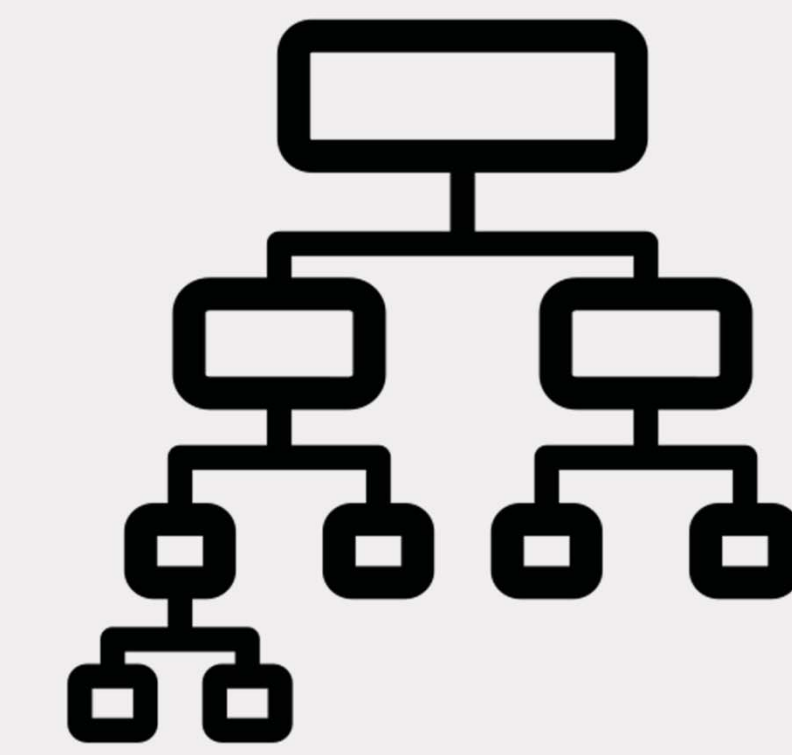


Prédiction de temps de rétention grâce à l'apprentissage automatique

N. Grenier ; P.A. Segura
Université de Sherbrooke



Introduction

- L'identification de composés inconnus dans les échantillons environnementaux nécessite l'utilisation d'étalons purs, ce qui peut être impossible ou irréaliste.
- L'apprentissage automatique (*machine learning*) est un outil d'intérêt pouvant permettre l'identification des pics chromatographiques sans utilisation de standard.

Méthode

Constituer un ensemble de données (training set)

- Une molécule peut être décrite par un ensemble de descripteurs moléculaires (*DM*) à partir de sa notation SMILES (Figure 1).
- Les descripteurs moléculaires sont obtenus grâce à la suite PaDELPy.
- Le temps de rétention (TR), étant le résultat des interactions entre la phase mobile, la phase stationnaire et l'analyte, est une fonction de ces *DM*.
- Le *training set* permettra d'établir la relation entre le TR et les descripteurs moléculaires :

$$TR = f(DM_1, DM_2, DM_3, DM_4 \dots DM_n)$$

Entraîner un modèle approprié

- Méthode de régression non linéaire : Forêt d'arbres décisionnels (*random forest*, *RF*).
- Évite le sur-ajustement comparativement à d'autres méthodes.
- Seuls les DM pertinents sont retenus (figure 2).

Évaluer la performance du modèle (test set)

