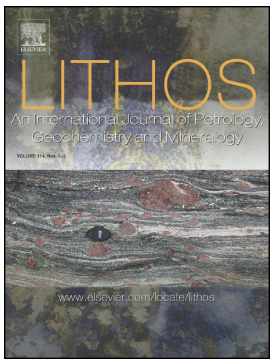
Journal Pre-proof




# Machine learning thermobarometry: Methods, applications, and a benchmarking protocol

Xirui Qiao, Yan Xia, Xisheng Xu

PII: S0024-4937(26)00266-5

DOI: https://doi.org/10.1016/j.lithos.2026.108665

Reference: LITHOS 108665

To appear in: *LITHOS*

Received date: 26 March 2026

Revised date: 6 June 2026

Accepted date: 6 June 2026

Please cite this article as: X. Qiao, Y. Xia and X. Xu, Machine learning thermobarometry: Methods, applications, and a benchmarking protocol, *LITHOS* (2024), https://doi.org/10.1016/j.lithos.2026.108665

# Machine Learning Thermobarometry: Methods, Applications, and a Benchmarking Protocol

Xirui Qiao, Yan Xia *, Xisheng Xu

State Key Laboratory of Critical Earth Material Cycling and Mineral Deposits, Frontiers Science Center for Critical Earth Material Cycling, School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China

*Corresponding author: Yan Xia (xiayan@nju.edu.cn)

# Abstract

Mineral thermobarometry is a fundamental tool for constraining the thermal evolution of magmatic systems. In recent years, machine learning (ML) thermobarometers have developed rapidly owing to their capacity to handle large, high-dimensional datasets. However, current ML models lack a standard, reproducible evaluation framework, making it difficult to identify the true bottlenecks in model performance. To address this limitation, a modular benchmarking protocol (MBP) is proposed that decomposes the ML thermobarometry workflow into three independently evaluable modules: data processing, algorithm architecture, and post-processing, supporting quantitative inversion of pressure–temperature (P–T) conditions during magmatic evolution.

Using the clinopyroxene-liquid system as a baseline, models were trained on 2079 experimental data pairs with P–T grid-stratified cross-validation to isolate the independent contribution of each module. Three principal findings emerged. First, effective thermodynamic information is the primary control on predictive accuracy: incorporating coexisting melt compositions reduced the root mean square error (RMSE) of temperature and pressure predictions by 40.9% and 16.1%, respectively, relative to the mineral-only (*NoLiquid*) baseline. Under the current feature set, pressure signals are predominantly encoded in mineral chemistry, whereas temperature predictions depend more strongly on melt composition. Second, physics-informed data augmentation improved model robustness, producing modest but statistically significant RMSE reductions. Third, the benefits of algorithmic complexity are bounded: because predictive precision closely approaches the physical limit (approximately ±16 °C for temperature and ±0.8 kbar for pressure) imposed by instrumental analytical errors, more complex Stacking ensembles did not outperform the simpler Extremely Randomized Trees algorithm.

At the current dataset scale, the principal bottleneck for thermobarometric model performance resides in the geological information encoded in the input variables, rather

than in algorithmic complexity. Future development of ML thermobarometers should prioritize engineering guided by thermodynamics and crystal chemistry, together with the explicit embedding of physical constraints into model architectures. The MBP provides a reproducible, reusable framework for optimizing thermobarometers across diverse mineral systems.

**Keywords:** Machine Learning Thermobarometry; Random Forest; Modular Benchmarking; Feature Engineering; Uncertainty Quantification; Clinopyroxene-Liquid

# 1. Introduction

Mineral thermobarometry is a fundamental tool in quantitative petrology for constraining the pressure–temperature (P–T) evolution of magmatic systems, characterizing volcanic plumbing systems, and reconstructing the thermodynamic conditions of ore-forming environments (e.g., Putirka, 2008; Blundy and Cashman, 2008; Richards, 2013; Cashman et al., 2017; Edmonds et al., 2019; Chiaradia and Caricchi, 2022; Zhao K. et al., 2023; Zhao S. et al., 2023; Taniwaki et al., 2025; Lin et al., 2026). Classical thermobarometers are grounded in mineral-melt or mineral-mineral equilibrium relationships, retrieving pressure and temperature (P–T) conditions through element partitioning or exchange equilibria (Holland and Blundy, 1994; Nimis and Ulmer, 1998; Ridolfi and Renzulli, 2012; Neave and Putirka, 2017). Natural geological systems, however, commonly evolve along open, dynamic pathways that depart from ideal equilibrium (Chicchi et al., 2023), and simplified empirical or thermodynamic models frequently fail to describe compositionally complex, strongly coupled natural assemblages. To overcome these limitations and reduce reliance on a single equilibrium criterion, machine learning (ML) thermobarometry has expanded rapidly in recent years owing to its capacity to handle complex, high-dimensional datasets (Petrelli et al., 2024; Weber and Blundy, 2024).

Since Petrelli et al. (2020) first applied ML systematically to the clinopyroxene-liquid system, data-driven thermobarometers have been extended to diverse mineral systems, including amphibole, biotite, pyroxenes, and garnet (Thomson et al., 2021; Higgins et al., 2022; Jorgenson et al., 2022; Li and Zhang, 2022; Qin et al., 2024). More recent studies have incorporated uncertainty quantification, data augmentation, and bias correction (Chicchi et al., 2023; Ágreda-López et al., 2024; Petrelli, 2024; Weber and Blundy, 2024). Collectively, these advances reflect a disciplinary shift from a narrow focus on fitting accuracy toward model robustness, interpretability, and broader geological applicability.

The rapid expansion of this field has, however, also produced a proliferation of divergent methodologies for constructing mineral thermobarometers. Published studies differ markedly in data preprocessing, feature engineering, validation protocols, and post-processing strategies, making it difficult to attribute observed performance differences to specific methodological choices. Whether reported accuracy gains stem from algorithmic optimization or from variations in training datasets and processing pipelines, therefore, remains unclear. When multiple modeling steps are modified simultaneously, the independent contribution of each module cannot be quantitatively isolated, hindering the identification of true performance bottlenecks and undermining model reproducibility, transferability, and cross-study comparability. Open-source frameworks such as Geochemistry π (ZhangZhou et al., 2024) and Orange-Volcanoes (Musu et al., 2025) have substantially lowered the technical barrier to implementing ML workflows on geochemical and petrological data. However, understanding why thermobarometric models succeed or fail in real geological settings remains a distinct methodological challenge.

To address these limitations, this study introduces a modular benchmarking protocol (MBP), developed through a systematic review of current ML thermobarometer workflows. The MBP decomposes the ML thermobarometry pipeline into three independently evaluable functional modules: data processing, algorithm architecture, and post-processing. Under a controlled experimental design, three core questions are addressed quantitatively: (1) To what extent does input information density improve predictive accuracy? (2) Can physics-informed data augmentation enable the model to extract more geologically meaningful information? (3) Given the constraints of limited geological datasets, does increasing algorithmic complexity deliver meaningful improvements? By isolating the independent contribution of each module to overall predictive performance, the MBP serves as a standardized baseline for cross-study comparisons and provides practical guidance for optimizing future ML thermobarometers.

# 2. Workflow for Machine Learning Thermobarometers

The construction pipeline of ML thermobarometers typically involves four main stages: data preparation and quality control, feature engineering and data transformation, model training and validation strategies, and model evaluation and interpretation.

## 2.1 Data Preparation and Quality Control

Training data for ML thermobarometers are predominantly drawn from high-quality experimental phase equilibrium databases, most notably the Library of Experimental Phase Relations (LEPR) database (Hirschmann et al., 2008; Petrelli et al., 2020; Jorgenson et al., 2022; Cutler et al., 2024; Qin et al., 2024; Weber and Blundy, 2024). Natural-sample databases such as GEOROC and PetDB, together with literature-compiled datasets, are also widely employed for external validation or to fill specific compositional gaps (Higgins et al., 2022; Wieser et al., 2023; Chicchi et al., 2023; Cutler et al., 2024).

Prior to training, raw data require rigorous quality control, although current studies differ considerably in how this is implemented. For analytical fidelity, researchers adopt different acceptable ranges for electron probe microanalyzer (EPMA) analytical totals (e.g., 98.5–101.5 wt%) and mineral stoichiometry thresholds (Li and Zhang, 2022; Chicchi et al., 2023; Cutler et al., 2024; Qin et al., 2024). Equilibrium filtering is the step that varies most across studies. Most studies employ the Fe–Mg exchange partition coefficient ($K_D$) as a baseline metric, but practical implementation differs widely: from the traditional fixed empirical window of 0.28 ± 0.08 (Putirka, 2008) to adaptive thresholds derived from dataset statistical characteristics (Jorgenson et al., 2022), multi-component thermodynamic constraints such as ΔDiHd (Wieser et al., 2022), and rigorous evaluation against theoretical predictions (Ágreda-López et al., 2024). This diversity reflects the inherent ambiguity of equilibrium criteria in natural systems. Rigid, single-value thresholds risk excluding genuine equilibrium assemblages that express natural compositional variation, whereas abandoning equilibrium filters entirely

introduces nonequilibrium noise (Chicchi et al., 2023). No universally accepted alternative has emerged to date.

Experimental petrological datasets also suffer from inherent sampling biases that can introduce systematic errors in nonuniformly sampled regions of P–T space (Wieser et al., 2023). As illustrated in Fig. 1, the calibration dataset is highly nonuniformly distributed: low-pressure samples ($P \leq 2.5$ kbar; n = 981) account for roughly half of the compilation, whereas high-pressure end-members ($P \geq 20$ kbar; n = 182) constitute less than 10%, indicating a substantial risk of sampling bias. Three distribution-adjustment strategies have been applied to mitigate this imbalance. Undersampling reduces the overrepresentation of low-pressure intervals by removing samples from high-density regions (Petrelli et al., 2020). Stratified sampling partitions the P–T space into discrete grid cells and constructs training and test sets in proportion within each cell, ensuring uniform thermodynamic coverage (Qin et al., 2024). Monte Carlo data augmentation synthesizes new samples in sparsely populated regions by explicitly propagating EPMA analytical errors, thereby expanding the training set while simultaneously encoding analytical uncertainties (Ágreda-López et al., 2024). Together, these strategies improve the representativeness of training data across P–T space and enhance model generalization at data-poor end-members, particularly under high-pressure or high-temperature conditions.

## 2.2 Feature Engineering and Data Transformation

Feature engineering transforms quality-screened raw data into model-ready input variables; studies differ considerably in their feature selection and combination strategies. The most widely adopted feature format consists of the major element oxide weight percentages (wt%) of minerals and their coexisting melt phases (Petrelli et al., 2020; Jorgenson et al., 2022; Qin et al., 2024). Guided by crystal-chemical principles, some researchers supplement these oxide compositions with cations normalized to a fixed number of oxygen atoms (atoms per formula unit, apfu) and petrologically

meaningful derived parameters such as Mg#, $Fe^{3+}/\Sigma Fe$ (Thomson et al., 2021), or plagioclase anorthite (An) content (Cutler et al., 2024), which provide a more direct representation of crystal-chemical substitution mechanisms and magmatic evolution. Depending on the target application, model inputs may consist of single-mineral compositions alone (e.g., CPX-only), mineral-melt pairs (e.g., CPX-LIQ), or multi-mineral assemblages that impose additional thermodynamic constraints on P–T inversion (Weber and Blundy, 2024).

