning the shared representation. Paired results are reported in Table A4b.

B.3 Evaluation Metrics and Range Geometry

The error metrics for the scalar tasks are

$$\mathrm{MAE} = \frac{1}{N}\sum_i |\widehat{y}_i - y_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (\widehat{y}_i - y_i)^2}. \tag{B6}$$

Pearson and Spearman correlations quantify linear and rank association, respectively. For an observed range $[y_{i\mathrm{L}}, y_{i\mathrm{U}}]$ and a predicted range $[\widehat{y}_{i\mathrm{L}}, \widehat{y}_{i\mathrm{U}}]$, define

$$\begin{aligned} W_i &= y_{i\mathrm{U}} - y_{i\mathrm{L}}, \qquad \widehat{W}_i = \widehat{y}_{i\mathrm{U}} - \widehat{y}_{i\mathrm{L}}, \\ I_i &= \max\left(0, \min(y_{i\mathrm{U}}, \widehat{y}_{i\mathrm{U}}) - \max(y_{i\mathrm{L}}, \widehat{y}_{i\mathrm{L}})\right). \end{aligned} \tag{B7}$$

The joint endpoint error is computed within each run as

$$\begin{aligned} \mathrm{JRMSE} &= \sqrt{\frac{1}{2N}\sum_i \left[(\widehat{y}_{i\mathrm{L}} - y_{i\mathrm{L}})^2 + (\widehat{y}_{i\mathrm{U}} - y_{i\mathrm{U}})^2\right]} \\ &= \sqrt{\frac{\mathrm{RMSE}_{\mathrm{L}}^2 + \mathrm{RMSE}_{\mathrm{U}}^2}{2}}. \end{aligned} \tag{B8}$$

For the published reference results, JRMSE is derived from the reported endpoint RMSE values. Local results are computed separately for each run and then averaged across the three seeds. For ordered intervals with positive width, we evaluate

$$\begin{aligned} &\mathrm{MOR} = \frac{1}{N}\sum_i \frac{I_i}{W_i}, \qquad \mathrm{IoU} = \frac{1}{N}\sum_i \frac{I_i}{W_i + \widehat{W}_i - I_i}, \\ &\mathrm{Precision} = \frac{1}{N}\sum_i \frac{I_i}{\widehat{W}_i}. \end{aligned} \tag{B9}$$

These metrics describe coverage of the observed range, intersection over union, and precision of the predicted range, respectively. All IoU values in this paper are arithmetic means over individual records. All observed and predicted intervals in the fixed prediction run have positive width. Let $c_i = (y_{i\mathrm{L}} + y_{i\mathrm{U}})/2$ and $\widehat{c}_i = (\widehat{y}_{i\mathrm{L}} + \widehat{y}_{i\mathrm{U}})/2$. Then

$$\text{C-MAE} = \frac{1}{N}\sum_i |\widehat{c}_i - c_i|, \qquad \text{W-MAE} = \frac{1}{N}\sum_i |\widehat{W}_i - W_i|. \tag{B10}$$

The endpoint error can be further decomposed as

$$\mathrm{JRMSE}^2 = \frac{1}{N}\sum_i (\widehat{c}_i - c_i)^2 + \frac{1}{4N}\sum_i (\widehat{W}_i - W_i)^2. \tag{B11}$$

This relation decomposes the error of the complete range into squared errors in location and span, complementing the per-record analysis of IoU and width error.

B.4 Complete Comparisons and Range Geometry

The overall gains extend to both endpoints, and the joint objective improves range geometry primarily through the span. Table A2 retains the complete published comparison. Table A3 adds correlations under the same protocol, range geometry across three seeds, and results from the fixed prediction run. Relative to local PatchET, RoCO-ET reduces RMSE at both endpoints, center MAE, and width MAE. Compared with Independent-t, RoCO-ET reduces width MAE from 12.950 to 10.697 °C, whereas center MAE changes from 9.928 to 10.029 °C, indicating that the joint objective improves the quality of the complete range primarily by improving its span.

Interval precision further characterizes the trade-off in coverage: RoCO-ET has a three-seed mean of 0.687, below 0.713 for local PatchET, while achieving higher IoU and MOR. Its range improvements therefore reflect better overall overlap and coverage of the observed range, while predictions that extend beyond the observed range still contribute to error.

Table A2: Published benchmark results.