- Mesures expérimentales par LC-QTOF

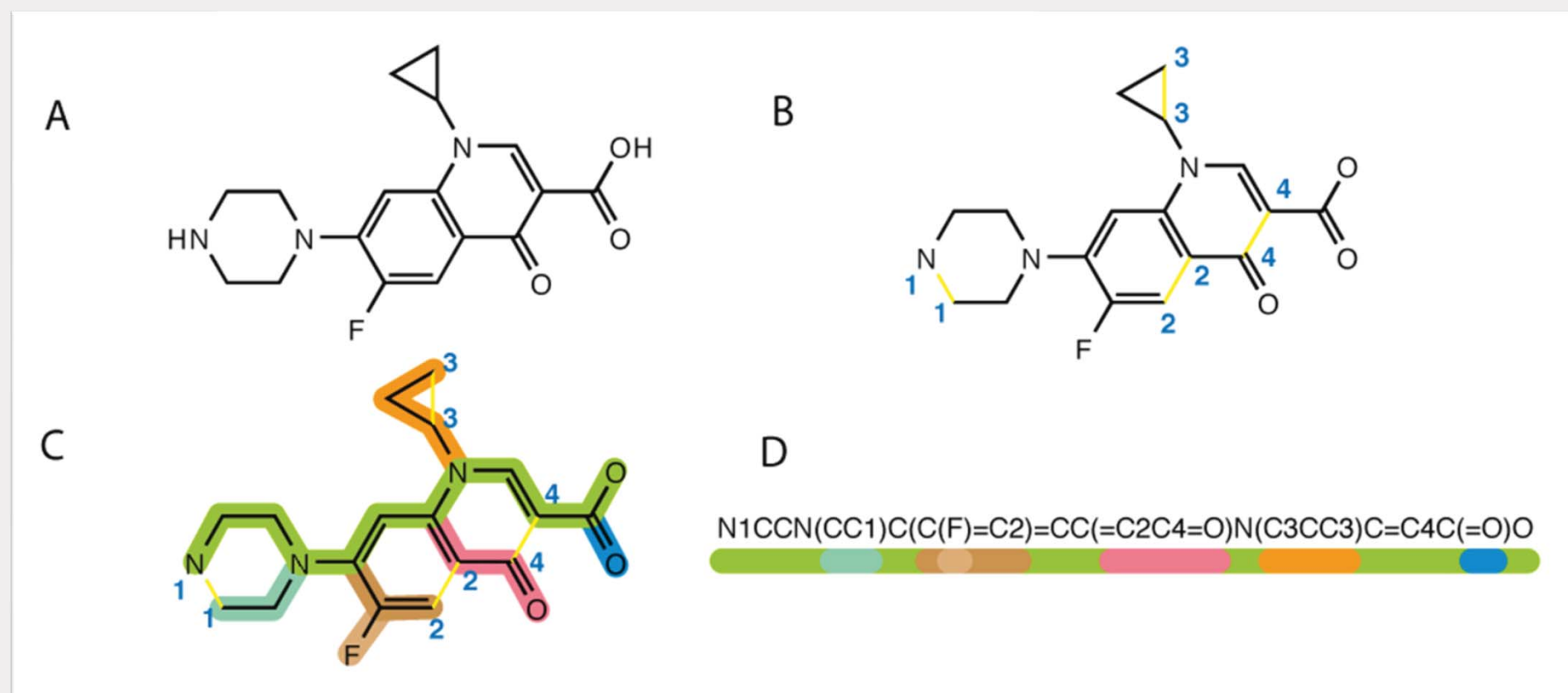
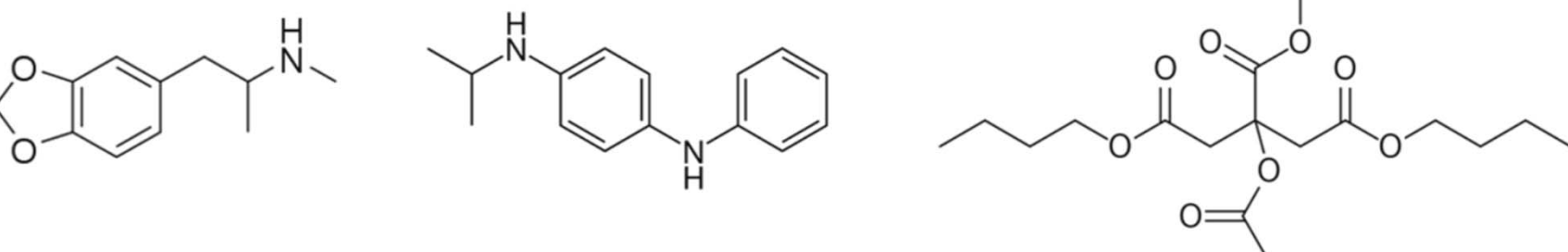


Figure 1. Exemple de notation SMILES [1] pour l'antibiotique ciprofloxacine.

Résultats

Molécule	TR exp. (min)	Prédiction RF	ACD	Err. Abs. RF	Err. Abs ACD
Irbesartan	11,6	15,38	13,61	3,8	2,0
Valsartan	16,1	14,88	14,26	1,2	1,8
Lauryl diethanolamide	19,1	11,78	15,17	7,3	3,9
Losartan	20,1	14,79	13,02	5,3	7,1
MDMA	12,3	6,58	4,37	5,7	7,9
Omeprazole	8,8	12,52	7,47	3,7	1,3
ATBC (Acetyl tributyl citrate)	21,2	18,54	17,68	2,7	3,5
Triton x-100	12,3	15,51	16,34	3,2	4,0
IPPD	12,2	12,38	8,61	0,2	3,6
Atenolol	12,2	5,82	1,50	6,4	10,7
Cocaine	7,1	9,67	9,09	2,6	2,0
Methadone	19,1	16,82	10,63	2,3	8,5
Sitagliptin	11,8	7,56	6,89	4,2	4,9
moyenne				3,7	4,7

Figure 2. Comparaison entre la performance de ACD/Chromgenius (méthode linéaire) et RF pour une même base de données.



