iScience, Volume *26*

Supplemental information

Optimal high-throughput virtual screening pipeline for efficient selection of redox-active organic materials

Hyun-Myung Woo, Omar Allam, Junhe Chen, Seung Soon Jang, and Byung-Jun Yoon

Contents

1 Performance comparison of various machine learning surrogate models in predicting properties

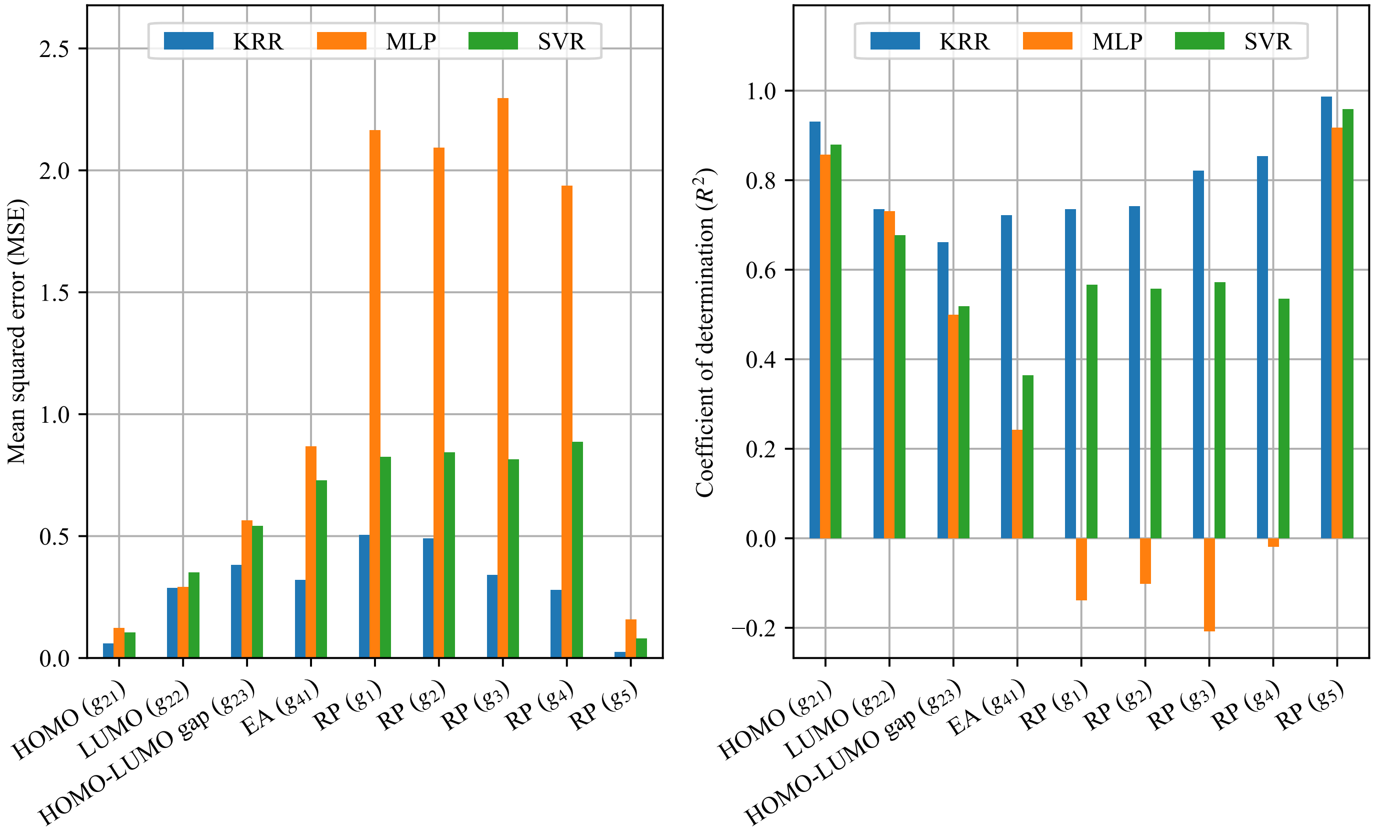


Figure S1. Performance comparison of several machine learning models in terms of mean squared error (MSE) and coefficient of determination, Related to STAR Methods.
For the kernel ridge regression (KRR) model, we used the hyperparameters summarized in Table S2. For multilayer perceptron (MLP) and support vector regression (SVR) models, we optimized hyperparameters based on a 5-fold cross-validation.

2 Specification of the optimized machine learning surrogate models of the HTVS pipeline

Surrogate	Descriptors	Predicting property
g_1	Primitive features (PFs): #C, #B, #O, #Li, #H, # of aromatic rings	RP
g_2	PFs, HOMO (pred.), LUMO (pred.), HOMO-LUMO gap (pred.)	
g_3	PFs, HOMO, LUMO, HOMO-LUMO gap	
g_4	PFs, HOMO, LUMO, HOMO-LUMO gap, EA (pred.)	
g_5	PFs, HOMO, LUMO, HOMO-LUMO gap, EA	
$g_{2,1}$	PFs	HOMO
$g_{2,2}$		LUMO
$g_{2,3}$		HOMO-LUMO gap
$g_{4,1}$	PFs, HOMO, LUMO, HOMO-LUMO gap	EA

Table S1. Specifications of the surrogate models (1 to 5) and sub-surrogate models (2.1 to 2.3 and 4.1), Related to STAR Methods.
Sub-surrogates predict intermediate properties used as virtual descriptors for the surrogate models to improve predictive capacity. A kernel ridge regressor with a radial basis function (RBF) kernel was used.

Surrogate	Machine learning model	Kernel	Hyperparameter (α)	Mean squared error	R^2
g_1	Kernel ridge regression	Radial basis function	0.1	0.5046	0.7346
g_2			0.1	0.4907	0.7419
g_3			0.1	0.3408	0.8208
g_4			0.1	0.2781	0.8538
g_5			0.1	0.0256	0.9865
$g_{2,1}$	Kernel ridge regression	Radial basis function	0.1	0.0595	0.9307
$g_{2,2}$			0.1	0.2870	0.7351
$g_{2,3}$			0.1	0.3808	0.6616
$g_{4,1}$			0.1	0.3194	0.7212

Table S2. Performance analysis of the optimized machine learning surrogate models utilized to construct the HTVS pipeline, Related to STAR Methods.
The hyperparameters–kernel function and α–were optimized via 5-fold cross-validation.

3 Performance evaluation at each stage of the optimized high-throughput virtual screening (HTVS) pipeline with structure $[S_1, S_2, S_3, S_4, S_5, S_6]$

3.1 Optimal computational campaign for selecting potential organic electrode materials with minimum target redox potential (RP) 2.5 V