(a) Scalar temperature

Method	Opt. P	Opt. S	Opt. MAE	Opt. RMSE	Stab. MAE	Stab. RMSE
BiLSTM	0.449	0.347	12.180	17.320	18.490	23.960
CNN	0.511	0.418	11.680	15.510	19.230	25.010
Transformer	0.521	0.416	12.380	16.720	19.100	24.870
Light Attention	0.515	0.413	11.590	15.450	19.980	25.170
RNN	0.459	0.358	13.510	18.200	21.020	27.530
DeepET	0.450	0.382	11.790	14.960	24.660	30.200
Seq2Topt	0.526	0.423	11.590	15.670	19.440	25.660
TemStaPro	0.501	0.414	11.790	15.540	19.310	25.030
Segment-Transformer	–	–	–	–	18.090	24.030
PatchET	**0.551**	**0.458**	**9.690**	**13.880**	**17.310**	**22.660**

(b) Temperature-range endpoints

Method	L-P	L-S	L-MAE	L-RMSE	U-P	U-S	U-MAE	U-RMSE	MOR
BiLSTM	0.470	**0.471**	11.530	14.960	0.402	0.437	15.500	19.410	0.565
CNN	0.469	0.409	11.720	15.260	0.478	0.431	14.000	17.070	0.650
Transformer	0.472	0.427	13.280	16.850	0.360	0.306	16.490	20.240	0.495
Light Attention	0.041	0.065	12.960	17.160	0.108	0.109	16.420	20.060	0.598
RNN	0.468	0.446	12.290	16.020	0.512	**0.520**	15.570	19.860	0.578
DeepET	0.471	0.468	18.190	23.390	0.433	0.403	40.670	44.810	0.059
Seq2Topt	0.455	0.396	11.640	15.600	0.484	0.474	13.830	17.700	0.683
TemStaPro	0.354	0.300	12.430	15.810	0.343	0.291	15.530	18.600	0.559
PatchET	**0.473**	0.418	**11.280**	**14.870**	**0.514**	0.470	**13.530**	**17.050**	**0.707**

Source: Zhang et al. (2026). P/S: Pearson/Spearman; L/U: lower/upper endpoint. Errors: °C. Reference JRMSE in Table 1 is derived from endpoint RMSEs.

C Observation Mechanisms and Output Structure

C.1 Intervention Settings and Recovery Behavior

The training interventions modified 20% of the range records. We used model seeds 1, 2, and 3 and perturbation seed 2701; methods received paired perturbations within each condition, and the test labels remained unchanged. Whole-range shifts had two nominal strength levels, with median absolute shifts of 5.700 and 10.800 °C and 90th percentiles of 10.900 and 20.400 °C, respectively. The span and single-endpoint conditions used strength levels of 1.5 and 2.0. Figure 3b evaluates changes relative to the reference without additional perturbation for the same seed:

$$\Delta\,\mathrm{JRMSE}_{M,h,s} = \mathrm{JRMSE}_{M,h,s} - \mathrm{JRMSE}_{M,0,s}, \tag{C1}$$

where M, h, and s denote the method, intervention condition, and model seed, respectively.

Table A3: Supplementary correlations and range metrics.

(a) Optimum-temperature correlation

Method	Pearson	Spearman
local PatchET	0.628 ± 0.011	0.631 ± 0.010
Huber	0.652 ± 0.016	0.656 ± 0.016
RoCO-ET	**0.682** ± 0.023	**0.686** ± 0.017

(b) Range geometry across three seeds

Method	L-RMSE	U-RMSE	C-MAE	W-MAE	Precision
local PatchET	14.851 ± 0.475	16.283 ± 0.635	10.779 ± 0.103	12.712 ± 0.217	**0.713** ± 0.005
Huber	15.164 ± 0.854	15.437 ± 0.996	10.306 ± 0.321	12.654 ± 0.236	**0.713** ± 0.022
Independent-t	15.047 ± 0.106	14.811 ± 0.172	**9.928** ± 0.205	12.950 ± 0.374	0.701 ± 0.016
RoCO-ET	**14.623** ± 0.389	**13.973** ± 0.539	10.029 ± 0.414	**10.697** ± 0.103	0.687 ± 0.008

(c) Fixed prediction instance in Figures 4 and A4

Method	JRMSE	IoU	MOR	C-MAE	W-MAE	Precision
local PatchET	15.490	0.522	0.674	10.728	12.866	**0.708**
RoCO-ET	**14.250**	**0.573**	**0.768**	**10.412**	**10.580**	0.682

(a,b): mean ± sample SD over three seeds; (c): seed 1, 182 records. Bold/underline: best/next distinct value. C/W: center/width; precision: overlap/predicted width.

