Eawag-Soil in enviPath: A new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data

Diogo A. R. S. Latino[a,]*, Jörg Wicker[b], Martin Gütlein[b], Emanuel Schmid[c], Stefan Kramer[b], Kathrin Fenner[a,d]

[a] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b] Institute of Computer Science, Johannes Gutenberg University Mainz, 55128 Mainz, Germany
[c] Scientific IT Services, ETH Zürich, 8092 Zürich, Switzerland
[d] Department of Chemistry, University of Zürich, 8057 Zürich, Switzerland

*Corresponding author: diogo.latino@eawag.ch, Tel.: +41 58 765 5737

Abstract

Developing models for the prediction of microbial biotransformation pathways and half-lives of trace organic contaminants in different environments requires as training data easily accessible and sufficiently large collections of respective biotransformation data that are annotated with metadata on study conditions. Here, we present the *Eawag-Soil* package, a public database that has been developed to contain all freely accessible regulatory data on pesticide degradation in laboratory soil simulation studies for pesticides registered in the EU (282 degradation pathways, 1535 reactions, 1619 compounds and 4716 biotransformation half-life values with corresponding metadata on study conditions). We provide a thorough description of this novel data resource, and discuss important features of the pesticide soil degradation data that are relevant for model development. Most notably, the variability of half-life values for individual compounds is large and only about one order of magnitude lower than the entire range of median half-life values spanned by all compounds, demonstrating the need to consider study conditions in the development of more accurate models for biotransformation prediction. We further show how the data can be used to find missing rules relevant for predicting soil biotransformation pathways. From this analysis, eight examples of reaction types were presented that should trigger the formulation of new biotransformation rules, e.g., Ar-OH methylation, or the extension of existing rules e.g., hydroxylation in aliphatic rings. The data were also used to exemplarily explore the dependence of half-lives of different amide pesticides on chemical class and experimental parameters. This analysis highlighted the value of considering initial transformation reactions for the development of meaningful quantitative-structure biotransformation relationships (QSBR), which is a novel opportunity offered by the simultaneous encoding of transformation reactions and corresponding half-lives in *Eawag-Soil*. Overall, *Eawag-Soil* provides an unprecedentedly rich collection of manually extracted and curated biotransformation data, which should be useful in a great variety of applications.

Introduction

When chemicals are released into the environment during or at the end of their product life cycle, their persistence in the environment is highly undesirable. Biotransformation by microbial communities in technical and environmental systems such as sewage treatment plants, aquatic sediments, and soils is a very efficient mechanism to reduce their environmental persistence, but might also lead to the formation of potentially hazardous transformation products [1-3]. Since the experimental assessment of chemical persistence and transformation product formation on a compound-by-compound basis is highly laborious and costly, so-called *in silico* or *non-testing* approaches that rely on computer-based algorithms to predict biotransformation have gained in importance for the evaluation of new and existing chemicals [4]. It has been suggested that such approaches would also be of use in the implementation of the "benign by design" concept where the environmental risk of a chemical is considered early in the development process or even before synthesis [5].

Quantitative structure-biodegradation relationships (QSBRs) predict chemical persistence, i.e., half-lives or readiness of biodegradation, based on chemical structure. They range from chemical class-specific to more broadly applicable models, and from simple regression models to models developed with machine learning methods [5-7]. Chemical class-specific models [8-10] typically yield reasonably accurate predictions of actual degradation half-lives, but are of limited use for risk assessment purposes due to their restricted applicability domain. In contrast, more widely applicable models are typically trained on a number of databases containing collections of ready biodegradability data from standardized tests carried out according to the OECD test guidelines for a wide variety of chemicals [11]. They usually show reasonable predictive power with approximately 80% correct binary classification as to whether a chemical is readily biodegradable or not (e.g., [12-14]). However, the accuracy of these more widely applicable models for quantitatively predicting biotransformation rates or half-lives under specific environmental conditions, which is what would actually be needed for risk assessment purposes, remains rather low [15-17].

Pathway prediction systems (e.g., PathPred [18], Catalogic [19], BNICE [20], and Eawag-PPS (former UM-PPS [21])) typically rely on dictionaries of biotransformation rules that recognize compound functional groups and transform them into product substructures. These biotransformation rules were designed to reflect known microbial transformation pathways of chemical contaminants. They are mostly based on the respective data collected in the Eawag Biodegradation/Biocatalysis Database (Eawag-BBD), formerly known as the University of Minnesota Biodegradation/Biocatalysis Database (UM-BBD) [22], which is considered the most extensive collection of manually curated biotransformation pathways of chemical contaminants [23] and, more recently, is available as Eawag-BBD from two online platforms (Eawag-BBD/PPS [24] and enviPath [25, 26]). Rule-based systems have been shown to predict transformation products observed in the environment fairly comprehensively (i.e., to display high sensitivity), but to notoriously predict many irrelevant products that are not likely to occur under specific environmental conditions (i.e., to display low selectivity) [27]. While application of machine learning methods to improve relative reasoning between rules has increased the selectivity for the training database (i.e., Eawag-BBD), selectivity on a set of pesticide soil degradation data used for external validation remained low [28]. This poses a problem if the models were to be used in a chemical risk assessment context where resources for assessing the risk associated with transformation products are limited.

We argue that the low accuracy of QSBRs and the low selectivity of pathway prediction have at least two common causes. First, almost all approaches to biotransformation prediction are based on chemical structure only and have so far mostly ignored the fact that half-lives for the same compound can vary strongly within the same type of environmental compartment [15]. This observed variance stems from the fact that slightly different environmental conditions shape different microbial communities that differ in their taxonomic composition and hence their pool of enzymes that catalyze biotransformation reactions of chemical contaminants. Thus, different enzyme-catalyzed reactions might occur at vastly different rates across different microbial communities [29, 30]. This suggests that biotransformation prediction could be greatly improved by not only considering chemical structure, but by also factoring in specific environmental conditions. Second, most of the data in Eawag-BBD stems from studies with pure cultures of microorganisms or laboratory cultures with elongated adaptation periods. Thus, while the organisms and degrading enzymes are typically well-characterized in these studies and hence reported in the database, the current data in Eawag-BBD cannot be used for understanding the influence of environmental factors on biotransformation pathways nor is the relevance of the reported pathways under actual environmental conditions known. In pure and enrichment culture systems, besides being known to be impacted by culturing artifacts [31], the compound of concern serves as sole growth substrate. The latter is most likely also true for the ready biodegradability tests that are run at high concentrations of the test chemicals as dominant carbon source [32]. Under actual environmental conditions, contaminant trace concentrations are likely transformed co-metabolically by mixed microbial communities alongside varying amounts of other, natural organic material. The determinants of such co-metabolic transformations are typically not of thermodynamic nature as in growth-related metabolism, but rather the available pool of catalytic enzymes of the microbial community as shaped by the prevailing environmental conditions [23]. Finally, it is worth noting that QSBRs and systems for the prediction of biotransformation pathways have so far mostly been developed independently. However, given the fact that observed biotransformation half-lives and transformation product spectra (i.e., the observed biotransformation pathways) both depend on the rates of individual enzyme-catalyzed biotransformation reactions, treating these two types of information separately may lead to a loss of information content.

In summary, we hypothesize that development of more accurate QSBRs and pathway prediction models is impeded by a lack of biotransformation data (i.e., half-lives and pathway information) from environmentally relevant mixed microbial communities and associated metadata on environmental and/or experimental conditions. The latter are needed to account for their influence on the observed biotransformation outcomes. Recently, we have introduced enviPath as a new database and pathways prediction system that is suited to approach these information gaps [25]. enviPath offers a database environment that, first, facilitates the annotation of biotransformation half-life and pathway information, and, second, allows for supplementing the half-life and pathway information with metadata, e.g., environmental and/or experimental conditions, through so-called scenarios. One fairly consistent and large resource of chemical biotransformation data is data submitted for regulatory risk assessments. These substance-specific dossiers typically contain information on biotransformation half-lives and pathways from so-called simulation studies conducted for different relevant environmental compartments (i.e., agricultural soil, aquatic sediments, activated sludge). Such data is currently mostly available for pesticides [33], but upon implementation of REACH should increasingly also become available for industrial chemicals [34]. However, these data are currently not readily available in electronic format [35], and, if so (e.g., PPDB [36]), do not contain pathway information, lack annotation with metadata on study conditions, or, to the best of our knowledge, are not publically available (e.g., MetaPath [35, 37]).