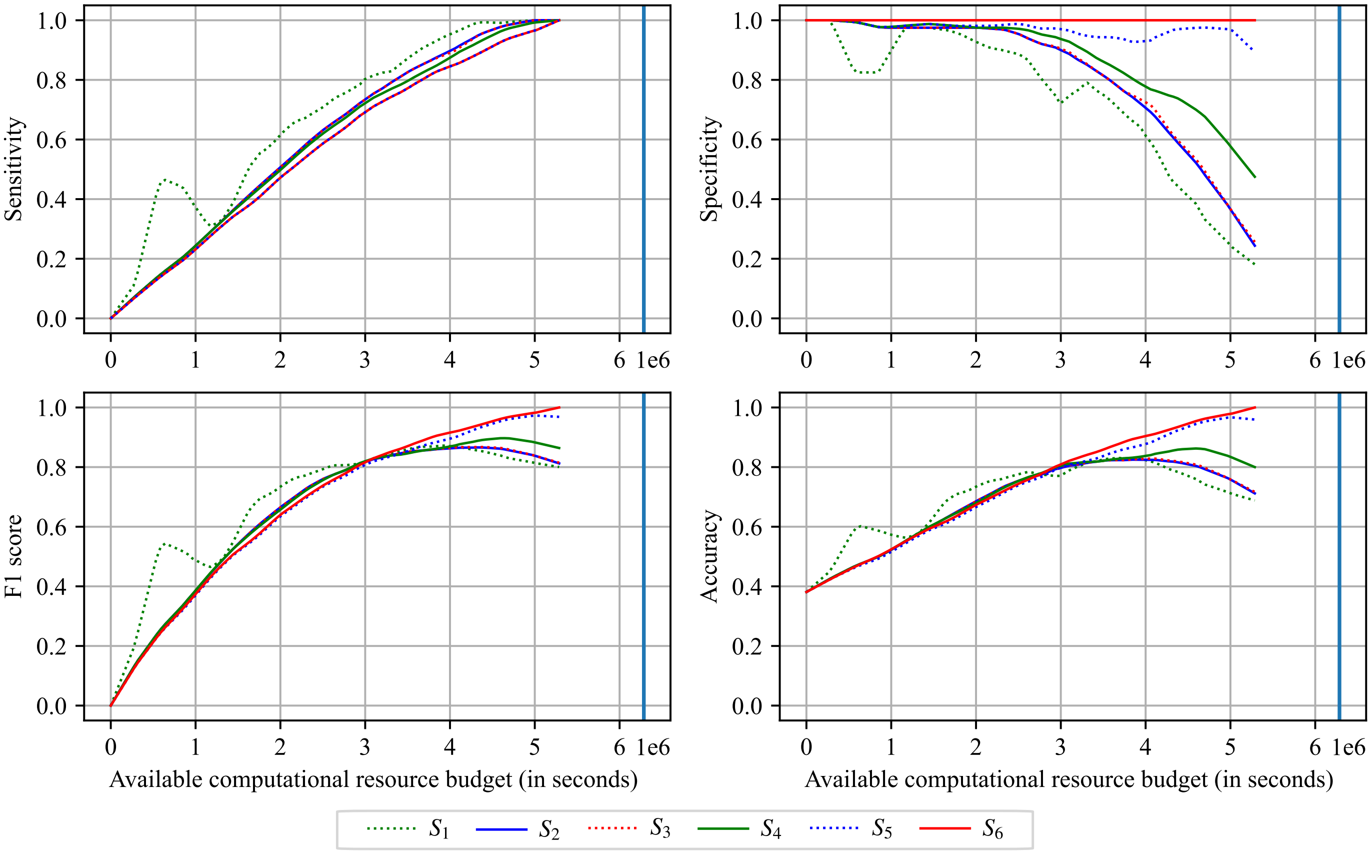


Figure S2. Performance evaluation of the individual stages constituting the HTVS pipeline based on a 5**-fold cross-validation, Related to Figs. 3 and 4.**
In general, sensitivity tended to increase as the allocated computational budget increased. On the other hand, the specificity of the stages (except for the last stage) tended to decrease as the allocated resource increased. This was because the earlier stages were designed to pass a larger number of candidates to later stages as the available budget grew, in order to evaluate and screen the materials with higher accuracy. For the same reason, the F1 score and the accuracy generally increased as the computational budget grew, but they eventually decreased due to the increasing false-positive rates as a result of passing too many candidates to subsequent stages.

3.2 Optimal computational campaign for selecting potential organic electrode materials with target RP screening range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$

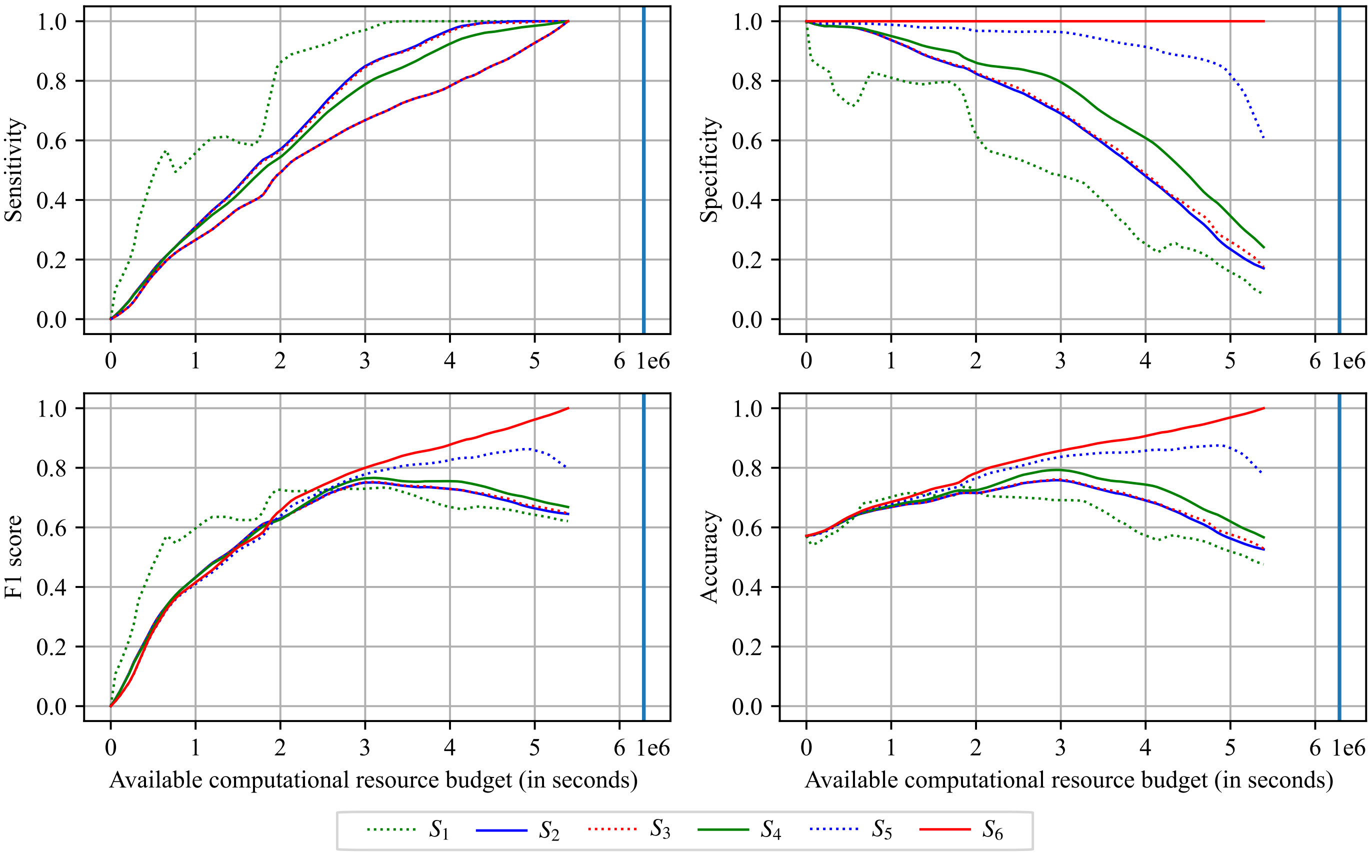


Figure S3. Performance evaluation of the screening stages in the optimized HTVS pipeline designed to detect the organic electrode materials according to target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ based on a 5-fold cross-validation, Related to Figs. 5 and 6.

As before, the sensitivity of the screening stages tended to increase as the computational budget (x-axis) grew. In general, the specificity decreased as the available resource rose (except for the last stage). The F1 score and accuracy improved as the available budget got larger, but they eventually decreased due to the increasing false-positive rates due to passing too many materials to the later stages.

4 Performance evaluation of the optimized HTVS pipeline with structure $[S_2, S_4, S_5, S_6]$

4.1 Optimal computational campaign for selecting potential organic electrode materials with minimum target RP 2.5 V

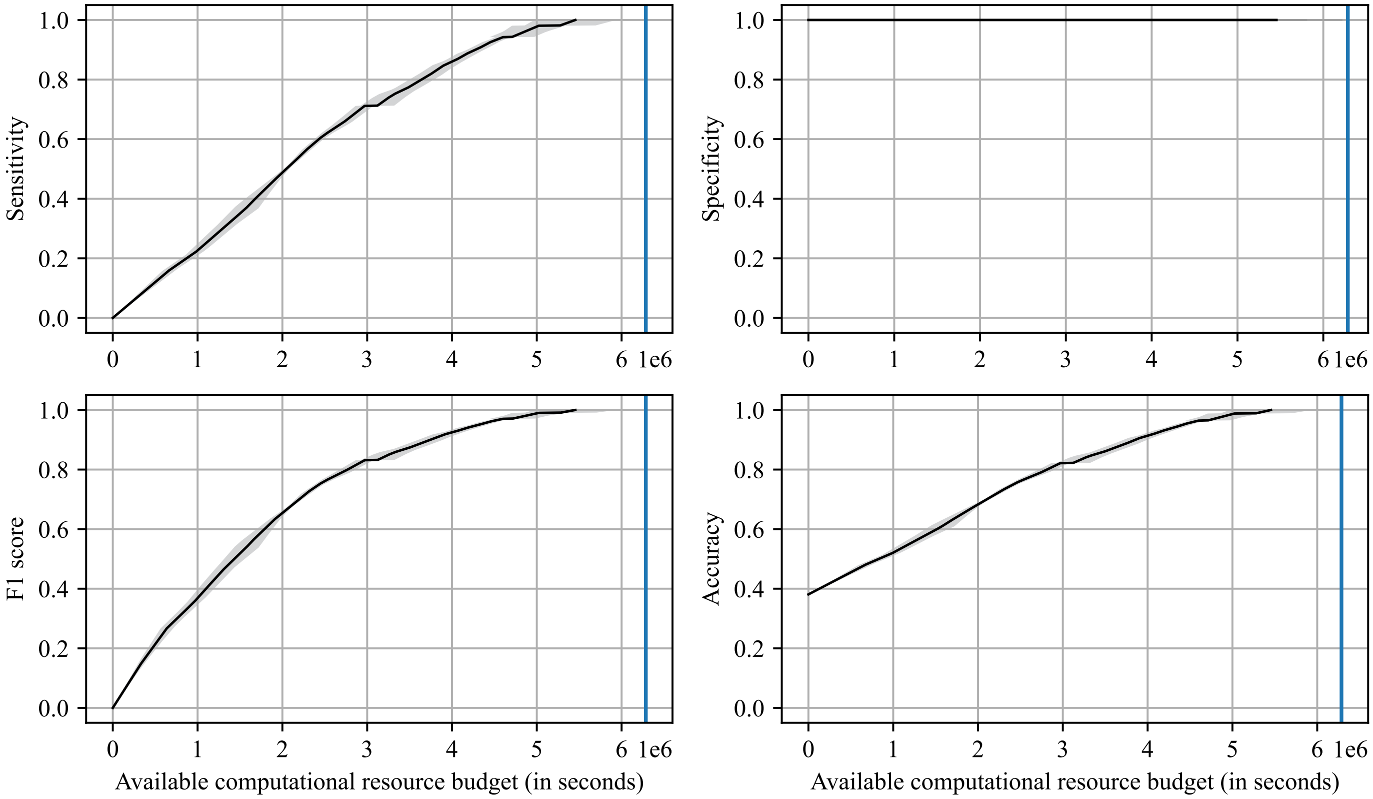


Figure S4. Performance evaluation of the optimized high-throughput virtual screening (HTVS) pipeline $[S_2, S_4, S_5, S_6]$ with minimum target RP 2.5 V under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 3 and 4.

