Current Opportunities and Limitations in Predicting Micropollutant Removal in Wastewater Treatment based on Molecular Structure

J.A. Cordero[1], J. Hafner[2], M. S. McLachlan[3], H. Singer, K. Fenner[1,2]

[1]Department of Environmental Chemistry, Swiss Federal Institute of Aquatic Science and Technology (Eawag),

[2]Department of Chemistry, University of Zürich

[3]Departmenf of Environmental Science, University of Stockholm

1 Introduction

Continuous release of harmful substances from WWTPs remains a global issue, emphasizing the urgent need for strategies to mitigate their impact on human health and the environment. Besides advanced treatment, these strategies involve alternatives assessment and safe-by-design initiatives.[1–3] Both strategies would highly benefit from models to predict removals of individual chemicals in WWTPs, but sufficiently accurate models to do so are missing.

Examples of state-of-the-art tools for predicting removals in WWTPs are the STP model in EPI Suite (i.e., STPWIN[4]) and SimpleTreat[5], both widely used in risk assessment and regulatory settings.[6] These models have a strong mechanistic foundation and effectively describe the various processes that influence the fate of chemicals within treatment plants.[7–10] However, their performance relies heavily on the accuracy with which the properties that govern these processes can be described. In particular, accurate predictions for polar substances depend on a sound knowledge of biodegradation rate constants but high-quality experimental biodegradation data are lacking for most chemicals in commerce, and would, in any case, not be available in a safe-by-design context.[11,12]

Accurately determining biodegradation rate constants is of great interest not only in the context of wastewater treatment but also because biodegradation is the the main transformation process reducing exposure to chemicals in different environmental compartments, including aerobic soils and water-sediment interfaces.[13,14] As a result, experimental procedures to assess microbial degradation in these environments have been extensively developed and are frequently used in risk assessment.[13,15–17] However, these experiments are costly and the number of substances to be evaluated is very large, which is why in silico approaches are increasingly promoted as an alternative.[18] The preferred approach is to build quantitative structure-activity relationship (QSAR) models based solely on features that can be calculated from chemical structures, rather than including properties that require physical measurement, thus enabling the evaluation of millions of existing and potentially new chemicals.[12,19,20] The currently most used QSAR models for biodegradation is the suite of BIOWIN models, which is provided through EPI Suite and also used to internally estimate biodegradation in STPWIN.

BIOWIN is a collection of models trained on small datasets and biodegradation ratings derived from expert judgment.[21] Many recent studies continue using the expert-judgment datasets to develop predictive models. [22,23] which limits their generalizability. Therefore, we believe that new, and more accurate models should be based on experimental data rather than expert judgment, a need that other researchers have also recognized. Notably, Wang et al.[24] developed linear regression-based QSAR models based on rate constants derived from batch experiments with activated sludge from a WWTP. Their models provided new insights into biodegradation mechanisms under both aerobic and anaerobic conditions, but they were validated with only 10 chemicals. Differently, Nolte et al.[25] built QSAR models but instead of data from batch experiments they used monitoring data from full-scale WWTPs to derive biodegradation rate constants, and built models for predicting removals of 69 compounds (51 train & 18 test). While these efforts using experimental data are valuable contributions, they are limited by the small number of substances covered. Later, Chirico et al.[26] , profiting from advances in high-resolution mass spectrometry (HRMS), were able to collect data for over 300 compounds, enabling the development of models for nearly 100 compounds (70 for training and 28 for testing). The authors acknowledged the challenge of creating a global model due to the heterogeneity of the data and the possibility that these relationships are non-linear.

Given the complex relationship between chemical structure, microbial degradation, and other processes occurring in wastewater treatment, machine learning (ML) has emerged as a promising alternative for predictive modeling.[20,27] ML methods help identify patterns between a large number of features and the target variable, making them well-suited for use within quantitative structure-activity relationship (QSAR) modeling frameworks.[20] Furthermore, many ML algorithms are designed to prevent overfitting, even when dealing with high-dimensional data, which is a common obstacle in traditional QSAR models that rely on multiple linear regression.[28] While there are numerous examples of ML-assisted QSAR models for various endpoints, mostly related to toxicity, biodegradation models are less common. A prominent example by Zhang et al.[29] demonstrated high performance (i.e., 85.1% accuracy) using machine learning to classify compounds as readily- or not readily-biodegradable. However, this model is not suited for more refined persistence and exposire assesment because it does not provide a continuous metric such as percent removal.

Building on the unprecedented opportunities provided by advances in HRMS and the potential of ML, we recognize a unique opportunity to address the limitations of previous studies by using ML and monitoring data from WWTPs to develop general predictive models for removal of chemicals in WWTPs. Our goal is to use to establish a robust benchmark for future modeling advancements. To this end, we explore a wide variety of algorithms and molecular representations to illustrate the applicability and current limits of ML for this task. We do this in a fully transparent way by developing an open-source library and providing a carefully curated database for others to use and explore alternative modeling approaches.

2 Methods

2.1 Description of available data

The data used in this study consist of information on 1153 unique chemical substances monitored in 44 WWTPs across Australia, Sweden, and Switzerland; all these plants employ conventional treatment with activated sludge. These data were compiled from four independent datasets (i.e., AMAR, AUS, SNF and SWE2), which do not cover the exact same chemical substances; that is, out of the 1153, 751 substances are unique to one of the datasets and 402 are found in two or more datasets. Further details about the sources and experimental procedures of each dataset are explained in the Supplemental Information (*SI) Section S1*. The datasets also vary in their chemical identification methods and hence certainty in structural annotation: specifically, SNF uses reference standards (level 1 confidence as described Schymanski et al.[30]), whereas for the other datasets structural annotation is done by library spectrum match (level 2 confidence[30]). These differences did not impact the model performance as further discussed in the SI Section 1. Moreover, a table summarizing key information on these datasets is provided in the SI document *WWTP_descriptions.xlsx*, and the complete database is available in the ERIC open repository (EAWAG Research Data Institutional Collection: https://opendata.eawag.ch/) and as part of the Renku project associated with this publication (renkulab.io/projects/fenner-labs/projects/pepper).

