

Supplementary materials

Effect of material properties on the thermal responses of the carbonization and pyrolysis layers of polymer matrix composites for charring-ablators

Yongxiang Li¹, Xiao Liu², Xiangdong Wang¹, Wei Xie¹, Di Qiu^{1,3,*}, Jiong Yang^{1,*}

¹Materials Genome Institute, Shanghai Engineering Research Center for Integrated Circuits and Advanced Display Materials, Shanghai University, Shanghai 200444, China.

²China Aerodynamics Research and Development Center, Mianyang 621000, Sichuan, China.

³Shanghai Frontier Science Center of Mechanoinformatics, Shanghai University, Shanghai 200444, China.

***Correspondence to:** Dr. Di Qiu, Prof. Jiong Yang, Materials Genome Institute, Shanghai Engineering Research Center for Integrated Circuits and Advanced Display Materials, Shanghai University, Shanghai 200444, China. E-mail: diqiu0319@shu.edu.cn; jiongy@t.shu.edu.cn

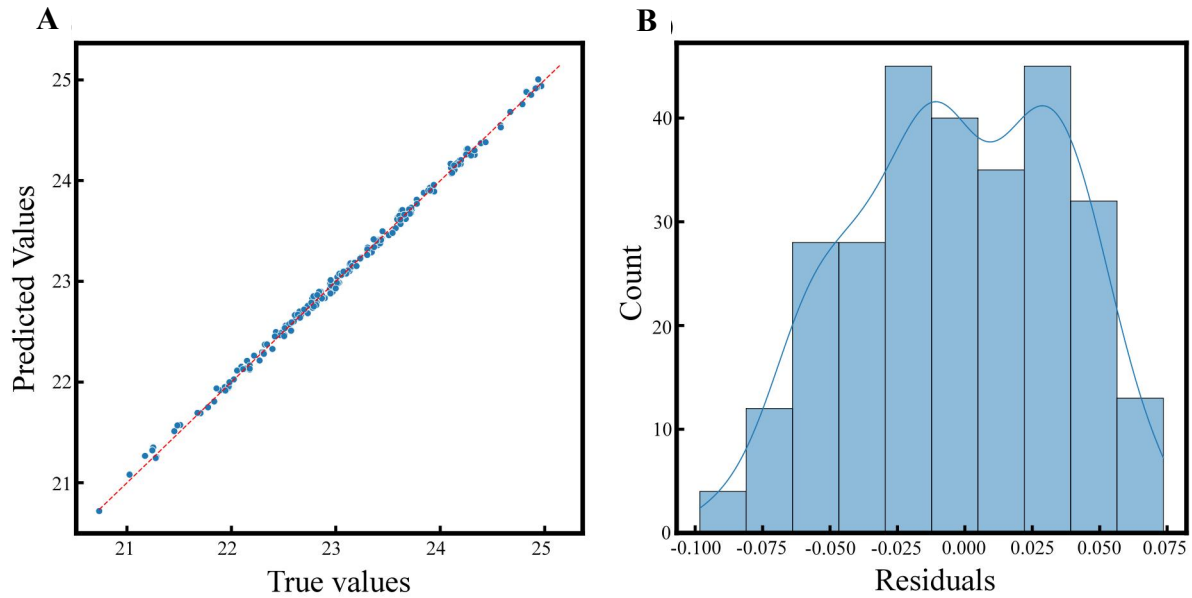
While feature selection methods such as random forests and principal component analysis (PCA) are widely used, they often fail to generate explicit models that can be directly applied by experimental researchers. This limitation motivated our adoption of the SISO algorithm, which not only selects features but also provides an explicit mathematical formulation for immediate practical use. To validate our methodological choice, we compared models generated by LASSO and SISO. Both linear regression (LASSO) and SISO-based regression yielded models with comparable accuracy. However, SISO incorporates nonlinear factors during regression, resulting in more concise models with fewer selected features. Given our objective to prioritize the physical interpretability of derived formulas alongside accuracy, we selected SISO for its ability to balance nonlinear complexity with explicit expression. The explicit models generated by symbolic regression (e.g., LASSO and SISO) enable exploration of the physical meaning underlying feature-target relationships, as demonstrated in our preliminary LASSO results (see Supplementary Materials for detailed comparisons). For further discussion on SISO's advantages in descriptor identification and compression, we direct readers to the foundational work by Ouyang et al. (Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L.M. SISO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* 2018, 2, 083802 ; Bartel, C.J.; Sutton, C.; Goldsmith, B.R.; Ouyang, R.; Musgrave, C.B.; Ghiringhelli, L.M.; Scheffler, M. New tolerance factor to predict the stability of perovskite oxides and halides. *Science advances* 2019, 5, eaav0693).