The concentration ranges of different oxides span several orders of magnitude, and feeding raw data directly into a model can allow features with larger absolute values to dominate the training process disproportionately. To address this, established studies routinely apply feature scaling or standardization to normalize all variables to comparable ranges (Petrelli et al., 2020; Jorgenson et al., 2022; Ágreda-López et al., 2024). Although tree-based ensemble algorithms such as Random Forest are, in principle, insensitive to feature scaling because their splits depend only on rank order, empirical evaluations by Petrelli et al. (2020) and Ágreda-López et al. (2024) demonstrated that standardization remains a recommended preprocessing step even for these architectures.

The constant-sum constraint inherent to geochemical data (the closure problem) generates spurious correlations among compositional components. In principle, this issue can be resolved through log-ratio transformations (alr, clr, and pwlr) within the framework of Compositional Data Analysis (CoDA). In practice, however, most ML studies apply standardized raw oxide features without log-ratio transformation (Petrelli et al., 2020; Jorgenson et al., 2022; Higgins et al., 2022). The systematic evaluation by Ágreda-López et al. (2024) confirmed that CoDA effectively eliminates statistical closure bias, but that performance gains for relatively low-dimensional thermobarometric models are negligible. For an optimal balance between model simplicity and predictive accuracy, working directly with raw chemical compositions therefore remains the preferred approach.

## 2.3 Model Training Strategies

Standard ML workflows partition datasets into training and test sets and employ cross-validation within the training set for hyperparameter optimization. Existing thermobarometric studies universally adopt repeated random partitioning to improve the robustness of performance evaluation (Petrelli et al., 2020; Li and Zhang, 2022). A critical pitfall is data leakage: all preprocessing transformations, such as standardization, must be fitted exclusively on the training set, with the learned parameters then applied to the test set (Zhu et al., 2023).

Among available algorithm families, tree-based ensemble methods demonstrate the most robust performance in thermobarometric applications (Petrelli et al., 2020; Li and Zhang, 2022; Cutler et al., 2024; Table 1). Random Forest (RF) and its variant Extremely Randomized Trees (ERT) capture high-dimensional nonlinear relationships by aggregating large ensembles of decision trees (Breiman, 2001; Bishop, 2006); ERT introduces additional stochasticity during node splitting, which further reduces model variance and yields superior predictive accuracy in several benchmarks (Petrelli et al., 2020; Cutler et al., 2024; Ágreda-López et al., 2024). Gradient boosting methods such as XGBoost handle missing values efficiently in compiled datasets drawn from heterogeneous literature sources (Qin et al., 2024). By contrast, deep learning methods possess greater fitting capacity in principle but are constrained by their opacity and the requirement for very large training datasets, limiting their applicability in petrology, where experimental data remain comparatively scarce (Chicchi et al., 2023).

Hyperparameter selection governs the learning dynamics of an algorithm. Several studies indicate that tree-based thermobarometers are largely insensitive to most hyperparameters, with the notable exception of the number of features considered at each split (mtry), which substantially influences predictive accuracy; exhaustive tuning of the remaining parameters typically yields only negligible marginal returns (Jorgenson et al., 2022). In practice, most studies therefore tune within established empirical ranges and apply the One Standard Error (One-SE) Rule to favor simpler

configurations whose performance is statistically indistinguishable from the optimum, thereby reducing the risk of overfitting (Li and Zhang, 2022; Weber and Blundy, 2024).

Tree-based ensemble algorithms are inherently susceptible to regression toward the mean in regression tasks. This statistical artifact generates systematic predictive bias at extreme P–T end-members, and the effect is particularly amplified by nonuniform data distributions (Zhang and Lu, 2012; Wieser et al., 2023; Ágreda-López et al., 2024). To address this limitation, several studies have adopted a two-step residual-based bias correction: a primary model is first trained and its systematic residuals computed; extreme predictions are then adjusted using supplementary models or piecewise functions (Zhang and Lu, 2012; Weber and Blundy, 2024; Ágreda-López et al., 2024).

ML methods provide sample-level uncertainty estimates that capture two distinct sources of prediction variance. The first is the dispersion of the tree-model prediction ensemble, quantified by metrics such as the interquartile range, which reflects intrinsic model confidence (Jorgenson et al., 2022; Higgins et al., 2022). The second is Monte Carlo error propagation, which explicitly incorporates EPMA analytical uncertainties into the probability distributions of P–T predictions (Li and Zhang, 2022; Ágreda-López et al., 2024; Cutler et al., 2024). Weber and Blundy (2024) demonstrated that even with a perfect model, EPMA errors alone constrain the maximum achievable predictive precision to approximately ±0.8 kbar and ±16 °C. Analytical error thereby defines the physical lower bound of thermobarometric precision (Wieser et al., 2023).

## 2.4 Model Evaluation and Result Interpretation

Quantitative performance evaluation is essential for establishing comparable baselines across studies, yet metric selection itself varies considerably between publications. Root mean square error (RMSE) is the most universally adopted performance indicator in ML thermobarometry (Petrelli et al., 2020; Jorgenson et al., 2022; Higgins et al., 2022; Li and Zhang, 2022; Ágreda-López et al., 2024), but it is intrinsically scaled to the P–T coverage of the calibration dataset, so cross-study

comparisons are meaningful only when evaluated over comparable condition ranges (Hyndman and Koehler, 2006; Weber and Blundy, 2024). The mean absolute error (MAE) is more robust to outliers, though its tolerance for extreme deviations risks masking systematic biases at the margins of the calibration space. The mean bias error (MBE) is specifically suited to diagnosing the direction and magnitude of predictive deviations, particularly for identifying regression-toward-the-mean bias; in practice, MBE should be interpreted alongside zone-specific bias analysis to avoid pseudo-unbiased states in which positive and negative errors cancel. The coefficient of determination (R2) is commonly reported as a supplementary indicator, but its sensitivity to P–T coverage range means it should not serve as the primary basis for cross-study performance comparison.

ML models are frequently criticized for the opacity of their decision-making processes. Feature importance and interpretability analyses are the principal tools for addressing this limitation. Impurity-based feature importance, an intrinsic metric of tree-based ensembles such as Random Forest, has been widely applied to identify the dominant compositional controls on thermobarometric predictions. Petrelli et al. (2020) identified $SiO_2$, $Al_2O_3$, and CaO as the primary drivers of clinopyroxene pressure predictions; Li and Zhang (2022) demonstrated that melt MgO content governs temperature predictions in biotite-bearing systems; and Thomson et al. (2021) established that Si and Na control pressure predictions in pyrope garnet barometry.

SHapley Additive exPlanations (SHAP) analysis provides sample-level quantification of feature contributions, linking algorithmic outputs to underlying geological mechanisms (Lundberg and Lee, 2017). Li et al. (2023) applied SHAP to a clinopyroxene hydrogen diffusion discrimination model and interpreted the resulting feature contributions in terms of hydrogen incorporation and diffusion mechanisms, illustrating the capacity of interpretability analyses to bridge data-driven outputs and petrological understanding.

The methodological diversity reviewed above, spanning data curation, feature

engineering, training strategies, and post-processing, is both a strength and a limitation of the current ML thermobarometry literature. Without a controlled framework to isolate the contribution of each step, identifying the specific design choices that drive predictive performance remains difficult. The MBP introduced in Section 3 addresses this gap by decomposing the workflow into independently evaluable modules within a unified experimental matrix.

# 3. MBP and Performance Analysis

The MBP formalizes the construction of ML thermobarometers as a decomposable, attributable composite function (Fig. 2). By systematically combining modules under controlled experimental conditions, the independent contribution of each module to overall predictive performance can be quantitatively isolated within a unified framework.

## 3.1 Experimental Design and Modular Framework

The mathematical framework is expressed as

$$\hat{y} = M_3\left(M_2\left(M_1(D_{\text{feature}})\right)\right) \quad \text{(Eq. 1)}$$

$$\sigma_{\hat{y}} = M_4\left(M_3\left(M_2\left(M_1(D_{\text{feature}})\right)\right); \Sigma_{\text{EPMA}}\right) \quad \text{(Eq. 2)}$$

where $D_{\text{feature}}$ denotes the input feature set and $M_1$, $M_2$, and $M_3$ denote the data preprocessing, algorithm, and post-processing correction modules, respectively, which together produce the point prediction $\hat{y}$. Complementing this point-prediction pipeline, the Monte Carlo propagation module $M_4$ propagates the input analytical covariance $\Sigma_{\text{EPMA}}$ through the trained model to yield the prediction uncertainty $\sigma_{\hat{y}}$. $M_1$–$M_3$ target point-prediction accuracy, $M_4$ characterizes uncertainty propagation and thereby probes the physical precision limit imposed by analytical errors (detailed in Section 3.4).

The clinopyroxene-liquid experimental dataset compiled and quality-controlled by Jorgenson et al. (2022) was adopted, based on the LEPR database (Hirschmann et al., 2008) together with supplementary published experiments. From an unfiltered

compilation of 2571 CPX-liquid pairs, samples were retained when their Fe–Mg exchange coefficient ($K_D$) fell within one standard deviation of the unfiltered-set mean. Pairs with $P$ > 30 kbar, melt $SiO_2$ < 35 wt%, or clinopyroxene $K_2O$ > 1.5 wt% were excluded, as these compositional end-members are too sparsely sampled to support reliable interpolation. $Fe_2O_3$ was recast to FeO, melt compositions were renormalized on an anhydrous basis to 100 wt%, and $H_2O$-in-melt was excluded as a feature owing to inconsistent reporting in the source literature. The final calibration dataset comprises 2079 phase-equilibrium pairs (1730 training and 349 fixed test samples) spanning 0–30 kbar and approximately 700–1600 °C.

To evaluate the effect of input information density, a dual feature-set comparative design was adopted. The *NoLiquid* feature set (9 dimensions, containing only clinopyroxene oxide compositions) simulates deployment scenarios without melt information, whereas the *Liquid* feature set (18 dimensions, supplemented with 9 melt oxides from the coexisting liquid) represents a theoretically information-complete configuration.

At the modular level, candidate configurations are defined as follows. $M_1$ (data processing) comprises three strategies: Raw (standardization only, serving as the baseline), Balanced (inverse-frequency reweighting of binned data), and Augmented (physics-informed data augmentation driven by EPMA error models). $M_2$ (algorithm) encompasses three architectures: ERT (Bagging), CatBoost (Boosting), and a Stacking ensemble that fuses ERT, CatBoost, and RF through ridge regression. $M_3$ (post-processing) evaluates two schemes: None (no correction) and Segmented (piecewise linear bias correction).

Combined with the dual feature sets, these modular configurations yield 24 controlled experimental groups (Table 2), organized into four categories: $G_1$ (raw baseline), $G_2$ (balanced comparison), $G_3$ (augmented main experiments), and $G_4$ (correction effectiveness evaluation). Detailed configurations of the full experimental matrix are provided in Supplementary Table S1.

