Journal of Zhejiang University-SCIENCE A 2026 27(8):852-870
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.A2500628

Check for updates

Mechanism-enhanced multitask distillation for predictive, interpretable design of biomass-based activated carbons

Peng ZHAO[1], Jiahui ZHOU[2], Guangyao LI[1,3], Hao YU[1,3]✉, Dongxu JI[3], Takahiko MIYAZAKI[1,4], Kyaw THU[1,4]

[1]*Department of Advanced Environmental Science and Engineering, Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan*
[2]*Department of Information Science and Technology, Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan*
[3]*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China*
[4]*Research Center for Next Generation Refrigerant Properties (NEXT-RP), International Institute for Carbon-Neutral Energy Research (I2CNER), Kyushu University, Fukuoka 819-0395, Japan*

Abstract: Designing biomass-derived activated carbons (ACs) is challenged by heterogeneous synthesis routes and multiobjective trade-offs among specific surface area (S_{BET}), total pore volume (V_T), and mass yield (M_{yield}). This study presents a mechanism-enhanced multitask distillation framework (AC-ResKD) built on a shared residual deep neural network (ResDNN) and dual priors from prediction-level and teacher-aware knowledge distillation (PD-KD and TA-KD). Process factors (agent, activation temperature/time, impregnation, heating rate, precursor composition) are modeled jointly with teacher predictions to learn an interpretable mapping across combined and separated one-step/two-step datasets. All results are reported as the mean and 95% confidence intervals over 20 repeated random 80/20 splits. On separated routes, TA-KD achieves robust accuracy for S_{BET} (one-step: R^2=0.806 [0.787, 0.824]; two-step: R^2=0.801 [0.773, 0.830]) and V_T (one-step: R^2=0.787 [0.764, 0.810]; two-step: R^2=0.817 [0.791, 0.843]). On the combined set, TA-KD yields the strongest gains for M_{yield} (R^2=0.816 [0.802, 0.831]; root mean squared error E_{RMS}=5.047 [4.853, 5.242]), while improving S_{BET} and V_T as well. Overall, M_{yield} is the most consistent beneficiary of distillation, and route-separated training exhibits improved monotonicity and reduced bias relative to combined training; the two-step route shows stronger V_T predictability consistent with a pre-carbonization priming effect. Explainability based on permutation feature importance (PFI) and Shapley additive explanations (SHAP) identifies teacher outputs, agent type, and thermal severity as dominant drivers. Impregnation governs V_T in one-step activation, while pyrolysis variables rise in importance in two-step activation. Robust Pareto screening with quantile-window extraction delivers agent- and route-specific operating envelopes (temperature–dose–time), enabling simultaneous S_{BET}/V_T improvement under bounded M_{yield} penalties. AC-ResKD thus provides accurate, interpretable, and actionable guidance for artificial intelligence (AI)-assisted AC design in heterogeneous, data-scarce settings.

Key words: Activated carbon (AC) synthesis; Pore structure prediction; Knowledge distillation (KD); Multitask learning (MTL); Process optimization

1 Introduction

The urgency of the transition to sustainable energy systems is underscored by escalating global carbon emissions, which reach approximately 37.5 Gt (fossil CO_2), necessitating the immediate deployment of effective capture and storage technologies (Lamb et al., 2022; Kundu et al., 2024). However, this transition is constrained by intermittency, storage inefficiency, and lifecycle emissions across production chains. High-performance porous materials provide essential functions that mitigate these bottlenecks, including adsorption, separation, catalysis, and energy storage (Yu et al., 2023; Li et al., 2025a). Among these, activated carbon (AC) occupies a central role due to its unique combination of large specific surface area (S_{BET}), tunable pore-size distributions, chemical stability, and low-cost manufacture from biomass precursors (Sevilla and Mokaya, 2014; Heidarinejad et al., 2020; Yudha et al., 2024). These properties make ACs indispensable for critical

✉ Hao YU, yuhao@cuhk.edu.cn
Hao YU, https://orcid.org/0000-0002-9146-4462

Received Nov. 30, 2025; Revision accepted Apr. 8, 2026; Crosschecked May 18, 2026; Online first June 18, 2026

technologies such as post-combustion CO_2 capture, solvent purification, supercapacitors, and metal–air batteries (Yu et al., 2024; Li et al., 2025b; Zhou et al., 2026). Despite the dominance of zeolites and metal-organic frameworks (MOFs) in the academic literature, ACs remain the industrial benchmark for stability and regeneration ease (Devi et al., 2023; Kaur et al., 2023).

AC production typically follows physical activation (e.g., steam or CO_2 at high temperature) or chemical activation (e.g., KOH, H_3PO_4, or $ZnCl_2$) routes, implemented via either one-step (coupled carbonization–activation) or two-step (pyrolysis followed by activation) protocols (Othman et al., 2013; Fu et al., 2023). Process severity—governed by temperature, residence time, heating rate, and impregnating agent dose—determines pore development, while precursor composition (volatile matter (VM), ash, fixed carbon (FC)) sets the structural stability of the evolving carbon skeleton (Muhammed et al., 2021; Sosa et al., 2023; Chairunnisa et al., 2024). Because application performance strictly depends on a delicate balance among S_{BET}, total pore volume (V_T), and mass yield (M_{yield}), the predictive control of synthesis remains a trial-and-error challenge, consuming significant experimental resources.

The paradigm of materials discovery is undergoing a fundamental shift from trial-and-error experimentation to data-driven inverse design. In the broader landscape of materials science, advanced architectures such as graph neural networks (GNNs) and generative transformers have achieved breakthroughs in crystalline and ordered systems (Reiser et al., 2022; Przybek et al., 2025; Zhao et al., 2025). For instance, Merchant et al. (2023) recently demonstrated the power of deep learning by discovering 2.2 million new stable inorganic crystals via the graph networks for materials exploration (GNoME) model, radically expanding the known materials horizon. Similarly, in the domain of MOFs, machine learning pipelines have been successfully deployed to screen millions of candidates for gas adsorption and catalysis by leveraging their well-defined crystallographic information files (CIFs) (Song et al., 2025). Kolbadinejad et al. (2022) successfully predicted the adsorption amounts of argon, xenon, krypton, and oxygen on AC and zeolite using artificial neural networks (ANNs) (especially multilayer perceptrons), verifying the high efficiency of deep learning in gas adsorption simulation.

However, translating these successes to AC presents unique challenges due to its turbostratic, amorphous structure and the complex, nonlinear kinetics of biomass pyrolysis, which lack simple structure-property descriptors (Giudicianni et al., 2021; Belluati et al., 2024). Consequently, artificial intelligence (AI) applications in AC synthesis have relied primarily on classical regressors on tabular experimental data rather than structural graph generation (Zhang et al., 2020; Shukla et al., 2024). Researchers have employed algorithms such as gradient boosting decision tree (GBDT), random forest (RF), and ANN to map synthesis conditions to textural outputs (Li et al., 2024; Ibrahim and Hussein, 2025). For instance, Yuan et al. (2021) successfully employed GBDT to map biomass properties to CO_2 adsorption capacity, achieving high predictive accuracy (R^2 approx 0.84 on test data). Similarly, recent work by Ibrahim and Hussein (2025) used ensemble learning to optimize texture properties.

Despite these initial successes, significant barriers prevent the widespread adoption of AI in robust AC design. First, regarding applicability and generalization, current predictive frameworks are heavily constrained by the heterogeneity of biomass precursors and the small size of experimental datasets. Most existing models, such as that of Liao et al. (2019), were often constrained by small-sample datasets (e.g., <200 points), limiting their generalization across diverse precursors and activation agents. Consequently, these local models suffer from poor transferability. A regressor optimized for KOH-activated coconut shell often fails catastrophically when applied to H_3PO_4-activated lignocellulosic residues due to fundamentally different reaction kinetics and functional group evolutions.

Second, and perhaps more critically, a fundamental methodological gap exists in how target properties are modeled. The prevailing paradigm, as seen in most current studies, relies on single-task learning strategies: separate, independent models are trained to predict S_{BET}, V_T, and M_{yield} individually (Yuan et al., 2021). This decoupled approach ignores the intrinsic physicochemical correlations and competitive trade-offs governing carbonization—most notably, the antagonistic relationship where increasing activation severity enhances porosity and S_{BET} but inevitably degrades the carbon skeleton (M_{yield}). By treating these outputs as isolated targets, independent models often generate thermodynamically inconsistent predictions—such as forecasting

ultrahigh surface areas cooccurring with high yields—which are physically unrealizable (Evans et al., 1999; Lu et al., 2017). Furthermore, standard models often function as black boxes that lack the capability to simultaneously optimize these conflicting objectives within a single differentiable framework. There is a scarcity of advanced multitask learning (MTL) or knowledge distillation (KD) architectures in this domain that can reconcile these trade-offs while compensating for data sparsity.

Beyond prediction accuracy, biomass-based AC synthesis data pose a distinct learning challenge: multiple activation mechanisms (agent identity and one-step vs. two-step routes) are mixed within a small tabular dataset, producing multimodal mappings, while S_{BET}, V_T, and M_{yield} are physically coupled and exhibit different noise/sensitivity levels. Existing MTL or KD pipelines for tabular material data often treat distillation as a generic regularizer, without explicitly addressing mechanism mixing or the need to translate multiobjective predictions into actionable operating guidance. This study addresses this gap by combining two complementary forms of teacher information—(1) prediction-level soft priors to stabilize multitarget learning under heterogeneous labels and (2) teacher-aware proxy features to help disambiguate mixed mechanisms—together with robust Pareto screening and quantile-window extraction for decision-support window identification.

To address these limitations, AC-ResKD (activated-carbon residual knowledge distillation) is proposed, a mechanism-aware multitask distillation framework that operationalizes the following design: prediction-level knowledge distillation (PD-KD) and teacher-aware proxy augmentation for mechanism-mixed synthesis data, coupled with Pareto-based operating-window extraction for decision support. The approach centers on a shared residual deep neural network (ResDNN) backbone and integrates dual-channel priors: PD-KD and teacher-aware knowledge distillation (TA-KD). Process factors, including activating agent, activation temperature/time, impregnation conditions, heating rate, and precursor composition, are embedded into an interpretable input–output mapping with multihead outputs for S_{BET}, V_T, and M_{yield}. Under unified training/evaluation across combined and separated one-step/two-step datasets, the models reproduce domain-consistent monotonicity and surpass the typical benchmarks observed in the literature (Zou et al., 2024). Beyond prediction, robust Pareto screening and quantile-window extraction are implemented to provide agent- and route-specific optimal operating intervals. These windows identify conditions under which S_{BET} and V_T can be jointly increased while controlling M_{yield} penalties, furnishing actionable guidance for process scale-up. In summary, AC-ResKD delivers a more accurate and comprehensive AI-based synthesis assistant capable of reconciling multitarget trade-offs and predicting optimal operating intervals under varied design requirements.

2 Method

2.1 Feature representation and targets

The dataset in this article is cited from Zou et al. (2024), which collects synthesis conditions and textural properties of ACs produced with multiple chemical activating agents. Three datasets consisted of three parts: the data of ACs by one-step activation; the data of ACs by two-step activation; the combined data from both one-step activation and two-step activation. Each record corresponds to one synthesis experiment and contains.