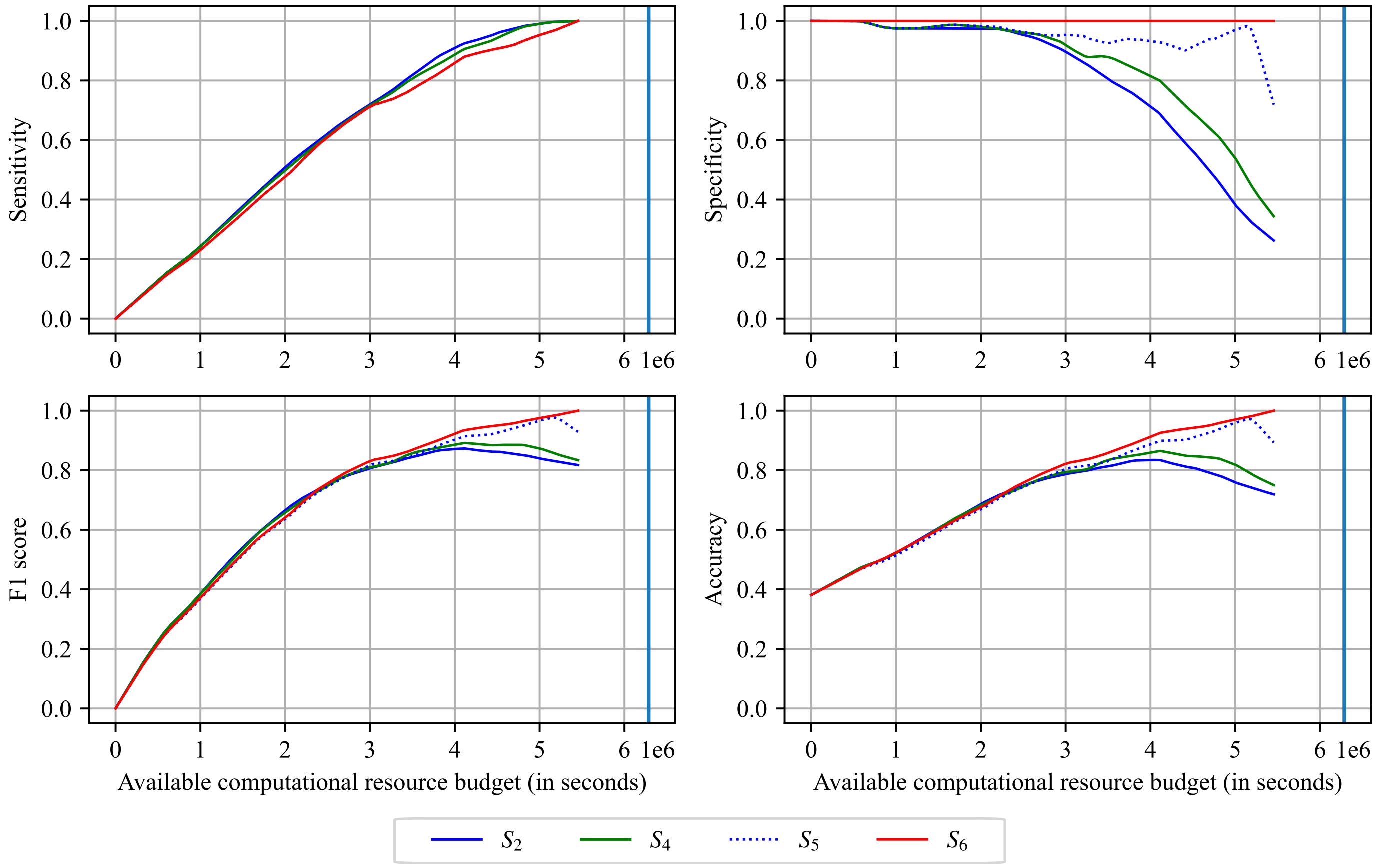


Figure S5. Performance evaluation at each stage in the optimized HTVS pipeline $[S_2, S_4, S_5, S_6]$ with minimum target RP 2.5 V under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 3 and 4.

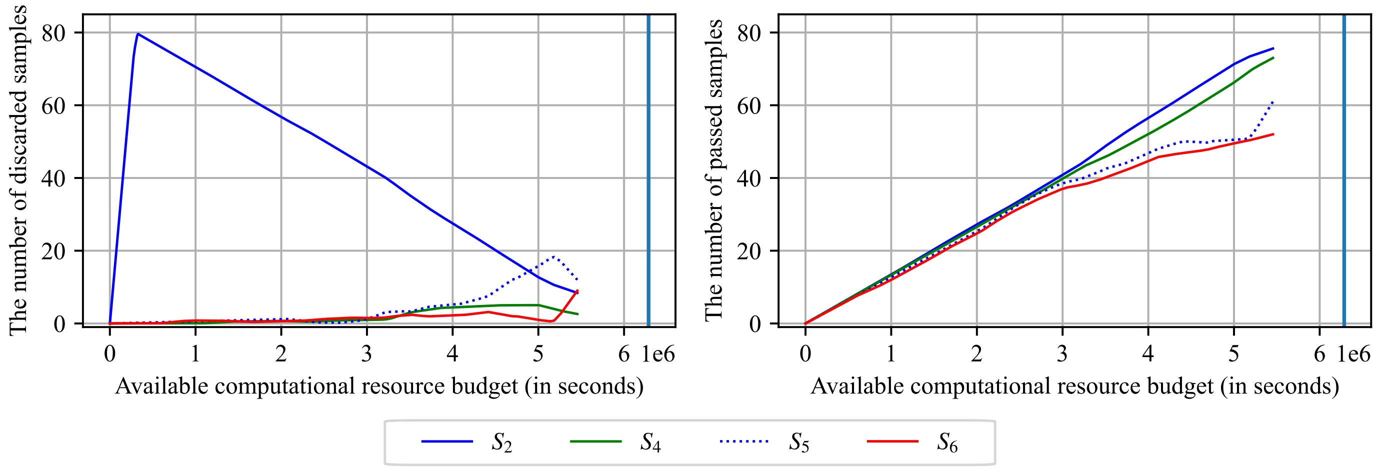


Figure S6. The number of samples discarded (left) or passed to the next stage (right) at each stage in the HTVS pipeline $[S_2, S_4, S_5, S_6]$ with minimum target RP 2.5 V under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 3 and 4.

α	Selected materials	Total cost (seconds)	Effective cost (seconds)	Sensitivity	Specificity	F1 score	Accuracy
0.25	21.2	$1,697,310.6$	$80,061.8$	0.4077	1	0.5562	0.6333
0.5	46.6	$4,211,703.2$	$90,379.9$	0.8962	1	0.9443	0.9357
0.75	48.2	$4,540,007.4$	$9,419.2$	0.9269	1	0.9616	0.9548

Table S3. Performance evaluation of the jointly optimized HTVS pipeline $[S_2, S_4, S_5, S_6]$ with minimum target RP 2.5 V based on a 5-fold cross-validation, Related to Figs. 3 and 4.

4.2 Optimal computational campaign for selecting potential organic electrode materials with target RP screening range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$

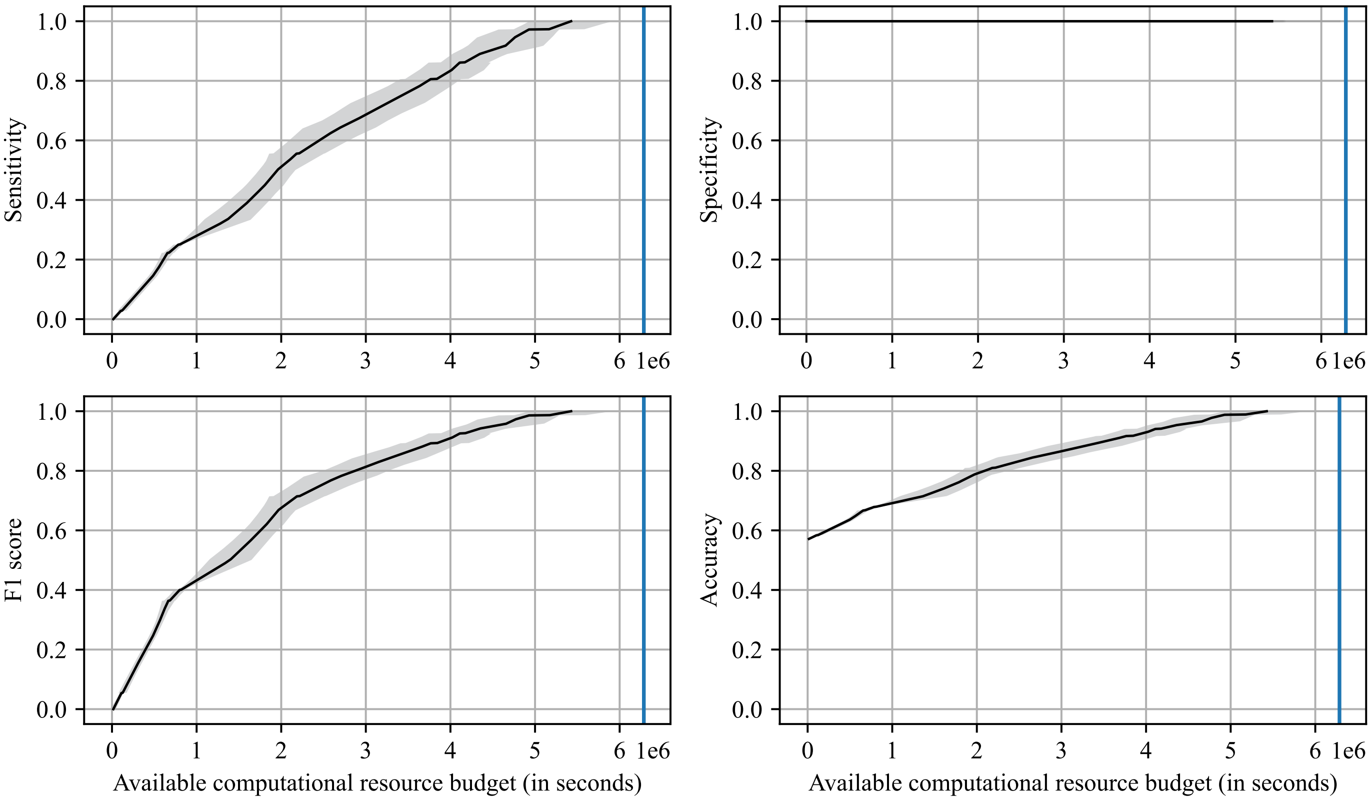


Figure S7. Performance evaluation of the jointly optimized HTVS pipeline $[S_2, S_4, S_5, S_6]$ with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 5 and 6.

