Holistic Evaluation of Biodegradation Pathway Prediction: Assessing Multi-Step Reactions and Intermediate Products

Jason Y C Tam[1,2]*, Tim Lorsbach*[2]*, Sebastian Schmidt*[3] *and Jörg Wicker*[1,2]

[1]School of Computer Science, University of Auckland, Auckland, New Zealand
[2]enviPath, Mainz, Germany
[3]Bayer AG, Crop Science Division, Environmental Safety, Monheim am Rhein, Germany
tam@envipath.com, lorsbach@envipath.com, sebastian.schmidt1@bayer.com,
j.wicker@auckland.ac.nz

Abstract

The prediction of metabolism and biotransformation pathways of xenobiotics is a highly desired tool in environmental and life sciences. There are several systems that currently predict single transformation steps or complete pathways as series of parallel and subsequent steps. Their accuracy is often evaluated on the level of a single transformation step. Such an approach cannot account for some specific challenges that are related to the nature of the biotransformation experiments. This is particularly true for missing transformation products in the reference data that occur only in low concentrations, e.g. transient intermediates or higher-generation metabolites. Furthermore, some rule-based prediction systems evaluate accuracy only based on the defined set of transformation rules. Therefore, the performance of different models cannot be directly compared.

In this paper, we introduce a new evaluation framework that extends the evaluation of biotransformation prediction to holistically evaluating predicted pathways, taking into account multiple generations of metabolites. We introduce a procedure to address transient intermediates and propose a weighted scoring system that acknowledges the uncertainty of higher-generation metabolites. We implemented this framework in enviPath and demonstrate its strict performance metrics on predictions of in vitro biotransformation and degradation of xenobiotics in soil. Our approach is model-agnostic and can be transferred to other prediction systems. It is also capable of revealing knowledge gaps in terms of incompletely defined sets of transformation rules.

1. Introduction

There have been rapidly growing data requirements for regulatory chemical risk assessment at the European (cf. REACH [1]) and global level, as well as for the development of new products with more benign profiles. This has lead to an increased need for prediction methods of metabolism and microbial biotransformation products, along with pathways of chemical substances. Existing methods for the prediction of biotransformation products and pathways can be categorized as either knowledge-based or machine learning-based approaches. Each of the two approaches has its strengths and weaknesses. Knowledge-based approaches take into account expert knowledge on the basis of sets of transformation rules. There are a number of popular knowledge-based systems:

- METEOR for the prediction of mammalian metabolism [2].
- PathPred for predictions of enzyme-catalyzed metabolic pathways [3].
- The EAWAG Pathway Prediction System (EAWAG-PPS) for microbial biodegradation [4].
- OASIS for prediction of chemical toxicity that incorporates metabolism in processes including skin sensitization, Ames mutagenicity, formation of micronuclei and estrogen receptor binding affinity [5].

Biotransformer [6] and enviPath [7] additionally refine the probability estimates of the rules on the basis of empirical data.

enviPath [7] is a database and prediction system for biotransformation of organic environmental contaminants. The database provides the possibility to store experimentally observed biotransformation pathways, as well as biotransformation rules derived from experimental observations recorded in the literature. Machine Learning-based relative reasoning models can be built to predict probabilities of individual transformation reactions. They are constructed using sets of selected biotransformation pathways and transformation rules as training data, such as the EAWAG-BBD [4] and EAWAG-SOIL [8] packages. These models predict which of the transformation rules that are applicable to a given compound will be observed for that compound.

The performance of such models is determined by comparing the predicted products for each compound, against the associated transformation products in the experimental validation set. This approach does not take into account the position of the compound/reaction in the pathway, and problems arise when:

- Multi-step reactions are represented as a single step in the experimental data.
- Intermediate metabolites were not observed or not elucidated.
- Transformation products were incorrectly assigned to the wrong educt.
- Concentrations of downstream metabolites become too low to be observed.
- Rule-based evaluation systems fail to address observed transformations not covered by the transformation rules.

To address such issues, we propose to consider the pathway holistically when evaluating predictions produced by models.

1.1. Biotransformation Pathway Comparison

Biotransfomation pathways are derived either from experimental studies or prediction systems. They are constructed from compounds connected by reactions, and represent chemical changes through transformations. Their structures can be represented as nodes and edges in graph objects. Figure 1 presents

the *Benzyl Sulfide* pathway from the EAWAG-BBD as an example.

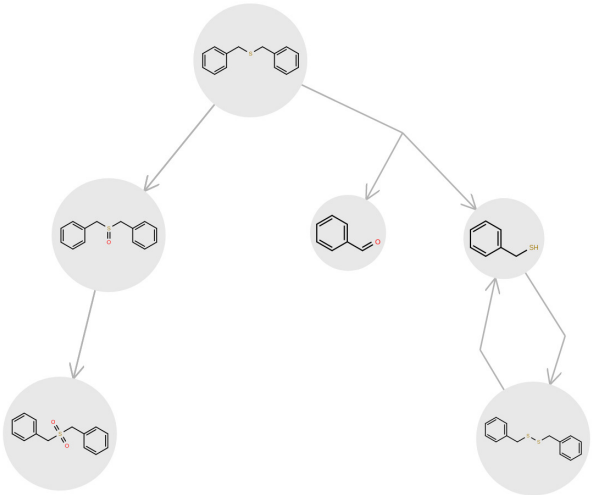

Figure 1: *The pathway* ***Benzyl Sulfide*** *from the EAWAG-BBD package.*

Experimental biotransformation or metabolism studies usually report only transformation products that were formed in high quantities, because lower concentration metabolites are considered less relevant and are more difficult to analyze and identify. Therefore, metabolites formed in low concentrations are less likely to be reported. This becomes more relevant for higher generation metabolites, because pathways typically diverge into multiple branches and transformations are not instantaneous but occur on different time scales. Both effects lead to decreasing maximum concentrations with increasing depth in the pathway. Thus, uncertainty about the actual formation of unreported metabolites increases for higher generation metabolites.