Known-reference recovery further links the benefits of joint modulation to the structure of the deviations. The two Student-t objectives perform similarly under homoscedastic conditions. The joint objective achieves lower recovery JRMSE under input-dependent conditions and shifts affecting entire records, with a gap of 0.763 °C under the record-shift condition. Under the single-endpoint condition, Independent-t performs better (Figure A1a).

Interventions on the actual labels show the same pattern: RoCO-ET achieves the lowest JRMSE under the strongest whole-range shift and span conditions. Under the strongest single-endpoint condition, Independent-t achieves 14.978 °C, below the 16.845 °C of RoCO-ET (Figure A1b). Under the stronger single-endpoint intervention, RoCO-ET has an MAE 1.212 °C higher than Independent-t on the unperturbed endpoint (Figure A1c). The independent objective allows the other endpoint to contribute separately to fitting, whereas the shared factor modulates both endpoints together. This distinction explains how the two objectives accommodate different deviation structures.

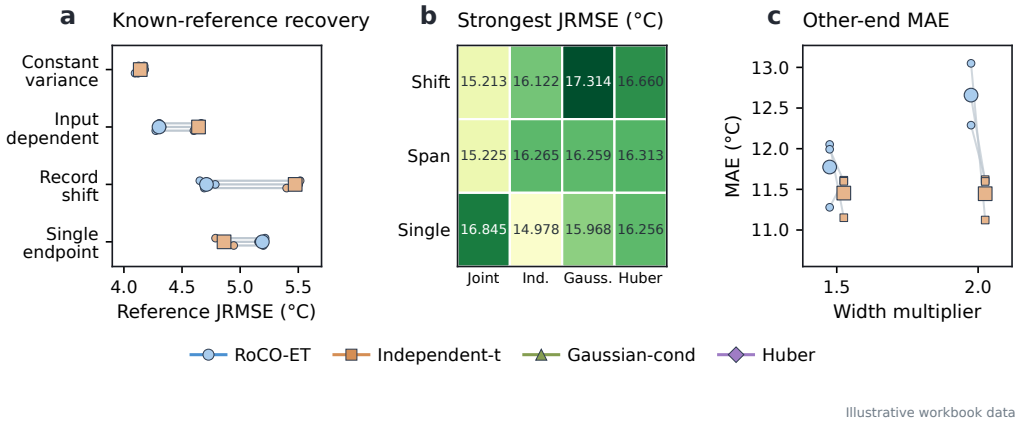

Figure A1: Known-reference recovery and training interventions. (a) Recovery JRMSE. (b) Mean test JRMSE at the strongest interventions. (c) Unperturbed-endpoint MAE under single-endpoint interventions. Small/large markers: three seed results/means; lines connect paired results. Test labels are unchanged in (b,c).

C.2 Output Structure and the Scale Gradient Path

The joint objective and the center–half-width readout play complementary roles. Table A4a compares combinations of two readout designs and two observation objectives. The direct-endpoint readout uses an ordering penalty, whereas the center–positive-half-width readout maintains ordering through its parameterization. With the center–half-width head, the joint term reduces JRMSE by 0.623 °C; with the direct-endpoint head, it reduces JRMSE by 0.333 °C. The center–half-width head also achieves lower JRMSE and higher IoU under both objectives. All variants have a zero endpoint-order violation rate on the test set, and the center–half-width readout additionally guarantees ordered endpoints through its parameterization.

The scale path contributes to learning the shared representation. Blocking gradients at the scale-head input while continuing to train the scale head increases optimum temperature RMSE and range JRMSE by 0.271 and 0.379 °C, respectively, and also lowers range IoU (Table A4b). This result is consistent with the update through both paths in Equation (A13), showing that conditional scales both enter the residual metric and influence the task representation.

Table A4: Readout, observation-objective, and gradient-path comparisons.