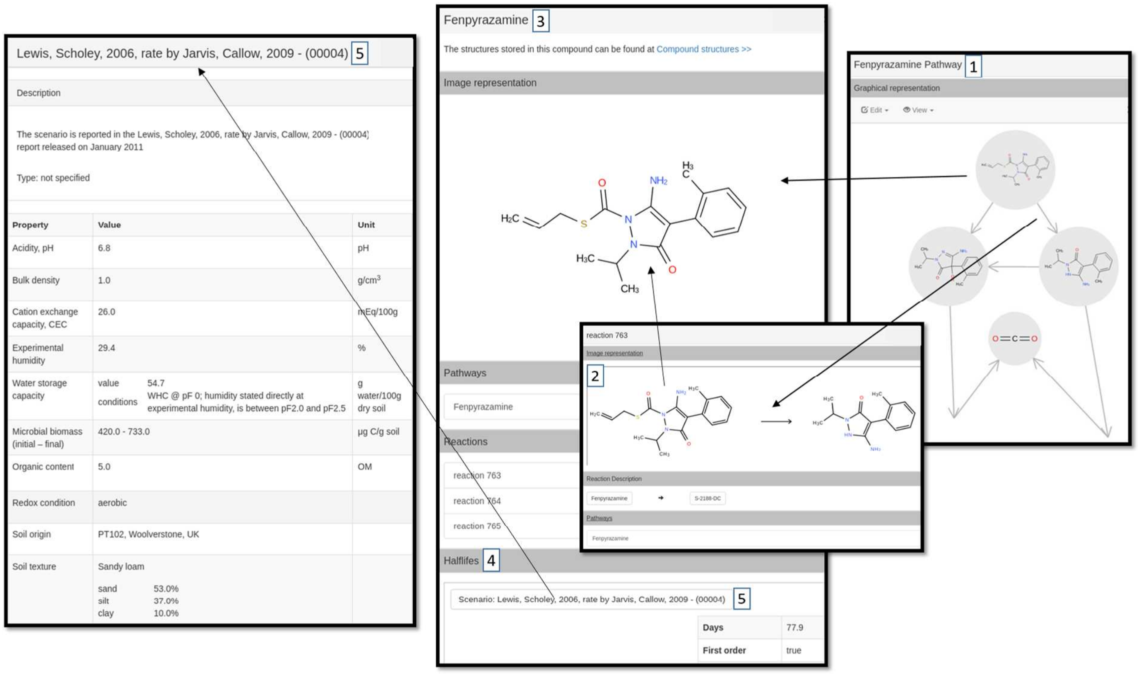
Therefore, the objective of the work presented here was to electronically encode all freely accessible regulatory data on pesticide degradation in laboratory soil simulation studies and to make these data publically available for the development of improved QSBR models. Here, we present a thorough description of this novel data resource, discuss characteristics of pesticide soil degradation data that are relevant to model development, and give two examples of explorative analyses that should support the further development of QSBRs and pathway prediction models.

Materials and Methods

Pesticide soil degradation data

We extracted pesticide soil degradation information from pesticide registration dossiers made publically available through the European Food Safety Authority (EFSA) [33]. Specifically, only results from laboratory studies conducted under aerobic conditions as reported in “Annex B.8: Fate and behavior, B8.1: Route and rate of degradation in soil” in the respective dossiers were considered. Initially, assessment reports, draft assessment reports (DARs) and additional reports available between 6/2015 and 6/2016 for 375 active substances were screened. Of these, dossiers for 93 active substances were not considered further because the pesticides agents were not actual chemicals (e.g., bacteria) or complex mixtures (e.g., clover oil), or because no degradation scheme was available or no aerobic degradation studies in soil had been submitted at all. For the remaining 282 pesticides and agriculture-related compounds, degradation information and accompanying metadata on study conditions were encoded as separate data package, *Eawag-Soil*, in enviPath [38].

In *the Eawag-Soil* package, pathway information is stored in a biotransformation reaction scheme in the entity *pathway* (see example in Figure 1). Compounds and reactions participating in a given pathway are stored separately in the entities *compound* and *reaction* (see Figure 1). Metadata on the experimental conditions (e.g., soil texture, soil moisture, pH, etc.) are stored in the entity *scenario* (see Figure 1). A detailed list and explanation of all experimental conditions considered in the *Eawag-Soil* package as well as the conventions used to store the data as standardized as possible are given in the Supporting Information (SI) (Section S1 *Eawag-Soil* metadata and conventions) When available, one or several biotransformation half-lives (in the form of dissipation half-lives, DT50) are additionally associated with a given compound in the pathway and a specific scenario. *Pathways* depict all reactions and compounds observed in aerobic soil experiments under any experimental condition. Since not all transformation products in a pathway scheme are always observed, compounds in the pathway are associated with a given scenario only when they have been experimentally observed under the specific experimental conditions (see example in Figure 1). The associated scenarios are listed on the compound page.



Lewis, Scholey, 2006, rate by Jarvis, Callow, 2009 - (00004)

Description

The scenario is reported in the Lewis, Scholey, 2006, rate by Jarvis, Callow, 2009 - (00004) report released on January 2011

Type: not specified

Property	Value	Unit
Acidity, pH	6.8	pH
Bulk density	1.0	g/cm^3
Cation exchange capacity, CEC	26.0	mEq/100g
Experimental humidity	29.4	%
Water storage capacity	value 54.7 conditions WHC @ pF 0; humidity stated directly at experimental humidity, is between pF2.0 and pF2.5	g water/100g dry soil
Microbial biomass (initial – final)	420.0 - 733.0	µg C/g soil
Organic content	5.0	OM
Redox condition	aerobic	
Soil origin	PT102, Woolverstone, UK	
Soil texture	Sandy loam sand 53.0% silt 37.0% clay 10.0%	

Fenpyrazamine

The structures stored in this compound can be found at Compound structures >>

Image representation

Pathways

Fenpyrazamine

Reactions

reaction 763

reaction 764

reaction 765

Halflifes

Scenario: Lewis, Scholey, 2006, rate by Jarvis, Callow, 2009 - (00004)

Days	77.9
First order	true

reaction 763

Image representation

Reaction Description

Fenpyrazamine → S-2188-DC

Pathways

Fenpyrazamine

Fenpyrazamine Pathway

Graphical representation

Figure 1: Scheme of assembled screenshots showing the most important elements of the *Eawag-Soil* package. 1: *Pathway* page, 2: *Reaction* page, 3: *Compound* page, 4: List of half-lives determined for the compounds and the associated scenario names, 5: *Scenario* page, containing the metadata on study conditions (i.e., experimental parameters).

Chemical space analysis

The chemical space covered by a set of compounds is defined by the multidimensional property space of the compounds and is used to define the applicability domain of a model. We compared the chemical spaces covered by *Eawag-BBD* and *Eawag-Soil*, and also compared them to the chemical space covered by a third set of 1024 pharmaceutical compounds prevalently used in Switzerland and the EU as extracted from Singer et al. [39]. This latter set of compounds was used to explore the hypothesis that the addition of the *Eawag-Soil* package extends the chemical space of enviPath towards more polar, multifunctional compounds such as pharmaceuticals.

The comparison was performed using a qualitative approach, i.e., the visualization of the top three principal components of the compounds in the three datasets, and a quantitative approach using a one-class support vector machine to identify objects that lie outside the chemical space. The top three principal components were calculated using the DataWarrior software [40] (http://www.openmolecules.org/datawarrior/) with the compounds represented using structural fragment fingerprints (i.e., binary structural features (ECFP4) [41]) calculated with the CDK software [42]. One-class support vector machines (SVM) [43] is a machine learning technique that was used to determine whether a compound belongs to the feature distribution space of an existing dataset or rather has to be considered as an outlier or a novel compound. One-class SVM models were trained on the *Eawag-BBD* dataset and the combined *Eawag-BBD* and *Eawag-Soil* datasets using the LIBSVM implementation [44]) with the compounds represented using structural fragment fingerprints as explained above. The ν-Parameter, which limits the number of predicted outliers in the training dataset, was set to a value of 2%.

Procedure for missing rule analysis (explorative analysis I)

The Eawag-PPS system, which is hosted and further maintained in our research group, currently uses a set of 249 biotransformation rules (btrules) that recognize specific functional groups in a molecule and transform them according to the generalized biotransformation reaction encoded in the rule [22]. These same rules also form the basis of the successor system enviPath [25]. When adding a new set of biotransformation data such as the *Eawag-Soil* package to enviPath, a first pre-requisite to use its information for improving pathway prediction models is to test the ability of the current rule set to cover the reactions in the new database. As a first explorative analysis, we therefore conducted a missing rule analysis.

Missing rule analysis was carried out in two steps: *i)* submission of the reactants of all chemical reactions in *Eawag-Soil* to the complete set of btrules contained in Eawag-PPS, and *ii)* comparison of the predicted reactions, i.e., reactant-product pairs, with the experimentally observed reactions. More specifically, the reactants of all reactions in *Eawag-Soil* were submitted to the Eawag-PPS system and three generations of transformation products were predicted through three times iterative application of the prediction cycle. Then all the predicted first-generation reactant-product pairs were compared with the experimentally observed reactant-product pairs and three different outcomes were noted: (*i*) an experimentally observed reaction is matched by a predicted reaction indicating that there is a rule that is correctly triggered by that reactant and that the system is therefore able to predict the biotransformation observed in the soil degradation studies; (*ii*) a predicted reaction does not match any of the experimentally observed reactions indicating that the system either predicts products that are not actually relevant for a given reactant, or that the product, although plausible, escaped analytical identification in the soil degradation studies; and (*iii*) an experimentally observed reaction is not matched by any of the predicted reactions pointing towards a missing rule for that specific kind of biotransformation reaction. While (*i*) gives the current sensitivity of the system towards pesticide active ingredients, and (*ii*) is indicative of its current selectivity (prior to the addition of new rules), reactions in (*iii*) were further explored to identify possible missing rules.