Our target variable for modeling is breakthrough, which is defined as the ratio of the concentration (C) detected in the effluent to the concentration detected in the influent for each substance in each of the WWTPs (eq. 1); thus, breakthrough may be interpreted as the fraction of each substance that is not removed during treatment.

$$Breakthrough\ (B) = \frac{\text{C (Effluent)}}{\text{C (Influent)}} \qquad \text{Eq. 1}$$

2.2 Model development

A computational workflow, PEPPER (pepper-lab · PyPI), was developed as an open-source library to build all models in this study. One of the benefits of PEPPER is that all data processing methods and models developed are documented in detail so users can reproduce our work entirely. Further details of PEPPER are provided in the SI section *S2*.

2.3 Curation of the database

We employed different levels of data curation and investigated their impact on model performance. The lowest level of curation consisted in careful treatment of duplicates, preprocessing of chemical structures and excluding substances with concentrations in influent below the limits of detection (even if they were found in effluent samples); further details in the SI Section S3.

The role of WWTP technology: WWTPs used in this study all employ conventional activated sludge processes. Of the 44 plants, 40 include a nitrifying-denitrifying step, which we refer to as nitrogen (N)-eliminating plants. The nitrifying-denitrifying step is known to significantly impact the removal of certain substances[7], raising concerns about how combining data from these two types of plants could affect model performance. We observed clear differences and concluded that restricting our model to only data from N-eliminating plants results in a more homogeneous training set. A deeper discussion is presented in the SI section S4.

Additional curation: We added five additional curation criteria that could influence prediction quality and tested them systematically. These criteria are: (I) exclude substances for which breakthrough values could be calculated for less than three WWTPs, because there is less confidence in whether these values are representative of a wider range of treatment plants. (II) Exclude substances with breakthrough values exceeding 120%, because these values could be the result of analytical errors or formation during treatment (i.e., transformation products). (III) Exclude substances with high variability in breakthrough values across plants, because we assume it is more challenging to establish a structure-activity relationship as breakthrough seems to be widely affected by subtle changes in treatment conditions. (IV) Exclude entries with effluent values below the limit of quantification (LOQ). (V) Exclude highly sorbing or highly volatile substances because we expect that the majority of substances in this study are removed via biodegradation so substances mainly removed by other mechanisms could introduce conflicting information in the models. For the latter purpose, we calculated organic carbon-water partition coefficients (K_{OC}) and Henry's law constants (H) of all substances using OPERA 2.9 (github.com/kmansouri/OPERA) and selected thresholds of 4000 L/kg and 10^{-5} atm-m^3/mol, respectively.

2.4 Descriptors

We calculated several molecular representations: as molecular descriptors: **PaDEL**[31] and **Mordred**[32], as fingerprints: **MACCS**, Extended Connectivity fingerprints (**ECFP**; using RDKit[33]) and **RDKit** Fingerprints. Additionally, we created a fingerprint (**ePFP**) by one-hot encoding to represent functional groups that trigger a biotransformation rule according to enviPath (i.e., a prediction system for microbial transformations.[34] Further details about the descriptors are provided in the SI Section S5.

As regressors we tested five linear models (Multiple Linear Regression (**MLR**), **Ridge** Regressor, Kernel Ridge Regressor (**KR**), Stochastic Gradient Descent Regressor (**SGD**), and Linear Support Vector Regressor (**LSVR**)), two ensemble regressors (Random Forest (**RF**) and AdaBoost (**AB**)), a Decision Tree Regressor (**DT**), a Multilayer Perceptron (**MLP**; a type of neural network), a Support Vector Regressor with a radial basis function kernel (**SVR**), and a K-Nearest Neighbors Regressor (**KNN**). These algorithms cover a wide range of robust linear and non-linear methods, frequently used in QSAR modeling.

All regressors were evaluated using 5-fold nested cross validation (CV) as explained in the SI Section S6. This workflow ensures three key aspects: i) the model never sees the test set during optimization, ii) optimization is validated over a wide range of molecules to prevent overfitting, iii) the performance of the model is not determined using a single test set but instead 5 different sets that cover the whole database. Model performance was assessed using the average coefficient of determination (R^2) and the average root mean squared error (RMSE). Statistical analysis to investigate differences in performance among models was performed using Pingouin,[35] a python library for statistical analyses.

3 Results and discussion

3.1 Systematic evaluation of each dataset

To understand the quality and contribution of the different data sets, we analyzed them in terms of measured breakthrough values and the chemical space covered. In Figure Figure 1.b, the box for each dataset contains all measurements from different wastewater treatment plants (WWTPs) for chemical substances unique to that dataset and *Multiple* refers to substances in at least two datasets. Notably, the SNF dataset contributed several substances that frequently escaped treatment across different WWTPs. Since substances in the SNF dataset were identified and quantified using reference standards, we also investigated potential systematic differences in breakthrough values due to variations in analytical methods. Figure S1 shows breakthrough values for substances shared between SNF and other datasets, using median values for comparison. The results demonstrate that despite different analytical methods, similar breakthrough values were obtained for the same substances. Even when differences occurred, there was no consistent trend of under- or overestimation when using semi-quantitative area-based methods to determine breakthrough. This finding has two key implications: (i) Consistent with recent studies,[26] area-based removal calculations are sufficient for monitoring chemical substance breakthroughs from WWTPs, and (ii) The target substances in the SNF dataset are more poorly removed, providing valuable examples for the model and enabling it to better identify which molecules tend to escape treatment. The sampling campaign that resulted in the SNF dataset was designed to focus on structurally complex micropollutants such as pesticides and pharmaceuticals, while the other datasets are not restrictive in this sense and include many molecules which are easier to degrade (e.g., organic acids, small peptides etc.).

