

Supplementary Information

Supplementary Notes, Results, Code and Prompts

S1 Data source and linear-homopolymer repeat-unit scope

All data were obtained from the NIMS PolyInfo database [1]. The usable experimental records comprise Tg (n = 7291), Td (n = 5571), Tm (n = 3612), EC (n = 863), ϵ -AC (n = 682), and TC (n = 78). The unequal sizes reflect the available experimental literature rather than selective data removal.

The present framework is intentionally applied to linear homopolymer repeat units with two connection sites. This representation gives a direct one-to-one mapping between one repeat-unit structure and its MACCS fingerprint. Copolymer composition or sequence, branched topology, and crosslinked connectivity require additional architectural descriptors and are therefore outside the present representation.

Canonical grouping is used only for the stricter split-robustness analyses. The 7,291 Tg observations correspond to 7,163 unique canonical repeat-unit structures; therefore, 128 observations are additional measurements associated with structures already represented elsewhere in the dataset. This difference does not imply 128 duplicate pairs and does not imply deletion. A structure observed twice contributes one additional observation, a structure observed three times contributes two, and so forth. The workflow performs the split at the canonical-group level and then maps every selected group back to all original records, so all 7,291 Tg measurements remain available for evaluation. The analogous EC audit retains all 863 records while grouping them into 850 canonical structures.

Reference

[1] Otsuka, S. et al. PoLyInfo: polymer database for polymeric materials design. In 2011 International Conference on Emerging Intelligent Data and Web Technologies 22–29 (IEEE, 2011).

S2 Industrial standard for performance threshold partitioning

S2.1 Engineering rationality of thermal performance thresholds

For the glass transition temperature Tg, the thresholds of 30%, 20%, and 10% correspond to 204 °C, 238 °C, and 275 °C for interpolation dataset, respectively, each possessing clear engineering significance. A Tg of 204 °C approximates that of epoxy resin-based composite materials, suitable for standard electronic packaging [1]. The 238 °C threshold satisfies the heat-resistance requirements for applications under automotive engine covers and aligns with high-performance polymers such as polyimide [2]. The 275 °C threshold reaches the aerospace-grade level of polymers such as PEEK [3]. It is noteworthy that the corresponding thresholds in the extrapolation dataset, 266 °C, 300 °C, and 334 °C, are systematically higher, indicating that chemical spatial edges tend to include high-performance but structurally unique materials.

For thermal conductivity TC, the thresholds of 0.26 W/(m·K), 0.32 W/(m·K), and 0.40 W/(m·K) correspond to critical benchmarks in polymer thermal management. A TC of 0.26 W/(m·K) characterizes amorphous polymers such as PMMA and PS [4], 0.32 W/(m·K) represents the level of engineering plastics like nylon 66 [5], while 0.40 W/(m·K) satisfies the minimum requirement for thermal interface materials [6].

The melting temperature T_m thresholds, 274 °C, 284 °C, and 340 °C, map to representative industrial polymers: PET (~260 °C), PBT (~280 °C), and liquid-crystal polymers (>330 °C) [7].

For decomposition temperature T_d , the 472, 500, and 537 °C thresholds represent progressively more demanding thermal-stability tiers. The lower and middle thresholds are consistent with engineering thermosets/plastics and high-temperature polyimides discussed in the cited reviews [8–9]. The 537 °C tier is retained as a data-defined extreme threshold rather than assigned to one named commercial polymer without direct supporting evidence.

S2.2 Application orientation of electrical performance thresholds

For the high-EC direction, the thresholds from $10^{-2.3}$ to $10^{0.9}$ S m⁻¹ are treated as semiconducting/conducting-polymer tiers rather than dielectric-insulation targets. Conjugated and doped polymers can span semiconducting through metallic-like transport regimes, depending strongly on backbone, doping, morphology, and processing [10–12]. The purpose of these tiers is to test bidirectional screening and potential charge-transport applications such as flexible electronics and sensors, not to imply that high conductivity is desirable for an insulator.

The corresponding ϵ -AC thresholds of 4.47, 5.01, 6.03, and 9.33 represent increasing permittivity within the available records. Higher permittivity can be useful in polarization- or capacitance-oriented materials, but ϵ -AC alone does not establish low dielectric loss, high breakdown strength, or superior insulation. EC-to- ϵ -AC transfer is therefore evaluated as statistical sharing of charge-response-related structural information rather than as a universal dielectric-design objective.

For the low-EC direction, thresholds from approximately $10^{-10.8}$ to 10^{-14} S m⁻¹ define increasingly stringent insulation tiers. These ranges are consistent with general polymer insulation, cable insulation, and high-voltage nanodielectric contexts [13–15]. They are used as progressively stricter data-defined screening boundaries rather than direct certification against a particular device standard.

The associated ϵ -AC values of 2.69-2.29 fall in a low-permittivity range relevant to microelectronic and fluoropolymer-based dielectric applications [16–17]. Low permittivity does not by itself guarantee low loss or high breakdown strength, which are not sufficiently populated in the present database. Collectively, the thermal and electrical tiers provide reproducible percentile-defined screening tasks with application context; they should not be interpreted as universal industrial standards or as simultaneous optimization of all dielectric properties.

Reference

- [1] Auvergne, R. et al. Biobased thermosetting epoxy: present and future. *Chem. Rev.* 114, 1082–1115 (2014).
- [2] Liaw, D.-J. et al. Advanced polyimide materials: Syntheses, physical properties and applications. *Progress in Polymer Science* 37, 907–974 (2012).
- [3] Díez-Pascual, A. M. et al. High-performance nanocomposites based on polyetherketones. *Prog. Mater. Sci.* 57, 1106–1190 (2012).
- [4] Mark, J. E. *Polymer Data Handbook*: Oxford University Press. New York (1999).
- [5] Choy, C. L. Thermal conductivity of polymers. *Polymer* 18, 984–1004 (1977).
- [6] Prasher, R. Thermal interface materials: historical perspective, status, and future directions. *Proc. IEEE* 94, 1571–1586 (2006).
- [7] Wunderlich, B. *Thermal analysis of polymeric materials*. Berlin, Heidelberg: Springer Berlin Heidelberg. (2005).
- [8] Levchik, S. V. & Weil, E. D. Thermal decomposition, combustion and flame-retardancy of epoxy resins—a review of the recent literature. *Polymer International* 53, 1901–1929 (2004).
- [9] Liaw, D.-J. et al. Advanced polyimide materials: Syntheses, physical properties and applications. *Progress in Polymer Science* 37, 907–974 (2012).
- [10] Sirringhaus, H. et al. Two-dimensional charge transport in self-organized, high-mobility conjugated polymers. *Nature* 401, 685–688 (1999).
- [11] MacDiarmid, A. G. "Synthetic metals": a novel role for organic polymers. *Current applied physics* 1, 269–279 (2001).
- [12] Heeger, A. J. Semiconducting and metallic polymers: the fourth generation of polymeric materials. *The Journal of physical chemistry B* 105, 8475–8491 (2001).
- [13] Dissado, L. A. & C, J. Fothergill. *Electrical degradation and breakdown in polymers*. Vol. 9. Iet. (1992).
- [14] Pleșa, I. et al. Polyethylene nanocomposites for power cable insulations. *Polymers* 11, 24 (2018).
- [15] Anandraj, J. & Joshi, G. M. Fabrication, performance and applications of integrated nanodielectric properties of materials—a review. *Composite Interfaces* 25, 455–489 (2018).
- [16] Maier, G. Low dielectric constant polymers for microelectronics. *Progress in polymer science* 26, 3–65 (2001).
- [17] Drobny, J. G. *Technology of fluoropolymers*. CRC Press. (2008).

S3 Sliding Threshold Correlation Analysis Positive & Negative Substructures

S3.1 Thermal substructures

The positive and negative substructures for STCA are derived from the Tg inter dataset, which contains a total of 5833 entries with Tg ranging from -123.0°C to 495.0°C. There are 53 positive and 96 negative substructures are found in total. The Figure S1 displays the MACCS keys corresponding to the top 1~12 positive and negative substructures and their linear correlation with changes in the Tg sliding threshold, where **r** represents the linear correlation coefficient between the sliding threshold and the distribution frequency of specific molecular substructures, while **p** indicates the p-value used to determine if the linear correlation is significant, with a threshold set at 0.05.

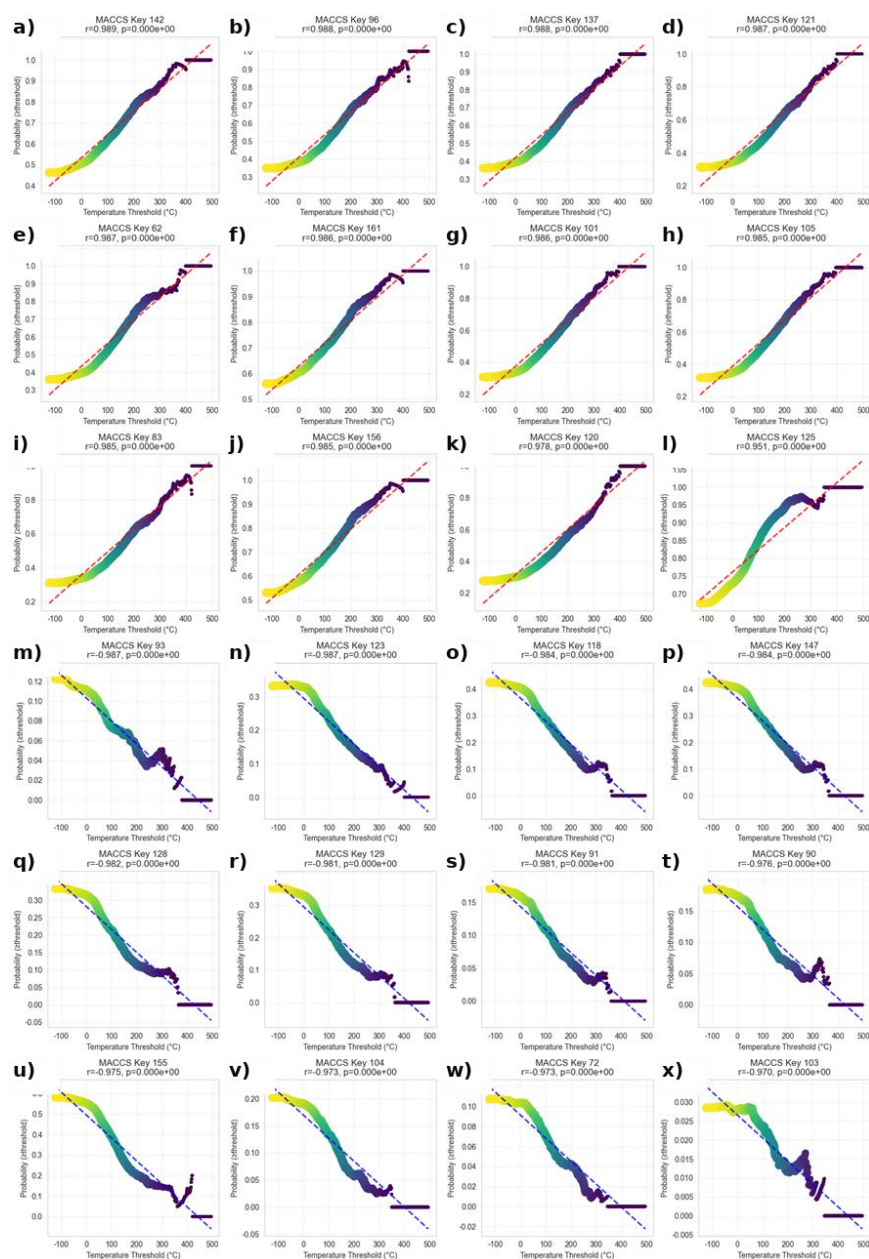


Figure S1 | Thermal STCA trajectories show strong positive and negative relationships between MACCS-key prevalence and the moving Tg threshold. a-l) The prevalence of the twelve highest-ranked positive Molecular ACCess System (MACCS) keys increases as the glass-transition-temperature (Tg) threshold is raised. m-x) The prevalence of the twelve highest-ranked negative keys decreases over the same scan. Each colored point denotes one threshold step, progressing from yellow at the lowest threshold to purple at the highest threshold. Red dashed lines denote positive least-squares fits and blue dashed lines denote negative fits. r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

S3.2 Electrical substructures

The positive and negative substructures identified by STCA were derived from the EC interpolation dataset, which comprises 691 entries with $\lg(\text{EC})$ values from -18.5 to 9.3. In total, 23 positive and 113 negative substructures were identified. Figures S2 and S3 report the twenty highest-ranked positive and negative MACCS keys, respectively, together with their correlations across the reverse sliding scan. Here, r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

Electrical-conductivity values are expressed as $\lg(\text{EC})$. Because records with positive $\lg(\text{EC})$ are sparse, the nominal correlations obtained from the forward high-EC scan deviate visibly from linearity and are not used as the primary electrical rules. Figures S4 and S5 report the corresponding positive and negative forward-scan patterns. In contrast, the reverse scan is strong and approximately linear; therefore, the low-EC combinations are presented directly, and the high-EC interpretation interchanges the positive and negative members of the same predeclared family. The horizontal axes in Figures S2 and S3 use the reverse $\lg(\text{EC})$ threshold.