In a next step, for the set of reactions in (*iii*), the 2nd and 3rd generation products associated with their respective substrates were explored to see whether any of them matched with the experimentally observed product of the reaction. If so, this reaction is predicted in Eawag-PPS through a series of multiple reactions where the intermediates might actually be readily transformed further and therefore were not necessarily analytically observed and identified. These reactions were assigned as multi-step reactions and added to the pool of reactions in (*i*).

To find a first set of missing rules, the remaining reactions in (*iii*) were, first, sorted according to mass differences between the reactant and the product and, second, for a given mass difference, further manually sorted into types of reactions based on our perceived similarity of the reaction center. This approach is potentially limited because it will not group together reactions of the same type if functional groups of different size are cleaved off. Therefore, in a next step, some groups of reactions were joined together to make the reaction type more general and to more easily implement it as a new rule later. For example, O-demethylations and O-deethylations were grouped together as O-dealkylations, and, similarly, N-demethylations and N-deethylations were grouped together as N-dealkylations. While fully manual, the approach was found appropriate for the identification of the most populated reaction types. This approach also provides well-curated information that can be used for validation of semi- and fully automated chemoinformatics methods for reaction classification, which we plan to explore in a follow-up study.

Univariate and multivariate analysis of half-life data (explorative analysis II)

We performed an exploratory analysis of relationships between the experimental parameters and the DT50 values for a group of six sulfonamide herbicides. We calculated Spearman rank correlation coefficients and their related significance of being different from zero; significance was tested using a two-tailed t-test on an approximation of the Student's t distribution equated by $abs(r((n-2)/(1-r^2))^{0.5})$ where r is the Spearman rank correlation coefficient and n the number of half-lives per compounds. Due to the multidimensional problem, the experimental parameters were also used to build multiple linear regression models. Selection of the most relevant descriptors was performed with the Correlation-based Feature Subset Selection (CFS) algorithm [45] implemented in Weka 3.8.1 [46]. The algorithm takes into account the usefulness of the individual parameters for predicting the DT50 together with the level of intercorrelation among them. The experiments were carried out using the AttributeSelectedClassifier routine of Weka with the CfsSubsetEval option for evaluator and BestFirst or LinearForwardSelection options for search. The final set of parameters selected to build the model will be in principle also the most relevant to explain the transformation of compounds across a structural class of compounds and transformation reaction.

Results and Discussion

Relevant characteristics of Eawag-Soil data

In its current form, the *Eawag-Soil* package contains 282 degradation pathways, 1535 reactions (excluding reactions leading to CO_2 through an unknown sequence of reactions) and 1619 compounds (282 parent pesticides and agriculture-related compounds and 1337 biotransformation products). Of these 1619 compounds, 777 (282 parent compounds and 495 biotransformation products) have at least one associated half-life value. Since multiple half-lives may be available for individual compounds, the *Eawag-Soil* package altogether contains 4716 biotransformation half-life values with corresponding scenarios. These numbers will increase over time as the package is being further developed. The size of the *Eawag-Soil* package in terms of numbers of pathways, reactions and compounds lies in a similar range as the current size of the *Eawag-BBD* package (i.e., 219 pathways, 1503 reactions, 1396 compounds). Introducing it thus not only doubles the amount of biotransformation pathway information to learn QSBRs and pathway prediction models from, but also extends the chemical space covered from mostly legacy chemicals (i.e., persistent organic pollutants and a few pesticides with long and extensive usage history) to modern, polar, and structurally more complex pesticide active ingredients (see section on Chemical Space Analysis for a detailed discussion).

A descriptive statistical analysis of the entire data set was performed. In the pre-processing of the data set, values that seemed to be physically implausible based on the frequency distribution of the parameters and on our knowledge about the different soil properties and their ranges were removed. For example, values for the three *soil texture* parameters, i.e., % *sand*, *silt* and *clay*, were removed if the sum of the three parameters was higher than 100%, for *humidity* parameter values higher than 100% were removed (5 values removed), or for soil *organic content* parameter values higher than 10 g OC/100g soil were removed (12 values removed). In general, only a few values per parameter were removed due to this analysis, corresponding to less than 1% of the values for most part of the parameters.

For the experimental parameter *organic content* the soil organic content reported as soil organic matter (OM) was transformed to organic carbon (OC) using the relationship OC = OM/1.724 [47]. For each parameter, the number of missing values was determined and the distribution of values was characterized by several statistical measures. A summary of results for DT50 and all experimental parameters is given in Table 1 and additional statistical measures and histogram plots of all parameters can be found in the SI (Table S1 and Figures S1 and S2).

The number of DT50s per compound in *Eawag-Soil* varies strongly. From the 777 compounds with associated half-lives, 113 have more than ten DT50s (each one for a set of unique experimental conditions) and more than half of the compounds, i.e., 419, have more than five DT50s (see SI Figures S3 for a plot of the frequency distribution of half-lives per compounds). Figure 2 gives the maximum, minimum and median DT50 values for all compounds in *Eawag-Soil* with at least 10 DT50 values per compound (N=113). A corresponding graph for all 777 compounds with one or more associated half-lives is given in the SI (Figure S4). The median half-lives across all of these compounds cover about three orders of magnitude, suggesting that these data should provide a valuable resource to develop QSBR models for soil half-life prediction. Also, the dataset separates out into 139 out of 777 compounds, i.e., 18%, having a median DT50 above 120 days, which is the persistence criterion for pesticides and industrial chemicals in soil [48, 49]. Using appropriate sampling procedures the dataset can be fairly balanced for the potential purpose of developing a persistence classification model from the data. The presence of half-lives values close to zero is due to the presence of some rapidly degradable or volatile compounds in the data set, e.g., metam-sodium or dazomet. On the opposite side, half-lives above 1000 days indicate the presence of stable compounds in the data set, e.g., flutriafol and butralin, where the DT50 values are extrapolated well beyond the study duration for most cases. For these cases of extreme behavior, the DT50 values should be considered merely approximate values of the behavior of the respective pesticides in soil.

Another important aspect that can be learned from the DT50 data shown in Figure 2 is the large variability of DT50 values for individual compounds observed across different experimental conditions. Considering the maximum and minimum half-lives, 244 out of 777 compounds, i.e., 31%, show a variability in the DT50 values of two orders of magnitude or more. This shows that the consideration of the experimental parameters is crucial for the development of QSBR models with improved accuracy for the prediction of soil half-lives. The extensive collection of metadata made available in *Eawag-Soil* should provide a useful resource for this purpose.

Table 1. Summary statistic of *DT50* and experimental parameters *% sand*, *% silt*, *% clay*, *pH*, *temperature*, *water storage capacity*, *% humidity*, *organic content* (*OC*), *cation exchange capacity* (*CEC*), *bulk density*, *biomass start*, *biomass end* and *spike concentration*.

Parameter	**Dimension**	**N**		**Mean**	**Median**	**Std. Deviation**	**Minimum**	**Maximum**
		Valid	**Missing**					
DT50	days	4716	0	88.8	24.0	206	0.003	3690
% Sand	–	4108	608	51.9	55.0	24.6	0.00	99.0
% Silt	–	4114	602	32.0	28.0	18.6	0.00	88.6
% Clay	–	4132	584	16.0	12.6	11.0	0.10	94.3
Acidity, pH	–	4641	75	6.62	6.70	0.888	3.60	8.80
Temperature	°C	4659	57	20.1	20.0	3.50	1.00	49.0
Water Storage Capacity	g water / 100 g dry soil	3998	718	40.2	38.9	18.1	1.54	137
% Humidity	–	4350	366	55.4	45.0	21.6	5.00	100
OC	g OC/100 g soil	4540	176	1.81	1.62	1.08	0.02	10.0
CEC	mEq/100 g soil	3785	931	14.0	12.2	7.92	1.20	60.0
Bulk Density	g/cm^3	1741	2975	1.34	1.40	0.282	0.00	2.66
Biomass Start	µg C/g	3147	1569	408	312	360	1.00	2445
Biomass End	µg C/g	2651	2065	345	257	339	0.05	2452
Spike Concentration	mg/kg dry soil	3917	799	1.28	0.400	2.55	0.003	25.0