We also analyzed the substances in each dataset in terms of their chemical structures. Figure 1.a shows a two-dimensional representation of the chemical space of all the substances included in our dataset. We used a t-stochastic-neighbors algorithm to group similar molecules, where similarity is based on the Morgan circular fingerprints. This is a common procedure to visualize the chemical space but often representations are not comparable. We therefore chose to reproduce the embedding of a recent publication[11] as shown in Figure 1.b, and mapped our substances within the same space of over 134’000 marketed chemicals. The figure shows that our dataset covers a wide chemical space with examples in all major classes of organic chemicals. Nonetheless it also evident that most classes are only sparsely covered. We also analyzed the individual datasets and found they covered similar sections of the space; as shown in Figure 1 there is no clear clustering of any of the sets.

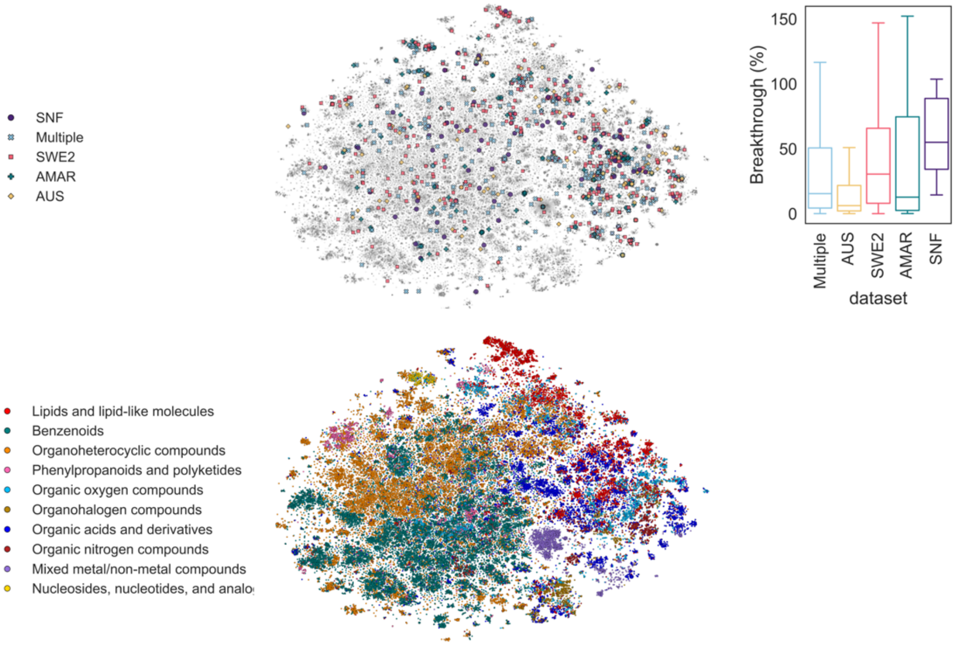

Figure 1. Two-dimensional representation of the chemical space covered by molecules included in this study (cyan) in comparison with 140000 marketed chemicals as described by von Borris et al.

3.2 Evaluation of model performance – Setting the benchmark by testing different regressor-feature pairs

We developed models to predict breakthrough values based on chemical features by exploring various combinations of regressors and feature sets. To ensure fair comparisons, each regressor-feature pair was evaluated using the same training and testing splits, following the nested CV method explained in the SI Section S6. The data used in this section was selected applying all the curation criteria described in Section 2.3 as these are the data with the highest confidence; this set contains 462 compounds.

Preliminary analysis showed very poor performance from DT, L-SVR, Ridge, KR, and MLR (see SI Section S7, Figure S3), so we only discuss the results obtained for RF, SVR, GBR, KNN, GPR and AB.

The performances on unseen data for the best regressor-descriptor pairs are summarized Figure 2. All regressors have on average comparable performance, with RF performing slightly better than the other regressors, regardless of the descriptors selected or the feature reduction methods. Further details about the statistical significance of these differences are explained in the SI Section S8. For any given regressor the differences in performance using different descriptors were also small but there is a tendency of MACCS to perform better, exceptions are KNN and GPR where models using Mordred were superior. Moreover, we found that overall ePFP was least effective for predicting breakthroughs, especially when compared to MACCS which are also expert-defined substructure-based fingerprints. This poor performance was unexpected as substructures included in the ePFP are known to trigger biological transformations.[36] However, since these rules are specific to enzymatic transformations, many substances trigger only a small subset of rules, while many other rules remain untriggered. Among the 1,037 substances with valid breakthrough values, only 577 had unique fingerprints. In comparison, the number of unique fingerprints using MACCS is significantly higher, with 909 unique fingerprints. We