(1) A categorical variable indicating the activating agent (e.g., KOH, H_3PO_4, and $ZnCl_2$), (2) several continuous process variables (such as precursor composition, soaking time and temperature, activation temperature (Act. T) and time, and heating rate), and (3) three regression targets: S_{BET} (m^2/g), V_T (cm^3/g), and M_{yield} (%).

For learning, a mixed feature vector $\boldsymbol{x}$ is constructed that combines the categorical agent label and the continuous synthesis variables. The activating agent is one-hot encoded into a fixed-length binary vector. All remaining continuous variables are scaled to [0, 1] using a min–max scaling fitted on the training portion only to avoid target leakage. The final input vector is

$$\boldsymbol{x} = [\boldsymbol{x}_{agent}, \boldsymbol{x}_{proc}] \in \mathbb{R}^F, \tag{1}$$

where F denotes the total number of input features after one-hot expansion, $\boldsymbol{x}_{agent}$ is the one-hot activating-agent vector, and $\boldsymbol{x}_{proc}$ is the scaled continuous process-variable vector. The three prediction targets are denoted by

$$y_1 = S_{BET},\ y_2 = V_T,\ y_3 = M_{yield}. \tag{2}$$

For numerical stability, each target is also individually rescaled to [0, 1] on the training set before

optimization. All performance metrics reported in this work, including the coefficient of determination (R^2), root mean square error (E_{RMS}), and Pearson correlation coefficient (r), are computed after inverting this scaling back to the original physical units so that they are directly interpretable by materials scientists.

Although several AC datasets with similar variables exist in many studies (Zou et al., 2024; Ibrahim and Hussein, 2025), this work deliberately adopts this dataset as a primary test bed for two reasons. First, it is relatively challenging: the synthesis conditions are highly heterogeneous across activating agents, and the experimental points are sparsely and irregularly distributed in the process-parameter space rather than forming dense grids (machine learning application for predicting key properties of AC produced from lignocellulosic biomass waste with chemical activation). In practice, the same activating agent is often explored at only a few discrete operating points, which makes it difficult for purely data-driven models to interpolate smoothly and capture global trends. Second, the dataset is provided in a clean tabular format with well-defined physical variables, which makes it a representative benchmark for evaluating tabular deep-learning architectures under realistic data scarcity. To address robustness concerns beyond a single curated source, we further report an external evaluation on the independent dataset (Ibrahim and Hussein, 2025), which differs in variable schema and targets and therefore provides a complementary generalization check.

In addition to the original combined dataset, which merges one-step and two-step activation protocols, this work also evaluates the models on two separate subsets: a one-step dataset and a two-step dataset containing only experiments with the corresponding processing route. As shown later in Section 3.1, the combined dataset leads to systematically lower performance than training separate models on one-step and two-step data. This suggests that pooling distinct synthesis routes into a single model introduces conflicting patterns and reduces the effective smoothness of the input–output mapping. In contrast, the proposed architecture remains reasonably accurate on this difficult combined setting and performs even better on the cleaner one-step and two-step subsets, indicating that the method can be transferred to other AC datasets by only adapting the input features and the number of output heads.

Finally, the feature-importance and partial-dependence analyses in Section 3.2 reveal physically meaningful trends: the model recovers the expected influence of activation temperature, precursor composition (VM, ash, FC), soaking conditions, and agent type on S_{BET}, V_{T}, and M_{yield}. This consistency with domain knowledge supports that the proposed network not only is predictive but also captures plausible latent structures in the synthesis–property relationship.

2.2 Model design and structure

The model structure is shown in Fig. 1, consisting of a teacher model and a student model, which are distilled through different methods. The specific method will be explained in the following sections.

2.2.1 Teacher model library and per-target selection

As a strong nonneural reference, a heterogeneous teacher library is constructed on top of the same feature representation described in Section 2.1. The candidate teachers include tree-based, kernel-based, instance-based, neural, and simple ensemble regressors so that different inductive biases can be explored for different property targets. Rather than assuming that a single teacher family is uniformly optimal for all outputs, this work selects the teacher separately for each target. Detailed definitions of the candidate teacher models and their ensemble forms are provided in Section S1 and Eqs. (S1)–(S7) of the electronic supplementary materials (ESM). For each outer train/test split and for each target y_k, model selection is performed only on the training set. Let $\mathcal{M}_k$ denote the set of all candidate teacher models, including both individual regressors and simple equal-weight. For any denoted candidate model $m \in \mathcal{M}_k$, its validation coefficient of determination is computed as

$$R_k^2(m) = 1 - \frac{\sum_{i \in \mathcal{V}} \left(y_{k,i} - \hat{y}_{k,i}^{(m)}\right)^2}{\sum_{i \in \mathcal{V}} \left(y_{k,i} - \dot{y}_k\right)^2}, \tag{3}$$

where $\mathcal{V}$ is the validation subset within the training portion, $y_{k,i}$ is the observed value of target t for validation sample i, $\hat{y}_{k,i}^{(m)}$ is the corresponding prediction from candidate teacher m, and $\dot{y}_k$ is the mean of $y_{k,i}$ over $\mathcal{V}$. A cross-validated estimation of $R_k^2(m)$ is then used to determine the final teacher for each target:

$$m_k^* = \underset{m \in \mathcal{M}_k}{\operatorname{argmax}} R_k^2(m). \tag{4}$$

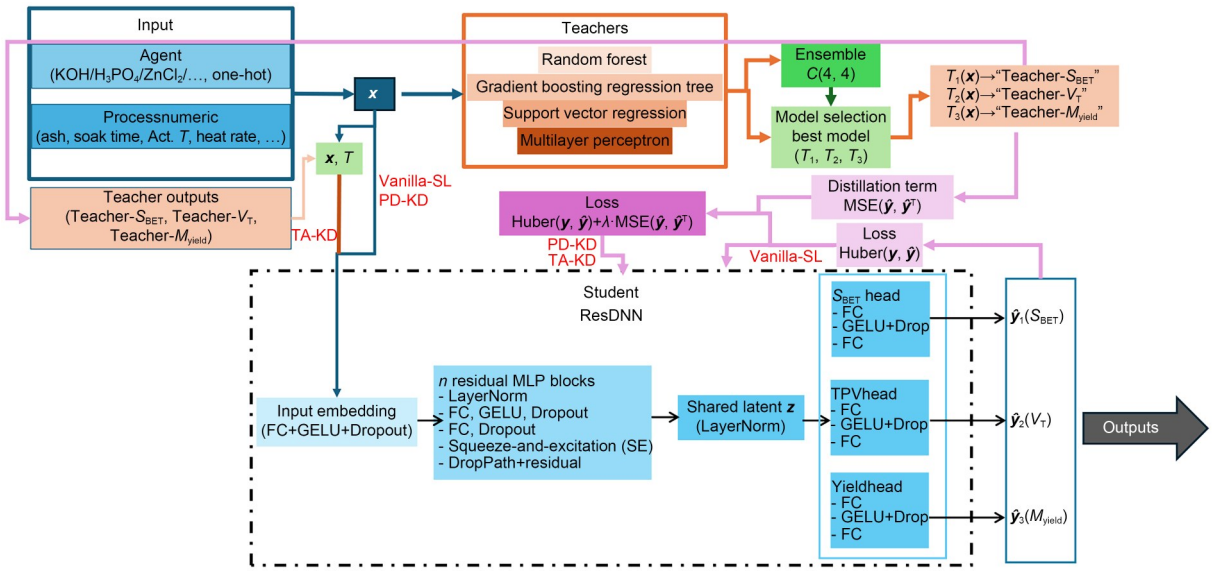

Fig. 1 Model structure. x denotes the input feature vector, z denotes the shared latent representation, and $\hat{y}$ denotes the predicted target vector. DropPath denotes stochastic depth, and $C(4, 4)$ denotes the equal-weight ensemble combination of four candidate teacher models. MSE denotes mean square error

This procedure yields three potentially different best teachers, one for S_{BET}, one for V_T, and one for M_{yield}, thereby allowing the teacher assignment to adapt to target-specific response characteristics rather than enforcing a single shared teacher across all outputs.

Once selected, each teacher is refitted on the full training subset of the corresponding split and then frozen. The resulting teacher predictions are generated for both training and test samples and subsequently used in two ways in the student framework described below: (1) as prediction-level soft supervision in PD-KD and (2) as high-level proxy features in TA-KD. Because teacher selection, refitting, and student training are all restricted to the same training portion in each split, the comparison among Vanilla-SL, PD-KD, and TA-KD remains strictly leakage-free and directly comparable under identical data partitions. Here, Vanilla-SL denotes Vanilla supervised learning.

2.2.2 Student architecture (ResDNN multitask regressor)

The student model is a shared-backbone, multi-head regressor operating on the same feature vector as the teacher models. A residual multilayer perceptron (MLP), referred to here as a ResDNN, is adopt for tabular data because S_{BET}, V_T, and M_{yield} are driven by the same synthesis variables and therefore benefit from hard parameter sharing (Caruana, 1997; Ruder, 2017; Gorishniy et al., 2021; Kadra et al., 2021). Given an input vector $\boldsymbol{x}$, the network first applies a linear stem followed by Gaussian error linear unit (GELU) activation and dropout:

$$\boldsymbol{h}^{(0)}=\mathrm{Dropout}\left(\mathrm{GELU}\left(\boldsymbol{W}_{\mathrm{stem}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{stem}}\right)\right), \tag{5}$$

where $\boldsymbol{h}^{(0)}$ is the initial hidden representation, $\boldsymbol{W}_{\mathrm{stem}}$ and $\boldsymbol{b}_{\mathrm{stem}}$ are the weight matrix and bias vector of the stem layer, GELU denotes the Gaussian error linear unit activation, and Dropout(·) denotes dropout regularization.

This is followed by preactivation residual blocks:

$$\begin{aligned}&\boldsymbol{u}^{(l)}=\mathrm{LayerNorm}\left(\boldsymbol{h}^{(l-1)}\right),\\&\boldsymbol{v}^{(l)}=\\&\mathrm{Dropout}\left(\phi\left(\boldsymbol{W}_2^{(l)}\,\mathrm{Dropout}\left(\phi\left(\boldsymbol{W}_1^{(l)}\boldsymbol{u}^{(l)}+\boldsymbol{b}_1^{(l)}\right)\right)+\boldsymbol{b}_2^{(l)}\right)\right),\\&\tilde{\boldsymbol{v}}^{(l)}=\mathrm{SE}\left(\boldsymbol{v}^{(l)}\right),\ \boldsymbol{h}^{(l)}=\boldsymbol{h}^{(l-1)}+\alpha_l\,\mathrm{DropPath}\left(\tilde{\boldsymbol{v}}^{(l)}\right),\end{aligned} \tag{6}$$

where $l=1, 2, \ldots, L$ indexes the lth residual block, $\boldsymbol{u}^{(l)}$ and $\boldsymbol{v}^{(l)}$ are intermediate hidden vectors in block l, $\boldsymbol{W}_1^{(l)}$, $\boldsymbol{W}_2^{(l)}$, $\boldsymbol{b}_1^{(l)}$, and $\boldsymbol{b}_2^{(l)}$ are block parameters, $\phi(\cdot)$ denotes the GELU activation, LayerNorm(·) denotes layer normalization, SE(·) denotes squeeze-and-excitation feature reweighting, DropPath(·) denotes stochastic depth, and α_l is the residual scaling coefficient. After the last block, a shared latent representation is obtained:

$$\boldsymbol{z}=\mathrm{LayerNorm}\left(\boldsymbol{h}^{(L)}\right)\in\mathbb{R}^{d_{\mathrm{model}}}. \tag{7}$$

Here, L is the total number of residual blocks, and d_{model} is the dimension of the shared latent representation $\boldsymbol{z}$.