P–T grid-stratified 10-fold cross-validation was adopted as the core evaluation protocol. The P–T space was discretized into a regular two-dimensional grid with a resolution of $k = \lceil \sqrt{n} \rceil \approx 46$ bins per axis (n = 2079), and a hold-out test set was constructed by drawing one sample from each nonempty cell, yielding 349 test samples and 1730 training samples. This scheme enforces uniform geometric coverage of the held-out partition across the full thermodynamic range and prevents test-set evaluation from being dominated by the densely populated low-pressure regime. Within the training partition, 10-fold splits were generated with stratified sampling on the grid-cell index, ensuring that the marginal P–T distribution of each fold remained consistent with the overall calibration set (Fig. 3). All analyses were implemented in Python 3, using the *scikit-learn* framework for tree-based ensembles and cross-validation, the *catboost* library for the gradient boosting architecture, and the *shap* package for interpretability analysis.

## 3.2 Dominant Role of Effective Thermodynamic Information

The comparison of the dual feature sets demonstrated that incorporating coexisting melt compositional information markedly improved model predictive accuracy. In the $G_1$ experimental group (Raw + ERT + None; Table 3), inclusion of melt components reduced the temperature RMSE from 54.35 to 32.13 °C (-40.9%) and the pressure RMSE from 2.43 to 2.04 kbar (-16.1%).

The pronounced asymmetry in performance gains between temperature and pressure reflects fundamental differences in their underlying thermodynamic constraints. Temperature is strongly governed by mineral-melt equilibrium and liquidus relations; without melt data, the model must infer temperature from indirect compositional proxies on the mineral side, such as variations in Mg#, which substantially amplifies predictive errors. Pressure measurements in experiments are typically less precise and less accurate than temperature determinations (Wieser et al., 2023; Wang et al., 2025), yet pressure signals are more robustly encoded in the intrinsic

crystal-chemical substitution mechanisms of the mineral phase itself. As pressure increases, clinopyroxene preferentially incorporates the jadeite ($NaAlSi_2O_6$) component, a substitution directly reflected in $Na_2O$ and $Al_2O_3$ mass fractions. This crystal-chemical encoding enables ML algorithms to extract pressure signals through coupled Na-Al variations even under the *NoLiquid* configuration, explaining the comparatively limited marginal gain for pressure relative to the substantial temperature improvement achieved when the *Liquid* feature set is introduced. The optimized E07 model confirms the disproportionate gain first observed in the $G_1$ baseline (Fig. 4).

SHAP interpretability analysis further supports this thermodynamic interpretation. In the E07b ERT model, $MgO_{liq}$ is the most important feature for temperature prediction (Fig. 5a), whereas the top two features governing pressure predictions are $Al_2O_{3,CPX}$ and $Na_2O_{CPX}$ (Fig. 5b). The corresponding analysis under the *NoLiquid* configuration (Figs. 5c and d) corroborates this picture from the complementary direction: $Al_2O_{3,CPX}$ and $Na_2O_{CPX}$ retain their top-ranked positions for pressure, confirming that removal of melt composition does not erode the mineral-side encoding of pressure, whereas no single mineral-side feature emerges as a dominant temperature predictor once melt composition is excluded. Temperature is therefore more sensitive to melt composition, whereas pressure relies predominantly on the crystal-chemical signals of the mineral phase. This pattern is consistent with thermobarometric models developed across diverse experimental systems (Petrelli et al., 2020; Li and Zhang, 2022).

The observed performance gains, therefore, reflect an increase in effective thermodynamic information density, rather than a simple expansion of feature dimensionality. In the feature engineering of ML thermobarometers, the thermodynamic discriminative capacity of compositional variables is fundamentally more critical than the total number of input features.

## 3.3 Encoding of Physical Priors and Model Generalization

The EPMA error model of Ágreda-López et al. (2024) was applied to generate 15 Gaussian-perturbed replicates per sample within the training folds, implementing physics-informed data augmentation. Table 4 compares performance between the Raw and Augmented configurations under the same base algorithm (ERT + None). Under the *Liquid* feature set, augmentation reduced the temperature RMSE from 32.13 to 30.98 °C (-3.6%) and the pressure RMSE from 2.04 to 1.93 kbar (-5.4%). Paired *t*-tests on the 10-fold cross-validation results confirmed the statistical significance of these improvements (T: $\Delta = -1.15$ °C, $p = 0.019$; P: $\Delta = -0.10$ kbar, $p = 0.012$), indicating that physics-informed augmentation enhances predictive stability and reduces the risk of overfitting to experimental noise. The Balanced strategy (G2) produced only negligible overall changes and yielded no consistent performance gains across algorithms or feature-set combinations (Supplementary Table S2), performing substantially below the physics-based augmentation in effectiveness. This confirms that, for this dataset, explicitly embedding analytical error constraints is far more effective than statistical sample reweighting alone. The test-set stability distributions of the E07b model across 1000 repeated iterations are shown in Fig. 6.

From a conventional ML perspective, data augmentation is typically regarded as a regularization technique for increasing sample diversity and preventing overfitting (Goodfellow et al., 2016). In thermobarometry, however, the essential function of EPMA-based augmentation is to encode known physical boundary conditions into the training process, transforming the regression task from unconstrained fitting across the full feature space into optimization within a physically bounded subspace. By internalizing geological prior knowledge into the training pipeline, this operation enables the model to extract authentic geological signals more effectively, thereby improving its adaptability and predictive credibility in real-world geological scenarios.

Hartmeier et al. (2025) calibrated a neural network biotite thermobarometer on a comparable premise of embedded geological priors. They explicitly incorporated the

topological sequence constraints of metamorphic mineral assemblages and the thermodynamic relationship of monotonically increasing Ti saturation with temperature into the training pipeline, reinforcing the relative P–T ranking among samples (Henry et al., 2005). Even with a comparatively limited training dataset (n = 2148), combining transfer learning with embedded thermodynamic priors constrained the temperature prediction RMSE to ±45 °C, an improvement magnitude comparable to the augmentation strategy presented here. Despite methodological differences in prior encoding, the underlying principle is shared: internalizing thermodynamic or crystal-chemical prior knowledge as effective training information, rather than relying solely on the intrinsic statistical patterns of the dataset.

Explicitly encoding physical information into the training process is a fundamental requirement for constructing reliable ML thermobarometers. Whether achieved through analytical error propagation, thermodynamic constraint embedding, or other integrations of geological prior knowledge, this strategy provides an effective and generalizable pathway for improving model generalization.

## 3.4 Marginal Returns of Algorithmic Complexity

Under the $G_3$ configuration (Augmented + *Liquid* + None), the three algorithm architectures exhibited remarkably similar overall performance (Table 5). Temperature RMSE differed by less than 1 °C among ERT, CatBoost, and Stacking, and the maximum difference in pressure RMSE was only 0.08 kbar. The more complex CatBoost and Stacking ensembles demonstrated no consistent performance advantage over the simpler ERT architecture.

An incremental data proportion experiment (Fig. 7) was conducted to investigate the underlying cause of this convergence. As the proportion of training samples increased, the rate of RMSE improvement decelerated continuously across all algorithms. Expanding the training subset from 80% to 100% reduced ERT temperature RMSE by only 1.68 °C and pressure RMSE by approximately 0.1 kbar. Even on the

complete dataset, the Stacking temperature RMSE (32.16 °C) remained higher than that of ERT (31.37 °C). These results confirm that, at the current data scale, the effective information extractable by the models is approaching its upper bound. Deploying more complex ensemble architectures does not unlock additional fitting potential under these conditions and may instead introduce additional model variance.

The $M_4$ Monte Carlo error propagation module was also applied to the E07b configuration (Augmented + ERT + None, *Liquid*). The results yielded average predictive standard deviations of $\sigma_T = 6.83$ °C for temperature and $\sigma_P = 0.45$ kbar for pressure, both attributable to input-side analytical uncertainties. This magnitude of propagated error agrees closely with the theoretical precision limits derived thermodynamically by Wang et al. (2025) and empirically through data-driven approaches by Weber and Blundy (2024). The agreement confirms that current tree-based models, such as ERT, are closely approaching the precision boundary imposed by existing data quality at the level of analytical error propagation.

Neither increasing model complexity nor optimizing hyperparameters delivers substantial marginal returns under the current feature set and dataset scale. Thermobarometer development should therefore shift away from architectural elaboration and prioritize instead the robust representation of geologically meaningful information, improvements in intrinsic data quality, and rigorous empirical validation against real geological scenarios.

## 3.5 Post-Processing Correction

Evaluation of the $G_4$ experimental group (Table 6) revealed that the overall benefit of piecewise linear correction is inherently limited. On the independent test set, RMSE improvement was marginal across all three algorithms. ERT exhibited the most pronounced correction gain (T: −1.96 °C; P: −0.16 kbar), whereas improvements for CatBoost and Stacking were negligible. This discrepancy is attributed to the precorrection systematic bias of each algorithm, as quantified by MBE. ERT displayed

a clear positive temperature MBE (+1.21 °C), and piecewise correction effectively reduced predictive deviations at extreme high and low end-members to near zero (Fig. 8). By contrast, CatBoost exhibited an MBE of only -0.04 °C, rendering it essentially unbiased and thereby providing no meaningful residual bias signal for post-processing to exploit.

These findings indicate that the practical effectiveness of post-processing correction depends strictly on the presence of systematic bias in the base model; correction is not a universal performance-enhancing tool. Constrained by the effective density of input information, post-processing can only address residual systematic biases locally, and cannot substitute for intrinsic data quality, robust feature representation, or the explicit encoding of physical information. The post-processing correction module should therefore be applied selectively, only when substantial systematic bias is confirmed (e.g., |MBE| exceeding ~5% of the model RMSE; equivalently |MBE| > 1 °C or 0.1 kbar in the present case study), rather than treated as a default component of the standard modeling workflow.

# 4. Conclusion

Using the clinopyroxene-liquid system as a baseline, a MBP was established for ML thermobarometry, designed to isolate the independent contributions of data processing, algorithm architecture, and post-processing modules to overall model performance. Incorporating coexisting melt compositional constraints and physics-informed data augmentation produced substantial, reproducible gains. At the current scale of available experimental datasets, the predictive capacity of thermobarometric models is governed primarily by the effective geological information encoded in the input variables, rather than by algorithmic complexity. Advancing ML thermobarometry will require prioritizing feature engineering guided by thermodynamics and crystal chemistry, improvements in model interpretability, and empirical validation against real geological systems. The proposed MBP established a reproducible experimental framework for driving systematic optimization of

thermobarometric workflows across diverse mineral systems.

## Acknowledgements

The authors are grateful for the constructive comments of Zhou Zhang. We also thank Zhuohao Zhao for his valuable assistance with the code design. We thank the editors and two anonymous reviewers for their helpful comments that greatly improved the manuscript. This work was financially supported by the National Key R&D Program of China (2022YFF0800404), the National Natural Science Foundation of China (Grant No. 42472086), and the open research fund (2025-Z01) of the State Key Laboratory of Critical Earth Material Cycling and Mineral Deposits, Nanjing University.