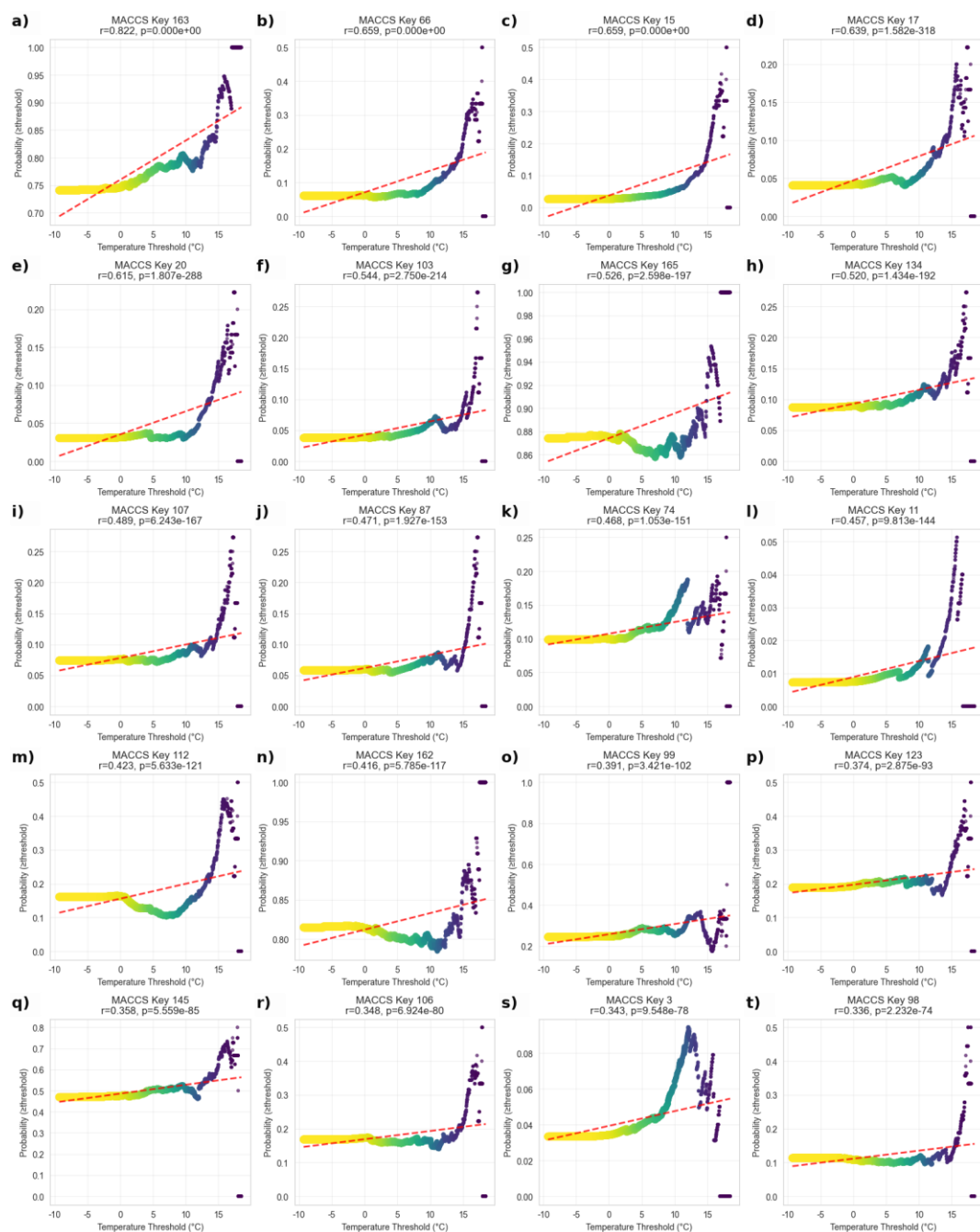


Figure S2 | The reverse electrical-conductivity scan identifies MACCS keys whose prevalence increases toward more stringent low-conductivity thresholds. a–t) The twenty highest-ranked positive Molecular ACCess System (MACCS) keys are shown for the EC-interpolation dataset. The horizontal axis is the reverse sliding threshold for the base-10 logarithm of electrical conductivity, $\lg(\text{EC})$, and the vertical axis is the fraction of records containing the key. Colored points progress from yellow to purple over the scan. Red dashed lines are least-squares fits. r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

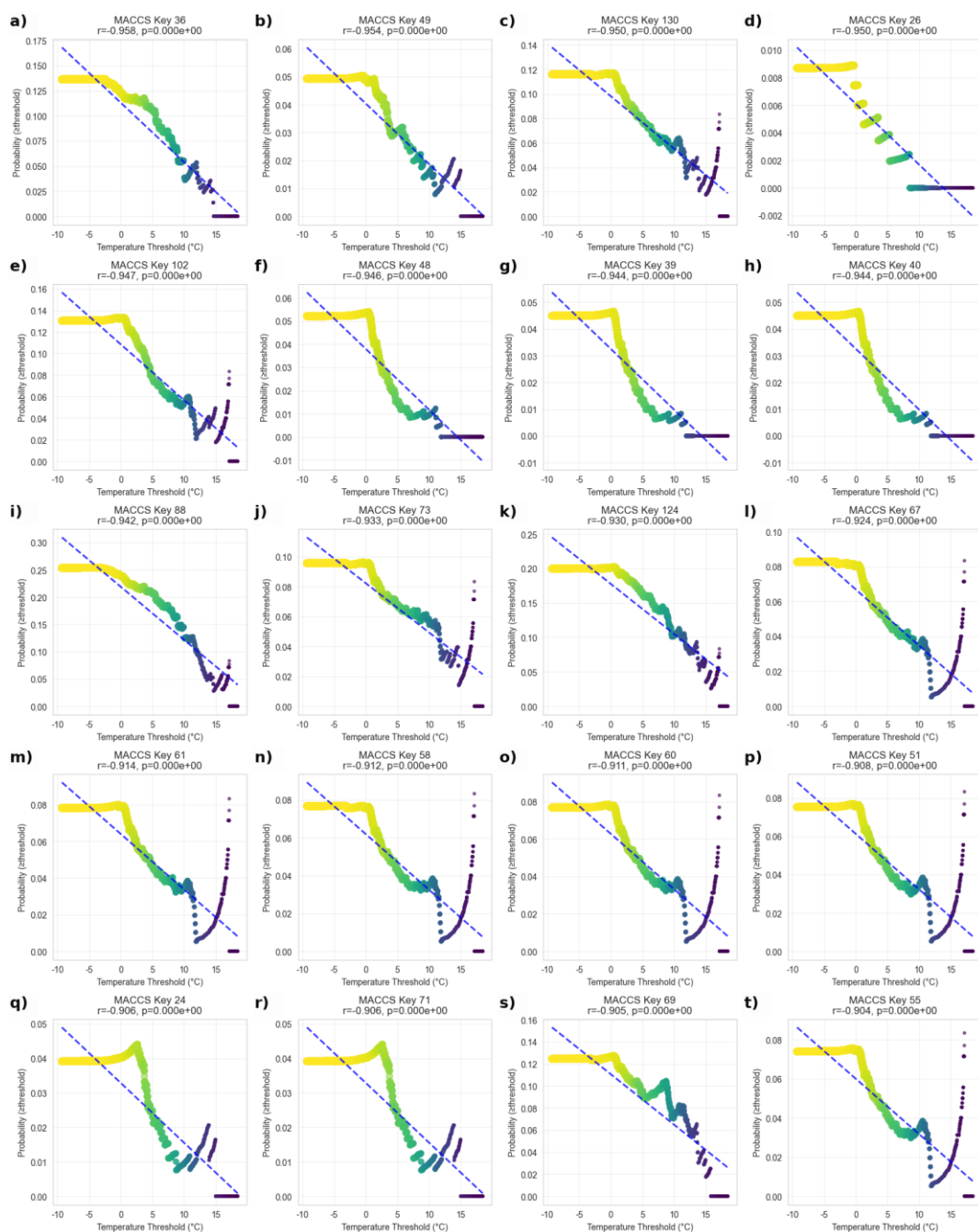


Figure S3 | The reverse electrical-conductivity scan identifies MACCS keys whose prevalence decreases toward more stringent low-conductivity thresholds. a–t) The twenty highest-ranked negative Molecular ACCess System (MACCS) keys are shown for the EC-interpolation dataset. The horizontal axis is the reverse sliding threshold for the base-10 logarithm of electrical conductivity, $\lg(\text{EC})$, and the vertical axis is the fraction of records containing the key. Colored points progress from yellow to purple over the scan. Blue dashed lines are least-squares fits. r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

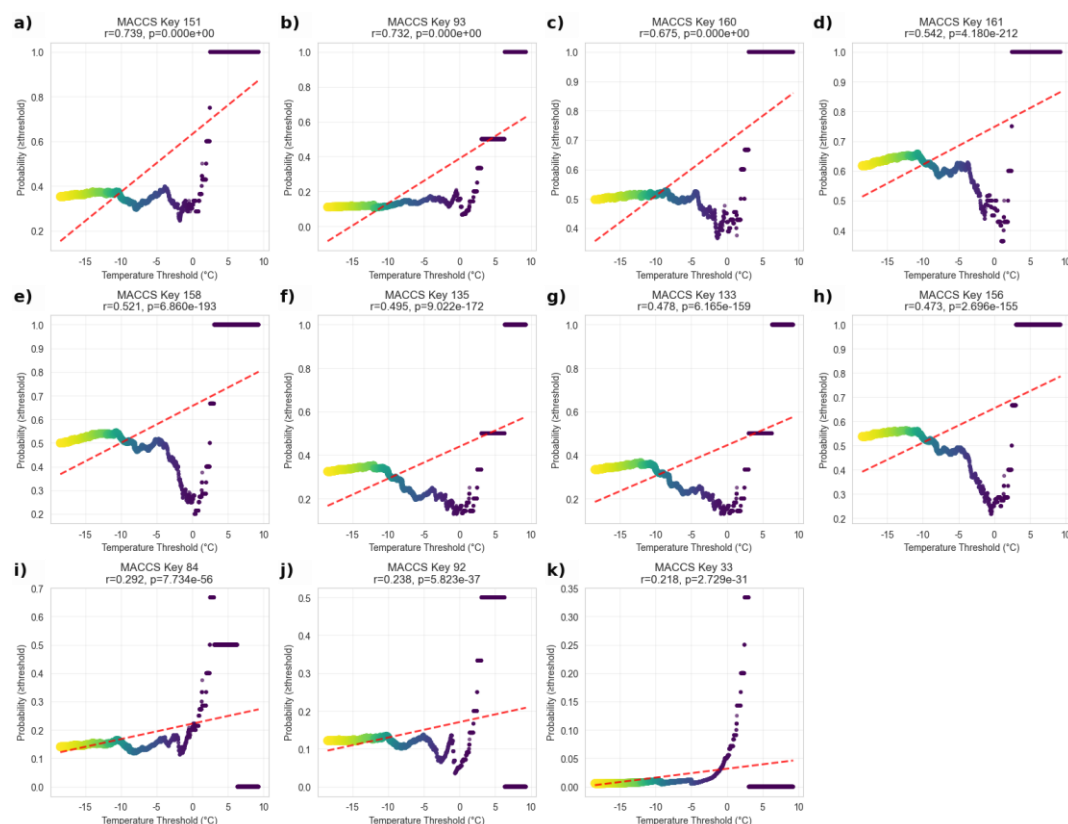


Figure S4 | The sparse high-conductivity forward scan yields weak positive correlations that are not used as the primary electrical rules. a–k) The eleven nominally positive Molecular ACCess System (MACCS) keys obtained from the forward scan of the base-10 logarithm of electrical conductivity, $\lg(\text{EC})$, are shown. Colored points progress from yellow to purple over the scan, and red dashed lines are least-squares fits. The visible deviations from linearity reflect the sparse positive- $\lg(\text{EC})$ regime. r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

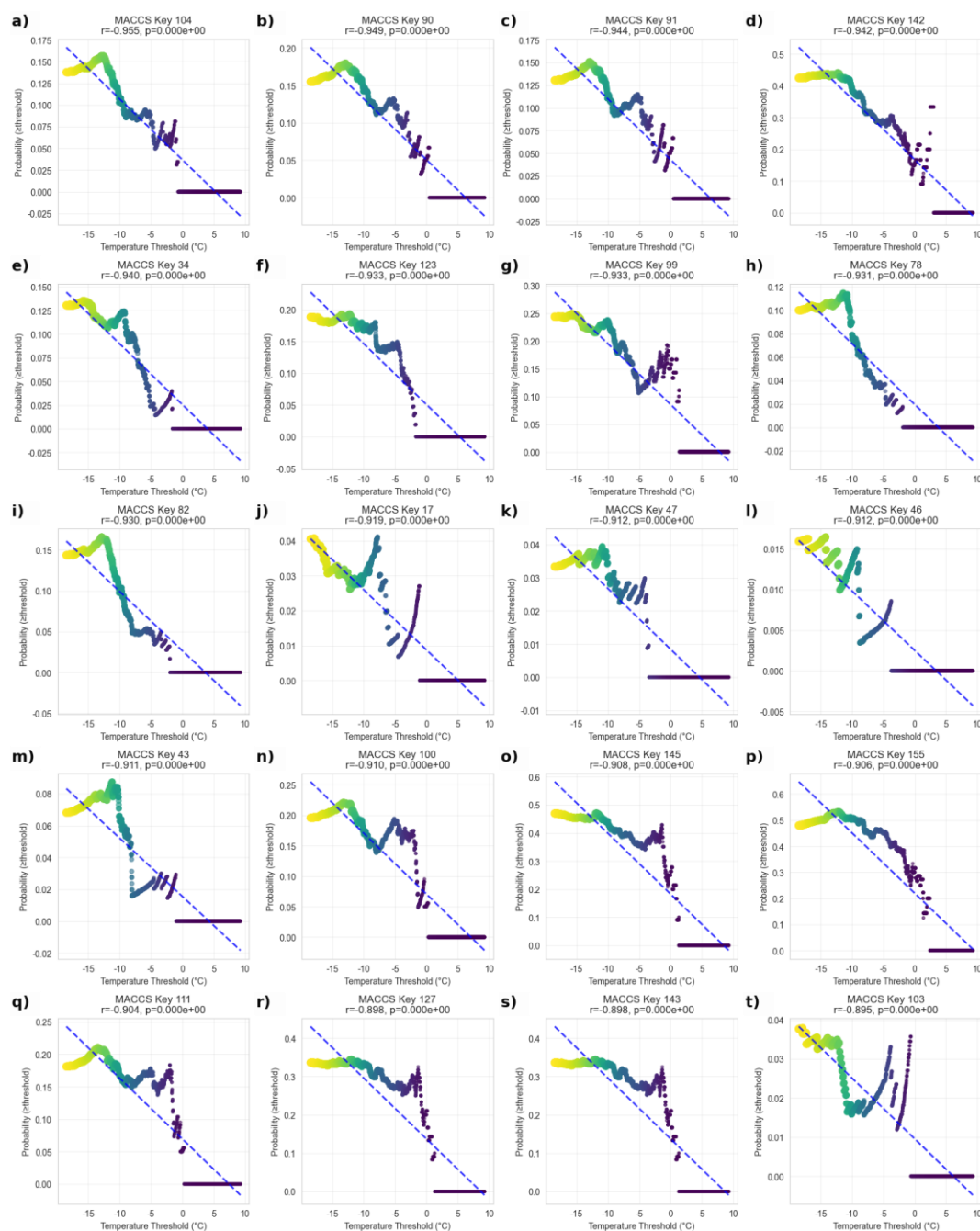


Figure S5 | The sparse high-conductivity forward scan yields weak negative correlations that are not used as the primary electrical rules. a–t) Twenty representative negative Molecular ACCESS System (MACCS) keys obtained from the forward scan of the base-10 logarithm of electrical conductivity, $\lg(\text{EC})$, are shown. Colored points progress from yellow to purple over the scan, and blue dashed lines are least-squares fits. The pronounced nonlinearity and threshold-end plateaus support the use of the reverse scan for the principal electrical rule extraction. r is the Pearson correlation coefficient and p is the two-sided probability value for the null hypothesis of zero linear correlation.