Three task-specific regression heads are then applied to predict the scaled targets for S_{BET}, V_{T}, and M_{yield}:

$$\hat{y}_k = f_k(\boldsymbol{z}) = \boldsymbol{w}_{k,2}^{\text{T}}\,\phi(\boldsymbol{W}_{k,1}\boldsymbol{z} + \boldsymbol{b}_{k,1}) + b_{k,2},\ k = 1,\ 2,\ 3, \quad (8)$$

where $f_k(\cdot)$ is the scalar regression head for target k, $\boldsymbol{W}_{k,1}$ and $\boldsymbol{b}_{k,1}$ are the weight matrix and bias vector of the hidden layer in head k, $\boldsymbol{w}_{k,2}$ and $b_{k,2}$ are the output weight vector and scalar bias, and $\hat{y}_k$ is the scaled scalar prediction.

All heads are optimized jointly through a summed regression loss. Regularization is implemented using dropout and stochastic depth, and hyperparameters are selected on the training set only.

2.2.3 KD variants

Three student-training paradigms are compared using the same ResDNN backbone. Let $\hat{y}_k(\boldsymbol{x})$ and $\hat{y}_k^{(\text{Tch})}(\boldsymbol{x})$ denote the student and teacher predictions (Tch) for target y_k, respectively. All losses are computed on individually scaled targets using the Huber loss. The student is trained only on the ground-truth labels:

(1) Vanilla-SL

$$L_{\text{Vanilla}} = \sum_{k=1}^{3} w_k\, l_\delta(\hat{y}_k(\boldsymbol{x}),\ y_k), \quad (9)$$

where w_k is the task weight for target k, $l_\delta(\cdot,\cdot)$ is the Huber loss with threshold δ, and $k = 1, 2, 3$ correspond to S_{BET}, V_{T}, and M_{yield}, respectively. This serves as the deep learning baseline.

(2) PD-KD

A distillation term is added to encourage agreement with the selected teacher:

$$L_{\text{PD-KD}} = \sum_{k=1}^{3} w_k \Big[l_\delta(\hat{y}_k(\boldsymbol{x}),\ y_k) + \lambda_{\text{KD}} \cdot \big\| \hat{y}_k(\boldsymbol{x}) - \hat{y}_k^{(\text{Tch})}(\boldsymbol{x}) \big\|_2^2 \Big], \quad (10)$$

where λ_{KD} is a scalar trade-off coefficient.

(3) TA-KD

The teacher predictions are further concatenated with the original input as auxiliary features:

$$\tilde{\boldsymbol{x}} = [\,\boldsymbol{x},\ \hat{y}_1^{(\text{Tch})},\ \hat{y}_2^{(\text{Tch})},\ \hat{y}_3^{(\text{Tch})}\,], \quad (11)$$

and the student is optimized with the same supervised-distillation objective:

$$L_{\text{TA-KD}} = \sum_{k=1}^{3} w_k \Big[l_\delta(\hat{y}_k(\tilde{\boldsymbol{x}}),\ y_k) + \lambda_{\text{KD}} \cdot \big\| \hat{y}_k(\tilde{\boldsymbol{x}}) - \hat{y}_k^{(\text{Tch})}(\boldsymbol{x}) \big\|_2^2 \Big]. \quad (12)$$

In all experiments, the distillation weight is fixed to $\lambda_{\text{KD}} = 0.5$ for PD-KD and $\lambda_{\text{KD}} = 0.7$ for TA-KD. These values are chosen empirically to provide a balanced teacher signal without overwhelming the supervised loss and are kept unchanged across all targets, datasets, and repeated splits for consistency.

2.2.4 Training and evaluation protocol

For each dataset variant (combined, one-step, and two-step), model performance is assessed using 20 repeated random 80/20 train/test splits. In each repeat, 80% of the samples are used for model development, and the remaining 20% are held out for final evaluation. The same set of repeated splits was reused across all learning paradigms (Vanilla-SL, PD-KD, TA-KD, and Teacher) to enable paired and directly comparable performance assessment. All reported R^2, E_{RMS}, and r values arecomputed on the held-out test subsets after inverse transformation to the original physical units and are summarized as the mean values with 95% confidence intervals (CIs) over the 20 repeats.

Within each training portion, all preprocessing and model-selection steps are performed without access to the test subset. Specifically, categorical encoding and min–max scaling are fitted using the training data only, and each target is scaled individually for optimization before being mapped back to physical units for evaluation. The teacher model for each target is selected only from the training portion, as described in Section 2.2.1, and then is refitted on the full training subset of the corresponding split. A fresh ResDNN student is subsequently initialized for each paradigm and trained under the same optimization protocol, differing only in whether teacher information is used in the loss term (PD-KD) and/or as auxiliary input features (TA-KD). This design ensures that performance differences among the student variants can be attributed to the use of teacher information rather than to differences in data partitioning, preprocessing, or training budget.

To examine robustness beyond the primary dataset, the same evaluation pipeline is also applied to an

independent AC dataset with different descriptors and targets. The corresponding benchmark results are reported in Section S2 and Table S1 of the ESM.

2.3 Robust Pareto screening and optimal-interval extraction

To translate the trained surrogate models into actionable synthesis guidance, a robust Pareto screening procedure followed by quantile-window extraction is performed for the most frequent activating agents. For each dataset and agent, a feasible candidate domain is constructed by restricting each continuous process variable to the empirical 2nd–98th percentile range observed in the training data while fixing categorical variables to the target agent (and dominant category where applicable). Monte Carlo candidate points are then generated:

$$\mathcal{X}_a = \{\boldsymbol{x}_a^{(i)}\}_{i=1}^{N}, \tag{13}$$

and transformed using the same preprocessing pipeline as in model training.

Each candidate is evaluated by two student models, Vanilla-SL and PD-KD, yielding predictions for S_{BET}, V_{T}, and M_{yield}:

$$M_k: \tilde{\boldsymbol{x}}_a^{(i)} \mapsto (\hat{y}_{k,1}^{(i)}, \hat{y}_{k,2}^{(i)}, \hat{y}_{k,3}^{(i)}) = (\widehat{S}_{\mathrm{BET}}, \widehat{V}_{\mathrm{T}}, \widehat{M}_{\mathrm{yield}})_k^{(i)}. \tag{14}$$

Here, M_k denotes the kth student surrogate model, with $k=1$ and $k=2$ corresponding to Vanilla-SL and PD-KD, respectively. The three components $\hat{y}_{k,1}^{(i)}$, $\hat{y}_{k,2}^{(i)}$, and $\hat{y}_{k,3}^{(i)}$ correspond to predicted S_{BET}, V_{T}, and M_{yield}, respectively. The hat symbol denotes a model prediction.

These outputs are combined into a six-dimensional objective vector,

$$\boldsymbol{f}(\boldsymbol{x}_a^{(i)}) = (\hat{y}_{1,1}^{(i)}, \hat{y}_{1,2}^{(i)}, \hat{y}_{1,3}^{(i)}, \hat{y}_{2,1}^{(i)}, \hat{y}_{2,2}^{(i)}, \hat{y}_{2,3}^{(i)})^{\mathrm{T}} \in \mathbb{R}^6, \tag{15}$$

with all objectives treated as larger-is-better. A candidate is considered dominant if it is no worse in all objectives and strictly better in at least one:

$$\boldsymbol{f}(\boldsymbol{x}_a) \succ \boldsymbol{f}(\boldsymbol{z}_a) \Leftrightarrow f_m(\boldsymbol{x}_a) \geqslant f_m(\boldsymbol{z}_a)\ \forall m \text{ and } f_m(\boldsymbol{x}_a) > f_m(\boldsymbol{z}_a) \text{ for at least one } m. \tag{16}$$

The robust Pareto set for each agent is then defined as the nondominated subset:

$$\mathcal{P}_a^{\mathrm{robust}} = \{\boldsymbol{x}_a^{(i)} \in \mathcal{X}_a \mid \nexists\, \boldsymbol{z}_a \in \mathcal{X}_a \text{ s.t. } \boldsymbol{f}(\boldsymbol{z}_a) \succ \boldsymbol{f}(\boldsymbol{x}_a^{(i)})\}. \tag{17}$$

When necessary, quasi-Pareto fallback and crowding-distance subsampling are used to maintain numerical stability and diversity.

To obtain practically usable operating ranges, the Pareto set is summarized by narrow quantile windows. For each continuous variable, the 45th and 55th percentiles over the Pareto set are extracted:

$$[L_j^*(a),\ U_j^*(a)] = \left[q_{45\%}(x_j \mid \boldsymbol{x} \in \mathcal{P}_a^{\mathrm{robust}}),\ q_{55\%}(x_j \mid \boldsymbol{x} \in \mathcal{P}_a^{\mathrm{robust}})\right], \tag{18}$$

where j indexes the continuous process variable, $L_j^*(a)$ and $U_j^*(a)$ are the lower and upper bounds of the extracted operating window for agent a, and $q_{45\%}$ and $q_{55\%}$ denote the 45th and 55th percentiles, respectively.

These quantile intervals define Pareto-consistent operating windows in which S_{BET}, V_{T}, and M_{yield} remain jointly favorable and predictions from different student models are mutually consistent. The resulting ranges are reported in Section 3.3 as agent- and route-specific best-interval recommendations.

3 Results and discussion

3.1 Comparison of model results

This study investigates the chemical activation of biomass-derived ACs and examines how key process variables—activating agent type, activation temperature and duration, impregnation conditions, heating rate, and precursor composition (VM; ash; FC)—jointly affect S_{BET}, V_{T}, and M_{yield}. A multitask ResDNN together with KD variants (PD-KD and TA-KD) is adopted, and findings are cross-validated against mechanistic expectations and model behavior. Across dataset configurations (combined one-/two-step activation; separated one-step and two-step routes), the model exhibits consistent alignment with expected physical trends, captures the directionality and strength of variable-property relationships, and reflects interpretable trade-offs among the three targets.