Temps expérimental	12,3	12,2	21,2
Temps calculé (ACD)	4,4	8,6	17,7
Temps calculé (RF)	6,6	12,4	18,5

Figure 3. Exemple de TR prédits à partir de données de structure.

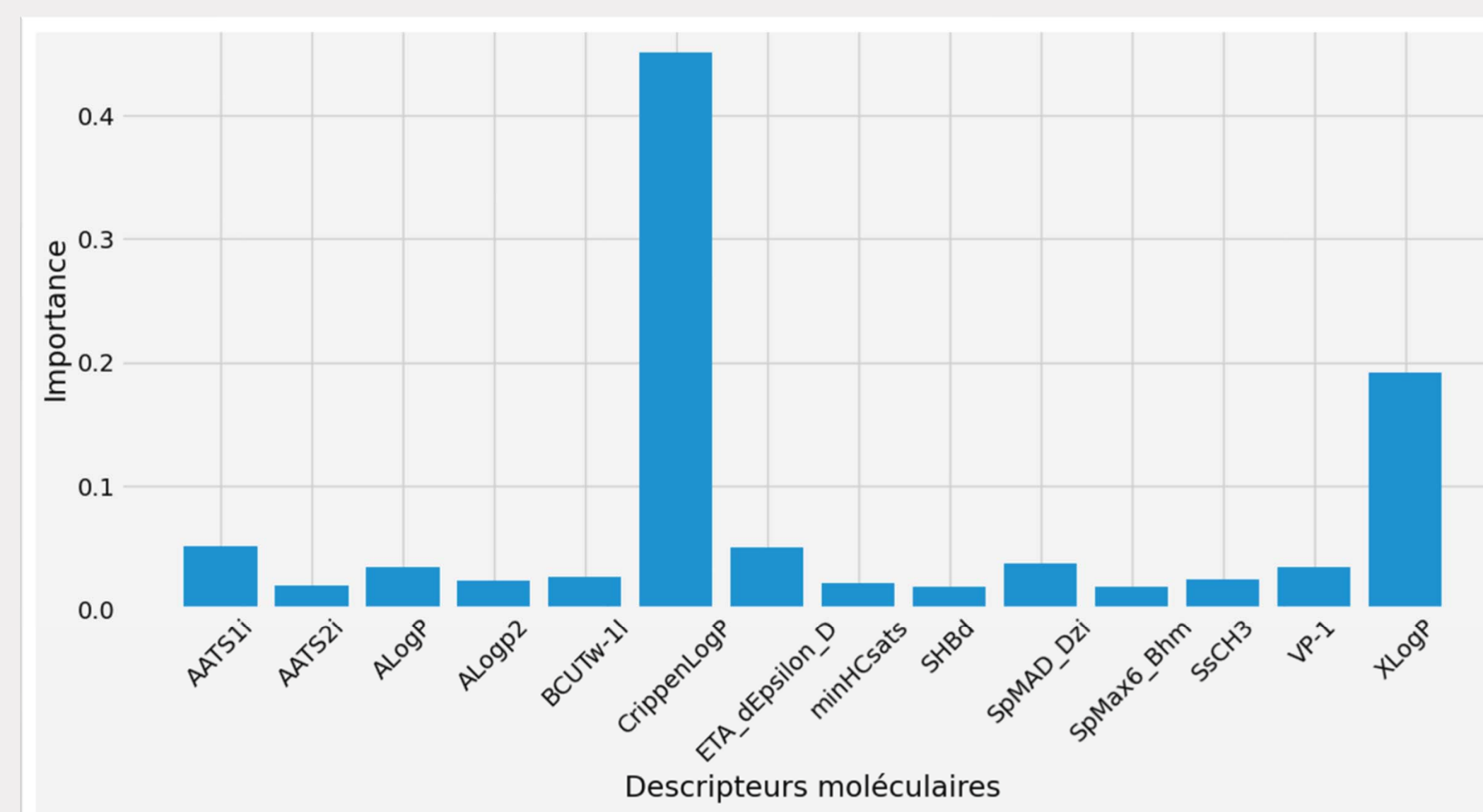


Figure 4. 14 descripteurs moléculaires desquels dépend le plus le TR.