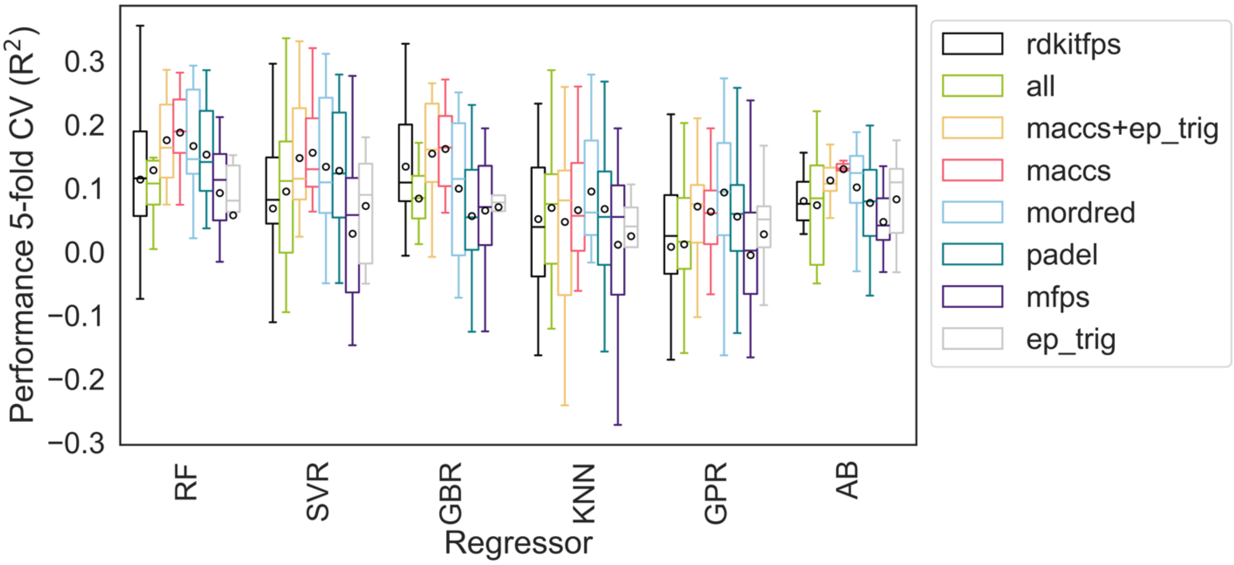

Figure 2. Performance of models in unseen data. a) models are grouped by regressor type showing that RF performs slightly better regardless of the descriptor of choice. b) models using RF as regressor are grouped by descriptor type showing improved performance using MACCS and worst performance using ePFP

therefore attribute the difference in performance between ePFP and MACCS fingerprints to the higher incidence of collisions in ePFP and hence consider ePFP by themselves as insufficient for modeling.

Our analyses revealed that similar performance may be obtained with different combinations of regressors and descriptors, however we conclude that there is a tendency of the RF-MACCS combination to do better. Moreover, RF offers additional benefits: it performs well without feature selection, is inherently robust, has an out-of-bag (OOB) option to reduce overfitting, and, as an ensemble method, provides a measure of confidence in predictions. MACCS also have several advantages: i) they are easy to interpret because they refer to substructures with clear definitions ii) they are calculated even faster than traditional descriptors because calculations are limited to matching substructures iii) training the models is also very fast because MACCS result in very short fingerprints (i.e., > 10 times smaller than ECFPs). Interpretability is important for added confidence in the developed models and fast calculations are particularly important when screening very large numbers of chemicals. Given these advantages, we further optimized the MACCS-RF model.

3.3 Can we improve the models by adding more data during training?

Since we had explored the performance of a wide range of algorithms and features and all models showed low predictive ability with unseen data, we explored whether performance could be improved by additional training data. Specifically, we decided to focus on the quality of the data and aimed to identify the criteria that defined the best training set. This is important for two main reasons: (i) to optimize and select features using the most informative data, and (ii) to provide guidelines on which types of data should or should not be included in future expansions of the database.

The results of retraining the RF-MACCS model with different training sets covering all possible combinations of our additional curation criteria are shown in the SI Figure S4 in Section S9. The models are tested on unseen subsets (size = 92) of the data with the highest confidence (size = 462). The concept is to discern whether additional data with a higher uncertainty improves model performance or rather introduces noise. Moreover, different criteria also mean a different training size which is also expected to have an important effect on performance. The best performances were achieved using all curation criteria and combination I+III, that is, the combination of only compounds for which data from at least three WWTPs were available and only those for which the variability across plants was low (i.e., standard deviation in log units < 0.7). Most other sets have similar or worse performance. We opted for combination (I+III) for further model development considering that a larger number of chemicals (856 compared to 369) in the training set leads to a more general model, and we consider that restricting the domain of applicability does not compensate the small gain in performance.

3.4 Performance of the final model – optimization and comparison to widely used regulatory models

Optimization. Models were finally optimized by 5-fold CV using the subset that follows the curation criteria explained in Section 3.3. We performed a randomized search over a large range of hyperparameter combinations and later a grid search over a smaller range close to the best combination found in the random search. When performance was the same, we gave priority to simpler models (i.e., smaller number of trees, smaller depth and larger number of minimum samples for a split). Further details are provided in the SI Section S10 and Figure S5. Finally, As previously described by Sheridan et al.[37] RF models often overestimate low values and underestimate high values. To address this systematic bias, we applied a linear regression model to adjust the RF's raw predictions, following Sheridan's method. In this approach, the linear model is fitted on training data only and uses the relationship between RF predictions and actual breakthrough values from the training set to produce adjusted predictions. When reporting the performance of our optimized model, we refer to this adjusted prediction rather than the raw RF output.

Definition of applicability domain (AD). As the final step to characterize our model, we aimed to define its domain of applicability. There is no established consensus on how AD must be defined but most approaches use either similarity metrics or ensemble agreement metrics.[37–41] Among similarity metrics, the Tanimoto Similarity Index is widely used and has proven effective.[42,43] This similarity measure is calculated pairwise; it can be defined as the similarity to the most similar molecule (i.e., ***SimiliratyNearest***) in the training set or as the average similarity to the five nearest molecules in the training set (i.e., ***SimiliratyNearest5***). A threshold is typically applied to determine whether a molecule falls within the model's domain of applicability. In our case, we tested both *SimilarityNearest* and *SimilarityNearest5*. Rather than introducing a threshold, we investigated whether similarity to the training set correlated with improved predictions by ranking our predictions based on similarity and recalculating the RMSE as we progressively excluded a fraction of the set with the lowest similarity. We evaluated the RMSE of 18 subsets, ranging from all data to the top 10% most similar data in 5% increments. Figure S6 in the SI Section S11, shows that when SimilarityNearest is used, improvement in RMSE is achieved only after removing nearly 80% of most dissimilar molecules. This suggests that, in our case, SimilarityNearest alone is not such a useful metric for identifying good predictions.

We observed better performance when using *SimilarityNearest5*, but it also required removing a very large portion of predictions before a clear improvement could be observed. These results indicate that, if we were to apply a similarity metric, an arbitrary threshold would lead to unreliable results and an optimized threshold (e.g., only top 20% most similar) would limit considerably the applicability of the model.

We repeated this analysis, but instead of ranking based on similarity, we ranked predictions by the standard deviation of the individual tree predictions in the ensemble (i.e., *TreeSD*). Here, we observed a steady decrease in RMSE and increase in R^2 as we removed predictions with the largest standard deviations, indicating that predictions were more accurate when the trees agreed more closely. Overall, standard deviation in the predictions of individual trees serves as a strong indicator of confidence in predictions.