## Data availability

The benchmark dataset and source code used in this study are publicly available at the GitHub repository (https://github.com/foxsplendid/ml-thermobarometer-benchmark) and have been archived with a permanent DOI on Zenodo (https://doi.org/10.5281/zenodo.20112388). This study is based on the processed clinopyroxene–liquid experimental dataset of Jorgenson et al. (2022). The repository contains the code framework, details of the grid-based test-set construction, and additional benchmark outputs. Complete experimental configurations and supplementary performance metrics are also provided in the Supplementary Material associated with this article.

## References

Ágreda-López, M., Parodi, V., Musu, A., Jorgenson, C., Carfì, A., Mastrogiovanni, F., Caricchi, L., Perugini, D., Petrelli, M., 2024. Enhancing machine learning thermobarometry for clinopyroxene-bearing magmas. Comput. Geosci. 193,

105707.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Blundy, J., Cashman, K., 2008. Petrologic reconstruction of magmatic system variables and processes. Rev. Mineral. Geochem. 69, 179–239.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cashman, K.V., Sparks, R.S.J., Blundy, J.D., 2017. Vertically extensive and unstable magmatic systems: A unified view of igneous processes. Science 355, eaag3055. https://doi.org/10.1126/science.aag3055.

Chiaradia, M., Caricchi, L., 2022. Supergiant porphyry copper deposits are failed large eruptions. Commun. Earth Environ. 3, 107. https://doi.org/10.1038/s43247-022-00440-7.

Chicchi, L., Bindi, L., Fanelli, D., Tommasini, S., 2023. Frontiers of thermobarometry: GAIA, a novel deep learning-based tool for volcano plumbing systems. Earth Planet. Sci. Lett. 620, 118352.

Cutler, K.S., Cassidy, M., Blundy, J.D., 2024. Plagioclase-saturated melt hygrothermobarometry and plagioclase-melt equilibria using machine learning. Geochem. Geophys. Geosyst. 25, e2023GC011357.

Edmonds, M., Cashman, K.V., Holness, M., Jackson, M., 2019. Architecture and dynamics of magma reservoirs. Philos. Trans. R. Soc. A 377, 20180298. https://doi.org/10.1098/rsta.2018.0298.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA.

Hartmeier, P., Forshaw, J.B., Lanari, P., 2025. Calibration, validation and evaluation of machine learning thermobarometers in metamorphic petrology: An application to biotite and outlook for future strategy. J. Metamorph. Geol. 43, 755–780.

Henry, D.J., Guidotti, C.V., Thomson, J.A., 2005. The Ti-saturation surface for low-to-medium pressure metapelitic biotites: Implications for geothermometry and Ti-substitution mechanisms. Am. Mineral. 90, 316–328.

https://doi.org/10.2138/am.2005.1498.

Higgins, O., Sheldrake, T., Caricchi, L., 2022. Machine learning thermobarometry and chemometry using amphibole and clinopyroxene: A window into the roots of an arc volcano (Mount Liamuiga, Saint Kitts). Contrib. Mineral. Petrol. 177, 10.

Hirschmann, M.M., Ghiorso, M.S., Davis, F.A., Gordon, S.M., Mukherjee, S., Grove, T.L., Krawczynski, M., Médard, E., Till, C.B., 2008. Library of Experimental Phase Relations (LEPR): A database and web portal for experimental magmatic phase equilibria data. Geochem. Geophys. Geosyst. 9, Q03011.

Holland, T.J.B., Blundy, J.D., 1994. Non-ideal interactions in calcic amphiboles and their bearing on amphibole-plagioclase thermometry. Contrib. Mineral. Petrol. 116, 433–447. https://doi.org/10.1007/BF00310910.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Jorgenson, C., Caricchi, L., Petrelli, M., 2022. A machine learning-based approach to clinopyroxene thermobarometry model optimization and distribution. J. Geophys. Res. Solid Earth 127, e2021JB022904.

Li, A., Wu, S., Chen, H., et al., 2023. Explainable machine learning to uncover hydrogen diffusion mechanism in clinopyroxene. Chem. Geol. 641, 121776.

Li, X., Zhang, C., 2022. Machine learning thermobarometry for biotite-bearing magmas. J. Geophys. Res. Solid Earth 127, e2022JB024137.

Lin, R., Wu, T., Zhu, H., Lu, J., Tian, L., Hong, Y., 2026. Dynamic magmatic evolution of boninites from the Challenger Deep, southernmost Mariana Trench: Insights from mineral thermobarometry and MELTS modeling. Lithos 522-523, 108381.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Musu, A., Parodi, V., Toplak, M., Carfì, A., Ágreda-López, M., Mastrogiovanni, F.,

ZhangZhou, J., Perugini, D., Belmonte, D., Wieser, P.E., Zupan, B., Petrelli, M., 2025. Orange-Volcanoes: A new open and collaborative platform to perform data-driven investigations and machine learning analyses in petrology and volcanology. Appl. Comput. Geosci. 27, 100270.

Neave, D.A., Putirka, K.D., 2017. A new clinopyroxene-liquid barometer, and implications for magma storage pressures under Icelandic rift zones. Am. Mineral. 102, 777–794. https://doi.org/10.2138/am-2017-5968.

Nimis, P., Ulmer, P., 1998. Clinopyroxene geobarometry of magmatic rocks. Part 1: An expanded structural geobarometer for anhydrous and hydrous, basic and ultrabasic systems. Contrib. Mineral. Petrol. 133, 122–135. https://doi.org/10.1007/s004100050442.

Petrelli, M., 2024. Machine learning in petrology: State-of-the-art and future perspectives. J. Petrol. 65, egae036. https://doi.org/10.1093/petrology/egae036.

Petrelli, M., Caricchi, L., Perugini, D., 2020. Machine learning thermo-barometry: Application to clinopyroxene-bearing magmas. J. Geophys. Res. Solid Earth 125, e2020JB020130.

Putirka, K.D., 2008. Thermometers and barometers for volcanic systems. Rev. Mineral. Geochem. 69, 61–120.

Qin, B., Ye, C., Liu, J., Huang, S., Wang, S., Zhang, Z., 2024. Mapping global lithospheric mantle pressure-temperature conditions by machine-learning thermobarometry. J. Geophys. Res. Solid Earth 129, e2023JB027961.

Richards, J.P., 2013. Giant ore deposits formed by optimal alignments and combinations of geological processes. Nat. Geosci. 6, 911–916. https://doi.org/10.1038/ngeo1920.

Ridolfi, F., Renzulli, A., 2012. Calcic amphiboles in calc-alkaline and alkaline magmas: Thermobarometric and chemometric empirical equations valid up to 1,130 °C and 2.2 GPa. Contrib. Mineral. Petrol. 163, 877–895. https://doi.org/10.1007/s00410-011-0704-6.

Taniwaki, Y., Fukui, T., Saito, S., Fukuyama, M., 2025. Hygrothermobarometry for granites using melt inclusions in zircon. Lithos 504-505, 108029.

Thomson, A.R., Walter, M.J., Kohn, S.C., Brooker, R.A., 2021. Evaluating the formation pressure of diamond-hosted majoritic garnets: A machine learning majorite barometer. J. Geophys. Res. Solid Earth 126, e2020JB020604.

Wang, X., Hou, T., Wieser, P.E., Zhang, Z., 2025. Thermodynamic insights into the reliability of mineral-based thermobarometers. Commun. Earth Environ. 6, 913. https://doi.org/10.1038/s43247-025-02831-y.

Weber, G., Blundy, J., 2024. A machine learning-based thermobarometer for magmatic liquids. J. Petrol. 65, egae020.

Wieser, P.E., Kent, A.J.R., Till, C.B., 2023. Barometers behaving badly II: A critical evaluation of Cpx-only and Cpx-Liq thermobarometry in variably-hydrous arc magmas. J. Petrol. 64, egad050.

Wieser, P.E., Petrelli, M., Lubbers, J., et al., 2022. Thermobar: An open-source Python3 tool for thermobarometry and hygrometry. Volcanica 5, 349–384.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39, 151–160.

Zhang, Q., Hardman, M.F., Stachel, T., Chinn, I., Seller, M., Kjarsgaard, B., Pearson, D.G., 2025. A machine learning approach to single garnet geothermometry and application to tracing the fingerprint of superdeep diamonds. Geochem. Geophys. Geosyst. 26, e2024GC012124. https://doi.org/10.1029/2024GC012124.

ZhangZhou, J., He, C., Sun, J., Zhao, J., Lyu, Y., Wang, S., Zhao, W., Li, A., Ji, X., Agarwal, A., 2024. Geochemistry π: Automated machine learning Python framework for tabular data. Geochem. Geophys. Geosyst. 25, 2023GC011324.

Zhao, H., Zhang, Y., Shao, Y., Liao, J., Song, S., Cao, G., Tan, R., 2024. A New Sphalerite Thermometer Based on Machine Learning with Trace Element Geochemistry. Nat. Resour. Res. 33, 2609–2626. https://doi.org/10.1007/s11053-024-10408-3.

Zhao, K., Xu, X., He, Z., Xia, Y., 2023. The obscuring effect of magma recharge on the connection of volcanic-plutonic rocks. Am. Mineral. 108, 2260–2282.

Zhao, S., Xia, Y., Xu, X., Zhao, K., 2023. Timescale of magma recharge for a crystal-rich mush: Perspectives from the compositional zoned quartz in Xiangshan caldera, SE China. Lithos 456-457, 107288.

Zhu, J.J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: Common pitfalls and best practices. Environ. Sci. Technol. 57, 17671–17689.

# Figure Captions

Fig. 1. P–T distribution of the calibration dataset (n = 2079; Jorgenson et al., 2022).
(a) Two-dimensional P–T density plot with hexagonal binning and a logarithmic color scale. Horizontal dashed lines denote P = 2.5 kbar (orange) and P = 20 kbar (red). (b) Marginal distribution (histogram) and cumulative distribution function (CDF, solid black line) of pressure. Vertical dashed lines denote P = 2.5 kbar (orange) and P = 20 kbar (red).

Fig. 2. Modular benchmarking protocol used in this study.
(a) Dual feature–set configurations. (b) $M_1$ compares three alternative data treatments: Raw, Balanced, and Augmented. (c) $M_2$ evaluates three representative algorithms: CatBoost, ERT, and Stacking. (d) $M_3$ applies a segmented post-processing correction to reduce systematic bias. (e) $M_4$ performs Monte Carlo propagation of input analytical uncertainty as an auxiliary module.

Fig. 3. Marginal P–T distributions of cross-validation folds.
Marginal kernel density distributions of (a) temperature and (b) pressure for each of the 10 cross-validation folds (thin grey curves) and for the complete calibration dataset (n = 2079; thick colored curve).

Fig. 4. Performance comparison between *NoLiquid* and *Liquid* feature sets (E07 configuration).

1:1 comparison plots for (a, b) temperature and (c, d) pressure predictions, generated by the E07a (NoLiquid; left column) and E07b (Liquid; right column) configurations (Augmented + ERT + None). Scatter points show aggregated out-of-fold predictions from the 10-fold cross-validation, and red dashed lines indicate the ideal 1:1 fit.

Fig. 5. SHAP interpretability analysis for the ERT model (E07 configuration).