1. LASSO Linear Regression Full Parameter Model of K_{start1}

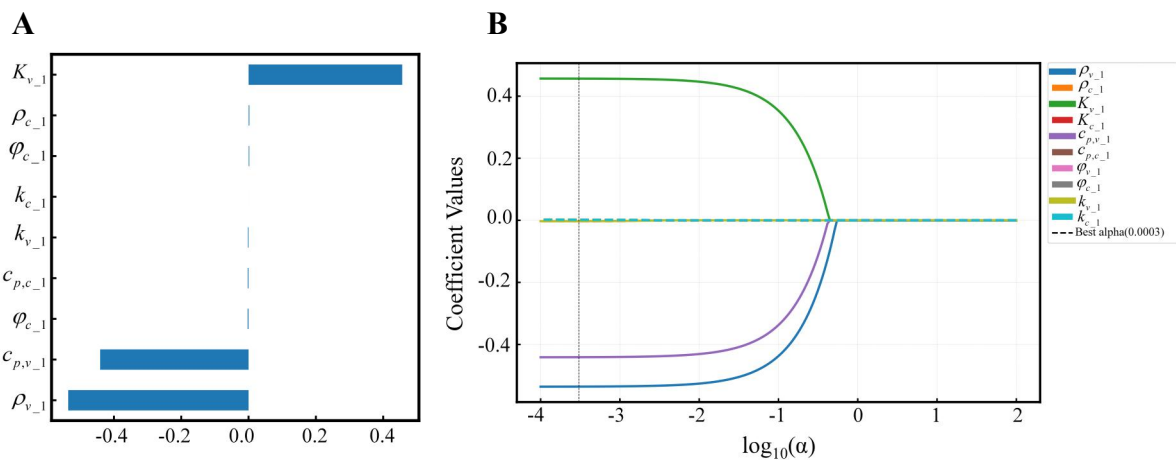
We employed the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm to conduct regression analysis on the dataset, yielding an explicit linear model as presented in Eq.(S-1).

$$K_{start1} = 45.7970 - 0.0175 * \rho_{v1} + 0.0001 * \rho_{c1} + 65.0191 * K_{v1} - 0.0281 * c_{p,v1} - 0.0001 * c_{p,c1} - 0.0214 * k_{v1} - 0.0004 * k_{c1} - 0.0713 * \varphi_{v1} + 0.0516 * \varphi_{c1} \quad (S-1)$$

The resulting model establishes a linear relationship between the target variable and nine parameters (e.g., ρ_{v1} , ρ_{c1} , K_{v1} , $c_{p,v1}$, $c_{p,c1}$, k_{v1} , k_{c1} , φ_{v1} , φ_{c1}). To assess the model's performance, we evaluated its predictive accuracy on the test set, with the results illustrated in Supplementary Figure 1. The full-feature model, derived through LASSO, exhibits excellent performance, with predicted values closely aligning with the true values, demonstrating robust predictive capability. Supplementary Figure 1 provides a detailed evaluation of the regression model's performance. Supplementary Figure 1A presents a scatter plot comparing the predicted values against the true values, where the x-axis corresponds to the true values and the y-axis corresponds to the predicted values. The data points, represented as blue dots, are tightly clustered around the red dashed line, which denotes the line of perfect prediction ($y = x$). This close alignment indicates a strong linear relationship between the predicted and true values, underscoring the model's high predictive accuracy. Supplementary Figure 1B displays a histogram of the residuals, with the x-axis representing the residuals and the y-axis representing the count. The residuals exhibit a symmetric distribution centered around zero, with the majority falling within the range of -0.050 to 0.050. The frequency decreases as the residuals deviate from zero, and a kernel density estimate overlaid on the histogram confirms that the residuals follow an approximately normal distribution. This pattern suggests that the model's errors are small, random, and free of systematic bias, further reinforcing its reliability.



Supplementary Figure 1. Performance Evaluation of the LASSO Regression Model on the Test Set (A) Scatter plot of predicted values versus true values for the test set. The red dashed line represents the line of perfect prediction ($y = x$). (B) Histogram of residuals (prediction errors) with a kernel density estimate (KDE, orange curve) overlaid.



Supplementary Figure 2. (A) Feature importance and (B) LASSO regularization path.