Relative to the dataset-agnostic reference study of Zou et al. (2024) (S_{BET}: R^2=0.78, E_{RMS}=293.76; V_{T}: R^2=0.76, E_{RMS}=0.19; M_{yield}: R^2=0.91, E_{RMS}=4.49), our repeated-split evaluation shows that route-separated

modeling achieves competitive—and for S_{BET}/V_T often higher—accuracy in original physical units. For example, as shown in Table 1 one-step TA-KD attains S_{BET} R^2= 0.806 [0.787, 0.824] and V_T R^2=0.787 [0.764, 0.810], while two-step TA-KD attains S_{BET} R^2=0.801 [0.773, 0.830] and V_T R^2=0.817 [0.791, 0.843], with V_T E_{RMS}= 0.130 [0.119, 0.141] below the reference E_{RMS} of 0.19. The combined dataset remains more challenging due to stronger mechanism heterogeneity; nevertheless, distillation still yields substantial improvements (e.g., combined TA-KD: S_{BET} R^2=0.768 [0.751, 0.784]; V_T R^2= 0.739 [0.720, 0.757]). For M_{yield}, our student models do not reach the R^2=0.91 reported in Zou et al. (2024) and exhibit higher E_{RMS} (e.g., combined TA-KD E_{RMS}=5.047 [4.853, 5.242]), highlighting the sensitivity of M_{yield} to heterogeneous reporting and the fact that cross-study comparisons can be affected by differences in preprocessing choices (e.g., target scaling/inverse-scaling, split protocol, and any filtering rules) that are not fully specified in the reference. Accordingly, we treat Zou et al., (2024) as a useful reference point rather than a strictly controlled baseline, and we report all results under a fully specified and repeatable protocol (20 repeated splits with mean and 95% CIs in original units). M_{yield} is the most consistently improved target under distillation. In the combined dataset, TA-KD increases R^2 from 0.577 [0.538, 0.616] (Vanilla-SL) to 0.816 [0.802, 0.831] and reduces E_{RMS} from 7.646 [7.325, 7.968] to 5.047 [4.853, 5.242], delivering the strongest student performance. In the one-step dataset, TA-KD is the best-performing student for M_{yield} (R^2=0.825 [0.807, 0.843]; E_{RMS}=4.815 [4.511, 5.119]) and approaches the teacher baseline (R^2=0.835 [0.813, 0.856]). In the two-step dataset, TA-KD again leads among students for M_{yield} (R^2=0.791 [0.770, 0.811]), while the teacher remains substantially stronger (R^2=0.861 [0.844, 0.878]), indicating that M_{yield} is particularly sensitive to route-specific reporting heterogeneity and benefits from stronger inductive bias. For V_T, distillation improves performance across all settings, with the most reliable gains under route-separated training: one-step TA-KD reaches R^2= 0.787 [0.764, 0.810] (E_{RMS}=0.159 [0.151, 0.167]), two-step TA-KD reaches R^2=0.817 [0.791, 0.843] (E_{RMS}= 0.130 [0.119, 0.141]), and the combined dataset remains more challenging (best student V_T R^2=0.739 [0.720, 0.757]). For S_{BET}, TA-KD provides the best or near-best student performance across dataset variants (combined R^2=0.768 [0.751, 0.784]; one-step R^2=0.806 [0.787, 0.824]; two-step R^2=0.801 [0.773, 0.830]). S_{BET} is more sensitive to microstructural detail and extreme values; compared with PD-KD, explicit conditioning in TA-KD is more robust under combined heterogeneity.

The influence of route differences (one-step vs. two-step) and data heterogeneity is corroborated by aggregated metrics and calibration behavior under repeated-split evaluation. In homogeneous routes (one-step and two-step), input–output mappings are nearly monotonic and better calibrated, and distillation consistently enhances V_T and M_{yield}. The combined dataset exhibits stronger heterogeneity; TA-KD provides the largest gains in this setting, consistent with the need for teacher-conditioned signals to mitigate multimechanism mismatches. Evidence for a precarbonization priming effect is most clearly reflected in V_T: the two-step route achieves higher V_T predictability (TA-KD R^2= 0.817 [0.791, 0.843]) and lower E_{RMS} of V_T than the one-step route, consistent with pyrolysis presetting the carbon skeleton and pore-size spectrum. The S_{BET} performance is comparable between the one-step and two-step methods within the CIs, suggesting that the priming effect manifests more strongly in pore volume formation than in surface area extremes.

To assess cross-source robustness, the framework is further tested on an independent AC dataset with different descriptors and targets. The distillation variants preserve the same performance ordering, and TA-KD remain the strongest student model (Table S1 of the ESM).

Fig. 2 presents parity plots for the combined, one-step, and two-step datasets. Each point represents the mean predicted value and the mean observed value within an equal-frequency bin of predicted scores. The in-panel metrics are calculated only for the plotted representative split, and therefore are intended as visual diagnostics of the displayed parity plots rather than as the final quantitative evaluation. The repeated-split performance statistics used for model comparison are reported in Table 1 as mean values and 95% confidence intervals. Proximity to the dashed line ($y=x$) indicates minimal bias and a more faithful monotonic mapping within that prediction interval. Systematic overestimation or underestimation appears as entire segments lying above or below the diagonal, respectively. Slope contraction and fan-shaped dispersion at the extremes signify regression-to-the-mean and heteroscedasticity.

Table 1 Model results of different datasets

Dataset	Method	Target	R^2 (mean [95% CIs])	r (mean [95% CIs])	E_{RMS} (mean [95% CIs])
Combined	Vanilla-SL	S_{BET} (m^2/g)	0.708 [0.679, 0.736]	0.847 [0.831, 0.864]	346.684 [328.088, 365.279]
		V_T (cm^3/g)	0.658 [0.624, 0.693]	0.816 [0.796, 0.837]	0.203 [0.192, 0.213]
		M_{yield} (%)	0.577 [0.538, 0.616]	0.771 [0.747, 0.794]	7.646 [7.325, 7.968]
	PD-KD	S_{BET} (m^2/g)	0.751 [0.728, 0.775]	0.871 [0.858, 0.884]	319.779 [303.708, 335.851]
		V_T (cm^3/g)	0.716 [0.690, 0.742]	0.853 [0.839, 0.866]	0.185 [0.176, 0.194]
		M_{yield} (%)	0.744 [0.719, 0.770]	0.869 [0.855, 0.882]	5.937 [5.651, 6.223]
	TA-KD	S_{BET} (m^2/g)	0.768 [0.751, 0.784]	0.882 [0.873, 0.890]	309.643 [299.841, 319.446]
		V_T (cm^3/g)	0.739 [0.720, 0.757]	0.866 [0.855, 0.877]	0.177 [0.170, 0.185]
		M_{yield} (%)	0.816 [0.802, 0.831]	0.909 [0.899, 0.918]	5.047 [4.853, 5.242]
	Teacher	S_{BET} (m^2/g)	0.760 [0.746, 0.774]	0.880 [0.872, 0.889]	315.384 [304.881, 325.887]
		V_T (cm^3/g)	0.725 [0.710, 0.741]	0.859 [0.849, 0.870]	0.183 [0.174, 0.191]
		M_{yield} (%)	0.809 [0.796, 0.821]	0.910 [0.901, 0.919]	5.164 [4.953, 5.375]
One-step	Vanilla-SL	S_{BET} (m^2/g)	0.758 [0.739, 0.776]	0.878 [0.867, 0.889]	322.039 [307.767, 336.312]
		V_T (cm^3/g)	0.737 [0.717, 0.757]	0.869 [0.857, 0.881]	0.177 [0.170, 0.185]
		M_{yield} (%)	0.705 [0.672, 0.737]	0.850 [0.832, 0.869]	6.263 [5.791, 6.735]
	PD-KD	S_{BET} (m^2/g)	0.796 [0.778, 0.814]	0.900 [0.889, 0.910]	294.821 [280.225, 309.417]
		V_T (cm^3/g)	0.785 [0.760, 0.810]	0.896 [0.882, 0.910]	0.159 [0.152, 0.167]
		M_{yield} (%)	0.782 [0.754, 0.810]	0.894 [0.879, 0.909]	5.376 [4.933, 5.819]
	TA-KD	S_{BET} (m^2/g)	0.806 [0.787, 0.824]	0.906 [0.896, 0.915]	286.991 [274.933, 299.050]
		V_T (cm^3/g)	0.787 [0.764, 0.810]	0.896 [0.884, 0.908]	0.159 [0.151, 0.167]
		M_{yield} (%)	0.825 [0.807, 0.843]	0.918 [0.908, 0.928]	4.815 [4.511, 5.119]
	Teacher	S_{BET} (m^2/g)	0.805 [0.790, 0.820]	0.903 [0.894, 0.911]	288.580 [276.628, 300.533]
		V_T (cm^3/g)	0.786 [0.769, 0.802]	0.893 [0.883, 0.903]	0.160 [0.153, 0.168]
		M_{yield} (%)	0.835 [0.813, 0.856]	0.923 [0.912, 0.934]	4.662 [4.313, 5.012]
Two-step	Vanilla-SL	S_{BET} (m^2/g)	0.746 [0.702, 0.791]	0.872 [0.847, 0.897]	268.615 [240.862, 296.367]
		V_T (cm^3/g)	0.782 [0.733, 0.830]	0.894 [0.869, 0.918]	0.138 [0.124, 0.153]
		M_{yield} (%)	0.734 [0.708, 0.760]	0.882 [0.863, 0.901]	6.465 [5.863, 7.066]
	PD-KD	S_{BET} (m^2/g)	0.781 [0.749, 0.812]	0.893 [0.874, 0.911]	250.140 [230.457, 269.824]
		V_T (cm^3/g)	0.790 [0.735, 0.845]	0.902 [0.878, 0.926]	0.135 [0.120, 0.150]
		M_{yield} (%)	0.754 [0.730, 0.778]	0.889 [0.870, 0.909]	6.215 [5.676, 6.755]
	TA-KD	S_{BET} (m^2/g)	0.801 [0.773, 0.830]	0.905 [0.890, 0.920]	238.235 [218.637, 257.832]
		V_T (cm^3/g)	0.817 [0.791, 0.843]	0.915 [0.901, 0.930]	0.130 [0.119, 0.141]
		M_{yield} (%)	0.791 [0.770, 0.811]	0.911 [0.895, 0.927]	5.766 [5.179, 6.353]
	Teacher	S_{BET} (m^2/g)	0.785 [0.763, 0.807]	0.899 [0.887, 0.911]	249.900 [233.873, 265.927]
		V_T (cm^3/g)	0.805 [0.779, 0.832]	0.910 [0.894, 0.926]	0.135 [0.121, 0.148]
		M_{yield} (%)	0.861 [0.844, 0.878]	0.936 [0.927, 0.945]	4.643 [4.215, 5.071]

The parity plots indicate that the one-step and two-step curves adhere more closely to the ideal diagonal, reflecting that a more homogeneous kinetic context facilitates learning near-linear, monotonic input–output mappings. Distillation yields clear gains for V_T and M_{yield}, with target-dependent benefits. In contrast, the combined dataset exhibits widespread slope contraction and high-end compression, attributable to the nonequivalence of route semantics, agent–operating-window covariation, unmodeled interactions between activation severity and composition, and compounded measurement heterogeneity. TA-KD substantially corrects bias for S_{BET} and M_{yield}, whereas PD-KD performs the best for V_T. Full mitigation of these structural challenges likely requires