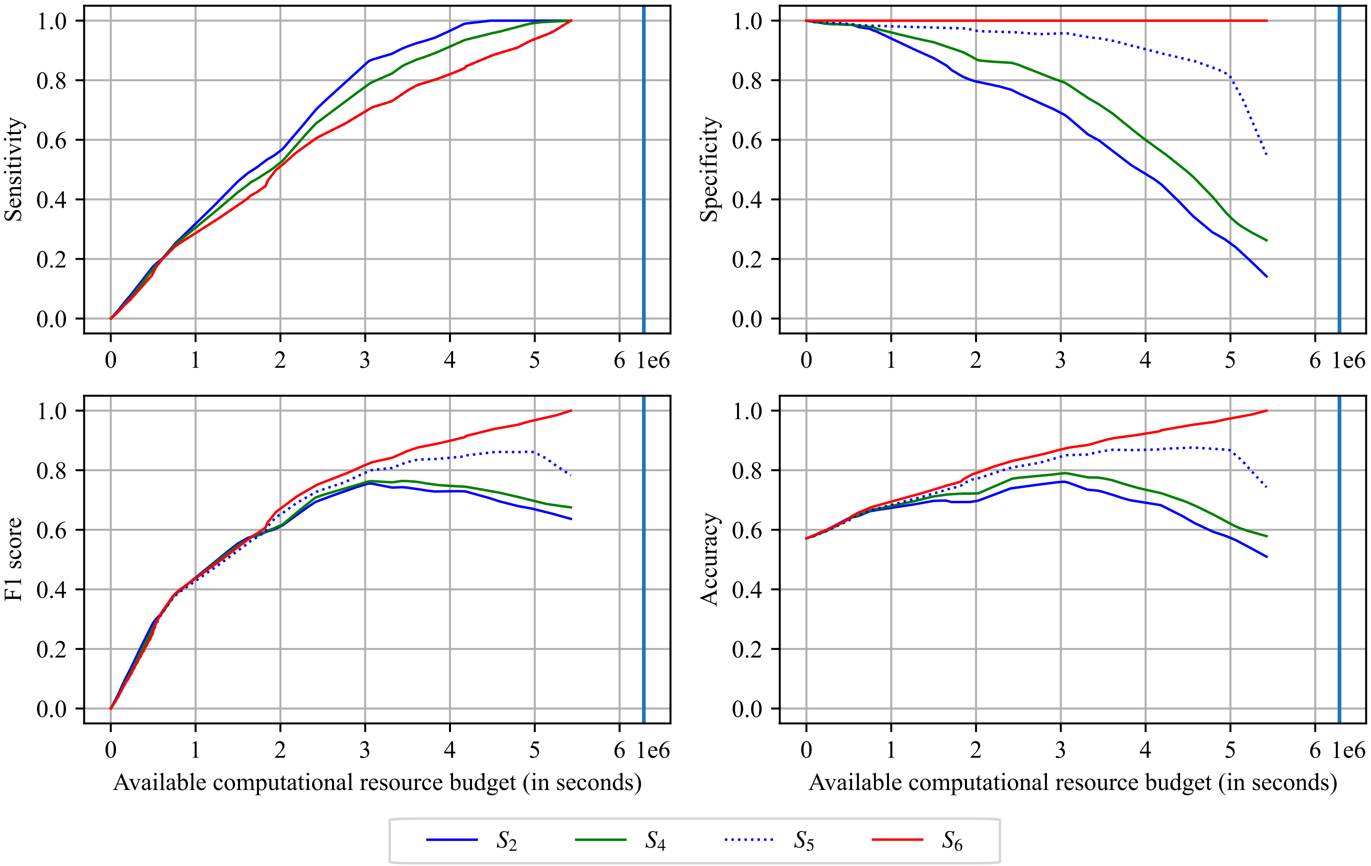


Figure S8. Performance evaluation at each stage in the optimized HTVS pipeline $[S_2, S_4, S_5, S_6]$ with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 5 and 6.

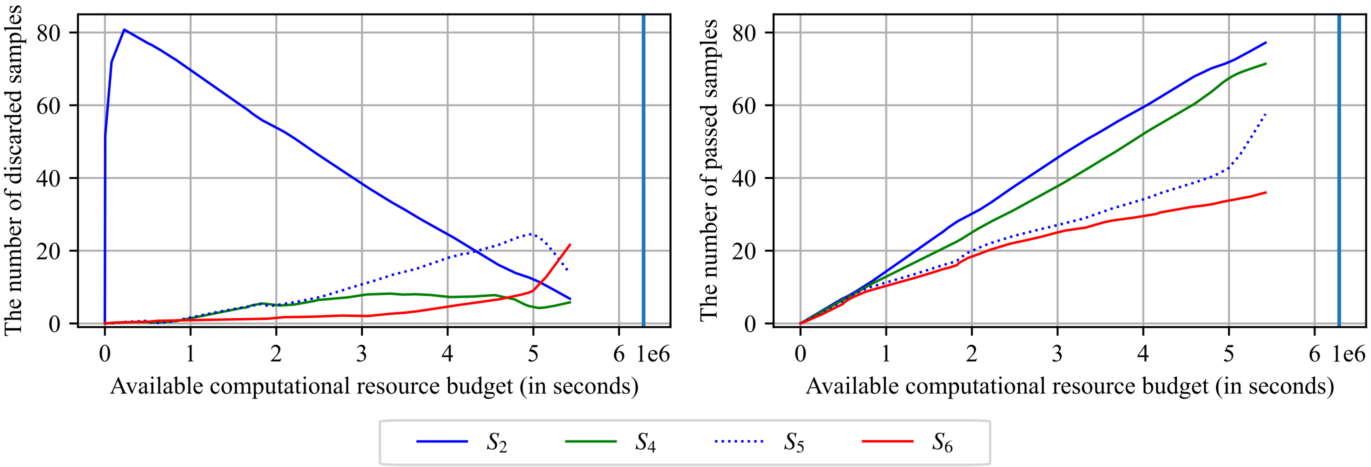


Figure S9. The number of samples discarded (left) or passed to the next stage (right) at each stage in the HTVS pipeline $[S_2, S_4, S_5, S_6]$ with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a 5-fold cross-validation, Related to Figs. 5 and 6.

α	Selected materials	Total cost (seconds)	Effective cost (seconds)	Sensitivity	Specificity	F1	Accuracy
0.25	12.2	$1,350,211.2$	$110,673$	0.3389	1	0.4732	0.7167
0.5	30.6	$3,645,767.6$	$119,142.7$	0.85	1	0.9054	0.9357
0.75	31.6	$4,307,546.6$	$136,314.8$	0.8778	1	0.9303	0.9476

Table S4. Performance evaluation of the jointly optimized HTVS pipeline $[S_2, S_4, S_5, S_6]$ with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ based on a 5-fold cross-validation, Related to Figs. 5 and 6.