The sparse positive- $\lg(\text{EC})$ regime produces visibly weak and nonlinear forward-scan patterns (Figures S4

and S5), in contrast to the strong reverse-scan relationships in Figures S2 and S3.

S4 XGBoost Model Configuration and Interpretability Analysis

The eXtreme Gradient Boosting (XGBoost) algorithm was employed as an interpretable ensemble learning framework to quantify the contribution of individual structural descriptors to polymer properties. In this study, the model was not primarily used for property prediction but for extracting feature importance metrics that reveal the relationship between MACCS key fingerprints and experimental properties. Two complementary interpretability indices were derived: Gini importance, representing the aggregated gain in model performance attributed to each feature across all trees, and SHAP values, which provide a game-theoretic decomposition of each individual prediction into additive feature contributions. These indices together form the benchmark for STCA, allowing both global and local evaluation of the influence of chemical substructures.

The XGBoost model was implemented using the open-source `xgboost` library. Model training was conducted using the regression objective function, optimized via gradient boosting on decision trees. The training, validation and testing dataset are split from Tg and EC interpolation dataset, with the ratio of 70%, 10% and 20%. We conducted automated hyperparameter tuning using the Optuna optimization framework (v3.4.0) with a Tree-structured Parzen Estimator (TPE) sampler. The objective function minimized the mean squared error (MSE) on the validation set. Optimized trial range is set from 50~100 for efficient search. Fifty independent trials were executed for each model, with garbage collection enabled between trials to reduce memory overhead. The search ranges were defined in Table S1. All models were trained with a fixed random seed (`random_state = 0`) to ensure reproducibility.

Table S1. Hyperparameter search range of XGBoost training with Optuna

Parameter	Search range	Description
n_estimators	100-1000	Number of boosting rounds
learning_rate	log-uniform [0.01, 0.3]	Step-size shrinkage
max_depth	3-10	Maximum tree depth
min_child_weight	log-uniform [1, 10]	Minimum child weight
subsample	uniform [0.5, 1.0]	Sample fraction per tree
colsample_bytree	uniform [0.5, 1.0]	Feature fraction per tree
gamma	log-uniform [1e-8, 1.0]	Minimum loss reduction
reg_alpha	log-uniform [1e-8, 1.0]	L1 regularization
reg_lambda	log-uniform [1e-8, 1.0]	L2 regularization

The optimal configuration typically corresponded to moderate tree depth and learning rates, ensuring smooth convergence and interpretability stability. Model fitting achieved reasonable RMSE and R^2 values on both training and test datasets, confirming internal consistency and preventing overfitting. However,

these metrics were reported solely for model validation, as the XGBoost framework in this context serves as a mechanistic interpretation tool rather than a predictive model. The visualized training and testing parity plots of Tg inter and EC inter is shown in Figure S6.

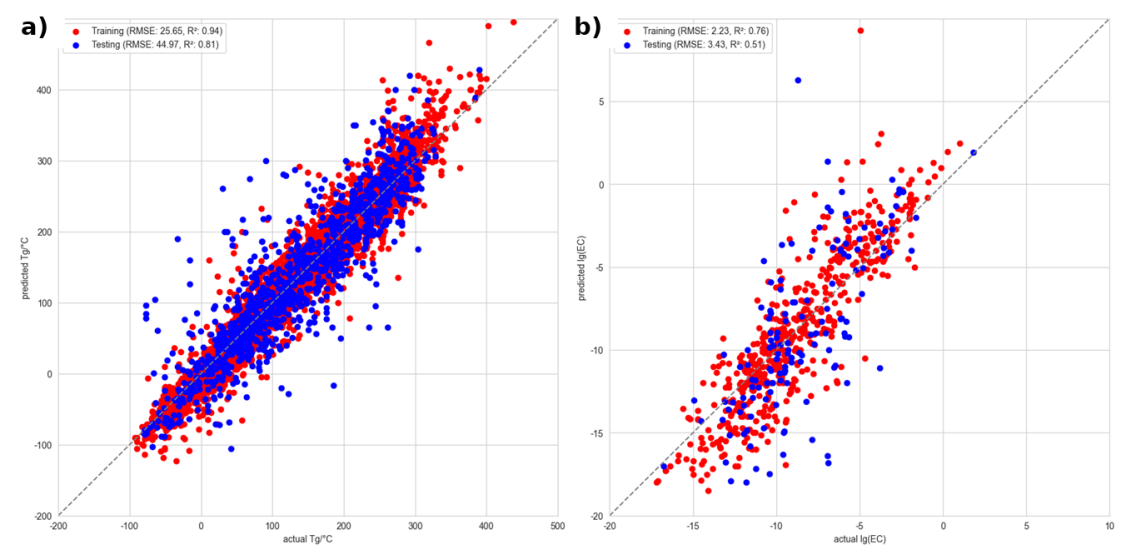


Figure S6 | XGBoost captures the broad Tg and EC trends while retaining nonzero predictive error. a) Predicted versus experimental glass-transition temperature (Tg) is shown in degrees Celsius. b) Predicted versus experimental base-10 logarithm of electrical conductivity, lg(EC), is shown. Red circles denote training records, blue circles denote testing records, and the gray dashed diagonal is the identity line. Root-mean-square error (RMSE) and coefficient of determination (R²) values are reported separately for training and testing in each panel.

S5 Screening results corresponding to Figure 3

To ensure legibility at publication size, this section reports the property-wise average score (AS) and comprehensive analysis (CA) values used in Figure 3. Section S6 reports the corresponding TP/FP recovery diagnostics. A dash indicates that the validity criteria are not satisfied.

S5.1 Thermal screening

Table S2. Tg inter/extra/all results used in Figure 3a.

Threshold	Method	Tg inter AS	Tg extra AS	Tg all AS	CA inter/extra/all
Top 30%	STCA	0.685	0.699	0.743	0.708
	WRAcc	0.708	0.661	0.721	0.695
	Average SHAP	0.291	0.282	0.306	0.293
	SHAP	0.613	0.489	0.588	0.558
	Gini Importance	0.613	0.489	0.588	0.558
Top 20%	STCA	0.584	0.518	0.623	0.572

Threshold	Method	Tg inter AS	Tg extra AS	Tg all AS	CA inter/extra/all
	WRAcc	0.580	0.320	0.530	0.445
	Average SHAP	0.235	0.198	0.239	0.222
	SHAP	0.514	0.356	0.472	0.436
	Gini Importance	0.514	0.356	0.472	0.436
	STCA	0.411	0.221	0.357	0.320
Top 10%	WRAcc	0.436	0.314	0.427	0.383
	Average SHAP	0.148	0.108	0.148	0.132
	SHAP	0.345	0.208	0.302	0.272
	Gini Importance	0.345	0.208	0.302	0.272
	STCA	0.411	0.221	0.357	0.320

Table S3. Thermal cross-property results used in Figure 3b.

Threshold	Method	Tg inter AS	Tg extra AS	Tg all AS	TC AS	Tm AS	Td AS	Total CA
Top 30%	STCA	0.685	0.699	0.743	0.399	0.529	0.596	0.582
	WRAcc	0.708	0.661	0.721	0.318	0.498	0.530	0.529
	Average SHAP	0.291	0.282	0.306	0.339	—	0.276	—
	SHAP	0.613	0.489	0.588	0.290	0.390	0.471	0.444
	Gini Importance	0.613	0.489	0.588	0.290	0.390	0.471	0.444
Top 20%	STCA	0.584	0.518	0.623	0.298	0.555	0.490	0.482
	WRAcc	0.580	0.320	0.530	0.151	0.420	0.376	0.326
	Average SHAP	0.235	0.198	0.239	0.232	—	0.218	—
	SHAP	0.514	0.356	0.472	0.188	0.391	0.384	0.346
	Gini Importance	0.514	0.356	0.472	0.188	0.391	0.384	0.346
Top 10%	STCA	0.411	0.221	0.357	0.320	0.469	0.292	0.326
	WRAcc	0.436	0.314	0.427	0.200	0.448	0.302	0.327
	Average SHAP	0.148	0.108	0.148	0.154	—	0.127	—
	SHAP	0.345	0.208	0.302	0.272	0.309	0.229	0.269
	Gini Importance	0.345	0.208	0.302	0.272	0.309	0.229	0.269

Figure 3 retains the original TC-inclusive thermal scope. The reviewer-requested with-TC and no-TC robustness results are reported separately under canonical-grouped alternative partitions in Section S12 (Table S18); they are not used to overwrite the displayed Figure 3 values.

S5.2 High-EC semiconductor/conduction screening

The high-EC direction represents semiconducting/conducting-polymer screening, not electrical insulation. Table S4 reports EC inter/extra/all results and Table S5 the corresponding ϵ -AC and total cross-property

scope.

Table S4. High-EC inter/extra/all results used in Figure 3a.

Threshold	Method	EC inter AS	EC extra AS	EC all AS	CA inter/extra/all
Top 20%	STCA	0.256	0.327	0.290	0.356
	WRAcc	0.383	0.341	0.374	0.365
	Average SHAP	0.174	0.312	0.194	0.213
	SHAP	0.495	0.402	0.320	0.393
	Gini Importance	0.423	0.407	0.383	0.404
Top 15%	STCA	0.202	0.297	0.239	0.300
	WRAcc	0.321	0.406	0.349	0.355
	Average SHAP	0.132	0.179	0.164	0.156
	SHAP	0.247	0.141	0.214	0.190
	Gini Importance	0.409	0.386	0.376	0.390
Top 10%	STCA	0.224	0.304	0.307	0.273
	WRAcc	0.265	0.156	0.225	0.205
	Average SHAP	0.095	0.107	0.094	0.098
	SHAP	0.106	0.109	0.091	0.130
	Gini Importance	0.270	0.160	0.199	0.200
Top 5%	STCA	0.227	0.164	0.203	0.194
	WRAcc	0.330	0.199	0.282	0.258
	Average SHAP	0.093	0.112	0.105	0.103
	SHAP	0.276	0.058	0.185	0.114
	Gini Importance	0.133	0.133	0.104	0.264

Table S5. High-EC cross-property results used in Figure 3b.

Threshold	Method	EC inter AS	EC extra AS	EC all AS	ϵ -AC AS	Total CA
Top 20%	STCA	0.256	0.327	0.290	0.273	0.284
	WRAcc	0.383	0.341	0.374	—	—
	Average SHAP	0.174	0.312	0.194	0.270	0.225
	SHAP	0.495	0.402	0.320	—	—
	Gini Importance	0.423	0.407	0.383	0.397	0.402
Top 15%	STCA	0.202	0.297	0.239	0.253	0.243
	WRAcc	0.321	0.406	0.349	0.121	0.239
	Average SHAP	0.132	0.179	0.164	0.232	0.170
	SHAP	0.247	0.141	0.214	—	—
	Gini Importance	0.409	0.386	0.376	0.367	0.384
Top 10%	STCA	0.224	0.304	0.307	0.113	0.201
	WRAcc	0.265	0.156	0.225	—	—

Threshold	Method	EC inter AS	EC extra AS	EC all AS	ϵ -AC AS	Total CA
Top 5%	Average SHAP	0.095	0.107	0.094	0.191	0.112
	SHAP	0.106	0.109	0.091	0.168	0.112
	Gini Importance	0.270	0.160	0.199	0.349	0.224
	STCA	0.227	0.164	0.203	0.148	0.180
	WRAcc	0.330	0.199	0.282	0.057	0.138
	Average SHAP	0.093	0.112	0.105	0.081	0.096
	SHAP	0.276	0.058	0.185	—	—
	Gini Importance	0.133	0.133	0.104	0.052	0.091

S5.3 Low-EC insulation screening

Table S6. Low-EC inter/extra/all results used in Figure 3a.