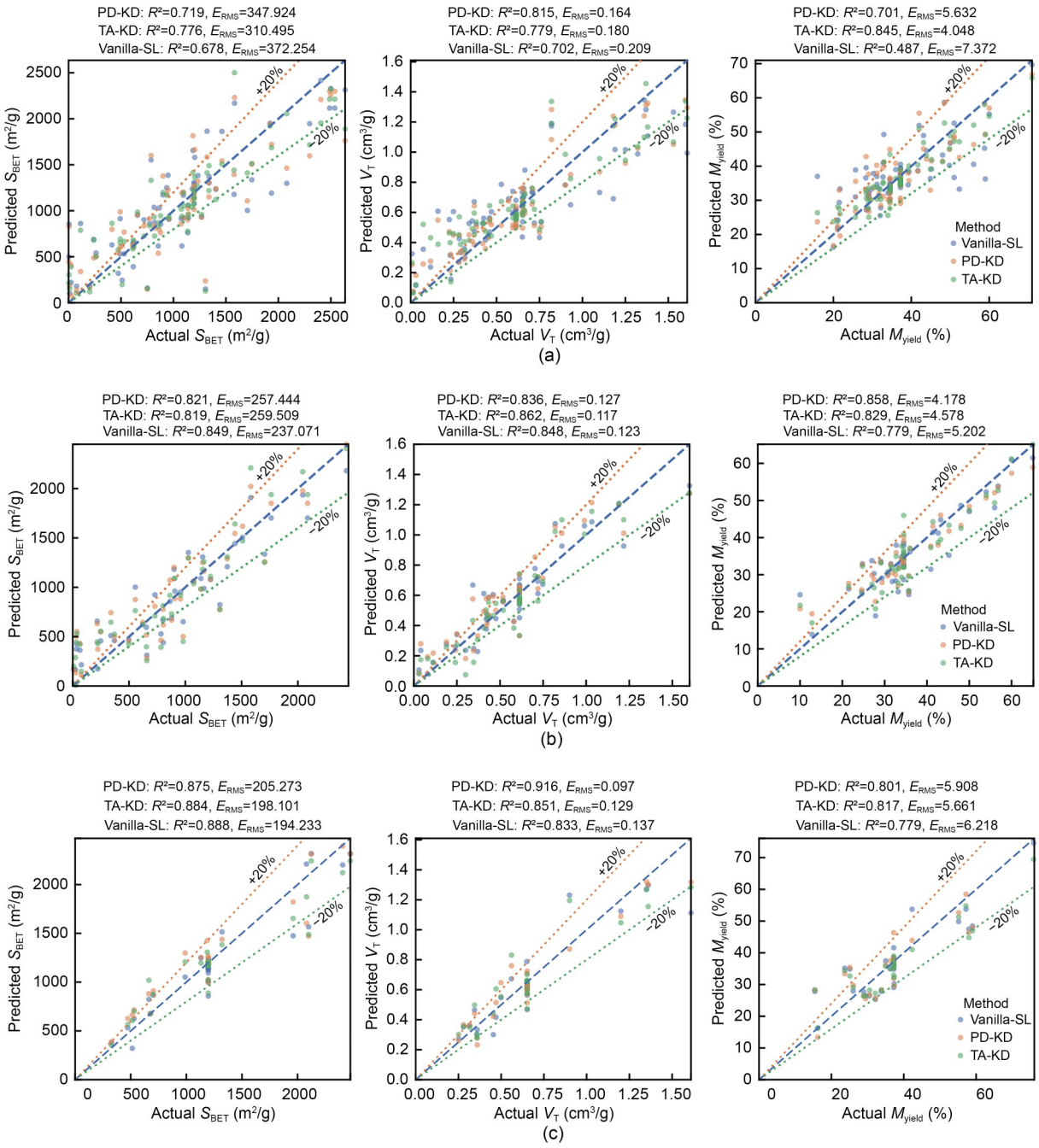

Fig. 2 Parity plots for the three dataset configurations: (a) combined dataset; (b) one-step dataset; (c) two-step dataset. The plots visualize one representative split for qualitative comparison. The in-panel metrics are calculated for the displayed split only and are used as visual diagnostics. Aggregated performance over 20 repeated splits, reported as mean values and 95% CIs in original physical units, is given in Table 1 and used for all quantitative claims

mechanistic severity features and explicit interaction terms on the feature side together with conditional or routed architectures on the model side. Following these findings, target- and route-specific distillation and post hoc calibration can render the three targets closer to $y=x$ across the full range, yield more uniform residuals, and improve generalization under extreme operating conditions and across activating agents.

Coupling and process trade-offs among the three targets follow expected physicochemical patterns. The S_{BET} and V_T are generally positively correlated but show a characteristic rise–peak–saturation or decline. The

M_{yield} is negatively correlated with both, reflecting the cost of deeper pore development. The updated analyses confirm that a higher activation temperature or longer duration, as well as more thorough impregnation, increases S_{BET} and V_T but reduces M_{yield}. The position on this trade-off depends on the activating agent: KOH induces stronger etching (S_{BET}/V_T increase, and M_{yield} decreases), H_3PO_4 promotes mid-temperature crosslinking that better preserves M_{yield}, and $ZnCl_2$ favors mesoporous development. Multitask parameter sharing helps the model learn these endogenous trade-offs and explains the consistent advantage of Vanilla-SL for S_{BET} on the separated datasets. Distillation, through smoothing and conditioning, further reduces variance and bias for V_T and M_{yield}. Similar route- and target-dependent bias patterns are observed in the calibration analysis (Fig. S1 of the ESM), further supporting the superior monotonicity of route-separated training.

Consistent with the physicochemical mechanisms of biomass activation, the AI models reproduce domain-aligned trends for S_{BET}, V_T, and M_{yield}. The multitask residual student, conditioned on teacher information, captures nonlinearities and interactions robustly in heterogeneous, data-scarce tabular settings. Under a unified training and evaluation protocol, the performance metrics exceed the upper bounds reported by the reference study, with especially strong gains for V_T and M_{yield}. Distillation gains for S_{BET} are less stable on separated routes, motivating target- and route-specific distillation strengths and structural constraints. These results provide quantitative tools and physical grounding for process-parameter optimization and offer practical strategies for model transfer to broader datasets.

3.2 Model interpretability analysis

Fig. 3 illustrates the calibration of variable importance under the one-step and two-step modeling frameworks and presents a unified interpretation of how process parameters affect the predicted outputs. Two complementary explainability metrics are employed to assess feature importance: permutation feature importance (PFI), which quantifies global performance contribution by the drop in test R^2 (ΔR^2) upon shuffling a single feature, and global mean Shapley additive explanations (SHAP) values, which capture the magnitude and direction of local, sample-wise contributions. As comparative baselines, three student training paradigms are considered: a purely supervised multitask residual network (Vanilla-SL), PD-KD, and TA-KD. Analyses span both one-step and two-step activation datasets and are cross-validated against other results in the document (combined-set behavior, calibration, and parity plots) to corroborate variable ranking and physical consistency.

The importance profiles reveal consistent global regularities. Teacher outputs are the most influential features. Teacher-$S_{BET}/V_T/M_{yield}$ produces the largest ΔR^2 in PFI and ranks among the top contributors in SHAP, indicating that teacher predictions serve not merely as loss-level corrections but also as high-information priors actively used by the student models. Secondary dominant factors comprise the activating agent and the key thermal processing parameters. Activating agent type, activation temperature, and activation time consistently rank near the top across both datasets. In the two-step setting, the pyrolysis heating rate/temperature/duration also contribute substantially. Distillation modifies the balance rather than the ordering of feature importance. Relative to Vanilla-SL, PD-KD and TA-KD reduce overreliance on a single teacher feature and elevate the weights of process variables without altering the underlying physical conclusion regarding which variables are most important.

For the two-step dataset, S_{BET} exhibits PFI with Teacher-S_{BET} as an unequivocal top feature, followed by the activating agent, activation time, pyrolysis heating rate, and sample/formulation (agent/sample) as a second tier. SHAP indicates that local variation is predominantly driven by agent choice and time/heating-rate parameters, with Teacher-S_{BET} refining global bias. The S_{BET} is governed by a balance between micropore development and pore-wall stability. The pyrolysis stage preconditions the skeleton's etchability, thereby elevating the importance of pyrolysis variables in two-step activation. For V_T, Teacher-V_T ranks the first in both PFI and SHAP, followed by agent/sample and the activating agent, with activation temperature/time and pyrolysis rate/time forming the second tier. This pattern indicates that the total pore volume is strongly influenced by the agent chemistry and the time scales of impregnation/pyrolysis release. In two-step activation, precarbonization-induced meso-/macrostructural features (e.g., cracks) also contribute to V_T. For M_{yield}, Teacher-M_{yield} remains the first in PFI but is comparable to the activating agent in SHAP, highlighting activation temperature, pyrolysis rate, and sample/formulation

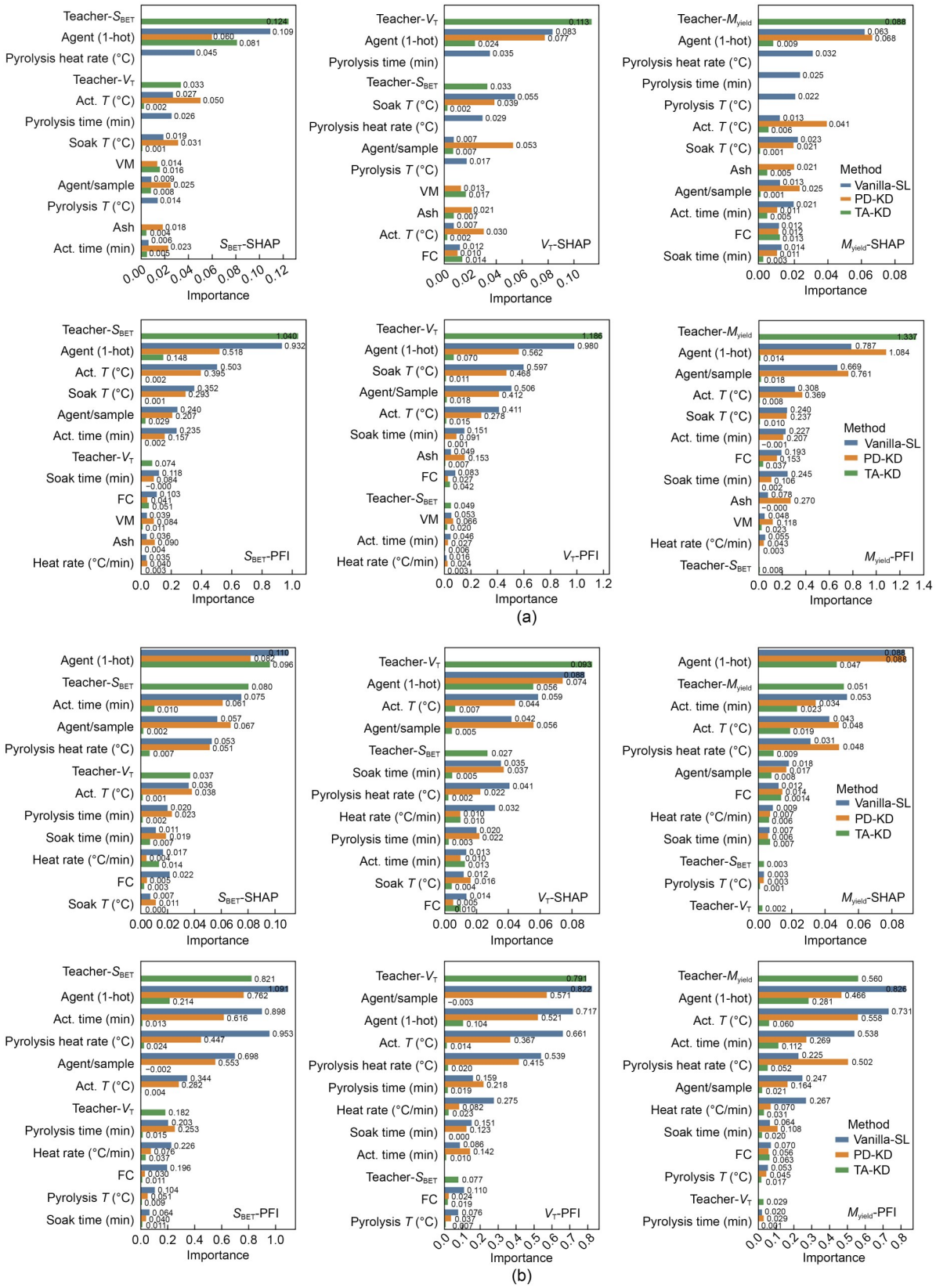

Fig. 3 Variable importance under the one-step and two-step modeling frameworks: (a) PFI and SHAP under dataset one-step; (b) PFI and SHAP under dataset two-step. Agent/sample indicates the number ratio of agents to samples