5 Performance evaluation of the optimized HTVS pipeline based on a strict 5-fold cross-validation

5.1 Optimal computational campaign for selecting potential organic electrode materials with minimum target RP 2.5 V

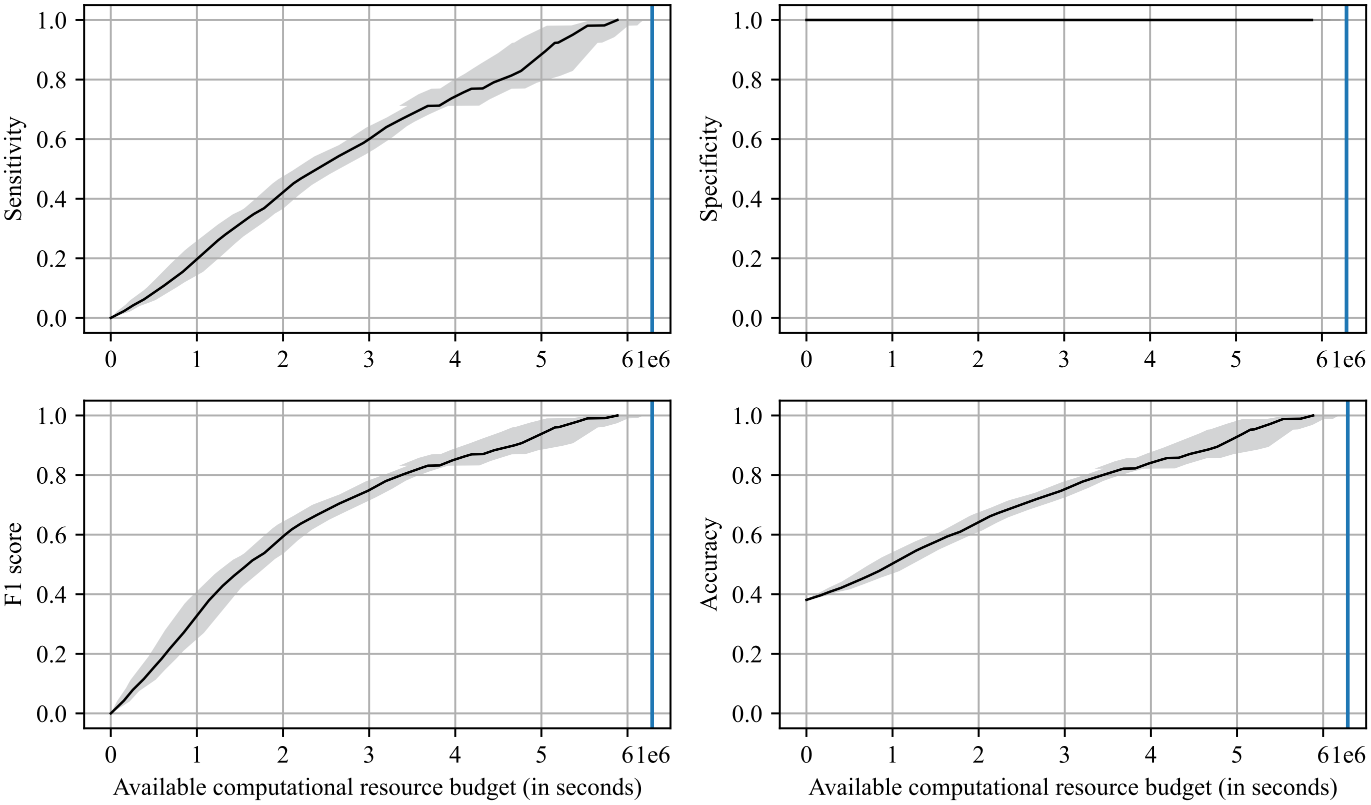


Figure S10. Performance evaluation of the optimized HTVS pipeline with minimum target redox potential (RP) 2.5 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

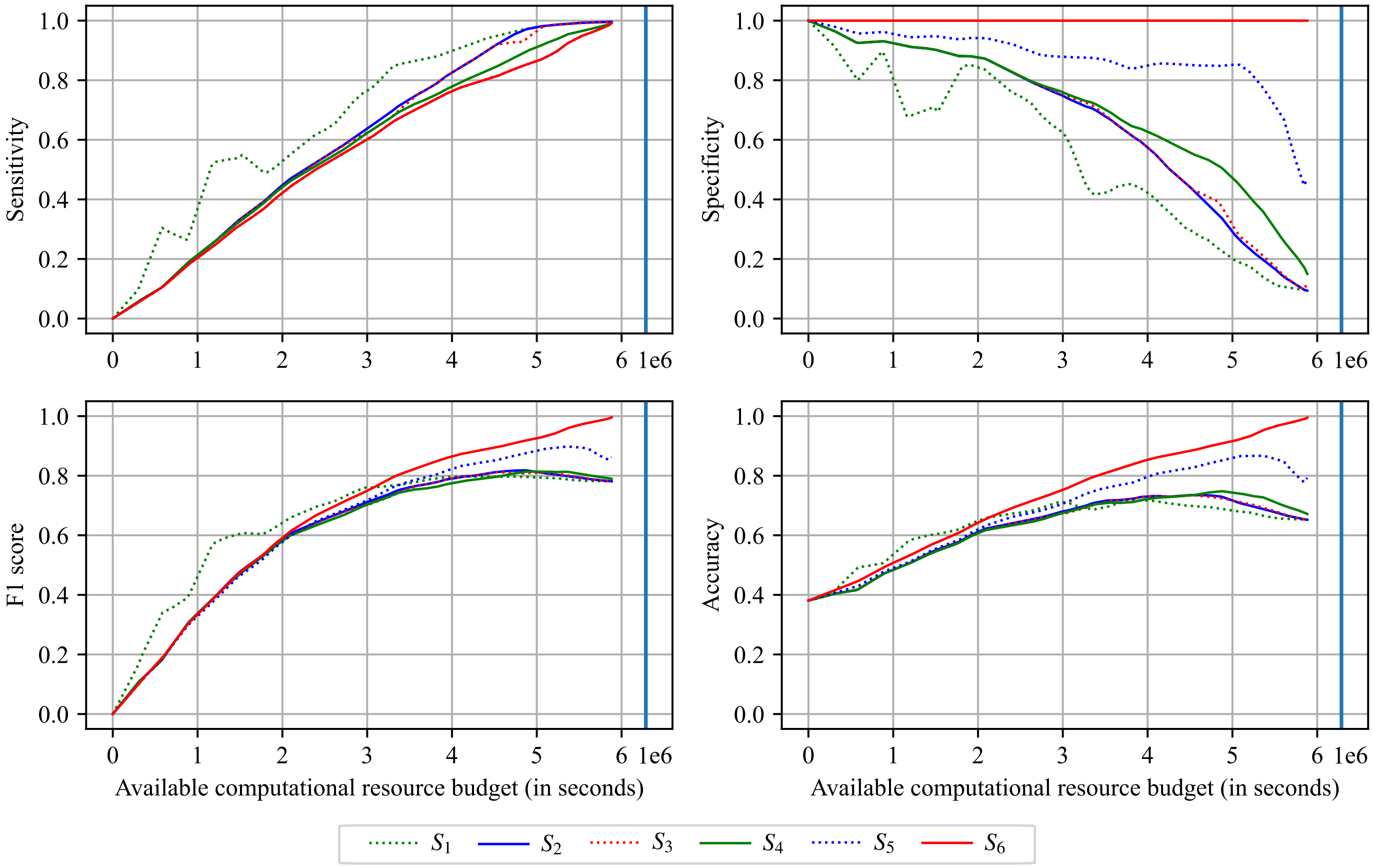


Figure S11. Performance evaluation at each stage in the optimized HTVS pipeline with minimum target RP 2.5 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

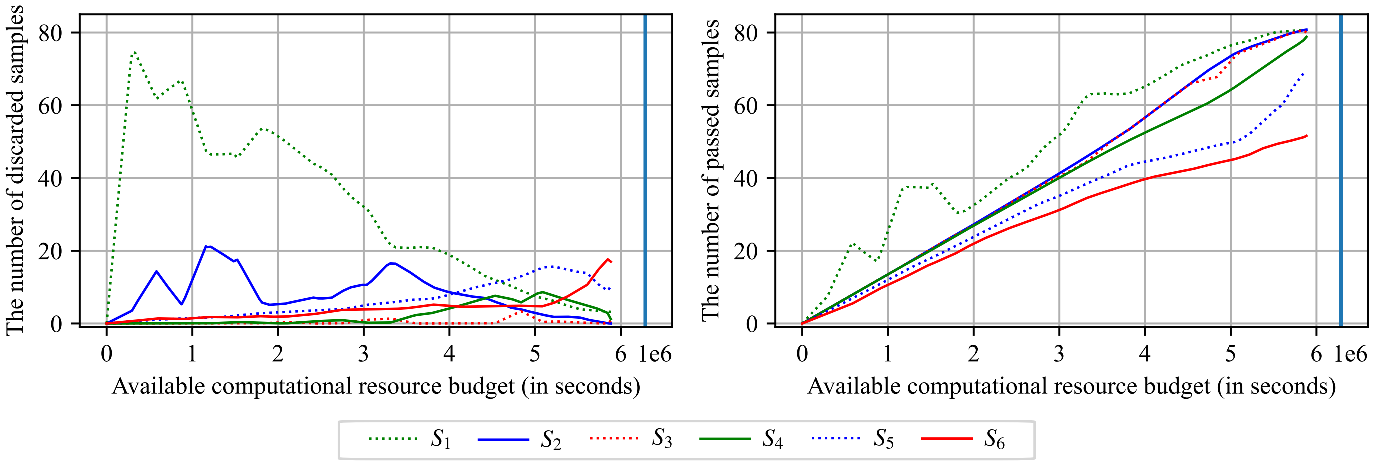


Figure S12. The number of samples discarded (left) or passed to the next stage (right) at each stage in the HTVS pipeline with minimum target RP 2.5 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

α	Selected materials	Total cost (seconds)	Effective cost (seconds)	Sensitivity	Specificity	F1 score	Accuracy
0.25	36.6	$3,506,408.8$	$95,803.5$	0.7038	1	0.8233	0.8167
0.5	43.8	$4,425,191.2$	$101,031.8$	0.8423	1	0.9134	0.9024
0.75	45.4	$4,605,138.8$	$101,434.8$	0.8731	1	0.9316	0.9214

Table S5. Performance evaluation of the jointly optimized HTVS pipeline with minimum target RP 2.5 V based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

5.2 Optimal computational campaign for selecting potential organic electrode materials with target RP screening range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$

