Electronic Supplementary Material (ESI) for Digital Discovery.

A Scientific Machine Learning Framework to Understand Flash Graphene Synthesis

Kianoosh Sattari,[1] Lucas Eddy,[2, 3] Jacob L. Beckham,[2] Kevin M. Wyss,[2] Richard Byfield,[1] Long Qian,[2] James M. Tour,[2, 4, 5*] and Jian Lin[1*]

[1]Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, USA

[2]Department of Chemistry, [3]Applied Physics Program and Smalley-Curl Institute, [4]Department of Materials Science and NanoEngineering, [5]Department of Computer Science Engineering, NanoCarbon Centre and the Welch Institute for Advanced Materials, Rice University, 6100 Main Street, Houston, Texas 77005, United States

*Emails: linjian@missouri.edu; tour@rice.edu

Supplementary Figures

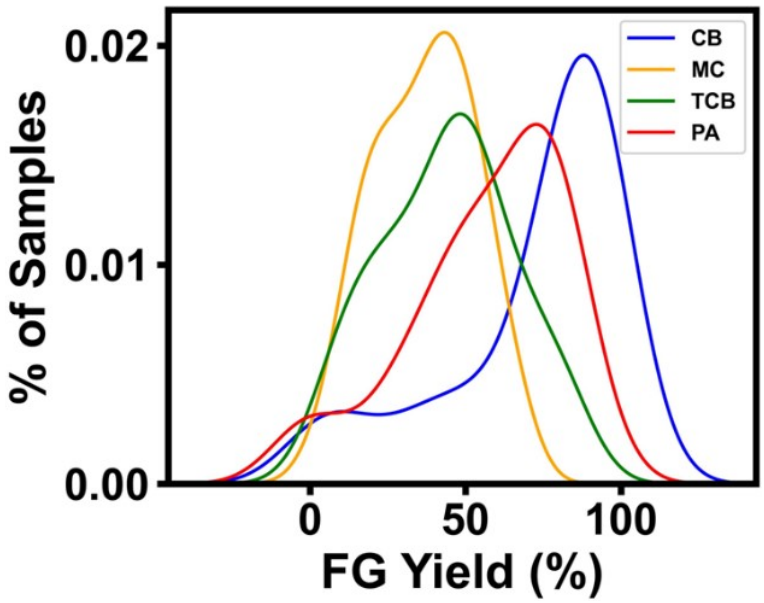


Figure S1. Distribution of the FG yield (%) synthesized from four starting materials of carbon black (CB), metallurgical coke (MC), plastic waste-derived pyrolysis ash (PA), and waste tire-based carbon black (TCB).

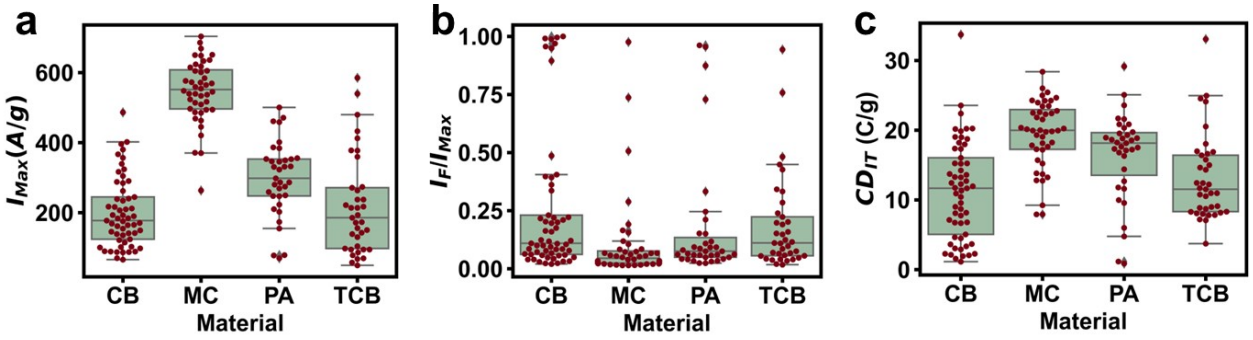


Figure S2. The mixed of box plot and swarm plot showing the distribution of (**a**) I_{Max}; (**b**) I_F/I_{Max}; and **c**) CD_{IT}. They are grouped based on the starting materials. The interquartile range (IQR=q_3-q_1) is shown as the boxes. The lower end of the box is the 1st quartile (q_1) and the upper end is the 3rd quartile (q_3). The horizontal line in the boxes show the median value. Lower and upper adjacent mark the first quartile minus 1.5 times of the IQR and third quartile plus 1.5 times of the IQR, respectively. The rest of the individual points beyond the whiskers are outliers according to the mentioned 1.5*IQR rule.[1]