as important drivers. Increasing activation temperature and duration intensifies activation severity and lowers yield. Strong bases such as KOH typically elevate S_{BET}/V_T at the expense of M_{yield}, making the agent a switch for performance–yield trade-offs.

In the one-step dataset, S_{BET} shows Teacher-S_{BET} and agent as leading features, with activation temperature third and soaking temperature/time together with agent/sample comprising the second tier. SHAP confirms the dominant roles of Teacher-S_{BET} and agent. This behavior reflects the absence of a dedicated pyrolysis stage in one-step activation. The S_{BET} is primarily controlled by agent selection and the peak activation temperature, while soaking conditions regulate agent penetration uniformity and depth as secondary but nonnegligible levers. For V_T, Teacher-V_T is a clear first, followed by the activating agent and soaking temperature/time, with activation temperature ranked next. This indicates heightened sensitivity of V_T to the impregnation stage in one-step activation. The volumetric distribution and uniformity of the activating agent predominantly determine the accessible pore volume, while the downstream activation temperature acts as an amplifier. In contrast, for M_{yield}, the activating agent often ties or slightly exceeds Teacher-M_{yield} in SHAP, whereas Teacher-M_{yield} retains the top spot in PFI. The activation temperature/time and soaking temperature/time contribute strongly. M_{yield} variation reflects the cost of deeper pore development: more thorough impregnation and more severe activation typically raise S_{BET}/V_T while lowering M_{yield}. The models project this trade-off onto temperature/time and agent choice.

These results support a physically interpretable synthesis narrative and offer practical guidance for process design. The activating agent acts as the "steering wheel", while temperature and time serve as the "throttle". Agent chemistry (strong base, acid, or salt) determines the micro/mesopore spectrum and the dominant etching pathway. The activation temperature and duration then tune the severity and enable a controllable trade-off among the S_{BET}, V_T, and M_{yield}. The impregnation stage sets both dose and uniformity, which is especially critical for V_T in one-step activation. A higher soaking temperature or longer soaking time generally increases the accessible pore volume but can also exacerbate downstream skeleton loss, thereby reducing M_{yield}. In the two-step activation, a clear pyrolysis "priming" effect is observed. The heating rate, temperature, and time shape the initial skeleton and crack network, presetting the upper bound and stability of subsequent pore opening. These factors exert indirect yet substantial impacts on S_{BET}, V_T, and M_{yield}. Distillation should be tailored to the target. TA-KD is preferable for structural properties (S_{BET} and V_T) under heterogeneous conditions because it strengthens conditioning. PD-KD is preferable for macroscopic targets (M_{yield}) because it suppresses interval bias. For microstructure-sensitive S_{BET}—particularly on separate datasets—a weaker distillation strength or selective disabling of teacher features may be warranted.

Cross-consistency is evident across importance rankings, calibration/parity plots, and performance metrics. Two-step activation exhibits more linear and monotonic behavior. In one-step activation, S_{BET} is more sensitive to teacher smoothing, and TA-KD slightly overestimates at lower quantiles. Under two-step activation with PD-KD, V_T exhibits near-perfect calibration. For M_{yield}, PD-KD is least biased in one-step activation, whereas TA-KD provides the most balanced calibration in two-step activation. These patterns align with the test R^2/E_{RMS} rankings: in one-step activation, V_T is best with TA-KD, M_{yield} with PD-KD, and S_{BET} with Vanilla-SL; in two-step activation, V_T is best with PD-KD, M_{yield} with TA-KD, and S_{BET} is more stable with Vanilla-SL/TA-KD.

Overall, the analysis indicates that teacher outputs, agent type, and key thermal parameters jointly dominate the input–output mapping for all three targets. Distillation improves robustness by redistributing information sources without altering the fundamental physical relationships. The agreement among importance ordering, calibration/parity alignment, and quantitative performance supports that the models achieve high predictive accuracy while remaining consistent with an 'agent–activation–pore structure–yield' synthesis logic. For process optimization, priority should be given to tuning agent selection, impregnation conditions, and activation temperature/time—and, in two-step activation, pyrolysis heating rate/temperature—supplemented by target-dependent distillation and mechanistic severity descriptors to achieve a quantifiable and interpretable balance among S_{BET}, V_T, and M_{yield}.

Importantly, this interpretability narrative is supported by cross-evidence stability rather than a single post-hoc view. The same small set of dominant drivers (teacher signals, route/agent descriptors, and thermal-severity variables) underpins the models that achieve the best averaged predictive performance across repeated

train/test splits (Table 1). Moreover, on an independent external dataset with a different descriptor schema and targets, namely surface area (SA), total pore volume (TPV), and micropore volume (MPV), the distillation variants preserve the same performance ordering and yield consistent improvements (Table S1 of the ESM), suggesting that the learned attributions reflect transferable synthesis regularities rather than split-specific artifacts.

3.3 Prediction of best interval

3.3.1 Prediction of one-step activation

Under the one-step configuration, robust Pareto screening identifies clear agent-specific operating envelopes that remain favorable across the coupled trade-offs among S_{BET}, V_T, and M_{yield}. Fig. 4 visualizes the Pareto-preferred regions in the activation-temperature design space, while the condensed operating windows are summarized in Table 3. Compared with the original full-variable table, Table 3 emphasizes the most directly controllable process parameters so that the recommended operating ranges can be interpreted more intuitively at the process-design level. Full variable bounds, including feedstock-composition constraints and additional auxiliary variables, are provided in Table S2 of the ESM.

For H_3PO_4, the preferred one-step window is characterized by a relatively high agent dose, long soaking duration, and moderate soaking temperature, followed by activation at comparatively high temperature with moderate residence time. This pattern indicates that in the coupled one-step environment, homogeneous

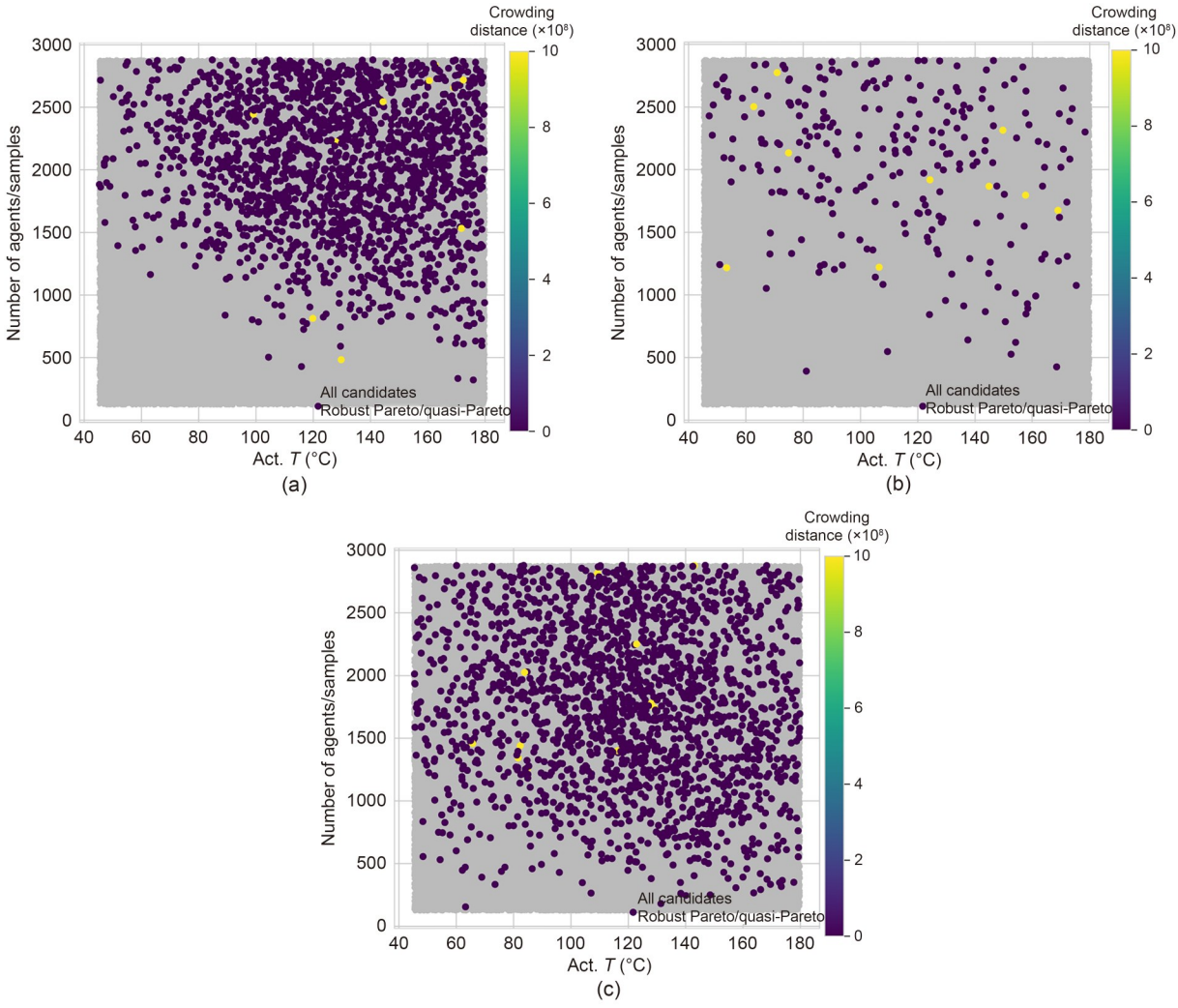

Fig. 4 Robust Pareto screening in the activation-temperature design space for each agent: (a) KOH; (b) H_3PO_4; (c) $ZnCl_2$

Table 3 Condensed optimal operating windows for one-step activation

Agent	Impregnation window	Activation window	Main process signature
H_3PO_4	Number ratio of agents to samples: 4.93–5.50; soaking time: 2100–2260 min; soaking temperature: 37–40 °C	Activation temperature: 638–682 °C; activation time: 109–118 min	High dose and thorough impregnation combined with relatively high activation temperature, favoring balanced S_{BET}–V_T enhancement with controlled M_{yield} penalty
KOH	Number ratio of agents to samples: 2.80–3.21; soaking time: 2017–2152 min; soaking temperature: 41–46 °C	Activation temperature: 452–480 °C; activation time: 124–133 min	Efficient low-to-mid-temperature etching, consistent with strong-base activation, but more sensitive to M_{yield} loss at excessive severity
$ZnCl_2$	Number ratio of agents to samples: 1.76–2.14; soaking time: 1725–1912 min; soaking temperature: 58–64 °C	Activation temperature: 521–599 °C; activation time: 114–124 min	More V_T-oriented operating envelope, with soaking temperature playing a stronger role than extreme dosage

upstream impregnation and sufficiently strong downstream activation jointly promote balanced improvement in S_{BET} and V_T while avoiding excessive M_{yield} loss. The Pareto band in Fig. 4 is therefore consistent with a "high-dose, thorough-impregnation, and moderate-duration activation" strategy.

