Supplementary Information (SI) for RSC Advances.
This journal is © The Royal Society of Chemistry 2026

Machine Learning Pipeline Workflow

Stage 1 — **SCAPS-1D Data Generation**

- Systematic variation of 7 parameters
- 3 device architectures: $TlPbI_3$-CdS, $TlPbI_3$-In_2S_3, $TlPbI_3$-SnS_2
- Target: Power Conversion Efficiency (PCE)

Stage 2 — **Data Preprocessing**

- Feature matrix ($\mathbf{X}$) & target ($\mathbf{y}$) separation
- StandardScaler normalization
- Outlier screening & data integrity check

Stage 3 — **Feature Selection**

7 Input Features:

- Bandgap
- Electron Affinity
- Doping Concentration
- Thickness
- Defect Density
- Electron Mobility
- Hole Mobility

Stage 4 — **Model Training (4 Algorithms)**

- Random Forest Regressor
- Gradient Boosting Regressor
- Decision Tree Regressor
- LightGBM Regressor
- Fixed `random_state=42` for reproducibility

Stage 5 — **Validation & Evaluation**

- 5-Fold Cross-Validation
- Metrics: R^2, RMSE, MAE
- Parity plots (Actual vs. Predicted)

Stage 6 — **Interpretation & Prediction**

- SHAP summary, waterfall & bar plots
- Feature importance ranking per model
- Rapid PCE prediction for new configurations