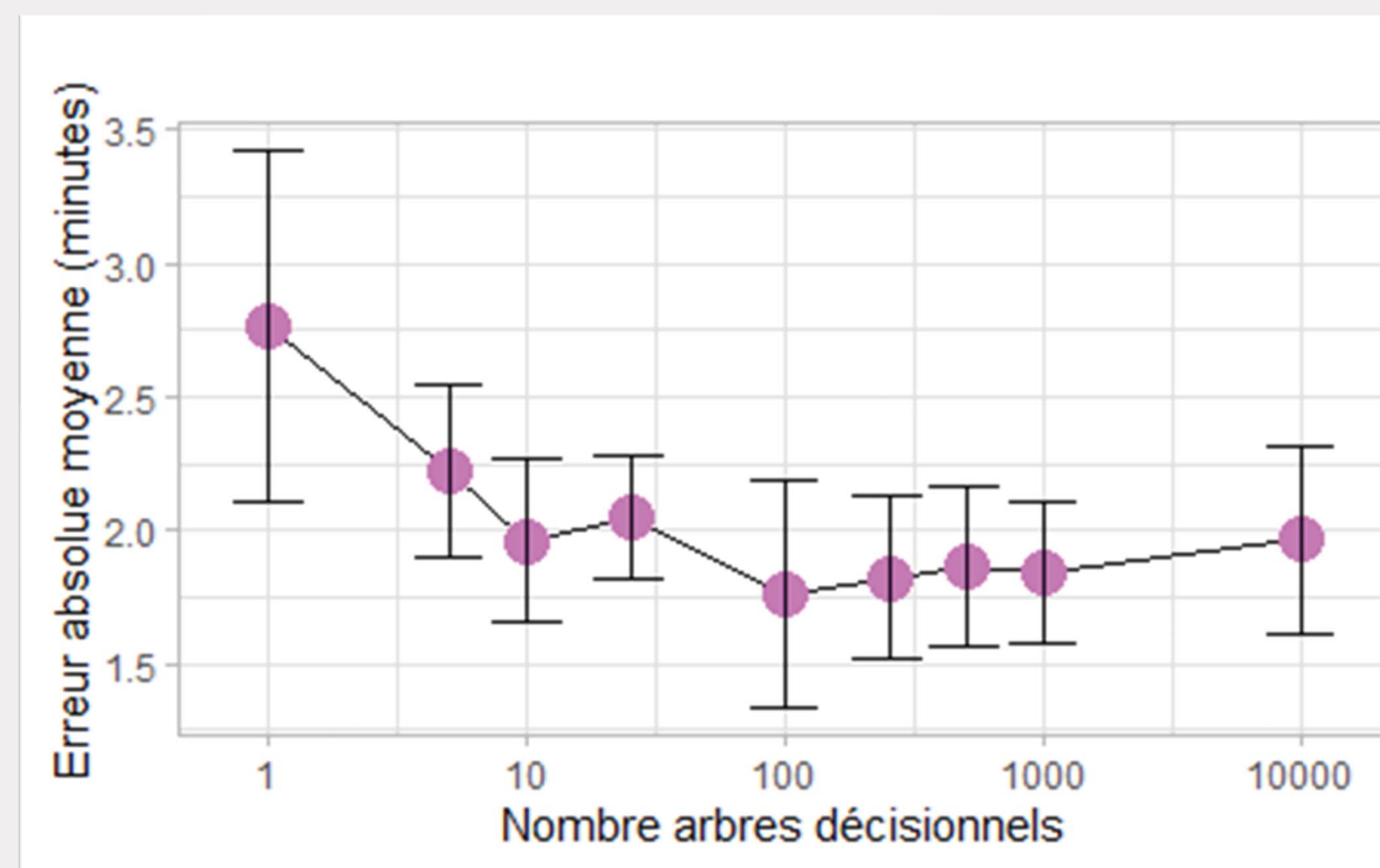


Figure 5. Influence du nombre d'arbres sur la précision de la prédiction.