In this paper, we introduce a new *Multi-Generation* approach for evaluating pathways that addresses some of the problems of the *Single-Generation* approach. It explicitly includes the compound positions in the graph. Instead of only evaluating the reactions for each of the compounds, entire pathways are predicted and evaluated against experimentally derived validation pathways. Compounds at higher depth in the pathway bear higher uncertainty, as their likelihood of being missed increases with decreasingly lower concentrations. Our new evaluation approach takes this into consideration by assigning reduced weights to compounds at higher depths.

Another motivation is the treatment of intermediate metabolites in the pathway. These metabolites are quickly transformed to downstream products and therefore exist only in very low concentrations. As a consequence, they are often neglected or not analyzed in experimental reference pathways. *Single-Generation* evaluation approaches tend to incorrectly penalize these intermediates. However, multi-generation approaches can take them into consideration when the downstream products are known. In our approach their prediction is not penalized during the scoring process, and the depths of other downstream compounds in the pathway are adjusted accordingly.

With our new evaluation approach we can evaluate whole pathway predictions more realistically than before and independent of the underlying set of transformation rules. The results produced are more in line with the expectations of experimentalists and more comparable across models. This is because some approaches restrict the evaluation on the set of defined rules, basically ignoring the undefined transformation space. This in turn enables further improvements of the prediction models. Our methodology is a special case of graph analysis that is particularly useful for (bio)degradation or metabolism pathways and chemical reaction networks.

The main contribution of this paper are:

1. Use of conditional probabilities for depth considerations in biotransformation pathway predictions.
2. A new scoring system that quantifies the agreement between two biotransformation pathways.
3. Consideration of compound position (pathway depth) information in evaluating pathway predictions.
4. Consideration of potential intermediate metabolites in evaluating pathway predictions.

2. Background & Related Work

Biochemical Network Integrated Computational Explorer (BNICE) [9] is a framework that generates all known reactions for compounds. It makes use of the set of enzyme reaction rules based on the enzyme commission (EC) classification system. Metabolic pathways are generated by first determining functional groups contained in the root compounds, and generate associated products on ones that can be acted upon by the reaction rules. The same process is repeated on each of the products in successive generations. The iteration terminates when a limit on the number of iterations has been reached, or when no new compounds are created.

METEOR [2] contains options of knowledge based prediction methods as well as machine learning approaches. The knowledge based option utilizes a combination of *Absolute* and *Relative Reasoning* in their predictions of reactions. The process commences by applying biotransformation rules to the starting compound, and this generate potential metabolites. The absolute reasoning process then assigns a level of belief to each biotransformation [10, 11].

Biotransformations that satisfy the absolute reasoning threshold preset by the user are then ranked in the relative reasoning process. The process uses a relative reasoning threshold to calculate the resulting relative hierarchy. *Static Scores* and *Site of Metabolism Scoring* are other prediction options that make use of machine learning techniques on experimental data. The first utilizes an occurrence ratio – actual occurrences over all possible occurrences. The latter further considers similarity on additional chemical properties – attributes from generated fingerprints and molecular weights. The processes in each of these options are repeated for all surviving biotransformations, until some preset stopping conditions are satisfied, such as reaching the maximum depth.

PathPred [3] executes predictions by first searching for compounds from the KEGG [12] COMPOUND database that are similar to the chosen starting compound. The results are then used as input to search through the KEGG REACTION database for matching RDM transformation patterns [13]. These patterns are defined as KEGG atom type changes at the reaction center (R), the difference region (D), and the matched region (M). Products of these matching reactant pairs are then used as input, and this process is repeated until stopping conditions are reached. The Jaccard coefficient between the query and matched compounds of each reaction is used as

the *reaction score* to indicate its plausibility. The average of all individual reaction scores in the pathway gives the *pathway score*.

EAWAG-PPS (formerly UM-PPS) [4] performs pathway prediction by first determining the functional groups in the starting compound, and applies biotransformation rules to determine the transformed products. Applying these rules iteratively to the educts would lead to combinatorial explosion, and known pathways were used to determine biotransformation priorities [14]. User input is used at the end of each transformation prediction, to determine whether prediction continues downstream of the predicted compound(s). The predicted pathway grows as this cycle is repeated.

Biotransformer [6] combines a rule or knowledge based approach in conjunction with a machine learning approach, to predict metabolic reactions for compounds. It makes use of experimentally confirmed biotransformations derived from literature, as well as precedence rules created based on reported observations. Many of them are from the EAWAG-PPS database. The Biotransformer Metabolism Prediction Tool (BMPT) then uses a set of random forest and ensemble prediction methods to predict reactions. For example ones related to Cytochrome P450 enzymes (CYP450). Additionally, it adds a filtering phase of molecules. Metabolic pathways are built progressively with likely predicted reactions from the starting compound, one reaction at a time.

OASIS [5] predicts chemical toxicity by integrating metabolism simulators into models assessing toxicity not only of the parent chemical, but also its transformation products. This has improved model performance significantly compared to the traditional approach. Predictions were based predominantly on the analysis of the structure and properties of the toxicant and had difficulties modelling some endpoints. OASIS incorporates metabolic logic in models which accounts for enzyme interactions, channeling effects and depletion of highly reactive intermediates. It simulates metabolism using a complex mathematical model rather than a rule-based approach. The metabolism simulator uses xenobiotic pathway data from MetaPath [15] as a reference and aims to reproduce the observed pathways.

PathPred computes the Jaccard index on compounds in each of the predicted reactions, and uses the average of all such values in a pathway as the overall score. Similar integrated scoring systems that attempt to quantify the quality of predictions are not found in other systems such as METEOR, BNICE and Biotransformer. Their prediction performances in published work are obtained only via one off independent tests, without anything integrated that indicates the quality of ongoing predictions. OASIS details the prediction evaluation for metabolic pathways in their work. They union the observed and predicted pathways, and tally true/false positives/negatives by comparing the metabolites. Only the first false positive in a sequence of false positives would be penalized, because the rest are conditioned from it. The system can also identify intermediates, and an option is provided to either reward, penalize or ignore them.

Another related field is the prediction of graph networks using machine learning techniques. Link predictions is a core component in many of the different approaches, such as analysing information directly from the graph. This includes common neighbours [16], using metadata of the nodes from the application domain [17], or making use of pre-existing information on the connections between nodes in the graph [18]. There are a lot of similar concepts between these approaches and the work in this paper, and we will explore them further for applicability in future work.

