Supporting Information - How to actively learn chemical reactions yields in real-time using stopping criteria

Vincent Delmas,*[a] Denis Jacquemin,[ab] Aymeric Blondel,[a] Morgane Vacher,[a] and Adèle D. Laurent*[a] [1]

[a] Nantes Université, CNRS, CEISAM, UMR 6230, F-44000, Nantes, France.

[b] Institut Universitaire de France (IUF), Paris, 75005, France.

Contents

1 Source code and datasets

The source code to compute the different SP scores as well as the datasets used in this work can be found here: `https://gitlab.univ-nantes.fr/modes/public/sp_score`. The datasets can also be found in the Open Reaction Database (ORD) here: `https://open-reaction-database.org` or on their GitHub: `https://github.com/open-reaction-database/ord-data`.

2 Data computed for the S-M dataset

In Table S1 are computed all the data for the S-M dataset, using different QPI (10 and 50), stop set size (15% and 30%) and descriptors with an SP score ≥ 0.99. For the supervised ML part (column on the right of Table S1), the F1 score was calculated on all the available data with a 5-fold cross-validation.

Table S1: All data (queried labels and F1 scores) computed for the S-M dataset with the AL procedure and compared with supervised ML (5-fold cross-validated).

Model	**Stop**	**10 QPI - SP $\geq$ 0.99**		**50 QPI - SP $\geq$ 0.99**		**Supervised ML**
	set size	**# Labels**	**F1 score**	**# Labels**	**F1 score**	**F1 score**
Descriptor: PCA MFP						
Ada	15%	1942 ± 636	0.712 ± 0.014	3490 ± 378	0.705 ± 0.016	0.694 ± 0.008
	30%	2324 ± 158	0.696 ± 0.010	2910 ± 120	0.698 ± 0.008	
k-NN	15%	1129 ± 258	0.648 ± 0.015	3445 ± 172	0.673 ± 0.018	0.652 ± 0.012
	30%	1623 ± 373	0.639 ± 0.011	2930 ± 33	0.647 ± 0.011	
RF	15%	2324 ± 158	0.696 ± 0.010	2910 ± 120	0.698 ± 0.008	0.802 ± 0.007
	30%	2900 ± 0	0.817 ± 0.011	2950 ± 0	0.815 ± 0.012	
SVC	15%	867 ± 365	0.728 ± 0.027	1765 ± 153	0.747 ± 0.019	0.722 ± 0.007
	30%	1010 ± 131	0.731 ± 0.012	1670 ± 65	0.734 ± 0.011	
Descriptor: One-hot encoder						
Ada	15%	1659 ± 401	0.713 ± 0.019	3115 ± 457	0.702 ± 0.012	0.690 ± 0.016
	30%	1828 ± 199	0.701 ± 0.011	2840 ± 182	0.697 ± 0.010	
k-NN	15%	1252 ± 304	0.717 ± 0.028	3615 ± 70	0.691 ± 0.014	0.655 ± 0.011
	30%	1605 ± 259	0.726 ± 0.016	2950 ± 0	0.674 ± 0.018	
RF	15%	3300 ± 0	0.840 ± 0.012	3650 ± 0	0.839 ± 0.015	0.826 ± 0.010
	30%	2910 ± 0	0.829 ± 0.006	2950 ± 0	0.830 ± 0.007	
SVC	15%	948 ± 252	0.786 ± 0.018	2110 ± 94	0.822 ± 0.011	0.816 ± 0.009
	30%	1375 ± 225	0.805 ± 0.011	1795 ± 82	0.810 ± 0.006	
Descriptor: Reduced 1D RDKit						
Ada	15%	1915 ± 603	0.708 ± 0.010	3475 ± 242	0.705 ± 0.015	0.697 ± 0.007
	30%	2167 ± 285	0.699 ± 0.007	2940 ± 30	0.698 ± 0.009	
k-NN	15%	632 ± 205	0.680 ± 0.016	2530 ± 180	0.738 ± 0.009	0.716 ± 0.006
	30%	849 ± 172	0.680 ± 0.016	2530 ± 180	0.738 ± 0.009	
RF	15%	3650 ± 0	0.827 ± 0.011	3650 ± 0	0.828 ± 0.011	0.817 ± 0.008
	30%	2950 ± 0	0.821 ± 0.009	2950 ± 0	0.822 ± 0.011	
SVC	15%	858 ± 277	0.750 ± 0.017	2190 ± 91	0.778 ± 0.017	0.753 ± 0.012
	30%	1184 ± 146	0.762 ± 0.012	1870 ± 74	0.778 ± 0.017	

