Supplement of Technical note: Towards atmospheric compound identification in chemical ionization mass spectrometry with machine learning

Federica Bortolussi[1], Hilda Sandström[2], Fariba Partovi[3,4], Joona Mikkilä[4], Patrick Rinke[2,5], and Matti Rissanen[1,3]

[1]Department of Chemistry, University of Helsinki, 00560 Helsinki, Finland
[2]Department of Applied Physics, Aalto University, Espoo, Finland
[3]Aerosol Physics Laboratory, Physics Unit, Tampere University, 33720 Tampere, Finland
[4]Karsa Ltd., A. I. Virtasen aukio 1, 00560 Helsinki, Finland
[5]Physics Department, TUM School of Natural Sciences, Technical University of Munich, Garching, Germany

Correspondence: Federica Bortolussi (federica.bortolussi@helsinki.fi)

Contents

S1 Description of S.I.

Section: S2

The dataset (Partovi et al., 2024) contains the measurements of 716 pesticides at 5 different sample concentrations. Figure S1 presents the number of molecules detected for each sample concentration for the four ionization schemes. Since the dataset size, we decided to focus the analysis in the main text on the highest concentration only (2,5 ng μl^{-1}), to provide a relatively balanced dataset (in the case of the classification model) and the highest amount of instances (in case of the regression model). Figure S2 presents the distribution of logarithmic signal intensities for the five different concentrations for each ionization scheme. Figure S3 presents the molecular weight distribution against the signal intensity. The pesticides giving a signal outside the red box were not considered for the analysis in the main text. Figure S4 presents the molecular weight distribution for both detected and undetected pesticides. Figure S5 presents the normalized distribution of different functional groups (with highlighted the frequency of occurrence), and the overall distribution of the number of functional groups for the dataset. To calculate the functional groups we employed the Python library developed by Ruggeri and Takahama (2016).

Tables S1 and S2 present the list of pesticides excluded from the analysis in the main text. Table S3 shows the three molecules containing tin. One of the molecules was excluded from the analysis, having a molecular weight above 600 u.

Section: S3

Table S4 presents all the properties present in the RDkitPROP model, with a description. All the properties are calculated with the class rdkit.Chem.rdMolDescriptors.Properties() from the library *rdkit* (Landrum, 2006). Figure S6 shows the learning curve of KRR for the four different ionization methods with MBTR as the molecular descriptor. The figure presents MBTR with three terms on the left (k1 = the atomic numbers, k2 = the distance and k3 = the angle) and MBTR with two terms on the right (k2 = the distance and k3 = the angle). The results obtained are comparative, so only two terms were used for all the ML models to decrease the computational costs.

Section: S4

Table S5 presents the tuning range list of the hyperparameters for both the ML models and the molecular descriptors (with additional information on the hyperparameter). The tuned hyperparameters for each ML model based on each ionization method data can be found in the following tables for each molecular descriptor and each random re-shuffle of the data. Tables S6, S7, S8, S9, S10 present the hyperparameter tuning for RF classifier. Tables S11, S12, S13, S14, S15 present the hyperparameter tuning for KRR with Gaussian kernel.

Section: S5

In the main text, RF presents the accuracy of the prediction in a table form, here Figure S7 presents the learning curve of the accuracy over an increase of training size. The following features present similar performance metrics: recall (Fig. S8), precision (Fig. S9) and f1 score (Fig. S10). In the main text, the learning curve of KRR was based on the mean absolute error value (MAE), here we will show the learning curve of similar performance metrics: mean squared error (MSE, Fig. S11) and correlation coefficient (R^2, Fig. S12). Tables S16 and S17 present the complete RDkitPROP RF best estimator feature importances % for Br^- and O_2^-, and H_3O^+ and $AceH^+$. Tables S18 and S19 present the complete MACCS-based RF best estimator feature importances% for Br^- and O_2^-, and H_3O^+ and $AceH^+$.

Section: S6

In this work we trained in total 6 ML models: 3 classifiers (RF, naive Bayes (NB) and support vector classifier (SVC)), 3 regressors (KRR with Gaussian kernel, KRR with linear kernel and RF regressor). For each model, we will report the hyperparameters tuning and the learning curves with accuracy, recall, precision and f1 score (classifier), and MAE, MSE and R^2 (regressor).

Table S20 presents the tuning range list of the hyperparameters for the additional ML models (with additional information on the hyperparameter). The tuned hyperparameters for each ML model based on each ionization method data can be found in the following tables for each molecular descriptor and each random re-shuffle of the data.

We run a NB classifier to show the performance of a simple classification model in comparison to the one reported in the main text (Tables S21, S22 and Fig. S13, S14, S15, S16). The NB model presents a poor performance compared to the RF classifier.

We run SVC to show the performance of a similarly complicated classifier model as the one used in the main text (Tables S23,S24, S25, S26, S27 and Fig. S17, S18, S19, S20). The performance appears similar to the results found with KRR with the Gaussian kernel.

We run KRR with a linear kernel to show the performance of a simple regression model in comparison to the one reported in the main text (Tables S28 S29 and Fig. S21, S22, S23). In fact, the linear kernel performs poorly compared to the Gaussian kernel.

We run a RF regressor to show the performance of a similarly complicated regression model as the one used in the main text (Tables S30,S31, S32, S33, S34 and Fig. S24, S25, S26). The performance appears similar to the results found with KRR with the Gaussian kernel.

S2 Dataset: additional analysis, excluded molecules and tin molecules

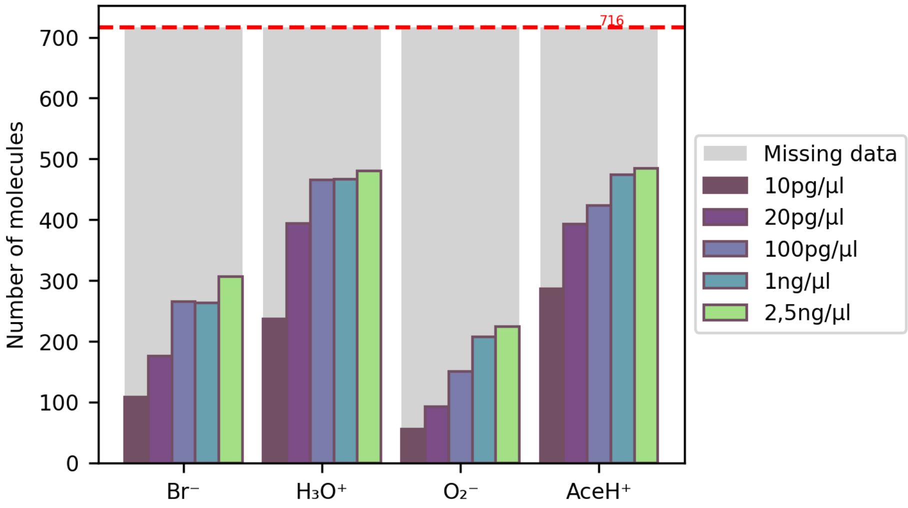


Figure S1. Number of detected pesticides for the experiments at five different concentrations and before filtering.

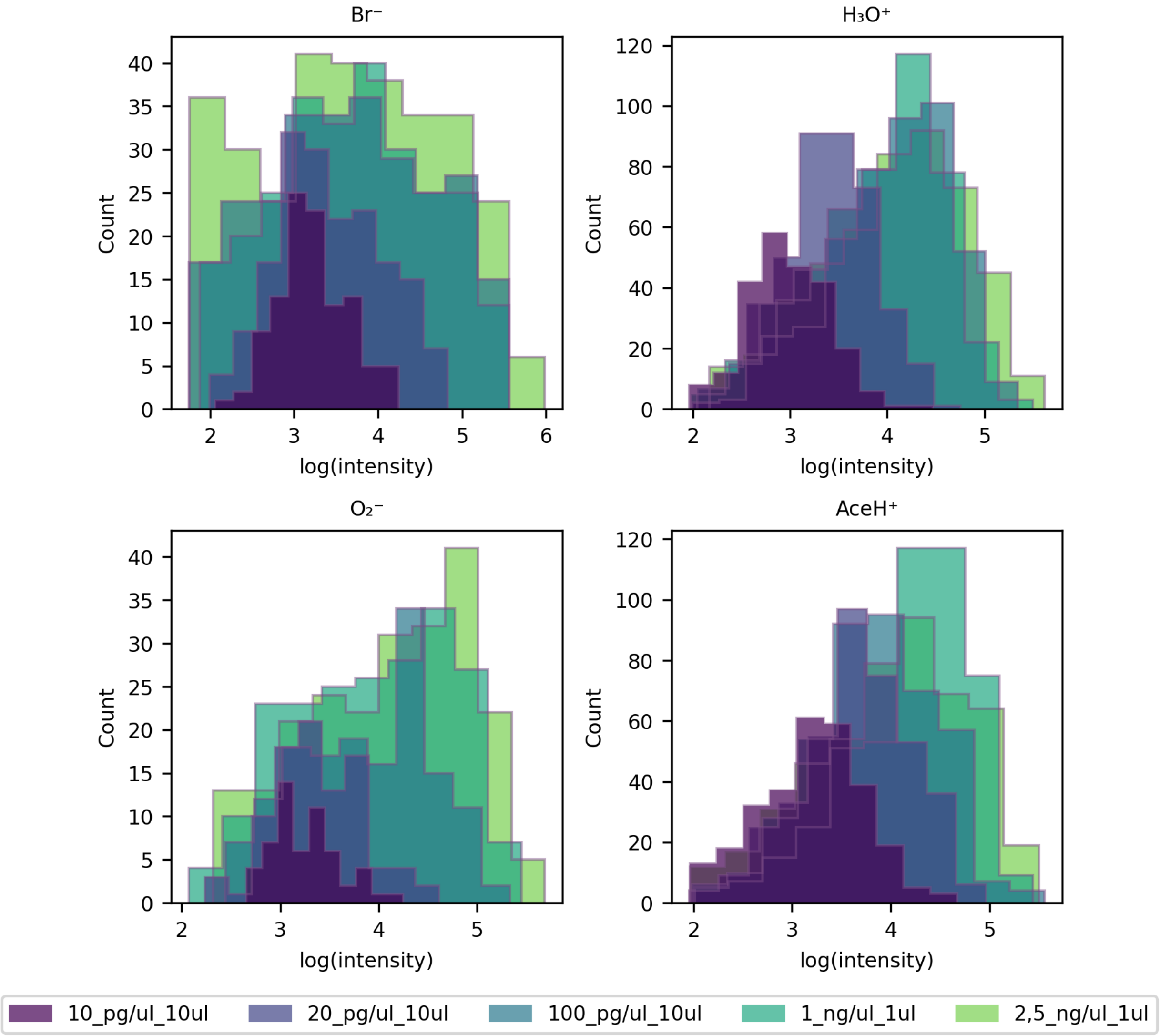


Figure S2. Detected molecules signal intensity distribution shown in a logarithmic scale for the four reagent ions studied for five different concentrations

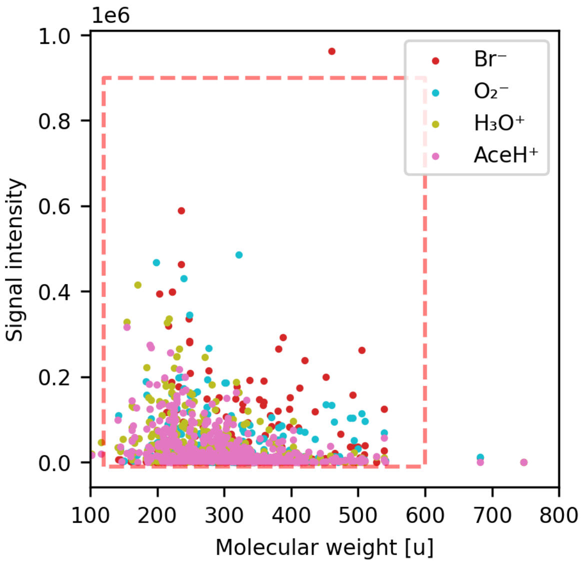


Figure S3. Scatter plot of the dataset prior to outlier filtering. The data points outside the red square are excluded from the analysis.

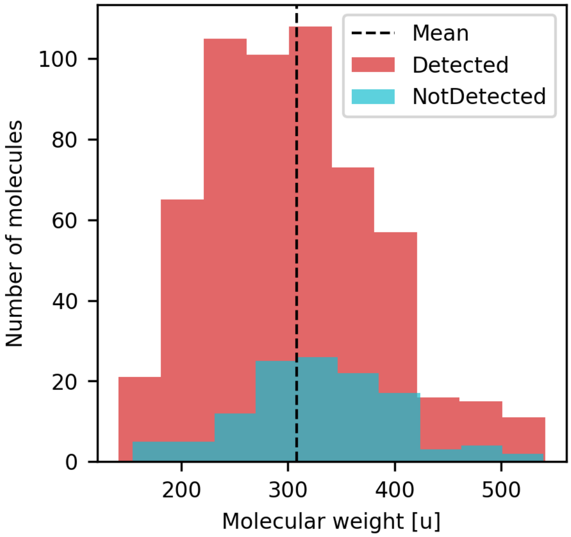


Figure S4. Distribution of molecular weight: the detected count is red, while the undetected count is light blue.

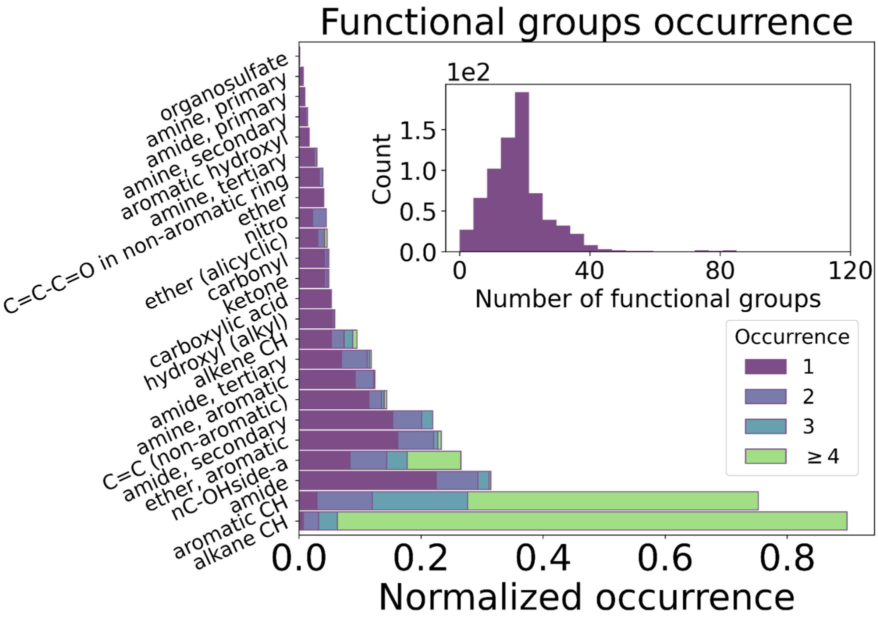


Figure S5. Normalized functional groups distribution of the dataset, with the occurrence highlighted and total number of functional groups distribution.

Table S1. List of pesticides excluded from the analysis: name, CAS, molecular weight (Mw), signal intensities for the four ionization methods and reason of exclusion. Measurements at a sample concentration of 2,5 ng μl^{-1}.

Name	CAS	Mw [u]	Br^- [a.u.]	$O_2{}^-$ [a.u.]	H_3O^- [a.u.]	$AceH^+$ [a.u.]	Reason
1-Naphthylacetamide (1)	86-86-2	185.08	0	0	17700	6460.0	Appears twice
1-Naphthylacetamide (2)	86-86-2	185.08	0	0	14300	18500.0	Appears twice
Abamectin B1a	65195-55-3	872.49	0	0	0	0	Mw > 600 u
Acibenzolar acid (1)	35272-27-6	180.00	0	0	0	0	Appears twice
Acibenzolar acid (2)	35272-27-6	180.00	0	0	0	0	Appears twice
Azadirachtin	11141-17-6	720.26	0	0	0	0	Mw > 600 u
Emamectin B1a	155569-91-8	885.52	0	0	0	0	Mw > 600 u
Ethylenethiourea	96-45-7	102.03	0	0	19700	16300.0	Mw < 120 u
Fenbutatin-oxide	13356-08-6	1054.41	0	0	0	0	Mw > 600 u
Fenpicoxamid	517875-34-2	614.25	0	0	0	0	Mw > 600 u
Fenpyrazamine (1)	473798-59-3	331.14	0	0	0	0	Appears twice
Fenpyrazamine (2)	473798-59-3	331.14	1070.0	2360	45100	21800.0	Appears twice
Flubendiamide	272451-65-7	682.02	10600.0	12500	0	189.0	Mw > 600 u
Hexadecyltrimethylammonium chlorid (1)	112-02-7	319.30	0	0	0	0	Appears twice
Hexadecyltrimethylammonium chlorid (2)	112-02-7	319.30	0	0	0	0	Appears twice
Hexaflumuron	86479-06-3	459.98	962000.0	134000	4600	6440.0	Br^- signal > 900000 [a.u.]
Imazapyr (1)	81334-34-1	261.11	1780.0	13800	16600	17600.0	Appears twice
Imazapyr (2)	81334-34-1	261.11	4310.0	39900	93000	8500.0	Appears twice
Propylen Thiourea	2122-19-2	116.04	0	0	46000	19200.0	Mw < 120 u
Pyridafol (1)	40020-01-7	206.02	0	0	0	0	Appears twice
Pyridafol (2)	40020-01-7	206.02	34900.0	60600	15600	12600.0	Appears twice
Spinetoram J	187166-40-1	747.49	289.0	0	301	266.0	Mw > 600 u
Spinosad A	131929-60-7	731.46	0	0	0	0	Mw > 600 u

Table S2. List of pesticides excluded from the analysis: name and 2D molecular structure

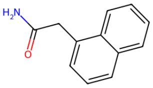
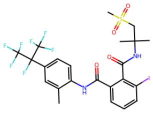
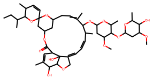
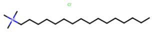
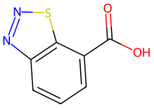
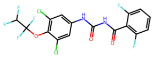
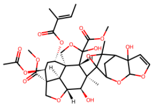
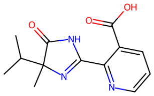
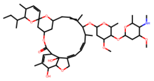
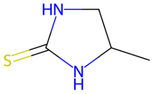
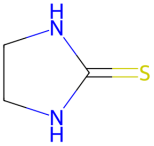
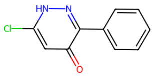
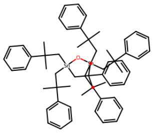
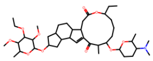
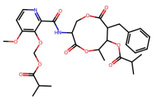
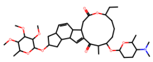
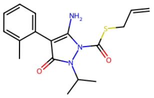
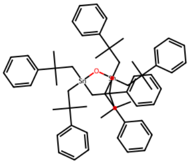
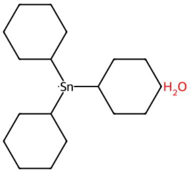
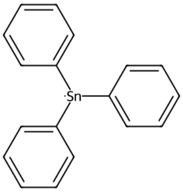
Name	Structure	Name	Structure
1-Naphthylacetamide		Flubendiamide	
Abamectin B1a		Hexadecyltrimethylammonium chlorid	
Acibenzolar acid		Hexaflumuron	
Azadirachtin		Imazapyr	
Emamectin B1a		Propylen Thiourea	
Ethylenethiourea		Pyridafol	
Fenbutatin-oxide		Spinetoram J	
Fenpicoxamid		Spinosyn A	
Fenpyrazamine			

Table S3. Molecules containing tin (Sn).