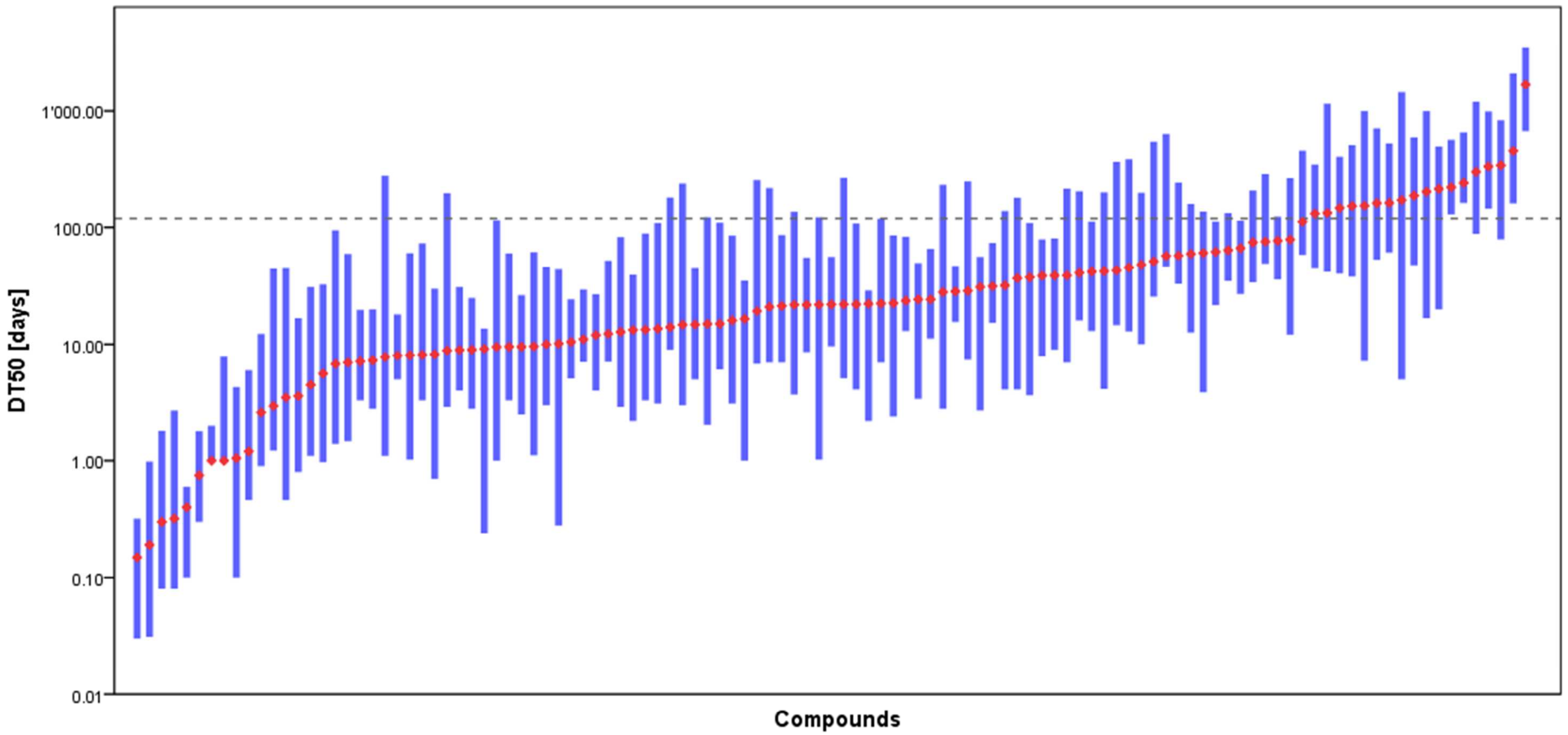


Figure 2: Median DT50 values (red diamonds) and DT50 distributions (minimum to maximum) for 113 compounds with more than 10 associated DT50 values in *Eawag-Soil* (data used to build the Figure is available in Table S2 in SI). The dashed line indicates a persistence criterion in soil of 120 days.

Regarding the collection of metadata, missing values may pose a problem in data analysis. The analysis of missing values in Table 1 shows that for 9 of the 13 numeric experimental parameters the number of missing values is reasonably low and varies between 1.2% for *% humidity* and 17% for *spike concentration*. The *soil texture*, a categorical parameter not shown in Table 1, is another parameter with a small number of missing values, i.e., only 134 out of 4716, or 3% (see Figures S1 in the SI for a frequency plot of the distribution of the 12 soil textural classes). The *water storage capacity* only shows 8% of missing values, but will need some further pre-processing to be used for modeling purposes. First, the *water storage capacity* is a property of the soil that could per se have an influence on observed degradation. However, it has been measured under slightly varying water tensions and therefore needs to be harmonized to one set of conditions. Second, the *water storage capacity* does not directly describe the experimental moisture content of the soil, which also potentially influences degradation. The latter could be obtained from multiplying the *water storage capacity* with the *% humidity*.

Considerably more problematic in terms of missing values are the *cation exchange capacity*, *bulk density*, *biomass start* and *biomass end* parameters, with 20%, 63%, 34% and 44% of missing values, respectively. Bulk density shows the highest number of missing values, with 2975 scenarios out of 4716 not containing any information on it. Preliminary experiments showed that the missing bulk densities could be imputed using the known relation between *bulk density*, soil texture and *organic content*. A k-nearest neighbor model trained using the available *bulk density* values and *% clay* and *organic content* as descriptors yielded results with a mean absolute error of 0.06 g/cm^3 in 10-fold cross validation. Similarly, *cation exchange capacity* could be imputed from *% clay*, *organic content* and *pH* with a mean absolute error of 1.12 mEq/100 g soil in 10-fold cross validation. While imputation of some parameters based on known relationships between soil properties thus seems feasible and useful for further model development, other missing values such as *biomass start* and *biomass end* will be more difficult to address in this way because of the unknown relationship between these parameters and other experimental parameters.

Overall, *Eawag-Soil* provides an unprecedentedly rich collection of half-lives and experimental parameters manually extracted and curated, which should be useful in a great variety of applications, some of which will be demonstrated in the following subsections.

Chemical space analysis

The projection of the top three principal components deduced from the structural fingerprints of the compounds in the *Eawag-BBD* and *Eawag-Soil* packages is shown in Figure 3A, and with the inclusion of the pharmaceuticals in Figure 3B. While *Eawag-Soil* overlaps with *Eawag-BBD* in some regions of the space, it clearly extends it to a point where it also allows for a better coverage of other relevant classes of compounds, e.g., the set of pharmaceuticals included for illustrative purposes (Figure 3B).

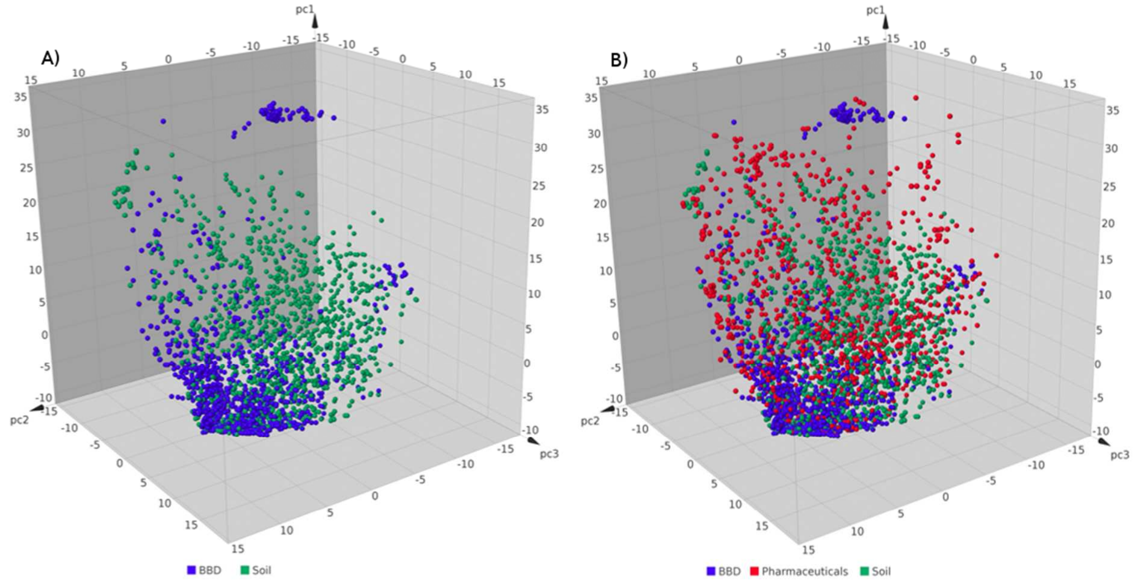


Figure 3. Projection of the top three principal components of A) *Eawag-BBD* and *Eawag-Soil*, and B) *Eawag-BBD*, *Eawag-Soil* and pharmaceuticals set.

To quantitatively confirm the results shown in Figure 3, an analysis was also carried out based on one-class SVM for outlier detection. The ν-Parameter, which limits the number of outliers in the training dataset, was initially set to a default value of 2%. To confirm that this value was reasonable, the 29 compounds identified as outliers in the *Eawag-BBD* package were visually checked and confirmed as reasonable outliers. The outlier analysis indeed showed that the combined set of *Eawag-BBD* and *Eawag-Soil* compounds covers a wider area of the relevant chemical space, compared to using only the *Eawag-BBD* compounds. First, the outliers for the set of parent pesticides from *Eawag-Soil* showed a reduction from 134 to 6 (Table 2). Similarly, the number of outliers for all compounds in *Eawag-BBD* and *Eawag-Soil* combined showed a reduction from 590 to 61 (Table 2). These results confirm that compounds such as pesticides would have been badly covered by the *Eawag-BBD* dataset alone and, at the same time, can be considered as an internal validation of the outlier detection method. More importantly, the number of outliers for the set of pharmaceuticals is reduced by 70% (152 instead of 515) when adding the *Eawag-Soil* dataset to the *Eawag-BBD* dataset. Three examples of pharmaceuticals that were outliers in *Eawag-BBD* but are considered inside the combined chemical space of *Eawag-BBD* and *Eawag-Soil* are shown, together with their corresponding three nearest neighbors, in Table S3 of the SI. With the inclusion of the *Eawag-Soil* package, 85% of the pharmaceuticals are now covered by the chemical space of the compounds in enviPath. Based on the combination of the *Eawag-BBD* and *Eawag-Soil* packages, it should therefore become possible to train models with a significantly enlarged application domain, and hence strongly increased prediction accuracy and reliability for structurally complex and polar compounds, which are of particular concern for water quality [50].