Threshold	Method	EC inter AS	EC extra AS	EC all AS	CA inter/extra/all
Bottom 20%	STCA	0.313	0.572	0.355	0.403
	WRAcc	0.444	—	0.367	—
	Average SHAP	—	—	—	—
	SHAP	0.163	0.433	0.236	0.236
	Gini Importance	0.223	0.235	0.226	0.228
Bottom 15%	STCA	0.262	0.535	0.304	0.355
	WRAcc	0.386	0.294	0.350	0.339
	Average SHAP	—	—	—	—
	SHAP	0.146	0.394	0.219	0.215
	Gini Importance	0.174	0.213	0.178	0.192
Bottom 10%	STCA	0.224	0.404	0.266	0.280
	WRAcc	0.300	—	0.177	—
	Average SHAP	—	—	—	—
	SHAP	0.085	0.258	0.188	0.143
	Gini Importance	0.124	0.134	0.126	0.159
Bottom 5%	STCA	0.152	0.260	0.186	0.190
	WRAcc	0.264	—	0.206	—
	Average SHAP	—	—	—	—
	SHAP	0.048	0.164	0.108	0.083

Threshold	Method	EC inter AS	EC extra AS	EC all AS	CA inter/extra/all
	Gini Importance	0.079	0.104	0.059	0.077

Table S7. Low-EC cross-property results used in Figure 3b.

Threshold	Method	EC inter AS	EC extra AS	EC all AS	ϵ -AC AS	Total CA
Bottom 20%	STCA	0.313	0.572	0.355	0.228	0.329
	WRAcc	0.444	—	0.367	0.329	—
	Average SHAP	—	—	—	—	—
	SHAP	0.163	0.433	0.236	—	—
	Gini Importance	0.223	0.235	0.226	0.212	0.224
Bottom 15%	STCA	0.262	0.535	0.304	0.169	0.268
	WRAcc	0.386	0.294	0.350	0.160	0.265
	Average SHAP	—	—	—	—	—
	SHAP	0.146	0.394	0.219	—	—
	Gini Importance	0.174	0.213	0.178	0.161	0.180
Bottom 10%	STCA	0.224	0.404	0.266	0.092	0.186
	WRAcc	0.300	—	0.177	0.195	—
	Average SHAP	—	—	—	—	—
	SHAP	0.085	0.258	0.188	—	—
	Gini Importance	0.124	0.134	0.126	0.116	0.124
Bottom 5%	STCA	0.152	0.260	0.186	—	—
	WRAcc	0.264	—	0.206	0.083	—
	Average SHAP	—	—	—	—	—
	SHAP	0.048	0.164	0.108	—	—
	Gini Importance	0.079	0.104	0.059	0.061	0.072

S6 Raw-split TP/FP recovery corresponding to Figure 3

This section reports confusion-count recovery for the fixed original all-record split used in Figure 3. It is kept adjacent to the Figure 3 AS/CA tables so that the raw record-level evidence is not conflated with the canonical-grouped t-SNE grid or the independently regenerated alternative partitions in Sections S11 and S12.

TP/FP is shown as the odds form of precision and is used only as a recovery diagnostic, not as an additional independent performance metric. Thermal no-TC counts aggregate the source-locked external Tg-extra, Td, and Tm evaluations; the with-TC panel adds the TC records. Electrical families are fixed from EC before held-out EC or ε -AC evaluation.

Table S8. Thermal external-record recovery on the fixed raw split.

Tier	Method	TP no-TC	FP no-TC	TP/FP	TC TP	CA no-TC	CA with-TC
Top 30%	STCA	1822	978	1.863	6	0.636	0.579
	WRAcc	2042	1547	1.320	5	0.610	0.543
Top 20%	STCA	1370	1430	0.958	4	0.547	0.480
	WRAcc	1482	1776	0.834	3	0.546	0.444
Top 10%	STCA	477	1293	0.369	2	0.340	0.312
	WRAcc	697	2000	0.348	2	0.379	0.339

Table S9. Low-EC insulation EC-extra recovery on the fixed raw split.

Tier	Method	Valid family	EC-extra TP	FP	TP/FP	Family CA
Bottom 20%	STCA	4	31	30	1.033	0.342
	WRAcc	0	—	—	—	—
Bottom 15%	STCA	4	28	33	0.848	0.295
	WRAcc	6	23	32	0.719	0.317
Bottom 10%	STCA	3	18	43	0.419	0.232
	WRAcc	3	15	27	0.556	0.245
Bottom 5%	STCA	2	10	51	0.196	0.135
	WRAcc	0	—	—	—	—

Table S10. EC-derived family transfer to ε -AC on the fixed raw split.

Regime	Tier	Method	Valid ε members	ε TP	ε FP	TP/FP	Family CA
High EC	Top 20%	STCA	8	44	81	0.543	0.311
		WRAcc	1	35	128	0.273	0.227
	Top 15%	STCA	8	36	89	0.404	0.273
		WRAcc	4	27	137	0.197	0.188

Regime	Tier	Method	Valid ϵ members	ϵ TP	ϵ FP	TP/FP	Family CA
Low EC / insulation	Top 10%	STCA	8	26	99	0.263	0.234
		WRAcc	4	8	16	0.500	0.273
	Top 5%	STCA	4	3	11	0.273	0.166
		WRAcc	8	4	24	0.167	0.167
	Bottom 20%	STCA	3	65	128	0.508	0.342
		WRAcc	8	56	74	0.757	—
	Bottom 15%	STCA	4	50	143	0.350	0.295
		WRAcc	8	46	121	0.380	0.317
	Bottom 10%	STCA	2	26	104	0.250	0.232
		WRAcc	8	33	144	0.229	0.245
	Bottom 5%	STCA	5	17	176	0.097	0.135
		WRAcc	8	8	83	0.096	—

The recovery tables do not imply a universal numerical ordering between STCA and WRAcc. WRAcc remains competitive in several moderate or strict cells; the STCA-specific evidence is limited to the rule-family availability and recovery conditions explicitly shown.

S7 Advantageous molecular substructure combinations

S7.1 Thermal properties performance

For STCA, the optimal substructure combinations vary adaptively with performance thresholds. For Tg, the *top 1 pos + top 8 neg* combination exhibits the highest screening performance at the top 30% and 20%, whereas *top 7–8 pos + top 3–4 neg* becomes optimal at the top 10%. For Tg + TC, *top 1 pos + top 3–4 neg* performs best at 30%, *top 1 pos + top 8 neg* is more effective at 20%, and *top 4 pos + top 3–4 neg* dominates at 10%. For comprehensive screening of Tg + TC + Td + Tm, *top 1 pos + top 8 neg* again shows superior performance at 30% and 20%, while *top 4 pos + top 3–4 neg* yields the best results at 10%. These trends demonstrate that STCA dynamically balances positive and negative substructure contributions across varying screening stringencies.

For average SHAP, only the *top 5 any present* combination yields effective screening for Tg and Tg + TC, whereas no valid combination is found for the comprehensive case.

For SHAP-based screening, effective results are obtained across Tg, Tg + TC, and comprehensive screening, but only the *top 1 neg* combination consistently exhibits meaningful performance among all candidates.

For Gini importance, the *top 1 all present* and *top 3 weighted threshold* combinations consistently show

superior performance for Tg and the comprehensive screening from 30% to 10%. For Tg + TC, ***top 5 weighted threshold*** performs best at 30% and 20%, while ***top 1 all present*** and ***top 3 weighted threshold*** become dominant at 10%.

STCA top 7 positive features:

- 142 — Presence of nitrogen atoms; used when the molecule contains more than one nitrogen atom.
- 96 — Five-membered ring; any closed loop of five connected atoms.
- 137 — Heterocyclic atom; any ring atom that is not carbon.
- 121 — Nitrogen atom in a ring; identifies nitrogen-containing heterocycles.
- 62 — Heterocyclic system; identifies atoms forming heterocyclic rings.
- 161 — Nitrogen atom; general presence of nitrogen in the molecule.
- 101 — Large ring (8-membered or more, up to 14 atoms); captures macrocycles or extended ring systems.

STCA top 4 negative features:

- 93 — Non-carbon/non-hydrogen atom connected to a methyl group; identifies QCH3 motifs.
- 123 — Oxygen-carbon-oxygen linkage.
- 118 — Two consecutive methylene groups (CH₂-CH₂), either in chain or ring; ACH₂CH₂A motif.
- 147 — Two consecutive methylene groups in a ring; detects CH₂-CH₂ pairs within cyclic structures.

Average SHAP top 5 features:

- 129 — Two pairs of methylene groups separated by two atoms (ACH₂A-ACH₂A); detects extended CH₂ motifs.
- 99 — Carbon-carbon double bond (C=C); identifies alkenes or unsaturated carbons.
- 87 — Halogen attached to a non-hydrogen atom (F, Cl, Br, I); identifies halogenated fragments.
- 118 — Two consecutive methylene groups (CH₂-CH₂), either in chain or ring; ACH₂CH₂A motif; occurs at least once.
- 105 — Topological pattern captures branched atom connectivity motifs in molecules.

SHAP top 1 negative features:

166 — Unknown fragment; used to flag molecular substructures not captured by any of the predefined MACCS keys. This key serves as a catch-all for rare or atypical fragments.

Gini importance top 5 features:

105 — Branched atom connectivity pattern (A\$A(\$A)\$A); captures branched substructures in molecules.

155 — Non-methylene atom connected to methylene (A!CH2!A); identifies atoms adjacent to CH2 units excluding CH2 itself.

165 — Ring atom; flags any atom that is part of a ring.

145 — Six-membered ring occurring more than once; detects repeated 6-membered ring motifs.

133 — Atom connected to another atom but not nitrogen (A\$A!N); captures connectivity excluding nitrogen atoms.

S7.2 Electrical conduction performance

For STCA, in the case of EC, the **top 1 pos** consistently delivers the best performance from the 20% to 5% thresholds. For the combined EC + ϵ -AC screening, **top 1 neg** is more effective at the 20% to 10% levels, whereas **top 1 pos** becomes optimal at the 10% and 5% levels.

For average SHAP, EC-only and EC + ϵ -AC exhibit identical screening trends. At the 20% threshold, **top 6 majority present** performs best; from 15% to 10%, **top 2 any present** becomes optimal; at 5%, **top 5 majority present** yields the highest performance.

For SHAP, EC screening is dominated by **top 2 neg** from 20% to 10%, while **top 3 neg** becomes optimal at 5%. For EC + ϵ -AC, a valid combination only emerges at the 10% threshold, where **top 4 pos** shows effective performance.

For Gini importance, EC and EC + ϵ -AC follow the same trend from 20% to 10%, with **top 1 present** consistently demonstrating superiority. At the 5% level, EC is best captured by **top 6 any present**, while EC + ϵ -AC remains dominated by **top 1 present**.

STCA top 1 positive feature:

36 — Sulfur atom in a ring; identifies sulfur-containing heterocycles.

STCA top 1 negative feature:

163 — Six-membered ring; flags any closed loop of six connected atoms.

Average SHAP top 6 features:

82 — Methylene group connected to a heteroatom or non-carbon/non-hydrogen atom (ACH2QH); detects CH2 adjacent to heteroatoms.

124 — Non-carbon/non-hydrogen atom connected to another non-carbon/non-hydrogen atom (QQ); identifies heteroatom–heteroatom linkages.

134 — Halogen atom (F, Cl, Br, I); general presence of halogen elements.

140 — Oxygen atoms occurring more than three times; flags molecules with multiple oxygen atoms.

81 — Sulfur atom connected via a specific pattern (SA(A)A); detects sulfur connectivity motifs.

76 — Carbon–carbon double bond with substituents (C=C(A)A); identifies substituted alkenes.

SHAP top 4 positive features:

83 — Heteroatom in a five-membered ring (QAAAA@1); identifies heteroatoms within 5-membered cyclic structures.

96 — Five-membered ring; any closed loop of five connected atoms.

142 — Nitrogen atom; used when the molecule contains more than one nitrogen atom.

15 — Carboxylic acid group (OC(O)O); flags molecules containing carboxyl functional groups.

SHAP top 3 negative features:

13 — Amide or related motif (ON(C)C); identifies nitrogen–oxygen–carbon connectivity.

46 — Bromine atom (Br); flags presence of Br.

98 — Heteroatom in a six-membered ring (QAAAAA@1); identifies heteroatoms within 6-membered cyclic structures.

Gini importance top 6 features:

48 — Oxygen atom connected to a heteroatom with two additional oxygen substituents (OQ(O)O); detects specific oxygen-containing motifs.

102 — Heteroatom connected to oxygen (QO); flags heteroatom–oxygen linkages.

83 — Heteroatom in a five-membered ring (QAAAA@1); identifies heteroatoms within 5-membered rings.

15 — Carboxylic acid group (OC(O)O); flags molecules containing carboxyl functional groups.

39 — Oxygen–sulfur–oxygen motif (OS(O)O); detects sulfonic acid or related structures.

88 — Sulfur atom (S); flags the presence of sulfur in the molecule.

S7.3 Electrical insulation properties

For STCA, for EC, the *top 1 pos* + *top 11–20 neg* combination consistently delivers the highest screening performance across the 20% to 5% thresholds. In comprehensive EC + ϵ -AC screening, *top 11–15 neg* shows superior performance at the 20% and 15% levels, while *top 1 pos* + *top 11–20 neg* becomes optimal at the 10% threshold. No comprehensive screening is performed at the 5% level.

For average SHAP, no effective screening combinations are identified for either EC or EC + ϵ -AC.

For SHAP, EC performance is dominated by *top 1 pos* from 20% to 10%, whereas *top 1 neg* becomes optimal at 5%. For EC + ϵ -AC, no valid screening combinations are obtained at any threshold.

For Gini importance, for EC, *top 1 present* performs best at 20%, followed by *top 3 majority* present at 15% to 10%, and *top 2 weighted threshold* at 5%. For EC + ϵ -AC, *top 1 present* dominates from 20% to 10%. No comprehensive screening is performed at the 5% level.

STCA top 1 positive feature:

163 — Six-membered ring; flags any closed loop of six connected atoms.

STCA top 20 negative features

36 — Sulfur atom in a ring; identifies sulfur-containing heterocycles.

49 — Atom with a formal charge; flags charged atoms in the molecule.

