Supporting Information-1 for Critical insights into data curation and label noise for accurate prediction of aerobic biodegradability of organic chemicals

Paulina Körner,[†] Juliane Glüge,[*,†] Stefan Glüge,[‡] and Martin Scheringer[†]

[†]*Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland*

[‡]*Institute for Computational Life Science, ZHAW, 8820 Wädenswil, Switzerland*

E-mail: juliane.gluege@usys.ethz.ch

Contents

S1 Definitions

S1.1 Chemical Abstracts Service Registry Number (CAS RN™)

CAS RN™ have been introduced as unique and unambiguous identifiers for chemical substances by the Chemical Abstracts Service[1]. A CAS RN™ is a distinct numeric identifier that represents a single substance and does not carry any chemical significance. CAS RN™ are assigned to substances as they enter the CAS Registry database. CAS RN™ typically consist of up to ten digits, divided into three parts by hyphens[1]. For example, the CAS RN™ of 7-Aminocephalosporanic acid is 957-68-6.

CAS RN™ have proven helpful in tracking and managing substances. They have the advantage that they are unique, easy to validate, and internationally accepted[1]. However, even though CAS RN™ are unique, more than one CAS RN™ can exist for a given substance because multiple CAS RN™ were assigned to a substance or due to deprecated CAS RN™. For example, 7-Aminocephalosporanic acid is also linked to the deprecated CAS RN™856652-38-5, 856652-39-6, 13256-42-3, 23241-25-0, 70035-93-7, and 26328-10-9. Even though all these CAS RN™ are unique for 7-Aminocephalosporanic acid, the fact that multiple CAS RN™ exist for substances can lead to issues if this is not considered. These deprecated CAS RN™ can often still be found in online references[2].

S1.2 Simplified Molecular Input Line Entry Specification (SMILES)

The SMILES representation was introduced by Weininger[3] and has since then become the most widely used line notation[3–6]. It is obtained by assigning a distinct number to each atom in the molecule and then traversing the molecular graph using that specific order. However, due to multiple possible atom numberings for a given molecule, different SMILES notations can be generated while maintaining the same graph traversal algorithm. Therefore, SMILES

are not unique[6].

SMILES can be described as isomeric and canonical. Isomeric SMILES allow for the specification of isotopism and stereochemistry. Canonical SMILES are supposed to ensure that the same SMILES is generated for a given molecule. However, various algorithms have been developed to generate canonical SMILES. Therefore, multiple different canonical SMILES can exist for a substance[7]. The canonical SMILES of 7-Aminocephalosporanic acid is `CC(=O)OCC1=C(N2C(C(C2=O)N)SC1)C(=O)O` and the isomeric SMILES is `CC(=O)OCC1=C(N2[C@@H]([C@@H](C2=O)N)SC1)C(=O)O` according to PubChem[2]. Furthermore, SMILES notation encounters challenges in describing certain complex structures that cannot be easily represented using molecular graphs, including organometallic compounds and ionic salts[6].

S1.3 International Chemical Identifier (InChI™)

The InChI™ string is a standardized, machine-readable string of symbols used to represent chemical compounds in an unambiguous manner[8]. It is a unique and automatically generated representation of a compound's molecular structure[8,9]. The InChI adopts a layered format to encompass all relevant structural information for compound identification. Each layer contains specific types of structural details, with successive layers adding additional information. The layered design allows for the inclusion of various structural levels. The layers in the InChI string are separated by slashes followed by lower-case letters, arranged in a predefined order[8]. Table S1 shows the structure of an example InChI™ and its layers and segments. The main layer of the InChI™ provides information about the core parent structure, including its chemical formula, atom connectivity (non-hydrogen), and hydrogen atom connectivity[10]. In addition to the full InChI™ strings, the InChI™-main-layer was be used to identify substances with the same parent structure, independent of charge and stereochemistry.

It is important to note that any ambiguities or uncertainties in the original structure representation persist in the InChI[8]. Furthermore, unlike SMILES notation, InChI™ do not

offer a guarantee of reversibility to reconstruct the original molecular graphs from which they originate[6].

Table S1: The structure of an InChI™ string and the meaning of each layer. The example InChI™ represents the substance 7-Aminocephalosporanic acid.