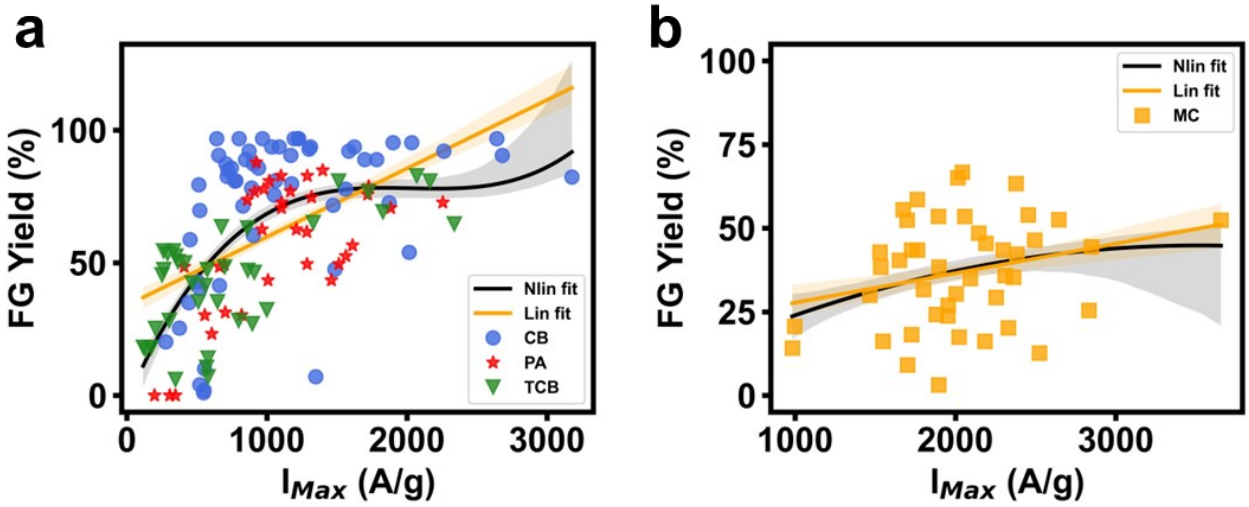


Figure S3. Plots of the FG yield vs. I_{Max} colored by the starting materials for (**a**) reactions with CB, PA, or TCB with a strong significant correlation and (**b**) MC with no significant correlation. The scattered data in both figures is fitted with both linear (orange) and non-linear curves (black) functions.