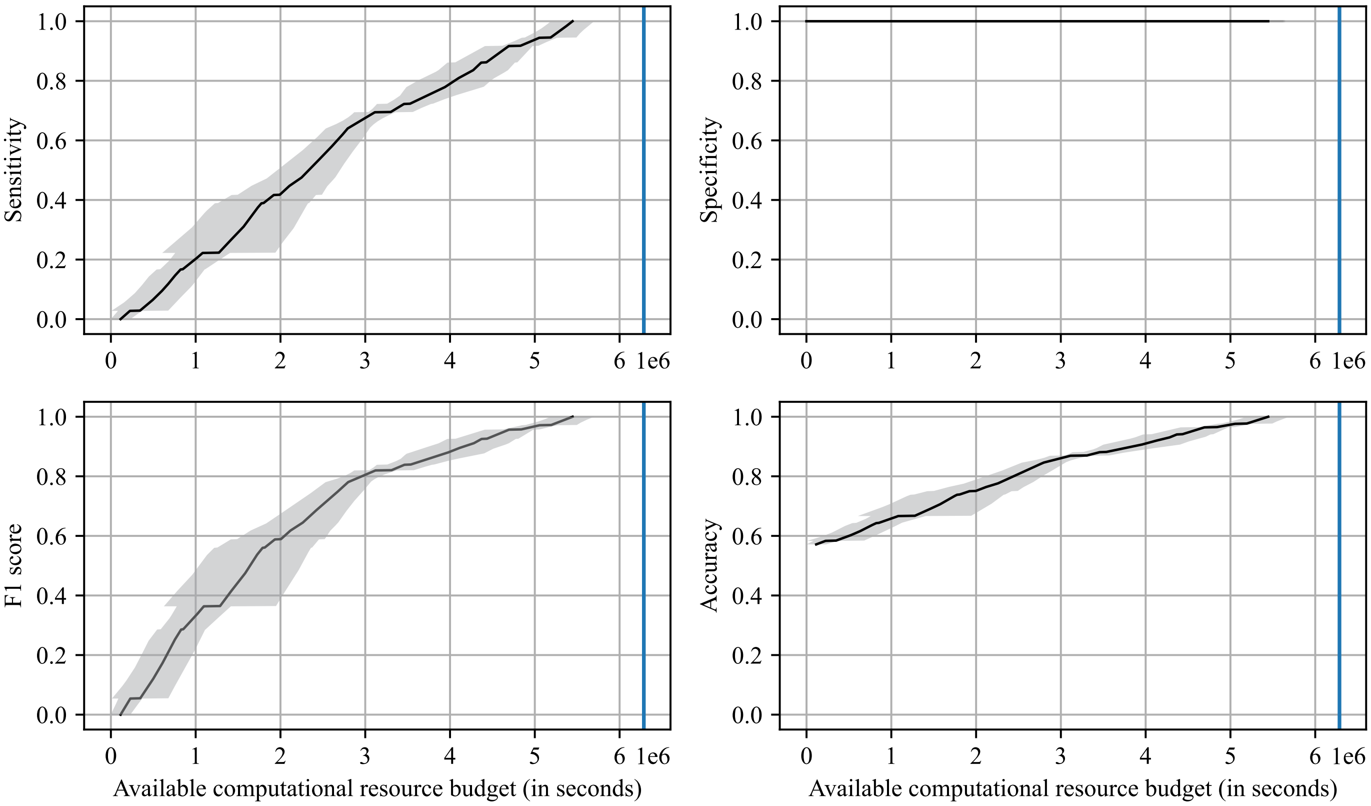


Figure S13. Performance evaluation of the optimized HTVS pipeline with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 5 and 6.

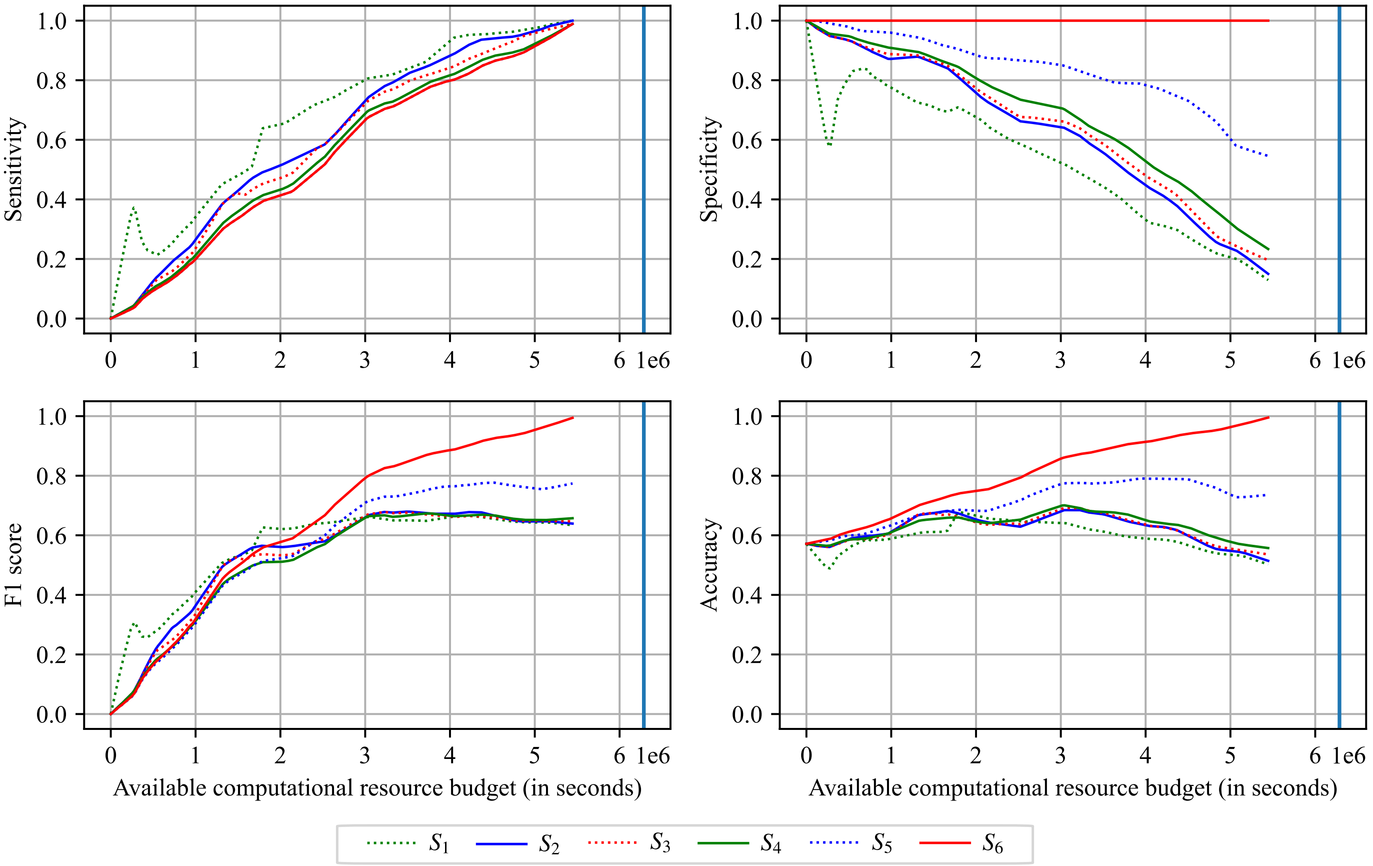


Figure S14. Performance evaluation at each stage in the optimized HTVS pipeline with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 5 and 6.

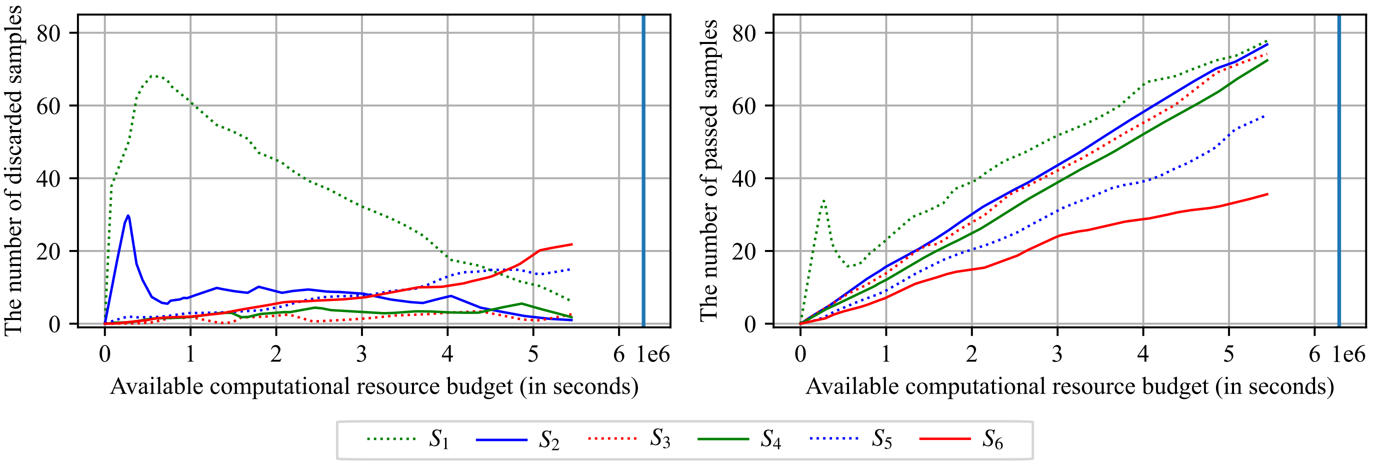


Figure S15. The number of samples discarded (left) or passed to the next stage (right) at each stage in the HTVS pipeline with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 5 and 6.