130 — Heteroatom–heteroatom linkage occurring more than once (QQ > 1); identifies repeated heteroatom connections.

26 — Branched alkene connectivity (C\$=C(\$A)\$A); captures substituted C=C motifs.

- 102 — Heteroatom connected to oxygen (QO); flags heteroatom–oxygen linkages.
- 48 — Oxygen connected to a heteroatom with two additional oxygens (OQ(O)O); detects oxygen-rich motifs.
- 39 — Oxygen–sulfur–oxygen motif (OS(O)O); detects sulfonic acid or related structures.
- 40 — Sulfur–oxygen single bond (S–O); identifies S–O connectivity.
- 88 — Sulfur atom (S); general presence of sulfur.
- 73 — Sulfur double-bonded to any atom (S=A); identifies sulfoxides, sulfones, etc.
- 124 — Heteroatom–heteroatom linkage (QQ); identifies heteroatom–heteroatom connections.
- 67 — Heteroatom connected to sulfur (QS); flags heteroatom–sulfur linkages.
- 61 — Sulfur connected via a branched pattern (AS(A)A); detects sulfur connectivity motifs.
- 58 — Heteroatom–sulfur–heteroatom linkage (QSQ); identifies sulfur bridging heteroatoms.
- 60 — Sulfur double-bonded to oxygen (S=O); detects sulfoxides or sulfones.
- 51 — Carbon–sulfur–oxygen motif (CSO); identifies thiocarbonyl or related structures.
- 24 — Nitrogen–oxygen single bond (N–O); flags N–O linkages.
- 71 — Nitrogen connected to oxygen (NO); general N–O connectivity.
- 69 — Heteroatom connected to another heteroatom excluding hydrogen (QQH); identifies partially substituted heteroatom linkages.
- 55 — Oxygen–sulfur–oxygen motif (OSO); detects sulfonate or related structures.

SHAP top 1 positive feature:

- 13 — Amide or related motif (ON(C)C); identifies nitrogen–oxygen–carbon connectivity.

SHAP top 1 negative feature:

- 83 — Heteroatom in a five-membered ring (QAAAA@1); identifies heteroatoms within 5-membered cyclic structures.

Gini importance top 3 features:

48 — Oxygen atom connected to a heteroatom with two additional oxygen substituents (OQ(O)O); detects specific oxygen-containing motifs.

102 — Heteroatom connected to oxygen (QO); flags heteroatom–oxygen linkages.

83 — Heteroatom in a five-membered ring (QAAAA@1); identifies heteroatoms within 5-membered rings.

All the substructures' SMARTS are resourced from <https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py>.

S8 Physical interpretation of rule evolution and cross-property transfer

The structural interpretations in this section are used as mechanistic hypotheses for the statistically selected MACCS combinations. Because each MACCS bit denotes a broad substructure pattern, the interpretation is based on the structures carrying the bit and is not treated as proof that one fragment alone causes a macroscopic property.

S8.1 Thermal-property rule evolution

The Tg families progressively enrich rigid and heterocyclic motifs while excluding flexible methylene-, methyl-, and oxygen-containing motifs. This trend is consistent with reduced segmental mobility and increased cohesive interactions, two established contributors to glass-transition behavior [1–3]. The stricter Top-10% families contain several co-occurring rigid motifs, indicating that the extreme tier is captured by a combination of structural constraints rather than by one isolated fragment.

For Td, exclusion of flexible or chemically labile motifs and enrichment of aromatic or heterocyclic motifs is consistent with delayed bond scission and improved radical stabilization, although the decomposition pathway remains chemistry-specific [3,4]. For Tm, the selected negative rules are plausibly associated with improved repeat-unit regularity and packing, whereas bulky, asymmetric, or flexible motifs can disrupt crystallization [3,5]. These interpretations explain why Tg-derived families can remain informative for Td and Tm without implying a one-to-one causal relation.

The partial transfer from Tg to TC is also physically plausible but is more morphology-sensitive. Molecular simulations and experiments show that extended, stiff polymer backbones can facilitate heat transport along covalent pathways, whereas bends, conformational disorder, side groups, and interchain packing introduce additional scattering and resistance [6–8]. Thus, rigidity-related Tg motifs can contribute to TC screening, but crystallinity, chain orientation, morphology, and processing remain important determinants not represented by the repeat-unit fingerprint.

Taken together, the thermal rule evolution indicates a recurring pattern in the present datasets: moderate tiers can be screened with a compact rigidity/flexibility balance, whereas the strict tier benefits from several complementary rigid motifs. This threshold-dependent organization is the key interpretable advantage of STCA and is not presented as a universal law for all polymers.

S8.2 Electrical-property rule evolution

For high-EC screening, the transition toward sulfur-containing heterocycles at stricter tiers is consistent with the electronic flexibility and polarizability of conjugated sulfur-containing systems [9,10]. The associated EC-to- ϵ -AC transfer is evaluated at the EC-derived family level: different members of one prespecified family may optimize conductivity and AC dielectric response. Shared sensitivity to conjugation, heteroatom chemistry, and polarizability motivates this transfer test, but it does not require one exact rule or establish a universal causal relation between DC conductivity and ϵ -AC.

MACCS key 163 identifies the presence of a six-membered ring and does not by itself distinguish aromatic, saturated, polar, or nonpolar rings. Its opposite roles in the high- and low-EC rankings must therefore be interpreted from the accompanying structures and co-occurring literals rather than from the bit alone. Likewise, key 36 specifically denotes sulfur in a ring; its enrichment in the high-EC direction and exclusion in low-EC families is consistent with, but does not independently prove, the proposed electronic mechanism.

For low-EC screening, the expanding negative family removes sulfur-containing, charged, oxygen-rich, heteroatom–heteroatom, and flexible motifs that can introduce polar or electronically active environments. The additional six-membered-ring literal at stricter tiers acts as a broad scaffold constraint, not as evidence that every matching ring is saturated or completely nonpolar. The resulting family-level rules remain valid over a broader low-EC range than the benchmark families and provide the clearest evidence for STCA under extreme insulation requirements.

The electrical results therefore support threshold-dependent, family-level structural transfer. They show how promotive and suppressive motifs reorganize between conductive and insulating objectives while preserving an auditable connection to the underlying MACCS literals.

Reference

- [1] Williams, M. L., Landel, R. F. & Ferry, J. D. The temperature dependence of relaxation mechanisms in amorphous polymers and other glass-forming liquids. *J. Am. Chem. Soc.* 77, 3701–3707 (1955).
- [2] Sperling, L. H. *Introduction to Physical Polymer Science*. Wiley. (2015).
- [3] Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*. Elsevier. (2009).
- [4] Grassie, N. & Scott, G. *Polymer Degradation and Stabilisation*. Cambridge University Press. (1988).
- [5] Wunderlich, B. *Macromolecular Physics: Crystal Melting*. Academic Press. (2013).

- [6] Henry, A. & Chen, G. High thermal conductivity of single polyethylene chains using molecular dynamics simulations. *Physical Review Letters* 101, 235502 (2008).
- [7] Zhang, T. & Luo, T. Role of chain morphology and stiffness in thermal conductivity of amorphous polymers. *J. Phys. Chem. B* 120, 803–812 (2016).
- [8] Zhang, T., Wu, X. & Luo, T. Polymer nanofibers with outstanding thermal conductivity and thermal stability: fundamental linkage between molecular characteristics and macroscopic thermal properties. *Journal of Physical Chemistry C* 118, 21148–21159 (2014).
- [9] Brédas, J.-L., Chance, R. R., Baughman, R. H. & Silbey, R. Chain-length dependence of electronic and electrochemical properties of conjugated systems. *J. Am. Chem. Soc.* 105, 6555–6559 (1983).
- [10] Roncali, J. Conjugated poly(thiophenes): synthesis, functionalization, and applications. *Chem. Rev.* 92, 711–738 (1992).

S9 Benchmark alignment and WRAcc comparison

WRAcc-guided subgroup discovery is included as the closest task-aligned native-rule comparator [1]. It directly generates supported present/absent conjunctions, whereas Average SHAP, signed SHAP, and Gini Importance require prespecified conversion from feature scores to executable rules. All methods use identical interpolation/extrapolation membership, source-only rule construction, matched primary candidate budgets, frozen extrapolation evaluation, and no external reranking.

Table S11. Benchmark alignment and rule-selection controls.

Method	Native score	Rule form	Primary/sensitivity budget	Selection data	External reranking
STCA	Threshold-frequency correlation	Present + absent conjunctions	8 displayed family members	Interpolation only	No
WRAcc	Weighted relative accuracy	Present + absent conjunctions	Native Top 8; Top 15/30/120 sensitivity	Interpolation only	No
Signed SHAP	Signed attribution ranking	Present + absent after prespecified conversion	8 displayed rules	Interpolation model only	No
Average SHAP	Mean absolute attribution	Presence-only all/any/majority/weighted rules	8 displayed rules	Interpolation model only	No
Gini Importance	Tree split importance	Presence-only all/any/majority/weighted rules	8 displayed rules	Interpolation model only	No

Reference

[1] Lavrač, N., Kavšek, B., Flach, P. A. & Todorovski, L. Subgroup discovery with CN2-SD. J. Mach. Learn. Res. 5, 153–188 (2004).

S10 Rule-search, support, and complexity sensitivity

Rule-pool size, support floor, and conjunction length were varied to confirm that the comparison was not produced by one favorable complexity choice. Table S12 records the prespecified settings. The interpretation is limited to task- and threshold-specific stability; it is not used to claim that STCA numerically exceeds WRAcc under every complexity setting.

Table S12. Rule-search and complexity sensitivity settings.

Method	Sensitivity factor	Evaluated range	Interpretation
STCA scan	scan step	0.5-2.0 property units	Source-selected families retained across tested scan settings
	minimum subset	1-20 records	Rule availability and selected-set masks reported
WRAcc	minimum support	prespecified support grid	Native ranks recomputed from interpolation only
	rule depth	matched literal lengths	Exact and ± 1 -literal complexity comparisons reported
	candidate pool	Top 8, 15, 30, and 120	Primary Top 8 conclusion checked against larger pools

S11 Canonical-grouped t-SNE sensitivity and all-benchmark raw-versus-canonical audit

Figures 3 and 5 retain the fixed original all-record t-SNE membership generated with `random_state = 42`, `perplexity = 30`, and a 20% held-out fraction. The reviewer-requested audit is stricter: identical canonical repeat-unit structures are indivisible groups, one representative per group is embedded, and every experimental observation is mapped back with its group.

For Tg, 7,291 observations form 7,163 canonical groups; for EC, 863 observations form 850 groups. Grouping changes membership construction but does not delete measurements. Table S13 uses one source-consistent interpolation-selected rule per method and freezes it while replacing the raw membership with the nominal canonical-grouped seed-42/perplexity-30/0.20 membership. This audit includes STCA, WRAcc, Average SHAP, signed SHAP, and Gini Importance.

Table S13 is a membership-sensitivity audit and does not overwrite the fixed Figure 3 display in Section S5. Tables S14-S17 then provide split-specific rule rediscovery for STCA and WRAcc across the complete 27-setting canonical-grouped t-SNE grid and the original-space Morgan distance diagnostics.

Table S13. Source-consistent frozen-rule thermal-all CA under the fixed raw membership and the nominal canonical-grouped membership. The same rule is re-evaluated; no method is reselected against the canonical held-out set.

Method	Top 30% raw	Top 30% canonical	Top 20% raw	Top 20% canonical	Top 10% raw	Top 10% canonical
STCA	0.579	0.582	0.479	0.483	0.288	0.293
WRAcc	0.543	0.548	0.443	0.448	0.336	0.344
Average SHAP	0.265	0.267	0.178	0.180	0.088	0.089
signed SHAP	0.347	0.349	0.255	0.257	0.165	0.168
Gini Importance	0.446	0.451	0.347	0.351	0.269	0.277

Table S14. Canonical-grouped t-SNE robustness averaged over the complete 27-setting seed/perplexity/held-out-ratio grid. STCA and WRAcc are independently rediscovered from each interpolation subset.

Tier	Scope	STCA mean $\pm$ SD	WRAcc mean $\pm$ SD
Top 30%	No-TC	0.600 $\pm$ 0.045	0.613 $\pm$ 0.012
Top 30%	TC-inclusive	0.536 $\pm$ 0.045	0.538 $\pm$ 0.013
Top 20%	No-TC	0.542 $\pm$ 0.026	0.539 $\pm$ 0.024
Top 20%	TC-inclusive	0.437 $\pm$ 0.046	0.427 $\pm$ 0.028
Top 10%	No-TC	0.395 $\pm$ 0.024	0.384 $\pm$ 0.017
Top 10%	TC-inclusive	0.355 $\pm$ 0.016	0.342 $\pm$ 0.012

Table S15. Canonical-grouped thermal STCA sensitivity to t-SNE seed, perplexity, and held-out ratio.