Interpretation of final model. Model performance on unseen data was lower than expected so understanding the model's decisions is important for confidence in predictions. We calculated feature importances across the different folds in 5-fold CV. Figure S10 in the SI Section shows the importance for the 10 most important features. Then we calculated the SHAP (SHapley Additive exPlanations) values for the fitted values (i.e., model "predictions" on the training set) in order to better understand the decisions taken by the model for molecules with a known breakthrough (Figure 3).

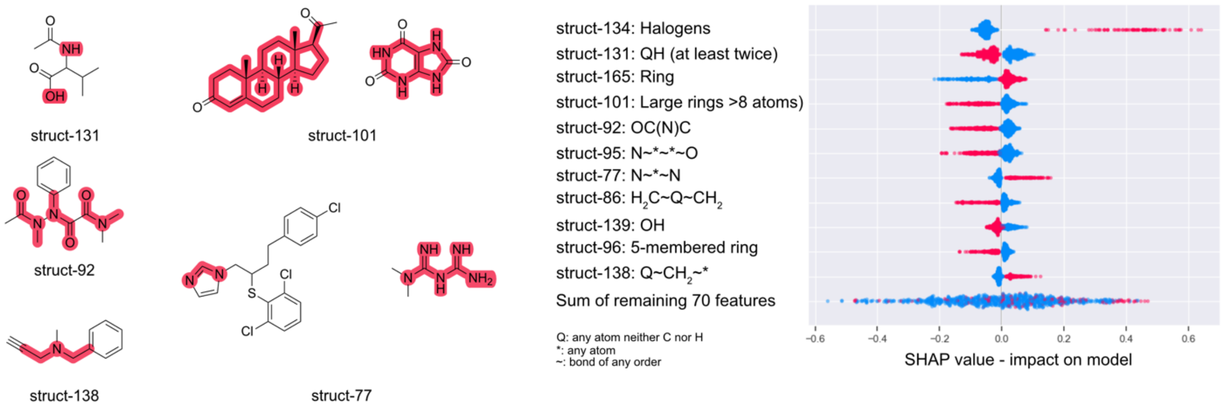

Figure 3. Feature importance. a) Shows the feature importance during 5-fold cross validation. b) shows the shapley additive values; red represents the substructure was present and blue that it was not. For example, the presence of struct-134 (i.e., halogens) always contributed to larger breakthroughs. The opposite is observed for struct-139 (i.e., -OH) where its presence always contributed to smaller breakthroughs.

Struct-134 (i.e., presence of -F, -Cl, -Br, -I) is both the most important feature during cross validation and the substructure with the largest impact on predictions when present. Figure 3 shows how every time a halogen is present the model assigns larger breakthrough values, which is in line with current understanding of biodegradability.[44,45] The absence of halogens (blue circles in struct-134 row) only moderately contributes to lower breakthroughs. A similar tendency is observed for struct-165, which refers to the presence of rings (both aromatic or aliphatic). If a ring structure is present, the model assigns larger breakthroughs, although to a lesser extent than if halogens are present. And differently from the halogens, the absence of rings highly contributes to lower breakthroughs, which can be observed from the large negative values of the blue circles for struct-165. Two more substructures have a similar tendency (i.e., higher breakthrough when present), i.e., struct-77, which refers to the pattern N~*~N where N is nitrogen, (*) is any atom and (~) is any bond, and struct-138, which refers to the pattern Q~CH_2~* where Q represents any atom which is not C or H. Struct-77 in many cases encodes the presence of imidazole, which is common in many pharmaceuticals (e.g., antifungals such as butoconazole) and is a moiety that is difficult to degrade. Guanidine-like substructures would also match struct-77 and these are often found in pharmaceuticals (e.g., metformin a treatment against diabetes). Struct-138 is more general, note that both Q and (*) could be nitrogen atoms too but the central atom must be a CH_2 which excludes imidazole, guanidine, carboxylic acids and amides. Common examples of compounds that contain struct-138 are tertiary amines.

Struct-131 normally matched terminal N and O atoms, like alcohols, carboxylic acids and primary amines which are expected to have low breakthroughs. Presence of struct-101 (8-membered rings or larger where adjacent rings are counted as single ring) contributing to smaller breakthroughs is observed and rather counterintuitive. Examples of compounds that match struct-101 and have low observed breakthroughs include compounds with guanine-like substructures. Intuitively these aromatic rings are not expected to be biodegradable but there are plenty of naturally occurring substances with these substructures such as nucleotides and nucleosides. Differently, smaller aromatic rings (e.g., 6-membered rings) which would encode struct-165 but not 101 are xenobiotic and presented larger breakthroughs, and this explains the tendency of struct-101 as a driver of lower breakthroughs. Finally, struct-139 which refers to the presence -OH is also selected as an important predictor and its presence always leads to smaller breakthroughs as expected.

Benchmarking model performance. Next, we compared the predictions of the optimized model with those of the STPWIN tool from the EPI Suite, a tool developed by the US EPA and widely used in regulation, alternatives assessment and even academic research.[46] The agreement between the monitoring data and predictions of both our models and STPWIN are shown in Figure 4.a.The RMSE for our predictions was 0.62, compared to 0.92 for STPWIN predictions. The coefficient of determination was 0.22 for our model and -0.70 for STPWIN. Recent assessment of the quality of predictions of STPWIN is missing, however, previous studies have reported prediction errors within 1-2 log units of magnitude for similar tools, such as SimpleTreat.[6,47]

We believe that the chosen test set of chemicals is particularly challenging for process-based models such as STPWIN and SimpleTreat, as most molecules in this subset are removed primarily via biodegradation, a complex mechanism that remains difficult to predict accurately.[23] Both STPWIN and SimpleTreat rely on physicochemical properties to estimate removal by various mechanisms, ultimately combining these individual predictions for a total removal value.[5] STPWIN in particular outputs values for each individual mechanism, enabling us to calculate the fraction attributed to biodegradation. As expected, nearly all predictions rely heavily on biodegradation, indicating that prediction accuracy depends largely on the accuracy of the primary biodegradation rate constant. Similarly, Lautz et al.[47] observed that errors were 10 times higher when using biodegradation rate constants predicted by BIOWIN in comparison to using measured rate constants. Their study also confirmed that using plant-specific reactor parameters did not improve predictions significantly compared to simply using default values, which also highlight the enormous weight of biodegradation rate constants in the prediction errors. Our modeling approach, which is purely data-driven but trained on a large set of actual WWTP monitoring data, better predicts removals across a large range of compounds, despite the fact that it does not explicitly account for different removal processes. We conclude that our model is an important alternative for predicting removal in WWTPs, particularly for substances that are mainly removed by biodegradation, where STPWIN and SimpleTreat are likely to fail as they need to rely on predictions by BIOWIN.