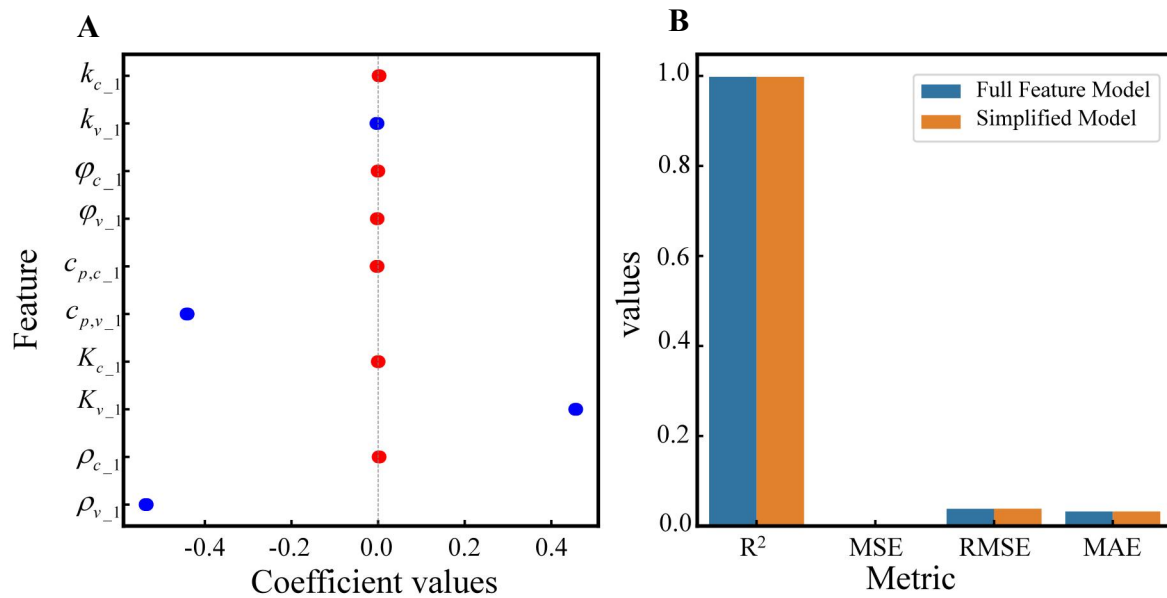
Supplementary Figure 2 illustrates the LASSO regression model's feature selection and regularization process. Supplementary Figure 2A displays the nine non-zero coefficients retained in the final model, ranked by their absolute magnitude. Among these, K_{v-1} exhibits the strongest positive association with the target variable (coefficient = 0.45), while ρ_{v-1} shows the most pronounced negative correlation (coefficient = -0.5). These directional

relationships quantitatively characterize the influence of each feature on the outcome. Supplementary Figure 2B provides a regularization path analysis, demonstrating how coefficients shrink toward zero as the regularization parameter increases from -3 to 2 . The vertical dashed line marks the optimal value (0.0003), determined via cross-validation to balance model complexity and predictive performance.

Notably, the coefficients of retained features (depicted as solid lines, e.g., $K_{v,1}$, $\rho_{v,1}$, $c_{p,v,1}$) remain stable across a range of values, indicating their robust contribution to the model. Conversely, features excluded from the final model (dashed lines) rapidly approach zero with increasing regularization strength.

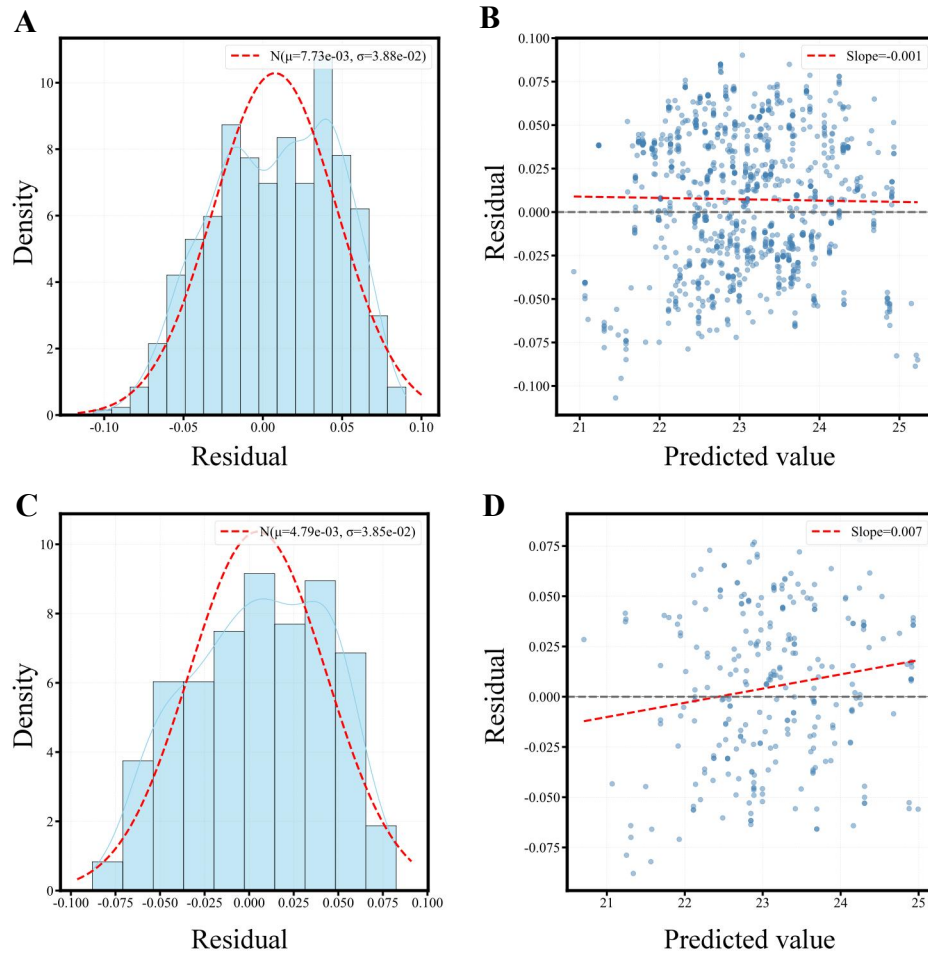
To ensure robust feature selection and validate the stability of coefficients, we applied a Bootstrap resampling approach to the full-feature LASSO model. This involved generating 500 bootstrap samples (with replacement) from the training data, retraining LASSO models on each subsample, and computing 95% confidence intervals (CIs) for each feature's coefficient. Features were retained only if their CIs excluded zero, ensuring statistically significant and stable associations with the target variable. This pruning process reduced the feature set from 9 to 4 significant variables (e.g., $K_{v,1}$, $\rho_{v,1}$, $c_{p,v,1}$, $\varphi_{v,1}$) while maintaining high predictive performance (test set vs. $R^2=0.998$ for the full model). The pruned model (Eq.(S-2)) achieved a 44% reduction in complexity without compromising accuracy.

