Supplementary information for

Accelerating materials discovery for water crisis: Multi-objective machine learning for atmospheric water harvesting by MOFs

Fatemeh Keshavarz,[a,*,†] Charalampos G. Livas,[b,†] Emmanuel Tylianakis,[c] Bernardo Barbiellini[a,d], and George E. Froudakis[b]

[a] Department of Physics, School of Engineering Science, Lappeenranta-Lahti University of Technology (LUT), Yliopistonkatu 34, FI-53850 Lappeenranta, Finland

[b] Department of Chemistry, Voutes Campus, University of Crete, Heraklion 71003, Greece

[c] Department of Materials Science and Engineering, Voutes Campus, University of Crete, Heraklion 71003, Greece

[d] Department of Physics, Northeastern University, Boston, MA 02115, USA

*Corresponding author: fatemeh.keshavarz@lut.fi

† F. K. and C. G. L. contributed equally.

Table S1 The performance of the tested regression models in prediction of water uptake at RH = 20% (Wup20cm3) and RH = 80% (Wup80cm3) and water selectivity (Wselectivity), prior to hyperparameter optimization. Their performance for each target is assessed based on the mean R^2 values of the test sets and their standard deviations ($R^2 \pm \sigma$) and the average R^2 (R^2_{avg}) value for each model averaged over the three targets.

	Wup20cm3		Wup80cm3		Wselectivity		*All*
Model	R^2	σ	R^2	σ	R^2	σ	R^2_{avg}
RandomForestRegressor	0.6823	0.0196	0.7033	0.0224	0.8173	0.0072	0.734
KNeighborsRegressor	0.5347	0.0278	0.5242	0.0369	0.3639	0.0167	0.474
LinearRegression	0.3529	0.0317	0.4257	0.0378	0.4325	0.0094	0.404
ExtraTreesRegressor	0.7101	0.0233	0.7183	0.0252	0.8143	0.0066	0.748
DecisionTreeRegressor	0.3885	0.0540	0.3975	0.0645	0.6377	0.0165	0.475
XGBRegressor	0.6943	0.0249	0.7143	0.0170	0.8126	0.0105	0.740
AdaBoostRegressor	-0.5964	0.1687	-0.2893	0.2712	0.7637	0.0062	-0.041
BaggingRegressor	0.6480	0.0292	0.6705	0.0305	0.8025	0.0086	0.707
BaggingRegressor	0.6641	0.0163	0.6920	0.0230	0.8199	0.0086	0.725
GradientBoostingRegressor	0.5797	0.0229	0.5979	0.0267	0.8129	0.0105	0.664
Ridge	0.3557	0.0292	0.4268	0.0357	0.4338	0.0096	0.405
LGBMRegressor	0.7086	0.0176	0.7154	0.0187	0.8557	0.0069	0.760
CatBoostRegressor	0.7213	0.0202	0.7270	0.0227	0.8295	0.0070	0.759
SVR	-0.0571	0.0069	-0.0735	0.0073	0.2406	0.0066	0.037
MLPRegressor	0.6304	0.0379	0.6287	0.0387	0.7526	0.0176	0.671
ElasticNet	0.1732	0.0110	0.2203	0.0117	0.3314	0.0021	0.242
HistGradientBoostingRegressor	0.7028	0.0170	0.7112	0.0202	0.8561	0.0048	0.757
NGBRegressor	0.5369	0.0198	0.5562	0.0284	0.8228	0.0092	0.639
HuberRegressor	-0.0556	0.0098	0.0849	0.0145	0.3415	0.0055	0.124

Table S2 Impact of hyperparameter optimization on the performance of the top four regression models in prediction of water uptake at RH = 20% (Wup20cm3) and RH = 80% (Wup80cm3) and water selectivity (Wselectivity). Their performance for each target is assessed based on the mean R^2 values of the test sets and their standard deviations ($R^2 \pm \sigma$).

Target	Model	R^2	σ	R^2	σ
		Default		Optimized	
Wup20cm3	LGBM	0.7086	0.0176	0.7328	0.0172
	CatBoost	0.7213	0.0202	0.7228	0.0204
	HGB	0.7028	0.0170	0.7105	0.0116
	ET	0.7101	0.0233	0.6474	0.0231
Wup80cm3	LGBM	0.7154	0.0187	0.7444	0.0228
	CatBoost	0.7270	0.0227	0.7259	0.0247
	HGB	0.7112	0.0202	0.7222	0.0209
	ET	0.7183	0.0252	0.6690	0.0237
Wselectivity	LGBM	0.8557	0.0069	0.8561	0.0080
	CatBoost	0.8295	0.0070	0.8314	0.0070
	HGB	0.8561	0.0048	0.8554	0.0065
	ET	0.8143	0.0066	0.6739	0.0046

[a] Optimized hyperparameters for each regressor: LGBM: {'num_leaves': 60, 'n_estimators': 400, 'max_depth': -1, 'learning_rate': 0.1}, CatBoost: {'learning_rate': 0.03, 'l2_leaf_reg': 5, 'iterations': 1000, 'depth': 10}, HGB: {'min_samples_leaf': 20, 'max_iter': 500, 'max_depth': None, 'learning_rate': 0.1}, and ET: {'n_estimators': 500, 'min_samples_split': 4, 'max_features': 'log2', 'max_depth': 25}