(a) Output structure and observation unit

Variant	JRMSE	IoU	MOR	W-MAE	Invalid rate
RoCO-ET	**14.307** ± 0.067	**0.581** ± 0.009	**0.774** ± 0.005	**10.697** ± 0.103	**0.000** ± 0.000
Independent-t	14.930 ± 0.075	0.545 ± 0.004	0.711 ± 0.014	12.950 ± 0.374	**0.000** ± 0.000
Direct-joint	<u>14.747</u> ± 0.106	<u>0.552</u> ± 0.019	<u>0.720</u> ± 0.037	12.965 ± 0.419	**0.000** ± 0.000
Direct-independent	15.080 ± 0.098	0.535 ± 0.007	0.679 ± 0.013	12.445 ± 0.525	**0.000** ± 0.000
Gaussian-fixed	15.550 ± 0.095	0.526 ± 0.007	0.663 ± 0.007	<u>12.286</u> ± 0.440	**0.000** ± 0.000

(b) Scale path into the shared representation

Gradient route	Opt. RMSE	Range JRMSE	Range IoU
Both paths	**12.382** ± 0.031	**14.273** ± 0.024	**0.580** ± 0.003
Stop scale path	<u>12.653</u> ± 0.052	<u>14.652</u> ± 0.032	<u>0.569</u> ± 0.002

Mean ± sample SD over three seeds; (b) uses separate paired controls. Variant definitions: Appendix B.2.

D Data Partitions, Representations, and Prediction Conditions

D.1 Sequence Grouping and Representation Interfaces

RoCO-ET retains its advantage under sequence-group holdout evaluation. With group holdout, absolute errors increase for all methods, and RoCO-ET's range IoU falls from 0.582 to 0.549. Nevertheless, its IoU remains higher than that of Huber and local PatchET under the same setting (Table A5a). Relative to local PatchET, range JRMSE decreases by 1.297 °C, showing that the full method continues to provide gains under this additional evaluation condition.

The benefits of observation learning also appear with a second representation interface. After switching to the ProtT5 projection interface, RoCO-ET maintains lower scalar and range prediction errors and higher IoU than Huber, with a range JRMSE gap of 1.209 °C (Table A5b). This indicates that the observation objective and sequence representation play complementary roles.

D.2 Temperature and Observed-Width Groups

Gains on the scalar tasks vary across temperature groups. In the fixed-seed run, RoCO-ET achieves lower RMSE for optimum temperature prediction in all four groups, Q1–Q4, with larger gaps in the two middle groups. Operational stability prediction improves in Q2–Q4, whereas RMSE in Q1 increases from 22.210 to 25.864 °C (Figure A2a, b). These results reveal differences across groups despite the improvement in the overall average.

Improvements in span quality extend across multiple width groups. Width MAE decreases in all groups W1–W4 (Figure A2c), and IoU increases in each corresponding group, with gains in group

Table A5: Performance across partitions and representation interfaces.

(a) Data partitions

Method	Opt. RMSE	Range JRMSE	Range IoU
Released split			
RoCO-ET	**12.382** ± 0.067	**14.301** ± 0.054	**0.582** ± 0.002
Huber	13.214 ± 0.014	15.325 ± 0.025	0.529 ± 0.003
local PatchET	13.891 ± 0.088	15.569 ± 0.045	0.513 ± 0.001
Sequence-group holdout			
RoCO-ET	**13.270** ± 0.034	**15.195** ± 0.042	**0.549** ± 0.002
Huber	14.294 ± 0.014	16.465 ± 0.066	0.490 ± 0.003
local PatchET	14.690 ± 0.012	16.492 ± 0.049	0.484 ± 0.003

(b) Representation interfaces

Method	Opt. RMSE	Range JRMSE	Range IoU
ESM2 interface			
RoCO-ET	**12.477** ± 0.049	**14.367** ± 0.060	**0.580** ± 0.000
Huber	13.293 ± 0.022	15.395 ± 0.078	0.527 ± 0.001
ProtT5 projection interface			
RoCO-ET	**12.950** ± 0.045	**14.942** ± 0.053	**0.560** ± 0.004
Huber	13.973 ± 0.035	16.151 ± 0.027	0.500 ± 0.002

Mean ± sample SD over three seeds, with separate paired controls for each setting.

means of 0.047–0.060. Center error increases in some groups, consistent with the more pronounced improvements in span seen in the overall geometric comparison.

Q and W denote ordered groups obtained by sorting records by observed temperature and observed range width, respectively; the number of records in each group is shown in the figure. Group metrics are calculated for the fixed seed. Overall RMSE is obtained by taking the square root of the sample-count-weighted mean of the group mean squared errors.