Model application to relevant chemical space. To illustrate the utility of our model, we predicted breakthrough for over 14’000 chemicals registered under REACH. This list of single organic chemicals registered under REACH was compiled and curated by Arp & Hale[12] In Figure 4.a, we present prediction outcomes as percentage breakthrough to better illustrate the environmental significance. In some cases, breakthrough values are largely overestimated leading to breakthrough up to 150%. For nearly half of chemicals, the predicted breakthrough is below 20% and for 15% of chemicals the breakthrough is predicted to be above 80%.

We also analyzed the relationship between the breakthrough and the confidence in predictions. Most examples with a large confidence in predictions also have a large predicted breakthrough. These are mostly substances with chlorine and fluorine as substituents. Substances with low breakthrough and high confidence are mostly carboxylic acids, alcohols, ethers and guanine-like metabolites.

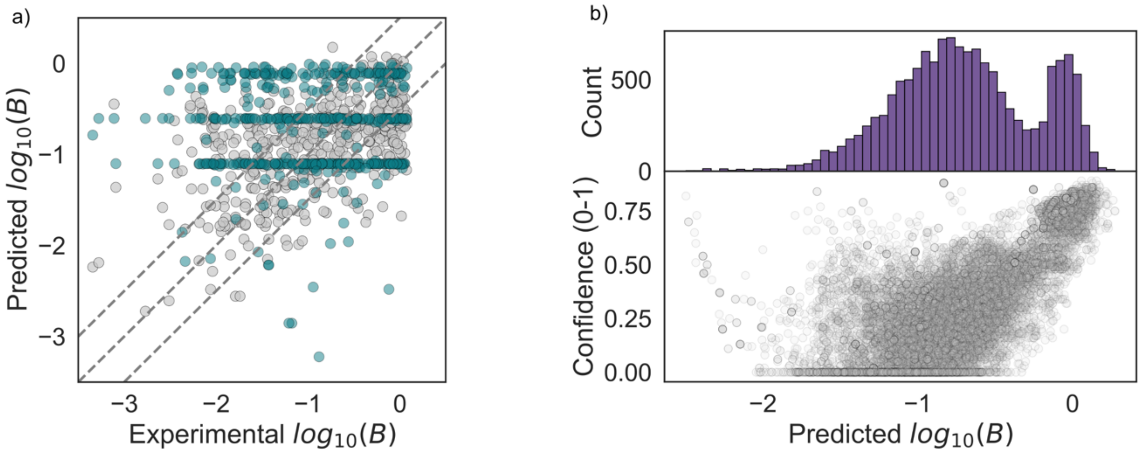

Figure 4. Predicted breakthrough for 14'000 chemical substances under REACH. a) distribution of predicted values b) relationship between predicted breakthrough and model confidence.

4 Conclusions

In this work, we present an approach to estimate the removal of micropollutants during wastewater treatment. With our approach, models learn from increasingly available, semi-quantitative monitoring data from WWTPs and are able to predict an expected breakthrough for a highly diverse set of organic molecules. These predictions proved more reliable than existing process-based models that are widely used in EU and US regulatory contexts, especially for molecules where no experimental biotransformation kinetic data are available. This suggests that our model could be an important and novel contribution to the toolbox of in silico models used for alternatives assessment, when evaluating new molecules in industrial research and development, or even for exposure modeling in a risk assessment context. We have here established a benchmark model, which is publicly available along with the training data and the scripts necessary to reproduce the data curation process (renkulab.io/projects/fenner-labs/projects/pepper). We anticipate that this benchmark and the highly transparently curated data set that we provide will facilitate further developments in the field.

5 Acknowledgements

This project was funded by the Swiss Federal Office for the Environment (FOEN).

The collection of the data used in this study was supported by two Australian Research Council Linkage Projects (LP150100364, LP190101124) and Unilever, Bedfordshire, U.K, and the Swiss National Science Foundation Project No.200021L_201006 .

6 References

(1) Coll, C.; Fenner, K.; Screpanti, C. Early Assessment of Biodegradability of Small Molecules to Support the Chemical Design in Agro & Pharma R&D. *CHIMIA* **2023**, *77* (11), 742–749. https://doi.org/10.2533/chimia.2023.742.

(2) Leder, C.; Suk, M.; Lorenz, S.; Rastogi, T.; Peifer, C.; Kietzmann, M.; Jonas, D.; Buck, M.; Pahl, A.; Kümmerer, K. Reducing Environmental Pollution by Antibiotics through Design for Environmental Degradation. *ACS Sustain. Chem. Eng.* **2021**, *9* (28), 9358–9368. https://doi.org/10.1021/acssuschemeng.1c02243.

(3) Zumstein, M. T.; Fenner, K. Towards More Sustainable Peptide- Based Antibiotics: Stable in Human Blood, Enzymatically Hydrolyzed in Wastewater? *CHIMIA* **2021**, *75* (4), 267–267. https://doi.org/10.2533/chimia.2021.267.

(4) US EPA, O. *EPI Suite^{TM}-Estimation Program Interface*. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface (accessed 2024-11-04).

(5) Struijs, J. SimpleTreat 4.0: A Model to Predict the Fate and Emission of Chemicals in Wastewater Treatment Plants.