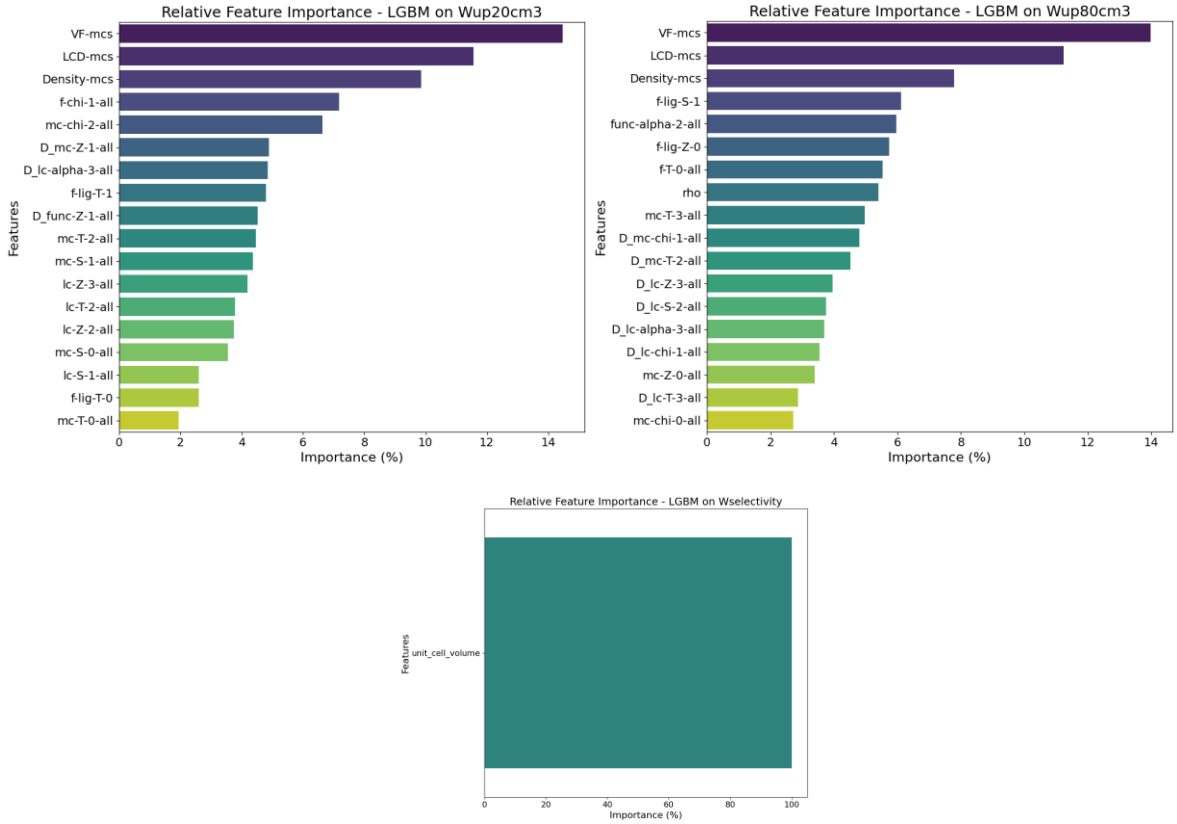

Fig. S1: The recursively selected features and their relative importances for the prediction of water uptake at RH = 20% (Wup20cm3) and RH = 80% (Wup80cm3) and water selectivity (Wselectivity) using LGBM with optimized hyperparameters.

Table S3 List, units and definitions of the structural features calculated by Zeo++.

Feature/descriptor	Unit	Definition
Di	Å	Diameter of the largest included sphere
Df	Å	Diameter of the largest free sphere
Dif	Å	Diameter of the largest included sphere along the free sphere path
unit_cell_volume	$Å^3$	The volume of the unit cell
rho	g cm^{-3}	Bulk density
VSA	m^2 cm^{-3}	Volumetric surface area
GSA	m^2 g^{-1}	Gravimetric surface area
VPOV	cm^3 cm^{-3}	Volumetric pore volume
GPOV	cm^3 g^{-1}	Gravimetric pore volume
POAV_vol_frac	unitless	Probe accessible volume fraction
PONAV_vol_frac	unitless	Probe non-accessible volume fraction
GPOAV	cm^3 g^{-1}	Gravimetric probe accessible volume
GPONAV	cm^3 g^{-1}	Gravimetric probe non-accessible volume
POAV	cm^3 cm^{-3}	Probe accessible volume
PONAV	cm^3 cm^{-3}	Probe non-accessible volume

Table S4: Description of the MOF-specific RAC features calculated by molSimplify and abbreviated as type-property-Depth for RAC calculation-scope. The features all averaged over all starting atoms and, then, all secondary building units (SBUs) or linkers. The "D" prefix before the types (D_type) indicates that the RAC feature is based on property differences. Its absence indicates that the RAC features are sums of products of the atomic properties. For example, the "D_lc-chi-3-all" feature is sum of the differences of electronegativities of the linker connecting atoms, starting from the metals, and including all linker atoms, and calculated at the depth of 3.

Type	Starting atom	Scope	Property	Depth
Full (f)	All SBU atoms	All SBU atoms	Nuclear charge (Z)	0
Metal centered (mc)	All metals	All unit cell atoms	Topology (T)	1
Linker connecting atom centered (lc)	Linker atoms bonded to metals	All linker atoms	Electronegativity (chi)	2
Full linker (f-lig)	All linker atoms	All linker atoms	Identity (I)	3
Functional group centered (func)	Non-hydrogen/carbon linker atoms not bonded to metals	All linker atoms	Covalent radius (S)	
			Polarizability (alpha)	

Table S5 The performance of the classification models for predicting 2-class acid, base and boiling stability and 4-class water stability using 5-fold cross validation and default hyperparameters.