Graph Isomorphism is a domain where quantification of similarity between graphs are studied in great detail. Many techniques focus on properties such as orientation or structural arrangements that share little relevancy with pathway objects. However, there are also commonly used metrics such as Graph Edit Distance (GED) [19], which can be useful in potential scoring systems or comparing predicted and observed pathways. Nevertheless, one has to keep in mind that for biotransformation studies, the resulting pathways are tentative manual assignments by experts. They do not always reflect the absolute ground truth of the underlying reaction mechanism.

In summary, the work related to predicting biodegradation pathways seem yet to have taken pathway structures into account. We present our work in this paper that aims to fill this gap, along with a new approach that evaluates the predictions accordingly.

3. Evaluation of Relative Reasoning Models

We extended the closely related *Single-Generation* evaluation in enviPath and used it as a baseline for direct comparison. This gives us a way to determine the improvement of the *Multi-Generation* evaluation approach.

3.1. Relative Reasoning Models

Standard enviPath Relative Reasoning models [7] were used to examine the new evaluation approach. They were built using a chosen set of biotransformation pathways as training data. The set of biotransformation rules consists of rules that were created by experts. All compounds present in these pathways are independently cross-referenced with the rules for their applicability, producing effectively a quasi Boolean Matrix [20] that describes their inter-relationships. The matrix would connect the compounds and rules in a manner similar to:

	r_1	r_2	r_3	r_4	r_5
c_1	1	1	1	1	1
c_2	0	1	0	-1	1
c_3	0	0	1	0	1
c_4	-1	-1	0	1	1
c_5	0	0	0	0	1

where r_n and c_n with $n = 1...5$ are rules and compounds used in the training procedure. Values -1, 0 and 1 in the matrix elements respectively represent *Not applicable*, *Applicable but not observed* and *Applicable and observed*. A machine learning approach will then be used to determine probabilities for each of the applicable transformations. A threshold value set by the user for the model will be used to convert the prediction matrix elements to boolean values.

3.2. Single-Generation Evaluation

Relative Reasoning models are evaluated using the resulting matrix. The matrix is used to generate predictions for applicable transformation rules for compounds in the chosen test set. These predictions are then compared to the applicable rules represented by the reactions observed in the reference data for each of the compounds. Fluctuations may arise due to some randomization components in the calculations. Repetitively applying this process in a holdout procedure will iron them out, providing a more representative estimation of the prediction performance.

3.3. Multi-Generation Evaluation

Compounds in the first generations naturally carry higher confidence in the experimental findings, compared to compounds occurring at higher depth in the pathway. This is due to the amount of test substance being divided into multiple reaction branches and only formed slowly over time. Such resulting product compounds would be in much lower concentrations, which are much more difficult to confirm experimentally. Here, we introduce a scoring system within our approach to account for increasing uncertainty when comparing predicted and observed pathways.

This scoring system assigns rewards and penalties with weights according to the generation of the respective compounds. The resulting score for a pathway represents the agreement between the predicted and observed pathways. The collective scores determined for each of the pathways in the validation set are used to compute conventional metrics such as recall-precision curves. This new approach evaluates the pathway as a whole across multiple generations of compounds. This is in contrast to approaches in other works where predicted reactions in each single generation are evaluated independently.

The prediction quality of Relative Reasoning models depends on the compatibility between the transformation rules and the training set, as well as the test set. Rule sets with low compatibility can lead to scenarios such as having no applicable rules to be applied to the target compound structure. In the *Single-Generation* evaluation process, such scenarios would result in all (if any) observed reactions from the educt being ignored. However, if there are further reactions for the product compound in the data, they would still be evaluated. Alternatively, in the *Multi-Generation* evaluation approach, the prediction would terminate at the initial educt and no further scores will be rewarded besides false negatives for the observed products. Figure 2 demonstrates this difference between the two evaluation approaches with a simple example. The *Multi-Generation* approach provides a better metric for the prediction accuracy on the overall pathway level.

3.4. Pathway Prediction

We predict pathways in the test/validation set starting from their root compound. Each of the possible reactions is determined using the supplied transformation rules, which can be represented as a possible branch evolving from the educt. The conditional probabilities approach reaction probabilities according to their position in the pathway. This procedure takes into account the relationships between the probabilities of prior/ upstream reactions with the current reaction. A depth-dependent adjusted version of the preset threshold value is used in the pruning process with the resulting conditional probability. This conditional probability is defined by the product of the probability value assigned to the current reaction, multiplied with values from all of the upstream reactions. An example pathway beginning from compound A is shown in Figure 3.

The example shows root compound A with probabilities P_B and P_C, for reactions that transform A into compounds B and C, respectively. A hypothetical probability threshold x of value $P_B > x > P_C$ is used in the example, to demonstrate the scenario where compound C is predicted to be not observed. The algorithm then continues to determine the possible reactions for compound B, transforming to compounds D and E at the second generation of the pathway, with respective probabilities P_D and P_E.

These values are combined with P_B using a conditional probabilities approach, to obtain the conditional probabilities

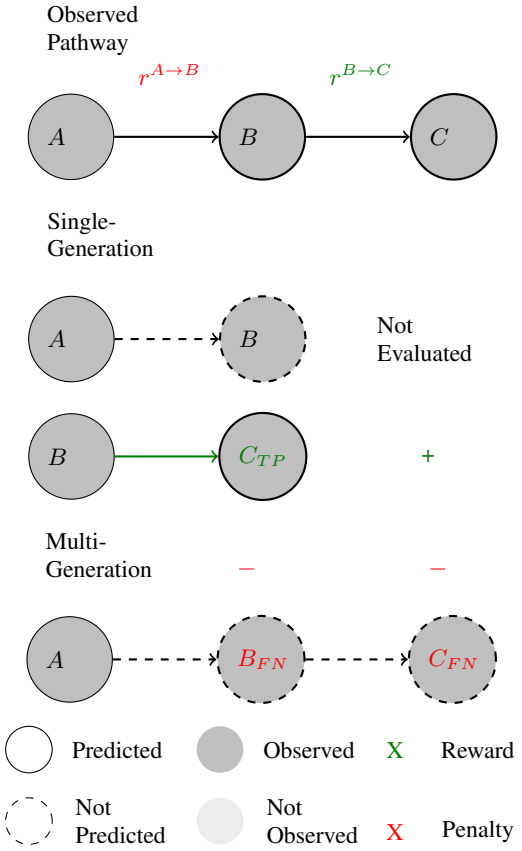

Figure 2: *Graphical representation of a scenario, where a reaction from an observed pathway is not described by any transformation rule used for training. The observed pathway has compound A transformed to B then to C, with the reaction from $B \to C$ described by a transformation rule ($r^{B\to C}$) but none for $A \to B$ ($r^{A\to B}$). The* Single-Generation *evaluation approach would only evaluate $B \to C$ (with a reward +) and ignore $A \to B$, since no applicable rule can be applied. The* Multi-Generation *evaluation approach would penalize both compounds B and C (−) for not being predicted. The additional penalty in the new approach reflects the knowledge gap in the rule set used, whereas it is simply ignored in the single-generation approach.*