Target	CAS	Molecular weight [u]	Structure	Additional information
Fenbutatin-oxide	13356-08-6	1054.41		Excluded from the analysis
Cyhexatin	13121-70-5	387.17		Included in the analysis
Triphenyltin hydride	892-20-6	351.02		Included in the analysis

S3 Molecular descriptors: properties info, MBTR comparison

Table S4. List of properties calculated with *rdkit* for the RDKitPROP descriptor.

Property	Info
Exactmw	Returns the exact molecular weight for a molecule ("Mw" in the main text)
Amw	Returns the average molecular weight for a molecule
LipinskiHBA	Calculates the standard Lipinski HBA definition (number of Ns and Os)
LipinskiHBD	Calculates the standard Lipinski HBD definition (number of N-H and O-H bonds)
NumRotatableBonds	Number of rotatable Bonds.
NumHBD	Calculates the number of H-bond donors ("n HBD" in the main text).
NumHBA	Calculates the number of H-bond acceptors ("n HBD" in the main text).
NumHeavyAtoms	Number of heavy atoms a molecule.
NumAtoms	Returns the number of atoms.
NumHeteroatoms	Returns the number of heteroatoms.
NumAmideBonds	Returns the number of amide bonds.
FractionCSP3	Returns the fraction of C atoms that are sp^3 hybridized.
NumRings	Returns the number of rings.
NumAromaticRings	Returns the number of aromatic rings for a molecule.
NumAliphaticRings	Returns the number of aliphatic (containing at least one non-aromatic bond) rings for a molecule.
NumSaturatedRings	Returns the number of saturated rings for a molecule.
NumHeterocycles	Returns the number of heterocycles for a molecule.
NumAromaticHeterocycles	Returns the number of aromatic heterocycles for a molecule.
NumSaturatedHeterocycles	Returns the number of saturated heterocycles for a molecule.
NumAliphaticHeterocycles	Returns the number of aliphatic (containing at least one non-aromatic bond) heterocycles for a molecule.
NumSpiroAtoms	Number of spiro atoms (atoms shared between rings that share exactly one atom).
NumBridgeheadAtoms	Number of bridgehead atoms (atoms shared between rings that share at least two bonds).
NumAtomStereoCenters	Calculates the total number of atom stereo centers
NumUnspecifiedAtomStereoCenters	Returns the number of atom stereo centers when the molecule has unspecified stereochemistry.
LabuteASA	Calculates Labute's Approximate Surface Area (Labute, 2000).
TPSA	Returns the topological Polar Surface Area (Ertl et al., 2000).
CrippenClogP	Uses an atom-based scheme based on the values in (Wildman and Crippen, 1999) to calculate the default Wildman-Crippen partition coefficient (LogP) for a molecule (Wildman and Crippen, 1999).
CrippenMR	Uses an atom-based scheme based on the values in (Wildman and Crippen, 1999) to calculate the default Wildman-Crippen molar refractivity (MR) estimate for a molecule.
chi0v, chi1v, chi2v, chi3v, chi4v	From equations (5),(9) and (10) of Hall and Kier (1991).
chi0n, chi1n, chi2n, chi3n, chi4n	From equations (5),(9) and (10) of Hall and Kier (1991) with nVal instead of valence.
HallKierAlpha	Calculate the Hall-Kier alpha value for a molecule from equation (58) of Hall and Kier (1991)
kappa1	Calculate the Hall-Kier kappa1 value for a molecule from equations (58) and (59) of Hall and Kier (1991)
kappa2	Calculate the Hall-Kier kappa2 value for a molecule from equations (58) and (60) of Hall and Kier (1991)
kappa3	Calculate the Hall-Kier kappa3 value for a molecule from equations (58), (61) and (62) of Hall and Kier (1991)
phi	Calculate the Kier Phi value for a molecule from Kier (1989)

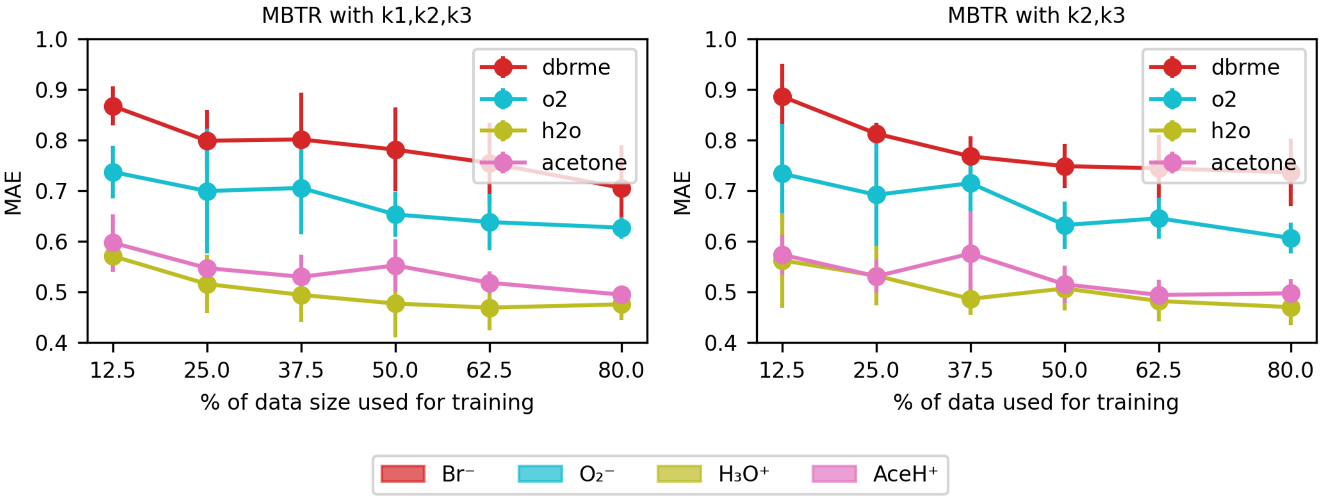


Figure S6. Comparison between MBTR results with terms k1,k2,k3 and k2,k3

S4 Tuned hyperparameters

Table S5. Hyperparameters tuning range list for each ML model and each molecular descriptor.

Model	Hyperparameters	Tuning range	Info
RDKitPROP	None		
TopFP	Fp size	[716,2048,4096,8192]	Final bit length of the fingerprint
	Max path	[7,8,9,10]	Maximum path considered for each bit
	N bits per hash	[2, 4, 8, 16]	Number of bit which are hashed together
MACCS	None		
CM	None		
MBTR	σ_2	[0.3, 0.1, 0.01, 0.001, 0.0001]	Broadening parameter for k=2
	w_2	[0.2, 0.4, 0.8, 1.2, 1.4]	Weighting parameter for k=2
	σ_3	[0.3, 0.1, 0.01, 0.001, 0.0001]	Broadening parameter for k=3
	w_3	[0.2, 0.4, 0.8, 1.2, 1.4]	Weighting parameter for k=3
KRR with Gaussian kernel	λ	np.logspace(-10, -1, 10)	Regularization strength
	σ	np.logspace(-10, 0, 10)	Length scale
RF classifier	Max depth	[20, 40, 60, 80, 100, None]	The length of each tree, from the root to the leaves
	Min samples leaf	[1, 2, 4]	Minimum number of samples per leaf
	Min samples split	[2, 5, 10]	Minimum number of samples per split
	N estimators	[100, 500, 1000, 1500, 2000]	Maximum number of estimators

Table S6. Hyperparameters tuned for RF classifier model with PROP as the molecular descriptor.

			Hyperparameters			
			RF classifier			
Reagent ion	Training size	Random seed	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	554	555	500	None	4	5
		8	1500	40	2	10
		52	2000	20	1	5
		1066	2000	20	2	2
		324	500	80	2	5
$O_2{}^-$	554	555	1000	None	1	5
		8	2000	None	1	2
		52	1000	80	1	2
		1066	2000	20	1	5
		324	100	60	1	10
H_3O^+	554	555	1000	100	2	5
		8	2000	60	1	5
		52	500	60	2	5
		1066	1500	40	2	2
		324	2000	40	1	5
$AceH^+$	554	555	1000	None	1	5
		8	100	40	4	5
		52	1500	100	2	5
		1066	2000	20	1	5
		324	500	40	2	10

Table S7. Hyperparameters tuned for RF classifier model with TopFP as the molecular descriptor.

			Hyperparameters						
Reagent ion	Training size	Random seed	TopFP			RF classifier			
			Fp size	Max path	N bits per hash	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	554	555	2048	7	4	1500	40	1	10
		8	8192	7	8	2000	None	2	2
		52	4096	7	4	1000	20	1	5
		1066	8192	8	8	2000	100	2	2
		324	8192	9	8	1000	20	1	5
$O_2{}^-$	554	555	8192	8	8	500	20	2	2
		8	4096	9	2	500	None	4	10
		52	2048	7	2	1000	40	2	5
		1066	8192	9	2	100	None	4	5
		324	4096	7	8	1500	None	4	10
H_3O^+	554	555	4096	7	8	100	60	2	2
		8	8192	7	4	1000	None	4	5
		52	4096	7	4	1500	100	2	2
		1066	8192	8	8	100	80	2	5
		324	8192	7	4	1000	100	1	5
$AceH^+$	554	555	8192	7	16	1500	20	4	5
		8	2048	8	4	1500	60	1	2
		52	8192	9	4	500	80	1	2
		1066	8192	7	8	100	40	4	10
		324	8192	8	4	100	80	4	10

Table S8. Hyperparameters tuned for RF classifier model with MACCS as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			RF classifier			
			N estimators	Max depth	Min samples leaf	Min samples split
Br^-	554	555	1000	100	2	2
		8	100	80	4	10
		52	100	20	2	5
		1066	100	40	1	5
		324	100	None	2	5
$O_2{}^-$	554	555	500	100	1	2
		8	1000	40	2	2
		52	100	80	1	2
		1066	500	60	1	2
		324	500	40	1	2
H_3O^+	554	555	100	20	1	2
		8	1500	80	1	2
		52	100	100	1	10
		1066	500	100	1	10
		324	500	20	1	5
$AceH^+$	554	555	500	40	1	2
		8	100	100	1	2
		52	100	40	1	2
		1066	100	60	2	2
		324	100	20	2	10

Table S9. Hyperparameters tuned for RF classifier model with CM as the molecular descriptor.

			Hyperparameters			
			RF classifier			
Reagent ion	Training size	Random seed	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	554	555	100	100	1	2
		8	100	80	2	2
		52	100	80	2	5
		1066	500	100	1	2
		324	100	100	1	2
$O_2{}^-$	554	555	500	60	1	2
		8	500	40	2	5
		52	100	60	2	2
		1066	100	40	2	10
		324	500	20	1	2
H_3O^+	554	555	500	40	1	2
		8	100	60	2	5
		52	100	100	2	5
		1066	1000	60	1	10
		324	100	40	1	2
$AceH^+$	554	555	500	60	1	5
		8	100	100	1	5
		52	500	80	1	2
		1066	100	20	2	2
		324	100	60	1	2

Table S10. Hyperparameters tuned for RF classifier model with MBTR as the molecular descriptor.

			Hyperparameters							
Reagent ion	Training size	Random seed	MBTR				RF classifier			
			σ_2	w_2	σ_3	w_3	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	554	555	0.0001	0.6	0.005	1.2	1000	100	1	2
		8	0.0005	0.4	0.1	1	500	40	1	10
		52	0.01	0.4	0.005	0.6	1500	60	2	10
		1066	0.005	1	0.1	1	2000	40	1	10
		324	0.005	0.6	0.1	0.8	2000	80	1	10
$O_2{}^-$	554	555	0.0001	1.2	0.1	0.8	1500	20	2	2
		8	0.001	0.8	0.1	1.2	100	None	2	2
		52	0.1	0.4	0.1	1	1500	60	2	10
		1066	0.01	0.6	0.1	1.2	2000	60	1	2
		324	0.01	0.6	0.0005	0.6	1000	40	1	5
H_3O^+	554	555	0.1	1.2	0.005	0.6	1500	40	4	10
		8	0.0001	0.2	0.005	0.8	500	100	1	5
		52	0.001	0.6	0.1	1	2000	None	1	2
		1066	0.001	1	0.1	1	1500	40	2	5
		324	0.1	0.8	0.0001	0.2	500	80	2	2
$AceH^+$	554	555	0.01	0.4	0.001	0.2	500	None	1	2
		8	0.001	1.2	0.01	0.6	1500	100	1	10
		52	0.1	1.2	0.1	1.2	2000	20	1	10
		1066	0.1	1	0.0005	1	1000	20	1	2
		324	0.01	0.8	0.0005	1	1500	20	1	10

Table S11. Hyperparameters tuned for KRR with Gaussian kernel model with PROP as the molecular descriptor.

			Hyperparameters	
Reagent ion	Training size	Random seed	KRR with Gaussian kernel	
			λ	σ
Br^-	240	555	1e-10	1.29e-09
		8	1e-08	1e-10
		52	1e-07	1.67e-08
		1066	0.001	2.78e-06
		324	1e-08	1.67e-08
$O_2{}^-$	174	555	1e-07	1.29e-09
		8	1e-08	1.67e-08
		52	0.001	2.15e-07
		1066	0.001	2.78e-06
		324	0.001	1.29e-09
H_3O^+	376	555	0.0001	2.15e-07
		8	1e-05	1.67e-08
		52	1e-05	2.78e-06
		1066	1e-08	1.29e-09
		324	1e-06	2.15e-07
$AceH^+$	379	555	1e-08	1e-10
		8	0.0001	2.15e-07
		52	0.0001	2.15e-07
		1066	1e-07	1.29e-09
		324	1e-07	1.67e-08

Table S12. Hyperparameters tuned for KRR with Gaussian kernel model with TopFP as the molecular descriptor.

			Hyperparameters				
Reagent ion	Training size	Random seed	TopFP			KRR with Gaussian kernel	
			Fp size	Max path	N bits per hash	λ	σ
Br^-	240	555	4096	7	4	1e-08	1.29e-09
		8	2048	7	16	0.0001	1.67e-08
		52	8192	8	4	0.0001	2.78e-06
		1066	8192	8	4	1e-05	2.78e-06
		324	8192	9	4	1e-06	1.67e-08
$O_2{}^-$	174	555	4096	7	8	1e-05	2.78e-06
		8	8192	8	2	0.01	3.59e-05
		52	8192	7	8	1e-05	0.000464
		1066	8192	7	16	1e-07	1e-10
		324	8192	7	2	1e-06	1.29e-09
H_3O^+	376	555	8192	7	16	1e-07	2.78e-06
		8	4096	7	2	1e-06	3.59e-05
		52	8192	9	4	1e-05	3.59e-05
		1066	8192	8	16	1e-09	1.29e-09
		324	8192	7	8	1e-06	0.000464
$AceH^+$	379	555	4096	7	8	1e-05	3.59e-05
		8	4096	8	8	1e-08	3.59e-05
		52	716	9	8	0.01	3.59e-05
		1066	8192	10	4	0.001	0.000464
		324	8192	7	4	0.001	2.15e-07

Table S13. Hyperparameters tuned for KRR with Gaussian kernel model with MACCS as the molecular descriptor.

			Hyperparameters	
Reagent ion	Training size	Random seed	KRR with Gaussian kernel	
			λ	σ
Br^-	240	555	0.01	0.00599
		8	0.1	0.00599
		52	0.01	0.00599
		1066	0.1	0.00599
		324	0.01	0.00599
$O_2{}^-$	174	555	0.1	0.00599
		8	0.1	0.00599
		52	0.1	0.00599
		1066	0.1	0.00599
		324	0.1	0.00599
H_3O^+	376	555	0.1	0.00599
		8	0.1	0.00599
		52	1e-07	1.29e-09
		1066	0.1	0.00599
		324	0.1	0.00599
$AceH^+$	379	555	1e-07	1.29e-09
		8	0.1	0.00599
		52	0.1	0.00599
		1066	0.1	0.00599
		324	0.1	0.00599

Table S14. Hyperparameters tuned for KRR with Gaussian kernel model with CM as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters	
			KRR with Gaussian kernel	
			λ	σ
Br^-	240	555	0.01	2.15e-07
		8	0.001	1.67e-08
		52	1e-05	1e-10
		1066	0.01	2.15e-07
		324	0.01	2.15e-07
${O_2}^-$	174	555	0.0001	1e-10
		8	0.01	1.67e-08
		52	0.01	2.15e-07
		1066	0.0001	1.29e-09
		324	0.01	1.67e-08
H_3O^+	376	555	1e-05	1e-10
		8	0.0001	1.29e-09
		52	1e-05	1e-10
		1066	1e-05	1e-10
		324	0.1	2.15e-07
$AceH^+$	379	555	0.01	1.67e-08
		8	0.001	1.29e-09
		52	1e-05	1e-10
		1066	0.0001	1e-10
		324	1e-05	1e-10

Table S15. Hyperparameters tuned for KRR with Gaussian kernel model with MBTR as the molecular descriptor.