(6) Comber, S.; Gardner, M.; Sörme, P.; Ellor, B. The Removal of Pharmaceuticals during Wastewater Treatment: Can It Be Predicted Accurately? *Sci. Total Environ.* **2019**, *676*, 222–230. https://doi.org/10.1016/j.scitotenv.2019.04.113.

(7) Jelic, A.; Gros, M.; Ginebreda, A.; Cespedes-Sánchez, R.; Ventura, F.; Petrovic, M.; Barcelo, D. Occurrence, Partition and Removal of Pharmaceuticals in Sewage Water and Sludge during Wastewater Treatment. *Water Res.* **2011**, *45* (3), 1165–1176. https://doi.org/10.1016/j.watres.2010.11.010.

(8) Ortiz de García, S.; Pinto Pinto, G.; García Encina, P.; Irusta Mata, R. Consumption and Occurrence of Pharmaceutical and Personal Care Products in the Aquatic Environment in Spain. *Sci. Total Environ.* **2013**, *444*, 451–465. https://doi.org/10.1016/j.scitotenv.2012.11.057.

(9) Franco, A.; Struijs, J.; Gouin, T.; Price, O. R. Evolution of the Sewage Treatment Plant Model SimpleTreat: Use of Realistic Biodegradability Tests in Probabilistic Model Simulations. *Integr. Environ. Assess. Manag.* **2013**, *9* (4), 569–579. https://doi.org/10.1002/ieam.1413.

(10) Kim, H.-J.; Lee, H.-J.; Lee, D. S.; Kwon, J.-H. Modeling the Fate of Priority Pharmaceuticals in Korea in a Conventional Sewage Treatment Plant. *Environ. Eng. Res.* **2009**, *14* (3), 186–194.

(11) von Borries, K.; Holmquist, H.; Kosnik, M.; Beckwith, K. V.; Jolliet, O.; Goodman, J. M.; Fantke, P. Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization. *Environ. Sci. Technol.* **2023**, *57* (46), 18259–18270. https://doi.org/10.1021/acs.est.3c05300.

(12) Arp, H. P. H.; Hale, S. E. Assessing the Persistence and Mobility of Organic Substances to Protect Freshwater Resources. *ACS Environ. Au* **2022**, *2* (6), 482–509. https://doi.org/10.1021/acsenvironau.2c00024.

(13) Fenner, K.; Screpanti, C.; Renold, P.; Rouchdi, M.; Vogler, B.; Rich, S. Comparison of Small Molecule Biotransformation Half-Lives between Activated Sludge and Soil: Opportunities for Read-Across? *Environ. Sci. Technol.* **2020**, *54* (6), 3148–3158. https://doi.org/10.1021/acs.est.9b05104.

(14) Fenner, K.; Elsner, M.; Lueders, T.; McLachlan, M. S.; Wackett, L. P.; Zimmermann, M.; Drewes, J. E. Methodological Advances to Study Contaminant Biotransformation: New Prospects for Understanding and Reducing Environmental Persistence? *ACS EST Water* **2021**, *1* (7), 1541–1554. https://doi.org/10.1021/acsestwater.1c00025.

(15) Escher, B. I.; Fenner, K. Recent Advances in Environmental Risk Assessment of Transformation Products. *Environ. Sci. Technol.* **2011**, *45* (9), 3835–3847. https://doi.org/10.1021/es1030799.

(16) Boethling, R.; Fenner, K.; Howard, P.; Klečka, G.; Madsen, T.; Snape, J. R.; Whelan, M. J. Environmental Persistence of Organic Pollutants: Guidance for Development and Review of POP Risk Profiles. *Integr. Environ. Assess. Manag.* **2009**, *5* (4), 539–556. https://doi.org/10.1897/IEAM_2008-090.1.

(17) Arp, H. P.; Hale, S. *REACH: Guidance and Methods for the Identification and Assessment of PMT/vPvM Substances: Second Edition*; 2023. https://doi.org/10.13140/RG.2.2.24980.48001.

(18) Kostal, J.; Voutchkova-Kostal, A. Going All In: A Strategic Investment in In Silico Toxicology. *Chem. Res. Toxicol.* **2020**, *33* (4), 880–888. https://doi.org/10.1021/acs.chemrestox.9b00497.

(19) Rücker, C.; Kümmerer, K. Modeling and Predicting Aquatic Aerobic Biodegradation – a Review from a User's Perspective. *Green Chem.* **2012**, *14* (4), 875–887. https://doi.org/10.1039/C2GC16267A.

(20) Soares, T. A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge

of Machine Learning Methods. *J. Chem. Inf. Model.* **2022**, *62* (22), 5317–5320. https://doi.org/10.1021/acs.jcim.2c01422.

(21) Posthumus, R.; Traas, T. P.; Peijnenburg, W. J. G. M.; Hulzebos, E. M. External Validation of EPIWIN Biodegradation Models. *SAR QSAR Environ. Res.* **2005**, *16* (1–2), 135–148. https://doi.org/10.1080/10629360412331319899.

(22) Acharya, K.; Werner, D.; Dolfing, J.; Barycki, M.; Meynet, P.; Mrozik, W.; Komolafe, O.; Puzyn, T.; Davenport, R. J. A Quantitative Structure-Biodegradation Relationship (QSBR) Approach to Predict Biodegradation Rates of Aromatic Chemicals. *Water Res.* **2019**, *157*, 181–190. https://doi.org/10.1016/j.watres.2019.03.086.

(23) Jiang, S.; Liang, Y.; Shi, S.; Wu, C.; Shi, Z. Improving Predictions and Understanding of Primary and Ultimate Biodegradation Rates with Machine Learning Models. *Sci. Total Environ.* **2023**, *904*, 166623. https://doi.org/10.1016/j.scitotenv.2023.166623.

(24) Wang, L.; Lei, Z.; Yun, S.; Yang, X.; Chen, R. Quantitative Structure-Biotransformation Relationships of Organic Micropollutants in Aerobic and Anaerobic Wastewater Treatments. *Sci. Total Environ.* **2024**, *912*, 169170. https://doi.org/10.1016/j.scitotenv.2023.169170.