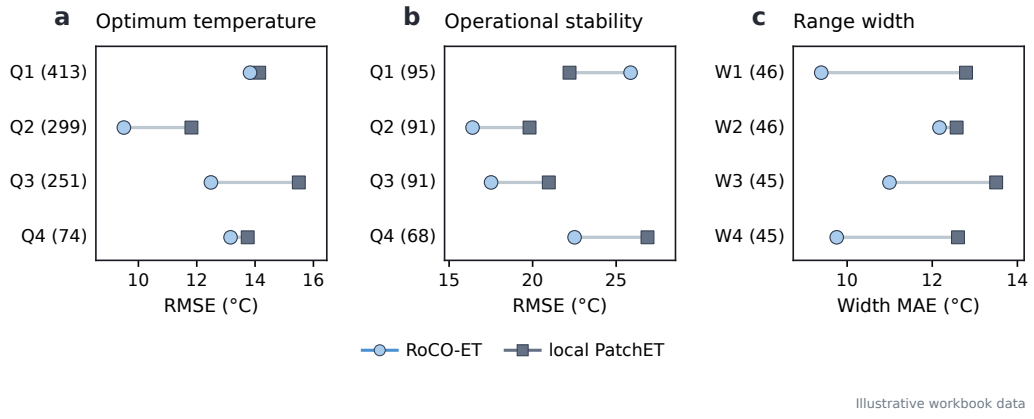

Figure A2: Grouped errors from the fixed-seed run. (a,b) Optimum-temperature and stability RMSE by temperature group. (c) Width MAE by observed-width group. Q/W: ordered temperature/width groups; parentheses: record counts. Lines connect methods within each group.

E Observation Parameters and Model Size

E.1 Local Sensitivity to Degrees of Freedom and Scale Bounds

The tested settings near the default yield small but noticeable changes in error. Figure A3 examines $\nu \in \{3, 4, 8\}$ and multipliers of $\{0.5, 1, 2\}$ for the lower and upper scale bounds relative to their defaults, varying one factor at a time. The default configuration achieves the lowest error among these candidates. The mean increases in RMSE/JRMSE under nondefault settings range from 0.148 to 0.399 °C, with a maximum increase of approximately 2.38% relative to the respective defaults.

The degrees of freedom control the form of the response to large standardized deviations, while the scale bounds constrain the range of scales used to normalize residuals. The candidate lower bounds are $0.025g_j$, $0.05g_j$, and $0.10g_j$, and the candidate upper bounds are $1.5g_j$, $3g_j$, and $6g_j$. Doubling the default upper bound leads to noticeable increases in error in all three tasks, indicating that a more permissive scale range does not improve fitting in these experiments. The degrees-of-freedom sweep keeps the scale parameterization unchanged.

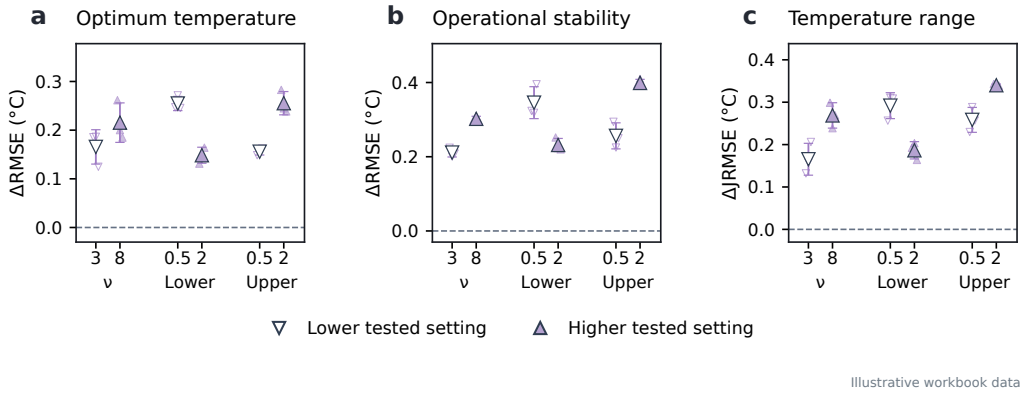

Figure A3: Paired error changes relative to the default settings for (a) optimum temperature, (b) operational stability, and (c) active temperature range. Small/large markers: three seed differences/means; error bars: sample SD. Defaults: $\nu = 4$ and scale-bound multipliers of 1.