			Hyperparameters					
Reagent ion	Training size	Random seed	MBTR				KRR with Gaussian kernel	
			σ_2	w_2	σ_3	w_3	λ	σ
Br^-	240	555	0.0001	1	0.01	0.2	1e-07	2.15e-07
		8	0.005	0.2	0.1	0.4	1e-09	1
		52	0.01	0.2	0.005	0.8	0.1	1
		1066	0.1	1	0.01	0.2	1e-06	2.78e-06
		324	0.01	0.4	0.0005	1.2	1e-08	1.29e-09
${O_2}^-$	174	555	0.1	0.8	0.01	0.2	1e-09	1.67e-08
		8	0.01	1	0.0005	0.6	0.01	0.0774
		52	0.01	0.4	0.001	1	0.1	0.0774
		1066	0.001	0.8	0.1	1.2	0.0001	0.000464
		324	0.01	0.2	0.1	1	1e-07	2.15e-07
H_3O^+	376	555	0.1	0.4	0.1	1.2	1e-07	2.78e-06
		8	0.0001	1	0.0005	0.8	1e-07	2.15e-07
		52	0.005	0.4	0.0005	0.2	1e-05	2.78e-06
		1066	0.1	1	0.1	0.8	1e-10	1.29e-09
		324	0.01	0.2	0.005	0.4	0.0001	3.59e-05
$AceH^+$	379	555	0.0005	0.8	0.005	0.8	0.001	0.000464
		8	0.01	0.8	0.01	0.6	1e-07	2.15e-07
		52	0.005	0.6	0.005	0.2	1e-08	2.15e-07
		1066	0.0001	0.6	0.001	0.2	0.001	0.000464
		324	0.1	0.2	0.1	0.4	0.001	0.00599

S5 Additional results: RF accuracy, recall, precision, f1 score; KRR MSE and R^2; chemical insight with RDkitPROP and MACCS descriptor

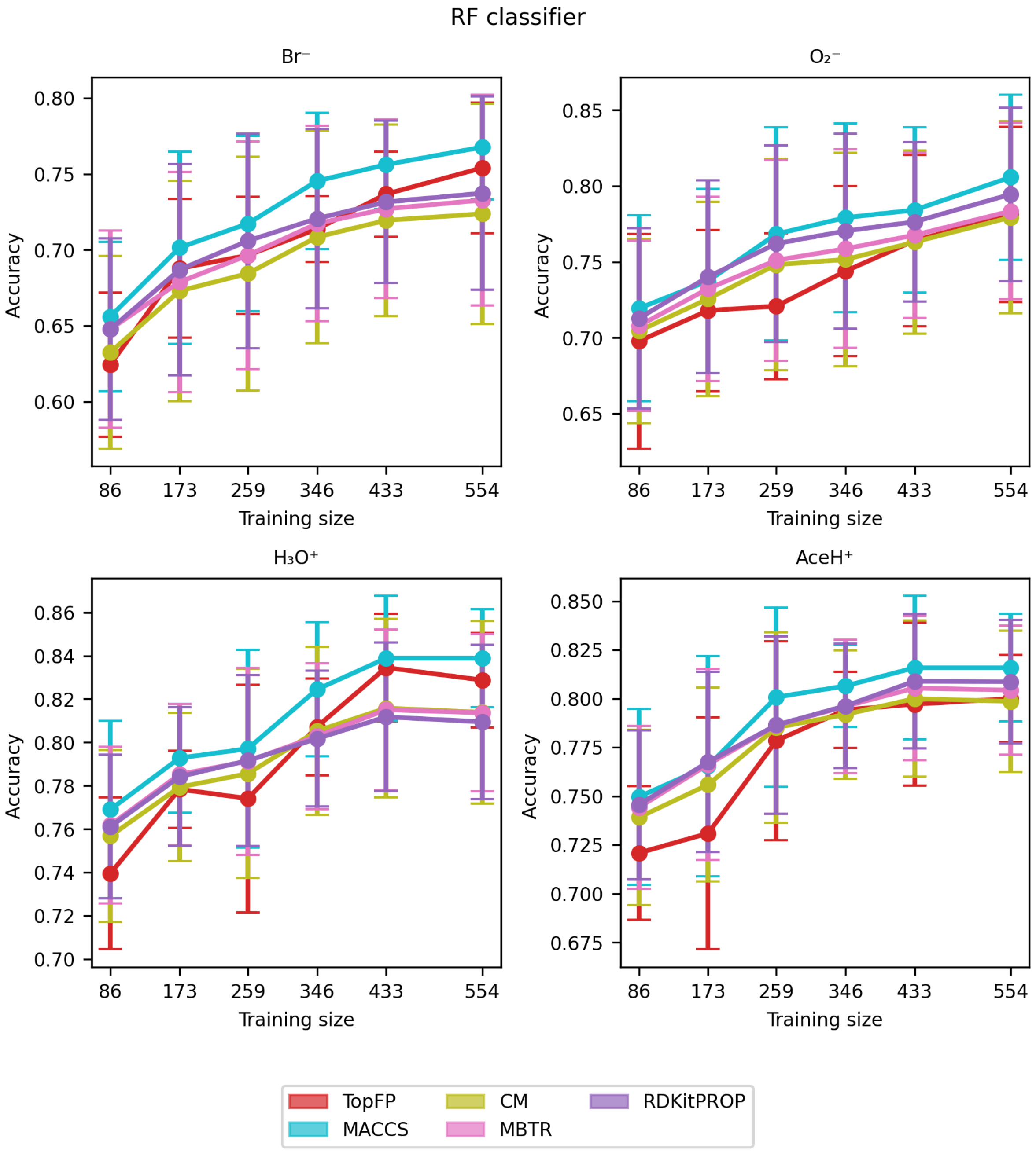


Figure S7. Learning curve of the RF with the accuracy of the classification of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification accuracy. We selected the largest training size (80% of the dataset) and obtained the mean value and standard deviation by repeating the training with five different random re-shuffling of the dataset.

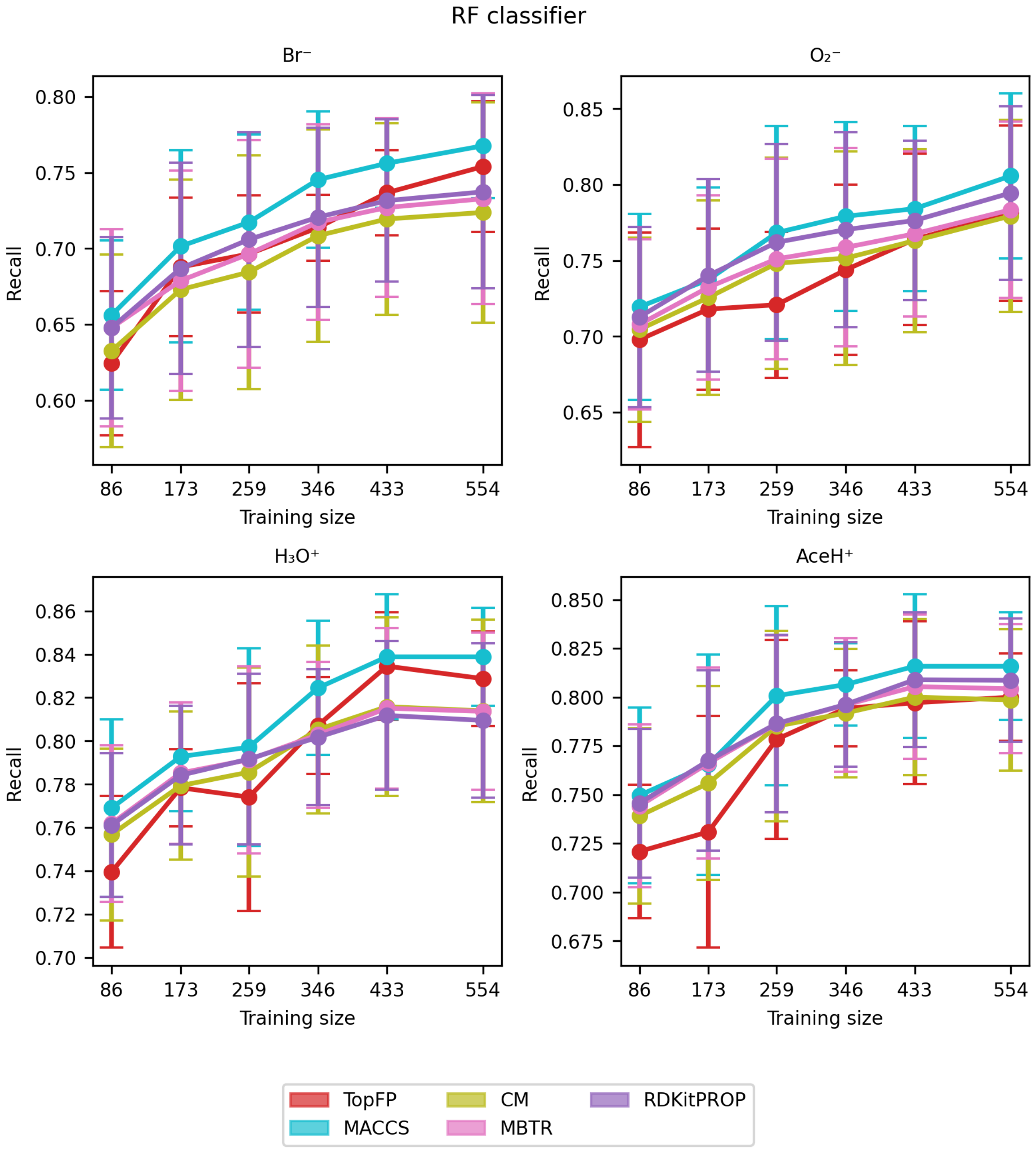


Figure S8. Learning curve of the ML with the recall of the classification of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification recall. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

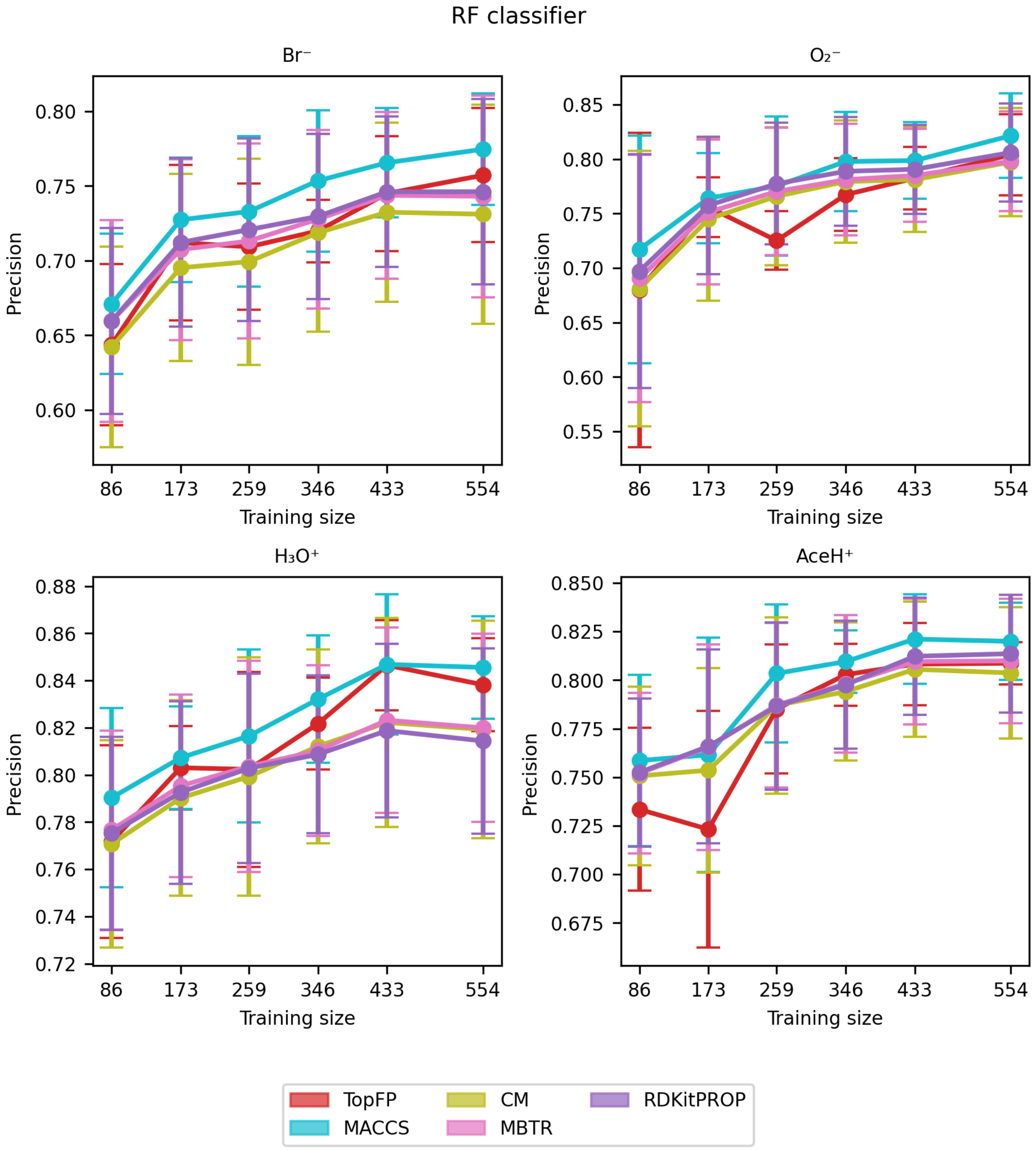


Figure S9. Learning curve of the ML with the precision of the classification of Br$^-$, O$_2{}^-$, H$_3$O$^+$ and AceH$^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification precision. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

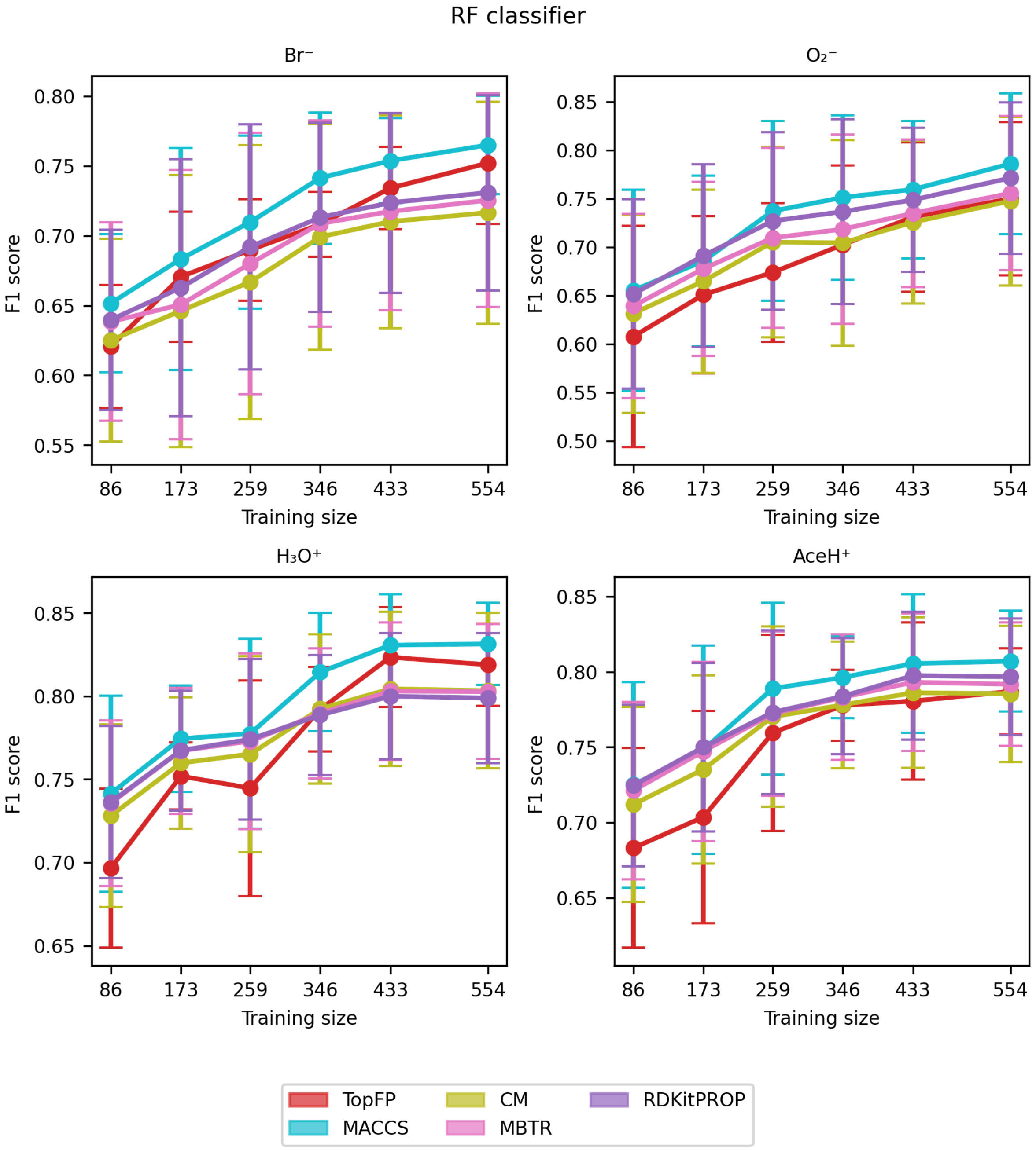


Figure S10. Learning curve of the ML with the F1 score of the classification of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification F1 score. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

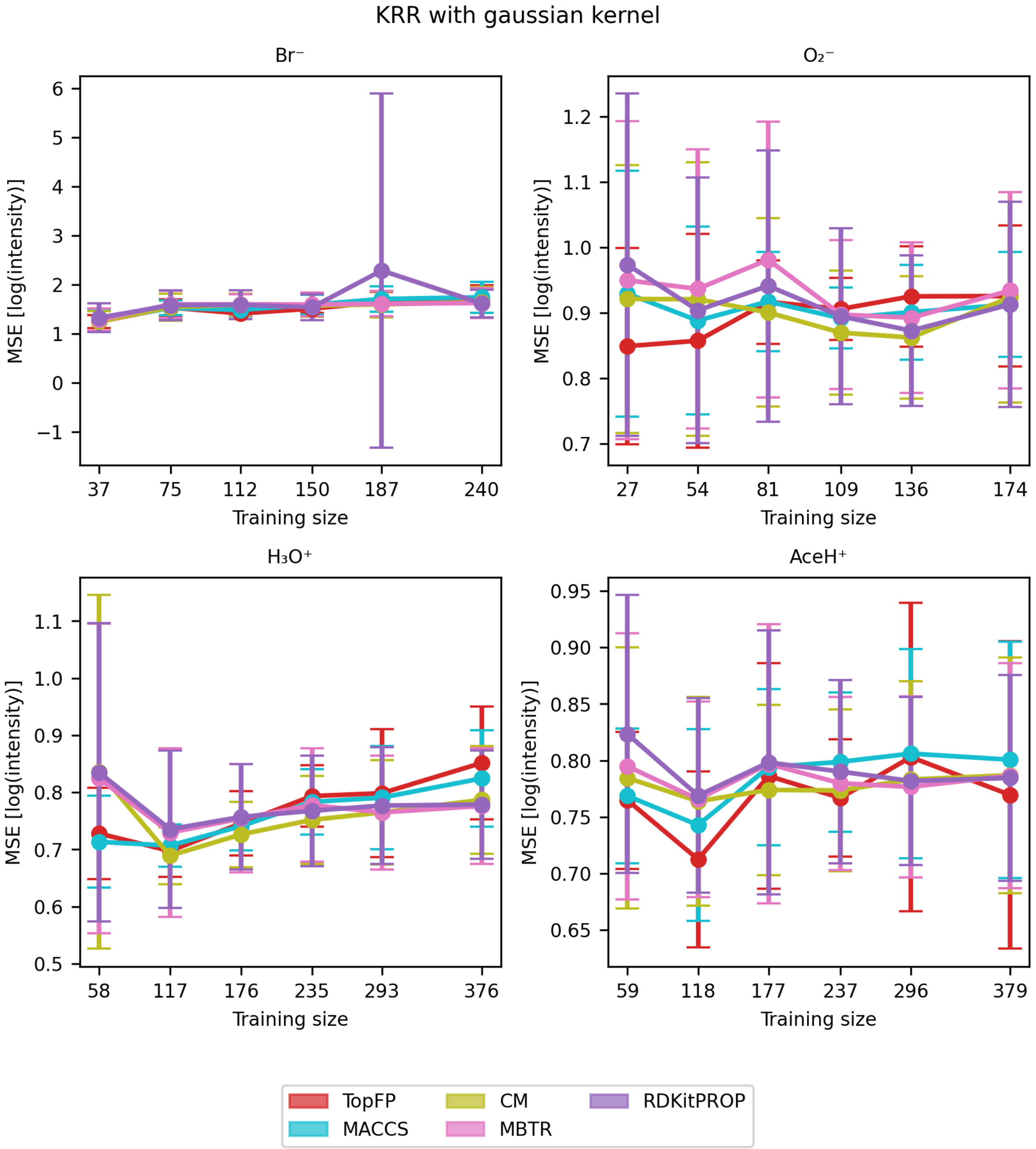


Figure S11. Learning curve with mean squared error (MSE) of the signal intensity values in logarithmic scale of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the MSE of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