Seed	Perp .	Held- out	NN sim.	NN dist.	no- TC 30	no- TC 20	no- TC 10	with- TC 30	with- TC 20	with- TC 10
0	10	15.0%	0.514	0.486	0.636	0.576	0.428	0.566	0.477	0.382
		20.0%	0.500	0.500	0.618	0.572	0.422	0.554	0.414	0.378
		25.0%	0.500	0.500	0.632	0.569	0.354	0.563	0.482	0.330
	30	15.0%	0.525	0.475	0.632	0.553	0.409	0.563	0.472	0.358
		20.0%	0.508	0.492	0.638	0.553	0.407	0.580	0.484	0.357
		25.0%	0.500	0.500	0.634	0.550	0.404	0.577	0.482	0.356
	50	15.0%	0.567	0.433	0.527	0.520	0.361	0.467	0.484	0.336
		20.0%	0.538	0.462	0.564	0.512	0.390	0.490	0.367	0.356
		25.0%	0.522	0.478	0.553	0.540	0.397	0.483	0.400	0.361
	100	15.0%	0.517	0.483	0.636	0.567	0.411	0.566	0.471	0.370
		20.0%	0.511	0.489	0.631	0.570	0.426	0.563	0.413	0.380
		25.0%	0.512	0.488	0.502	0.516	0.354	0.450	0.368	0.330
42	30	15.0%	0.526	0.474	0.632	0.553	0.393	0.564	0.472	0.348
		20.0%	0.507	0.493	0.631	0.552	0.410	0.575	0.483	0.359
		25.0%	0.493	0.507	0.636	0.562	0.410	0.534	0.387	0.359
	50	15.0%	0.578	0.422	0.587	0.511	0.358	0.555	0.467	0.333
		20.0%	0.561	0.439	0.551	0.518	0.380	0.482	0.369	0.349
		25.0%	0.523	0.477	0.535	0.503	0.397	0.472	0.382	0.360
	100	15.0%	0.515	0.485	0.636	0.570	0.422	0.566	0.413	0.378
		20.0%	0.508	0.492	0.614	0.567	0.419	0.552	0.412	0.376
		25.0%	0.507	0.493	0.503	0.503	0.352	0.423	0.363	0.329
	30	15.0%	0.514	0.486	0.634	0.559	0.404	0.565	0.476	0.355
		20.0%	0.508	0.492	0.631	0.553	0.409	0.575	0.484	0.358
		25.0%	0.500	0.500	0.636	0.552	0.410	0.579	0.483	0.359
100	50	15.0%	0.538	0.462	0.616	0.541	0.394	0.547	0.441	0.349
		20.0%	0.542	0.458	0.571	0.510	0.367	0.533	0.466	0.340
		25.0%	0.519	0.481	0.578	0.492	0.382	0.527	0.377	0.350

Table S16. Canonical-grouped electrical STCA t-SNE sensitivity for high-EC conductivity screening.

Seed	Perp.	Held-out	NN sim.	NN dist.	T20%	T15%	T10%	T5%
0	10	15.0%	0.388	0.612	0.324	0.271	0.226	0.138
		20.0%	0.379	0.621	0.304	0.265	0.223	0.155
		25.0%	0.356	0.644	0.294	0.276	0.236	0.157
	30	15.0%	0.413	0.587	0.323	0.298	0.247	0.173
		20.0%	0.365	0.635	0.310	0.286	0.239	0.112
		25.0%	0.343	0.657	0.391	0.393	0.383	0.162
	50	15.0%	0.421	0.579	0.320	0.301	0.239	0.164
		20.0%	0.365	0.635	0.318	0.300	0.249	0.168
		25.0%	0.336	0.664	0.323	0.298	0.239	0.174
42	10	15.0%	0.358	0.642	0.324	0.290	0.219	—
		20.0%	0.366	0.634	0.317	0.284	0.243	0.175
		25.0%	0.345	0.655	0.315	0.282	0.233	0.166
	30	15.0%	0.390	0.610	0.267	0.260	0.236	0.166
		20.0%	0.345	0.655	0.315	0.298	0.253	0.175
		25.0%	0.356	0.644	0.327	0.302	0.250	0.170
	50	15.0%	0.442	0.558	0.339	0.254	0.192	0.136
		20.0%	0.364	0.636	0.336	0.247	0.192	0.101
		25.0%	0.328	0.672	0.336	0.247	0.191	0.117
100	10	15.0%	0.429	0.571	0.304	0.297	0.208	0.169
		20.0%	0.375	0.625	0.321	0.290	0.243	0.229
		25.0%	0.364	0.636	0.322	0.291	0.244	0.225
	30	15.0%	0.390	0.610	0.324	0.302	0.253	0.178
		20.0%	0.341	0.659	0.320	0.293	0.248	0.176
		25.0%	0.344	0.656	0.323	0.247	0.197	0.123
	50	15.0%	0.453	0.547	0.332	0.296	0.250	0.181
		20.0%	0.387	0.613	0.323	0.283	0.243	0.178
		25.0%	0.340	0.660	0.333	0.246	0.190	0.116

Table S17. Canonical-grouped electrical STCA t-SNE sensitivity for low-EC insulation screening.

Seed	Perp.	Held-out	NN sim.	NN dist.	B20%	B15%	B10%	B5%
0	10	15.0%	0.388	0.612	0.329	0.282	0.218	0.126
		20.0%	0.379	0.621	0.323	0.268	0.204	0.121
		25.0%	0.356	0.644	0.334	0.289	0.232	0.133
	30	15.0%	0.413	0.587	0.343	0.281	0.210	0.123
		20.0%	0.365	0.635	0.330	0.275	0.213	0.130
		25.0%	0.343	0.657	0.325	0.270	0.215	0.128
	50	15.0%	0.421	0.579	0.344	0.294	0.228	0.133
		20.0%	0.365	0.635	0.349	0.303	0.238	0.140
		25.0%	0.336	0.664	0.274	0.222	—	—
42	10	15.0%	0.358	0.642	0.323	0.279	0.215	0.125
		20.0%	0.366	0.634	0.321	0.266	0.200	0.119
		25.0%	0.345	0.655	0.275	0.225	—	—
	30	15.0%	0.390	0.610	0.319	0.276	0.213	0.123
		20.0%	0.345	0.655	0.320	0.270	0.207	0.123
		25.0%	0.356	0.644	0.341	0.281	0.218	0.128
	50	15.0%	0.442	0.558	0.320	0.272	0.210	0.132
		20.0%	0.364	0.636	0.338	0.281	0.216	0.131
		25.0%	0.328	0.672	0.260	0.229	—	—
100	10	15.0%	0.429	0.571	0.334	0.275	0.207	0.126
		20.0%	0.375	0.625	0.329	0.277	0.209	0.121
		25.0%	0.364	0.636	0.326	0.274	0.209	0.120
	30	15.0%	0.390	0.610	0.340	0.279	0.211	0.127
		20.0%	0.341	0.659	0.313	0.260	0.200	0.119
		25.0%	0.344	0.656	0.339	0.281	0.225	0.135
	50	15.0%	0.453	0.547	0.352	0.294	0.228	0.138
		20.0%	0.387	0.613	0.346	0.281	0.214	0.131
		25.0%	0.340	0.660	0.331	0.266	0.202	0.109

Reference

[1] van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

- [2] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. J. Cheminform. 7, 20 (2015).

S12 Full-benchmark robustness under alternative structure-based partitions

All alternative partitions are generated at the canonical-structure group level. Rule discovery and interpolation-only family selection are repeated independently for PCA global distance, Morgan-Tanimoto low density, Butina cluster holdout, Bemis-Murcko scaffold holdout, and canonical grouped-random partitions. No canonical structure crosses train/test, and no held-out score is used to rerank a rule.

The uploaded split-specific calculation files provide direct alternative-partition results for the two native-rule methods, STCA and WRAcc. Attribution-derived methods remain fully reported for the fixed Figure 3 split and the all-benchmark frozen-membership audit in Sections S5 and S11, but they were not recomputed in the alternative-split output files; missing split-specific attribution values are therefore not imputed.

Thermal results are reported separately with and without TC because TC contains only 78 observations. Electrical results use the family-level dual-capability harmonic reported by the v15 structure-split output. The tables preserve ties and WRAcc-favorable conditions; they are not summarized as a universal STCA win.

Table S18. Source-locked thermal CA under each non-t-SNE alternative partition, with no-TC and with-TC scopes reported separately. Cells show mean $\pm$ SD and valid instances/total instances; single-instance splits show the value and 1/1.

Split	Method	Top 30% no-TC	Top 30% with-TC	Top 20% no-TC	Top 20% with-TC	Top 10% no-TC	Top 10% with-TC
PCA global distance	STCA	0.605 (1/1)	0.557 (1/1)	0.524 (1/1)	0.465 (1/1)	0.388 (1/1)	0.345 (1/1)
	WRAcc	0.597 (1/1)	0.507 (1/1)	0.526 (1/1)	0.432 (1/1)	0.365 (1/1)	0.329 (1/1)
Morgan-Tanimoto low density	STCA	0.557 (1/1)	0.453 (1/1)	0.533 (1/1)	0.397 (1/1)	0.425 (1/1)	0.380 (1/1)
	WRAcc	0.590 (1/1)	0.530 (1/1)	0.539 (1/1)	0.378 (1/1)	0.421 (1/1)	0.366 (1/1)
Butina cluster holdout	STCA	0.605 $\pm$ 0.027 (15/15)	0.500 $\pm$ 0.030 (15/15)	0.560 $\pm$ 0.028 (15/15)	0.423 $\pm$ 0.036 (15/15)	0.423 $\pm$ 0.029 (15/15)	0.383 $\pm$ 0.023 (15/15)
	WRAcc	0.628 $\pm$ 0.016 (15/15)	0.544 $\pm$ 0.016 (15/15)	0.569 $\pm$ 0.015 (15/15)	0.413 $\pm$ 0.032 (15/15)	0.416 $\pm$ 0.015 (15/15)	0.360 $\pm$ 0.012 (15/15)
Bemis-Murcko scaffold holdout	STCA	0.604 $\pm$ 0.030 (5/5)	0.506 $\pm$ 0.043 (5/5)	0.574 $\pm$ 0.005 (5/5)	0.427 $\pm$ 0.027 (5/5)	0.433 $\pm$ 0.013 (5/5)	0.385 $\pm$ 0.008 (5/5)
	WRAcc	0.622 $\pm$ 0.004 (5/5)	0.537 $\pm$ 0.002 (5/5)	0.574 $\pm$ 0.021 (5/5)	0.405 $\pm$ 0.027 (5/5)	0.415 $\pm$ 0.014 (5/5)	0.360 $\pm$ 0.007 (5/5)
Canonical grouped-random	STCA	0.620 (1/1)	0.525 (1/1)	0.576 (1/1)	0.477 (1/1)	0.425 (1/1)	0.411 (1/1)
	WRAcc	0.623 (1/1)	0.537 (1/1)	0.568 (1/1)	0.367 (1/1)	0.407 (1/1)	0.357 (1/1)

Table S19. High-EC electrical family-level CA under each non-t-SNE alternative partition.

Split	Method	Top 20%	Top 15%	Top 10%	Top 5%
PCA global distance	STCA	0.321 (1/1)	0.280 (1/1)	0.231 (1/1)	0.172 (1/1)
	WRAcc	—	0.341 (1/1)	0.342 (1/1)	0.207 (1/1)
Morgan–Tanimoto low density	STCA	0.264 (1/1)	0.233 (1/1)	0.171 (1/1)	0.146 (1/1)
	WRAcc	0.251 (1/1)	0.235 (1/1)	0.258 (1/1)	0.120 (1/1)
Butina cluster holdout	STCA	0.283 ± 0.011 (3/3)	0.270 ± 0.011 (3/3)	0.221 ± 0.012 (3/3)	0.168 ± 0.004 (3/3)
	WRAcc	0.268 ± 0.010 (3/3)	0.305 ± 0.034 (3/3)	0.279 ± 0.055 (3/3)	0.118 ± 0.040 (3/3)
Bemis–Murcko scaffold holdout	STCA	0.334 ± 0.031 (5/5)	0.301 ± 0.033 (5/5)	0.256 ± 0.042 (5/5)	0.174 ± 0.037 (5/5)
	WRAcc	0.258 ± 0.026 (4/5)	0.301 ± 0.021 (4/5)	0.313 ± 0.021 (5/5)	0.185 ± 0.037 (5/5)
Canonical grouped-random	STCA	0.302 ± 0.006 (5/5)	0.273 ± 0.008 (5/5)	0.228 ± 0.004 (5/5)	0.165 ± 0.005 (5/5)
	WRAcc	0.256 ± 0.021 (5/5)	0.331 ± 0.006 (5/5)	0.332 ± 0.005 (5/5)	0.206 ± 0.012 (5/5)

Table S20. Low-EC insulation family-level CA under each non-t-SNE alternative partition.

Split	Method	Bottom 20%	Bottom 15%	Bottom 10%	Bottom 5%
PCA global distance	STCA	0.296 (1/1)	0.261 (1/1)	0.213 (1/1)	0.169 (1/1)
	WRAcc	0.320 (1/1)	0.286 (1/1)	0.164 (1/1)	—
Morgan–Tanimoto low density	STCA	0.293 (1/1)	0.232 (1/1)	—	—
	WRAcc	0.355 (1/1)	0.295 (1/1)	0.179 (1/1)	—
Butina cluster holdout	STCA	0.311 ± 0.015 (3/3)	0.253 ± 0.017 (3/3)	0.179 ± 0.011 (2/3)	0.117 ± 0.042 (2/3)
	WRAcc	0.337 ± 0.018 (3/3)	0.285 ± 0.019 (3/3)	0.180 ± 0.016 (2/3)	0.094 (1/3)
Bemis–Murcko scaffold holdout	STCA	0.285 ± 0.011 (5/5)	0.228 ± 0.008 (5/5)	0.156 (1/5)	0.119 (1/5)
	WRAcc	0.344 ± 0.008 (5/5)	0.299 ± 0.016 (5/5)	0.180 ± 0.011 (3/5)	—
Canonical grouped-random	STCA	0.288 ± 0.007 (5/5)	0.226 ± 0.007 (5/5)	0.155 (1/5)	0.127 ± 0.029 (3/5)
	WRAcc	0.325 ± 0.035 (5/5)	0.303 ± 0.018 (5/5)	0.189 ± 0.010 (5/5)	0.094 (1/5)

Table S21. Thermal alternative-split TP/FP recovery without TC. Values are means ± SD across split instances; single-instance partitions are reported directly.