$$K_{start_1} = 45.8413 - 0.0175 * \rho_{v_1} + 65.0387 * K_{v_1} - 0.0281 * c_{p,v_1} - 0.0799 * \varphi_{v_1} \text{(S-2)}$$



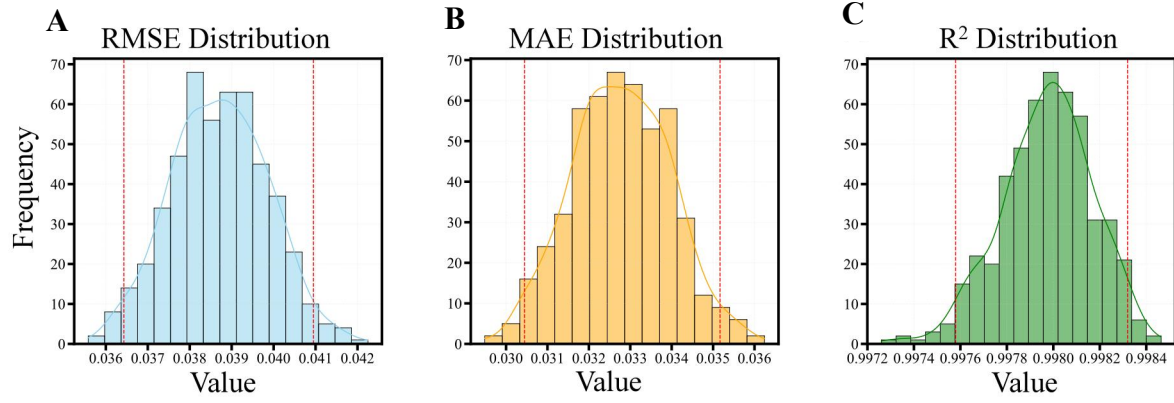
Supplementary Figure 3. (A) LASSO Coefficients Distribution Based on 500 Bootstrap Resamples, (B) Model Performance Metrics.

Supplementary Figure 4 comprehensively validates the error characteristics and predictive robustness of the pruned LASSO model. Supplementary Figure 4A and C display residual histograms with superimposed normal distribution curves (red dashed lines), quantifying minimal deviation from Gaussian assumptions (Supplementary Figure 4A, training residuals: $\mu = 7.73 \times 10^{-3}$, $\sigma = 3.88 \times 10^{-2}$; Supplementary Figure 4C, test residuals: $\mu = 0.005 \pm 0.01$, $\sigma = 0.038$). In the residual vs. predicted value scatter plots, the LOWESS trend line slopes for the training and testing sets are -0.001 and 0.007, respectively, with residuals equally distributed around the zero line (Supplementary Figure 4B, range: -0.1 to +0.1 for training dataset, Supplementary Figure 4D, range: -0.075 to +0.075 for testing dataset). This indicates stable error variance (homoscedasticity) and no systematic overestimation or underestimation. In summary, the model demonstrates no significant systematic bias within the data coverage range, with well-controlled error fluctuations.



Supplementary Figure 4. Combined Residual Diagnostics for Training and Test Sets, (A) Histogram of training set residuals with normal fit, (B) Training residuals vs. predicted values with LOWESS smoothing, (C) Histogram of test set residuals with normal fit, (D) Training residuals vs. predicted values with LOWESS smoothing.

To quantify robustness against data variability, 500 bootstrap iterations were conducted. Supplementary Figure 5 displays the distributions of (a) RMSE (blue), (b) MAE (orange), and (c) R^2 (green) calculated from 500 resampled iterations. RMSE: Tightly distributed around 0.039 (95% CI (Confidence Interval [62]): 0.0365–0.041), reflecting minimal prediction deviation. MAE: Narrow range (0.030–0.035, 95% CI). R^2 : Values cluster near 0.99 (95% CI: 0.9976–0.9983), indicating near-perfect explanatory power.