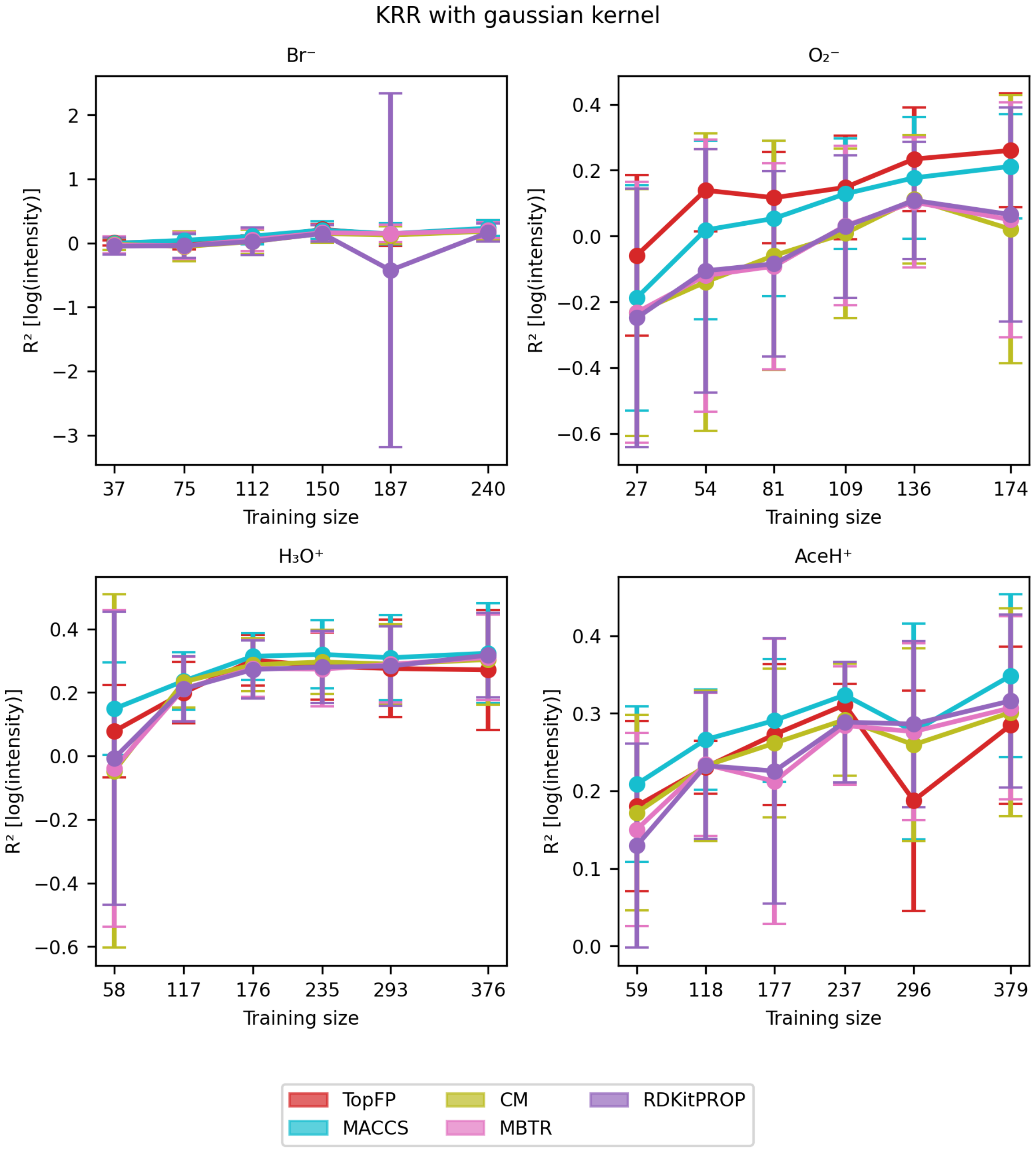


Figure S12. Learning curve with correlation coefficient (R^2) of the signal intensity values in logarithmic scale of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the R^2 of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

Table S16. RDKitPROP RF best estimator feature importances % for Br^- and $O_2{}^-$. For each property, the importance value (IMP %) and the average value (avg) for detected molecules (D) and undetected molecules (ND) are reported.

Property	Reagent ions					
	Br^-			$O_2{}^-$		
	IMP (%)	D_{avg}	ND_{avg}	IMP (%)	D_{avg}	ND_{avg}
TPSA	3.44	69.62	63.68	3.08	69.11	64.93
NumHBD (n HBD)	6.55	0.86	0.44	10.0	1.06	0.42
NumHBA (n HBA)	1.58	3.94	3.78	1.68	3.68	3.92
CrippenClogP	3.12	3.24	3.39	3.65	3.08	3.44
FractionCSP3	3.06	0.37	0.43	3.73	0.33	0.44
Exact Mw (Mw)	3.66	319.99	299.6	3.04	307.2	308.99
amw	3.53	320.72	300.33	3.16	307.87	309.75
lipinskiHBA	2.15	4.58	3.96	1.45	4.57	4.07
lipinskiHBD	6.41	0.91	0.49	9.25	1.12	0.46
NumRotatableBonds	1.85	4.04	4.61	1.16	3.78	4.63
NumHeavyAtoms	2.0	20.57	18.86	1.78	19.94	19.44
NumAtoms	2.44	35.2	34.83	2.39	33.74	35.56
NumHeteroatoms	3.32	7.05	6.01	2.17	6.92	6.25
NumAmideBonds	1.02	0.67	0.4	1.32	0.75	0.41
NumRings	0.77	1.82	1.61	0.47	1.72	1.69
NumAromaticRings	0.9	1.46	1.19	0.52	1.45	1.24
NumAliphaticRings	0.32	0.36	0.42	0.45	0.27	0.45
NumSaturatedRings	0.28	0.23	0.28	0.23	0.17	0.3
NumHeterocycles	0.79	0.63	0.49	0.43	0.59	0.53
NumAromaticHeterocycles	1.01	0.48	0.3	0.51	0.49	0.32
NumSaturatedHeterocycles	0.1	0.08	0.11	0.09	0.05	0.12
NumAliphaticHeterocycles	0.23	0.15	0.19	0.28	0.1	0.21
NumSpiroAtoms	0.06	0.02	0.01	0.05	0.01	0.01
NumBridgeheadAtoms	0.01	0.05	0.09	0.02	0.02	0.1
NumAtomStereoCenters	0.53	0.61	0.56	0.53	0.45	0.64
NumUnspecifiedAtomStereoCenters	0.71	0.52	0.44	0.48	0.41	0.5
labuteASA	3.08	126.05	118.74	2.65	120.96	122.34
CrippenMR	2.46	78.11	75.71	2.91	74.75	77.67
chi0v	2.61	12.67	12.36	2.83	12.05	12.7
chi1v	2.83	7.36	7.36	2.99	6.92	7.56
chi2v	3.43	4.13	4.32	3.28	3.75	4.47
chi3v	3.29	4.13	4.32	3.3	3.75	4.47
chi4v	2.92	2.8	3.01	3.12	2.45	3.14
chi0n	3.29	11.36	10.88	2.78	10.94	11.15
chi1n	2.98	6.18	5.93	2.71	5.93	6.09
chi2n	2.48	2.91	2.75	2.24	2.77	2.85
chi3n	2.49	2.91	2.75	2.32	2.77	2.85
chi4n	3.1	1.88	1.82	2.52	1.76	1.88
hallKierAlpha	4.11	-1.58	-1.11	3.94	-1.69	-1.14
kappa1	3.14	15.75	14.96	2.67	15.17	15.36
kappa2	2.62	6.48	6.54	2.57	6.17	6.67
kappa3	2.66	3.94	4.13	2.55	3.75	4.19
Phi	2.68	5.06	5.36	2.69	4.78	5.43

Table S17. RDKitPROP RF best estimator feature importances % for H_3O^+ and $AceH^+$. For each property, the importance value (IMP %) and the average value (avg) for detected molecules (D) and undetected molecules (ND) are reported.

Property	Reagent ions					
	H_3O^+			$AceH^+$		
	IMP (%)	D_{avg}	ND_{avg}	IMP (%)	D_{avg}	ND_{avg}
TPSA	7.29	70.5	57.29	8.51	69.66	58.87
n HBD	1.01	0.68	0.49	1.16	0.68	0.5
n HBA	3.8	4.2	3.11	4.57	4.15	3.18
CrippenClogP	4.1	3.15	3.69	4.05	3.22	3.54
FractionCSP3	3.91	0.43	0.36	2.4	0.41	0.4
Mw	3.31	302.17	321.61	3.22	304.33	317.3
amw	3.44	302.69	322.8	3.08	304.86	318.45
lipinskiHBA	7.03	4.59	3.48	7.05	4.55	3.54
lipinskiHBD	0.92	0.73	0.55	1.08	0.72	0.55
NumRotatableBonds	2.72	4.53	4.0	1.72	4.47	4.13
NumHeavyAtoms	1.28	19.87	19.02	1.35	20.03	18.67
NumAtoms	4.81	36.29	32.24	4.35	36.26	32.24
NumHeteroatoms	1.76	6.47	6.45	1.81	6.44	6.5
NumAmideBonds	1.97	0.61	0.32	1.82	0.61	0.32
NumRings	0.59	1.67	1.75	0.51	1.73	1.64
NumAromaticRings	0.59	1.35	1.2	0.54	1.4	1.09
NumAliphaticRings	0.31	0.32	0.55	0.34	0.32	0.55
NumSaturatedRings	0.17	0.2	0.37	0.3	0.19	0.4
NumHeterocycles	1.87	0.66	0.32	1.7	0.66	0.32
NumAromaticHeterocycles	2.71	0.47	0.17	2.25	0.47	0.18
NumSaturatedHeterocycles	0.15	0.11	0.08	0.13	0.1	0.09
NumAliphaticHeterocycles	0.23	0.19	0.14	0.14	0.19	0.14
NumSpiroAtoms	0.01	0.01	0.0	0.01	0.01	0.01
NumBridgeheadAtoms	0.16	0.01	0.22	0.23	0.01	0.22
NumAtomStereoCenters	0.59	0.49	0.78	0.57	0.51	0.74
NumUnspecifiedAtomStereoCenters	0.46	0.46	0.5	0.37	0.45	0.52
labuteASA	2.49	121.28	123.23	2.47	122.23	121.21
CrippenMR	2.1	77.09	76.02	2.34	77.63	74.85
chi0v	2.11	12.47	12.54	2.13	12.52	12.44
chi1v	2.61	7.37	7.34	3.04	7.37	7.33
chi2v	2.47	4.05	4.64	2.46	4.03	4.69
chi3v	2.46	4.05	4.64	2.47	4.03	4.69
chi4v	2.39	2.75	3.29	2.36	2.74	3.32
chi0n	3.23	11.41	10.4	3.66	11.45	10.3
chi1n	2.61	6.19	5.73	2.93	6.22	5.65
chi2n	2.51	2.82	2.82	2.71	2.85	2.76
chi3n	2.45	2.82	2.82	2.6	2.85	2.76
chi4n	2.25	1.83	1.88	2.13	1.84	1.84
hallKierAlpha	3.58	-1.51	-0.88	5.41	-1.55	-0.78
kappa1	2.28	15.37	15.16	2.29	15.39	15.1
kappa2	3.08	6.58	6.37	2.66	6.56	6.4
kappa3	3.64	4.04	4.08	2.64	4.0	4.16
Phi	2.51	5.21	5.27	2.45	5.16	5.37

Table S18. MACCS-based RF best estimator feature importances % of a subset of structural keys (groups) for Br^- and $O_2{}^-$. The subset contains the keys that reach 1% of importance for at least one ionization scheme (either positive or negative). For each key, the structure, the importance value (IMP %) and the proportion of presence (PP%) with, in addition, the average group count per molecule (Avg) for detected (D) and undetected (ND) molecules are stated. In the name of the structures, the special characters stand for: "A": any element, "X": halogen, "!": chain or non-ring bond, "$": ring bond, "%": aromatic query bond, and "not%": atom is at an aromatic/nonaromatic boundary.

		Br^-					$O_2{}^-$				
			D		ND			D		ND	
Structure	Key	IMP %	PP %	Avg	PP %	Avg	IMP %	PP %	Avg	PP %	Avg
NH	151	3.56	50.33	1.28	25.95	1.38	5.64	64.22	1.26	23.79	1.41
OH	139	2.29	21.0	1.02	6.11	1.25	3.19	24.77	1.02	6.95	1.18
N	161	0.66	83.67	2.39	71.5	2.05	0.57	87.16	2.41	72.0	2.1
NA(A)A	156	0.5	77.33	3.1	64.12	2.83	0.53	79.36	3.32	65.47	2.76
X	134	1.7	66.67	2.9	48.09	2.68	1.19	67.43	2.86	50.95	2.75
XA(A)A	107	1.87	64.33	3.43	43.0	3.6	1.38	64.68	3.33	46.53	3.63
X!A$A	87	1.59	54.0	4.13	36.39	4.94	1.19	53.67	3.79	39.58	4.95
Cl	103	1.18	53.33	1.94	40.2	2.37	0.99	50.46	1.72	43.79	2.38
F	42	1.56	25.67	3.34	10.43	2.56	1.81	28.9	3.48	11.58	2.6
C=O	154	1.8	68.33	1.34	50.38	1.32	1.52	72.48	1.31	51.58	1.34
CH_3	160	0.76	74.67	2.58	82.7	2.87	1.16	70.18	2.55	83.37	2.83
CH_3 > 1	149	1.39	56.0	3.1	71.5	3.16	1.14	53.21	3.04	70.11	3.17
CH_3 > 2 (&...)	141	0.94	35.0	3.76	43.77	3.9	1.21	29.82	3.86	44.63	3.84
NC(O)N	37	1.13	17.0	1.2	6.36	1.24	0.71	19.72	1.16	6.95	1.27
NC(C)N	38	0.58	10.33	1.16	3.31	1.08	1.12	13.3	1.14	3.16	1.13
Charge	49	0.4	3.0	2.33	10.43	2.66	0.17	5.96	2.69	7.79	2.57
NN	52	1.24	23.33	1.04	7.63	1.2	0.41	17.89	1.08	12.84	1.1
C%N	65	0.9	42.33	3.33	25.95	3.87	0.71	42.2	3.36	28.84	3.72
NAN	77	1.02	41.67	1.99	23.16	3.04	0.76	42.66	2.08	25.89	2.71
NAAN	79	1.1	28.33	1.87	12.98	1.75	0.73	25.69	1.91	16.84	1.76
CN(C)C	85	1.02	16.33	1.14	25.95	1.41	0.69	16.06	1.06	24.42	1.41
QHAACH$_2$A	90	1.09	13.67	1.17	4.07	1.31	0.71	13.76	1.17	5.68	1.26
QHAAACH$_2$A	91	1	15.0	1.04	4.33	1.29	1.98	19.27	1.07	4.21	1.2
OC(N)C	92	0.89	35.0	1.14	25.45	1.26	1.05	38.07	1.18	25.68	1.21
QCH$_3$	93	0.85	37.33	1.63	39.44	1.85	1.1	37.16	1.63	39.16	1.81
NAAO	95	1.05	34.33	1.45	20.87	1.51	0.7	28.44	1.44	25.89	1.5
NAAAO	97	1.77	49.0	2.58	26.97	2.94	0.92	48.17	2.76	31.16	2.71
QHACH$_2$A	104	0.92	14.0	1.19	5.34	1.33	1.53	16.51	1.22	5.68	1.26
ACH$_2$O	109	0.78	26.33	1.34	29.77	1.56	1.02	18.35	1.23	32.84	1.54
NCO	110	0.81	53.0	2.01	40.46	1.82	0.89	54.13	1.97	42.11	1.88
NAO	117	1.02	57.67	2.32	42.75	2.1	0.93	59.17	2.32	44.63	2.15
Heterocyclic atom > 1 (&...)	120	0.62	40.67	2.71	29.52	2.57	0.62	35.32	2.56	33.89	2.68
N heterocycle	121	0.49	47.67	2.14	35.37	1.98	0.65	46.33	2.01	38.11	2.09
OCO	123	0.91	30.67	1.11	31.04	1.14	1.22	26.61	1.05	32.84	1.15
Aromatic ring > 1	125	1.44	49.33	nan	32.06	nan	1.05	46.33	nan	36.42	nan
A!O!A	126	0.66	56.0	1.58	59.29	2.0	0.9	48.62	1.45	62.11	1.96
ACH$_2$AAACH$_2$A	128	0.71	20.0	3.13	29.26	4.46	1.02	12.84	3.39	30.95	4.12
A$A!N	133	0.68	38.0	2.47	33.59	2.92	1.04	47.71	2.54	29.89	2.85
Nnot%A%A	135	0.75	35.67	2.41	32.57	2.86	1.04	46.33	2.51	28.21	2.76
O=A > 1	136	1.22	35.33	2.51	25.19	2.36	0.81	34.86	2.57	27.16	2.36
Heterocycle	137	0.61	53.0	2.31	42.75	2.08	0.56	50.46	2.09	45.68	2.25
QCH$_2$A > 1 (&...)	138	0.54	22.0	2.59	30.03	2.53	1.17	12.84	2.61	32.84	2.54
O > 3 (&...)	140	0.83	27.33	4.59	29.77	4.62	1.09	24.31	4.6	30.74	4.61
N > 1	142	1.34	58.0	3.01	36.13	3.08	0.9	60.09	3.04	38.95	3.04
Anot%A%Anot%A	144	1.01	59.33	1.98	45.29	2.01	1.36	61.01	2.1	46.95	1.94
6M RING > 1	145	0.87	38.67	2.19	33.33	2.24	1.09	38.53	2.17	34.32	2.25
O > 2	146	1.09	51.0	3.85	48.85	3.99	0.74	46.79	3.83	51.16	3.97
A!A$A!A	150	1.04	70.0	2.43	54.45	2.87	0.98	69.27	2.43	57.47	2.78
QCH$_2$A	153	0.66	44.67	1.78	50.89	1.9	1.37	33.94	1.61	54.74	1.92
C-N	158	0.94	70.33	2.73	57.0	2.97	0.75	72.02	2.87	58.53	2.84

Table S19. MACCS-based RF best estimator feature importances % of a subset of structural keys (groups) for H_3O^+ and $AceH^+$. The subset contains the keys that reach 1% of importance for at least one ionization scheme (either positive or negative). For each key, the structure, the importance value (IMP %) and the proportion of presence (PP%) with, in addition, the average group count per molecule (Avg) for detected (D) and undetected (ND) molecules are stated. In the name of the structures, the special characters stand for: "A": any element, "X": halogen, "!": chain or non-ring bond, "$": ring bond, "%": aromatic query bond, and "not%": atom is at an aromatic/nonaromatic boundary.