$P_B \times P_D$ and $P_B \times P_E$. They are then tested against the threshold value adjusted for reactions at second generation, at x^2. This part of the example demonstrates the scenario where $P_B \times P_E > x^2 > P_B \times P_D$, and compound D is predicted to be not observed. This approach steers the pathway prediction such that, branches with high probabilities will be longer, while less likely branches will be cut earlier.

3.5. Pathway Scores and Model Performance

First, we determined a set of compounds that are present in both the observed and predicted pathway by calculating the overall score between both. This set is used to further determine a set of intermediate metabolites, and adjust the node depths

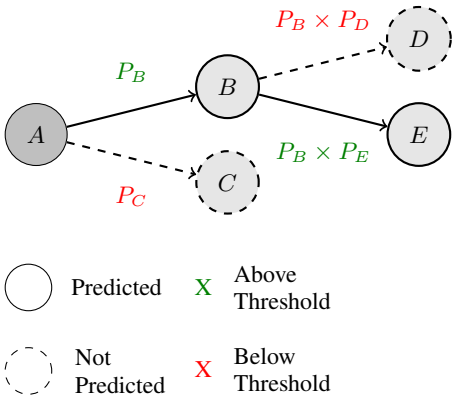

Figure 3: *Graphical representation of the prediction process for an example pathway. P_B and P_C are probabilities of reactions that would transform compound A to compounds B and C, respectively. P_D and P_E are probabilities of reactions that would transform compound B to compounds D and E, respectively. A hypothetical probability threshold x is used to demonstrate how compounds C and D are pruned from the pathway.*

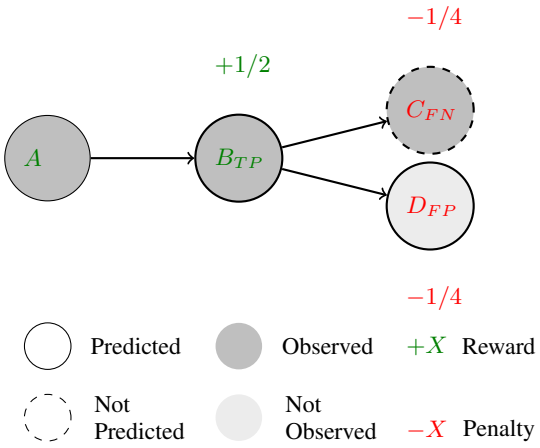

Figure 4: *Graphical representation of the pathway combined from a prediction and an observed pathway in the comparison process. True/false positives/negatives are determined in the comparison, and weights are assigned according to their depths for rewards and penalty calculations.*

accordingly in the predicted pathway. With the intermediate compounds ignored, the quantities TP, FP and FN are then computed as follows:

- TP - Common compounds count as true positives, with weights according to their depth in the observed pathway.
- FP - Compounds that only exist in the predicted pathway count as false positives, with weights according to their depth in the predicted (adjusted) pathway.
- FN - Compounds that only exist in the observed pathway count as false negatives, with weights according to their depth in the observed pathway.

These definitions are used with the *Weighting System* and treatment of intermediate metabolites.

3.5.1. Weighting System

We have constructed a simple mathematical model to compare two pathways with multiple generations. In accordance to the natural decrease in experimental certainty along the pathways in the datasets, the compounds are assigned decreasing weights as their generation/depth level increases. These weight values start at $\frac{1}{2}$ for compounds at generation/depth level one, and decrease by 50% for each increasing level. The weights are used as multipliers to the conventional classification metrics such as true/false positives/negatives. The multipliers are then used to quantify the agreement between predicted and experimental pathways. We use the **Jaccard Index** as metric for pathway similarity. It is defined as:

$$Sim = \frac{\sum(TP \times W_D)}{\sum(TP \times W_D) + \sum(FP \times W_D) + \sum(FN \times W_D)} \quad (1)$$

where TP and FP are the weight-tallied **True** and **False** positives, respectively, and FN represents the **False** negatives. W_D represents the weight multiplier that is dependent on the depth of the metabolite in the pathway. This approach avoids the infinite number of potential true negatives, and gives equal weight to each pathway in the validation set independent of the pathway length. For an example see Figure 4.

3.5.2. Intermediate Metabolites

Intermediate metabolites are compounds with enhanced reactivity. They are quickly transformed to downstream metabolites, and therefore exist only in very low concentrations. These intermediates are sometimes included in the experimental data and sometimes not. This depends on the choice of the author of the experimental study report or the data package and the underlying experimental evidence. If they are not included, the transformation of the educt is reported to lead directly to the downstream metabolite. While prediction of such an intermediate would be mechanistically correct, they might not be present in the experimental data. Such a scenario would incorrectly inflate the count of false positives during the *Single-Generation* evaluation, and would be even more detrimental in the *Multi-Generation* evaluation procedure. The intermediate metabolite would be penalized, along with all metabolites downstream to it, as they would appear at an incorrect depth in the pathway.

In order to correctly accommodate the intermediate metabolites in the evaluation procedure, we have designed a process that accordingly adjusts the depth level of other compounds. The process first determines a list of compounds that are present in both the predicted and observed pathways. Then it determines if any of them are immediately downstream to one another in the observed pathway. The compound pairs that fit this criteria are examined if additional compounds are between them in the predicted pathway. Such compounds are hence added to the list of intermediates. Such intermediate metabolites might still be correctly predicted without the downstream node from the observed pathway. However, the use of a correctly predicted downstream node is required to identify them in a reliable manner and treat them properly. In other words, we

can correct the evaluation of intermediates if and only if they have downstream products in the reference pathway that were correctly predicted.