Upper row (a, b): *Liquid* configuration (18 features); lower row (c, d): *NoLiquid* configuration (9 CPX-only features). Left column (a, c): temperature; right column (b, d): pressure. Each panel combines a SHAP beeswarm with a mean |SHAP| bar chart (upper x-axis, semi-transparent). In the beeswarm, each point represents one sample, with horizontal position indicating signed contribution and color indicating feature value (red = high; blue = low). Features are ordered top to bottom by descending importance, and the mean |SHAP| scale differs between rows.

Fig. 6. Predictive error distributions assessing the stability of the E07b model.

Test set error distributions of (a, b) RMSE, (c, d) MAE, and (e, f) MBE, derived from 1000 repeated experiments. Left column (a, c, e): temperature; right column (b, d, f): pressure. Histograms are overlaid with kernel density estimation (KDE) curves; black dashed lines denote the mean values.

Fig. 7. Learning curves for ERT and Stacking models under E07b configuration.

Cross-validation RMSE for (a) temperature and (b) pressure is plotted against the number of training samples. Error bars indicate the standard deviation across repeated experiments.

Fig. 8. Correction magnitudes of the segmented linear post-processing (E10b

configuration).

Out-of-fold prediction results for (a) temperature and (b) pressure. The y-axis shows the correction magnitude ($\Delta = Pred_{corr} - Pred_{raw}$). The red dashed line indicates the smoothed trend, with the 95% confidence interval shown by gray shading. Two vertical dotted lines (q33 and q67) mark the quantile boundaries used for segmented linear correction. Marginal histograms show the distributions of true values (top) and correction magnitudes (right).

Table 1. Performance comparison of machine learning algorithms in recent thermobarometer studies

| Reference | Mineral System | Algorithm | Pressure Range | Temperature Range | Reported Pressure Error (RMSE/SEE) | Reported Temperature Error (RMSE/SEE) |
|---|---|---|---|---|---|---|
| Petrelli et al., 2020 | cpx–liq | ERT | 0–40 kbar | 952–1882 K(≈ 679–1610 °C) | RMSE 2.6 kbar | RMSE 40 K |
| Petrelli et al., 2020 | cpx-only | ERT | 0–40 kbar | 952–1882 K(≈ 679–1610 °C) | RMSE 3.2 kbar | RMSE 66 K |
| Jorgenson et al., 2022 | cpx–liq | ERT | 0–30 kbar | 679–2180°C | SEE 2.7 kbar | SEE 44.9 °C |
| Jorgenson et al., 2022 | cpx-only | ERT | 0–30 kbar | 679–2180°C | SEE 3.2 kbar | SEE 72.5 °C |
| Ágreda-López et al., 2024 | cpx-only | ERT | 0–30 kbar | 700–1600 °C | CV 2.5 kbar; Test 2.3 kbar | CV 57°C; Test 84°C |
| Ágreda-López et al., 2024 | cpx–liq | ERT | 0–30 kbar | 700–1600 °C | CV 2.1 kbar; Test 2.2 kbar | CV 36°C; Test 44°C |
| Higgins et al., 2022 | amp-only | ERT | 0.002–12 kbar | 750–1250°C | SEE 1.6 kbar | SEE 40 °C |
| Higgins et al., 2022 | cpx-only | ERT | 0.002–12 kbar | 750–1250°C | SEE 2.3 kbar | SEE 57 °C |

| | | | | | | |
|---|---|---|---|---|---|---|
| Li & Zhang, 2022 | bt-only | ERT | 1–48 kbar | 625–1325°C | RMSE 4.36 kbar; Mean model error 4.7 kbar | RMSE 54°C; Mean model error 65°C |
| Li & Zhang, 2022 | bt–liq | ERT | 1–48 kbar | 625–1325°C | RMSE 2.38 kbar; Mean model error 3.2 kbar | RMSE 35°C; Mean model error 38°C |
| Thomson et al., 2021 | Diamond-Hosted Majoritic Garnets | Random Forest | 60–250 kbar | — | RMSE 21.24 kbar | — |
| Weber & Blundy, 2024 | melt-assemblage | ERT | 0.2–15 kbar | 675–1400 °C | RMSE 1.7–1.9 kbar | RMSE 36–42 °C |
| Cutler et al., 2024 | plag–liq | Random Forest | ⩽5 kbar | 664–1355 °C | 0.76 kbar（T/$H_2O$ dependent） | 25 °C（T/$H_2O$ dependent） |
| Qin et al., 2024 | multi-mineral | XGBoost | 20–120 kbar | 800–1,350°C | ±5 kbar | ±56 °C |
| Chicchi et al., 2023 | cpx-only | FNN | 1 bar–10 kbar | 0–1500 °C | 0.27 kbar | 12°C |
| Chicchi et al., 2023 | cpx–liq | FNN | 1 bar–10 kbar | 0–1500 °C | 0.15 kbar | 5 °C |
| Zhang et al., 2025 | garnet-only | XGBoost | — | 700–1450 °C | — | Mean RMSE ≈79°C; RMSE ≈61°C (900–1400°C) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Zhao et al., 2024 | Sphalerite trace elements | Random Forest | — | 75–430 °C | — | 25 °C（5-fold CV ） |
| E07b (this study) | cpx–liq | ERT | 0–30 kbar | 700–1600 °C | CV 1.93 kbar; Test 3.23 kbar | CV 30.98 °C; Test 54.65 °C |

**Note:** Direct numerical comparisons between studies should be made with caution because reported error metrics (e.g., RMSE, SEE), validation strategies, and calibration ranges differ. Temperatures originally reported in Kelvin are converted to °C. "—" indicates data not reported. *$T/H_2O$-dependent* denotes that the errors vary as a function of temperature and water content. FNN denotes Feedforward Neural Network; SEE denotes Standard Error of Estimate.

Table 2. Summary of the modular experimental configurations

| Group | M1 (Data) | M2 (Algorithm) | M3 (Correction) | Feature Set |
|---|---|---|---|---|
| G1 | Raw | ERT / CatBoost / Stacking | None | NoLiquid / Liquid |
| G2 | Balanced | ERT / CatBoost / Stacking | None | NoLiquid / Liquid |
| G3 | Augmented | ERT / CatBoost / Stacking | None | NoLiquid / Liquid |
| G4 | Augmented | ERT / CatBoost / Stacking | Segmented | NoLiquid / Liquid |

Table 3. Predictive performance of the *NoLiquid* versus *Liquid* feature sets

| Target Variable | Feature Set | RMSE | $R^2$ | Relative Reduction |
|---|---|---|---|---|
| Temperature (°C) | NoLiquid | 54.35 ± 4.49 | 0.795 ± 0.027 | – |
| | Liquid | 32.13 ± 3.87 | 0.928 ± 0.017 | -40.9% |
| Pressure (kbar) | NoLiquid | 2.43 ± 0.20 | 0.870 ± 0.022 | – |
| | Liquid | 2.04 ± 0.21 | 0.908 ± 0.021 | -16.1% |

**Note:** Results show the mean ± standard deviation across the 10-fold P-T grid-stratified cross-validation using the G1 baseline configuration (*Raw* + ERT + *None*). Relative reduction evaluates the percentage drop in RMSE upon adding melt features.

Table 4. Impact of physics-informed data augmentation on predictive performance

| Target Variable | Data Strategy | RMSE | $R^2$ | Relative Reduction |
|---|---|---|---|---|
| Temperature (°C) | Raw (G1) | 32.13 ± 3.87 | 0.928 ± 0.017 | - |
| | Augmented (G3) | 30.98 ± 3.29 | 0.933 ± 0.014 | -3.6% |
| Pressure (kbar) | Raw (G1) | 2.04 ± 0.21 | 0.908 ± 0.021 | - |
| | Augmented (G3) | 1.93 ± 0.25 | 0.917 ± 0.024 | -5.4% |

**Note:** Results show the mean ± standard deviation across the 10-fold cross-validation for ERT models using the *Liquid* feature set.

Table 5. Performance comparison of different algorithmic architectures

| Algorithm | Temperature RMSE (°C) | Pressure RMSE (kbar) | Model Parameters |
|---|---|---|---|
| ERT | 30.98 ± 3.29 | 1.93 ± 0.25 | 200 trees × 15 depth |
| CatBoost | 31.43 ± 2.99 | 1.96 ± 0.24 | 1000 iterations × 6 depth |
| Stacking | 31.93 ± 3.40 | 2.01 ± 0.27 | ERT+CatBoost+RF |

**Note:** Results (mean ± standard deviation) are based on the 10-fold cross-validation under the G3 configuration (*Augmented* + *Liquid* + *None*). Reported values represent theoretical model performance rather than actual high-temperature and high-pressure experimental results.

Table 6. Evaluation of the segmented linear post-processing correction

| Algorithm | G3 T_CV | G4 T_CV | Δ CV | G3 T_Test | G4 T_Test | Δ Test | G3 P_CV | G4 P_CV | Δ CV | G3 P_Test | G4 P_Test | Δ Test |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ERT | 30.98± 3.29°C | 30.51± 3.13°C | −0.47°C | 54.65°C | 52.69°C | −1.96°C | 1.93 ±0.25 kbar | 1.89 ±0.25 kbar | −0.04 kbar | 3.23 kbar | 3.07 kbar | −0.16 kbar |
| CatBoost | 31.43± 2.99°C | 31.37± 2.96°C | −0.06°C | 54.44°C | 53.78°C | −0.66°C | 1.96 ±0.24 kbar | 1.93 ±0.24 kbar | −0.03 kbar | 3.11 kbar | 3.04 kbar | −0.07 kbar |
| Stacking | 31.93± 3.40°C | 31.64± 3.36°C | −0.29°C | 57.05°C | 56.11°C | −0.94°C | 2.01 ±0.27 kbar | 1.99 ±0.26 kbar | −0.02 kbar | 3.50 kbar | 3.42 kbar | −0.08 kbar |

**Note:** Results compare the uncorrected G3 and corrected G4 configurations using the *Liquid* feature set. Performance is evaluated on both the 10-fold cross-validation (CV) and a fixed independent test set (n = 349). Δ denotes the change in RMSE (G4 − G3), where negative values indicate a performance improvement. Complete MBE statistics are provided in Supplementary Table S2.