Model		Accuracy	Precision	Recall	F1	ROC AUC		Accuracy	Precision	Recall	F1	ROC AUC
TabNet		0.3259	0.3860	0.3259	0.3216	-		0.2052	0.8879	0.2052	0.2675	0.5562
EasyEnsemble		0.4377	0.5226	0.4377	0.4566	-		0.6465	0.9189	0.6465	0.7382	0.7136
BalancedRF		0.5723	0.6122	0.5723	0.5825	-		0.8590	0.9164	0.8590	0.8830	0.7604
RandomForest		0.6520	0.6497	0.652	0.6269	-		0.9451	0.9389	0.9451	0.9267	0.7892
KNeighbors		0.5686	0.5647	0.5686	0.5559	-		0.9322	0.8857	0.9322	0.9080	0.6589
LogisticRegression		0.4661	0.5366	0.4661	0.4814	-		0.7692	0.9107	0.7692	0.8237	0.7119
ExtraTrees	Water stability	0.6566	0.6539	0.6566	0.6371	-	Boiling stability	0.9451	0.9352	0.9451	0.9306	0.7786
DecisionTree		0.5448	0.5549	0.5448	0.5448	-		0.9048	0.9116	0.9048	0.9076	0.6111
XGBoost_Balanced		0.6593	0.6538	0.6593	0.6435	-		0.9414	0.9328	0.9414	0.9267	0.7316
AdaBoost		0.4917	0.4379	0.4917	0.4425	-		0.9350	0.8957	0.9350	0.9110	0.6152
Bagging_DT		0.6089	0.6056	0.6089	0.5957	-		0.9414	0.9248	0.9414	0.9239	0.7074
Bagging_RF		0.6465	0.6549	0.6465	0.6162	-		0.9432	0.9326	0.9432	0.9214	0.7444
GradientBoosting		0.6263	0.6198	0.6263	0.6036	-		0.9359	0.9171	0.9359	0.9164	0.6863
RidgeClassifier		0.5494	0.5254	0.5494	0.5171	-		0.9368	0.8879	0.9368	0.9105	-
LGBM		0.6712	0.6688	0.6712	0.6548	-		0.9423	0.9307	0.9423	0.9244	0.7427
CatBoost_Balanced		0.6538	0.6547	0.6538	0.6352	-		0.9441	0.9421	0.9441	0.9255	0.7688
SVC		0.5009	0.5626	0.5009	0.5165	-		0.7957	0.9032	0.7957	0.8400	0.6820
MLP		0.6025	0.6026	0.6025	0.5971	-		0.9313	0.9251	0.9313	0.9226	0.7013
TabNet		0.2070	0.8225	0.2070	0.2553	0.5241		0.1960	0.8819	0.1960	0.2618	0.5544
EasyEnsemble		0.7408	0.9110	0.7408	0.7980	0.8203		0.6804	0.9238	0.6804	0.7657	0.7486
BalancedRF		0.8800	0.9194	0.8800	0.8951	0.8310		0.8818	0.9279	0.8818	0.9009	0.7998
RandomForest		0.9267	0.9183	0.9267	0.9099	0.8597		0.9432	0.9112	0.9432	0.9216	0.7995
KNeighbors		0.9249	0.9111	0.9249	0.9080	0.7920		0.9377	0.9106	0.9377	0.9173	0.7109
LogisticRegression		0.8343	0.9083	0.8343	0.8625	0.7956		0.8077	0.9141	0.8077	0.8516	0.6903
ExtraTrees	Acid stability	0.9295	0.9220	0.9295	0.9176	0.8565	Base stability	0.9414	0.9139	0.9414	0.9230	0.8255
DecisionTree		0.8938	0.8939	0.8938	0.8938	0.6377		0.8984	0.9128	0.8984	0.9048	0.5929
XGBoost_Balanced		0.9304	0.9230	0.9304	0.9241	0.8486		0.9451	0.9311	0.9451	0.9338	0.7881
AdaBoost		0.9139	0.8866	0.9139	0.8909	0.8113		0.9377	0.9032	0.9377	0.9182	0.7453
Bagging_DT		0.9313	0.9207	0.9313	0.9197	0.8008		0.9368	0.9046	0.9368	0.9179	0.7307
Bagging_RF		0.9258	0.9112	0.9258	0.9081	0.8326		0.9432	0.9189	0.9432	0.9218	0.7773
GradientBoosting		0.9277	0.9160	0.9277	0.9124	0.8301		0.9368	0.9161	0.9368	0.9199	0.7710
RidgeClassifier		0.9222	0.9107	0.9222	0.9000	-		0.9386	0.9008	0.9386	0.9164	-
LGBM		0.9267	0.9151	0.9267	0.9145	0.8574		0.9423	0.9203	0.9423	0.9235	0.8066
CatBoost_Balanced		0.9295	0.9188	0.9295	0.9172	0.8618		0.9423	0.9088	0.9423	0.922	0.8272
SVC		0.8736	0.9124	0.8736	0.8889	0.8466		0.8645	0.9218	0.8645	0.8885	0.7911
MLP		0.9304	0.9247	0.9304	0.9262	0.8046		0.9377	0.9271	0.9377	0.9292	0.7492

Table S6 10-fold vs. 5-fold cross validation for predicting 2-class acid, base, boiling and 4-class water stability using the top four classification models with default hyperparameters.