3 Reduced 2D descriptors from RDKit

All the 2D descriptors from RDkit were reduced (about 1600 descriptors) using a LinearSVC from the SKlearn python library to retain 190 and 144 features for the B-H (Table S2) and S-M dataset (Table S3), respectively.

Table S2: All the 190 RDKit 2D descriptors retained for the B-H dataset. The descriptors are presented for each reaction component.

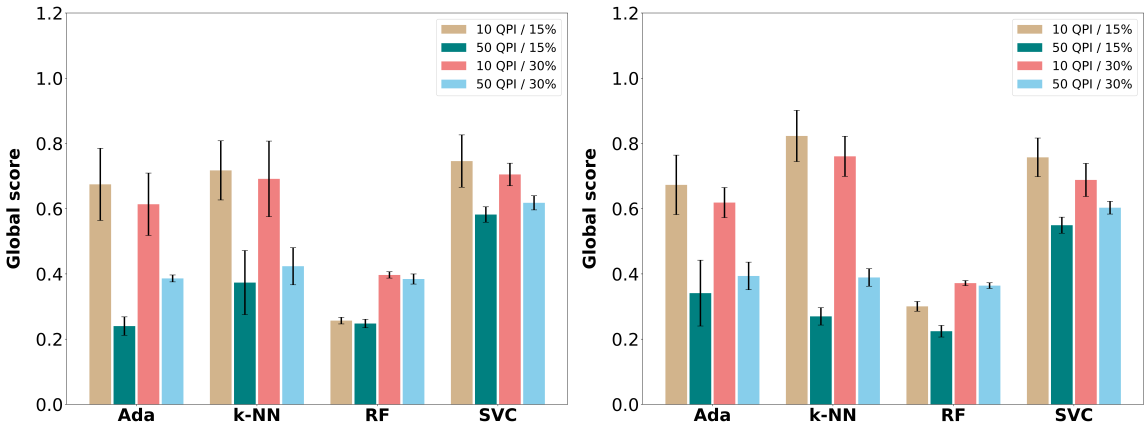
Catalyst	**Base**	**Aryl halide**		**Additive**		**Product**	
ATSC6dv	ATS4m	ATS0Z	ATS4i	ATS2dv	ATS0i	ATS2Z	ATS7v
ATSC6Z	ATS0i	ATS1Z	ATS5i	ATS3dv	ATS1i	ATS3Z	ATS8v
ATSC8Z	ATSC6m	ATS2Z	ATS6i	ATS4dv	ATS2i	ATS4Z	ATS0i
ATSC3m	ATSC7m	ATS3Z	ATS7i	ATS5dv	ATS3i	ATS8Z	ATS1i
ATSC4m	ATSC3v	ATS4Z	AATS0m	ATS4s	ATS4i	ATS0m	ATS2i
ATSC5m	ATSC4v	ATS5Z	ATSC0Z	ATS5s	ATS5i	ATS1m	ATS4i
ATSC6m	ATSC6v	ATS6Z	ATSC3Z	ATS2Z	ATS6i	ATS2m	ATS5i
ATSC7m	ATSC7v	ATS7Z	ATSC4Z	ATS3Z	ATS7i	ATS3m	ATS6i
ATSC8m		ATS0m	ATSC5Z	ATS4Z	ATS8i	ATS4m	ATS7i
ATSC3v		ATS1m	ATSC6Z	ATS5Z	AATS4m	ATS5m	ATSC0m
ATSC4v		ATS2m	ATSC7Z	ATS6Z	AATS4v	ATS6m	ATSC3m
ATSC5v		ATS3m	ATSC0m	ATS8Z	AATS3i	ATS7m	ATSC6m
ATSC6v		ATS4m	ATSC2m	ATS0m	ATSC2dv	ATS8m	ATSC8m
ATSC7v		ATS5m	ATSC3m	ATS1m	ATSC2Z	ATS0v	ATSC2v
ATSC8v		ATS6m	ATSC4m	ATS2m	ATSC0m	ATS1v	ATSC3v
MPC10		ATS7m	ATSC5m	ATS3m	ATSC2m	ATS2v	ATSC4v
TMPC10		ATS0v	ATSC6m	ATS4m	ATSC3m	ATS3v	ATSC5v
		ATS1v	ATSC7m	ATS5m	ATSC4m	ATS4v	ATSC7v
		ATS2v	ATSC0v	ATS6m	ATSC5m	ATS5v	ATSC8v
		ATS3v	ATSC2v	ATS7m	ATSC6m	ATS6v	
		ATS4v	ATSC3v	ATS8m	ATSC8m		
		ATS5v	ATSC5v	ATS0v	ATSC0v		
		ATS7v	ATSC7v	ATS1v	ATSC2v		
		ATS0i	AATSC5m	ATS2v	ATSC3v		
		ATS1i	PEOE_VSA9	ATS3v	ATSC4v		
		ATS2i	EState_VSA4	ATS4v	ATSC5v		
		ATS3i	EState_VSA8	ATS5v	ATSC6v		
				ATS6v	ATSC7v		
				ATS7v	ATSC8v		
				ATS8v	ATSC2i		
				ATS4se	ATSC4i		
				ATS5se	AATSC4m		
				ATS4pe	BertzCT		
				ATS5pe	PEOE_VSA10		
				ATS4are	PEOE_VSA11		
				ATS5are	TopoPSA		