		H_3O^+					$AceH^+$				
			D		ND			D		ND	
Structure	Key	IMP %	PP %	Avg	PP %	Avg	IMP %	PP %	Avg	PP %	Avg
NH	151	1.44	44.68	1.29	19.28	1.49	1.82	44.94	1.28	18.26	1.58
OH	139	0.81	10.43	1.0	17.04	1.18	0.82	10.34	1.0	17.35	1.18
N	161	3.65	88.3	2.28	52.47	1.97	4.0	88.19	2.26	52.05	2.04
NA(A)A	156	3.8	82.98	2.87	42.15	3.31	3.93	82.91	2.88	41.55	3.29
X	134	1.33	49.15	2.41	70.85	3.36	1.27	50.42	2.44	68.49	3.36
XA(A)A	107	1.27	45.53	2.79	66.37	4.55	1.35	47.05	2.81	63.47	4.64
X!A$A	87	2.07	37.02	3.23	58.74	6.21	1.24	39.66	3.26	53.42	6.52
Cl	103	1.81	38.51	1.5	61.43	3.01	1.34	40.51	1.51	57.53	3.13
F	42	0.63	19.15	3.04	12.56	3.14	0.48	18.99	3.1	12.79	2.96
C=O	154	0.82	63.62	1.31	46.64	1.39	1.57	64.77	1.32	43.84	1.38
CH_3	160	1.44	85.96	2.9	65.02	2.34	1.15	84.81	2.86	67.12	2.44
$CH_3 > 1$	149	2.28	72.55	3.25	48.43	2.8	1.69	70.68	3.23	52.05	2.86
$CH_3 > 2$ (&...)	141	1.85	47.66	3.9	23.77	3.62	1.34	45.36	3.92	28.31	3.58
NC(O)N	37	0.3	12.77	1.18	7.17	1.31	0.33	12.24	1.12	8.22	1.5
NC(C)N	38	0.21	7.23	1.15	4.48	1.1	0.15	7.59	1.17	3.65	1.0
Charge	49	1.14	3.83	2.83	14.35	2.47	0.99	3.8	2.72	14.61	2.53
NN	52	0.27	18.51	1.1	5.83	1.0	0.36	18.99	1.1	4.57	1.0
C%N	65	1.8	41.91	3.52	14.35	3.88	1.38	41.35	3.48	15.07	4.09
NAN	77	1.12	38.51	2.45	15.7	2.37	0.98	38.4	2.36	15.53	2.85
NAAN	79	0.64	25.11	1.83	8.07	1.78	0.54	24.68	1.84	8.68	1.74
CN(C)C	85	0.59	25.74	1.13	13.45	2.1	0.44	23.84	1.14	17.35	1.87
$QHAACH_2A$	90	0.18	10.43	1.18	3.59	1.38	0.27	10.34	1.18	3.65	1.38
$QHAAACH_2A$	91	0.23	10.64	1.12	5.38	1.08	0.27	10.76	1.1	5.02	1.18
OC(N)C	92	0.77	35.53	1.14	17.04	1.45	1.16	36.92	1.15	13.7	1.47
QCH_3	93	0.97	43.4	1.78	28.25	1.67	0.9	41.77	1.76	31.51	1.75
NAAO	95	0.94	32.98	1.45	13.45	1.63	1.53	33.76	1.46	11.42	1.6
NAAAO	97	1.2	44.89	2.54	18.83	3.69	1.02	44.09	2.45	20.09	4.07
$QHACH_2A$	104	0.25	10.0	1.28	7.17	1.12	0.33	10.34	1.22	6.39	1.29
ACH_2O	109	0.52	31.28	1.46	21.97	1.53	0.65	31.22	1.43	21.92	1.6
NCO	110	1.78	55.96	1.82	24.66	2.36	1.54	55.7	1.8	24.66	2.48
NAO	117	2.5	60.21	2.03	26.01	3.1	2.36	59.92	2.0	26.03	3.3
Heterocyclic atom > 1 (&...)	120	1.01	42.13	2.65	17.94	2.62	0.74	41.14	2.66	19.63	2.56
N heterocycle	121	2.2	50.21	2.08	20.63	1.96	1.56	49.79	2.07	21.0	2.0
OCO	123	0.65	30.0	1.13	32.74	1.11	0.75	31.01	1.14	30.59	1.1
Aromatic ring > 1	125	0.73	42.77	nan	32.74	nan	1.03	44.73	nan	28.31	nan
A!O!A	126	0.88	61.28	1.85	50.67	1.77	1.01	61.6	1.82	49.77	1.83
ACH_2AAACH_2A	128	0.46	27.45	3.37	20.63	5.78	0.55	26.79	3.24	21.92	6.02
A$A!N	133	0.55	38.09	2.76	30.04	2.6	0.64	39.03	2.75	27.85	2.62
Nnot%A%A	135	0.62	35.74	2.69	30.04	2.57	0.57	36.71	2.68	27.85	2.59
O=A > 1	136	0.85	29.36	2.37	30.04	2.58	0.77	29.75	2.3	29.22	2.75
Heterocycle	137	1.58	56.17	2.23	28.25	2.03	1.34	55.91	2.22	28.31	2.08
$QCH_2A > 1$ (&...)	138	0.44	29.36	2.59	20.63	2.41	0.5	27.85	2.57	23.74	2.5
O > 3 (&...)	140	1.53	29.36	4.51	27.35	4.82	1.44	28.48	4.47	29.22	4.89
N > 1	142	1.94	55.53	3.03	24.66	3.07	1.64	55.06	3.02	25.11	3.15
Anot%A%Anot%A	144	0.73	50.43	1.95	53.36	2.09	0.88	51.27	1.95	51.6	2.1
6M RING > 1	145	0.73	35.32	2.2	36.32	2.25	1.03	37.97	2.21	30.59	2.25
O > 2	146	0.79	50.85	3.87	47.53	4.05	1.32	51.27	3.82	46.58	4.19
A!A$A!A	150	0.66	57.66	2.26	68.61	3.35	0.69	58.65	2.32	66.67	3.29
QCH_2A	153	0.79	54.47	1.86	34.98	1.83	0.64	52.11	1.84	39.73	1.9
C-N	158	1.4	72.13	2.9	43.05	2.68	1.03	71.73	2.86	43.38	2.8

S6 Additional ML models: NB, SVC (classifiers); linear KRR and RF regressor (regression)

Table S20. Hyperparameters tuning range list for each ML model and each molecular descriptor.

Model	Hyperparameters	Tuning range	Info
RF regressor	Max depth	[20, 40, 60, 80, 100, None]	The length of each tree, from the root to the leaves
	Min samples leaf	[1, 2, 4]	Minimum number of samples per leaf
	Min samples split	[2, 5, 10]	Minimum number of samples per split
	N estimators	[100, 500, 1000, 1500, 2000]	Maximum number of estimators
SVC	C	[0.1, 1, 10, 100]	Regularization parameter
	Kernel	[’rbf’,’poly’,’sigmoid’]	Kernel type
	γ	[’scale’,’auto’]	Kernel coefficient
KRR with Linear kernel	None		
NB	None		

Table S21. Hyperparameters tuned for NB model with TopFP as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters		
			TopFP		
			Fp size	Max path	N bits per hash
Br^-	554	555	8192	7	2
		8	4096	8	2
		52	8192	7	2
		1066	4096	7	4
		324	8192	7	2
$O_2{}^-$	554	555	4096	7	4
		8	8192	9	2
		52	8192	8	2
		1066	2048	7	2
		324	8192	7	2
H_3O^+	554	555	8192	8	2
		8	716	9	8
		52	2048	10	16
		1066	716	10	16
		324	8192	10	2
$AceH^+$	554	555	2048	10	16
		8	8192	7	4
		52	716	10	8
		1066	8192	7	2
		324	2048	10	16

Table S22. Hyperparameters tuned for NB model with MBTR as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			MBTR			
			σ_2	w_2	σ_3	w_3
Br^-	554	555	0.1	0.4	0.0001	0.8
		8	0.0005	0.4	0.01	0.8
		52	0.0005	0.6	0.001	0.6
		1066	0.005	1	0.0001	0.4
		324	0.0001	1.2	0.005	0.8
${O_2}^-$	554	555	0.1	0.6	0.0005	1
		8	0.005	0.2	0.0005	1
		52	0.001	0.6	0.0005	0.8
		1066	0.001	0.6	0.001	0.8
		324	0.01	0.4	0.0005	0.8
H_3O^+	554	555	0.1	0.8	0.005	1
		8	0.0005	1.2	0.001	0.8
		52	0.001	0.2	0.005	0.8
		1066	0.001	1.2	0.0001	0.8
		324	0.0005	0.6	0.005	0.8
$AceH^+$	554	555	0.0005	0.8	0.0001	1
		8	0.001	0.6	0.0001	1
		52	0.01	0.4	0.1	0.8
		1066	0.001	0.8	0.005	1
		324	0.001	0.6	0.1	0.2

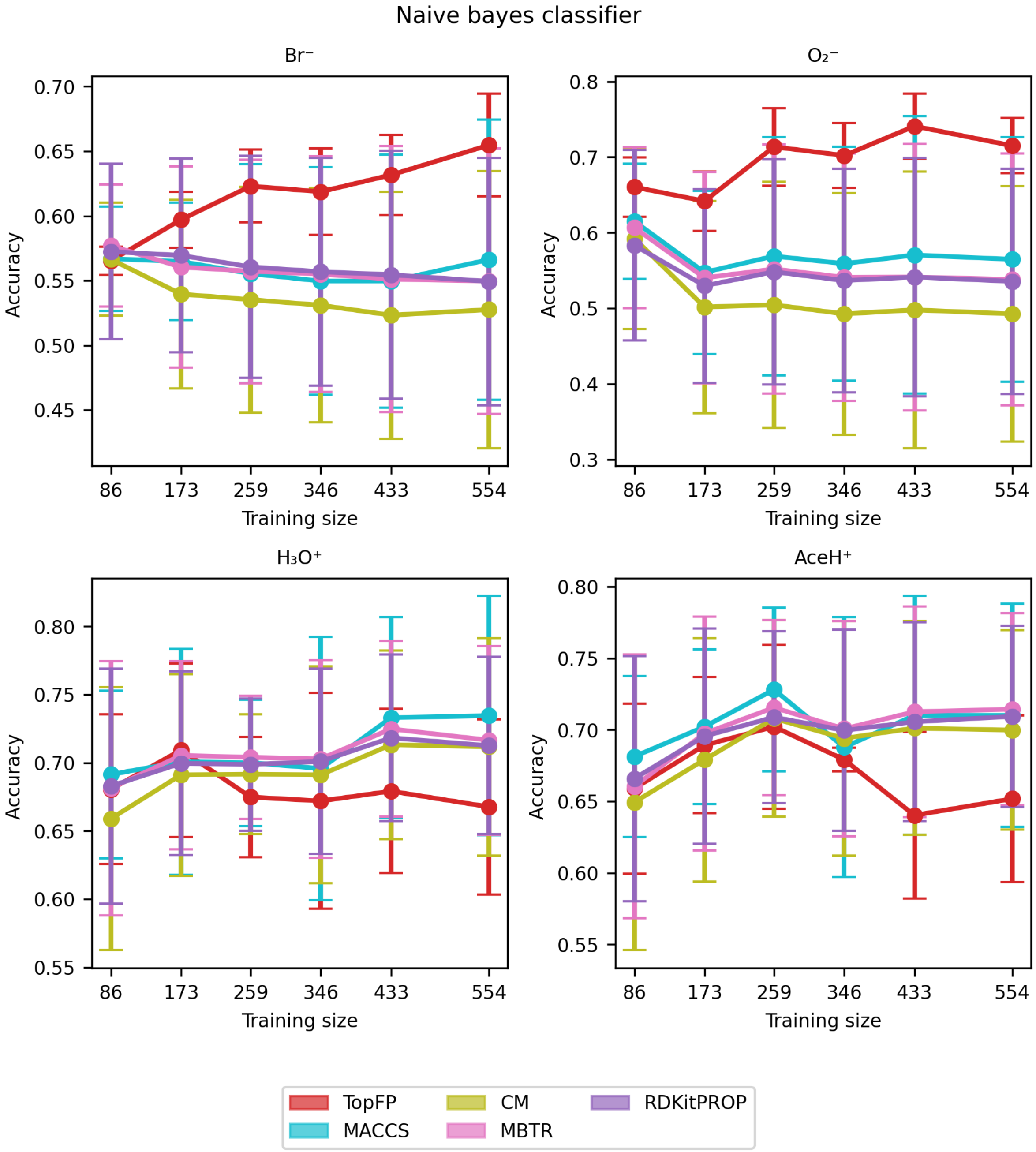


Figure S13. Learning curve of the naive bayes with the accuracy of the classification of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification accuracy. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

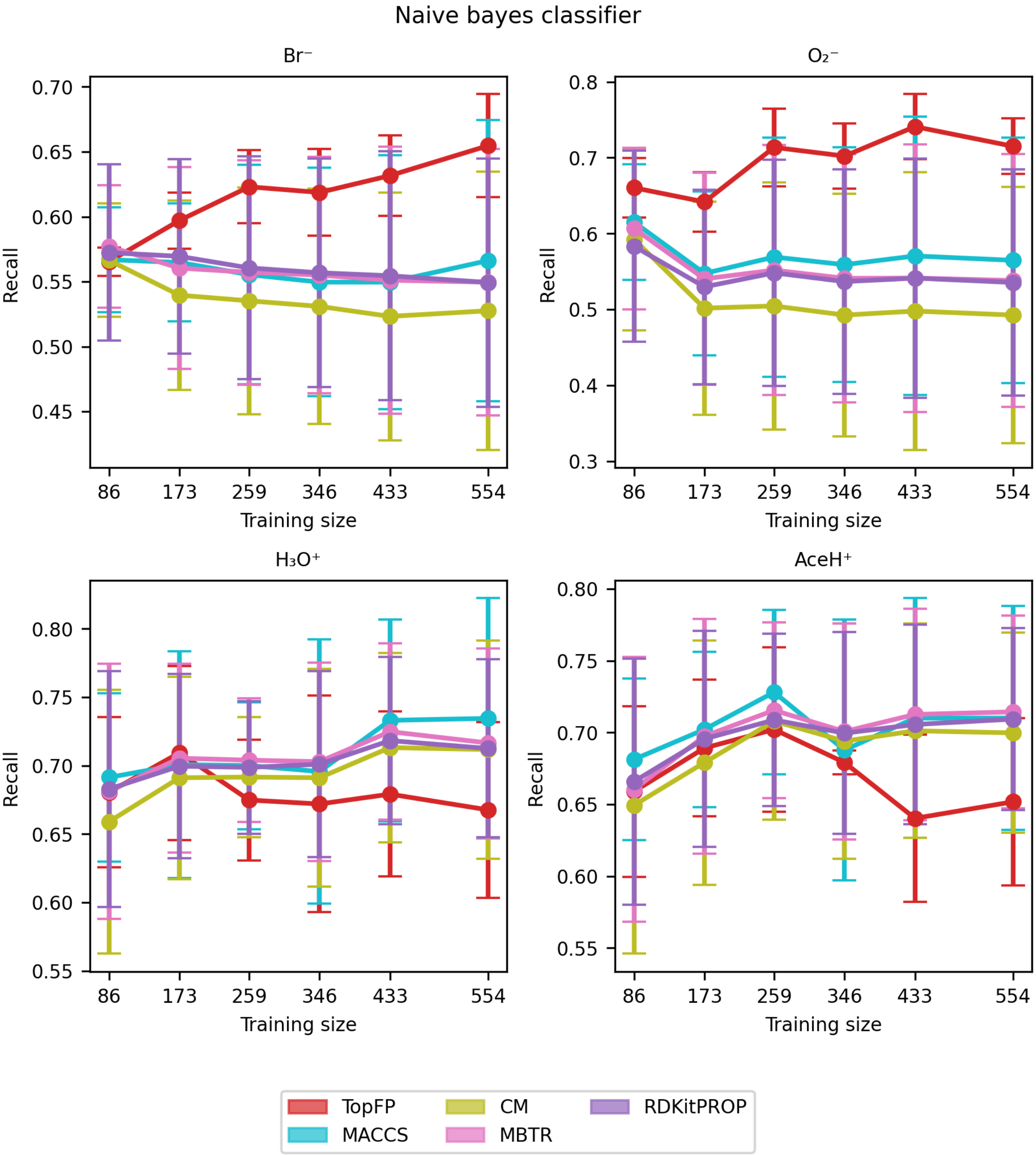


Figure S14. Learning curve of the naive bayes with the recall of the classification of Br$^-$, O$_2$$^-$, H$_3O^+$ and AceH$^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification recall. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

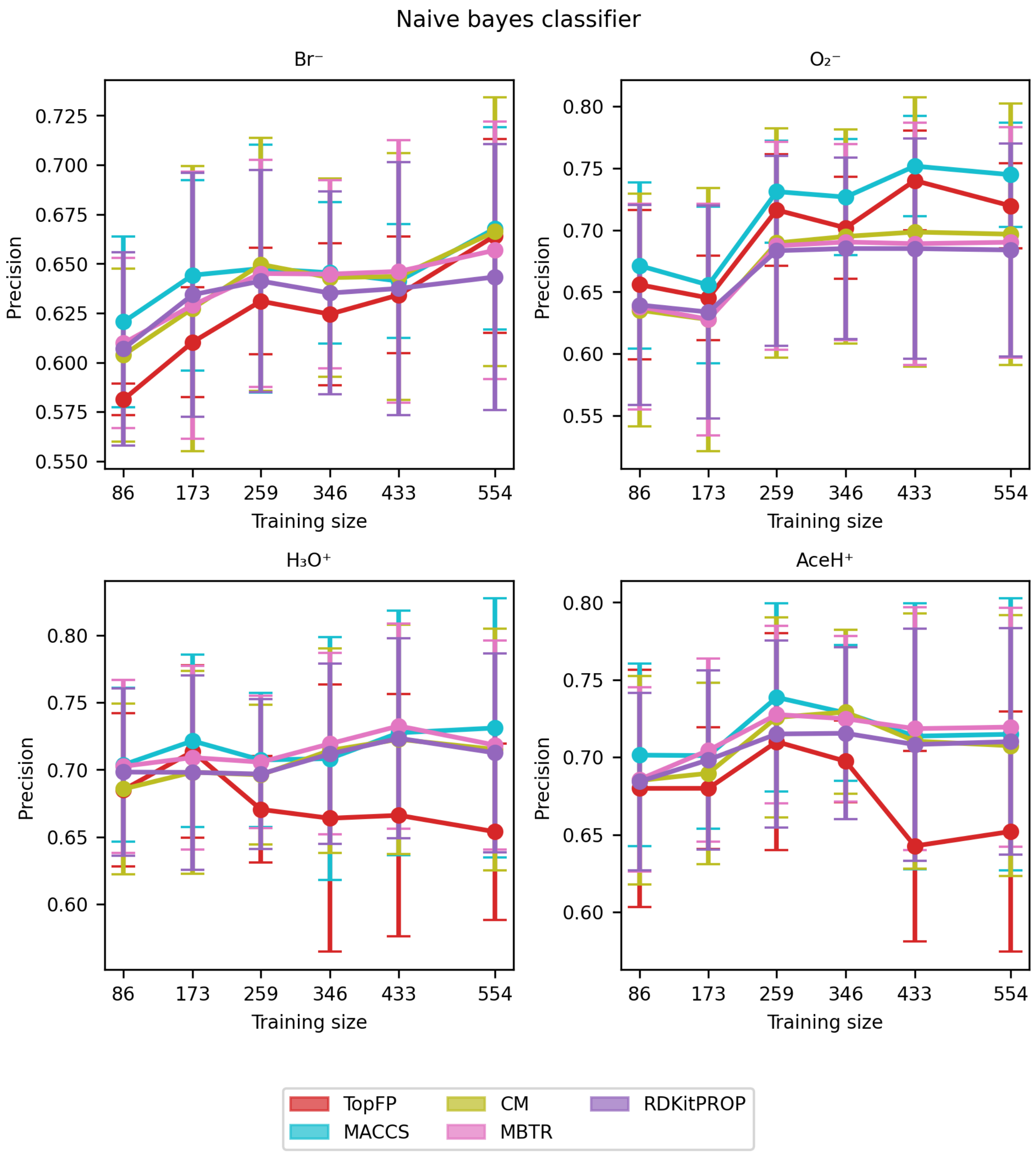


Figure S15. Learning curve of the naive bayes with the precision of the classification of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification precision. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