α	Selected materials	Total cost (seconds)	Effective cost (seconds)	Sensitivity	Specificity	F1	Accuracy
0.25	25	$3,026,165.6$	$121,046.6$	0.6944	1	0.8155	0.8690
0.5	31.6	$4,490,452.4$	$142,102.9$	0.8778	1	0.9336	0.9476
0.75	33.2	$5,112,623.8$	$153,994.7$	0.9222	1	0.9576	0.9667

Table S6. Performance evaluation of the jointly optimized HTVS pipeline with target RP range $[2.5\ \mathbf{V}, 3.2\ \mathbf{V}]$ based on a strict 5-fold cross-validation, Related to Figs. 5 and 6.

6 Performance evaluation of the optimized HTVS pipeline with minimum target RP 4.3 V based on a strict 5-fold cross-validation

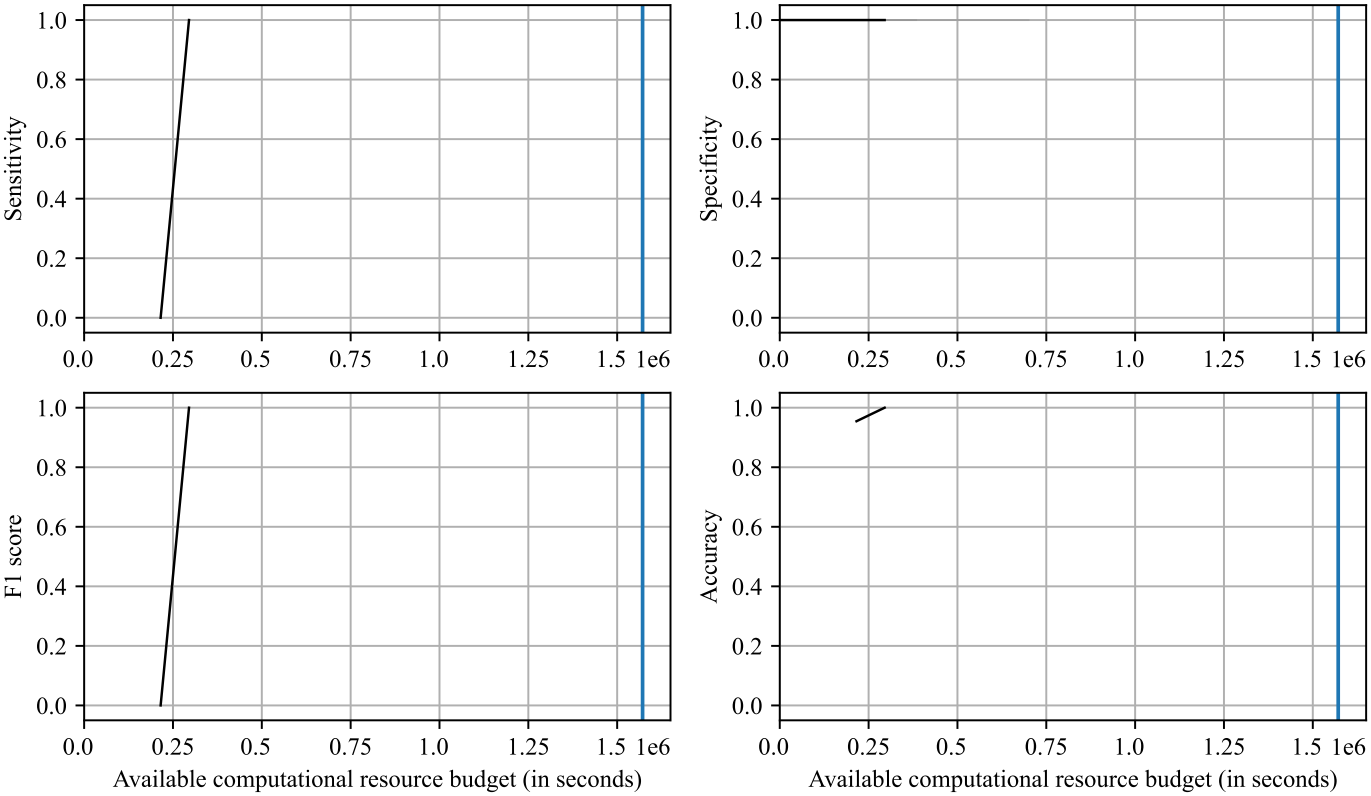


Figure S16. Performance evaluation of the optimized high-throughput virtual screening (HTVS) pipeline with minimum target redox potential (RP) 4.3 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

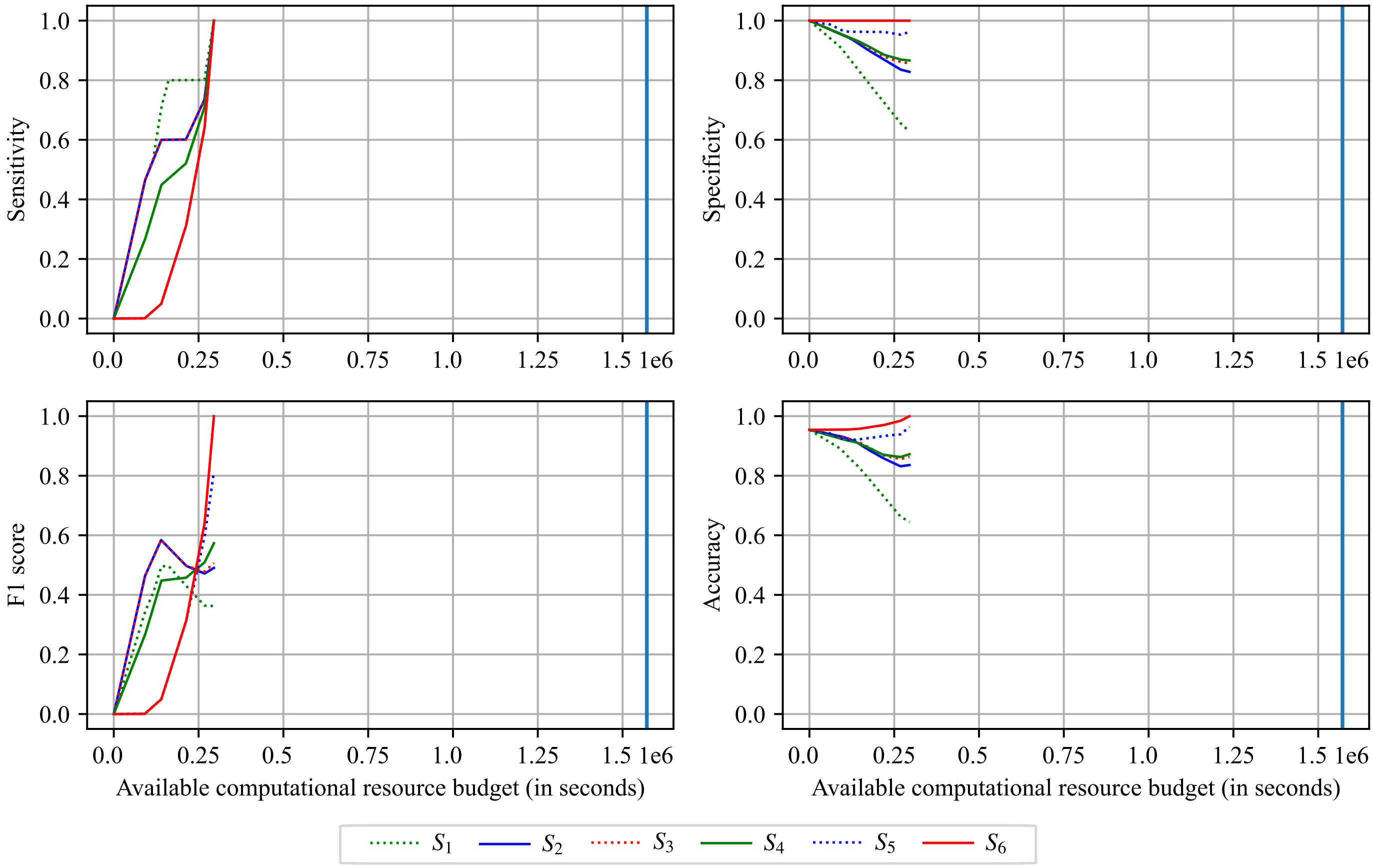


Figure S17. Performance evaluation at each stage in the optimized HTVS pipeline with minimum target RP 4.3 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

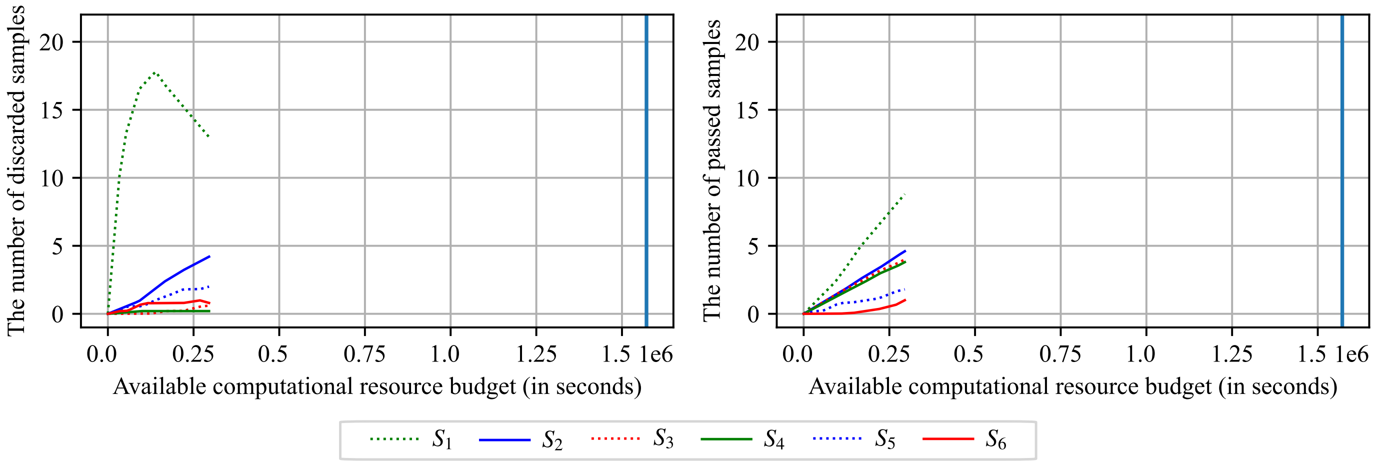


Figure S18. The number of samples discarded (left) or passed to the next stage (right) at each stage in the HTVS pipeline with minimum target RP 4.3 V under a computational resource budget constraint (x-axis) based on a strict 5-fold cross-validation, Related to Figs. 3 and 4.