	Model	Accuracy	Precision	Recall	F1	ROC AUC	Accuracy	Precision	Recall	F1	ROC AUC
		10-fold					5-fold				
Water stability	RF	0.6603	0.6625	0.6603	0.6362	-	0.6520	0.6497	0.6520	0.6269	-
	ET	0.6703	0.6632	0.6703	0.6475	-	0.6566	0.6539	0.6566	0.6371	-
	XGBB	0.6713	0.6747	0.6713	0.6578	-	0.6593	0.6538	0.6593	0.6435	-
	LGBM	0.6814	0.6818	0.6814	0.6661	-	0.6712	0.6688	0.6712	0.6548	-
Acid stability	RF	0.9313	0.9243	0.9313	0.9156	0.8675	0.9267	0.9183	0.9267	0.9099	0.8597
	ET	0.9295	0.9205	0.9295	0.9155	0.8696	0.9295	0.9220	0.9295	0.9176	0.8565
	XGBB	0.9350	0.9273	0.9350	0.9284	0.8578	0.9304	0.9230	0.9304	0.9241	0.8486
	LGBM	0.9341	0.9289	0.9341	0.9212	0.8536	0.9267	0.9151	0.9267	0.9145	0.8574
Base stability	RF	0.9423	0.9101	0.9423	0.9240	0.8377	0.9432	0.9112	0.9432	0.9216	0.7995
	ET	0.9405	0.9220	0.9405	0.9246	0.8638	0.9414	0.9139	0.9414	0.923	0.8255
	XGBB	0.9441	0.9288	0.9441	0.9325	0.8075	0.9451	0.9311	0.9451	0.9338	0.7881
	LGBM	0.9423	0.9145	0.9423	0.9244	0.8185	0.9423	0.9203	0.9423	0.9235	0.8066
Boiling stability	RF	0.9451	0.9371	0.9451	0.9266	0.8214	0.9451	0.9389	0.9451	0.9267	0.7892
	ET	0.9442	0.9398	0.9442	0.9292	0.8267	0.9451	0.9352	0.9451	0.9306	0.7786
	XGBB	0.9424	0.9364	0.9424	0.9308	0.7911	0.9414	0.9328	0.9414	0.9267	0.7316
	LGBM	0.9378	0.9118	0.9378	0.9179	0.8128	0.9423	0.9307	0.9423	0.9244	0.7427

Table S7 Impact of hyperparameter optimization on the performance of the top four classification models for predicting acid, base, boiling and 4-class water stability.

Target Stability	Model	Accuracy	Precision	Recall	F1	ROC AUC	Accuracy	Precision	Recall	F1	ROC AUC
		Optimized hyperparameters					Default hyperparameters				
Water	RF	0.6393	0.6379	0.6393	0.6326	-	0.6603	0.6625	0.6603	0.6362	-
Water	ET	0.6786	0.6714	0.6786	0.6583	-	0.6703	0.6632	0.6703	0.6475	-
Water	XGBoost	0.6768	0.6787	0.6768	0.6602	-	0.6713	0.6747	0.6713	0.6578	-
Water	LGBM	0.6786	0.6773	0.6786	0.6639	-	0.6814	0.6818	0.6814	0.6661	-
Acid	RF	0.9323	0.9300	0.9323	0.9271	0.8677	0.9313	0.9243	0.9313	0.9156	0.8675
Acid	ET	0.9313	0.9241	0.9313	0.9186	0.8617	0.9295	0.9205	0.9295	0.9155	0.8696
Acid	XGBoost	0.9341	0.9269	0.9341	0.9275	0.8552	0.9350	0.9273	0.935	0.9284	0.8578
Acid	LGBM	0.9322	0.9259	0.9322	0.9198	0.8679	0.9341	0.9289	0.9341	0.9212	0.8536
Base	RF	0.9441	0.9285	0.9441	0.9318	0.8446	0.9423	0.9101	0.9423	0.924	0.8377
Base	ET	0.9395	0.9153	0.9395	0.9216	0.8630	0.9405	0.922	0.9405	0.9246	0.8638
Base	XGBoost	0.9441	0.9285	0.9441	0.9338	0.8084	0.9441	0.9288	0.9441	0.9325	0.8075
Base	LGBM	0.9441	0.9200	0.9441	0.9268	0.8292	0.9423	0.9145	0.9423	0.9244	0.8185
Boiling	RF	0.9442	0.9385	0.9442	0.9284	0.8009	0.9451	0.9371	0.9451	0.9266	0.8214
Boiling	ET	0.9414	0.9345	0.9414	0.9270	0.8116	0.9442	0.9398	0.9442	0.9292	0.8267
Boiling	XGBoost	0.9378	0.9256	0.9378	0.9262	0.7989	0.9424	0.9364	0.9424	0.9308	0.7911
Boiling	LGBM	0.9396	0.9154	0.9396	0.9201	0.7971	0.9378	0.9118	0.9378	0.9179	0.8128

Table S8 The impact of the SMOTE technique on the stability predictions of the LGBM and RF models with default hyperparameters.

Target Stability	Model	Accuracy	Precision	Recall	F1	ROC AUC	Accuracy	Precision	Recall	F1	ROC AUC
		Without SMOTE treatment/ 4-class					With SMOTE treatment/ 4-class				
Water	RF	0.6603	0.6625	0.6603	0.6362	-	0.6567	0.6670	0.6567	0.6576	-
Water	LGBM	0.6814	0.6818	0.6814	0.6661	-	0.6685	0.6713	0.6685	0.6647	-
		Without SMOTE treatment/ 2-class					With SMOTE treatment/ 2-class				
Water	RF	0.7637	0.7669	0.7637	0.7570	0.8360	0.7463	0.7530	0.7463	0.7469	0.8304
Water	LGBM	0.7839	0.7851	0.7839	0.7801	0.8375	0.7775	0.7801	0.7775	0.7774	0.8423
Acid	RF	0.9313	0.9243	0.9313	0.9156	0.8675	0.9158	0.9159	0.9158	0.9116	0.8516
Acid	LGBM	0.9295	0.9205	0.9295	0.9155	0.8696	0.9222	0.9209	0.9222	0.9192	0.8649
Base	RF	0.9423	0.9101	0.9423	0.9240	0.8377	0.9395	0.9352	0.9395	0.9337	0.8512
Base	LGBM	0.9423	0.9145	0.9423	0.9244	0.8185	0.9405	0.9315	0.9405	0.9337	0.8484
Boiling	RF	0.9451	0.9371	0.9451	0.9266	0.8214	0.9259	0.9163	0.9259	0.9180	0.8242
Boiling	LGBM	0.9378	0.9118	0.9378	0.9179	0.8128	0.9304	0.9227	0.9304	0.9233	0.8221

Table S9 Impact of Recursive Feature Elimination (RCE) with 10-fold cross-validation on the performance of the LGBM and RF classification models for predicting acid, base, boiling and 4-class water stability.

