1 Input columns used in data cleaning

- Cell_area_measured
- ETL_stack_sequence
- HTL_stack_sequence
- Substrate_stack_sequence
- Backcontact_stack_sequence
- Backcontact_thickness_list
- Perovskite_composition_a_ions
- Perovskite_composition_a_ions_coefficients
- Perovskite_composition_b_ions
- Perovskite_composition_b_ions_coefficients
- Perovskite_composition_c_ions
- Perovskite_composition_c_ions_coefficients
- Perovskite_additives_compounds
- Perovskite_single_crystal
- Perovskite_composition_perovskite_ABC3_structure
- Perovskite_composition_perovskite_inspired_structure
- Perovskite_band_gap
- Perovskite_deposition_quenching_induced_crystallisation
- Perovskite_deposition_quenching_media
- Perovskite_deposition_solvent_annealing
- Perovskite_deposition_number_of_deposition_steps
- Perovskite_deposition_solvents
- Perovskite_deposition_solvents_mixing_ratios
- Perovskite_deposition_thermal_annealing_temperature
- Perovskite_deposition_thermal_annealing_time
- JV_reverse_scan_Jsc

2 Columns removed for duplicate check

- Ref_ID_temp
- Ref_ID
- Ref_name_of_person_entering_the_data
- Ref_date_entered_by_author

3 Device configurations for distribution plots

Distribution A

Column	*Value*	*Unit*
a_MA_L0	1.0	
b_Pb_L0	1.0	
c_I_L0	3.0	
bandgap_L0	1.6	eV
Back_contact_Ag_L0	80.0	nm
Depo_solvent_GBL_L0	0.7	
Depo_solvent_DMSO_L0	0.3	
Temp_100_L0	10.0	min
Substrate_stack_SLG_L0	1.0	
Substrate_stack_FTO_L1	1.0	
ETL_stack_sequence_PCBM-60_L0	1.0	
ETL_stack_sequence_BCP_L1	1.0	
HTL_stack_sequence_NiMgLiO_L0	1.0	
Quenching_media_Toluene	1.0	
Cell_area_measured	0.09	cm^2
Perovskite_quenching_induced_crystallisation	1.0	
Perovskite_deposition_number_of_steps	1.0	
Perovskite_composition_perovskite_ABC3_structure	1.0	
JV_reverse_scan_Jsc	18.32	$\frac{mA}{cm^2}$

Distribution B

Column	*Value*	*Unit*
a_MA_L0	1.0	
b_Pb_L0	1.0	
c_I_L0	3.0	
bandgap_L0	1.6	eV
Back_contact_Au_L0	80.0	nm
Depo_solvent_DMF_L0	08	
Depo_solvent_DMSO_L0	0.2	
Temp_100_L0	10.0	min
Substrate_stack_SLG_L0	1.0	
Substrate_stack_FTO_L1	1.0	
ETL_stack_sequence_TiO2-c_L0	1.0	
ETL_stack_sequence_TiO2-mp_L1	1.0	
HTL_stack_sequence_Spiro-MeOTAD_L0	1.0	
Quenching_media_Ethyl_acetate	1.0	
Cell_area_measured	0.1	cm^2
Perovskite_quenching_induced_crystallisation	1.0	
Perovskite_deposition_number_of_steps	1.0	
Perovskite_composition_perovskite_ABC3_structure	1.0	
JV_reverse_scan_Jsc	21.12	$\frac{mA}{cm^2}$

Distribution C

Column	*Value*	*Unit*
a_MA_L0	1.0	
b_Pb_L0	1.0	
c_I_L0	3.0	
bandgap_L0	1.6	eV
Back_contact_Au_L0	80.0	nm
Depo_solvent_DMSO_L0	1.0	
Temp_100_L0	20.0	min
Substrate_stack_SLG_L0	1.0	
Substrate_stack_FTO_L1	1.0	
ETL_stack_sequence_TiO2-c_L0	1.0	
ETL_stack_sequence_TiO2-mp_L1	1.0	
HTL_stack_sequence_PTAA_L0	1.0	
Quenching_media_Chlorobenzene	1.0	
Cell_area_measured	0.16	cm^2
Perovskite_quenching_induced_crystallisation	1.0	
Perovskite_deposition_number_of_steps	1.0	
Perovskite_composition_perovskite_ABC3_structure	1.0	
JV_reverse_scan_Jsc	22.41	$\frac{mA}{cm^2}$

Distribution D

Column	*Value*	*Unit*
a_MA_L0	1.0	
b_Pb_L0	1.0	
c_I_L0	3.0	
bandgap_L0	1.6	eV
Back_contact_Au_L0	60.0	nm
Depo_solvent_IPA_L0	1.0	
Temp_70_L0	30.0	min
Substrate_stack_SLG_L0	1.0	
Substrate_stack_FTO_L1	1.0	
ETL_stack_sequence_TiO2-c_L0	1.0	
ETL_stack_sequence_TiO2-mp_L1	1.0	
HTL_stack_sequence_Spiro-MeOTAD_L0	1.0	
Cell_area_measured	0.04	cm^2
Perovskite_deposition_number_of_steps	2.0	
Perovskite_composition_perovskite_ABC3_structure	1.0	
JV_reverse_scan_Jsc	16.42	$\frac{mA}{cm^2}$

4 Optuna search space

XGBoost

Parameter	*Value Range*
Max depth	1 - 15
Learning rate	0.01 - 1.0
Number of estimators	50 - 1000
Min child weight	1 - 10
Gamma	10^{-8} - 1.0
Subsample	0.5 - 0.9
colsample_bytree	0.5 - 0.9

Neural Network

Parameter	*Value Range*
Number of layers	1 - 3
Number of neurons	input size - 2*input size
Dropout percentage	0.2 - 0.5
Optimizer	Adam, RMSprop, SGD
Learning Rate	10^{-5} - 0.1

Gaussian Process Regressor

Parameter	*Value Range*
Number of kernels	1 - 3
Kernel types	Matern, Rational Quadratic, White
Matern length scale	10^{-5} - 10^{5}
Matern nu	0.5 - 5
Rational quadratic length scale	10^{-5} - 10^{5}
Rational quadratic alpha	10^{-5} - 10^{5}
White noise level	10^{-5} - 10^{5}
Alpha	10^{-3} - 10^{3}
Number of restarts	0 - 10

5 Final model configurations

XGBoost

Parameter	*Value*
Max depth	8
Learning rate	0.325
Number of estimators	525
Min child weight	6
Gamma	6.15×10^{-5}
Subsample	0.89
colsample_bytree	0.73

Neural Network

Parameter	*Value*
Number of layers	1
Number of neurons L1	4147
Dropout percentage L1	0.361
Optimizer	RMSprop
Learning Rate	8.021×10^{-5}

Gaussian Process Regressor

Parameter	*Value*
Number of kernels	1
Kernel type L1	Matern
Length scale L1	87087.9
Nu L1	4.475
White kernel noise level	45167.8
Alpha	988
Number of restarts	0