Table 2. Number of outliers detected in the different datasets.

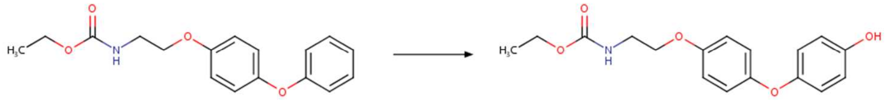
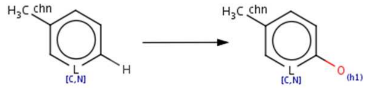
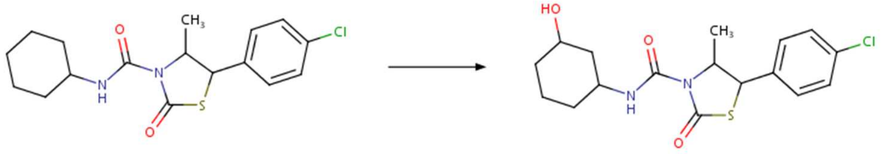
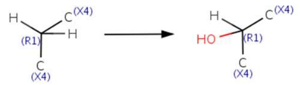
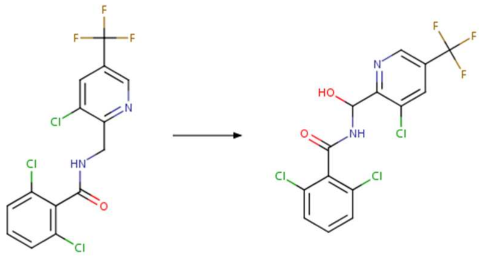
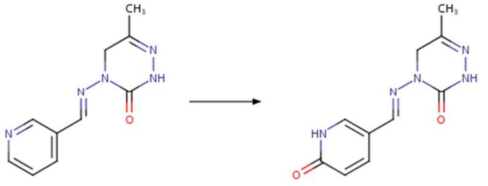
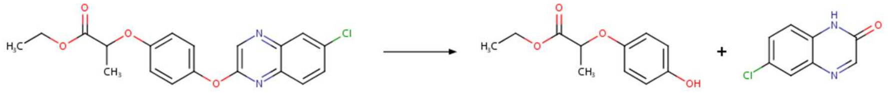
	Number of compounds detected as outliers			
	Eawag-BBD (N = 1399)	*Eawag-BBD* and *Eawag-Soil* (N = 3018)	Parent Pesticides in *Eawag-Soil* (N =280)	Pharmaceuticals (N = 1006)
Eawag-BBD[a]	29	590	134	515
Eawag-BBD & Eawag-Soil[a]	21	61	6	152

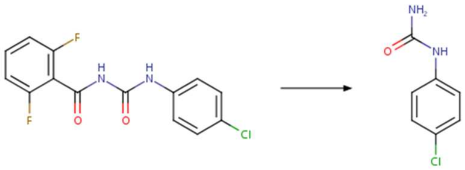
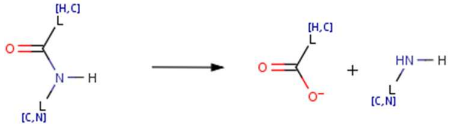
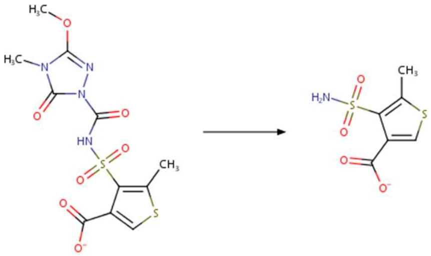
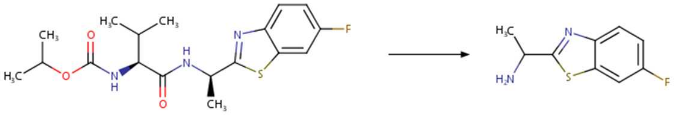
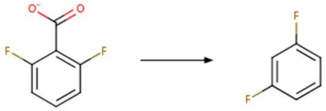
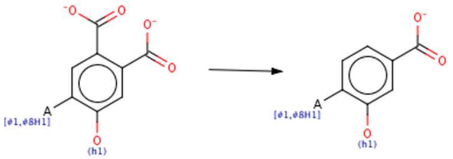
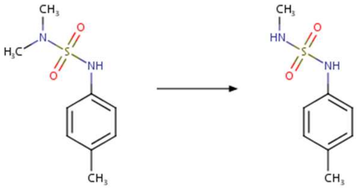
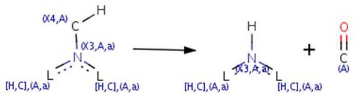
[a]Compound sets used to define the chemical space

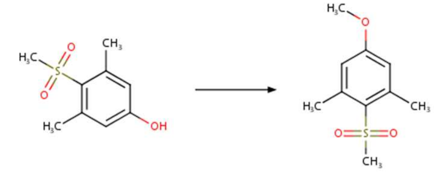
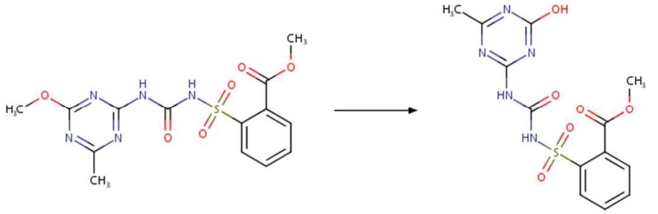
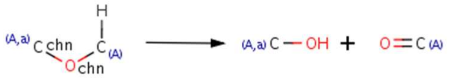
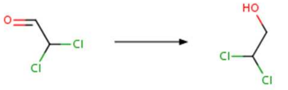
Missing rules analysis

From the set of 1535 experimental reactions, 711 reactions (i.e., 46%) are not predicted by the Eawag-PPS system over a three-generation prediction cycle. Thus, the current sensitivity of Eawag-PPS to predict transformation of the compounds contained in *Eawag-Soil* is only 54%. For the 711 reactions not predicted, the respective type of transformation might be missing completely, or corresponding rules exist but their specificity does not fully cover the substrate spectrum of compounds in *Eawag-Soil*. Table 3 shows the eight reaction types that contain at least eight reactions that were not predicted by Eawag-PPS. Together they cover 20% of the 711 reactions not predicted by Eawag-PPS. Another 253 reactions were classified into 48 more reaction types or combination of reaction types of smaller size. 315 reactions could so far not be classified at all, amongst other reasons because of incompletely documented biotransformation pathways with missing intermediates and improbable reported structures of intermediates. These cases will need further attention, but might not be fully resolvable.

Table 3: Most populated reaction types not predicted by Eawag-PPS with example reactions.

Reaction Type	Subclass of Reaction Type	N. #	Example reaction	Related *btrule*
Hydroxylation (52 reactions)	Hydroxylation in (hetero)aromatic ring	33		bt0013 
	Hydroxylation in aliphatic ring	7		bt0242 
	Hydroxylation in alpha-position to allyl/aryl/carbonyl group	6		
	Hydroxylation in (hetero)aromatic ring followed by keto enol tautomerism	6		
Scission of aryl-heteroaryl ether bond		19		

Amide hydrolysis (16 reactions)	(Cyclic) N-acylurea hydrolysis	5		bt0067 
	(Cyclic) Sulfonyl urea hydrolysis	7		
	Aliphatic amide hydrolysis	4		
Decarboxylation		13		bt0051 
N-dealkylation		11		bt0063 

Ar-OH methylation		9		
O-dealkylation		8		bt0023 
Reduction of ketone to alcohol		8		

The most populated reaction types cover hydrolysis, oxidative N- and O-dealkylations, decarboxylations, reductions and also addition reactions. For most of these reaction types, i.e., hydroxylations, amide hydrolyses, decarboxylations, and oxidative N- and O-dealkylations, similar *btrules* already exist but are too specific to predict the observed reactions. For instance, they may be too specific in the definition of the neighborhood atoms of the reaction center. This is case for rules bt0011, bt0012, bt0013 and bt0014, which predict hydroxylations of monosubstituted benzene and pyridine rings in *o*-, *m*-, and *p*-position, but do not cover multiply substituted aromatic rings and N-heteroaromatic rings other than pyridine, which, however, are present as observed reactions in the *Eawag-Soil* dataset. Another case of existing rules being too specific are rules that have restrictions related to the presence of specific functional groups in the molecule. Here, hydroxylations in aliphatic rings and in alpha position to allyl, aryl and carbonyl groups are an interesting example since the existing rule bt0242 handles hydroxylation of secondary aliphatic carbon atoms in a ring, adjacent to a carbon that is sp2 hybridized, or bound to N or O, and should thus cover the observed reactions. However, rule bt0242 has been prevented from acting on these compounds due to "functional group restrictions" that state that the rule should not act on esters and amides, which, however, are present in the substrates of the respective, not predicted reactions. In these and similar cases, existing rules should be extended to cover the structural patterns observed in the *Eawag-Soil* reaction set, and functional group restrictions removed where they are in contradiction to evidence from *Eawag-Soil*. The latter point suggests itself even more since approaches have been developed to learn relative priorities between rules (e.g., between hydroxylation and amide or ester hydrolysis) from the data and implement them in terms of relative reasoning models, rather than hardcode them into the rules [25, 27, 28].