**Declaration of interests**

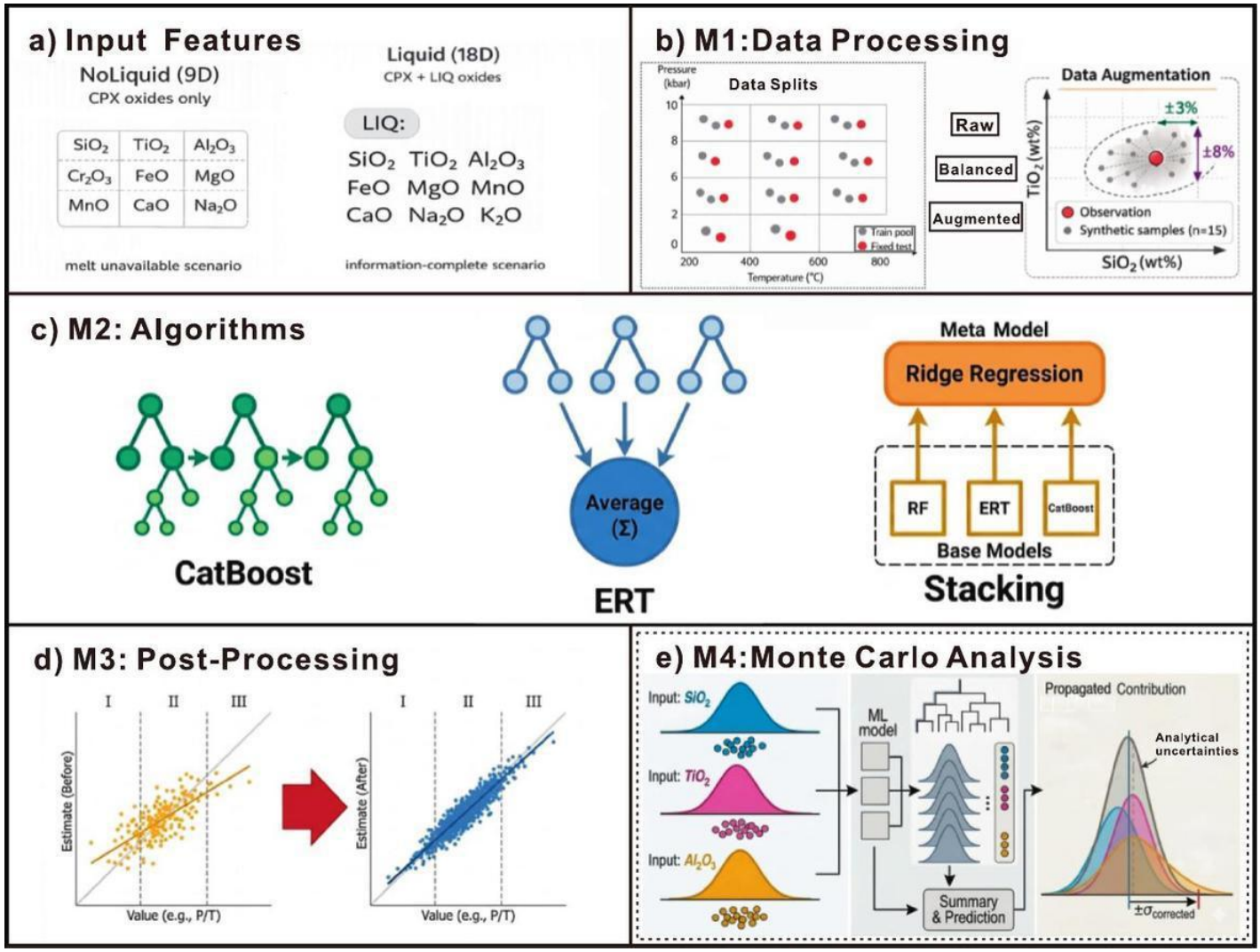
☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Graphical abstract

## Highlights

- A Modular Benchmarking Protocol (MBP) is proposed for ML thermobarometry.
- Coexisting melt compositions reduce T/P RMSE by 40.9%/16.1% vs. NoLiquid baseline.
- Physics-informed data augmentation significantly enhances model robustness.
- Predictive precision approaches the analytical limit.
- Input information density is the primary bottleneck for thermobarometric accuracy.




Graphics Abstract

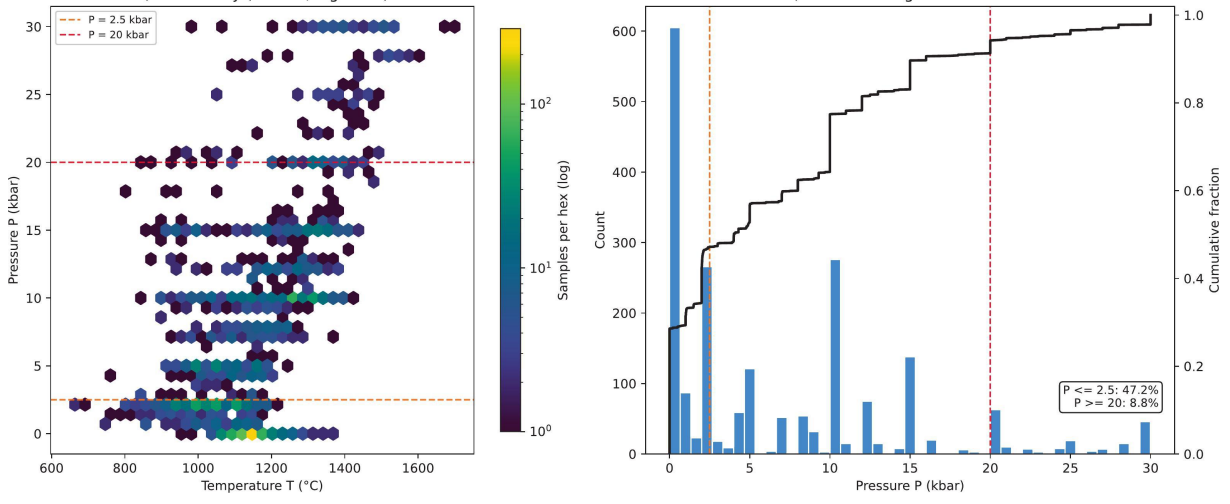

a) P-T Density (Hexbin, log scale)
P = 2.5 kbar
P = 20 kbar
Pressure P (kbar)
Temperature T (°C)
Samples per hex (log)
b) Pressure Marginal Distribution
Count
Cumulative fraction
Pressure P (kbar)
P <= 2.5: 47.2%
P >= 20: 8.8%

Figure 1

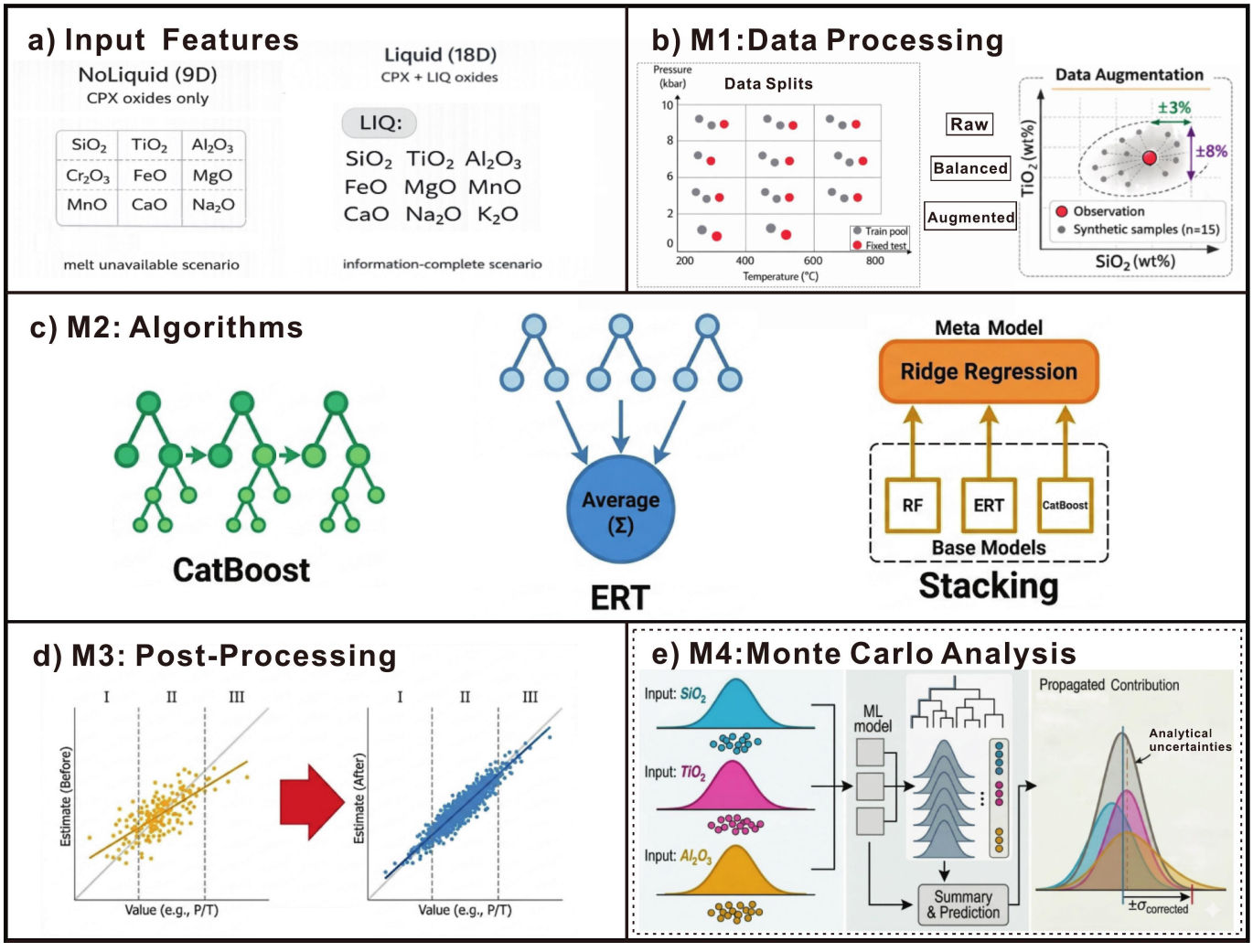

a) Input Features
NoLiquid (9D)
CPX oxides only
$SiO_2$ $TiO_2$ $Al_2O_3$
$Cr_2O_3$ FeO MgO
MnO CaO $Na_2O$
melt unavailable scenario
Liquid (18D)
CPX + LIQ oxides
LIQ:
$SiO_2$ $TiO_2$ $Al_2O_3$
FeO MgO MnO
CaO $Na_2O$ $K_2O$
information-complete scenario
b) M1:Data Processing
Pressure (kbar)
Data Splits
Train pool
Fixed test
Temperature (°C)
Raw
Balanced
Augmented
Data Augmentation
±3%
±8%
$TiO_2$ (wt%)
Observation
Synthetic samples (n=15)
$SiO_2$ (wt%)
c) M2: Algorithms
CatBoost
Average (Σ)
ERT
Meta Model
Ridge Regression
RF
ERT
CatBoost
Base Models
Stacking
d) M3: Post-Processing
I
II
III
Estimate (Before)
Value (e.g., P/T)
Estimate (After)
e) M4:Monte Carlo Analysis
Input: $SiO_2$
Input: $TiO_2$
Input: $Al_2O_3$
ML model
Summary & Prediction
Propagated Contribution
Analytical uncertainties
$\pm\sigma_{corrected}$