7 Schematic illustration of families of organic moieties

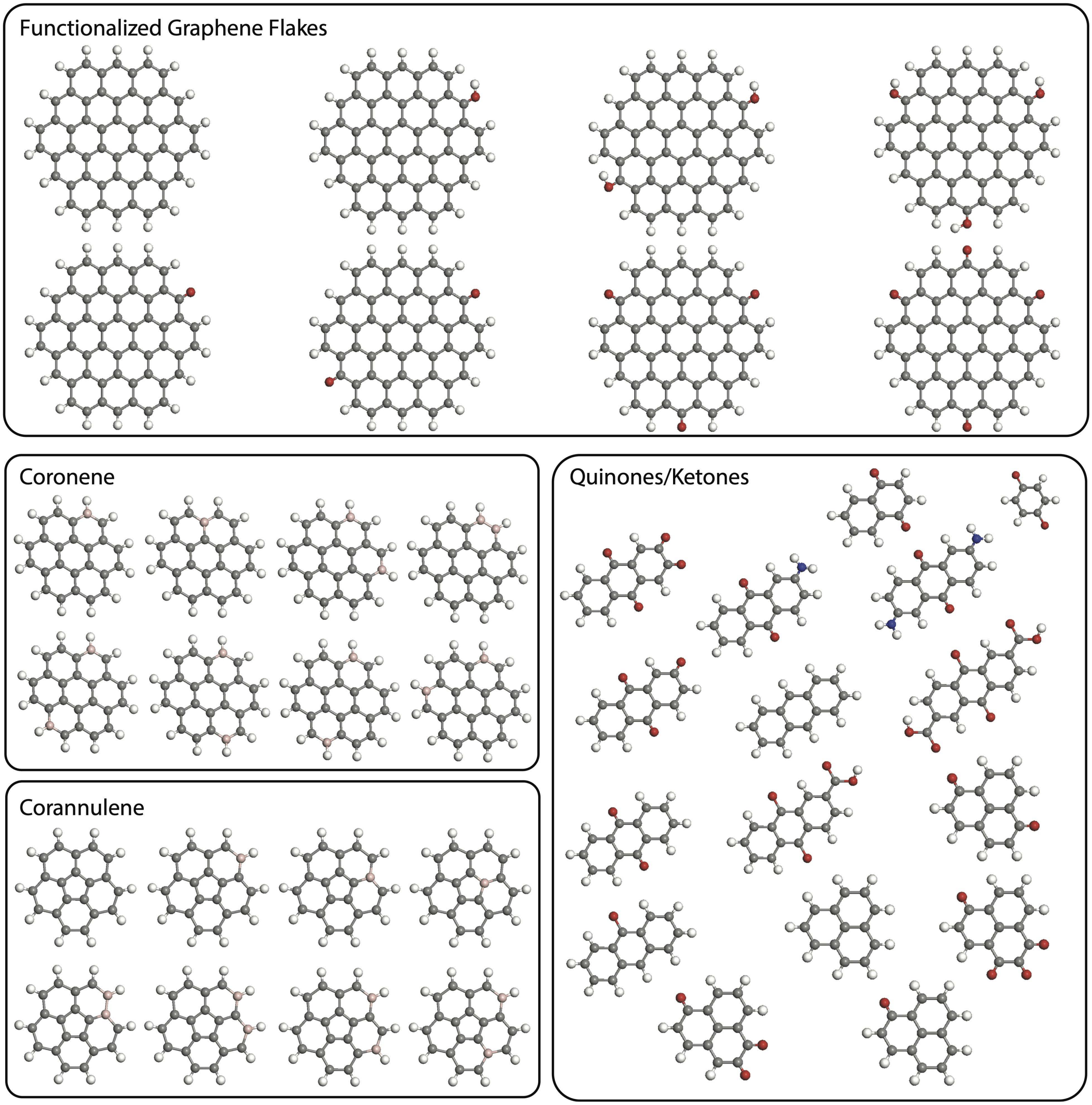
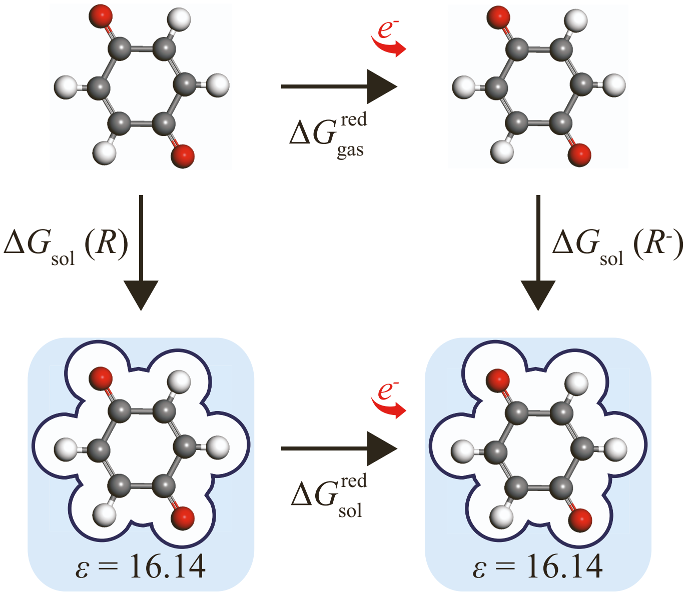


Figure S19. Schematic illustration of some of the different families of organic moieties that compose the dataset used in training the machine learning models in this study, Related to STAR Methods.

8 Thermodynamic cycle



e^-
$\Delta G^{\text{red}}_{\text{gas}}$
$\Delta G_{\text{sol}}\ (R)$
$\Delta G_{\text{sol}}\ (R^-)$
e^-
$\Delta G^{\text{red}}_{\text{sol}}$
$\varepsilon = 16.14$
$\varepsilon = 16.14$

Figure S20. The thermodynamic cycle used to calculate the RP in the condensed phase in this study, Related to STAR Methods.
The solvation-free energies ΔG_{sol} were evaluated using an implicit solvation model with dielectric constant ϵ approximating the carbonate mixture commonly used in the experimentation.

9 Illustration of constructing the skeleton structure of the HTVS pipeline

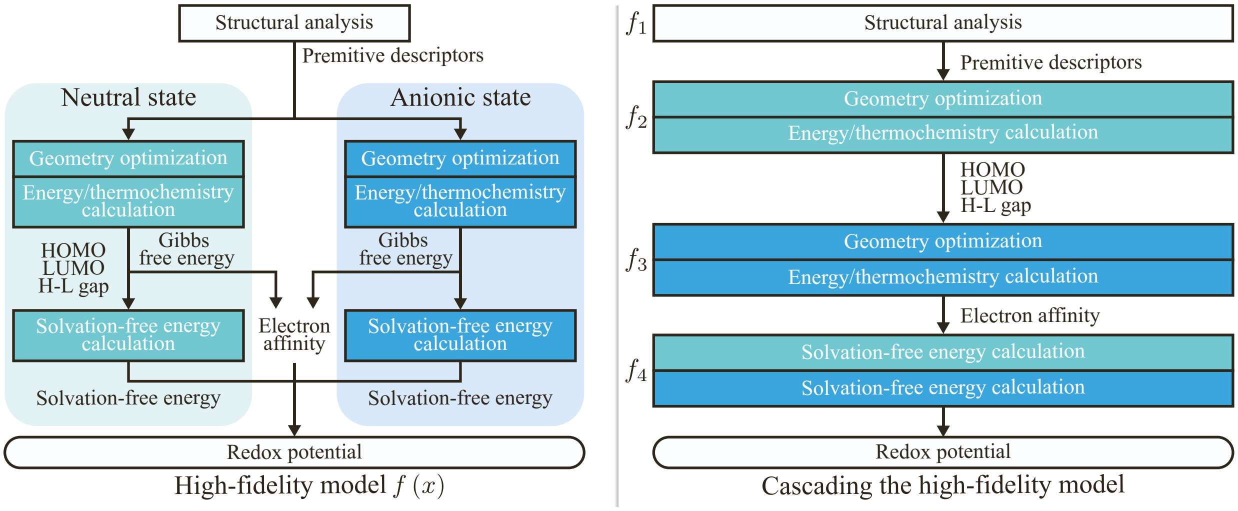


Figure S21. Illustration of constructing the skeleton structure of the HTVS pipeline based on the high-fidelity DFT model, Related to STAR Methods.