Inchi layer	Letter	Inchi segment
Main layer		
		InChI version number
		chemical formula
	c	atom connectivity
	h	hydrogen atom connectivity
Charge layer		
	q	charge
	p	proton balance
Stereochemical layer		
	b	double bonds
	t	tetrahedral parity
	m	parity inverted to obtain relative stereo (1=inverted, 0=not inverted)
	s	stereo type (1=absolute, 2=relative, 3=racemic)
Isotopic layer	i	isotopic enrichment

InChI=1S/C10H12N2O5S/c1-4(13)17-2-5-3-18-9-6(11)8(14)12(9)7(5)10(15)16/h6,9H,2-3,11H2,1H3,(H,15,16) /p-1/t6-,9-/m1/s1

S1.4 Label

In the context of supervised machine learning (ML), labels are the target variable that the ML model is trying to predict. The label is the ground truth or the correct answer of each data point in the training and test set.

The data points used here to train a classifier to predict whether a substance was ready biodegradable or not were labeled with a 0 or 1. The label 0 indicates that the substance was not readily biodegradable (NRB), and the label 1 indicates that the substance is readily biodegradable (RB) according to the experimental study results.

S1.5 Feature

In order to train machine learning models, the data needs to be prepared in a format that the ML model can use for training and testing. The data is given to the ML model in form of features. Features are individual characteristics of the input data and represent the information that the model analyzes to learn patterns, relationships, and structures within the data. Features can be either categorical, for example, gender or nationality, or numerical[11]. The quality and relevance of the features can significantly impact the performance of a ML model. Therefore, properly selecting, engineering, and managing features is a fundamental aspect of the ML process[12].

S1.6 Data leakage

In ML, models are typically evaluated on test data that the model has not seen during training. This provides a reliable indication of how well the model will perform when deployed in real-world scenarios and helps detect potential issues such as overfitting or lack of generalization. If data appears both in the train and the test set, it is called data leakage. Data leakage can lead to overly optimistic model performance or invalid assessments of the model's generalization ability[13].

S1.7 Molecular descriptors: Molecular Access System key (MACCS key)

Molecular descriptors can be categorized into two main classes: structural keys and hashed fingerprints (FPs). Structural keys are represented as bit strings, encoding the presence (1) or absence (0) of specific chemical groups. In contrast, chemical FPs are vectors containing indexed elements that encode various physicochemical or structural properties. The distinctive characteristic of hashed FPs is that each element is generated from the molecule itself, while in structural keys, predefined patterns are used[6].

A widely used example of a key-based molecular descriptor is the MACCS key. In the MACCS keys, each bit corresponds to the presence or absence of a specific structural fragment. Different variants of the MACCS keys have been developed, with the most commonly utilized version being 167 bits long. This version encodes for the presence or absence of 166 structural fragments[6,14].

S1.8 Balanced accuracy

The balanced accuracy is the arithmetic mean of the scores of sensitivity (SE) and specificity (SP)[15]. The latter two are defined as:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

where TP are the true positives (substances that are ready biodegradable in the experiment and in the prediction), TN are the true negatives (substances that are not-readily biodegradable in the experiment and in the prediction), FP are the false positives (substances that are not-readily biodegradable in the experiment but ready biodegradable in the prediction), and FN the false negatives (substances that are ready biodegradable in the experiment but not-readily biodegradable in the prediction). The balanced accuracy is then:

$$\text{Balanced accuracy} = \frac{\mathrm{SE} + \mathrm{SP}}{2}$$

S2 The SMILES-Retrieval-Pipeline

During the analysis of the Huang-Datasets, inconsistencies were found in the CAS RN™ – SMILES pairings. The CAS RN™ was treated as the definitive substance identifier because they were present in the original eChemPortal dataset. Therefore, the correct SMILES

corresponding to a given CAS RN™ had to be found. To accomplish this, the SMILES-Retrieval-Pipeline was developed, which is shown in Figure S1.

The initial step of the SMILES-Retrieval-Pipeline involved retrieving the unique CAS RN™ entries from the Huang-Regression-Dataset. These CAS RN™ and their corresponding SMILES were then split into two groups based on whether they were verified by Glüge et al.[9]. In case a CAS RN™ was included in the Gluege-Dataset, the verified and valid SMILES for this CAS RN™ from the Gluege-Dataset was added to the data point. For the substances in the Gluege-Dataset, and therefore the substances registered under Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), it was also checked if the experimental study was conducted for the registered substance or if it was based on read-across. The goal was to only have studies carried out for the registered substance, and, therefore, studies based on read-across were removed.