Of the most populated reactions in Table 3 only the scission of the aryl-heteroaryl ether bond, Ar-OH methylation and the reduction of the ketone group are not covered by any existing rule. The scission of the ether bond is an interesting case because the ether linkage is a common feature that, due to the high energy of the ether bond, confers stability and consequently renders these compounds typically rather resistant to microbial degradation. However, abiotic hydrolysis data [36, 51] demonstrate that at least some of these aryl-heteroaryl ether bonds can be hydrolyzed at acidic pH, resulting in the formation of a phenol and a 2-pyridone derivative. This must be due to the presence of the N-atom in the aromatic heterocycle, which renders it more e-deficient and draws electron density from the carbon atom that is attached to the ether bridge, making it more vulnerable for nucleophilic attack. These considerations suggest that the observed scission of aryl-heteroaryl ether bonds is due to hydrolysis, but based on the data available in *Eawag-Soil* it remains difficult to judge whether it is a purely abiotic or enzyme-catalyzed hydrolysis.

The Ar-OH methylation is another interesting case since it is an addition reaction. Methylation of phenol to anisol is a common transformation in many biosynthetic pathways (e.g., lignol, hormone, and flavonoid biosynthesis) and may be catalyzed by enzymes from the methyltransferase class (EC numbers of class 2.1.1.-). The substrates of six out of the nine Ar-OH methylations in *Eawag-Soil* are halogenated substituted phenols. Therefore, likely candidate enzymes that could perform this type of biotransformation are EC 2.1.1.136, a halogenated phenol O-methyltransferase that acts on mono-, bi- and trichlorophenols, and EC 2.1.1.25, a phenol O-methyltransferase that acts on a wide variety of simple alkyl-, methoxy- and halophenols. Interestingly, EC 2.1.1.136 has so far only been found to occur in fungi, which suggests that the *Eawag-Soil* data set might also highlight some fungal transformations that are only scarcely covered in *Eawag-BBD* and hence not only extend the coverage of enviPath towards new types of compounds but also other types of catalyzing enzymes and microbial organisms. There are also other addition reactions such as formylations, acetylations or conjugations with more complex groups that are increasingly observed in microbial communities [52-54], and, at least for the case of N-formylation (1 reaction) and N-acetylation (3 reactions), have also been observed in *Eawag-Soil*. Addition reactions have typically not been implemented in pathway prediction systems so far because their focus was on catabolic reactions. As a consequence, as is the case for Ar-OH methylation, a *btrule* for the reverse reaction, i.e., the oxidative O-dealkylation, often exists. Therefore, if addition reactions were to be implemented in the future, care has to be taken that the system does not predict any products from a given substrate that are identical with any of its precursor compounds in the pathway to avoid “dead cycles” in the prediction. The same is true if rules for reductions such as the reduction of ketones to secondary alcohols (Table 3) were to be implemented. Since these are more likely to proceed under anaerobic conditions, assigning them a low aerobic likelihood within the Eawag-PPS system [55] could further restrict the application of such rules.

Exploring half-life variability

As discussed in a previous section, median half-lives across all pesticides cover about three orders of magnitude, yet variability in half-lives for individual pesticides also spans about two orders of magnitude. Thus, improved QSBR models need to account for both inter- and intra-compound variability in half-lives. This should become more achievable if mechanistic understanding about the fate of the pesticides in soil and about the ongoing transformation processes is included in the model development as much as possible. Here, we therefore explore whether the consideration of initial transformation reactions can support such an endeavor. The underlying hypothesis is that if a set of structurally similar compounds, i.e., from the same pesticide class, showed the same initial transformation reaction, this transformation is likely catalyzed by the same type of enzyme. The two ensuing hypotheses, for which we did some initial exploration here, are *(i)* that inter-compound half-life variability is considerably smaller within compounds that belong to the same pesticide class and undergo the same transformation than across all compounds, and *(ii)* that intra-compound half-life variability shows similar, characteristic dependencies on environmental conditions across compounds that belong to the same pesticide class and undergo the same transformation. Hypothesis (*ii*) is restricted to compounds belonging to the same pesticide class because we assume that direct effects of the environmental conditions on their bioavailability and abiotic stability are consistent within a class of pesticides but not necessarily across classes.

We explored these two hypotheses for all amide pesticides in the *Eawag-Soil* package. Amides were selected because they constitute a large class in the *Eawag-Soil* package (i.e., 40 out of 282 parent compounds are classified as amides according to ref [56]). The class also contains several compounds that have particularly large numbers of half-lives and scenarios associated with them (i.e., 10 half-lives/amide on average compared to 6 half-lives/compound on average across all of *Eawag-Soil*). The amides were then further grouped into consistent sub- and subsubclasses, first according to ref [56] and later through further manual curation. Finally, every amide was annotated manually with its initial transformation reaction(s) according to the pathway maps in the *Eawag-Soil* package. This resulted in three amide subsubclasses (sulfonamides, chloroacetanilides, anilides with N-substituted pyrazole ring) that contained four or more structurally similar compounds undergoing the same type of initial transformation reaction (Table 4). All other subsubclasses were either smaller or their members underwent different initial transformation reactions. In the following, the three groups in Table 4 form the basis for testing hypotheses (*i*) and (*ii*).

Table 4. Initial transformation reactions and half-lives (range and median) for sulfonamides, chloroacetanilides, and anilides with N-substituted pyrazole ring.

Pesticide class	Compound	Initial transformation reaction	Median DT50 [days]	DT50 range [days]	Number of DT50 values
Sulfonamides	Penoxsulam	O-demethylation	24.5	15-137	7
	Pyroxsulam	O-demethylation	3.6	1-17	25
	Florasulam	O-demethylation	3.5	0-45	17
	Metosulam	O-demethylation	9.15	4-25	4
	Asulam [a]	*Several, O-demethylation not possible*	*3.89*	*3-10*	*5*
	Oryzalin [a]	*Several, O-demethylation not possible*	*182*	*63-468*	*8*
Chloroacetanilides	Acetochlor	Substitution with GSH & hydrolysis; reductive dechlorination	10.2	0.28-44	50
	Dimethachlor	Substitution with GSH; reductive dechlorination; others	7.15	3.31-19.8	12
	Dimethenamide	Substitution with GSH & hydrolysis	13	7.8-43.4	5

	Metazachlor	Substitution with GSH; other	13.6	3.1-109	49
	Propisochlor	Substitution with GSH & hydrolysis	11.4	8.4-40.2	5
Anilides with N-substituted pyrazole ring	BAS 700	N-dealkylation; amide hydrolysis; other	326	72.7-810	6
	Benzovindiflupyr	N-dealkylation; amide hydrolysis; hydroxylation	550	349-1000	7
	Bixafen	N-dealkylation; amide hydrolysis	>365	n.a.	4 [b]
	Isopyrazam	N-dealkylation; amide hydrolysis; hydroxylation	231	29.8-976	9
	Penflufen	Hydroxylation	231	117-434	6
	Penthiopyrad	Hydroxylation; N-dealkylation; amide hydrolysis	146	60.5-413	6
	Sedaxane	N-dealkylation; other	74.2	57.6-138	8

[a] These two sulfonamides were included to demonstrate the effect of adding structurally similar compounds that undergo different initial transformation reactions.

[b] For all four scenarios, half-lives were given as >365 d.

The median half-life data given in Table 4 and the distribution of individual half-lives of the three groups as compared to the entirety of all amides shows that the distribution of half-lives for the chloroacetanilides and the sulfonamides are overlapping, whereas the median half-life distribution for the anilides is quite distinct. More importantly, however, the distributions for the three groups in Table 4 are more narrow than the half-life distribution for all amides (Figure 4), which is also demonstrated by the coefficients of variation (CV) of the median half-lives, which are 0.97, 0.23, and 0.64 for the sulfonamides, the chloroacetanilides and the anilides, respectively, as compared to 1.24 for all amides. For demonstration purposes, we also considered half-life data for the remaining two sulfonamides in the *Eawag-Soil* package (i.e., asulam and oryzalin), which, however, do not contain the methoxypyrimidine moiety that is subject to O-demethylation in penoxsulam, pyroxsulam, florasulam and metosulam and therefore undergo different initial transformation reactions. This resulted in a CV across all six sulfonamides of 1.89. Altogether, these observations lend some support to hypothesis *(i)* in that they demonstrate smaller inter-compound variability within groups of compounds undergoing the same initial transformation reaction amongst the class of amide pesticides.