For KOH, the optimal window shifts toward a lower activation-temperature regime with a slightly longer activation time, despite retaining a substantial impregnation requirement. This agrees with the known strong-base etching behavior of KOH, which can efficiently develop microporosity at lower thermal loads than the acidic route. However, the corresponding Pareto region also indicates a stronger sensitivity of M_{yield} to excessive activation severity, implying that the KOH window should be controlled more carefully to prevent overetching.

For $ZnCl_2$, the preferred envelope is comparatively less dose-intensive but more dependent on elevated soaking temperature, suggesting that solute mobility and impregnation effectiveness play an important role in the resulting pore-volume development. In this route, the activation window remains in a mid-to-high temperature range with moderate duration, and the resulting Pareto region is more clearly aligned with stable V_T enhancement than with extreme S_{BET} maximization. Overall, the one-step results reveal distinct agent-specific operating signatures rather than a single universal optimum, highlighting that the recommended process window must be conditioned on activation chemistry.

3.3.2 Prediction of two-step activation

In the two-step route, the predicted optimal windows more clearly reflect a precarbonization priming effect, whereby the pyrolysis stage reshapes the feasible downstream activation regime. Fig. 5 shows the corresponding Pareto-preferred regions in the activation-temperature design space, and the condensed route-specific operating windows are listed in Table 4. Compared with one-step activation, the two-step process introduces an additional layer of process control through pyrolysis temperature and duration, so a compact main-text summary is particularly useful for linking the graphical Pareto patterns to practically interpretable operating conditions. The full variable bounds, including feedstock-related constraints and heating-rate ranges, are provided in Table S3 of the ESM.

For $(NH_3)_2PO_4$, the two-step optimum is associated with moderate impregnation intensity, a relatively strong pyrolysis stage, and subsequent activation at high temperature but controlled duration. This combination suggests that the carbon skeleton is first effectively preconditioned during pyrolysis and then further opened during activation without relying on a prolonged residence time. The result is a coherent process window that supports simultaneous gains in the S_{BET} and V_T while moderating the M_{yield} penalty, consistent with the more stable Pareto structure observed in Fig. 5.

For KOH, the two-step route differs noticeably from the one-step case. The preferred window combines an elevated soaking temperature with an extended pyrolysis stage, followed by a shorter activation step at intermediate temperature. This shift indicates that stronger precarbonization partly replaces the need for prolonged downstream etching, yielding a more controlled pathway toward pore development. In practical terms, the KOH optimum under two-step processing is better interpreted as a "preform first, activate briefly" strategy rather than the longer one-step etching pattern.

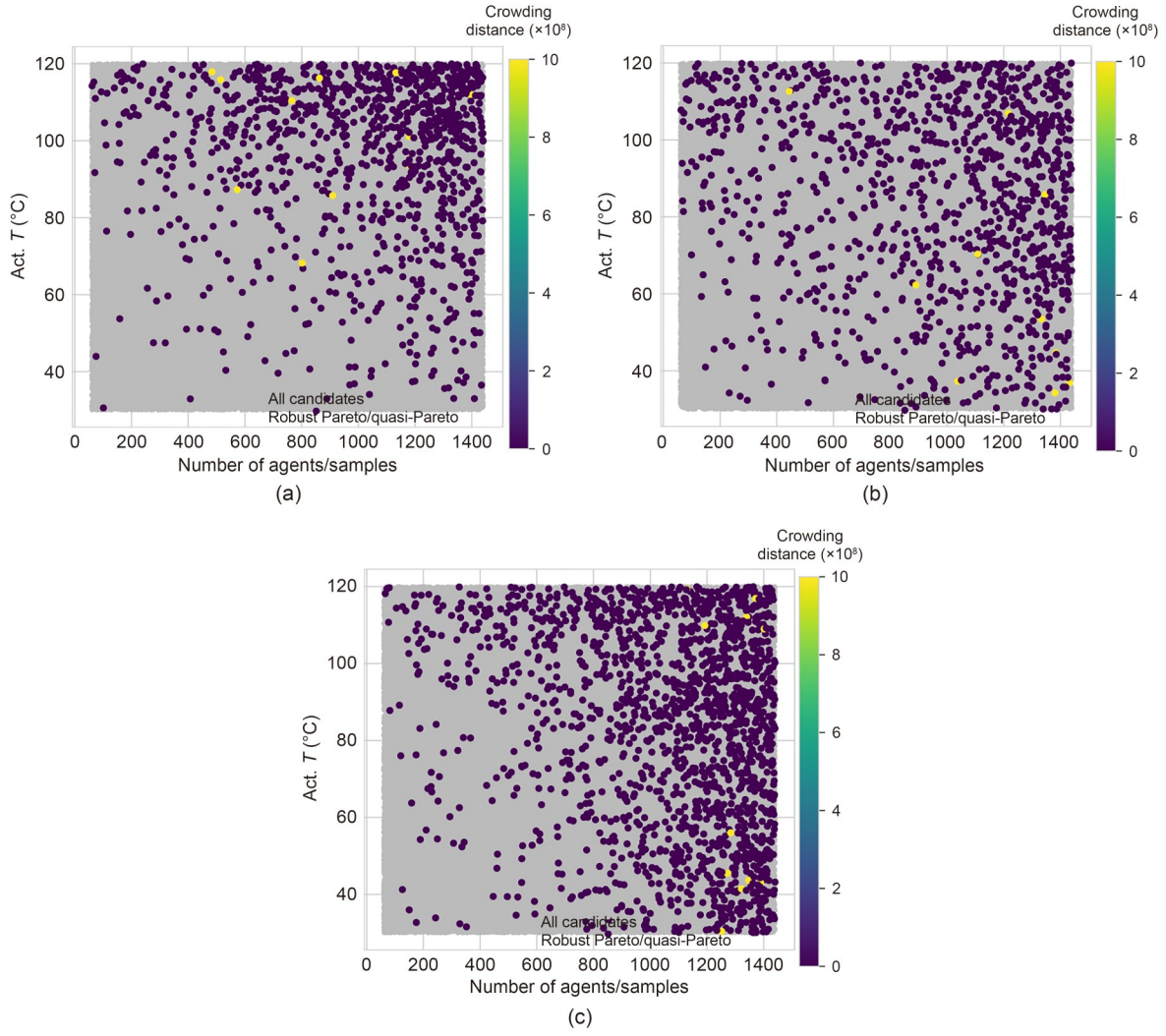

Fig. 5 Robust Pareto screening in the activation-temperature design space for each agent: (a) KOH; (b) $(NH_3)_2PO_4$; (c) $ZnCl_2$

Table 4 Condensed optimal operating windows for two-step activation

Agent	Impregnation window	Pyrolysis window	Activation window	Main process signature
$(NH_3)_2PO_4$	Number ratio of agents to samples: 1.21–1.60; soaking time: 979–1106 min; soaking temperature: 68–79 °C	Pyrolysis temperature: 698–786 °C; pyrolysis time: 162–204 min	Activation temperature: 786–835 °C; activation time: 86–95 min	Pyrolysis preforming followed by high-temperature but controlled-duration activation, supporting concurrent S_{BET} and V_T development
KOH	Number ratio of agents to samples: 1.85–2.42; soaking time: 1023–1128 min; soaking temperature: 76.5–84.9 °C	Pyrolysis temperature: 650–743 °C; pyrolysis time: 306–353 min	Activation temperature: 633–715 °C; activation time: 100–105 min	Stronger precarbonization reduces the need for prolonged downstream etching and shifts the optimum toward shorter activation
$ZnCl_2$	Number ratio of agents to samples: 0.78–0.96; soaking time: 1115–1191 min; soaking temperature: 68.6–77.8 °C	Pyrolysis temperature: 746–807 °C; pyrolysis time: 188–236 min	Activation temperature: 799–832 °C; activation time: 81–90 min	A stable V_T-oriented compromise characterized by moderate pyrolysis conditioning and short, high-temperature activation

For $ZnCl_2$, the Pareto-preferred region again reflects a stable V_T-oriented compromise, but here, the role of pyrolysis becomes more explicit. The recommended window combines relatively modest impregnation with mid-to-high pyrolysis conditioning and a short, high-temperature activation stage. This implies that once a suitable mesostructural precursor state is established during pyrolysis, the subsequent activation can be kept short while still preserving favorable pore-volume development. Taken together, the two-step results reinforce that the optimal process window is jointly controlled by the activating agent and route and that pyrolysis acts as a key upstream lever for shaping the final activation requirements.

These conclusions are consistent with the calibration and parity evidence. They indicate that the modeling framework not only achieves robust predictive performance but also offers coherent, implementable guidance on the kinetic trade-offs among activating agent, activation severity, pore structure, and yield.

4 Conclusions

The objective of this work is to develop a robust and interpretable modeling framework for predicting key AC properties (S_{BET}, V_T, and M_{yield}) from tabular, small-sample, and heterogeneous process data while enabling route-/agent-specific process-window screening and multiobjective optimization under practical trade-offs. This objective is achieved by proposing AC-ResKD, a mechanism-enhanced multitask distillation model built on a shared ResDNN backbone and dual-channel priors from PD-KD and TA-KD. By embedding key process factors (activating agent, temperature/time, impregnation, heating rate, and precursor composition) together with high-level teacher predictions, AC-ResKD establishes an interpretable input–output mapping and supports calibrated multihead learning for S_{BET}, V_T, and M_{yield}.

The results indicate that KD significantly enhances target-dependent performance under repeated-split evaluation (mean and 95% CIs over 20 random 80/20 splits). Route-separated training consistently exhibits higher monotonicity and better calibration than the combined training, and the two-step route shows the clearest precarbonization "priming" signatures in V_T (higher R^2 and lower E_{RMS}), consistent with pyrolysis presetting the carbon skeleton and pore-size spectrum. For S_{BET} and V_T, AC-ResKD achieves competitive—often higher—accuracy than the literature reference (Zou et al., 2024) when evaluated in original units, while M_{yield} remains the most challenging target, motivating route-specific modeling and uncertainty-aware screening for actionable recommendations.