E.2 Model Size and Runtime Settings

ESM2 is frozen and excluded from downstream parameter updates. The trainable components comprise the residue projection, local residual modules, dual aggregation and Fusion, and the location and scale heads. The models for the scalar and range tasks have 2,488,771 and 2,489,093 trainable parameters, respectively, both approximately 2.49 M (Table A6). The two task types use the same representation interface; the difference in parameter count arises from the output dimensionality of the prediction heads.

The matched supplementary experiments were implemented with PyTorch and Transformers and run on a single NVIDIA GeForce RTX 4090. Forward passes used FP16 automatic mixed precision, while temperature location outputs, residual scales, and the observation loss were computed in FP32. Frozen residue representations were cached in advance for training the task networks.

Table A6: Trainable parameters of the task networks.

Task	Output dimension	Trainable parameters
Scalar temperature	1	2,488,771
Active temperature range	2	2,489,093

Counts exclude the frozen ESM2 encoder.

F Representative Ranges and Per-Record Changes

The examples distinguish location correction from span recovery as sources of improved range predictions. In C1, RoCO-ET moves a prediction shifted toward lower temperatures back to the observed range, increasing IoU from 0.001 to 0.755. In C3 and C5, RoCO-ET mainly recovers spans that the baseline compresses excessively. In C2, it corrects a marked shift toward higher temperatures and increases overlap, but the predicted range is too wide. In C4, the entire prediction is shifted toward higher temperatures, and IoU falls from 0.908 to 0.383 (Figure A4). These patterns show that location and span jointly determine range overlap.

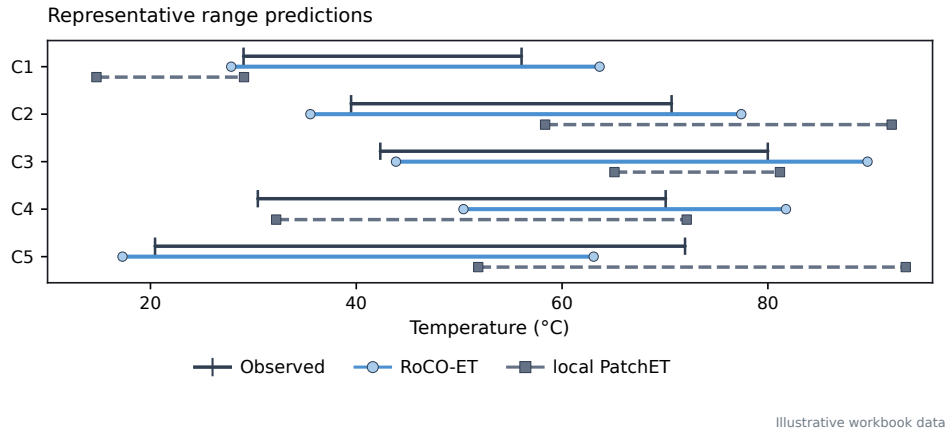

Figure A4: Observed ranges and predictions from local PatchET and RoCO-ET, ordered by observed width (seed 1). Case identifiers match Figure 4a.

In the comparison across all records in Figure 4a, 69 records show both higher IoU and lower width error, while another 29 show higher IoU together with higher width error. For predictions from the fixed seed-1 run, RoCO-ET achieves higher mean IoU and lower width MAE than local PatchET, whereas interval precision decreases from 0.708 to 0.682 (Table A3c). These metrics provide complementary views of how coverage, span, and extension beyond the observed range are related.

The widening control keeps the baseline center $\widehat{c}_i$ fixed. Let $\widehat{d}_i = (\widehat{y}_{i\mathrm{U}} - \widehat{y}_{i\mathrm{L}})/2$; then

$$\widehat{y}_i^{(\alpha)} = [\widehat{c}_i - \alpha\widehat{d}_i,\ \widehat{c}_i + \alpha\widehat{d}_i]^\top, \qquad \alpha \in \{0.75, 1.00, 1.25, 1.50, 2.00\}. \tag{F1}$$

The value $\alpha = 1.00$ was selected on the validation set; the other multipliers were used to compare changes in coverage and overlap for predictions from the fixed seed-1 run. Increasing the half-width can increase MOR, but further widening reduces IoU, consistent with the pattern in Figure 4b in the main text.