The list of intermediate compounds is used to adjust depth levels in the predicted pathway accordingly. The shortest path between each of the compounds in the pathway to the root compound is determined using **Breadth-first search** (BFS), and the list of in-between compounds is determined. The depth level of the end compound is then decreased by the number of intermediate compounds that are in this list of in-between compounds. The intermediate compounds are ignored by the *Multi-Generation* evaluation scoring algorithm.

An evaluation example incorporating concepts from both the *Weighting System* and the treatment of intermediate metabolites is presented in Figure 5.

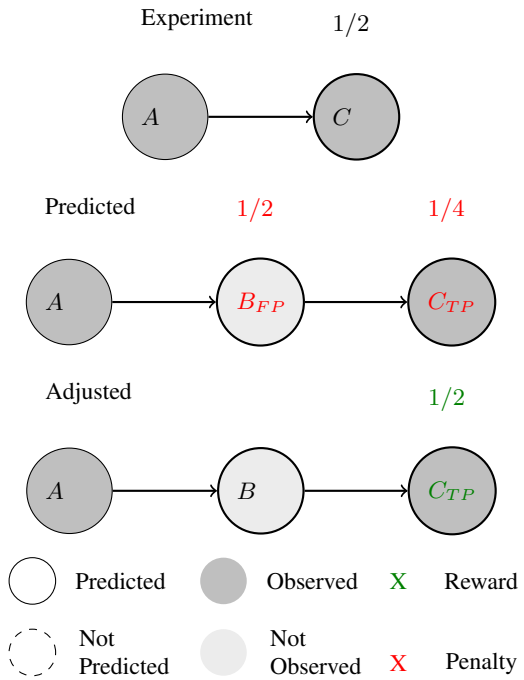

Figure 5: *Graphical representation of the depth adjustment process according to intermediate metabolites determined in the predicted pathway. Compounds A and C are present in both observed and predicted pathways, which allows compound B to be identified as an intermediate metabolite. It can be ignored and the depth-associated-weight for scoring can be adjusted accordingly for compound C.*

The average of comparison scores from all pathways in the validation set represents the accuracy of the model. The precision and recall values are determined from the tallies of true/false positives/negatives obtained. The process is repeated over a range of threshold values to collect the precision and recall values at each threshold, and in turn a representative precision-recall curve is constructed.

4. Experiments

Several experiments are set up to examine the effectiveness of the proposed evaluation approach compared to the conventional approaches. They also test the validity of the mathematical model used to measure similarity between predicted and observed pathways.

4.1. Biotransformation Pathway Data

Several defined sets of biotransformation pathways were used to run the designed experiments:

4.1.1. EAWAG-BBD

The set of biodegradation pathways contained in the EAWAG Biocatalysis / Biodegradation Database package [4]. It contains primarily xenobiotic chemical compounds and microbial biocatalytic reactions. Information on such microbial enzyme-catalyzed reactions carries great importance in the field of biotechnology.

4.1.2. EAWAG-SOIL

The set of biodegradation pathways in the EAWAG-SOIL package [8] contains pesticide degradation pathways compiled from laboratory soil degradation studies. These pesticides are registered in the EU, and their degradation pathways are freely accessible regulatory data.

From the **EAWAG-SOIL** package we selected diverse subsets of pathways as training and test sets that evenly cover the chemical space. This is done to obtain a representative set without over-representation of certain compound clusters. The selection is based on the Tanimoto similarities from Morgan2 fingerprints [21]:

$$T(a,b) = \frac{N_{ab}}{N_a + N_b - N_{ab}} \tag{2}$$

where N_a and N_b are the numbers of "1" bits present in the fingerprints of compounds a and b, and N_{ab} is the number of "1" bits occurring in both fingerprints. The MaxMin algorithm [22] was used to incrementally pick compounds with the least similarity to the most similar compound from the already selected set. We selected 80% of the EAWAG-SOIL pathways to become the **TRAIN-SOIL** package for model training purposes. The remaining 20% make up the **TEST-SOIL** package which is to be used as a test set. We excluded pathways that are not representative for typical organic chemistry, because their root compounds are inorganic salts, too big, or contain heavy metal elements.

4.2. Experiment Setup

We use the set of validated biotransformation rules from the EAWAG-BBD package to build relative reasoning models with compound structures from pathways inside the specified training package. Several experimental setups are designed to examine various aspects of the proposed evaluation approach. We have set the probability threshold for reactions to a low value of 0.1 for all experiments, in order to efficiently capture the difference between the two evaluation approaches.

4.2.1. Train and evaluate

A procedure where the entire chosen list of compounds is used to train a relative reasoning model once. Then we carry out the

evaluation on the nominated test set TEST-SOIL. This procedure is performed on these pathway set combinations: TRAIN-SOIL, EAWAG-BBD + TRAIN-SOIL.

4.2.2. 100-Holdout

This procedure uses a random process to select approximately 66% of the chosen molecules to train a relative reasoning model. The model is then evaluated on the remaining 34% of data. A list of compounds extracted from all selected pathways is used for selection for the *Single-Generation* evaluation approach, and the list of pathways is used for the *Multi-Generation* approach. The process is repeated 100 times, and the results of each individual run is averaged. This approach additionally allows an opportunity to also repeatedly examine the model's prediction ability on data that is new to the training set. This procedure is performed on these pathway set combinations: EAWAG-BBD, EAWAG-SOIL, TRAIN-SOIL, EAWAG-BBD + EAWAG-SOIL, EAWAG-BBD + TRAIN-SOIL.

4.2.3. Validation Test

A procedure to strictly validate the accuracy of the proposed mathematical approach that compares biotransformation pathways. Three sub-procedures are performed:

Full Pathway Evaluate each pathway against itself. The result is expected to be 1.

Empty Pathway Evaluate each pathway against only its starting compound. As the comparison of the pathway starting compound is ignored in the scoring system, the result is expected to be 0.