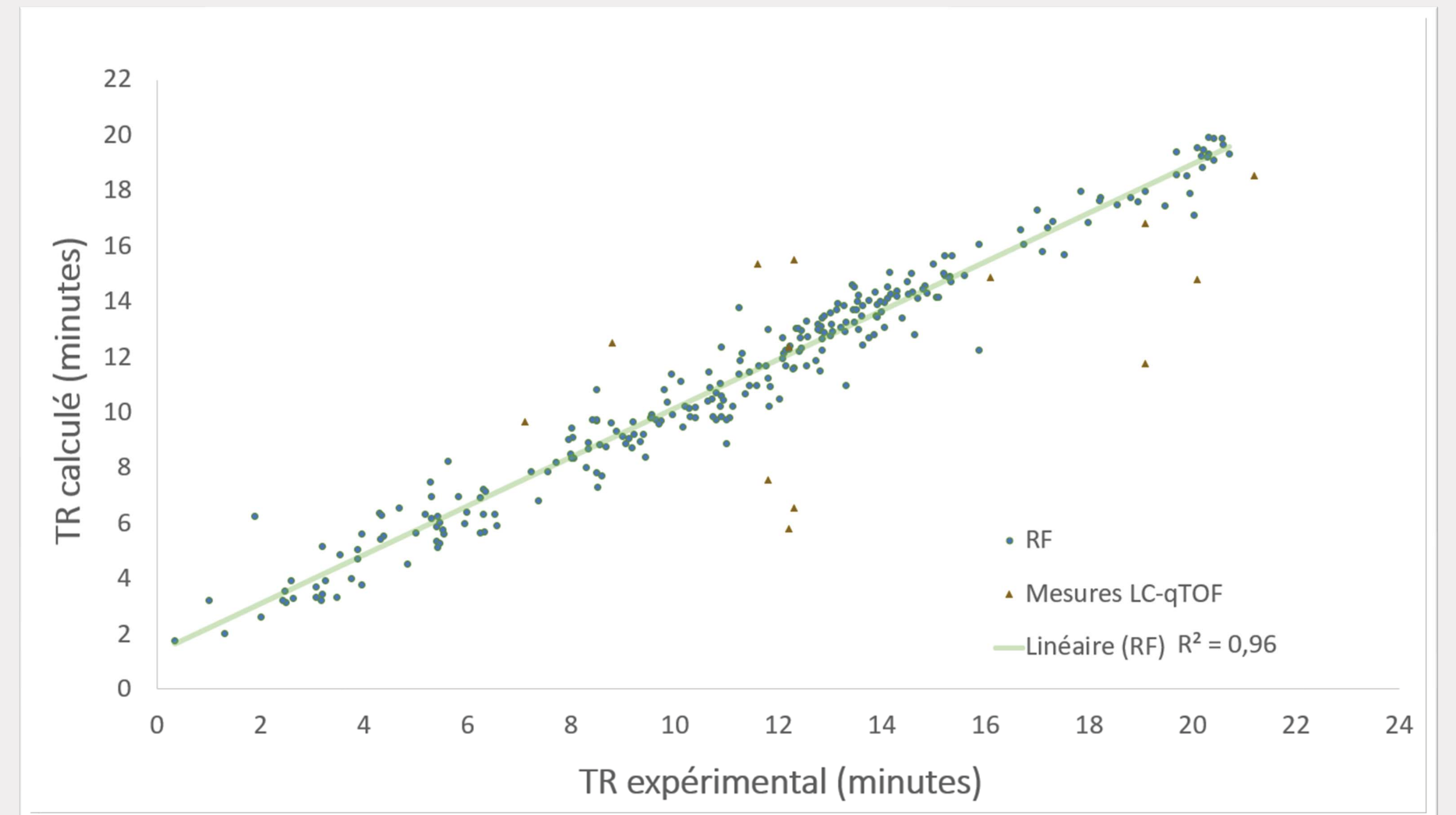


Figure 6. Comparaison des TR obtenus expérimentalement et prédits par RF.

Conclusions

- Meilleure performance que le logiciel ACD/Chromgenius pour une même base de données (*training set*).
- L'acuité des prédictions dépend de la richesse de la base de données.
- Avec une base de données appropriée, il est possible d'appliquer cette approche pour prédire plusieurs autres propriétés d'intérêt environnemental (dose létale médiane, concentration effective médiane, etc.).
- Méthode **verte** très accessible

Remerciement

Je remercie Mégane Moreau pour son soutien.

Références

Palmer, D. S., O'Boyle, N. M., Glen, R. C. & Mitchell, J. B. O. (2006, 2 décembre). Random Forest Models To Predict Aqueous Solubility. *Journal of Chemical Information and Modelling*, 47(1), 150-158. <https://doi.org/10.1021/ci060164k>

Randazzo, G. M., Bileck, A., Danani, A., Vogt, B. & Groessl, M. (2020, février). Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry. *Journal of Chromatography A*, 1612, 460661. <https://doi.org/10.1016/j.chroma.2019.460661>

Debus, B., Parastar, H., Harrington, P. & Kirsanov, D. (2021, décembre). Deep learning in analytical chemistry. *TrAC Trends in Analytical Chemistry*, 145, 116459. <https://doi.org/10.1016/j.trac.2021.116459>

[1] Wikipedia contributors. (2022, 6 novembre). *Simplified molecular-input line-entry system*. Wikipedia. https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