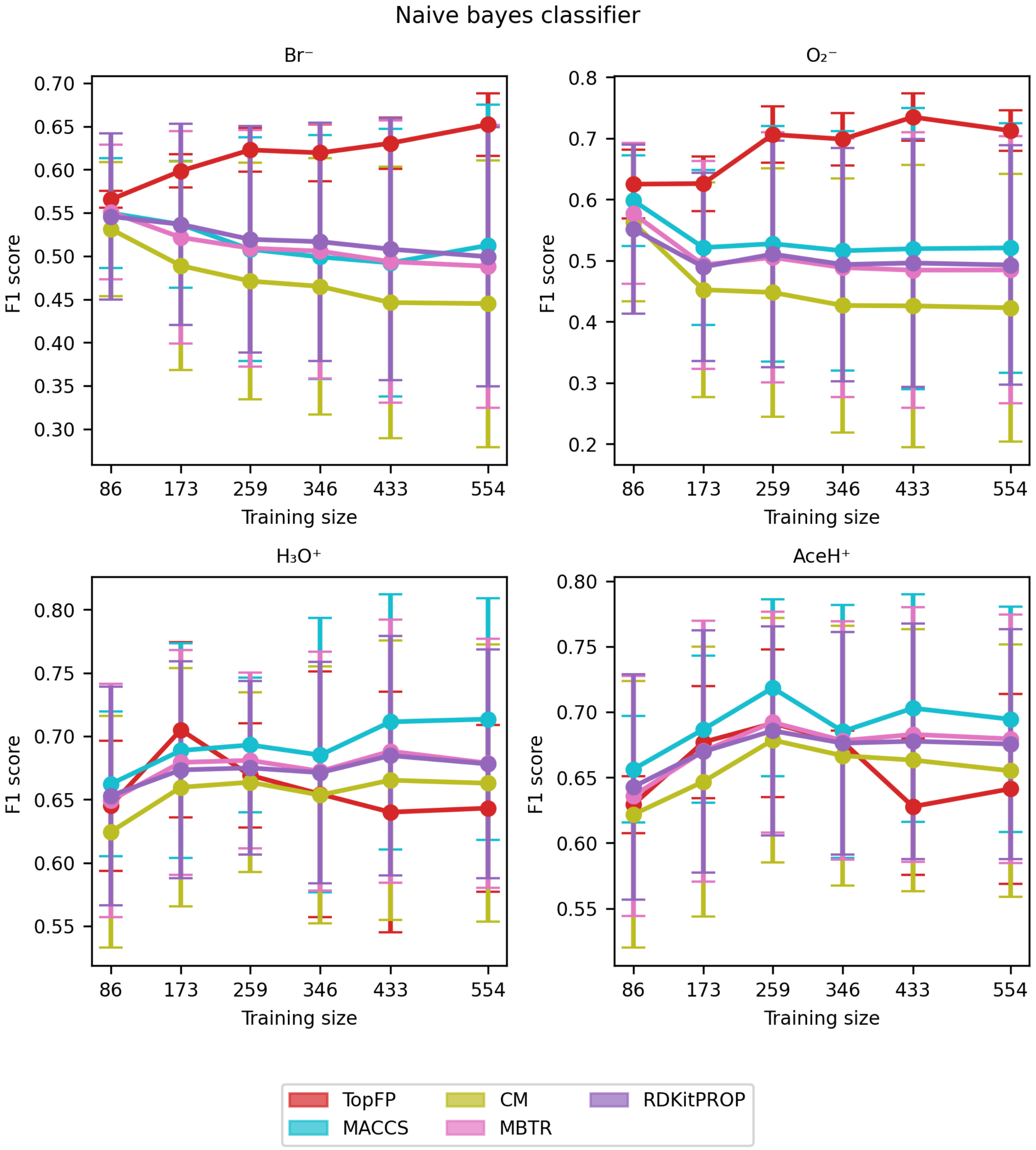


Figure S16. Learning curve of the naive bayes with the F1 score of the classification of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification F1 score. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

Table S23. Hyperparameters tuned for SVC model with RDKitPROP as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters		
			SVC		
			C	Kernel	γ
Br^-	554	555	100	rbf	auto
		8	100	rbf	auto
		52	1	rbf	auto
		1066	10	rbf	auto
		324	100	rbf	auto
$O_2{}^-$	554	555	100	rbf	auto
		8	10	rbf	auto
		52	10	rbf	auto
		1066	100	rbf	auto
		324	10	rbf	auto
H_3O^+	554	555	10	rbf	scale
		8	100	rbf	scale
		52	10	rbf	scale
		1066	1	rbf	auto
		324	10	rbf	scale
$AceH^+$	554	555	100	rbf	scale
		8	100	rbf	scale
		52	100	rbf	scale
		1066	100	rbf	scale
		324	100	rbf	scale

Table S24. Hyperparameters tuned for SVC model with TopFP as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters					
			TopFP			SVC		
			Fp size	Max path	N bits per hash	C	Kernel	γ
Br^-	554	555	4096	9	8	1	poly	scale
		8	8192	7	4	10	rbf	scale
		52	8192	7	16	100	poly	auto
		1066	8192	8	8	10	rbf	auto
		324	4096	7	16	0.1	poly	scale
$O_2{}^-$	554	555	2048	9	2	10	rbf	auto
		8	2048	7	2	100	rbf	scale
		52	2048	7	2	10	rbf	auto
		1066	4096	10	4	100	rbf	scale
		324	2048	9	4	1	poly	scale
H_3O^+	554	555	8192	8	4	100	rbf	scale
		8	2048	9	2	1	poly	scale
		52	8192	9	4	10	sigmoid	auto
		1066	4096	9	4	100	rbf	auto
		324	8192	7	8	10	poly	scale
$AceH^+$	554	555	4096	8	8	10	poly	scale
		8	8192	7	2	1	rbf	scale
		52	4096	7	4	100	rbf	scale
		1066	4096	9	2	100	rbf	auto
		324	2048	9	4	10	poly	auto

Table S25. Hyperparameters tuned for SVC model with MACCS as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters		
			SVC		
			C	Kernel	γ
Br^-	554	555	10	rbf	scale
		8	10	rbf	scale
		52	1	rbf	scale
		1066	1	poly	scale
		324	100	rbf	scale
${O_2}^-$	554	555	10	rbf	auto
		8	100	rbf	scale
		52	10	rbf	auto
		1066	100	rbf	scale
		324	100	rbf	scale
H_3O^+	554	555	10	rbf	scale
		8	10	sigmoid	auto
		52	10	rbf	auto
		1066	10	rbf	auto
		324	10	rbf	auto
$AceH^+$	554	555	10	sigmoid	auto
		8	10	rbf	auto
		52	10	rbf	scale
		1066	100	poly	auto
		324	1	poly	scale

Table S26. Hyperparameters tuned for SVC model with CM as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters		
			SVC		
			C	Kernel	γ
Br^-	554	555	100	rbf	auto
		8	100	rbf	auto
		52	10	rbf	auto
		1066	100	rbf	scale
		324	100	rbf	scale
${O_2}^-$	554	555	1	rbf	auto
		8	1	rbf	auto
		52	1	rbf	auto
		1066	10	rbf	auto
		324	10	rbf	auto
H_3O^+	554	555	1	rbf	scale
		8	100	rbf	scale
		52	1	poly	scale
		1066	100	rbf	auto
		324	10	rbf	scale
$AceH^+$	554	555	0.1	poly	scale
		8	10	rbf	scale
		52	10	poly	scale
		1066	100	rbf	scale
		324	1	sigmoid	scale

Table S27. Hyperparameters tuned for SVC model with MBTR as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters						
			MBTR				SVC		
			σ_2	w_2	σ_3	w_3	C	Kernel	γ
Br^-	554	555	0.1	0.4	0.005	0.6	10	rbf	scale
		8	0.1	0.2	0.1	1.2	10	poly	scale
		52	0.1	1.2	0.001	0.6	10	poly	scale
		1066	0.001	0.2	0.1	0.6	10	rbf	scale
		324	0.01	0.8	0.0001	0.6	100	rbf	scale
${O_2}^-$	554	555	0.0001	0.4	0.01	1.2	10	rbf	scale
		8	0.1	0.6	0.0005	0.6	100	rbf	scale
		52	0.0001	0.2	0.01	1	1	sigmoid	scale
		1066	0.01	0.4	0.1	1.2	100	rbf	scale
		324	0.001	1.2	0.1	1.2	10	rbf	scale
H_3O^+	554	555	0.005	0.4	0.0001	1.2	1	poly	scale
		8	0.1	1.2	0.0005	0.6	1	poly	scale
		52	0.1	0.8	0.01	0.8	100	rbf	scale
		1066	0.1	0.4	0.005	0.2	10	poly	scale
		324	0.001	1.2	0.001	0.8	100	rbf	scale
$AceH^+$	554	555	0.001	0.8	0.005	0.8	10	rbf	scale
		8	0.01	1.2	0.01	0.2	1	poly	scale
		52	0.005	1	0.001	1.2	10	poly	scale
		1066	0.005	0.2	0.001	1	100	rbf	scale
		324	0.01	1.2	0.1	0.4	0.1	sigmoid	scale

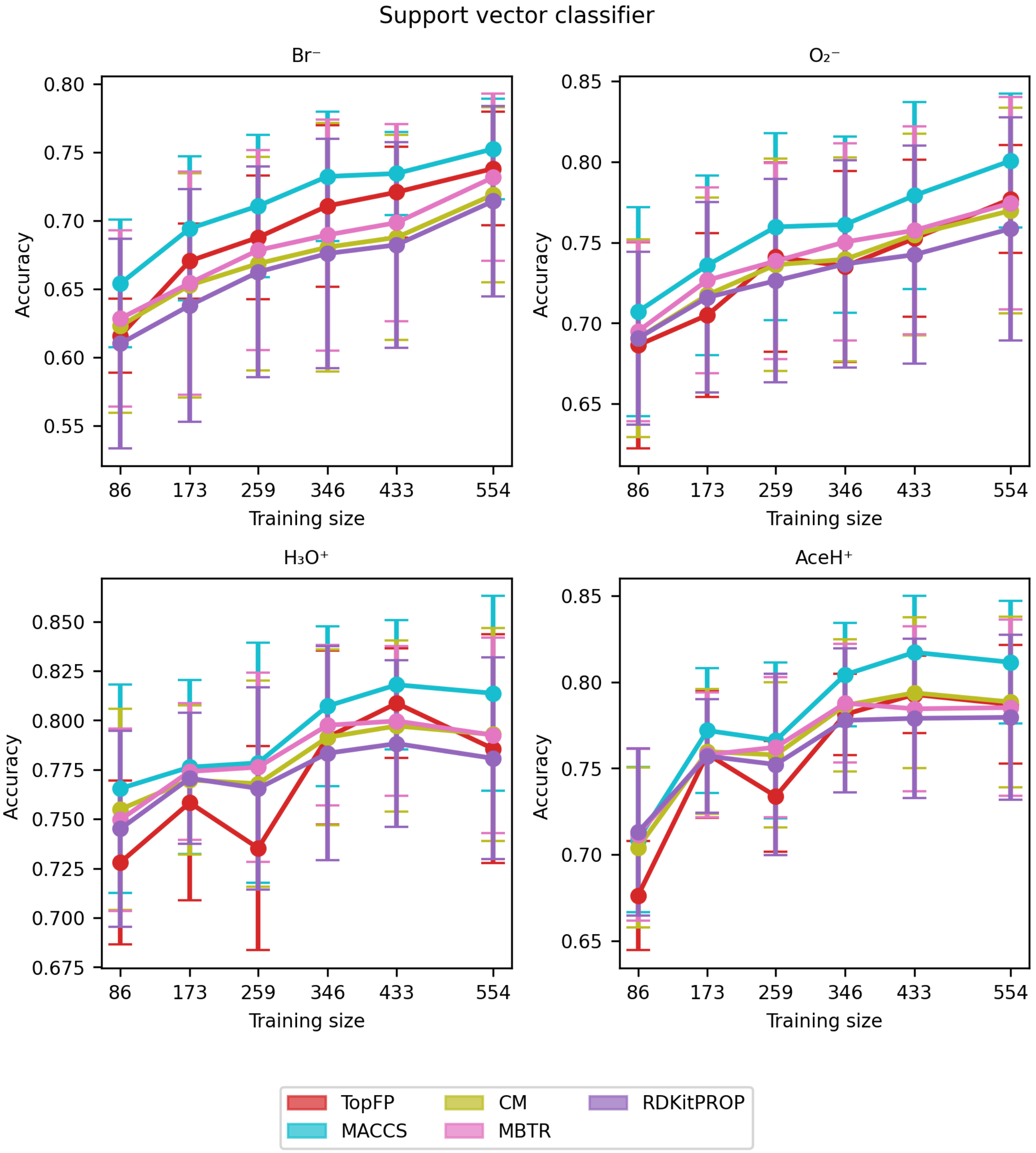


Figure S17. Learning curve of the support vector classifier with the accuracy of the classification of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification accuracy. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

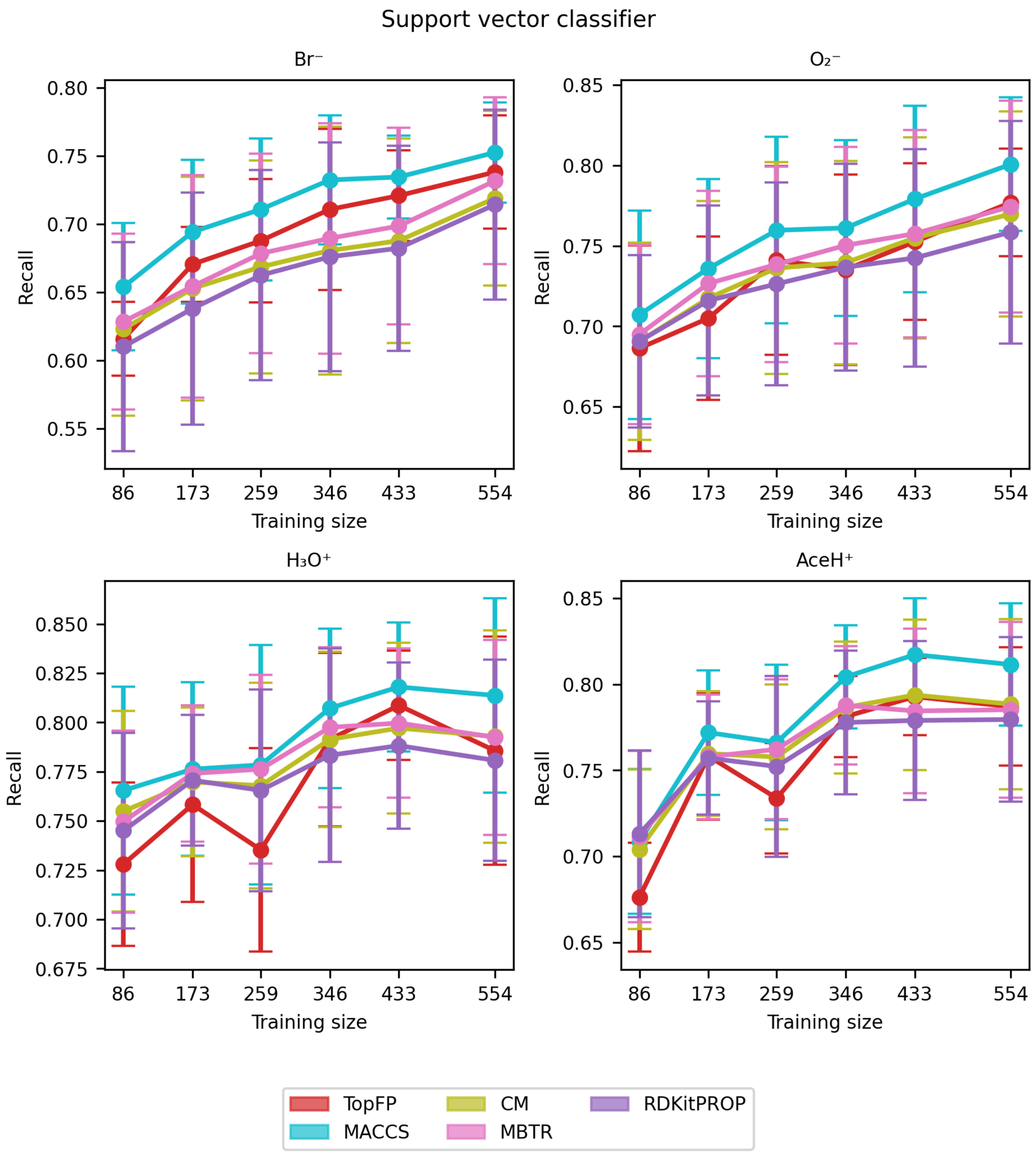


Figure S18. Learning curve of the support vector classifier with the recall of the classification of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification recall. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