Split	Tier	Method	TP no-TC	FP no-TC	TP/FP no-TC
PCA global distance	Top 30%	STCA	1838.0	1209.0	1.520
		WRAcc	2321.0	1914.0	1.213

Split	Tier	Method	TP no-TC	FP no-TC	TP/FP no-TC
	Top 20%	STCA	1362.0	1599.0	0.852
		WRAcc	1471.0	1907.0	0.771
	Top 10%	STCA	542.0	1430.0	0.379
		WRAcc	652.0	2032.0	0.321
	Top 30%	STCA	1218.0	542.0	2.247
		WRAcc	1614.0	982.0	1.644
	Top 20%	STCA	803.0	478.0	1.680
		WRAcc	1109.0	936.0	1.185
Morgan–Tanimoto low density	Top 10%	STCA	426.0	855.0	0.498
		WRAcc	557.0	1260.0	0.442
	Top 30%	STCA	1387.0 ± 62.6	655.8 ± 164.8	2.223 ± 0.465
		WRAcc	1862.4 ± 133.6	1189.3 ± 228.2	1.608 ± 0.241
	Top 20%	STCA	1015.5 ± 57.9	776.3 ± 174.2	1.361 ± 0.259
		WRAcc	1327.8 ± 56.7	1317.9 ± 228.0	1.030 ± 0.142
	Top 10%	STCA	494.7 ± 32.1	989.0 ± 119.2	0.505 ± 0.052
		WRAcc	639.7 ± 28.2	1548.7 ± 169.6	0.416 ± 0.031
Butina cluster holdout	Top 30%	STCA	1434.6 ± 38.0	706.6 ± 128.9	2.085 ± 0.384
		WRAcc	1983.6 ± 43.8	1389.6 ± 53.1	1.428 ± 0.030
	Top 20%	STCA	1021.4 ± 19.6	686.4 ± 31.0	1.490 ± 0.045
		WRAcc	1315.8 ± 67.2	1244.2 ± 262.9	1.087 ± 0.180
	Top 10%	STCA	514.2 ± 20.6	1017.6 ± 80.3	0.508 ± 0.047
		WRAcc	649.6 ± 17.1	1572.8 ± 106.2	0.414 ± 0.025
Bemis–Murcko scaffold holdout	Top 30%	STCA	1448.0	665.0	2.177
		WRAcc	1999.0	1398.0	1.430
	Top 20%	STCA	1042.0	724.0	1.439
		WRAcc	1294.0	1229.0	1.053
	Top 10%	STCA	558.0	1208.0	0.462
		WRAcc	676.0	1686.0	0.401
Canonical grouped-random	Top 30%	STCA	1448.0	665.0	2.177
		WRAcc	1999.0	1398.0	1.430
	Top 20%	STCA	1042.0	724.0	1.439
		WRAcc	1294.0	1229.0	1.053
	Top 10%	STCA	558.0	1208.0	0.462
		WRAcc	676.0	1686.0	0.401

Table S22. Thermal alternative-split TP/FP recovery with TC included.

Split	Tier	Method	TP with-TC	FP with-TC	TP/FP with-TC
PCA global distance	Top 30%	STCA	1844.0	1213.0	1.520
		WRAcc	2326.0	1920.0	1.211
	Top 20%	STCA	1366.0	1605.0	0.851
		WRAcc	1474.0	1913.0	0.771
	Top 10%	STCA	544.0	1436.0	0.379
		WRAcc	654.0	2038.0	0.321
Morgan–Tanimoto low density	Top 30%	STCA	1221.0	546.0	2.236
		WRAcc	1619.0	986.0	1.642
	Top 20%	STCA	805.0	483.0	1.667
		WRAcc	1111.0	942.0	1.179
	Top 10%	STCA	428.0	860.0	0.498
		WRAcc	559.0	1266.0	0.442
Butina cluster holdout	Top 30%	STCA	1390.7 ± 62.6	659.9 ± 165.0	2.213 ± 0.461
		WRAcc	1867.1 ± 133.9	1193.6 ± 228.3	1.606 ± 0.239
	Top 20%	STCA	1017.8 ± 58.1	781.3 ± 174.3	1.355 ± 0.256
		WRAcc	1330.2 ± 57.0	1324.0 ± 227.9	1.027 ± 0.140
	Top 10%	STCA	496.7 ± 32.1	993.6 ± 119.5	0.505 ± 0.052
		WRAcc	641.7 ± 28.2	1555.0 ± 169.7	0.416 ± 0.031
Bemis–Murcko scaffold holdout	Top 30%	STCA	1438.4 ± 38.7	710.6 ± 128.9	2.078 ± 0.382
		WRAcc	1988.6 ± 43.8	1394.6 ± 53.1	1.427 ± 0.030
	Top 20%	STCA	1023.6 ± 19.7	691.4 ± 31.0	1.482 ± 0.044
		WRAcc	1318.0 ± 67.5	1250.2 ± 262.9	1.084 ± 0.178
	Top 10%	STCA	516.2 ± 20.6	1022.6 ± 80.3	0.507 ± 0.047
		WRAcc	651.6 ± 17.1	1579.0 ± 106.0	0.414 ± 0.025
Canonical grouped- random	Top 30%	STCA	1452.0	669.0	2.170
		WRAcc	2004.0	1403.0	1.428
	Top 20%	STCA	1045.0	729.0	1.433
		WRAcc	1296.0	1236.0	1.049
	Top 10%	STCA	561.0	1213.0	0.462

Split	Tier	Method	TP with-TC	FP with-TC	TP/FP with-TC
		WRAcc	678.0	1692.0	0.401

For Tables S21 and S22, TP and FP are reconstructed from the exact selected count and full-precision value in the source-locked calculation output (TP = selected n $\times$ precision, rounded to the corresponding integer count). No counts are pooled across overlapping split instances for inference; the tables report descriptive per-instance means.

Table S23. Electrical alternative-split rule-family applicability. Each cell counts split-instance $\times$ tier combinations with at least one valid family member or one valid same-rule member.

Split	Regime	Method	Family-valid cells	Same-rule-valid cells
PCA global distance	High EC	STCA	4/4 (100.0%)	4/4 (100.0%)
		WRAcc	3/4 (75.0%)	3/4 (75.0%)
	Low EC	STCA	4/4 (100.0%)	3/4 (75.0%)
		WRAcc	3/4 (75.0%)	3/4 (75.0%)
Morgan–Tanimoto low density	High EC	STCA	4/4 (100.0%)	4/4 (100.0%)
		WRAcc	4/4 (100.0%)	4/4 (100.0%)
	Low EC	STCA	2/4 (50.0%)	2/4 (50.0%)
		WRAcc	3/4 (75.0%)	3/4 (75.0%)
Butina cluster holdout	High EC	STCA	12/12 (100.0%)	12/12 (100.0%)
		WRAcc	12/12 (100.0%)	12/12 (100.0%)
	Low EC	STCA	10/12 (83.3%)	10/12 (83.3%)
		WRAcc	9/12 (75.0%)	9/12 (75.0%)
Bemis–Murcko scaffold holdout	High EC	STCA	20/20 (100.0%)	20/20 (100.0%)
		WRAcc	18/20 (90.0%)	18/20 (90.0%)
	Low EC	STCA	12/20 (60.0%)	11/20 (55.0%)
		WRAcc	13/20 (65.0%)	13/20 (65.0%)
Canonical grouped-random	High EC	STCA	20/20 (100.0%)	20/20 (100.0%)
		WRAcc	20/20 (100.0%)	20/20 (100.0%)
	Low EC	STCA	14/20 (70.0%)	14/20 (70.0%)
		WRAcc	16/20 (80.0%)	16/20 (80.0%)

Reference

[1] Butina, D. Unsupervised data base clustering based on Daylight’s fingerprint and Tanimoto similarity. Journal of Chemical

Information and Computer Sciences 39, 747–750 (1999).

- [2] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* 39, 2887–2893 (1996).
- [3] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *J. Cheminform.* 7, 20 (2015).

S13 Full-pipeline bootstrap confidence intervals and label-permutation tests

The resampling analysis repeats STCA rule discovery, eligible-family construction, and interpolation-AS selection before reporting bootstrap intervals or a maximum-score label-permutation null. The displayed observed score, interval, empirical probability, null threshold, and permutation count therefore arise from one internally consistent repeated pipeline.

The supplied resampling files reproduce the fixed source split used by the reviewer pipeline; they do not encode a canonical-grouped resampling unit and are therefore not relabelled as a canonical-split test. Canonical-grouped robustness is addressed independently in Sections S11 and S12. This section is placed after the alternative-split analysis so that membership robustness and resampling uncertainty remain conceptually separate.

Table S24. Full-pipeline bootstrap intervals and label-permutation tests for STCA-selected rule families. The displayed observed score, interval, empirical probability, null threshold, and permutation count arise from the same repeated discovery-and-selection pipeline.

Scope	Tier	Pipeline statistic	Bootstrap 95% CI	Permutation p	Null 95%	n
Thermal: High Tg / Tg-core	Top 30%	0.700	0.688-0.712	0.0012	0.270	1000
	Top 20%	0.565	0.551-0.579	0.0012	0.196	1000
	Top 10%	0.328	0.317-0.339	0.0012	0.115	1000
Electrical: High EC / EC_core	Top 20%	0.341	0.283-0.414	0.0019	0.231	1000
Electrical: High EC / Electrical_all	Top 20%	0.280	0.265-0.406	0.0032	0.210	1000
Electrical: High EC / EC_core	Top 15%	0.289	0.233-0.388	0.0062	0.193	1000
Electrical: High EC / Electrical_all	Top 15%	0.238	0.225-0.385	0.0035	0.174	1000
Electrical: High EC / EC_core	Top 10%	0.260	0.176-0.282	0.0042	0.154	1000
Electrical: High EC / Electrical_all	Top 10%	0.196	0.178-0.302	0.0037	0.133	1000
Electrical: High EC / EC_core	Top 5%	0.190	0.089-0.249	0.0197	0.137	1000
Electrical: High EC / Electrical_all	Top 5%	0.178	0.097-0.178	0.0063	0.092	1000

Scope	Tier	Pipeline statistic	Bootstrap 95% CI	Permutation p	Null 95%	n
Electrical: Low EC / insulation / EC_core	Bottom 20%	0.399	0.244- 0.444	0.0012	0.251	1000
Electrical: Low EC / insulation / Electrical_all	Bottom 20%	0.327	0.254- 0.372	0.0018	0.230	1000
Electrical: Low EC / insulation / EC_core	Bottom 15%	0.354	0.205- 0.414	0.0025	0.226	1000
Electrical: Low EC / insulation / Electrical_all	Bottom 15%	0.268	0.214- 0.314	0.0018	0.197	1000
Electrical: Low EC / insulation / EC_core	Bottom 10%	0.280	0.154- 0.343	0.0037	0.190	1000
Electrical: Low EC / insulation / Electrical_all	Bottom 10%	—	—	—	—	—
Electrical: Low EC / insulation / EC_core	Bottom 5%	0.185	0.118- 0.263	0.0336	0.171	1000
Electrical: Low EC / insulation / Electrical_all	Bottom 5%	—	—	—	—	—

Reference

- [1] Efron, B. & Tibshirani, R. J. An Introduction to the Bootstrap. Chapman & Hall/CRC. (1993).
- [2] Good, P. I. Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd ed.; Springer. (2005).

S14 Candidate-generation screening and prioritization

All candidates generated in the one-turn audit are passed through the same RDKit parse, exact STCA-rule compliance, canonical uniqueness, synthetic-accessibility, and Polymer Genome trend-screen stages. The funnel and trend results quantify how frequently the prompt yields candidates suitable for subsequent physics-based prioritization. Synthetic accessibility and Polymer Genome are used only as prioritization descriptors [1,2].

Table S25. Candidate-generation audit funnel.

Stage	Count	Rate vs raw
Raw candidates	70	100.0%
RDKit-parseable	67	95.7%
Exact STCA-compliant	54	77.1%
Canonical-unique compliant	44	62.9%
Br-capped SA < 6	44	62.9%

Table S26. Synthetic-accessibility and Polymer Genome trend-screen summary.

Audit item	Result	Interpretation
Canonical-unique compliant candidates	44	After RDKit and exact MACCS constraint checking
Br-capped SA range	1.92-3.64	Synthetic-complexity heuristic only
Numerical chain Eg and Tg predictions	43	Polymer Genome point estimates with reported uncertainty
Point estimates in internal target quadrant	22/43	Eg $\geq$ 3.0 eV and Tg $\geq$ 477.15 K; qualitative trend screen

Reference

- [1] Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J. Cheminform. 1, 8 (2009).
- [2] Doan Tran, H. et al. Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. 128, 171104 (2020).

S15 Supplementary code, prompt design, and one-turn commercial-LLM audit

The frozen prompt combines the standardized instruction below with the target-specific positive and negative MACCS key/SMARTS constraints reported in Section S7. All positive constraints must be present, all negative constraints must be absent, and each candidate must contain exactly two polymerization placeholders. The complete conversational transcript was supplied to the journal as a reviewer-only audit file and is not intended for publication.

Supplementary Prompt. Frozen molecular-generation instruction

Prompt template: “Please read the following MACCS keys SMARTS, and generate 5 potential molecule SMILES that meet all the requirements. All pos must be present in the SMILES and monomer, and all neg must be absent. Set the placeholder as ‘’. Each monomer must contain two placeholders.”*

Supplementary Code. RDKit validation of prompt constraints

The following minimal validator implements the syntactic and MACCS-bit compliance checks used in the prompt audit. The positive_keys and negative_keys arguments are populated with the target-specific key numbers listed in Section S7.

```
from rdkit import Chem

from rdkit.Chem import MACCSkeys

def validate_candidate(smiles, positive_keys, negative_keys):

    if smiles.count("*") != 2:

        return False

    mol = Chem.MolFromSmiles(smiles)

    if mol is None:

        return False

    fp = MACCSkeys.GenMACCSKeys(mol)

    has_all_positive = all(bool(fp[key]) for key in positive_keys)

    has_no_negative = all(not bool(fp[key]) for key in negative_keys)

    return has_all_positive and has_no_negative
```

Table S27. One-turn prompt audit by web-interface model/mode.