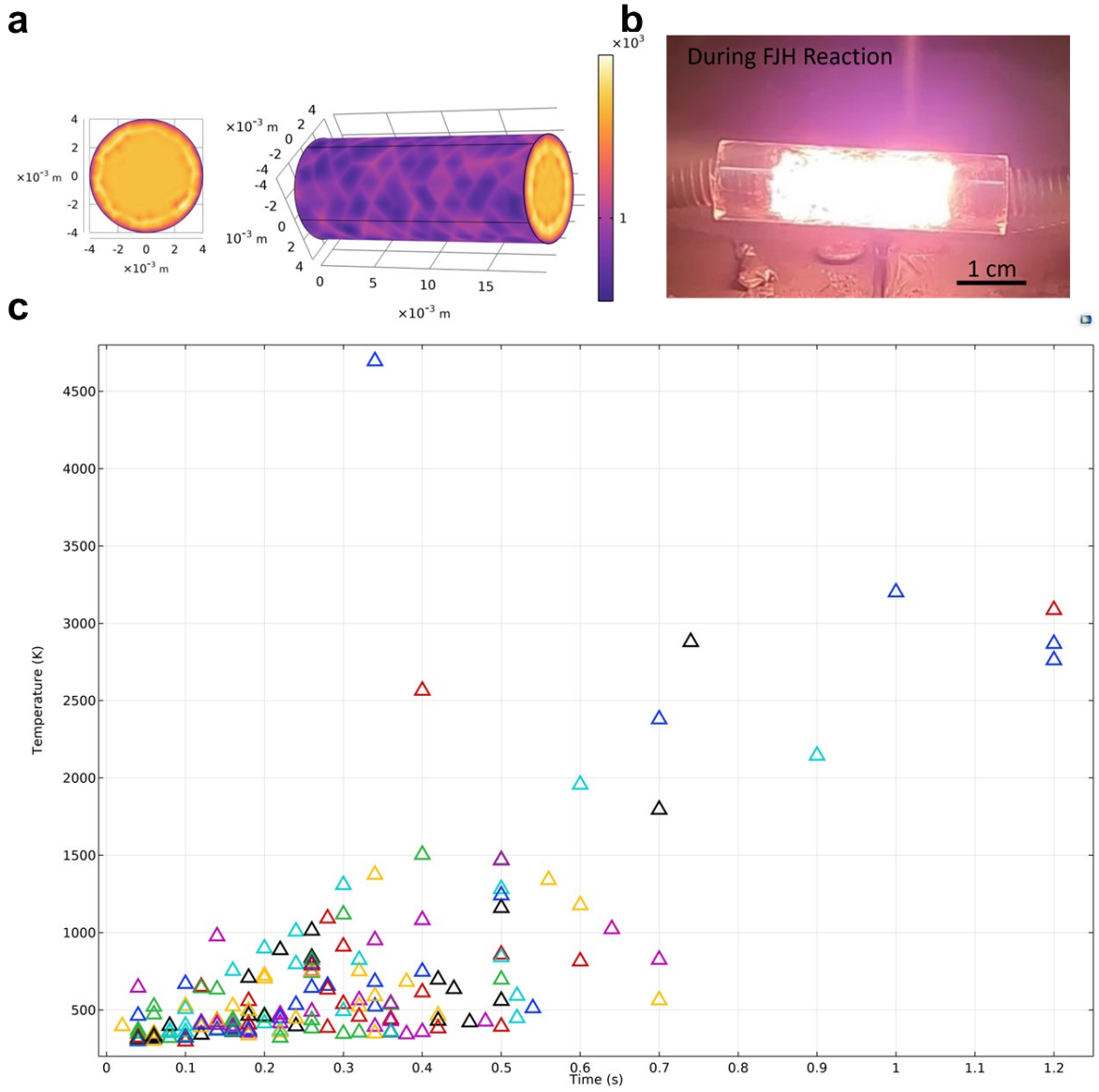


Figure S4. Multi-physics simulation of the temperature in the FJH process. (**a**) A reaction cylinder of diameter ~8 mm with length ~20 mm was used for simulation. The length of the simulated area was modified based on the starting materials' mass and particle size. (**b**) A photograph of FJH apparatus during the flashing. (**c**) Simulated temperatures of all 173 reactions over their pulse time.