Figure 2

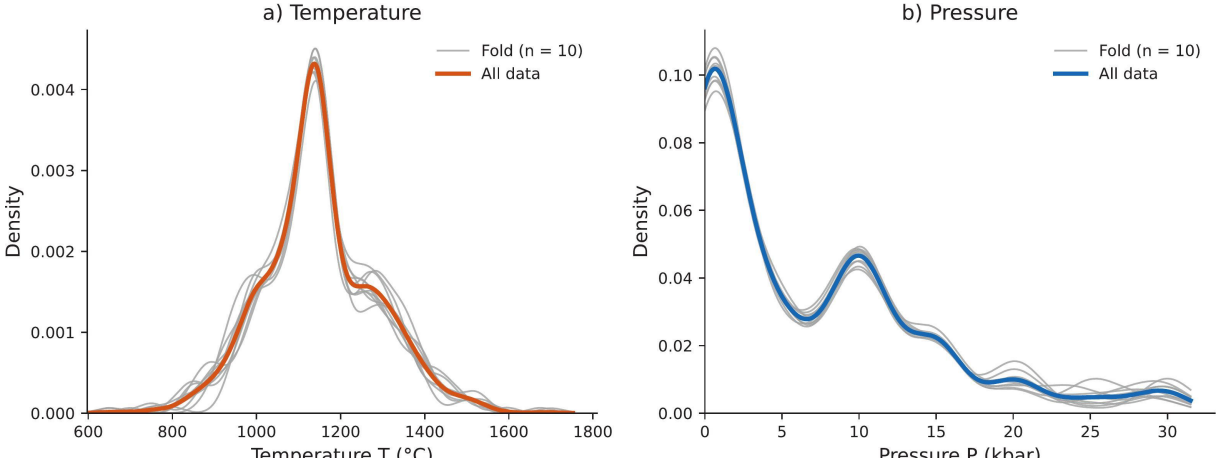

a) Temperature
Fold (n = 10)
All data
Density
0.000
0.001
0.002
0.003
0.004
600
800
1000
1200
1400
1600
1800
Temperature T (°C)
b) Pressure
Fold (n = 10)
All data
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
Pressure P (kbar)

Figure 3

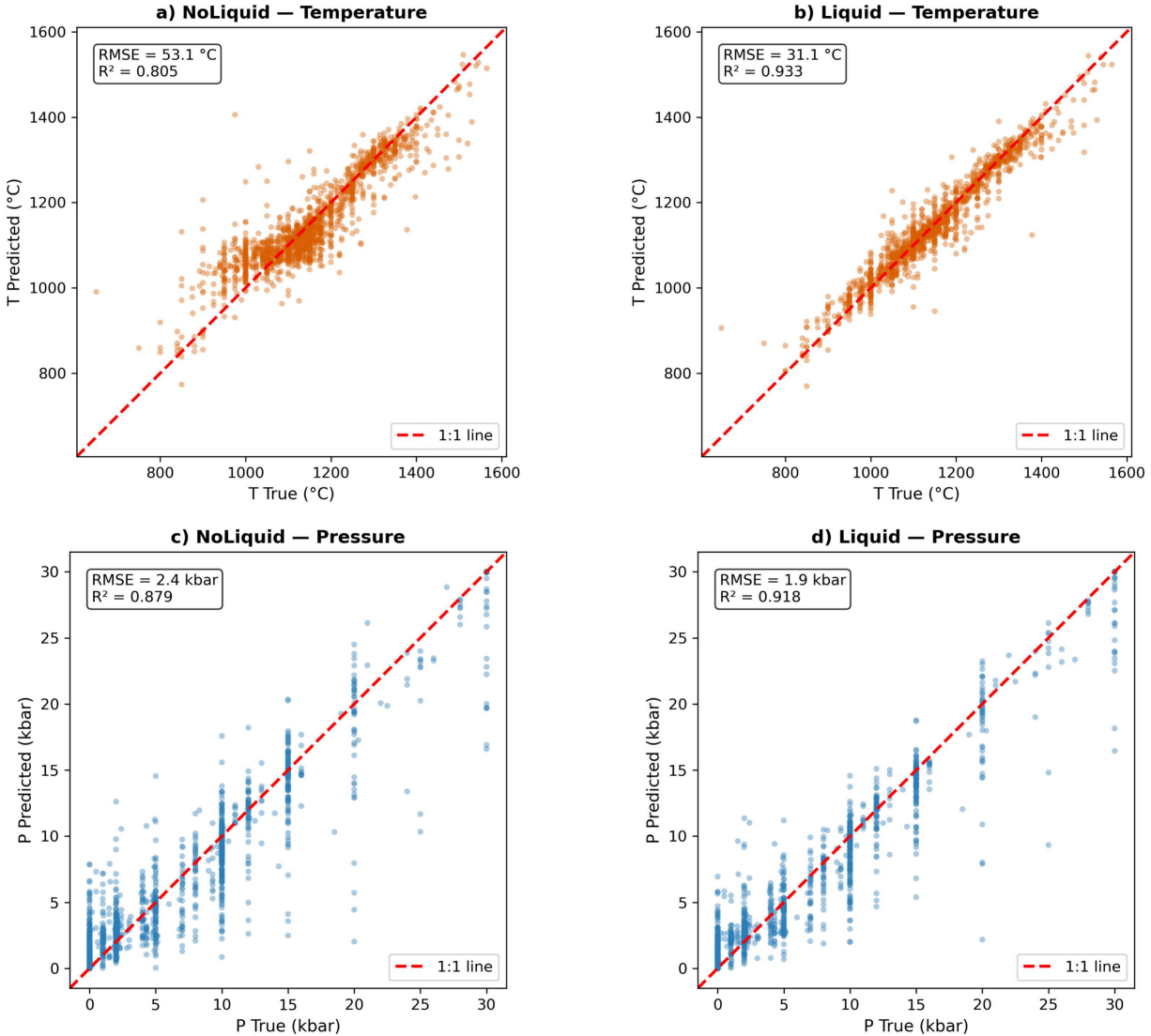

a) NoLiquid — Temperature
RMSE = 53.1 °C
R² = 0.805
1:1 line
T Predicted (°C)
T True (°C)
b) Liquid — Temperature
RMSE = 31.1 °C
R² = 0.933
1:1 line
T Predicted (°C)
T True (°C)
c) NoLiquid — Pressure
RMSE = 2.4 kbar
R² = 0.879
1:1 line
P Predicted (kbar)
P True (kbar)
d) Liquid — Pressure
RMSE = 1.9 kbar
R² = 0.918
1:1 line
P Predicted (kbar)
P True (kbar)

Figure 4

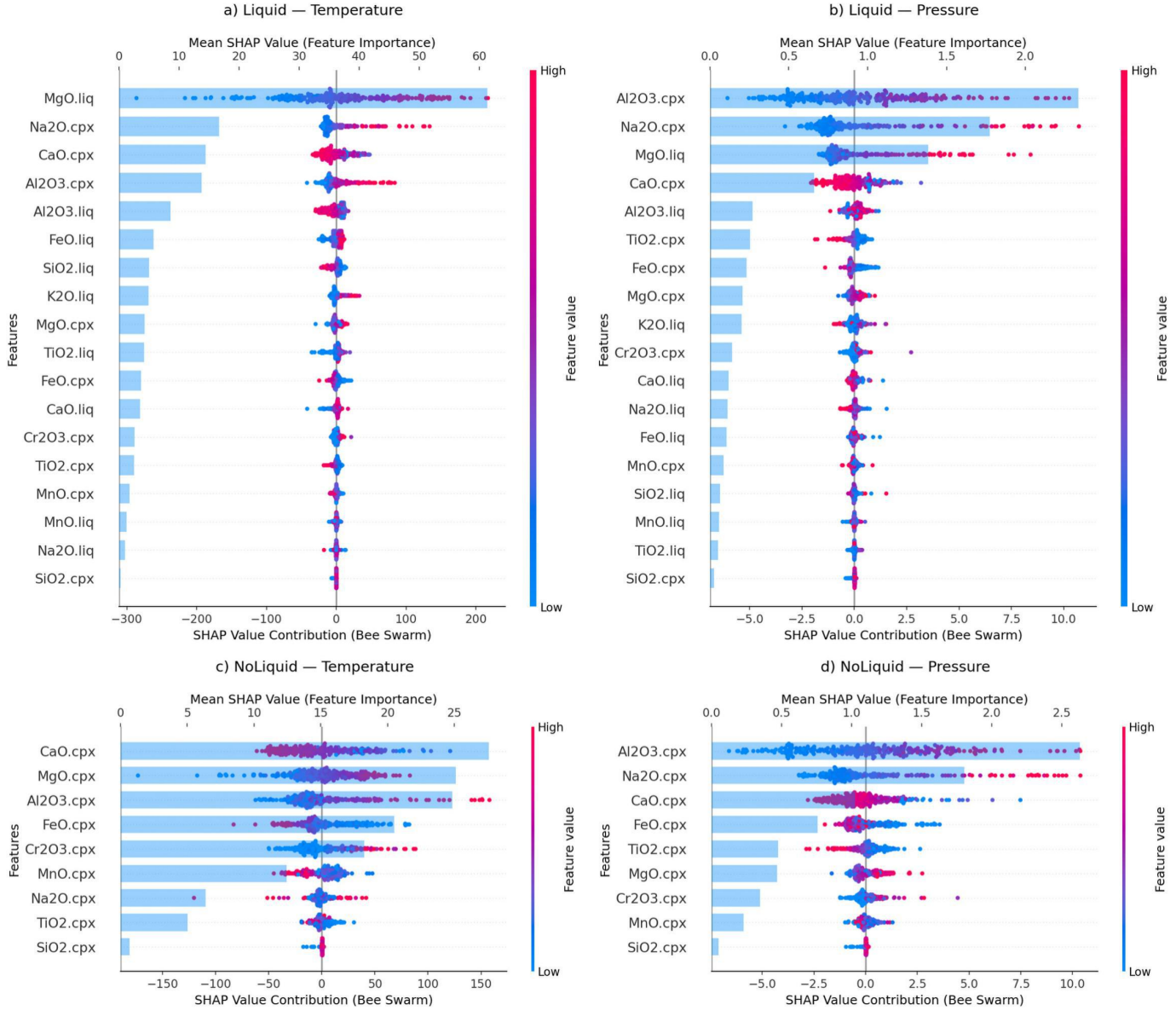

a) Liquid — Temperature
Mean SHAP Value (Feature Importance)
MgO.liq
Na2O.cpx
CaO.cpx
Al2O3.cpx
Al2O3.liq
FeO.liq
SiO2.liq
K2O.liq
MgO.cpx
TiO2.liq
FeO.cpx
CaO.liq
Cr2O3.cpx
TiO2.cpx
MnO.cpx
MnO.liq
Na2O.liq
SiO2.cpx
Features
SHAP Value Contribution (Bee Swarm)
Feature value
High
Low
b) Liquid — Pressure
Mean SHAP Value (Feature Importance)
Al2O3.cpx
Na2O.cpx
MgO.liq
CaO.cpx
Al2O3.liq
TiO2.cpx
FeO.cpx
MgO.cpx
K2O.liq
Cr2O3.cpx
CaO.liq
Na2O.liq
FeO.liq
MnO.cpx
SiO2.liq
MnO.liq
TiO2.liq
SiO2.cpx
Features
SHAP Value Contribution (Bee Swarm)
Feature value
High
Low
c) NoLiquid — Temperature
Mean SHAP Value (Feature Importance)
CaO.cpx
MgO.cpx
Al2O3.cpx
FeO.cpx
Cr2O3.cpx
MnO.cpx
Na2O.cpx
TiO2.cpx
SiO2.cpx
Features
SHAP Value Contribution (Bee Swarm)
Feature value
High
Low
d) NoLiquid — Pressure
Mean SHAP Value (Feature Importance)
Al2O3.cpx
Na2O.cpx
CaO.cpx
FeO.cpx
TiO2.cpx
MgO.cpx
Cr2O3.cpx
MnO.cpx
SiO2.cpx
Features
SHAP Value Contribution (Bee Swarm)
Feature value
High
Low