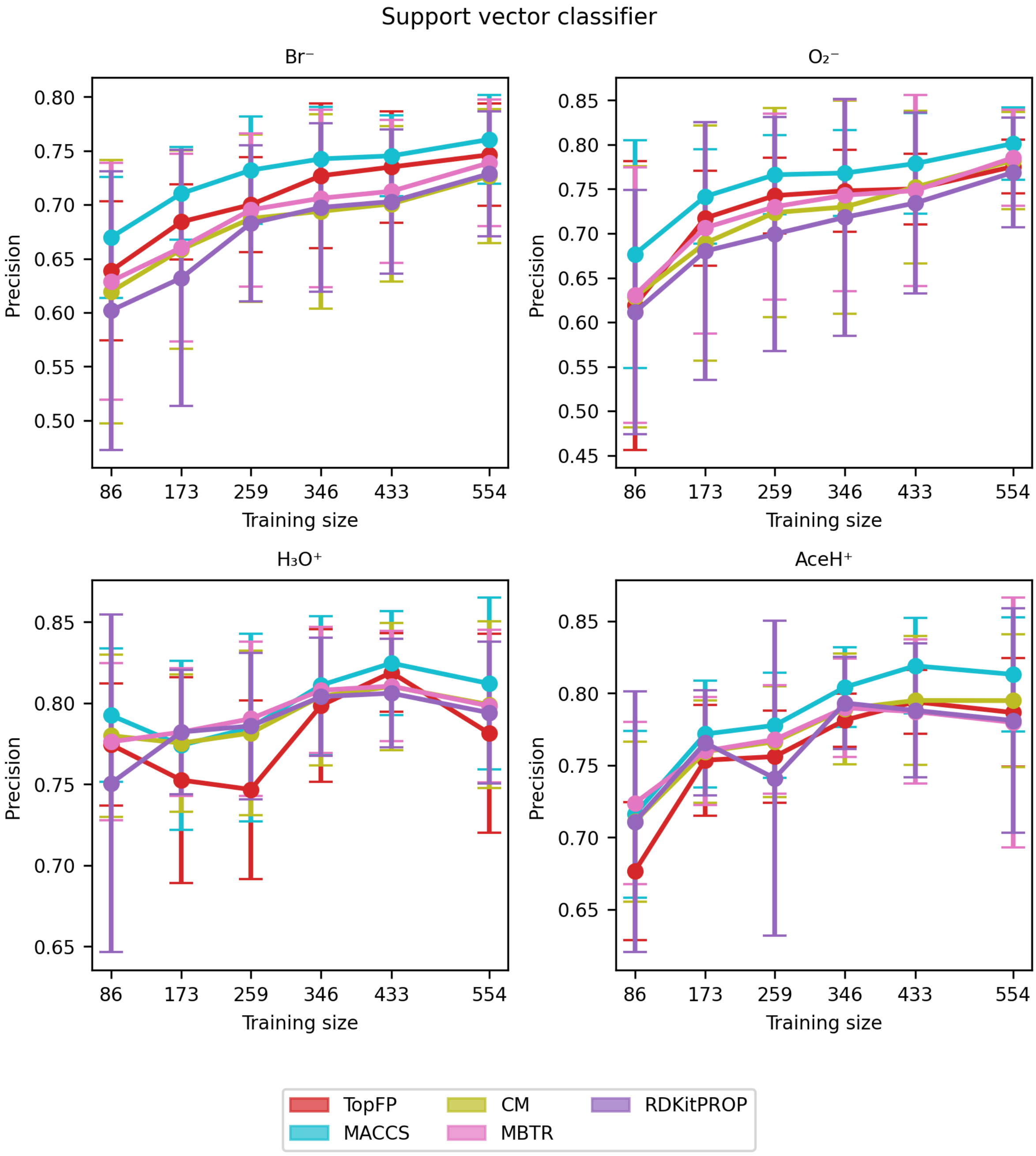


Figure S19. Learning curve of the support vector classifier with the precision of the classification of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification precision. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

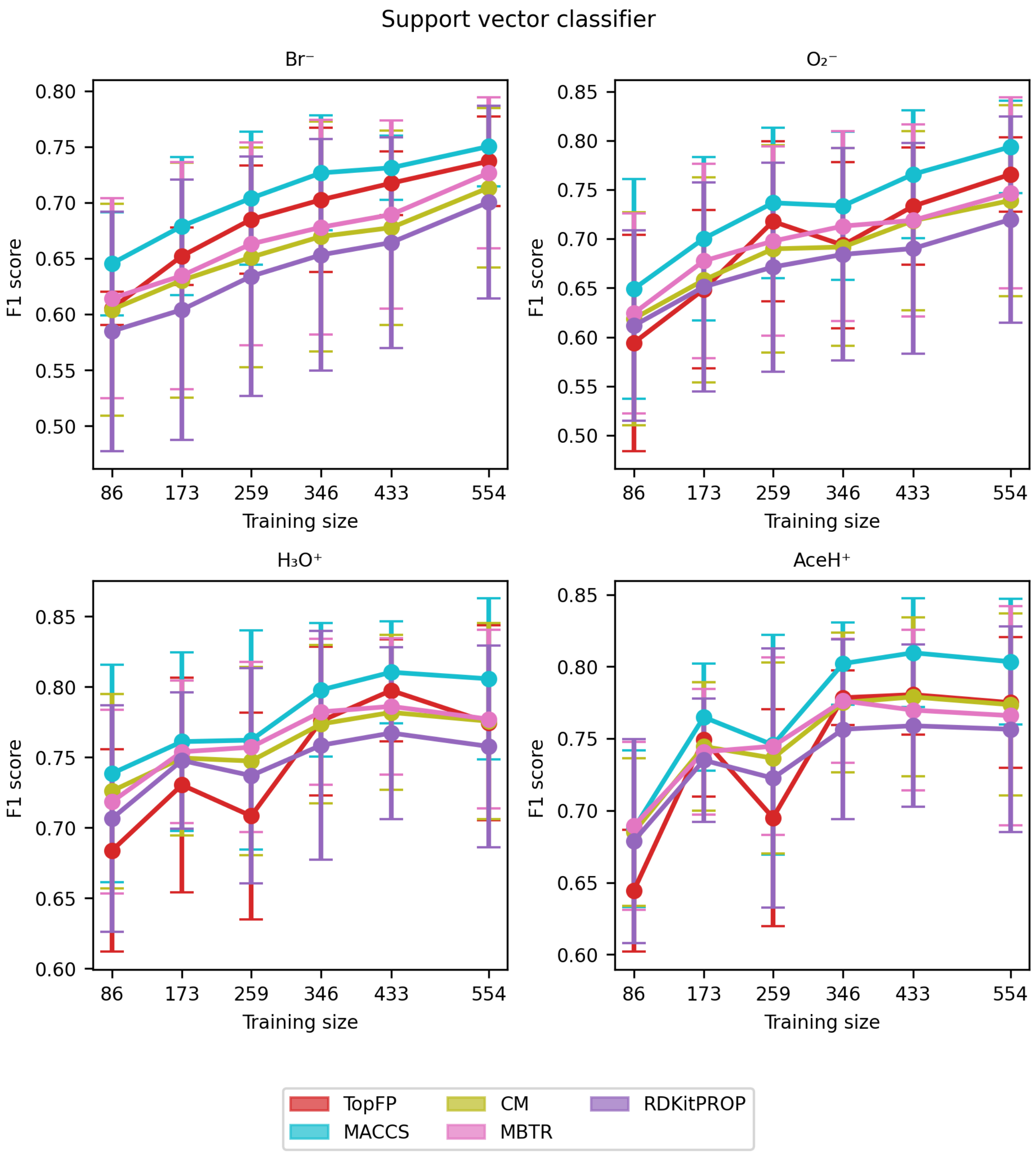


Figure S20. Learning curve of the support vector classifier with the F1 score of the classification of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the classification F1 score. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

Table S28. Hyperparameters tuned for KRR with Linear kernel model with TopFP as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters		
			TopFP		
			Fp size	Max path	N bits per hash
Br^-	240	555	716	8	2
		8	2048	9	2
		52	2048	7	8
		1066	2048	7	4
		324	8192	7	16
${O_2}^-$	174	555	716	8	16
		8	2048	8	4
		52	2048	9	16
		1066	716	9	16
		324	716	7	16
H_3O^+	376	555	4096	7	16
		8	716	9	16
		52	8192	7	16
		1066	716	8	16
		324	2048	7	16
$AceH^+$	379	555	8192	8	16
		8	4096	8	8
		52	716	8	16
		1066	4096	7	16
		324	4096	7	8

Table S29. Hyperparameters tuned for KRR with Linear kernel model with MBTR as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			MBTR			
			σ_2	w_2	σ_3	w_3
Br^-	240	555	0.01	0.2	0.001	0.4
		8	0.001	0.2	0.01	0.8
		52	0.0001	0.4	0.1	0.4
		1066	0.0001	0.2	0.1	1.2
		324	0.01	0.4	0.001	0.2
$O_2{}^-$	174	555	0.1	0.2	0.001	0.2
		8	0.001	0.2	0.01	0.2
		52	0.01	0.4	0.3	0.4
		1066	0.01	0.4	0.0001	0.8
		324	0.01	0.2	0.3	1.4
H_3O^+	376	555	0.0001	0.4	0.3	0.2
		8	0.3	0.2	0.01	0.2
		52	0.1	0.2	0.0001	0.2
		1066	0.1	0.4	0.01	0.2
		324	0.3	0.4	0.001	0.2
$AceH^+$	379	555	0.01	0.2	0.001	0.2
		8	0.01	0.2	0.0001	0.4
		52	0.01	0.2	0.0001	0.2
		1066	0.01	0.2	0.01	0.4
		324	0.01	0.2	0.001	0.8

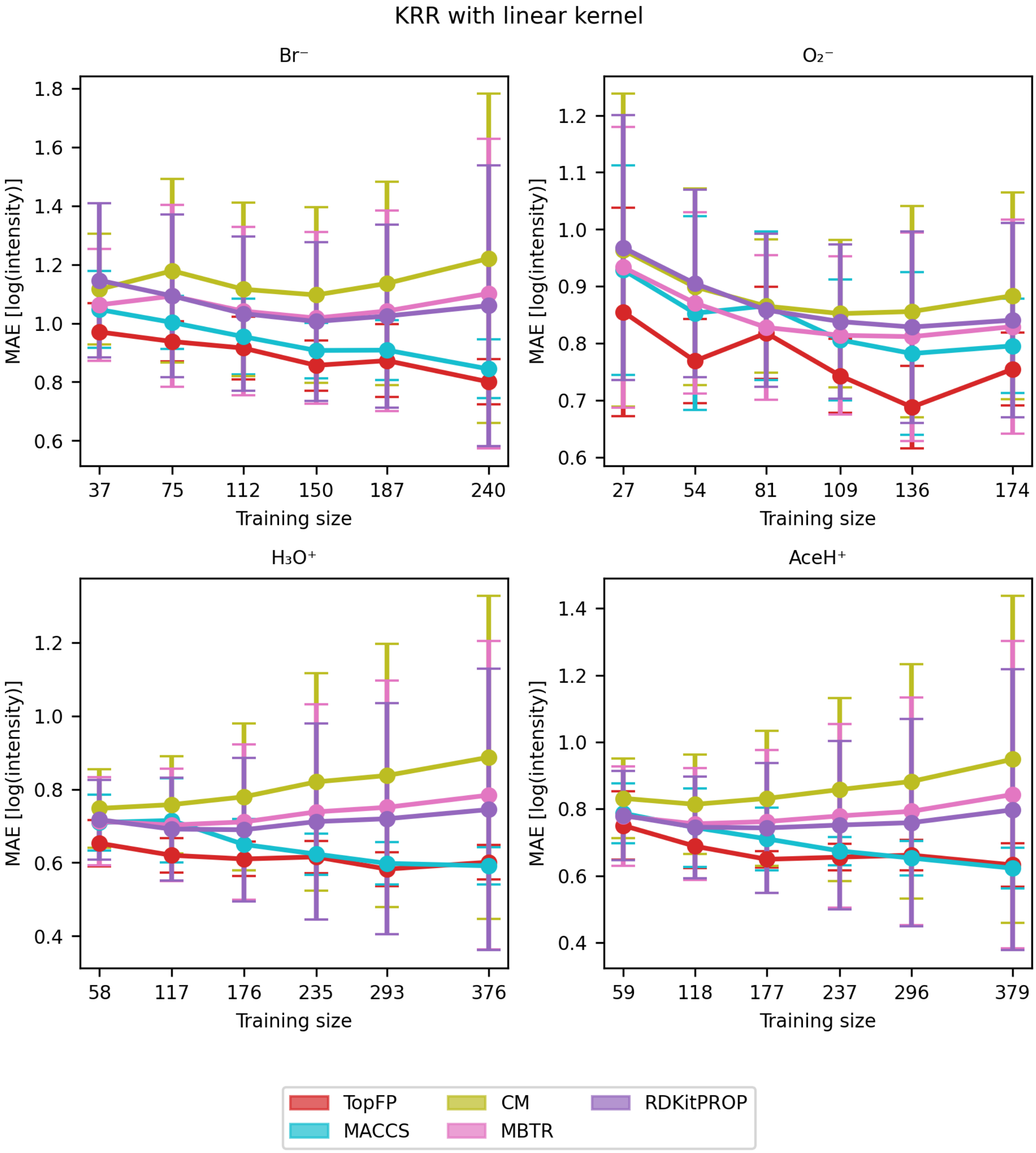


Figure S21. Linear KRR learning curve with mean absolute error (MAE) of the signal intensity values in logarithmic scale of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the MAE of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

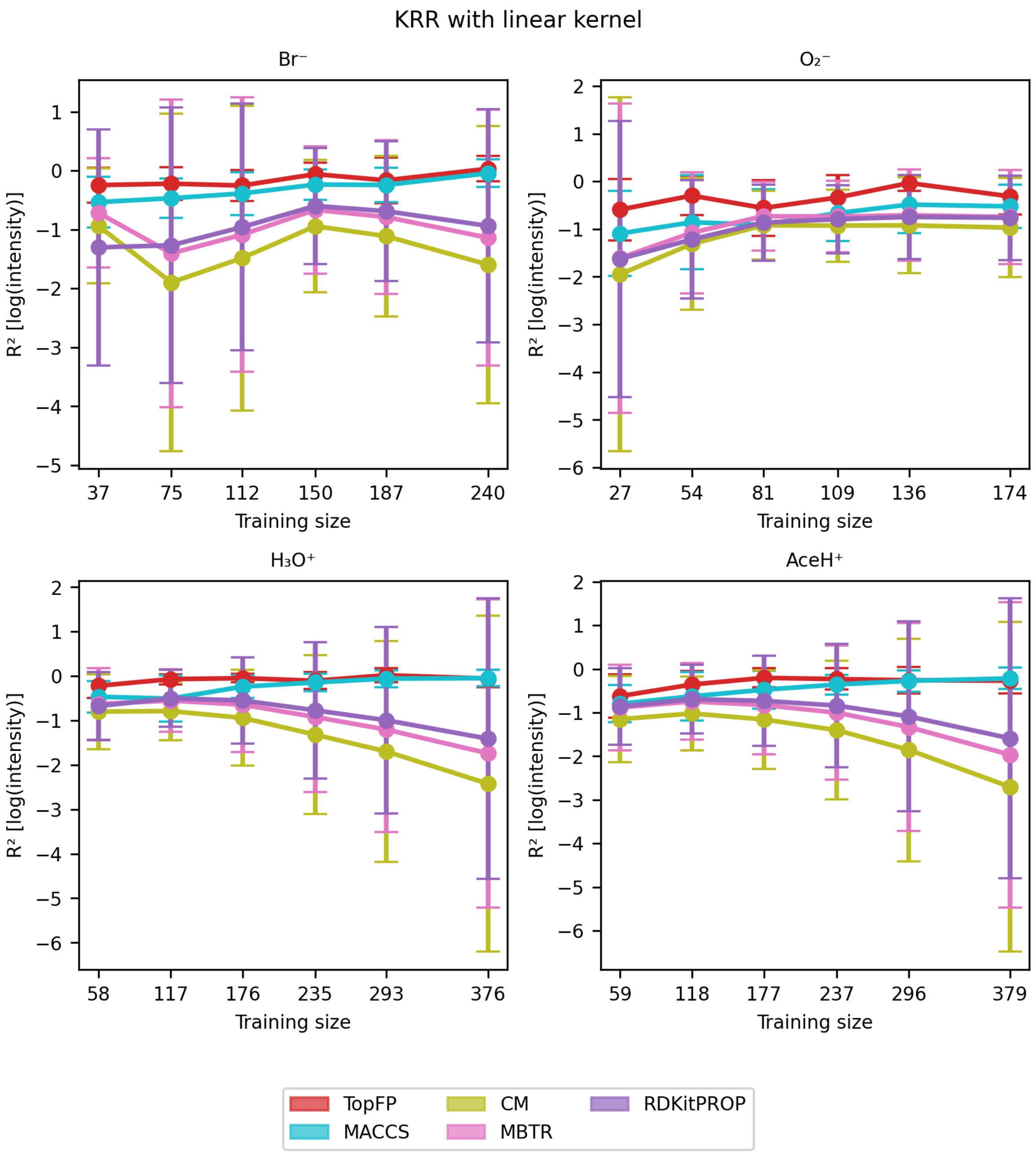


Figure S22. Linear KRR learning curve with correlation coefficient (R^2) of the signal intensity values in logarithmic scale of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the R^2 of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

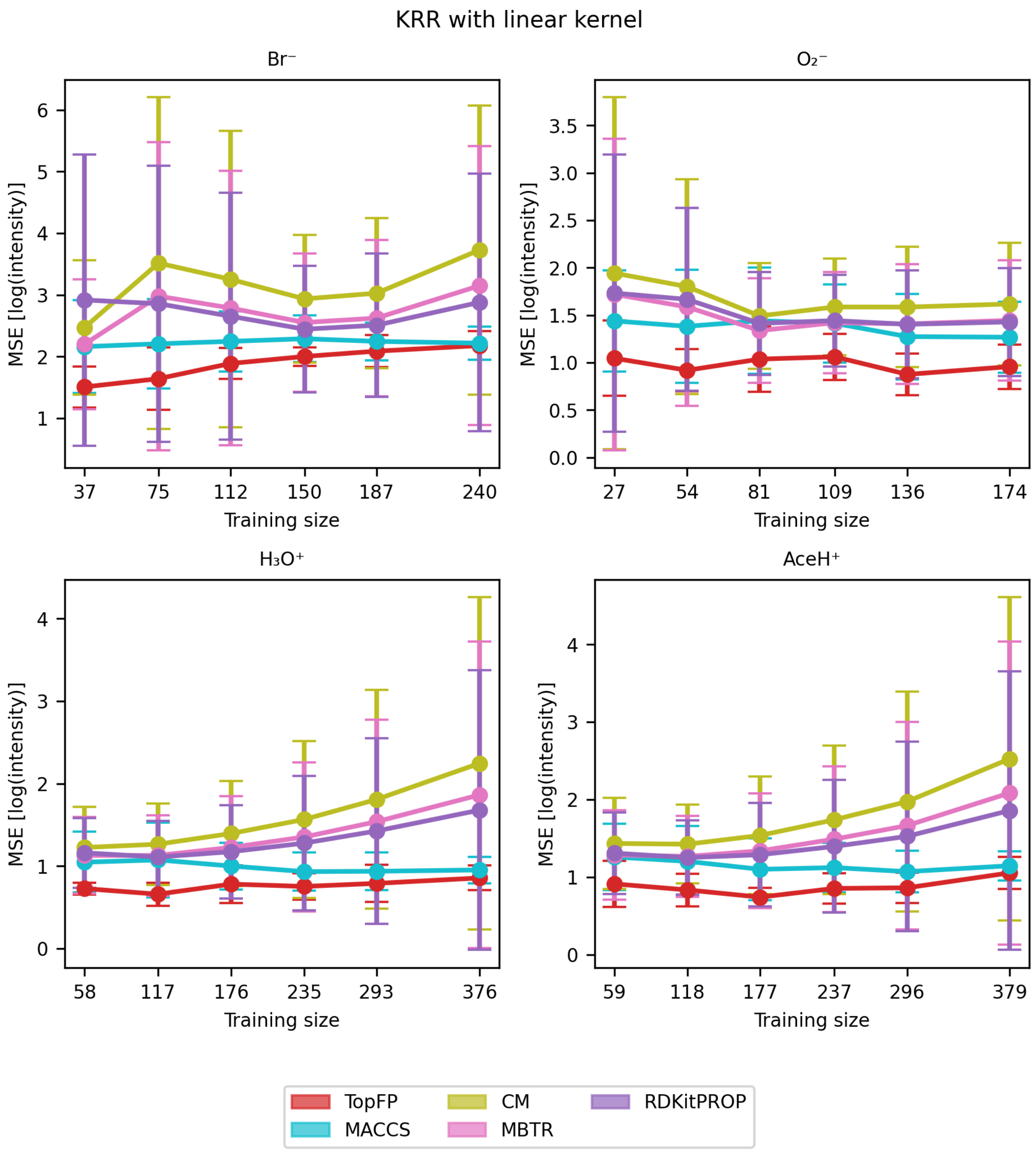


Figure S23. Linear KRR learning curve with mean squared error (MSE) of the signal intensity values in logarithmic scale of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the MSE of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

Table S30. Hyperparameters tuned for RF regressor model with PROP as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			RF regressor			
			N estimators	Max depth	Min samples leaf	Min samples split
Br^-	240	555	2000	60	2	5
		8	1500	None	1	5
		52	1000	100	1	5
		1066	500	60	1	2
		324	2000	None	1	5
$O_2{}^-$	174	555	1000	40	1	2
		8	1500	100	1	2
		52	1500	20	1	2
		1066	100	40	4	5
		324	500	40	2	2
H_3O^+	376	555	2000	60	2	2
		8	1000	None	4	5
		52	1000	60	4	10
		1066	100	80	4	2
		324	500	80	2	5
$AceH^+$	379	555	500	20	4	5
		8	1500	40	1	2
		52	2000	80	4	2
		1066	1000	20	2	2
		324	1500	40	1	2

Table S31. Hyperparameters tuned for RF regressor model with TopFP as the molecular descriptor.