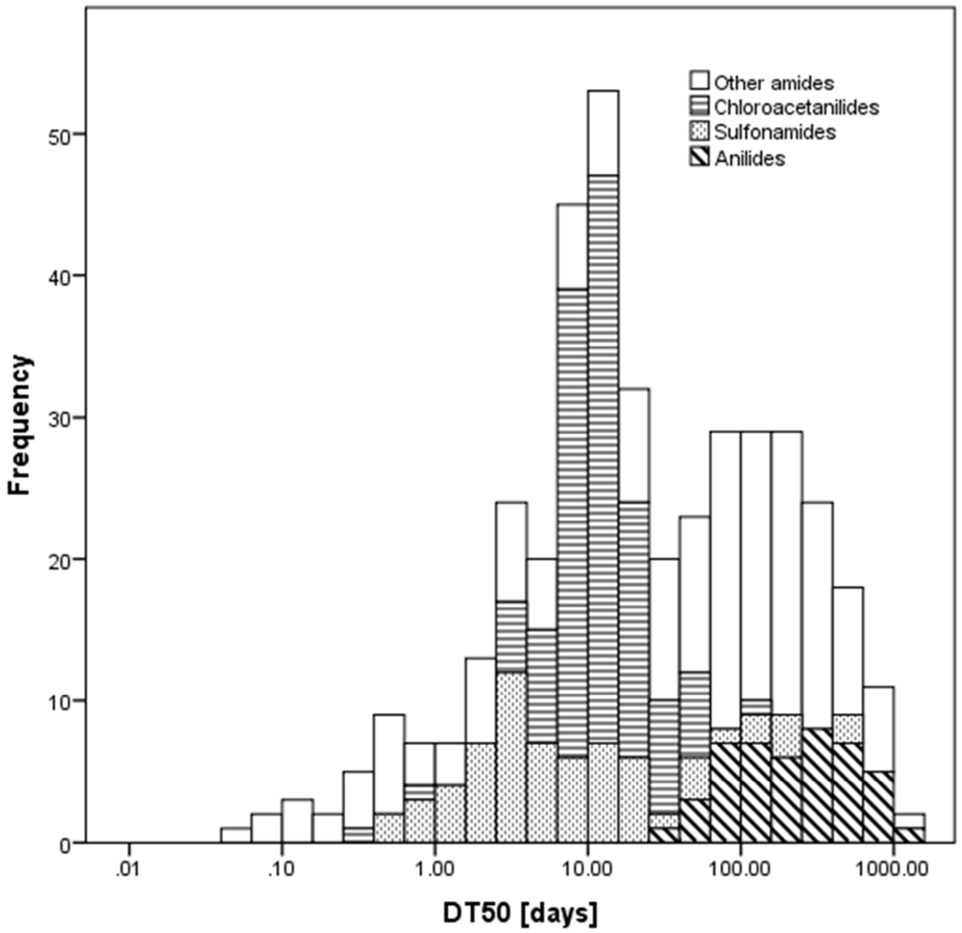


Figure 4. Half-life distribution (frequency plot) for chloroacetanilides, sulfonamides, anilides with N-substituted pyrazole ring, and remaining compounds classified as amides in the *Eawag-Soil* package.

To explore hypothesis (*ii*), first, Spearman rank correlation coefficients for the relationship between the DT50 values of the six sulfonamides and nine experimental parameters were calculated (Table 5). For some combinations of compounds and parameters the analysis was not possible because of missing values. A consistent negative relationship with temperature was observed (albeit only significant for one compound), which could be expected because of previous reports of Arrhenius-type dependence of pesticide soil degradation [57]. Only few additional univariate relationships were found that are significant at the 5% level. The case of OC seems most interesting because contradictory, yet significant dependencies are found for pyroxsulam (negative dependence, O-demethylation) and oryzalin (positive dependence, transformations other than O-demethylation). Additionally, a close to significant negative dependence on OC was also found for florasulam (O-demethylation). These results lend some support to the hypothesis that dependencies on environmental conditions are more similar if compounds undergo the same initial transformation reaction.

Table 5. Spearman rank correlation coefficients for relationships between DT50 values of six sulfonamides and selected experimental parameters. Significance of rank correlation coefficients are given in parenthesis and coefficients that are significant at the 5% significance level are highlighted in bold.

	Log(DT50)					
	Penoxsulam N=7	Pyroxsulam N=25	Florasulam N=17	Metosulam N=4	Asulam N=5	Oryzalin N=8
% Sand	0.213 (0.685)	0.095 (0.651)	0.204 (0.661)	-0.738 (0.262)	0.205 (0.741)	0.627 (0.070)
% Silt	-0.213 (0.685)	0.005 (0.981)	-0.204 (0.661)	0.000 (1.00)	-0.205 (0.741)	**-0.814 (0.008)**
% Clay	-0.213 (0.685)	-0.172 (0.411)	0.337 (0.460)	0.800 (0.200)	-0.308 (0.614)	-0.627 (0.070)
pH	0.273 (0.601)	**-0.503 (0.010)**	0.163 (0.727)	-0.400 (0.600)	0.718 (0.172)	-0.579 (0.102)
Temperature	**-0.845 (0.034)**	-0.283 (0.170)	-0.628 (0.131)	-	-0.707 (0.182)	-0.548 (0.127)
Organic Content	0.395 (0.438)	**-0.523 (0.007)**	-0.738 (0.058)	-1.00 (-)	-0.205 (0.741)	**0.848 (0.004)**
CEC	0.030 (0.955)	-0.312 (0.129)	0.535 (0.216)	-0.80 (0.200)	0.103 (0.869)	0.034 (0.931)
Biomass Start	0.516 (0.295)	**-0.611 (0.001)**	-0.553 (0.198)	-1.00 (-)	-0.205 (0.741)	-
Spike concentration	-	0.248 (0.232)	-	-	-	-0.402 (0.283)

Because univariate relationships are strongly confounded by the influence of all other experimental parameters on the observed half-lives, multiple linear regression models were developed to further explore the validity of hypothesis *(ii)* for the example of the sulfonamide herbicides. Considering the fact that the DT50 ranges of all the sulfonamide herbicides are in the same range with exception of the oryzalin, multiple linear regression were developed using as training set a combination of all DT50 values and corresponding scenarios for only those four sulfonamides undergoing O-demethylation (experiment 1) or for all six sulfonamides (experiment 2). In Table 6, a summary of the resulting models is given.

Table 6. Multiple linear regression models developed for DT50 values of sulfonamides.

			Training			Test (10-fold cross validation)		
Compounds	N desc	Desc. Selected	R	MAE	RMSE	R	MAE	RMSE
Penoxsulam Florasulam Metosulam Pyroxsulam (N=53)	5	pH; T; OC; CEC; Biomass	0.813	0.252	0.315	0.729	0.312	0.377
Penoxsulam Florasulam Metosulam Pyroxsulam Oryzalin Asulam (N=66)	4	T; CEC; Biomass; Spike Concentration	0.743	0.371	0.476	0.627	0.426	0.563

After parameter selection, the final MLR model for experiment 1 yielded the following equation:

$$Log(DT50) = -0.166 * pH - 0.0467 * T - 0.166 * OC + 0.0249 * CEC - 0.0005 * Biomass\ Start + 2.75$$

The final MLR model had a mean absolute error of 0.312 (corresponding to roughly a factor of two) in 10-fold cross validation and showed an only minor decrease of R between training and cross-validation, indicating that the data were not over fitted by the model. Also, at least two of the observed dependencies are plausible based on our understanding of the fate of pesticides in soil: The temperature-dependence again follows the logic of an Arrhenius relationship, and the negative dependence on biomass follows the logic of a second-order rate constant that depends on biomass and compound concentration. The fact that a model could be built that encompassed all four sulfonamides in experiment 1 and that remained robust in rigorous 10-fold cross-validation, without the need to include information on the structure of the compound or other molecular descriptors, clearly support both hypotheses *(i)* and *(ii)*. Hypothesis *(ii)* is further supported by the fact that the regression model from experiment 2 performed worse than for experiment 1, suggesting that adding data for structurally similar compounds that, however, undergo a different type of transformation weakens the observed dependences on experimental parameters. The finding that experiment 2 performs worse than experiment 1 still holds true when the half-lives for each compound are z-normalized to account for the different half-life range of oryzalin compared to the other sulfonamides (i.e., R values of 0.680 and 0.648 are obtained for training the model on only those sulfonamides containing the methoxypyrimidine moiety and on all six sulfonamides, respectively). Overall, the most encouraging outcome from this exploration of half-life variability is that for groups of structurally similar compounds undergoing the same transformation models can be built that capture a relevant part of the observed variability (i.e., >60%).