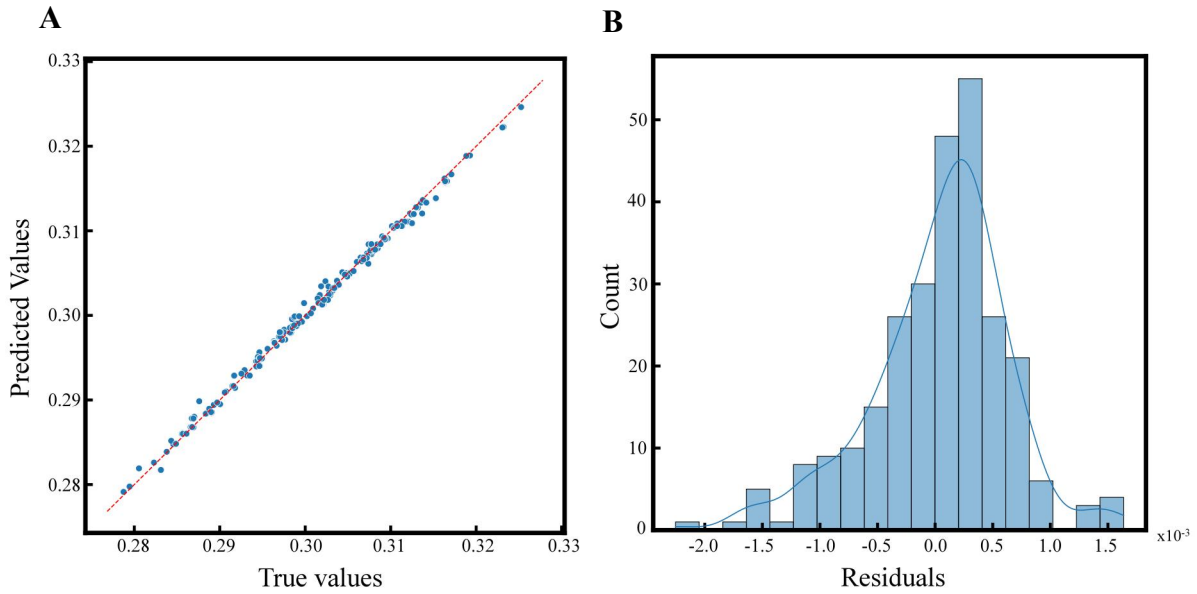
Supplementary Figure 5. Bootstrap Stability Analysis: Error Distributions and Model Explanatory Power, (A) RMSE distribution (mean=0.039, 95% CI [0.0365–0.041]), (B)MAE distribution (range=0.005, 95% CI [0.030–0.035]), (C)R² distribution (median=0.99).

2. LASSO Linear Regression Full Parameter Model of K_{end_2}

For the K_{end_2} model, the same LASSO linear regression procedure was implemented following our established methodology, yielding the formulation presented in Eq.(S-3).

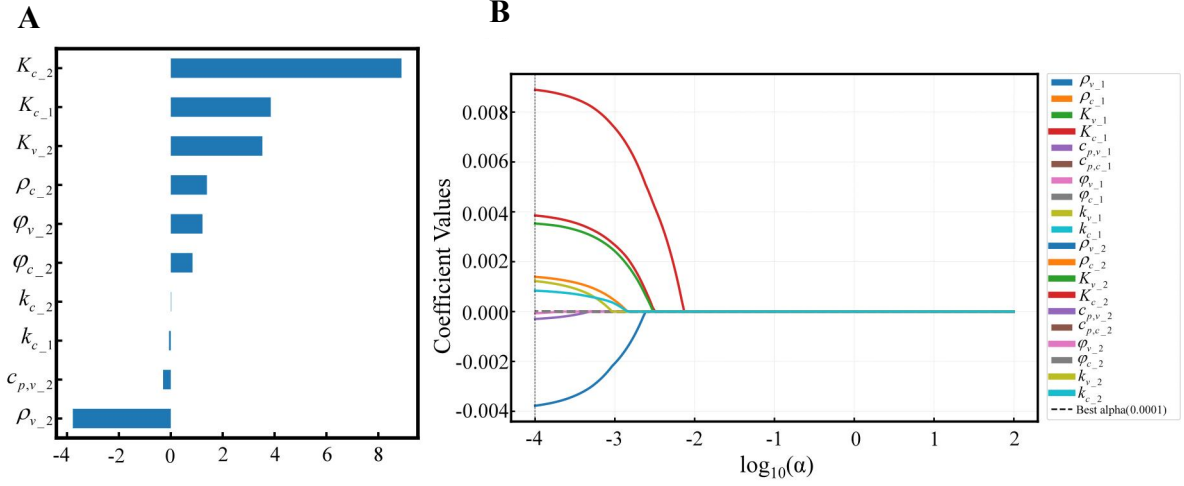
$$K_{end_2} = 0.1668235 + 0.2338069 * K_{c_1} - 0.0002202 * \rho_{v_2} + 0.0001879 * \rho_{c_2} + 1.2694184 * K_{v_2} + 0.5544020 * K_{c_2} - 0.0000159 * c_{p,v_2} - 0.0007415 * k_{v_2} + 0.0001300 * k_{c_2} + 0.0332039 * \varphi_{v_2} + 0.0210698 * \varphi_{c_2} \quad (S-3)$$