			Hyperparameters						
Reagent ion	Training size	Random seed	TopFP			RF regressor			
			Fp size	Max path	N bits per hash	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	240	555	4096	7	2	100	80	4	5
		8	8192	7	16	100	40	1	2
		52	8192	10	2	100	None	1	2
		1066	4096	7	16	1000	80	1	2
		324	2048	8	2	2000	100	2	10
$O_2{}^-$	174	555	4096	7	4	1000	20	1	10
		8	4096	8	4	500	80	4	10
		52	8192	7	2	1000	40	1	5
		1066	8192	7	8	1500	20	4	5
		324	716	7	2	100	60	4	5
H_3O^+	376	555	8192	8	2	1500	60	2	2
		8	4096	7	2	100	20	1	10
		52	716	7	2	1000	None	2	2
		1066	8192	9	4	100	60	1	5
		324	4096	9	4	500	40	2	5
AceH$^+$	379	555	8192	7	2	1500	60	4	5
		8	4096	8	4	1000	80	4	2
		52	4096	7	4	2000	80	4	10
		1066	8192	7	4	100	None	2	10
		324	8192	7	4	1500	40	1	2

Table S32. Hyperparameters tuned for RF regressor model with MACCS as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			RF regressor			
			N estimators	Max depth	Min samples leaf	Min samples split
Br^-	240	555	100	60	1	2
		8	1500	None	1	2
		52	1500	20	1	2
		1066	500	None	2	2
		324	500	80	1	5
${O_2}^-$	174	555	1000	60	1	5
		8	1500	20	1	2
		52	2000	20	1	2
		1066	100	40	1	2
		324	100	None	1	2
H_3O^+	376	555	2000	20	1	2
		8	500	100	1	2
		52	500	None	1	2
		1066	1500	80	2	5
		324	1000	40	2	2
$AceH^+$	379	555	2000	None	1	2
		8	100	40	1	2
		52	2000	20	1	2
		1066	500	60	2	2
		324	1000	80	1	2

Table S33. Hyperparameters tuned for RF regressor model with CM as the molecular descriptor.

Reagent ion	Training size	Random seed	Hyperparameters			
			RF regressor			
			N estimators	Max depth	Min samples leaf	Min samples split
Br^-	240	555	1500	20	2	2
		8	1500	None	1	2
		52	2000	40	1	5
		1066	100	None	2	2
		324	100	80	1	10
$O_2{}^-$	174	555	1500	60	1	2
		8	500	40	1	10
		52	100	None	4	10
		1066	500	100	1	2
		324	500	None	1	5
H_3O^+	376	555	100	80	2	5
		8	100	20	4	5
		52	1000	80	1	5
		1066	100	60	1	5
		324	1000	None	4	10
$AceH^+$	379	555	1500	80	4	2
		8	100	60	2	5
		52	1000	80	1	2
		1066	2000	40	2	2
		324	500	40	4	2

Table S34. Hyperparameters tuned for RF regressor model with MBTR as the molecular descriptor.

			Hyperparameters							
Reagent ion	Training size	Random seed	MBTR				RF regressor			
			σ_2	w_2	σ_3	w_3	N estimators	Max depth	Min samples leaf	Min samples split
Br^-	240	555	0.1	0.4	0.1	0.8	500	None	1	5
		8	0.1	0.4	0.3	1.2	1000	20	4	5
		52	0.3	0.2	0.1	0.8	500	100	2	5
		1066	0.01	0.8	0.01	1.2	1000	40	4	2
		324	0.01	0.2	0.1	1.2	100	20	1	2
$O_2{}^-$	174	555	0.01	0.8	0.3	0.4	500	None	4	2
		8	0.1	0.4	0.3	0.4	1500	80	2	5
		52	0.01	0.4	0.1	1.4	500	40	1	2
		1066	0.1	0.2	0.1	0.8	100	100	2	5
		324	0.1	0.2	0.01	0.8	1000	40	2	10
H_3O^+	376	555	0.1	1.4	0.3	1.2	2000	80	1	5
		8	0.001	0.8	0.001	1.2	1000	40	1	2
		52	0.01	1.4	0.1	0.4	500	80	4	2
		1066	0.1	1.2	0.3	0.2	2000	40	1	5
		324	0.01	0.8	0.1	0.8	100	None	2	10
$AceH^+$	379	555	0.01	1.2	0.3	1.4	2000	80	2	2
		8	0.1	0.4	0.1	0.2	1500	None	1	10
		52	0.3	0.2	0.3	1.2	500	40	1	2
		1066	0.1	0.2	0.1	0.4	2000	60	2	5
		324	0.01	0.8	0.1	1.2	100	80	1	5

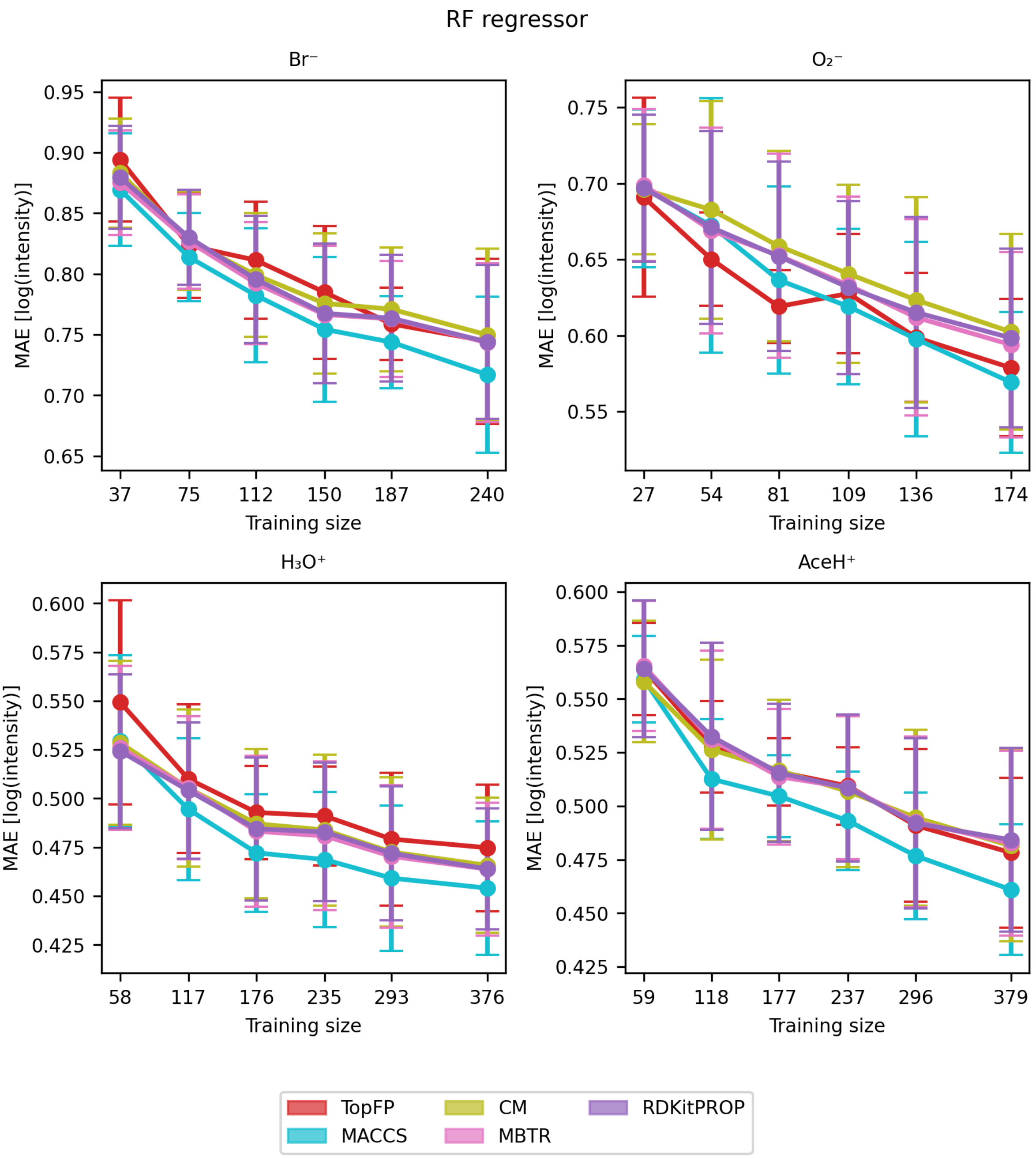


Figure S24. RF regressor learning curve with mean absolute error (MAE) of the signal intensity values in logarithmic scale of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the MAE of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

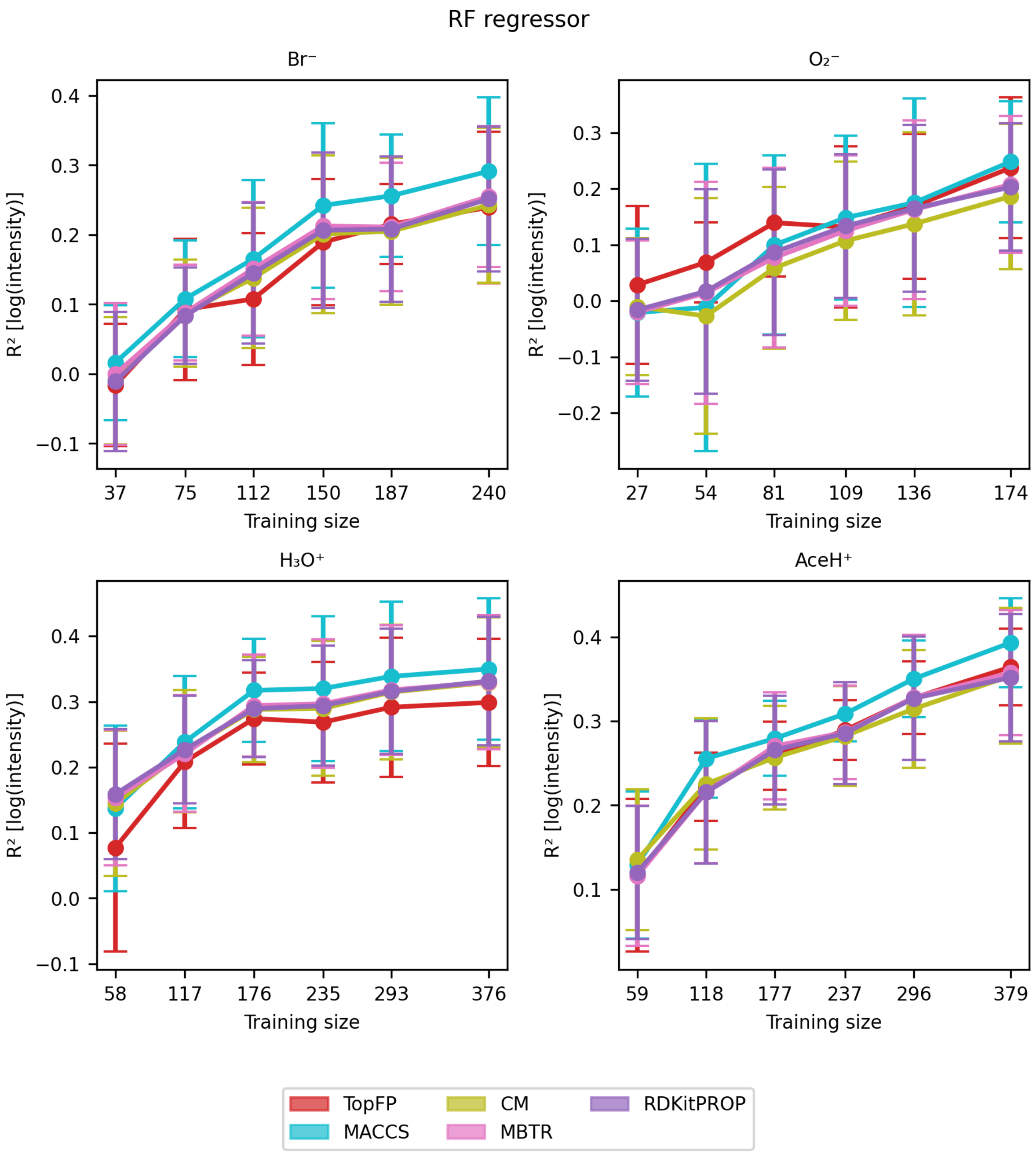


Figure S25. RF regressor learning curve with correlation coefficient (R^2) of the signal intensity values in logarithmic scale of Br^-, $O_2{}^-$, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the R^2 of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

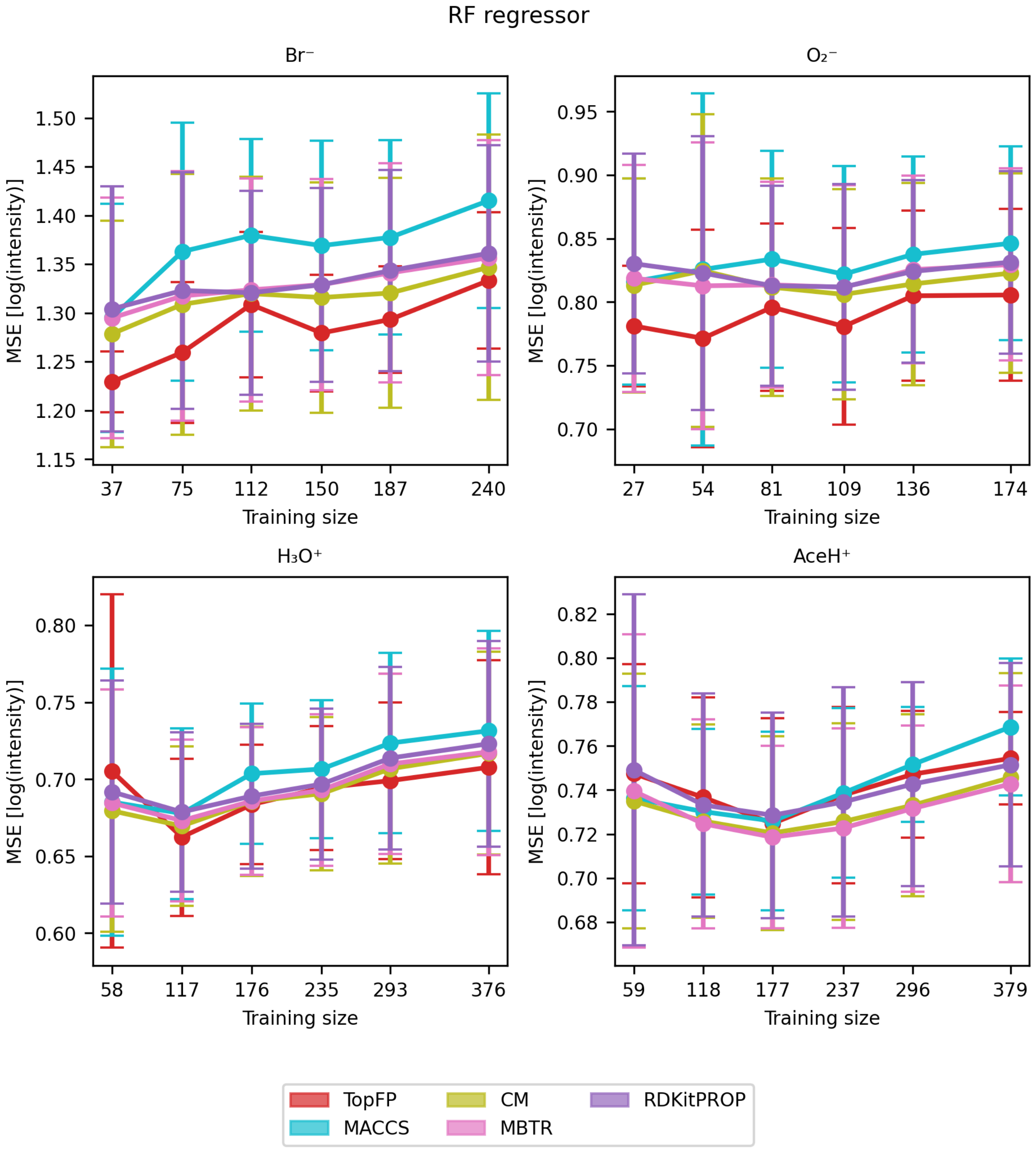


Figure S26. RF regressor learning curve with mean squared error (MSE) of the signal intensity values in logarithmic scale of Br^-, O_2^-, H_3O^+ and $AceH^+$ datasets, based on the TopFP, MACCS, CM, MBTR and properties as the descriptors. The x-axis reports the training set size, the y-axis reports the MSE of the logarithmic signal intensity. The mean value and standard deviation are obtained by repeating the training with five different random re-shuffling of the dataset.

References

Ertl, P., Rohde, B., and Selzer, P.: Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties, Journal of Medicinal Chemistry, 43, 3714–3717, https://doi.org/10.1021/jm000942e, 2000.

Hall, L. H. and Kier, L. B.: The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, chap. 9, pp. 367–422, Wiley-VCH, Inc., New York, 1991.

Kier, L. B.: An Index of Molecular Flexibility from Kappa Shape Attributes, Quantitative Structure-Activity Relationships, 8, 218–221, https://doi.org/10.1002/qsar.19890080307, 1989.

Labute, P.: A Widely Applicable Set of Descriptors, Journal of Molecular Graphics and Modelling, 18, 464–477, https://doi.org/10.1016/S1093-3263(00)00068-1, 2000.

Landrum, G.: RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org, accessed: 2024-06-04, 2006.

Partovi, F., Bortolussi, F., Mikkilä, J., and Rissanen, M.: Organic pesticide database with 716 molecules analyzed with chemical ionization mass spectrometry. Reagent ions: bromide, protonated acetone, hydronium ion, dioxide., https://doi.org/10.5281/zenodo.11208543, 2024.

Ruggeri, G. and Takahama, S.: Technical Note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization, Atmospheric Chemistry and Physics, 16, 4401–4422, https://doi.org/10.5194/acp-16-4401-2016, 2016.

Wildman, S. A. and Crippen, G. M.: Prediction of Physicochemical Parameters by Atomic Contributions, Journal of Chemical Information and Computer Sciences, 39, 868–873, https://doi.org/10.1021/ci990307l, 1999.