Interpretability analyses provide convergent evidence that teacher outputs, activating agent, and thermal parameters jointly dominate the input–output mapping for S_{BET}, V_T, and M_{yield}. In one-step activation, the impregnation stage emerges as the primary determinant of accessible pore volume; in two-step activation, pyrolysis variables rise markedly in importance, aligning with mechanistic expectations. Calibration and parity plots corroborate these rankings, indicating higher linearity and lower bias for the two-step route and target-specific advantages of TA-KD and PD-KD.

Moreover, robust Pareto screening combined with quantile-window extraction identifies agent- and route-specific optimal operating ranges under multiobjective trade-offs. These intervals are subsequently adjusted to executable bounds, delineating coherent temperature–dose–time subdomains within which S_{BET}/V_T can be jointly improved while controlling yield penalties, thereby providing actionable guidance for scale-up.

The proposed approach satisfies the request for robust prediction of AC properties in complex process spaces, supports coordinated optimization of specific surface area and pore volume under yield constraints, enables screening of optimal process windows and route selection across activating agents, and can be further leveraged for parameter tuning and inverse design in AC manufacturing.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 52576233) and the Shenzhen Science and Technology Innovation Commission (Nos. KCXFZ20240903093459001 and KCXST20221021111609024), China.

Author contributions

Peng ZHAO and Hao YU designed the research. Peng ZHAO and Jiahui ZHOU processed the corresponding data. Hao YU wrote the first draft of the manuscript. Peng ZHAO helped to organize the manuscript. Hao YU revised and edited the final version.

Conflict of interest

Peng ZHAO, Jiahui ZHOU, Guangyao LI, Hao YU, Dongxu JI, Takahiko MIYAZAKI, Kyaw THU declare that they have no conflict of interest.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Reference

Belluati M, Tabasso S, Calcio Gaudino E, et al., 2024. Biomass-derived carbon-based catalysts for lignocellulosic biomass and waste valorisation: a circular approach. *Green Chemistry*, 26(15):8642-8668. https://doi.org/10.1039/D4GC00606B

Caruana R, 1997. Multitask learning. *Machine Learning*, 28(1): 41-75. https://doi.org/10.1023/A:1007379606734

Chairunnisa, Yu H, Saren S, et al., 2024. A review of recent advances in sustainable preparation of high-performing activated carbon for dehumidification technology. *Journal of Materials Science*, 59(43):20121-20156. https://doi.org/10.1007/s10853-024-10265-8

Devi R, Kumar V, Kumar S, et al., 2023. Recent advancement in biomass-derived activated carbon for waste water treatment, energy storage, and gas purification: a review. *Journal of Materials Science*, 58(30):12119-12142. https://doi.org/10.1007/s10853-023-08773-0

Evans MJB, Halliop E, MacDonald JAF, 1999. The production of chemically-activated carbon. *Carbon*, 37(2):269-274. https://doi.org/10.1016/S0008-6223(98)00174-2

Fu J, Kang QH, Ao WY, et al., 2023. Comparison and analysis of one- and two-step activation for preparation of activated carbon from furfural residues. *Biomass Conversion and Biorefinery*, 13(6):4681-4694. https://doi.org/10.1007/s13399-021-01439-4

Giudicianni P, Gargiulo V, Grottola CM, et al., 2021. Inherent metal elements in biomass pyrolysis: a review. *Energy & Fuels*, 35(7):5407-5478. https://doi.org/10.1021/acs.energyfuels.0c04046

Gorishniy Y, Rubachev I, Khrulkov V, et al., 2021. Revisiting deep learning models for tabular data. Proceedings of the 35th International Conference on Neural Information Processing Systems, article 1447.

Heidarinejad Z, Dehghani MH, Heidari M, et al., 2020. Methods for preparation and activation of activated carbon: a review. *Environmental Chemistry Letters*, 18(2):393-415. https://doi.org/10.1007/s10311-019-00955-0

Ibrahim AF, Hussein MA, 2025. Leveraging machine learning for prediction and optimization of texture properties of sustainable activated carbon derived from waste materials. *Scientific Reports*, 15(1):11313. https://doi.org/10.1038/s41598-025-95061-3

Kadra A, Lindauer M, Hutter F, et al., 2021. Well-tuned simple nets excel on tabular datasets. Proceedings of the 35th International Conference on Neural Information Processing Systems, article 1832.

Kaur M, Mittal N, Charak A, et al., 2023. Rice husk derived activated carbon for the adsorption of scarlet RR an anionic disperse dye. *Evergreen*, 10(1):438-443. https://doi.org/10.5109/6782146

Kolbadinejad S, Mashhadimoslem H, Ghaemi A, et al., 2022. Deep learning analysis of Ar, Xe, Kr, and O_2 adsorption on activated carbon and zeolites using ANN approach. *Chemical Engineering and Processing - Process Intensification*, 170:108662. https://doi.org/10.1016/j.cep.2021.108662

Kundu S, Khandaker T, Anik MAAM, et al., 2024. A comprehensive review of enhanced CO_2 capture using activated carbon derived from biomass feedstock. *RSC Advances*, 14(40):29693-29736. https://doi.org/10.1039/D4RA04537H

Lamb WF, Grubb M, Diluiso F, et al., 2022. Countries with sustained greenhouse gas emissions reductions: an analysis of trends and progress by sector. *Climate Policy*, 22(1):1-17. https://doi.org/10.1080/14693062.2021.1990831

Li GY, Xue B, Yu H, et al., 2025a. Performance improvement of waste heat upgrading adsorption heat pump by employing copper oxide-loaded composite zeolites for high-temperature steam generation. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 711:136330. https://doi.org/10.1016/j.colsurfa.2025.136330

Li GY, Yu H, Ji DX, et al., 2025b. Pine cone-based activated carbon via dual physical activation for efficient carbon dioxide capture: experimental and molecular simulation studies. *Energy*, 328:136506. https://doi.org/10.1016/j.energy.2025.136506

Li XH, Huang ZH, Shao SS, et al., 2024. Machine learning prediction of physical properties and nitrogen content of porous carbon from agricultural wastes: effects of activation and doping process. *Fuel*, 356:129623. https://doi.org/10.1016/j.fuel.2023.129623

Liao MC, Kelley SS, Yao Y, 2019. Artificial neural network based modeling for the prediction of yield and surface area of activated carbon from biomass. *Biofuels, Bioproducts and Biorefining*, 13(4):1015-1027. https://doi.org/10.1002/bbb.1991

Lu YH, Zhang SL, Yin JM, et al., 2017. Mesoporous activated carbon materials with ultrahigh mesopore volume and effective specific surface area for high performance supercapacitors. *Carbon*, 124:64-71. https://doi.org/10.1016/j.carbon.2017.08.044

Merchant A, Batzner S, Schoenholz SS, et al., 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80-85. https://doi.org/10.1038/s41586-023-06735-9

Muhammed F, Lavaggi T, Advani S, et al., 2021. Influence of material and process parameters on microstructure evolution during the fabrication of carbon–carbon composites: a

review. *Journal of Materials Science*, 56(32):17877-17914. https://doi.org/10.1007/s10853-021-06401-3

Othman NAB, Kim T, Imamura A, et al., 2013. Investigation on H_2S removal factors of activated carbons derived from waste palm trunk. *Journal of Novel Carbon Resource Sciences*, 7:7-11.

Przybek A, Setlak K, Piątkowski J, et al., 2025. Alkaline-activated materials for CO_2 capture–literature review, own observations, and future perspectives. *Evergreen*, 12(3): 1664-1679. https://doi.org/10.5109/7388857

Reiser P, Neubert M, Eberhard A, et al., 2022. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93. https://doi.org/10.1038/s43246-022-00315-6

Ruder S, 2017. An overview of multi-task learning in deep neural networks. arXiv:1706.05098. https://doi.org/10.48550/arXiv.1706.05098

Sevilla M, Mokaya R, 2014. Energy storage applications of activated carbons: supercapacitors and hydrogen storage. *Energy & Environmental Science*, 7(4):1250-1280. https://doi.org/10.1039/C3EE43525C

Shukla AK, Alam J, Mallik S, et al., 2024. Optimization and prediction of dye adsorption utilising cross-linked chitosan-activated charcoal: response surface methodology and machine learning. *Journal of Molecular Liquids*, 411:125745. https://doi.org/10.1016/j.molliq.2024.125745

Song YH, Li JK, Chi DZ, et al., 2025. AI-driven advances in metal–organic frameworks: from data to design and applications. *Chemical Communications*, 61(82):15972-16001. https://doi.org/10.1039/D5CC04220H

Sosa JA, Laines JR, García DS, et al., 2023. Activated carbon: a review of residual precursors, synthesis processes, characterization techniques, and applications in the improvement of biogas. *Environmental Engineering Research*, 28(3):220100. https://doi.org/10.4491/eer.2022.100

Yu H, Seo SW, Mikšík F, et al., 2023. Effects of temperature and humidity ratio on the performance of desiccant dehumidification system under low-temperature regeneration. *Journal of Thermal Analysis and Calorimetry*, 148(8):3045-3058. https://doi.org/10.1007/s10973-022-11368-7

Yu H, Mikšík F, Thu K, et al., 2024. Characterization and optimization of pore structure and water adsorption capacity in pinecone-derived activated carbon by steam activation. *Powder Technology*, 431:119084. https://doi.org/10.1016/j.powtec.2023.119084

Yuan XZ, Suvarna M, Low S, et al., 2021. Applied machine learning for prediction of CO_2 adsorption on biomass waste-derived porous carbons. *Environmental Science & Technology*, 55(17):11925-11936. https://doi.org/10.1021/acs.est.1c01849

Yudha CS, Hutama AP, Gustiana HSEA, et al., 2024. An efficient Nypa fiber waste conversion to activated carbon for Li-ion battery anode material. *Evergreen*, 11(3):2193-2201. https://doi.org/10.5109/7236863

Zhang K, Zhong SF, Zhang HC, 2020. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. *Environmental Science & Technology*, 54(11): 7008-7018. https://doi.org/10.1021/acs.est.0c02526

Zhao P, Zhou JH, Li GY, et al., 2025. Prediction of physisorption on zeolites using graph integrated adsorption network and spline data augmentation. *Thermal Science and Engineering Progress*, 66:104047. https://doi.org/10.1016/j.tsep.2025.104047

Zhou ZY, Yu H, Li GY, et al., 2026. Material–scenario coupled techno-economics comparison of carbon capture adsorbents: feasibility and suitability across various contexts. *Energy Conversion and Management*, 348:120589. https://doi.org/10.1016/j.enconman.2025.120589

Zou RG, Yang ZB, Zhang JH, et al., 2024. Machine learning application for predicting key properties of activated carbon produced from lignocellulosic biomass waste with chemical activation. *Bioresource Technology*, 399:130624. https://doi.org/10.1016/j.biortech.2024.130624

Electronic supplementary materials

Sections S1–S5, Fig. S1, Tables S1–S3, Eqs. (S1)–(S7)