Conclusions & outlook

In this article we presented the *Eawag-Soil* package as a novel biotransformation data package made available through the enviPath environment. *Eawag-Soil* contains a comprehensive collection of all freely accessible regulatory data on pesticide degradation in laboratory soil simulation studies under aerobic conditions for pesticides registered in the EU. This data resource has been developed in order to respond to the need for more environmentally relevant training data sets to develop models for the microbial biotransformation of polar, structurally complex trace organic contaminants such as pesticides and pharmaceuticals. An analysis of the chemical spaces covered by the existing *Eawag-BBD* dataset and *Eawag-Soil* confirmed a strongly improved coverage of these types of chemicals, suggesting that through the combination of the *Eawag-BBD* and *Eawag-Soil* packages it should become possible to train models with an increased prediction accuracy and reliability for structurally complex and polar compounds.

We have further explored two lines of research that can greatly profit from the data in *Eawag-Soil*: *(i)* the formulation of new rules or adaptation of existing rules to obtain a better coverage for the prediction of soil biotransformation of structurally complex trace organic contaminants, and *(ii)* the elucidation of the dependency of observed half-life variability on the study conditions as expressed by the experimental parameters. Based on the analysis of missing rules, eight examples of reaction types were presented that should trigger the formulation of new biotransformation rules, e.g., Ar-OH methylation, or the extension of existing rules e.g., hydroxylation in aliphatic rings. The exploration of the half-lives of different amide pesticides not only showed that different subsubclasses of structurally similar amides have significantly different median half-lives, but also yielded some first evidence that the consideration of initial transformation reactions within groups of structurally similar amides seems to support a more accurate description of how half-lives depend on environmental conditions. Based on these results, we argue that the consideration of the type of initial transformation reactions in the development of QSBRs should greatly facilitate the consideration of the influence of experimental parameters on half-lives in such models. Doing so is a novel opportunity offered by the simultaneous encoding of transformation reactions and corresponding half-lives in *Eawag-Soil*. Ultimately, a combined pathway prediction system could be developed where the reactivity pattern of the compound (as encoded by an extended set of btrules) is used as one type of descriptor in combination with molecular descriptors and experimental conditions to predict half-lives. To work towards this end, a more complete analysis of the reactions not predicted by Eawag-PPS will be sought, and an automated procedure for defining new rules based on a chemoinformatics approach for semi-automatic analysis and assignment of reaction types will be implemented.

Overall, *Eawag-Soil* makes an unprecedented amount of manually curated soil biotransformation information available to the public in an easily accessible manner. This should not only be of high interest for researchers developing QSBR-type models and pathway prediction systems, but also for regulators and the general public as an information resource.

Acknowledgments

We would like to thank six student assistants (Lukas Allemann, Oskar Jönsson, Cyrill Kern, Julian Kornprobst, Adelene Lai and David Stadler) from ETH Zürich who electronically encoded and quality checked the information in the *Eawag-Soil* package. We would further like to thank three student assistants (Tim Lorsbach, Steven Lang and Kevin Kaster) from the Johannes Gutenberg University Mainz for writing code to accommodate pesticide data-specific features in enviPath. We acknowledge financial support from the European Research Council under the European Union's Seventh Framework Programme (ERC grant agreement no. 614768, PROduCTS), and from the Swiss National Science Foundation (project no. CR22I2L 149711).

References

1. D. M. Cwiertny, S. A. Snyder, D. Schlenk and E. P. Kolodziej, *Environmental Science & Technology* **2014,** *48*, 11737-11745.

2. B. Escher and K. Fenner, *Environmental Science & Technology* **2011,** *45*, 3835 - 3847.

3. A. B. A. Boxall, C. J. Sinclair, K. Fenner, D. Kolpin and S. J. Maud, *Environmental Science & Technology* **2004,** *38*, 368A-375A.

4. ECHA *Guidance on information requirements and chemical safety assessment, Chapter R.7b: Endpoint specific guidance*; European Chemicals Agency, Helsinki, Finland, 2016.

5. C. Rücker and K. Kümmerer, *Green Chemistry* **2012,** *14*, 875-887.

6. L. Mamy, D. Patureau, E. Barriuso, C. Bedos, F. Bessac, X. Louchart, F. Martin-Laurent, C. Miege and P. Benoit, *Critical Reviews in Environmental Science and Technology* **2015,** *45*, 1277-1377.

7. M. Pavan and A. P. Worth, *QSAR Comb. Sci.* **2008,** *27*, 32-40.

8. S. Banerjee, P. H. Howard, A. M. Rosenberg, A. E. Dombrowski, H. Sikka and D. L. Tullis, *Environmental Science & Technology* **1984,** *18*, 416-422.

9. D. F. Paris and N. L. Wolfe, *Appl. Environ. Microbiol.* **1987,** *53*, 911-916.

10. Y. Urushigawa, S. Masunaga and Y. Yonezawa, *Water Sci. Technol.* **1988,** *20*, 459-461.

11. OECD *guidelines for testing of chemicals; 301, Ready Biodegradability* OECD: Paris, 1992.

12. R. S. Boethling, P. H. Howard, W. Meylan, W. Stiteler, J. Beauman and N. Tirado, *Environmental Science & Technology* **1994,** *28*, 459-465.

13. S. Vorberg and I. V. Tetko, *Mol. Inf.* **2014,** *33*, 73-85.

14. K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini and V. Consonni, *J. Chem Inf. Model.* **2013,** *53*, 867-878.

15. D. Aronson, R. S. Boethling, P. H. Howard and W. Stiteler, *Chemosphere* **2006,** *63*, 1953-1960.

16. K. Fenner, S. Canonica, B. I. Escher, L. Gasser, S. Spycher and H. C. Tulp, *Chimia* **2006,** *60*, 683-690.

17. R. Kuhne, R. U. Ebert and G. Schuurmann, *QSAR Comb. Sci.* **2007,** *26*, 542-549.

18. Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto and M. Kanehisa, *Nucleic Acids Res.* **2010,** *38*, W138-W143.

19. S. Dimitrov, T. Pavlov, N. Dimitrova, D. Georgieva, D. Nedelcheva, A. Kesova, R. Vasilev and O. Mekenyan, *SAR QSAR Environ. Res.* **2011,** *22*, 719-755.

20. S. D. Finley, L. J. Broadbelt and V. Hatzimanikatis, *Biotechnology and Bioengineering* **2009,** *104*, 1086-1097.

21. L. B. M. Ellis, J. Gao, K. Fenner and L. P. Wackett, *Nucleic Acids Res.* **2008,** *36*, W427-W432.

22. J. Gao, L. B. M. Ellis and L. P. Wackett, *Nucleic Acids Res.* **2010,** *38*, D488-D491.

23. V. de Lorenzo, *Current Opinion in Biotechnology* **2008,** *19*, 579-589.

24. Eawag Biocatalysis/Biodegradation Database. http://eawag-bbd.ethz.ch (accessed November 2016).

25. J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D. Latino, S. Kramer and K. Fenner, *Nucleic Acids Res.* **2016,** *44*, D502-D508.

26. enviPath – The environmental contaminant biotransformation pathway resource. https://envipath.org (accessed November 2016).

27. K. Fenner, J. F. Gao, S. Kramer, L. Ellis and L. Wackett, *Bioinformatics* **2008,** *24*, 2079-2085.

28. J. Wicker, K. Fenner, L. Ellis, L. Wackett and S. Kramer, *Bioinformatics* **2010,** *26*, 814-821.

29. G. D. Bending, S. D. Lincoln and R. N. Edmondson, *Environmental Pollution* **2006,** *139*, 279-287.

30. D. E. Helbling, D. R. Johnson, M. Honti and K. Fenner, *Environmental Science & Technology* **2012,** *46*, 10579-10588.

31. T. Kaeberlein, K. Lewis and S. S. Epstein, *Science* **2002,** *296*, 1127-1129.

32. A. Kowalczyk, T. J. Martin, O. R. Price, J. R. Snape, R. A. van Egmond, C. J. Finnegan, H. Schafer, R. J. Davenport and G. D. Bending, *Ecotox. Environ. Safe.* **2015,** *111*, 9-22.

33. EFSA Draft Assessment Reports (DAR). dar.efsa.europe/dar-web/provision (accessed November 2016)

34. EU, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union* **2006,** *L 396*, 1-849.

35. R. C. Kolanczyk, P. Schmieder, W. J. Jones, O. G. Mekenyan, A. Chapkanov, S. Temelkov, S. Kotov, M. Velikova, V. Kamenska, K. Vasilev and G. D. Veith, *Regul. Toxicol. Pharmacol.* **2012,** *63*, 84-96.

36. PPDB: Pesticide Properties DataBase. http://sitem.herts.ac.uk/aeru/ppdb/en/index.htm (accessed November 2016).

37. S. Dimitrov, D. Nedelcheva, N. Dimitrova and O. Mekenyan, *Science of the Total Environment* **2010,** *408*, 3811-3816.

38. Eawag-Soil package. https://envipath.org/package/5882df9c-dae1-4d80-a40e-db4724271456 (accessed December 2016).

39. H. P. Singer, A. E. Wössner, C. S. McArdell and K. Fenner, *Environmental Science & Technology* **2016,** *50*, 6698-6707.

40. T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem Inf. Model.* **2015,** *55*, 460-473.

41. D. Rogers and M. Hahn, *J. Chem Inf. Model.* **2010,** *50*, 742-754.