50% Full Pathways A random process is performed to remove all but the starting compound in approximately 50% of a cloned set of pathways. Each pathway in the original set is evaluated against the associated one in the cloned set. Results of some metrics such as Accuracy and Recall are expected to be close to the ratio of unmodified pathways in the cloned set.

5. Results

To determine the effectiveness and validity of our *Multi-Generation* evaluation approach, results from the procedures detailed in the **Experiments** section were gathered and analyzed. The results include Accuracy, Precision, Recall and Area under the Precision-Recall Curve (AUPRC). Due to the nature of the *Multi-Generation* evaluation approach, where pathways have an infinite number of true negatives, the false positive rate can not be computed. In the *Single-Generation* evaluation approach, the number of true negatives can be calculated from the applicable transformation rules, which are neither predicted (i.e. below the threshold) nor observed experimentally. The Area under the Receiver Operating Characteristic curve (AUROC) is hence only computable for the *Single-Generation* approach and is provided as an indicator.

5.1. Validation Tests

Results of the validation tests performed on the EAWAG-BBD compounds are presented in Table 1. As expected, the evaluated full pathways from both packages achieve 1.0 for **Accuracy**, **Precision** and **Recall**, as there are only true positives and no false positives or negatives. The expected values for evaluated empty pathways from both packages are also 0 for all three metrics, as there are only false positives or negatives without any true positives. The "Half Full" pathways from both packages achieve 1.0 for Precision, and a value that is proportional to the amount of empty pathways (see Table 2) for Accuracy and Recall. The empty pathways will contribute with false negatives while the full pathways will contribute to the true positive score.

Table 1: *Results of validation tests performed for Multi-Generation Evaluation. The validation process was performed on three different modified versions of the training data itself.*

Pathway	Accuracy		Precision		Recall	
	BBD	SOIL	BBD	SOIL	BBD	SOIL
Full	1.0	1.0	1.0	1.0	1.0	1.0
Half Full	0.51	0.47	1.0	1.0	0.5	0.47
Empty	0	0	0	0	0	0

Table 2: *Counts of the full and empty pathways in the validation test process where a random 50% of pathways are emptied.*

	Count	
Pathways	BBD	SOIL
Full	113	153
Empty	105	165
Ratio	0.52	0.48

5.2. Train and Evaluate

Relative reasoning models were trained with the TRAIN-SOIL package and the combination of EAWAG-BBD + TRAIN-SOIL packages. In both cases we evaluated the models on the TEST SOIL package. Tables 3 and 4 show the results, and Figure 6 gives the associated Precision-Recall curves.

Table 3: *Statistics of the **Train and Evaluate** experiments for threshold 0.1*

	Accuracy		Precision		Recall	
	SG	MG	SG	MG	SG	MG
BBD+TRAIN_SOIL	0.53	0.09	0.34	0.1	0.66	0.36
TRAIN_SOIL	0.6	0.15	0.4	0.21	0.71	0.38

Table 4: *Statistics of the **Train and Evaluate** experiments for the whole range of thresholds*

	AUPRC		AUROC
	SG	MG	SG
BBD+TRAIN_SOIL	0.43	0.04	0.8
TRAIN_SOIL	0.41	0.09	0.82

The numerical values of each metric are noticeably lower for the *Multi-Generation* evaluation approach compared to the *Single-Generation* approach. This has mainly two reasons: First, the *Single-Generation* evaluation is based only on defined transformation rules, whereas *Multi-Generation* evaluates

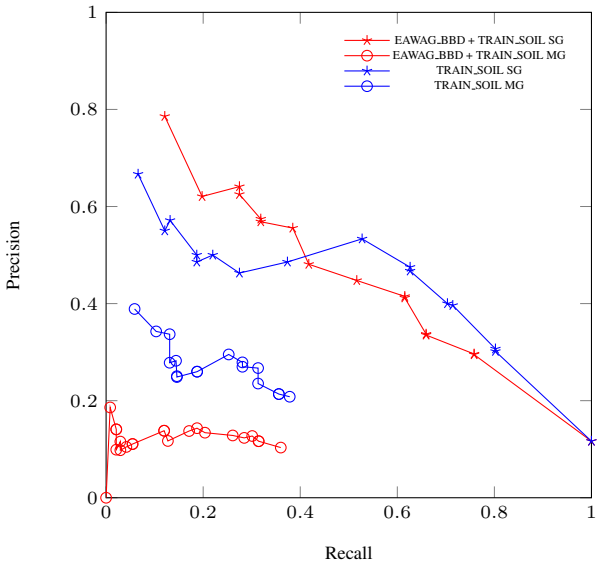

Figure 6: *Precision-Recall curves for the* ***Train and Evaluate*** *experiments. We can see that the* Multi-Generation *evaluation approach better reflects the compatibility between compound structures and the transformation rules used to train the model.*

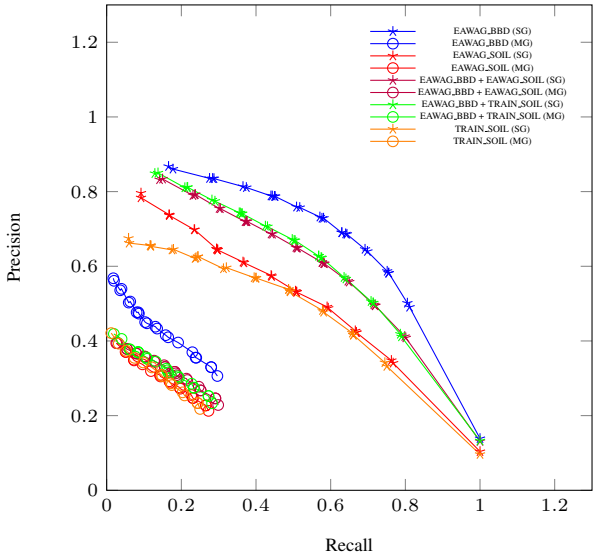

Figure 7: *Precision-Recall curves for the* ***100-Holdout*** *experiments. The repeat-and-average component of this training approach quite effectively smooth out the kinks observed from the* ***Train and Evaluate*** *experiments. The differences in the gap that indicates the compatibility between the transformation rules and the observed compound structures are also visible. The expected relationships between threshold, precision and recall are better reflected in the results.*

all nodes in the reference pathway and thus penalizes incomplete sets of transformation rules. Second, a wrong prediction in the *Multi-Generation* approach is more detrimental, because all the downstream nodes from this branch will be wrong as well. In other words, for a true positive to be tallied, all upstream nodes also have to be predicted correctly. Additionally, a false positive will lead to even more false positives downstream. These two reasons make the *Multi-Generation* approach a much harder evaluation criterion.