The resultant model revealed a significant linear association between the target variable and 10 physical parameters: (e.g., K_{c_1} , ρ_{v_2} , ρ_{c_2} , K_{v_2} , K_{c_2} , c_{p,v_2} , k_{v_2} , k_{c_2} , φ_{v_2} , φ_{c_2}). Model performance was quantitatively assessed through predictive accuracy evaluation on the independent test set, with detailed visualization provided in Supplementary Figure 6. Supplementary Figure 6A demonstrates the agreement between predicted and observed values through a bivariate scatter plot, where the ordinate represents model predictions and the abscissa denotes ground truth measurements. The observed data points (blue markers) show tight clustering along the 1:1 reference line (red dashed line, $y = x$), indicating strong predictive consistency. This alignment is quantitatively supported by a coefficient of determination (R^2) approaching unity. A complementary analysis of residual distributions is presented in Supplementary Figure 6B. The histogram reveals a symmetric, unimodal residual distribution centered at zero, with >95% of residuals confined within ± 0.50 units. Frequency distribution follows a monotonic decay pattern with increasing distance from zero, consistent with Gaussian error distribution as confirmed by the superimposed kernel density estimate. This error structure demonstrates that model inaccuracies are predominantly random fluctuations rather than systematic deviations, thereby validating the model's statistical robustness and absence of significant bias in predictions.



Supplementary Figure 6. Performance Evaluation of the LASSO Regression Model on the Test Set (A) Scatter plot of predicted values versus true values for the test set. The red dashed line represents the line of perfect prediction ($y = x$). (B) Histogram of residuals (prediction errors) with a kernel density estimate (KDE, orange curve) overlaid.

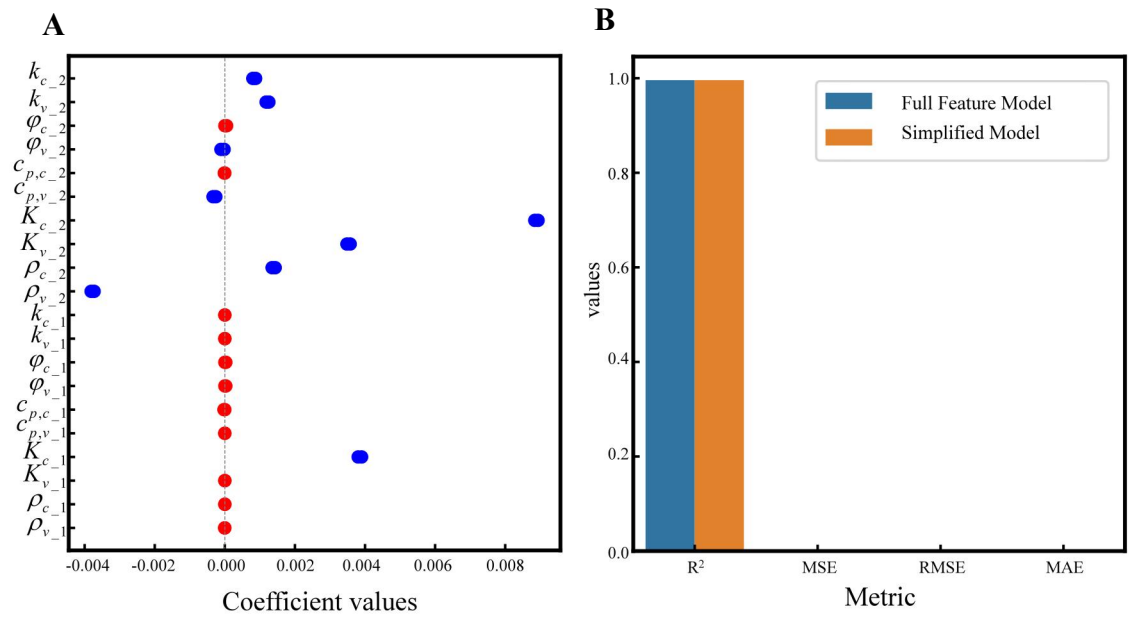
Supplementary Figure 7 delineates the LASSO regression model's feature selection mechanism and regularization dynamics. In Supplementary Figure 7A, 10 features with non-zero coefficients are retained in the final model, ordered by their absolute magnitudes. The analysis reveals that K_{c_2} , demonstrates the most substantial positive association with the target variable, whereas ρ_{v_2} exhibits the strongest negative correlation. These directional trends quantitatively parameterize each feature's influence on the predictive outcome. Supplementary Figure 7B traces the regularization path, illustrating coefficient evolution as the regularization parameter α spans from 10^{-4} to 10^2 (logarithmic scale). The vertical dashed line denotes the optimal α value (0.0001), selected through cross-validation to optimize the bias-variance tradeoff.