For the substances not checked by Glüge et al.[9], valid SMILES had to be retrieved. For one-component substances, the SMILES were retrieved via an Application Programming Interface (API) based on the CAS RN™ from CAS Common Chemistry. CAS Common Chemistry is a reliable source for SMILES of substances with one component but contains some systematic errors in SMILES for multiple-component substances[9]. For one-component substances not found on CAS Common Chemistry and for multiple-component substances, a weight-of-evidence approach was taken. The SMILES had to be found from at least two independent sources. If this was not possible, the substance was removed from the dataset (Figure S1).

Once the SMILES were found by CAS RN™, further processing steps were performed. First, substances containing multiple CAS RN™ were identified. This was necessary to avoid data leakage later on. Identification of compounds in the Huang-Regression-Dataset with multiple CAS RN™ was performed using InChI™. If more than one CAS RN™ was associated with the data points in each group, then the shortest CAS RN™ was taken and assigned to all data points with the same InChI™.

For all ionizable substances, the retrieved SMILES was replaced with the SMILES of the substance's dominant species at pH 7.4 and 298 K. Substances were removed when no dominant species existed under the specified conditions. This step was not performed by Huang and Zhang[16]. Instead, they introduced two extra features (pK_a and α-values) that represent the chemical specification of the substances. They reported a performance increase in the accuracy from 85.1% to 87.6% when including pK_a and α-values as extra features. However, using the same model, we could not reproduce this performance increase (see Table S6). Therefore, we did not include information on chemical specification directly as features. However, this information is reflected in the SMILES.

Furthermore, substances that were mixtures were removed, and all counterions were removed from the SMILES representations. For stereoisomers, the SMILES of one stereoisomer was randomly selected. Lastly, organometallic substances were removed.