Another point worth noting from the *Multi-Generation* evaluation results is that the values for recall do not reach 1 (see Precision-Recall curve, Figure 6). The gap between the maximum recall and the value of 1.0 is caused by transformations in the reference pathways which are not covered by transformation rules and their downstream nodes. The products of such reactions can therefore never be predicted correctly, no matter how low the probability threshold is and will always count as false negatives. Moreover, as discussed above, any downstream nodes won't be predicted either. In contrast, missing transformation rules have no effect on the *Single-Generation* performance, since *Single-Generation* is only evaluated on the existing rules. Thus, the maximum recall value at probability threshold zero can be used as an indicator for the completeness of the rules for the test set.

The data in the TRAIN-SOIL package is naturally closer to the evaluated TEST-SOIL package in terms of chemical and biological properties compared to the EAWAG-BBD package. Therefore, the relative reasoning model trained without the EAWAG-BBD package is more compatible with the evaluation data set. This can be observed in the statistics from the *Single-Generation* evaluation approach. However, the difference is evidently more obvious in the *Multi-Generation* evaluation results, particularly in the Precision-Recall curve. The differences in the areas under the *Multi-Generation* Precision-Recall curves are evidently larger than the *Single-Generation* evaluation counterpart.

5.3. 100-Holdout

Relative Reasoning models were trained with the EAWAG-BBD package, EAWAG-SOIL package, TRAIN-SOIL package, EAWAG-BBD + EAWAG-SOIL, and EAWAG-BBD + TRAIN-SOIL packages. For all cases, we repeated a holdout evaluation 100 times. The results are presented in Tables 5 and 6, and the associated Precision-Recall curves are presented in Figure 7.

Table 5: *Statistics of the* ***100-Holdout*** *experiments.*

Packages	Accuracy		Precision		Recall	
	SG	MG	SG	MG	SG	MG
BBD	0.65	0.17	0.58	0.31	0.76	0.3
SOIL	0.65	0.13	0.42	0.21	0.67	0.27
TRAIN_SOIL	0.65	0.13	0.42	0.22	0.66	0.25
BBD+SOIL	0.62	0.15	0.49	0.22	0.72	0.3
BBD+TRAIN_SOIL	0.63	0.15	0.5	0.233	0.71	0.28

Table 6: *Statistics of the* ***100-Holdout*** *experiments.*

Packages	AUPRC		AUROC
	SG	MG	SG
BBD	0.64	0.12	0.87
SOIL	0.47	0.07	0.83
TRAIN_SOIL	0.43	0.08	0.82
BBD+SOIL	0.56	0.09	0.85
BBD+TRAIN_SOIL	0.57	0.09	0.85

Observations from the **Train and Evaluate** results are

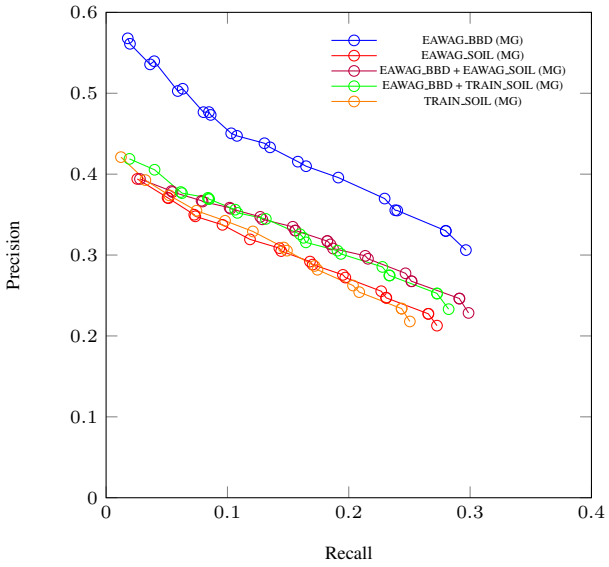

Figure 8: *Precision-Recall curves from the* Multi-Generation *results for the* ***100-Holdout*** *experiments. The thresholds used for the curve are derived from the distribution of probability values from all reactions evaluated. The regular gaps between data points on each line is a result of probability provided by Meka [23], which are rounded to the nearest 0.05 step.*

also notably present in these results from a more repetitive-averaging process. The Precision-Recall curves are also notably smoother from the averaging process, for results from both approaches. The curves from the *Multi-Generations* approach also distinguish more clearly between results from only using the EAWAG-BBD package and other configurations. This is partially due to the fact that transformation rules from the EAWAG-BBD package are used to train all of these relative reasoning models. These rules were optimized for EAWAG-BBD but not for the other packages for which they are less suitable. Also note that soil is a more complex system. The outcome of such an experiment is more difficult to predict [8] than for a pure culture study from EAWAG-BBD.

Both Figure 6 and 7 indicate the Precision-Recall curves from each of the evaluation approach occupy a different region in this phase space. Numerical values of Precision and Recall from the *Multi-Generation* approach are less ideal by conventional standards. This is due to the difference in nature for an agreement to be registered in both approaches. The *Single-Generation* approach is analogous to evaluating each individually predicted word written by a columnist. The multi-generation approach on the other hand, is analogous to extending this to sentences and paragraphs. That is, correct predictions in the former may be penalized in the latter for being in the wrong place. Such a relationship between the two approaches indicate that it is natural to expect this difference in resulting numerical values between the two approaches. An enlarged version of the *Multi-Generation* Precision-Recall curves from the **100-holdout** experiments are presented in Figure 8.

6. Summary

In this paper, we presented a new *Multi-Generation* approach for evaluating relative reasoning prediction models, that are used to predict biodegradation pathways. It includes methodology as well as performance in specifically designed experiments. The new approach evaluates predicted pathways with multiple generations of compounds as a whole, in contrast to considering each reaction independently. Our approach additionally takes into consideration the increased uncertainty of observing compound at higher depths in the pathways. We also propose an algorithm to account for intermediate metabolites, which would otherwise be incorrectly penalized during evaluation.