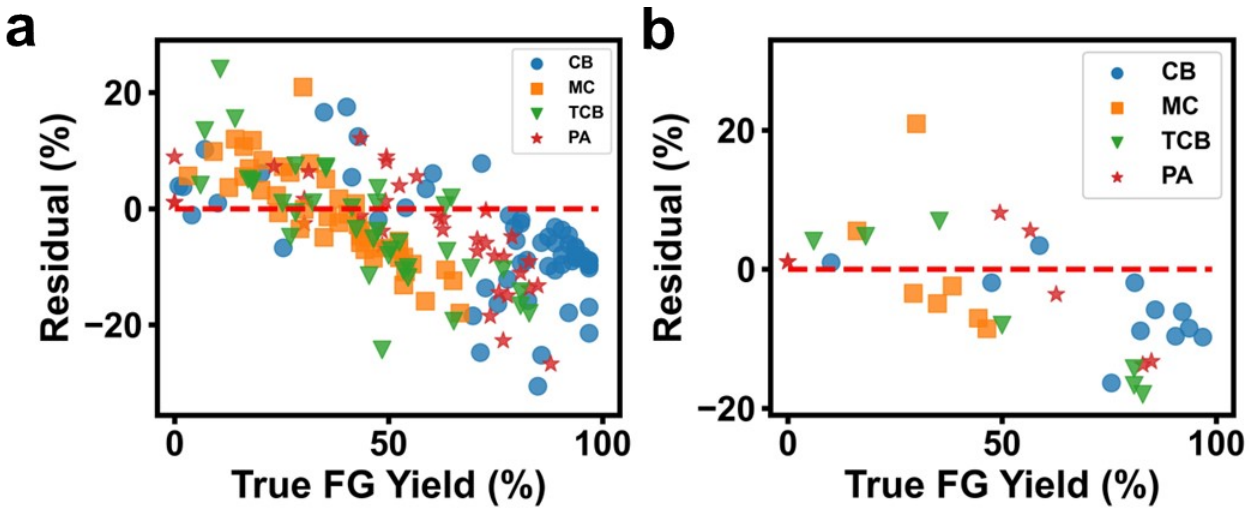
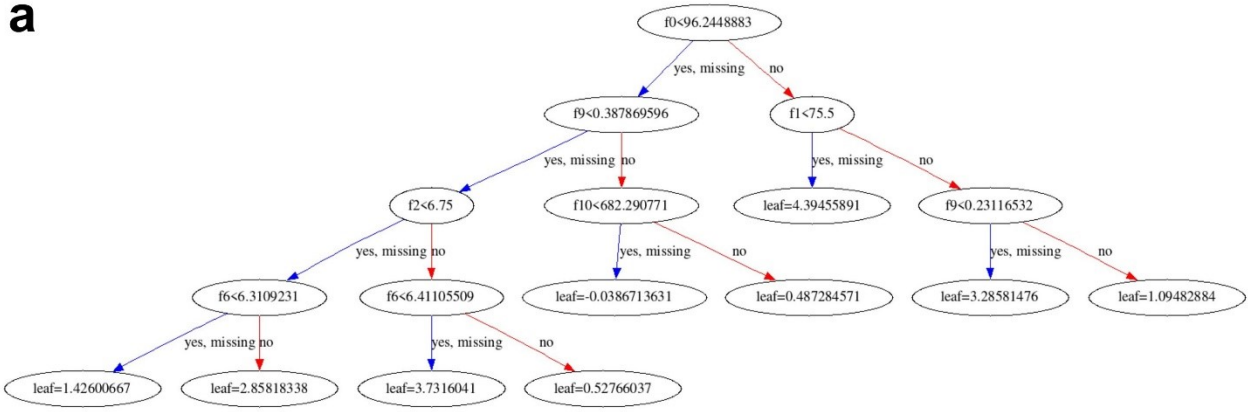


Figure S5. Error distribution of the final model when comparing the predicted versus the experimental FG yields for (**a**) all the samples and (**b**) testing samples.

a



b

Symbol	Feature	Range	
		Min	Max
f0	CD_0 (*C/g*)	19.9	204.4
f1	M_{PS} (*μm*)	45.0	150.0
f2	M_R (*Ω*)	0.4	7.2
f3	M_{Sp2} (*%*)	30.6	45.9
f4	*t* (*ms*)	30	1200
f5	V_{Pre} (*V*)	0.0	400
f6	*Log-scaled* T_{Sim} (*K*)	2.5	3.8
f7	*Atm==air (dummy)*	0 or 1	
f8	*Atm==argon (dummy)*	0 or 1	
f9	I_F/I_{Max}	0.1	0.7
f10	I_{Max} (*A/g*)	253.9	2182.4
f11	CD_{IT} (*C/g*)	11.3	96.8
f12	*H* (*J*)	80	60500

Figure S6. (**a**) An example of a decision tree used in the XGBoost model. In our case, 36 decision trees were assembled to predict the final FG yield. (**b**) The index defining features and their ranges used in Fig. S6a.