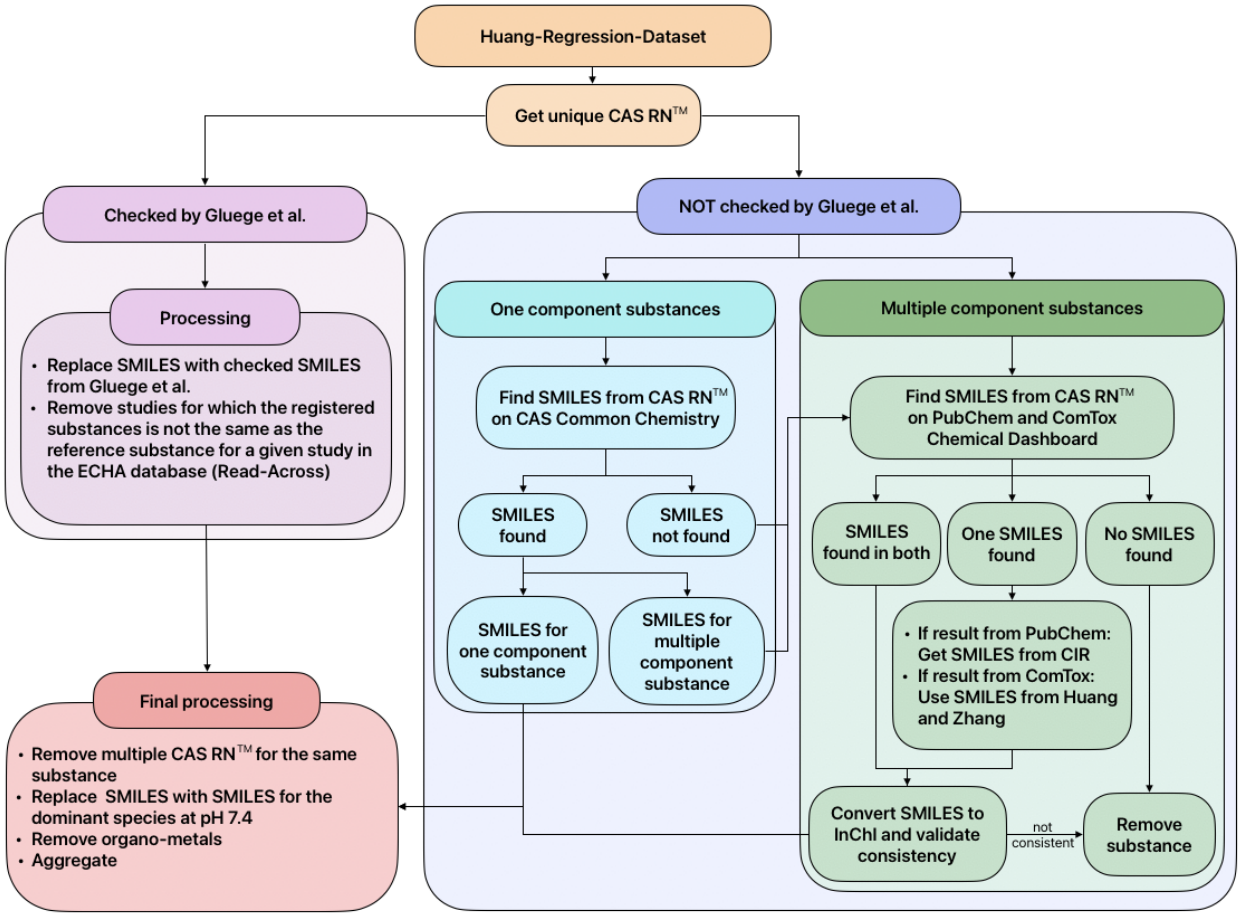


Figure S1: SMILES-Retrieval-Pipeline used for finding the SMILES corresponding to the CAS RN™ for each substance in the Huang-Regression-Dataset.

S3 Label noise filtering

Label or class noise filtering is a technique used in ML to mitigate the impact of incorrect labels in the data. Noisy labels refer to mislabeled or inaccurately labeled instances in the train and test dataset, which can adversely affect the performance of the ML model[17,18]. Label noise filtering aims to identify and correct or remove these instances to improve the model's performance and generalization.

Identifying label noise can be tricky in the case of experimental Ready-Biodegradability Tests (RBT) results. It is impossible to manually check the labels or check the labels of

all substances in a database or literature. Therefore, we used BIOWIN5 and BIOWIN6 to identify potentially noisy labels. Substances with potentially noisy labels were removed. The resulting $\text{CURATED}_{\text{BIOWIN}}$ dataset was then used for training and testing of the ML models.

BIOWIN5 and BIOWIN6 are part of Estimation Program Interface Suite (EPI Suite™) and widely used. However, both models have a reported accuracy of 83% on an external test set. This means that a significant number of data points might have been falsely removed. Since not only data quality but also data quantity is important for ML model performance, a third classifier was used to readd some of the potentially falsely removed data points. This third classifier was a XGBClassifier trained on the $\text{CURATED}_{\text{BIOWIN}}$ dataset and had a balanced accuracy of $94.2 \pm 0.9\%$. When this third classifier agreed with the experimental label, then the label was not considered noisy, and the data point was added back. This resulted in the creation of the $\text{CURATED}_{\text{FINAL}}$ dataset. The process used for the label noise filtering is also shown in Table S2.

Table S2: Label noise filtering system to create the $\text{CURATED}_{\text{BIOWIN}}$, $\text{CURATED}_{\text{PROBLEMATIC}}$, $\text{CURATED}_{\text{FINAL}}$, and the $\text{CURATED}_{\text{REMOVED}}$ datasets. The first vote decides whether a substance is in the $\text{CURATED}_{\text{BIOWIN}}$ or the $\text{CURATED}_{\text{PROBLEMATIC}}$ dataset. If the BIOWIN models agree with the label, then the vote is positive (✓), and the substance is in the $\text{CURATED}_{\text{BIOWIN}}$ dataset. If one or both of the BIOWIN models do not agree with the label (✕), then the substance is in the $\text{CURATED}_{\text{PROBLEMATIC}}$ dataset, and the third classifier is consulted. If the third classifier agrees with the label (✓), then the substance is moved from the $\text{CURATED}_{\text{PROBLEMATIC}}$ to the $\text{CURATED}_{\text{FINAL}}$ dataset. Otherwise (✕), the substance is placed in the $\text{CURATED}_{\text{REMOVED}}$ dataset. NaN means that BIOWIN was not able to make a prediction.

