Modèles d'apprentissage automatique basés sur des descripteurs moléculaires pour prédire des facteurs de caractérisation toxicologique et écotoxicologique

Rémi Servien, Eric Latrille, Dominique Patureau, Arnaud Hélias

▶ **To cite this version:**

Rémi Servien, Eric Latrille, Dominique Patureau, Arnaud Hélias. Modèles d'apprentissage automatique basés sur des descripteurs moléculaires pour prédire des facteurs de caractérisation toxicologique et écotoxicologique. Congrès Management du Cycle de Vie, Nov 2024, Lille, France. ⟨hal-04782159⟩

HAL Id: hal-04782159

https://hal.science/hal-04782159v1

Submitted on 14 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Modèles d'apprentissage automatique basés sur des descripteurs moléculaires pour prédire des facteurs de caractérisation toxicologique et écotoxicologique

Rémi Servien[1,2], Eric Latrille[1,2], Dominique Patureau[1], Arnaud Hélias[3,4]

[1]INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France
[2]ChemHouse Research Group, Montpellier, France Univ Montpellier,
[3]ITAP, INRAE, Institut Agro, Montpellier, France
[4]ELSA, Research group for environmental LCSA and ELSA-Pact industrial chair, Montpellier, France
E-mail contact: remi.servien@inrae.fr

Des données (éco)-toxicologiques robustes et rapides à obtenir sont nécessaires pour prendre des décisions éclairées sur la manière de réglementer de nouveaux produits chimiques. Ces données doivent également être couplées avec des données d'exposition environnementale, afin de mieux comprendre l'impact sur l'environnement. Pour la toxicité humaine et l'écotoxicité des eaux douces, USEtox® [1] a été développé pour produire un modèle de caractérisation transparent et consensuel. Malheureusement, pour déterminer les facteurs de caractérisation (CF) d'une molécule, de nombreux paramètres physico-chimiques (tels que la solubilité, la dégradabilité …) ainsi que des données (éco)toxicologiques détaillées doivent être fournies ce qui est coûteux et chronophage. Cela reste un problème pour un certain nombre de composés chimiques existants, ce qui laisse leurs impacts potentiels largement inconnus. Des modèles sont donc nécessaires pour compléter les approches expérimentales, afin de réduire les coûts expérimentaux et de prioriser les composés chimiques. De tels modèles existent déjà, comme les modèles QSAR pour prédire des données écotoxicologiques ou des algorithmes d'apprentissage automatique pour prédire des paramètres environnementaux sur la base de différentes variables, notamment physico-chimiques [2]. Cependant, les variables d'entrée de ces modèles ne sont pas uniquement des descripteurs moléculaires (DM) pouvant être facilement collectés et les variables de sortie ne sont pas directement les CF. Nous proposons donc de tester différentes méthodes (linéaires et non linéaires) pour construire des modèles robustes qui pourraient prédire directement les CF manquants (CFET écotoxicologiques et CFHT toxicologie humaine) dans les eaux douces continentales, sur la base de DM faciles à obtenir.

Les différents modèles ont été construits et testés en utilisant la base de données USEtox® 2.12. La base de données TyPol [3] a été utilisée pour collecter 40 DM sur les différents composés. Ces DM faciles à obtenir étaient constitutionnels (nombre d'atomes...), géométriques (surface moléculaire de Conolly), topologiques (indices de connectivité...) ou quantiques (polarisabilité...). Un total de 274 composés étaient communs entre TyPol et USEtox®. Ils ont été utilisés pour construire les modèles et évaluer leurs performances. Nous avons comparé une méthode linéaire (PLS) et deux méthodes d'apprentissage automatique non linéaires : les forêts aléatoires (RF) et les machines à vecteurs de support (SVM) [4]. Nous avons également testé des approches de classification-puis-prédiction : une première classification a été effectuée sur l'ensemble de la base de données TyPol, puis 3 modèles (PLS, RF et SVM) ont été optimisés sur chaque cluster. Nous avons testé 6 méthodes concurrentes par cluster : 3 globales et 3 basées sur la classification-puis-prédiction. Les différents modèles ont été testés en utilisant une approche d'entraînement et de test classique en calculant les erreurs absolues de la prédiction sur l'ensemble de test.

Les méthodes d'apprentissage automatique basées sur la classification-puis-prédiction ont montré les meilleures performances pour les CFET, soulignant le besoin de modèles locaux non linéaires. Les médianes des erreurs absolues étaient inférieures à un logarithme, ce qui peut être considéré comme une marge d'erreur acceptable. Pour les CFHT, les mêmes bonnes performances ont été observées, principalement pour les méthodes d'apprentissage automatique globales. À noter que, pour les deux CF, les méthodes linéaires basées sur la PLS sont surpassées par les méthodes non linéaires. Les résultats détaillés peuvent être consultés dans [4].

En conclusion, ces résultats sont prometteurs car les performances des modèles prédictifs sont en dessous du niveau d'incertitude communément admis pour ces CF (1 log), et ils sont basés sur des DM faciles à obtenir. Cela permet de calculer des valeurs de CF de manière rapide et simple, ce qui peut être utilisé tant que les CF conventionnels ne sont pas disponibles. Ces modèles prédictifs ont ensuite été utilisés pour compléter l'évaluation de l'impact potentiel des micropolluants provenant des stations d'épuration des eaux

usées [5], où un manque de CF avait été observé [6]. Il convient de noter que cette stratégie de modélisation pourrait être appliquée à tout autre compartiment et/ou CF, à condition qu'une base de données d'apprentissage suffisamment grande existe.

Références

[1] USEtox® 2020. USEtox® database system, https://USEtox®.org/model/download.
[2] Hou et al. 2020. Environ Int. https://doi.org/10.1016/j.envint.2019.105393
[3] Servien et al. 2014. Chem. https://doi.org/10.1016/j.chemosphere.2014.05.020
[4] Servien et al. 2022. Peer Com. J. https://doi.org/10.24072/pcjournal.90.
[5] Servien et al. 2022. CSCEE. https://doi.org/10.1016/j.cscee.2021.100172.
[6] Aemig et al. 2021. Water Res. https://doi.org/10.1016/j.watres.2020.116524.