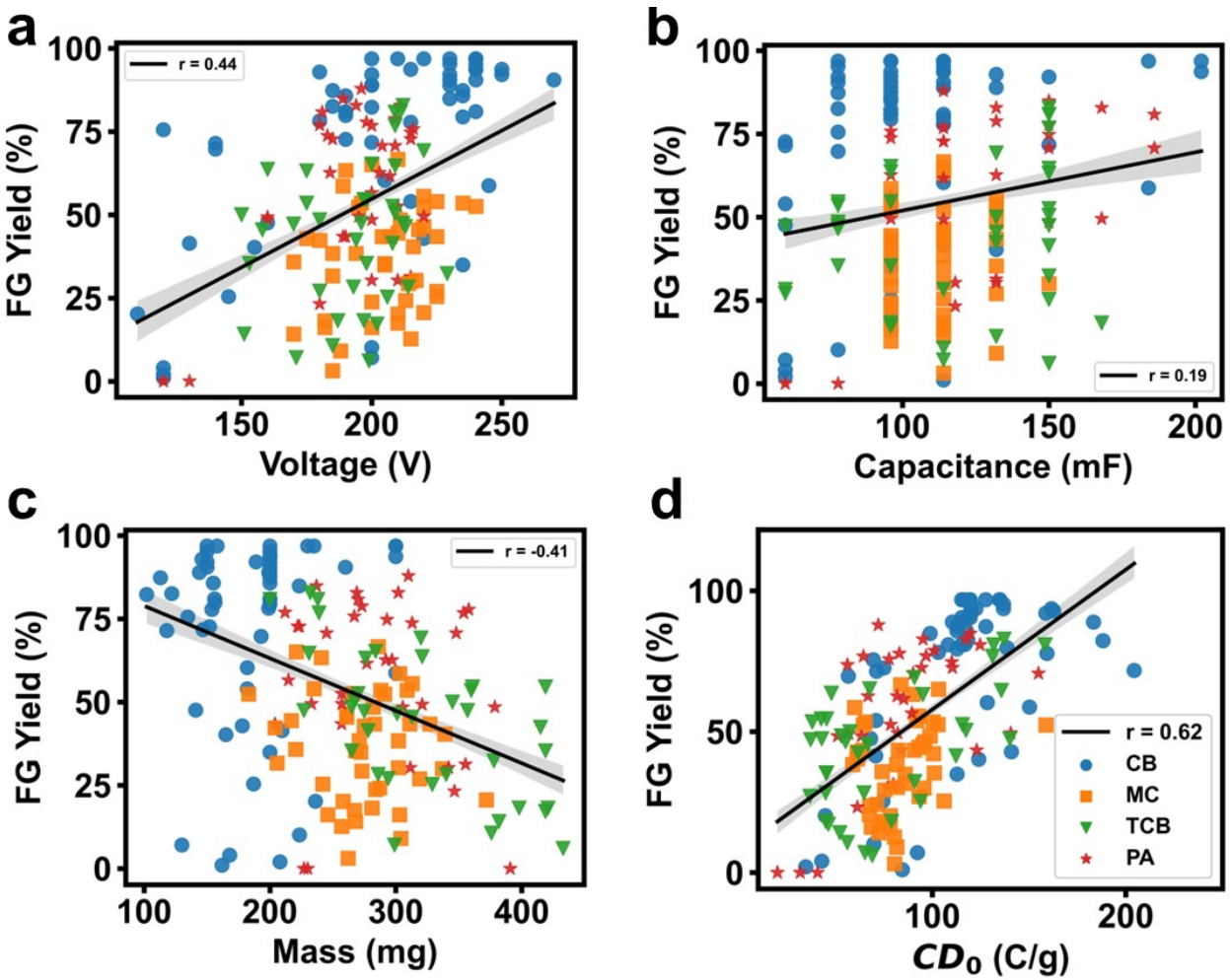


Figure S7. Distribution of parameters that define charge density (CD_0): (**a**) voltage; (**b**) capacitance; (**c**) mass of starting materials. (**d**) Correlation of the FG yield with CD_0. The line shows fitted linear central tendency and the margins show their confidence interval. The high correlation of the graphene yield with CD_0 shows its importance in the accuracy of the final model prediction.

Supplementary Tables

Table S1. Physical properties of the starting materials.

Starting material	**Particle size (PS) (μm)**	**Sample resistance (R) (Ω)**	**Surface area (SA) (m2/g)**	**Percent SP^2 (%)**
Carbon black (CB) BP-2000 raw	45	2.8	1750	41.2
Pyrolysis ash (PA) raw	125	7.2	62	42.4
Tire-based CB (TCB) raw	106	6.3	74	30.6
Metallurgical coke (MC) raw	150	0.4	18	45.9

Note: Particle size was measured by sieving. Resistance was measured by a simple multimeter. Brunauer-Emmett-Teller (BET)[2] was applied to measure the surface area. SP^2 percentage was measured from fitting the CKLL edge in the XPS spectra, known as the D-parameter.