Target Stability	Model	NF	Accuracy	Precision	Recall	F1	ROC AUC	Accuracy	Precision	Recall	F1	ROC AUC
			Selective number of features (NF)					Full features				
Water	RF	44	0.6493	0.6463	0.6493	0.6258	-	0.6603	0.6625	0.6603	0.6362	-
Water	LGBM	73	0.6795	0.6800	0.6795	0.6619	-	0.6814	0.6818	0.6814	0.6661	-
Acid	RF	138	0.9295	0.9215	0.9295	0.9143	0.8637	0.9313	0.9243	0.9313	0.9156	0.8675
Acid	LGBM	148	0.9350	0.9288	0.9350	0.9228	0.8623	0.9341	0.9289	0.9341	0.9212	0.8536
Base	RF	22	0.9469	0.9303	0.9469	0.9301	0.8408	0.9423	0.9101	0.9423	0.9240	0.8377
Base	LGBM	112	0.9414	0.9145	0.9414	0.9235	0.8314	0.9423	0.9145	0.9423	0.9244	0.8185
Boiling	RF	113	0.9433	0.9227	0.9433	0.9239	0.8308	0.9451	0.9371	0.9451	0.9266	0.8214
Boiling	LGBM	13	0.9506	0.9483	0.9506	0.9391	0.8569	0.9378	0.9118	0.9378	0.9179	0.8128

Table S10 Impact of Recursive Feature Addition (RCA) with 10-fold cross-validation on the performance of the LGBM and RF classification models (with default hyperparameters) for predicting acid, base, boiling and 4-class (Water/4) and 2-class (Water/2) water stability.

		Selective number of features (NF)						**Full features**				
Target Stability	Model	NF	Accuracy	Precision	Recall	F1	ROC AUC	Accuracy	Precision	Recall	F1	ROC AUC
Water/4	RF	7	0.6768	0.6710	0.6768	0.6619	-	0.6603	0.6625	0.6603	0.6362	-
Water/4	LGBM	8	0.6740	0.6778	0.6740	0.6633	-	0.6814	0.6818	0.6814	0.6661	-
Water/2	RF	8	0.7875	0.7850	0.8879	0.8323	0.8459	0.7637	0.7669	0.7637	0.7570	0.8360
Water/2	LGBM	10	0.7976	0.8073	0.8682	0.8357	0.8427	0.7839	0.7851	0.7839	0.7801	0.8375
Acid	RF	8	0.9433	0.9380	0.9433	0.9366	0.8274	0.9313	0.9243	0.9313	0.9156	0.8675
Acid	LGBM	9	0.9405	0.9361	0.9405	0.9342	0.8428	0.9341	0.9289	0.9341	0.9212	0.8536
Base	RF	1	0.9414	0.8898	0.9414	0.9148	0.4966	0.9423	0.9101	0.9423	0.9240	0.8377
Base	LGBM	7	0.9515	0.9454	0.9515	0.9390	0.7635	0.9423	0.9145	0.9423	0.9244	0.8185
Boiling	RF	7	0.9534	0.9481	0.9534	0.9417	0.8049	0.9451	0.9371	0.9451	0.9266	0.8214
Boiling	LGBM	10	0.9533	0.9523	0.9533	0.9418	0.7701	0.9378	0.9118	0.9378	0.9179	0.8128