(25) Nolte, T. M.; Chen, G.; van Schayk, C. S.; Pinto-Gil, K.; Hendriks, A. J.; Peijnenburg, W. J. G. M.; Ragas, A. M. J. Disentanglement of the Chemical, Physical, and Biological Processes Aids the Development of Quantitative Structure-Biodegradation Relationships for Aerobic Wastewater Treatment. *Sci. Total Environ.* **2020**, *708*, 133863. https://doi.org/10.1016/j.scitotenv.2019.133863.

(26) Chirico, N.; McLachlan, M. S.; Li, Z.; Papa, E. In Silico Approaches for the Prediction of the Breakthrough of Organic Contaminants in Wastewater Treatment Plants. *Environ. Sci. Process. Impacts* **2024**. https://doi.org/10.1039/D3EM00267E.

(27) Siemers, F. M.; Feldmann, C.; Bajorath, J. Minimal Data Requirements for Accurate Compound Activity Prediction Using Machine Learning Methods of Different Complexity. *Cell Rep. Phys. Sci.* **2022**, *3* (11), 101113. https://doi.org/10.1016/j.xcrp.2022.101113.

(28) Qiliang ‘Luke’ Wang; G. Apul, O.; Xuan, P.; Luo, F.; Karanfil, T. Development of a 3D QSPR Model for Adsorption of Aromatic Compounds by Carbon Nanotubes: Comparison of Multiple Linear Regression, Artificial Neural Network and Support Vector Machine. *RSC Adv.* **2013**, *3* (46), 23924–23934. https://doi.org/10.1039/C3RA43599G.

(29) Huang, K.; Zhang, H. Classification and Regression Machine Learning Models for Predicting Aerobic Ready and Inherent Biodegradation of Organic Chemicals in Water. *Environ. Sci. Technol.* **2022**, *56* (17), 12755–12764. https://doi.org/10.1021/acs.est.2c01764.

(30) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097–2098. https://doi.org/10.1021/es5002105.

(31) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474. https://doi.org/10.1002/jcc.21707.

(32) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(33) *Python API Reference — The RDKit 2024.03.4 documentation*. http://rdkit.org/docs/api-docs.html (accessed 2024-07-01).

(34) Hafner, J.; Lorsbach, T.; Schmidt, S.; Brydon, L.; Dost, K.; Zhang, K.; Fenner, K.; Wicker, J. Advancements in Biotransformation Pathway Prediction: Enhancements, Datasets, and Novel Functionalities in enviPath. *J. Cheminformatics* **2024**, *16* (1), 93. https://doi.org/10.1186/s13321-024-00881-6.

(35) Vallat, R. Pingouin: Statistics in Python. *J. Open Source Softw.* **2018**, *3* (31), 1026. https://doi.org/10.21105/joss.01026.

(36) Ellis, L. B. M.; Gao, J.; Fenner, K.; Wackett, L. P. The University of Minnesota Pathway Prediction System: Predicting Metabolic Logic. *Nucleic Acids Res.* **2008**, *36* (suppl_2), W427–W432. https://doi.org/10.1093/nar/gkn315.

(37) Sheridan, R. P. Using Random Forest To Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53* (11), 2837–2850. https://doi.org/10.1021/ci400482e.

(38) Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2019**, *59* (1), 181–189. https://doi.org/10.1021/acs.jcim.8b00597.

(39) Wang, Z.; Chen, J.; Hong, H. Applicability Domains Enhance Application of PPARγ Agonist Classifiers Trained by Drug-like Compounds to Environmental Chemicals. *Chem. Res. Toxicol.* **2020**, *33* (6), 1382–1388. https://doi.org/10.1021/acs.chemrestox.9b00498.

(40) Sheridan, R. P. The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity. *J. Chem. Inf. Model.* **2015**, *55* (6), 1098–1107. https://doi.org/10.1021/acs.jcim.5b00110.

(41) Korolev, V.; Mitrofanov, A.; Korotcov, A.; Tkachenko, V. Graph Convolutional Neural Networks as “General-Purpose” Property Predictors: The Universality and Limits of Applicability. *J. Chem. Inf. Model.* **2020**, *60* (1), 22–28. https://doi.org/10.1021/acs.jcim.9b00587.

(42) Reymond, J.-L. Molecular Similarity for Drug Discovery, Target Prediction and Chemical Space Visualization. *CHIMIA* **2022**, *76* (12), 1045–1045.

(43) Gini, G. The QSAR Similarity Principle in the Deep Learning Era: Confirmation or Revision? *Found. Chem.* **2020**, *22* (3), 383–402. https://doi.org/10.1007/s10698-020-09380-6.

(44) Boethling, R. S.; Howard, P. H.; Meylan, William.; Stiteler, William.; Beauman, Julie.; Tirado, Nestor. Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation. *Environ. Sci. Technol.* **1994**, *28* (3), 459–465. https://doi.org/10.1021/es00052a018.

(45) Straub, J. O.; Le Roux, J.; Tedoldi, D. Halogenation of Pharmaceuticals Is an Impediment to Ready Biodegradability. *Water* **2023**, *15* (13), 2430. https://doi.org/10.3390/w15132430.

(46) van Dijk, J.; Flerlage, H.; Beijer, S.; Slootweg, J. C.; van Wezel, A. P. Safe and Sustainable by Design: A Computer-Based Approach to Redesign Chemicals for Reduced Environmental Hazards. *Chemosphere* **2022**, *296*, 134050. https://doi.org/10.1016/j.chemosphere.2022.134050.

(47) Lautz, L. S.; Struijs, J.; Nolte, T. M.; Breure, A. M.; van der Grinten, E.; van de Meent, D.; van Zelm, R. Evaluation of SimpleTreat 4.0: Simulations of Pharmaceutical Removal in Wastewater Treatment Plant Facilities. *Chemosphere* **2017**, *168*, 870–876. https://doi.org/10.1016/j.chemosphere.2016.10.123.