Table S2. Hyperparameters of the trained three proxy models.

Proxy models	**Hyperparameters**
XGB predicting I_{Max}	max_depth=5, min_child_weight=12, n_estimator=25, learning_rate=0.099223, gamma=0.001, subsample=0.7
XGB predicting I_F/I_{Max}	max_depth=3, min_child_weight=9, n_estimator=29, learning_rate=0.099444, gamma=0.001, subsample=0.77
XGB predicting CD_{IT}	max_depth=4, min_child_weight=3, n_estimator=30, learning_rate=0.09947, gamma=0.001, subsample=0.75

Table S3. Hyperparameters of the trained six models used for the FG yield prediction.

Final models	**Hyperparameters**	**Range**
XGBoost (XGB)	max_depth=6, min_child_weight=3, n_estimator=40, learning_rate=0.086036, gamma=0.001, subsample=0.775789	Max_depth=np.arange(3, 7, 1), min_samples_split=np.arange(2, 5, 1), n_estimators=np.arange(35, 45, 1) learning_rate=np.arange(0.07, 0.2, 0.005) gamma=np.arange(0.0001, 1, 0.0001) subsample=np.arange(0.6, 0.85, 0.01)
Random Forest (RF)	max_depth=8, n_estimators=12, min_samples_split=4	Max_depth=np.arange(3, 10, 1), n_estimators=np.arange(10, 65, 2), min_samples_split=np.arange(2, 5, 1)
Decision Tree (DT)	max_depth=3, min_sample_split=10	Max_depth=np.arange(3, 10, 1), min_samples_split=np.arange(2, 11, 1)
Linear Regression (LR)	fit_intercept=True	N/A
Multilayer Perceptron (MLP)	hidden_layer_size=(200, 200, 200), activation=”relu”, alpha=0.12 learning_rate=”constant”, solver=”adam”, n_iter_no_change=6	Hidden_layer_sizes=[(100, 100, 100), (10,10,10), (50,50,50), (200, 200, 200)], activation=['relu','tanh','logistic'], alpha=[0.05, 0.10, 0.12, 0.15], learning_rate=['constant', 'adaptive'], solver=’adam’ n_iter_no_change=np.arange(3, 7, 1)
Bayesian Regression (BR)	Default	N/A

Note: the reported np.arange (start, stop, step) consists of a sequence of numbers that starts counting at start (inclusive) and increments according to step until it reaches stop (non-inclusive).

Table S4. The selected input features and their source of calculation

Parameters	Name	Symbol (unit)	Source
Direct	Nominal input charge density	CD_0 (C/g)	Calculated from FJH input parameters as: $CD_0 = V_0 \times C/m$
	Nominal input heat	H (J)	Calculated from FJH input parameters as: $H = V_0^2/(M_R \times t)$
	Sample particle size	M_{PS} (μm)	Measured from precursor materials
	Sample resistance	M_R (Ω)	Measured from precursor materials
	Sample Surface area	M_{SA} (m2/g)	Measured from precursor materials
	Sample Percent SP^2	M_{SP2} (%)	Measured from precursor materials
	Pulse time	t (ms)	FJH input parameters
	Pretreatment voltage	V_{Pre} (V)	FJH input parameters
	Atmosphere type	Atm. (categorical)	FJH input parameters
Physics-informed feature	Simulated temperature	$T_{Sim.}$ (K)	Simulated from FEM using Electrical-thermal multi-physics
Indirect: predicted from proxy ML models	Charge density	CD_{IT} (C/g)	Area under the current-time curve normalized by the mass
	Maximum current	I_{Max} (A/g)	Maximum current from the current-time curve normalized by the mass
	Ratio of final current to maximum current	I_F/I_{Max}	Ratio of the final to maximum current from the current-time curve

Supplementary References

1 Grubbs, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11**, 1-21, (1969).

2 Brunauer, S., Emmett, P. H. & Teller, E. Adsorption of Gases in Multimolecular Layers. *J. Am. Chem. Soc.* **60**, 309-319, (1938).