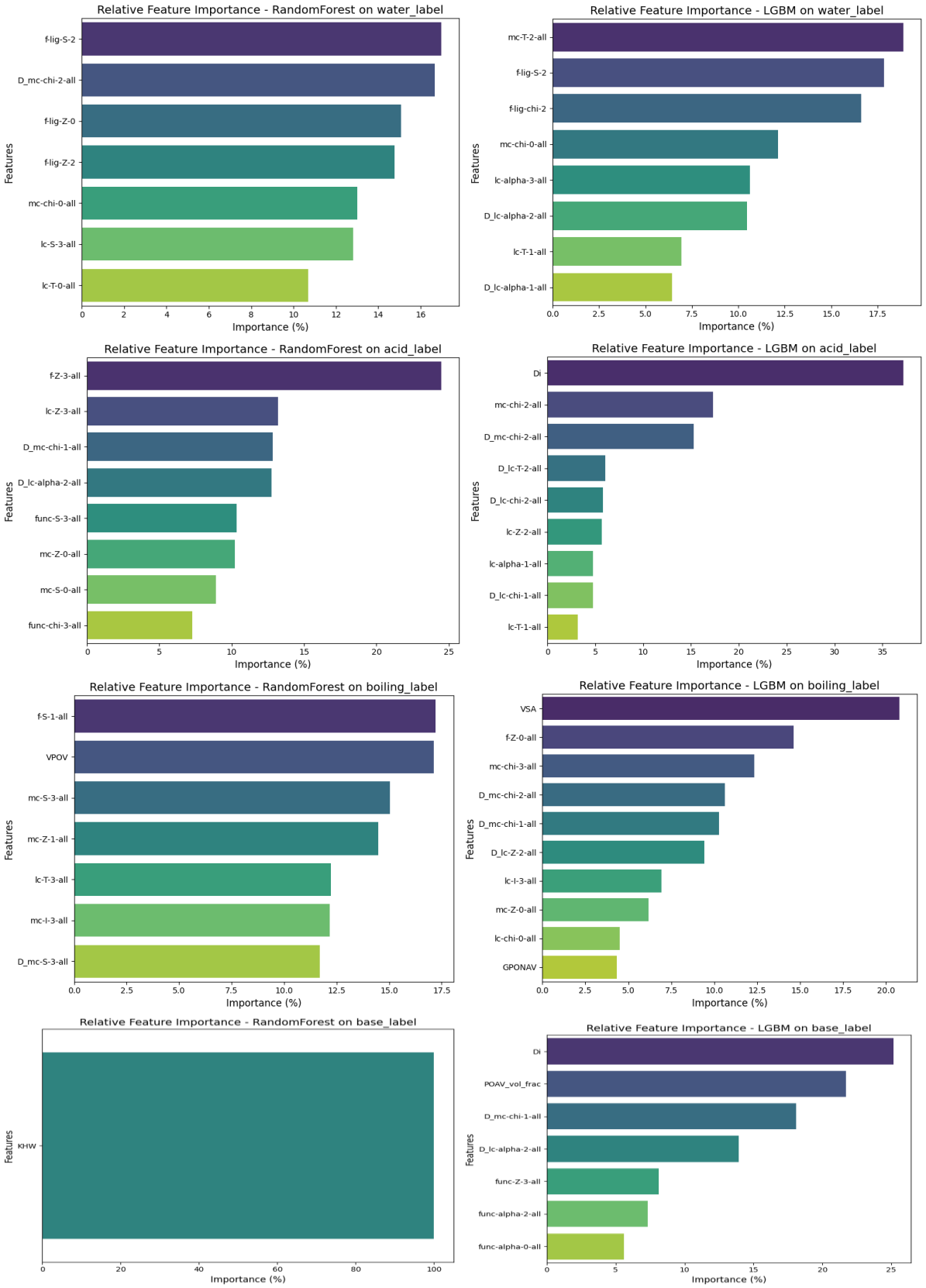

Fig. S2: The features selected based on recursive feature addition for application of the Random Forest and Light Gradient Boosting Machine (LGBM) classification models to base,

acid, boiling and 4-class water stability, and their relative importances. The features are explained in Tables S3 and S4, and the “Datasets” sub-section.

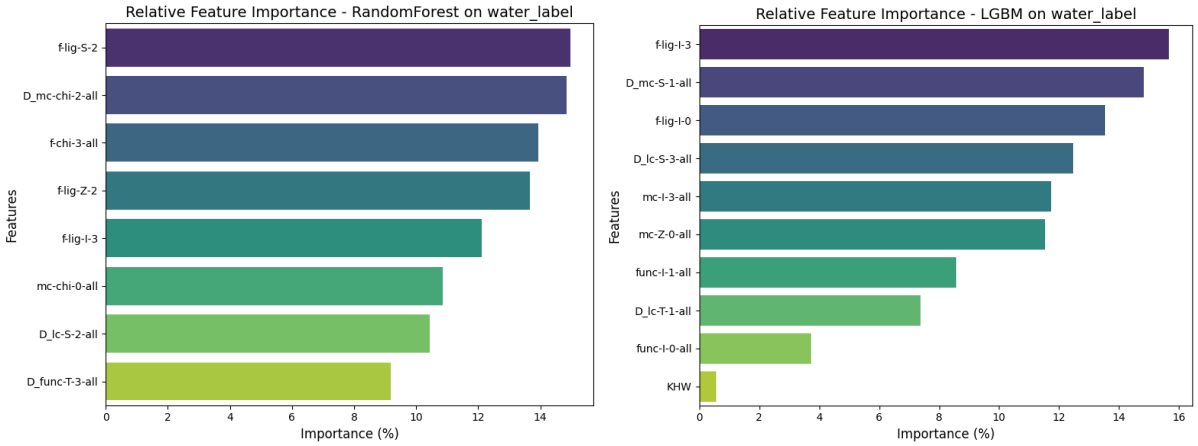

Fig. S3: The features selected based on recursive feature addition for application of the Random Forest and Light Gradient Boosting Machine (LGBM) classification models to 2-class water stability, and their relative importances. The features are explained in Tables S3 and S4, and the “Datasets” sub-section.

Table S11 The evaluated hyperparameter grids for the classification models.

Model	Hyperparameter grid
RandomForest	n_estimators: [100, 200, 300, 500]
	max_depth: [None, 10, 20, 50]
	min_samples_split: [2, 5, 10]
	min_samples_leaf: [1, 2, 4]
	bootstrap: [True, False]
ExtraTrees	n_estimators: [100, 200, 300]
	max_depth: [None, 10, 20, 50]
	min_samples_split: [2, 5, 10]
	min_samples_leaf: [1, 2, 4]
	bootstrap: [False]
XGBoost_Balanced	'n_estimators: [100, 200, 300]
	max_depth: [3, 6, 10]
	learning_rate: [0.01, 0.1, 0.2]
	subsample: [0.8, 1.0]
	colsample_bytree: [0.8, 1.0]
	gamma: [0, 1, 5]
LGBM	n_estimators: [100, 200, 300]
	num_leaves: [31, 40, 50]
	learning_rate: [0.01, 0.1, 0.2]
	max_depth: [-1, 10, 20]
	min_child_samples: [10, 20, 30]
	subsample: [0.8, 1.0]