Table S3: All the 144 RDKit 2D descriptors retained for the S-M dataset. The descriptors are presented for each reaction component.

Catalyst	**Base**	**Aryl halide**		**Additive**		**Product**	
ATS0m	ATS0m	ATS8dv	AATS5v	ATS0Z	ATS6v	ATS0Z	ATS0i
ATS0v	ATS1m	ATS0Z	AATS6v	ATS2Z	ATS0i	ATS1Z	ATS1i
ATS1v	ATS2m	ATS1Z	ATSC4Z	ATS3Z	ATS1i	ATS2Z	ATS2i
ATS2v	ATS1v	ATS2Z	ATSC1m	ATS4Z	ATS2i	ATS3Z	ATS3i
ATS3v	ATS3v	ATS3Z	ATSC2m	ATS5Z	ATS3i	ATS4Z	ATS4i
ATS4v	ATS4v	ATS4Z	ATSC3m	ATS6Z	ATS4i	ATS5Z	ATS6i
ATS0i	ATS0i	ATS5Z	ATSC4m	ATS0m	ATS5i	ATS0m	ATSC0m
ATS3i	ATS1i	ATS7Z	ATSC5m	ATS1m	ATS6i	ATS1m	ATSC3m
ATS4i	ATS3i	ATS8Z	ATSC6m	ATS2m	ATS7i	ATS2m	ATSC4m
ATSC2m	ATS4i	ATS0m	ATSC7m	ATS3m	ATSC0Z	ATS3m	ATSC5m
ATSC3m	ATS5i	ATS1m	ATSC8m	ATS4m	ATSC3Z	ATS4m	ATSC7m
ATSC2v	AATS0m	ATS2m	ATSC1v	ATS5m	ATSC0m	ATS5m	ATSC0v
	ATSC2v	ATS3m	ATSC2v	ATS6m	ATSC2m	ATS8m	ATSC5v
	ATSC3v	ATS4m	ATSC3v	ATS0v	ATSC3m	ATS0v	ATSC7v
		ATS5m	ATSC4v	ATS1v	ATSC4m	ATS1v	ATSC8v
		ATS7m	ATSC5v	ATS2v	ATSC5m	ATS0p	ATSC0p
		ATS8m	ATSC6v	ATS3v	ATSC7m		
		ATS8v	ATSC7v	ATS4v	ATSC0v		
		ATS6pe	ATSC8v	ATS5v	AATSC7m		
		ATS7pe	BertzCT				
		ATS6are	fragCpx				
		ATS7are	MPC10				
		ATS8are	TMPC10				
		AATS4v					