Our experiments show that the *Multi-Generation* evaluation metrics are much harder criteria. On the other hand, they also provide a more realistic view on the prediction quality of whole pathways and the completeness of the transformation rules. It also provides the possibility to directly compare the performance of different model approaches independent of their underlying transformation rules. *Single-Generation* evaluation on the other hand is more useful for determining the predictivity for individual (defined) transformation rules. Another applications include some steps during model development like hyperparameter optimization, for which computational efficiency is important.

Overall, our experiments demonstrate that it is still a long way until biotransformation prediction models can achieve top accuracy. However, the *Multi-Generation* approach addresses some of the challenges with pathway evaluation, and thus will facilitate the development of better models in the future. We plan to further improve our models by quantifying and improving the compatibility of the biotransformation rules and by integrating the new knowledge about likely intermediates into model training.

7. Interests Disclosure

Sebastian Schmidt is an employee of Bayer AG, a manufacturer of pharmaceutical, agricultural, and consumer health chemicals. Jörg Wicker and Tim Lorsbach are founders of enviPath UG & Co. KG a scientific software development company that develops and maintains the enviPath system. Along with Jason Tam who are also employees of the company.

8. Acknowledgements

This research was supported by the Nectar Research Cloud and The University of Auckland. The Nectar Research Cloud is a Collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy. We would also like to thank Kathrin Fenner from EAWAG for valuable feedback on a draft version of this manuscript, and patronizing the curation of the EAWAG-SOIL data package.

9. Datasets

The datasets used in this study are publicly available at:

- EAWAG-BBD: https://envipath.org/package/32de3cf4-e3e6-4168-956e-32fa5ddb0ce1
- EAWAG-SOIL: https://envipath.org/package/5882df9c-dae1-4d80-a40e-db4724271456

10. References

[1] Council of European Union, "Council regulation (EU) no 1907/2006," 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02006R1907-20200824.

[2] N. Greene, P. Judson, J. Langowski, and C. Marchant, "Knowledge-based expert systems for toxicity and metabolism prediction: Derek, star and meteor," *SAR and QSAR in Environmental Research*, vol. 10, no. 2-3, pp. 299–314, 1999.

[3] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, and M. Kanehisa, "Pathpred: an enzyme-catalyzed metabolic pathway prediction server," *Nucleic acids research*, vol. 38, no. suppl_2, pp. W138–W143, 2010.

[4] L. B. Ellis, D. Roe, and L. P. Wackett, "The university of minnesota biocatalysis/biodegradation database: the first decade," *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D517–D521, 2006.

[5] O. Mekenyan, S. Dimitrov, T. Pavlov, G. Dimitrova, M. Todorov, P. Petkov, and S. Kotov, "Simulation of chemical metabolism for fate and hazard assessment. v. mammalian hazard assessment," *SAR and QSAR in Environmental Research*, vol. 23, no. 5-6, pp. 553–606, 2012.

[6] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la Fuente, R. Greiner, C. Manach, and D. S. Wishart, "Biotransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification," *Journal of cheminformatics*, vol. 11, no. 1, pp. 1–25, 2019.

[7] J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D. Latino, S. Kramer, and K. Fenner, "envipath–the environmental contaminant biotransformation pathway resource," *Nucleic acids research*, vol. 44, no. D1, pp. D502–D508, 2016.

[8] D. A. Latino, J. Wicker, M. Gütlein, E. Schmid, S. Kramer, and K. Fenner, "Eawag-soil in envipath: a new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data," *Environmental Science: Processes & Impacts*, vol. 19, no. 3, pp. 449–464, 2017.

[9] K. C. Soh and V. Hatzimanikatis, "Dreams of metabolism," *Trends in biotechnology*, vol. 28, no. 10, pp. 501–508, 2010.

[10] P. N. Judson and J. D. Vessey, "A comprehensive approach to argumentation," *Journal of chemical information and computer sciences*, vol. 43, no. 5, pp. 1356–1363, 2003.

[11] P. N. Judson, C. A. Marchant, and J. D. Vessey, "Using argumentation for absolute reasoning about the potential toxicity of chemicals," *Journal of chemical information and computer sciences*, vol. 43, no. 5, pp. 1364–1370, 2003.

[12] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "Kegg for representation and analysis of molecular networks involving diseases and drugs," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D355–D360, 2010.

[13] M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa, "Computational assignment of the ec numbers for genomic-scale analysis of enzymatic reactions," *Journal of the American Chemical Society*, vol. 126, no. 50, pp. 16 487–16 498, 2004.

[14] K. Fenner, J. Gao, S. Kramer, L. Ellis, and L. Wackett, "Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction," *Bioinformatics*, vol. 24, no. 18, pp. 2079–2085, 2008.

[15] R. C. Kolanczyk, P. Schmieder, W. J. Jones, O. G. Mekenyan, A. Chapkanov, S. Temelkov, S. Kotov, M. Velikova, V. Kamenska, K. Vasilev *et al.*, "Metapath: an electronic knowledge base for collating, exchanging and analyzing case studies of xenobiotic metabolism," *Regulatory Toxicology and Pharmacology*, vol. 63, no. 1, pp. 84–96, 2012.

[16] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 5165–5175.

[17] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[18] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical science*, vol. 10, no. 2, pp. 370–377, 2019.

[19] A. Sanfeliu and K.-S. Fu, "A distance measure between attributed relational graphs for pattern recognition," *IEEE transactions on systems, man, and cybernetics*, no. 3, pp. 353–362, 1983.

[20] J. Wicker, B. Pfahringer, and S. Kramer, "Multi-label classification using boolean matrix decomposition," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, pp. 179–186.

[21] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[22] M. Ashton, J. Barnard, F. Casset, M. Charlton, G. Downs, D. Gorse, J. Holliday, R. Lahana, and P. Willett, "Identification of diverse database subsets using property-based and fragment-based molecular descriptions," *Quantitative Structure-Activity Relationships*, vol. 21, no. 6, pp. 598–604, 2002.

[23] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, "Meka: a multi-label/multi-target extension to weka," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 667–671, 2016.