Provider	Mode	Model type	Access	Raw	RDKit	Compliant	Pass
GPT5.5	Instant	Fast/default	Free-capable, limited quota	5	4	2	40.0%
	Medium	Thinking, standard effort	Paid model picker (Plus or higher)	5	5	0	0.0%
	High	Thinking, extended effort	Paid model picker (Plus or higher)	5	3	2	40.0%

Provider	Mode	Model type	Access	Raw	RDKit	Compliant	Pass
	Extra High	Thinking, highest effort	Paid premium (Pro or eligible workspace)	5	5	5	100.0%
	Pro	Pro model	Paid premium (Pro or eligible workspace)	5	5	5	100.0%
Gemini	3.5 Flash-lite	Flash-lite / fast	Free/general-access class	5	5	0	0.0%
	Flash 3.6	Flash / fast	Free/general-access class	5	5	5	100.0%
	Pro	Pro / reasoning	Free-limited; paid plans have higher limits	5	5	5	100.0%
	Pro-extend	Extended Pro mode	Paid/expanded-access mode used in audit	5	5	5	100.0%
	Low	Opus, low effort	Paid premium mode used in audit	5	5	5	100.0%
	Medium	Opus, medium effort	Paid premium mode used in audit	5	5	5	100.0%
Claude Opus 4.8	High	Opus, high effort	Paid premium mode used in audit	5	5	5	100.0%
	Extra	Opus, extra effort	Paid premium mode used in audit	5	5	5	100.0%
	Max	Opus, Max mode	Paid Max plan	5	5	5	100.0%

Table S28. Descriptive prompt-audit summary by access category.

Access group	Modes	Raw	RDKit valid	Valid rate	Compliant	Pass rate
Strictly free/default modes	3	15	14	93.3%	7	46.7%
Free-capable including limited Pro access	4	20	19	95.0%	12	60.0%
Paid-only modes	10	50	48	96.0%	42	84.0%

To determine whether the rule-compliant monomers also follow the intended high-T_g and large-band-gap trend, the canonical-unique candidates with numerical Polymer Genome predictions were evaluated using the same single-chain E_g and T_g outputs (Figure S7) [1]. The dashed references (E_g = 3.0 eV and T_g = 477 K, equivalent to 204 °C) were applied uniformly after the one-turn audit. Twenty-two of the 43 point estimates fall in the upper-right target-consistent quadrant. The displayed uncertainty bars are the uncertainty values returned by Polymer Genome and are retained to show prediction spread; the point

estimates are used as a low-cost prioritization screen rather than as experimental measurements.

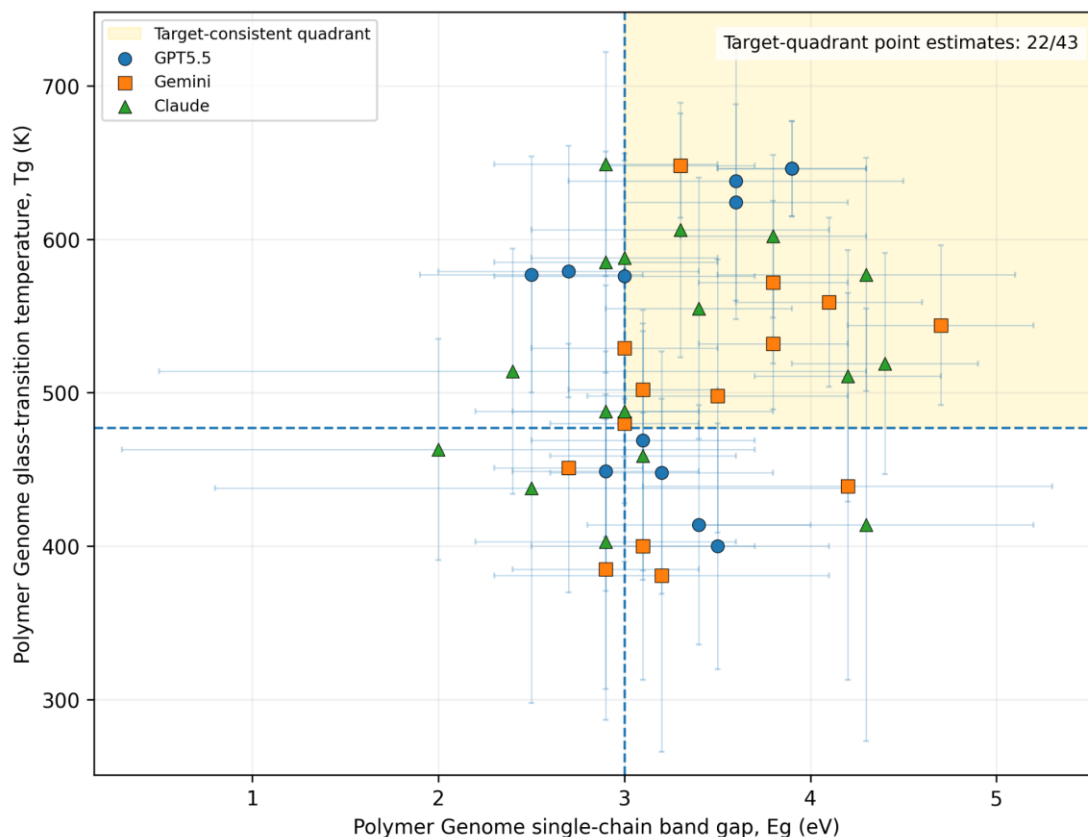


Figure S7 Polymer Genome trend screening places a subset of STCA-compliant candidates in the joint high- T_g , wide-gap region. The horizontal axis is the predicted single-chain band gap, E_g , in electronvolts; the vertical axis is the predicted glass-transition temperature, T_g , in kelvin. Vertical and horizontal dashed lines mark the internal screening references $E_g = 3.0$ eV and $T_g = 477$ K (204 °C), respectively, and the shaded upper-right region denotes the target-consistent quadrant. Error bars are the uncertainties returned by Polymer Genome. Marker shapes denote the commercial web interface used for candidate generation, as defined in the accompanying audit table. Twenty-two of the 43 canonical-unique candidates with numerical predictions lie in the target-consistent quadrant.

Reference

- [1] Doan Tran, H. et al. Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. 128, 171104 (2020).

S16 Materials Studio molecular dynamics, Tg fitting, and MD-derived trimer conformational sampling

All-atom MD simulations were carried out using the Forcite Plus module in BIOVIA Materials Studio with the COMPASS II force field. For each candidate polymer, three independently constructed amorphous models were prepared, and each periodic simulation cell contained ten trimer (3-mer) chains. The initial cells were geometry-optimized using the Ultra-fine quality setting, with convergence criteria of 1.0×10^{-4} kcal mol⁻¹ for energy, 5.0×10^{-3} kcal mol⁻¹ Å⁻¹ for force, and 5.0×10^{-5} Å for atomic displacement; the maximum number of optimization iterations was 2.0×10^5 . [1]

The optimized systems were equilibrated sequentially for 1000 ps in the NPT ensemble at 300 K and 1.01×10^{-4} GPa and for 1000 ps in the NVT ensemble at 300 K. A Nosé thermostat with Q ratio 0.01 and relaxation time 0.1 ps controlled temperature, a Berendsen barostat with relaxation time 0.1 ps controlled pressure during the NPT stages, and a 1 fs time step was used throughout. The same force field, optimization criteria, equilibration sequence, thermostat/barostat settings, and time step were applied to all candidate systems.

Thermal annealing was performed under the NPT ensemble over 300–800 K. Each system was heated to the high-temperature state to reduce dependence on the initial packed configuration and subsequently cooled to 300 K at an effective cooling rate of 2.4–2.6 K ps⁻¹. Density and simulation-cell volume were monitored, and only cooling-stage density–temperature data were used to determine Tg. The raw density observations are retained as gray points in Figure 4b and smoothed with a third-order Savitzky–Golay filter. Each smoothed trajectory is fitted by nonlinear least squares with the same unconstrained hyperbolic function; the black dashed curve is the fit, and one solid point in each design panel marks the fitted Tg. The fitted values are 570.5, 587.0, and 541.8 K for Designs 1–3, respectively, and all fits give $R^2 > 0.9$. The independent amorphous-model construction, raw-to-smoothed trajectory comparison, common fitting form, fitted transition point, and R^2 provide complementary checks on packing, trajectory noise, and fitting quality within the finite-trimer protocol.

For the electronic calculations, five individual trimer chains were randomly selected from the equilibrated ten-chain amorphous cell. Each selected trimer retained the conformation generated in the condensed-phase packing environment and was extracted as an individual molecular structure. No geometry optimization was performed after extraction. The fixed MD-derived trimers were passed to the Gaussian 09 single-point workflow described in S17.

Reference

- [1] Afzal, M. A. F. et al. High-throughput molecular dynamics simulations and validation of thermophysical properties of polymers for various applications. *ACS Appl. Polym. Mater.* 3, 620–630 (2021).

S17 Gaussian 09 monomer and MD-derived trimer frontier-orbital calculations and reproducibility

All reported monomer and trimer frontier-orbital calculations were performed using Gaussian 09 Revision D.01 [1]. Monomer geometries were initially optimized in the gas phase at B3LYP/6-31G(d) and subsequently refined at B3LYP/6-311G(d,p) using the checkpoint geometry and wavefunction; the reported monomer HOMO and LUMO energies were obtained at the refined B3LYP/6-311G(d,p) level. Frequency and polarizability calculations were also performed at B3LYP/6-311G(d,p). For the trimer analysis, the five fixed MD-derived conformations of each design were subjected to independent Gaussian 09 single-point calculations at B3LYP/6-311G(d,p) without further geometry optimization. The same electronic-structure level was used for all fifteen trimer calculations. The systems were treated as neutral singlets, and no empirical-dispersion or implicit-solvent keyword was used.

The monomer frontier energies are reported in Table S29.

Figure 4a contains the trimer results: the plotted HOMO and LUMO energies are the arithmetic means across the five fixed conformations, and the error bars show the full minimum-to-maximum ranges. For each conformation i , the gap was calculated as Table S30 separately reports the mean, standard deviation, and 95% t interval. The five gaps of Designs 1 and 2 all remain above 3.0 eV, whereas the five gaps of Design 3 all remain below 2.0 eV, preserving the intended insulation-versus-semiconductor differentiation across the sampled conformations. *Eg, $i = E_{LUMO,i} - E_{HOMO,i}$.*

The five-configuration trimer statistics, monomer frontier energies, and representative Gaussian reproducibility settings are tabulated below.

Table S29. Monomer frontier-orbital energies and HOMO–LUMO gaps.

Design	Role	HOMO (eV)	LUMO (eV)	Gap (eV)
Design 1	Strict high-Tg insulator	-6.875	-3.075	3.800
Design 2	Loose high-Tg insulator	-5.823	-2.174	3.649
Design 3	Loose high-Tg semiconductor	-4.614	-2.116	2.498

Table S30. Five-configuration frontier-energy summary from Gaussian 09 single-point calculations on fixed MD-derived trimers.

Design	Quantity	Mean (eV)	SD (eV)	95% low	95% high	n
Design 1	HOMO	-6.861	0.057	-6.932	-6.789	5
	LUMO	-3.407	0.036	-3.451	-3.363	5
	Gap	3.454	0.049	3.392	3.515	5
Design 2	HOMO	-5.770	0.032	-5.810	-5.730	5

Design	Quantity	Mean (eV)	SD (eV)	95% low	95% high	n
Design 3	LUMO	-2.517	0.057	-2.588	-2.447	5
	Gap	3.253	0.047	3.194	3.312	5
	HOMO	-4.122	0.043	-4.175	-4.068	5
	LUMO	-2.701	0.015	-2.720	-2.682	5
	Gap	1.421	0.056	1.351	1.490	5

The representative Gaussian input/output files document the monomer optimization, frequency, and polarizability workflow. The trimer calculations use the same Gaussian 09/B3LYP/6-311G(d,p) electronic-structure level as fixed-geometry single-point calculations on the MD-derived conformations.

Table S31. Representative Gaussian monomer reproducibility details extracted from the supplied input/output files.

Item	Recorded setting
Software	Gaussian 09, Revision D.01 (EM64W)
Checkpoint / resources	16 GB memory; 12 processors for initial optimization and 16 for subsequent Link1 jobs
Initial optimization	#p B3LYP/6-31G(d) opt
Refined optimization	#p B3LYP/6-311G(d,p) opt geom=check guess=read
Frequency calculation	#p B3LYP/6-311G(d,p) freq geom=check guess=read
Polarizability calculation	#p B3LYP/6-311G(d,p) polar geom=check guess=read
Charge / multiplicity	0 / 1
Final basis implementation	6-311G(d,p) (5D, 7F) for the refined optimization, frequency, and polarizability steps
Empirical dispersion	No empirical-dispersion keyword is present in the supplied route sections (uncorrected B3LYP).
Solvation / phase	No SCRF or solvent keyword is present; the supplied molecular calculations are gas-phase calculations.
Integration grid	No explicit Int/Grid keyword is present; the Gaussian 09 default numerical integration grid was used.
SCF convergence	RMS density 1.00D-08, maximum density 1.00D-06., energy 1.00D-06.; maximum 128 cycles
Geometry optimizer	Berny optimization in redundant internal coordinates; Gaussian default force/displacement convergence thresholds.
Final maximum force	0.000148 (threshold 0.000450)
Final RMS force	0.000017 (threshold 0.000300)

Item	Recorded setting
Final maximum displacement	0.001080 (threshold 0.001800)
Final RMS displacement	0.000218 (threshold 0.001200)
Stationary-point check	Frequency calculation at B3LYP/6-311G(d,p); first three harmonic modes 43.8086, 71.8466, and 109.0786 cm ⁻¹ ; no imaginary vibrational mode listed.

Reference

[1] Frisch, M. J. et al. Gaussian 09, Revision D.01 (Gaussian, Inc., Wallingford, CT, 2013).