4 Global Scores for One-hot encoder and Reduced RDKit 2D descriptors

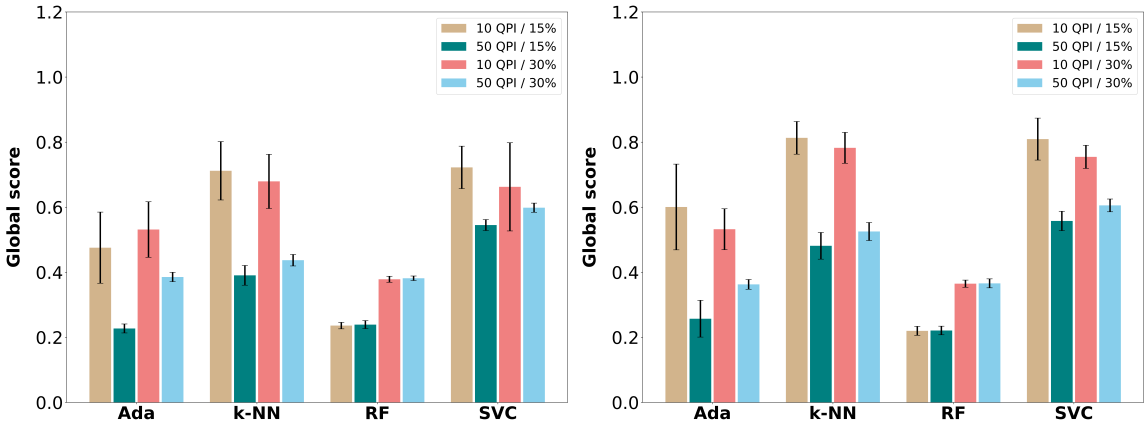
In Figure S1 and Figure S2 are shown the Global Scores (GS) for each dataset (B-H and S-M) and for the One-hot encoder and the reduced RDkit 2D descriptors, respectively.



(a) B-H dataset

(b) S-M dataset

Figure S1: Global score (GS) for each model with each combination of QPI and stop set size using the one-hot encoder descriptor for the B-H (left) and S-M (right) dataset, respectively. The bar shows the standard deviations of 10 runs.



(a) B-H dataset

(b) S-M dataset

Figure S2: Global score (GS) for each model with each combination of QPI and stop set size using the reduced RDKit 2D descriptors for the B-H (left) and S-M (right) dataset, respectively. The bar shows the standard deviations of 10 runs.

5 Simplified AL loop algorithm

In Figure S3 is represented a simplified pseudo-algorithm to perform Active Learning (AL).

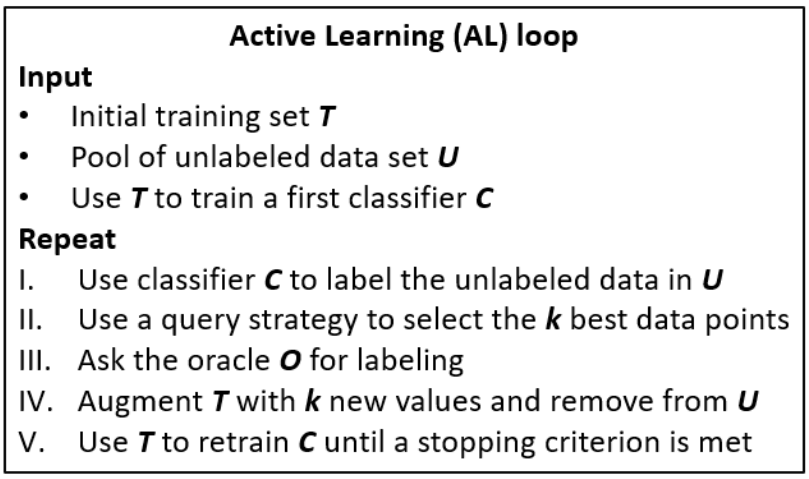


Figure S3: Simplified algorithm of the AL loop