Figure 5

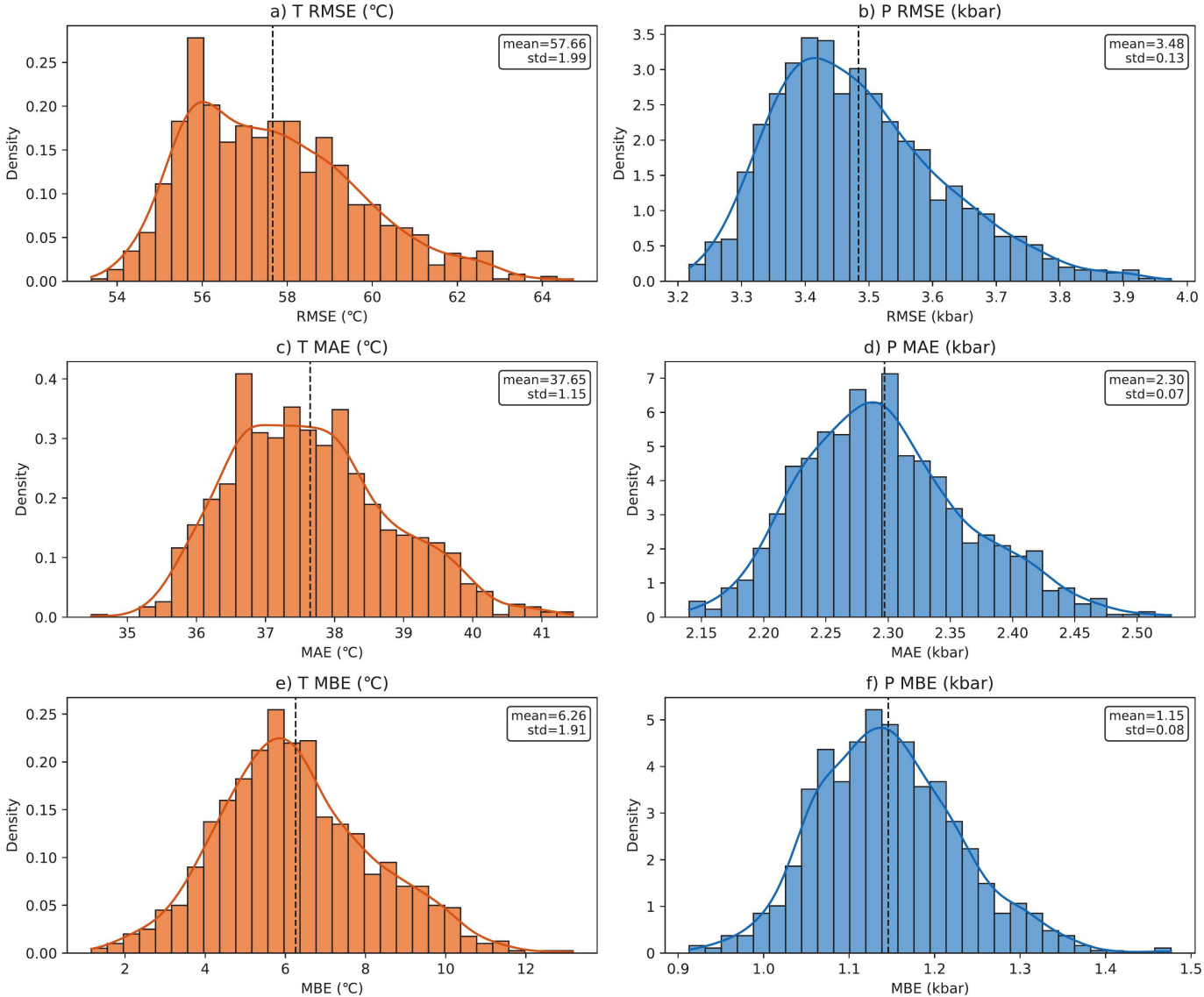

a) T RMSE (℃)
mean=57.66
std=1.99
Density
RMSE (℃)
b) P RMSE (kbar)
mean=3.48
std=0.13
Density
RMSE (kbar)
c) T MAE (℃)
mean=37.65
std=1.15
Density
MAE (℃)
d) P MAE (kbar)
mean=2.30
std=0.07
Density
MAE (kbar)
e) T MBE (℃)
mean=6.26
std=1.91
Density
MBE (℃)
f) P MBE (kbar)
mean=1.15
std=0.08
Density
MBE (kbar)

Figure 6

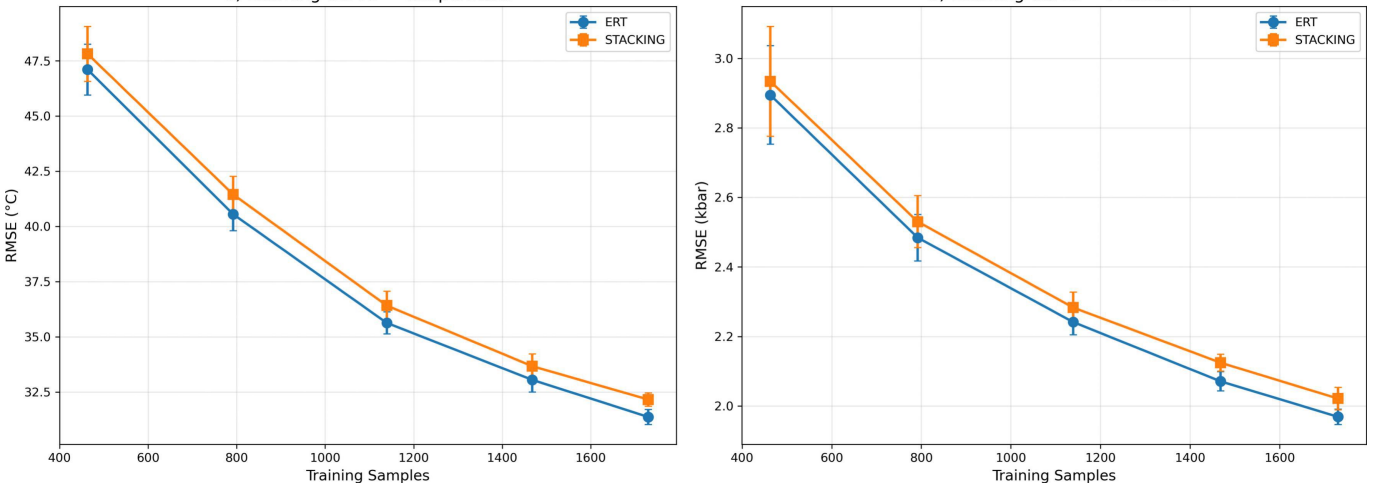

a) Learning Curve — Temperature
ERT
STACKING
RMSE (°C)
Training Samples
b) Learning Curve — Pressure
ERT
STACKING
RMSE (kbar)
Training Samples

Figure 7

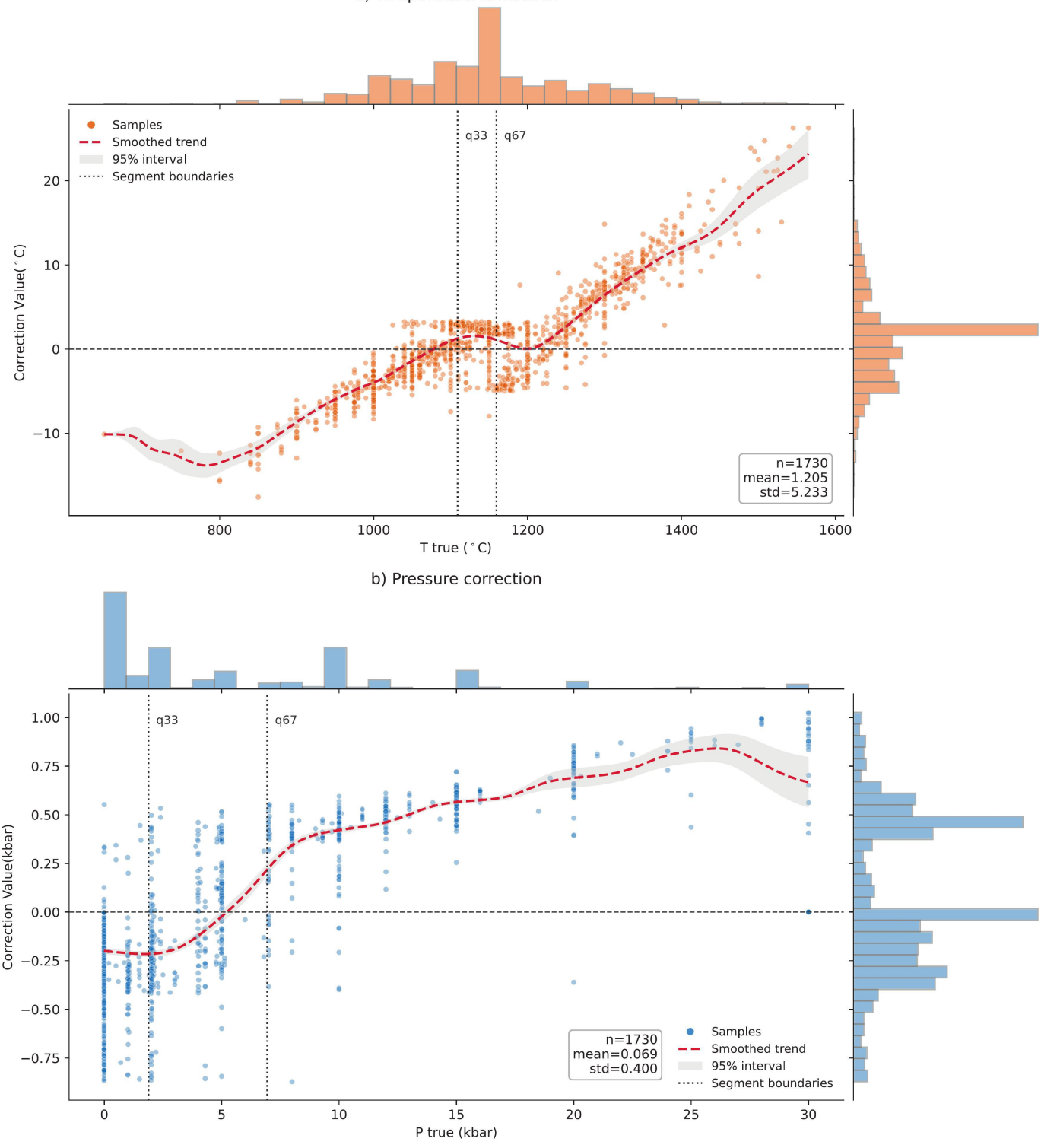

a) Temperature correction
Samples
Smoothed trend
95% interval
Segment boundaries
q33
q67
n=1730
mean=1.205
std=5.233
Correction Value(°C)
T true (°C)
b) Pressure correction
q33
q67
n=1730
mean=0.069
std=0.400
Samples
Smoothed trend
95% interval
Segment boundaries
Correction Value(kbar)
P true (kbar)

Figure 8