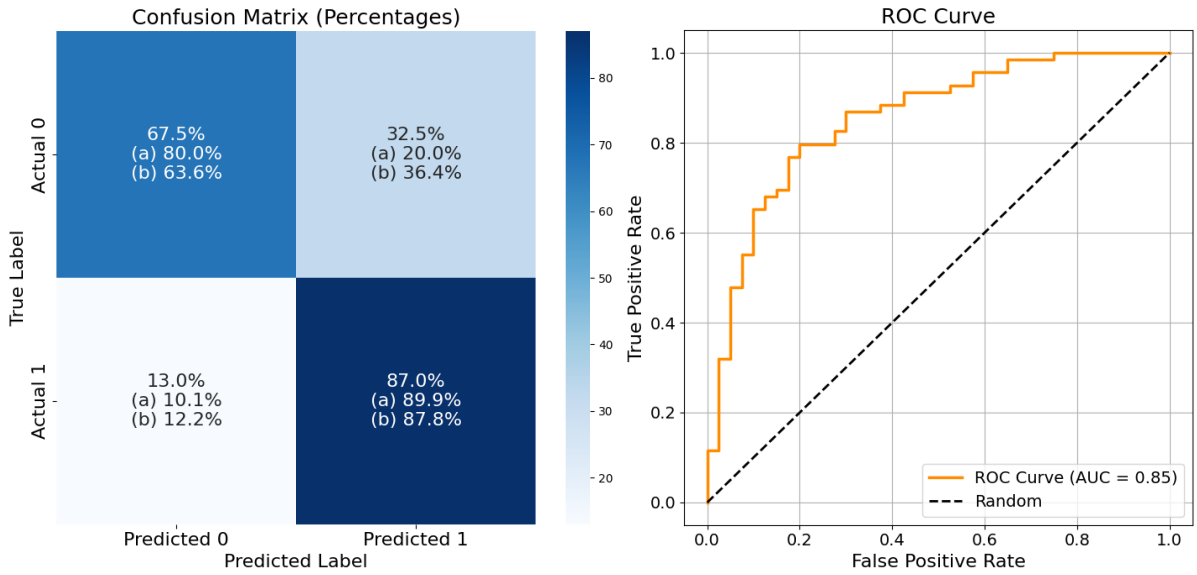

Fig. S4: The confusion matrix and ROC curve obtained from the 2-class water stability model trained using the features identified through recursive feature addition and the Light Gradient Boosting Machine (LGBM) classifier. In the confusion matrix, the (a) and (b) percentages are calculated based on the values reported by Song et al.[1] and Terrones et al.,[2] respectively.

Table S12 Evaluation of the feature generation (FGen) code and the model for prediction of the probability of water stability (WSprob) of several MOFs unseen by the model, and water harvesting MOFs.

MOF	CCDC code	FGen (error)[a]	WSprob	Water stable[REF]	Structural notes
Ga-MIL-53	704888	F (U)	–	Yes[3]	Uncoordinated water molecules (water missing hydrogens)
MOF-808	1002672	F (OA)	–	Yes[4]	Partial occupancies (metal node and uncoordinated water molecules)
MOF-808	1871192	F (OA)	–	Yes[4]	Partial occupancies (metal node and uncoordinated water molecules)
Fe-MIL-88B	1892483	Done	0.957	No[3]	Partial occupancies (ligand) and metal coordinated water molecules
MOF-303	2078717	Done	0.776	Yes[5]	–
CaSyr-2	2298353	Done	0.081	No[6]	Metal coordinated water molecules
Cr-mos-MOF-1	2362841	Done	0.847	Yes[7]	Insignificant partial occupancies (metal node's functional group)
Fe-mos-MOF-1	2362840	F (OA)	–	No[7]	Significant partial occupancies (metal node's functional group)
$Cu(INA)_2(H_2O)_4$	2369691	F (U)	–	Yes[8]	Metal coordinated water molecules
Cu_2TBAPy	2403517	F (OA)	–	Yes[9]	Extensive partial occupancies (ligand)
AR-2	2430163	F (Q)	–	Yes[10]	Metal coordinated water molecules
Tb-H4btc	2431923	Done	0.910	No[b,7]	Partial occupancies (ligand) and missing hydrogen
Tb-H4btc-Asp	2449649	Done	0.797	No[b,7]	Missing hydrogens
UoC-13(1F)	–	Done	0.060	No[11]	Partial occupancies (ligand)
Zn-UoC-7(2F)	–	F (U)	–	No[11]	Partial occupancies (metal node)

[a] F: failed, U: unknown (no specific error outlined), OA: overlapping atoms, and Q: "?" string to float conversion failure.
[b] Stable in the presence of a protective ion-imprinted polymer layer.