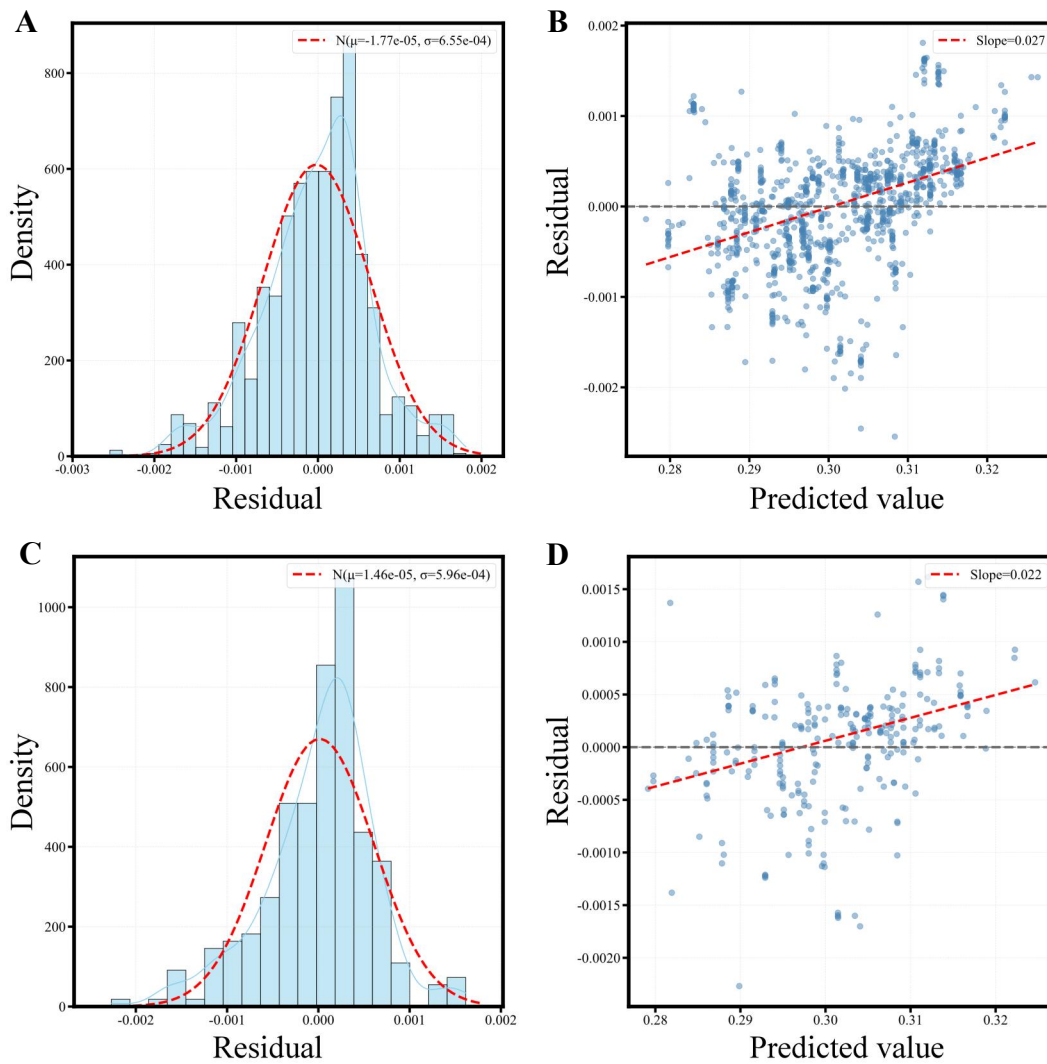
Supplementary Figure 7. (A) Feature importance and (B) LASSO regularization path.

$$K_{end2} = 0.1669953 + 0.2338261 * K_{c,1} - 0.0002202 * \rho_{v,2} + 0.0001878 * \rho_{c,2} + 1.2700623 * K_{v,2} + 0.5547655 * K_{c,2} - 0.0000159 * c_{p,v,2} - 0.0007433 * k_{v,2} + 0.0332509 * \varphi_{v,2} + 0.0210573 * \varphi_{c,2} \quad (S-4)$$

Eq. (S-4) represents the model after pruning. Supplementary Figure 8 delineates the systematic feature pruning process and its impact on model performance metrics. In panel (a), coefficient stability analysis is quantified through bootstrap resampling (n =500 iterations), where features are retained only if their 95% confidence intervals (CIs) exclude zero. This statistical filtering reduced the feature set from 10 to 9 critical variables with absolute coefficients ranging from -0.004 to 0.009. Supplementary Figure 8B validates the pruning efficacy through comparative performance metrics. The simplified model achieves parity with the full-feature model in predictive accuracy (test $R^2=0.995$). Error metrics remain statistically indistinguishable between configurations: RMSE (0.000595 vs. 0.000595), MSE (3.543491×10^{-7} vs. 3.542766×10^{-7}), and MAE (0.000452 vs. 0.000451) for full and simplified models, respectively. The preserved performance profile across regularization intensities ($\log_{10}(\alpha)$ from -4 to 2) confirms that feature elimination specifically targeted non-essential variables without distorting the underlying predictive structure.



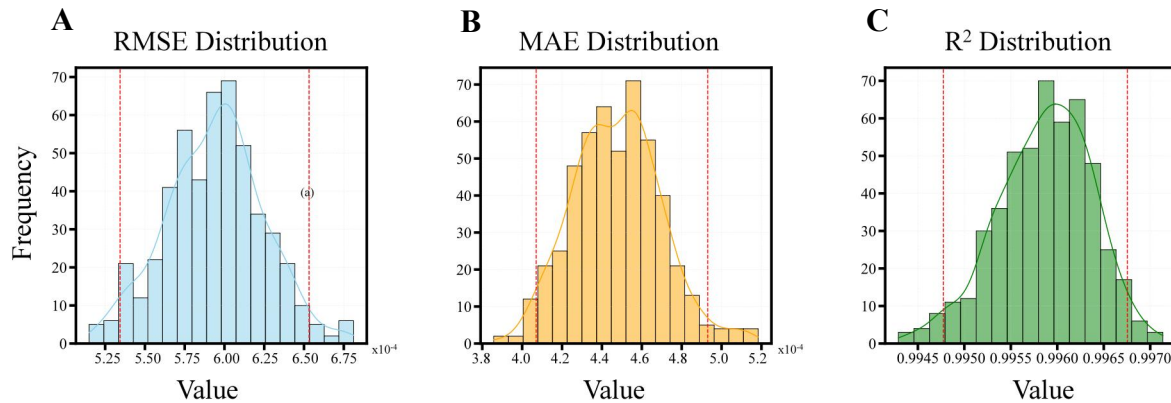
Supplementary Figure 8. (A) LASSO Coefficients Distribution Based on 500 Bootstrap Resamples (B) Model Performance Metrics.



Supplementary Figure 9. Combined Residual Diagnostics for Training and Test Sets, (A) Histogram of training set residuals with normal fit, (B) Training residuals vs. predicted values with LOWESS smoothing, (C) Histogram of test set residuals with normal fit, (D) Training residuals vs. predicted values with LOWESS smoothing.

Supplementary Figure 9 comprehensively validates the error characteristics and predictive robustness of the pruned LASSO model. Supplementary Figure 9A and C display residual histograms with superimposed normal distribution curves (red dashed lines), quantifying minimal deviation from Gaussian assumptions (Supplementary Figure 9A, training residuals: $\mu = 1.77 \times 10^{-5}$, $\sigma = 6.55 \times 10^{-4}$; Supplementary Figure 9C, test residuals: $\mu = 1.46 \times 10^{-5}$, $\sigma = 5.96 \times 10^{-4}$). In the residual vs. predicted value scatter plots, the LOWESS trend line slopes for the training and testing sets are 0.027 and 0.022, respectively, with residuals equally

distributed around the zero line (Supplementary Figure 9B, range: 0.28 to +0.32 for training dataset, Supplementary Figure 9D, range: 0.28 to 0.32 for testing dataset). This indicates stable error variance (homoscedasticity) and no systematic overestimation or underestimation. In summary, the model demonstrates no significant systematic bias within the data coverage range, with well-controlled error fluctuations.



Supplementary Figure 10. Bootstrap Stability Analysis: Error Distributions and Model Explanatory Power, (A) RMSE distribution (mean=0.039, 95% CI [5.30×10^{-4} – 6.50×10^{-4}]), (B)MAE distribution (range=0.005, 95% CI [4.10×10^{-4} – 4.95×10^{-4}]), (C)R² distribution (median=0.99)

To quantify robustness against data variability, 500 bootstrap iterations were conducted. Supplementary Figure 10 displays the distributions of (a)RMSE (blue), (b)MAE (orange), and (c)R² (green) calculated from 500 resampled iterations. RMSE: Tightly distributed around 6.0×10^{-4} (95% CI(Confidence Interval): 5.30×10^{-4} – 6.50×10^{-4}), reflecting minimal prediction deviation. MAE: Narrow range (4.10×10^{-4} – 4.95×10^{-4} , 95% CI). R²: Values cluster near 0.99 (95% CI: 0.9949–0.9969), indicating near-perfect explanatory power.