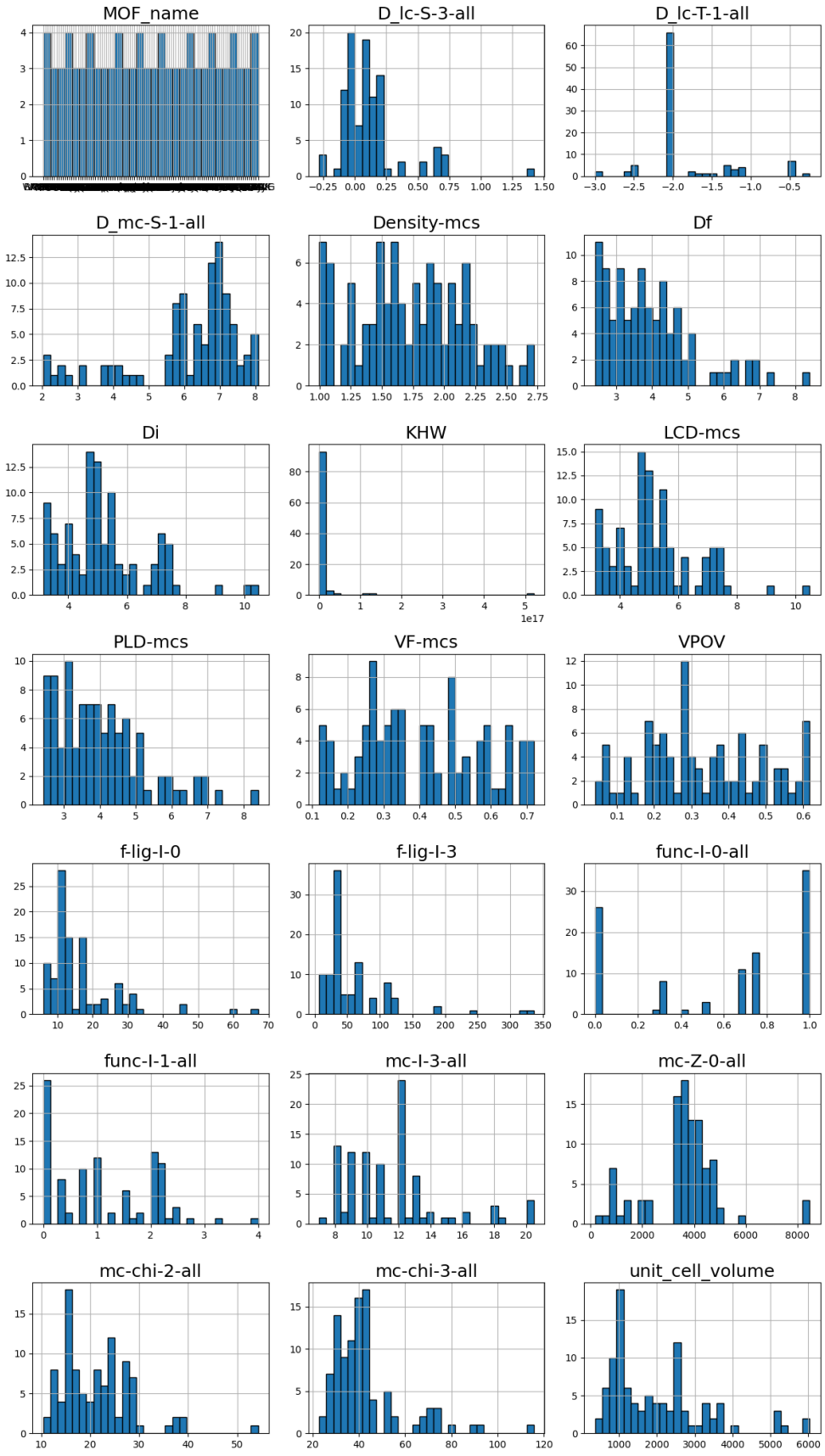

Fig. S5: The feature distribution of the top 100 MOF candidates for atmospheric water harvesting.

Table S13 The evaluated hyperparameter grids for the top four regression models.

Model	Hyperparameter grid
LGBMRegressor	n_estimators: [100, 200, 300, 400, 500] learning_rate: [0.01, 0.05, 0.1, 0.15, 0.2] num_leaves: [20, 31, 40, 60, 80, 100] max_depth: [3, 5, 7, 9, 11, -1]
CatBoostRegressor	iterations: [100, 200, 300, 400, 500, 1000] depth: [4, 6, 8, 10] learning_rate: [0.01, 0.03, 0.05, 0.1, 0.15, 0.2] l2_leaf_reg: [1, 3, 5, 7, 9]
HistGradientBoostingRegressor	max_iter: [100, 200, 300, 400, 500] learning_rate: [0.01, 0.05, 0.1, 0.15, 0.2] max_depth: [None, 3, 5, 7, 9] min_samples_leaf: [10, 20, 30, 50, 70, 90]
ExtraTreesRegressor	n_estimators: [100, 200, 300, 400, 500] max_features: ['auto', 'sqrt', 'log2'] max_depth: [None, 5, 10, 15, 20, 25] min_samples_split: [2, 4, 6, 8, 10]

References

1. Song, Z. et al. Accelerated screening of water-stable MOF structures using digital reticular chemistry method. *J. Mater. Chem. A* **13**, 17087–17101 (2025). doi:10.1039/D5TA01097G

2. Terrones, G. G. et al. Metal–organic framework stability in water and harsh environments from data-driven models trained on the diverse WS24 data set. *J. Am. Chem. Soc.* **146**, 20333–20348 (2024).

3. Bitencourt, G. F. et al. Bimetallic (Fe–Ga) metal–organic frameworks for tailoring peroxidase-like activity: an approach for methane partial oxidation. *ACS Mater. Au* (2025). doi:10.1021/acsmaterialsau.5c00045

4. Hastings, J. et al. High-temperature water adsorption isotherms and ambient temperature water diffusion rates on water harvesting metal–organic frameworks. *J. Phys. Chem. C* **128**, 11328–11339 (2024).

5. Wakekar, S. & Das, C. Deciphering the functions of metal–organic frameworks and their derived composites towards atmospheric water harvesting: a comprehensive review. *Sustain. Mater. Technol.* (2024). doi:10.1016/j.susmat.2024.e01120

6. Rosado, A. et al. Hydrophobic post-functionalization of a water instable bioMOF: effect on CO_2 and water adsorption. *Appl. Mater. Today* **42**, 102573 (2025).

7. Cao, H. et al. Two-periodic MoS2-type metal–organic frameworks with intrinsic intralayer porosity for high-capacity water sorption. *Adv. Mater.* **37**, 2414362 (2025).

8. Saha, R. et al. Simultaneous electron and proton conduction in a stable metal–organic material with highly selective electrocatalytic oxygen reduction reaction to water. *Chem. Sci.* **16**, 9501–9508 (2025).

9. Yang, H. et al. Ultrastable copper carboxylate metal–organic frameworks. *J. Am. Chem. Soc.* (2025).

10. Akmal, S. A., Khalid, M., Shahid, M. & Ahmad, M. Engineering an amine-functionalized Zn(II) interpenetrated metal–organic framework for comprehensive iodine uptake across diverse phases and selective azo dye adsorption. *Dalton Trans.* (2025).

11. Wenzel, S. *Metal–Organic Frameworks (MOFs) Based on Fluorinated Trimesate Linkers*. PhD thesis, Univ. zu Köln (2025).