Label	**BIOWIN5**	**BIOWIN6**	**First vote**	**Third classifier**	**Final vote**
NRB	NaN	NaN	✓	–	✓
NRB	NRB	NRB	✓	–	✓
NRB	RB	NRB	✕	NRB	✓
NRB	RB	NRB	✕	RB	✕
NRB	RB	RB	✕	NRB	✓
NRB	RB	RB	✕	RB	✕
RB	NaN	NaN	✓	–	✓
RB	RB	RB	✓	–	✓
RB	RB	NRB	✕	RB	✓
RB	RB	NRB	✕	NRB	✕
RB	NRB	NRB	✕	RB	✓
RB	NRB	NRB	✕	NRB	✕

S4 Applicability domain

Defining the Applicability Domain (AD) of ML models that predict the activity of chemicals based on their structure is crucial to assess the reliability and relevance of predictions made by the model for new and unseen data[19,20]. Huang and Zhang[16] used the Tanimoto index, which calculates similarities between two chemicals based on the number of common molecular fragments, to define the AD of their models[16,21]. Here, the Tanimoto index is calculated using adapted code from the cheminformatics package Python Applicability Domain Analyzer (pyADA)[22].

The Tanimoto index was used to calculate the similarities between each substance in the

testing set and all the substances in the training set. The Tanimoto index was calculated using the MACCS keys of the substances. The similarity calculation results in a value between 0 and 1 for each substance, with a 0 indicating no similarity at all between the two substances and a value of 1 indicating that they have identical MACCS keys[23,24].

S4.1 Defining the similarity threshold

To define the AD of a model, first, a similarity threshold must be set for the dataset the model is trained on. This threshold represents the minimum required similarity between a substance in the test set and the substances in the training set to be within the AD. The similarity threshold is set manually based on the number of substances in the test set that are between defined similarity ranges, the expected model performance for substances in these similarity ranges, and domain knowledge.

First, it was calculated how many of the substances in the test set have a maximum Tanimoto similarity to the training data within a certain similarity range. For example, it was calculated how many substances in the test set have a maximum similarity to the training set between 0.8 and 0.9. Furthermore, the expected mean model performance for all substances in that similarity range was calculated. These values were generated by training and testing the model again only on the substances within a given similarity range. For example, it was calculated that for all substances with a maximum Tanimoto similarity between 0.8 and 0.9, the classifier is expected to predict the biodegradability of these substances with an accuracy of 95%.

Here, the similarity threshold is set by training and testing the model again using the five different data splits. The mean of the reported performance metrics and the number of substances in a certain similarity range were calculated. Based on this, the similarity threshold was determined. Once the similarity threshold was set, the AD of the model was defined.

S4.2 Applying the AD to the DSSTox database

The defined similarity threshold could then be used to check whether a new and unseen substance is within the AD of the model. To do so, the Tanimoto similarities between the new substance and all substances in the training set must be calculated. If the maximum Tanimoto similarity is above the defined similarity threshold, then the test compound is considered to fall within the model's AD. If the new compound's similarity with the training set compounds is below the threshold, then it is considered outside the AD of the model. This suggests that the compound is dissimilar to the compounds on which the model was trained, and predictions may be less reliable.

To evaluate the broadness of the AD of the models, Huang and Zhang[16] evaluated how many of the substances in the Distributed Structure-Searchable Toxicity (DSSTox) database are in the AD. The DSSTox database is operated by the United States Environmental Protection Agency (U.S. EPA) and contains more than 850 000 environmentally relevant chemicals[16]. To check if the substances in this database were in the AD of the models, Huang and Zhang[16] calculated the Tanimoto similarities between all substances in the DSSTox database and the substances in the Huang-Regression-Dataset and Huang-Classification-Dataset. Here, the same procedure was replicated for the $\text{Curated}_{\text{SCS}}$, $\text{Curated}_{\text{BIOWIN}}$, and $\text{Curated}_{\text{FINAL}}$ datasets.

S5 Huang-Dataset analysis examples

Table S3: Examples of data points in the Huang-Regression-Dataset for which the SMILES added by Huang and Zhang[16] did not match with the SMILES according to CAS RN™ checked by Glüge et al.[9].

Category	**Value**
CAS RN™:	**1742-79-6**
SMILES Huang and Zhang[16]	O=S(=O)([O-])c1cccc(/N=N/c2ccc(Nc3ccccc3)cc2)c1.[Na+]
Molecular formula from SMILES	C18H15N3O3S.Na
Molecular formula from CAS RN™	C7H7NO3
SMILES according to CAS RN™	CC(=O)ON1C=CC=CC1=O
CAS RN™	**181525-38-2**
SMILES Huang and Zhang[16]	O=C(ON1C(=O)CCC1=O)ON1C(=O)CCC1=O
Molecular formula from SMILES	C9H8N2O7
Molecular formula from CAS RN™	C11H12N2O3
SMILES according to CAS RN™	CC1=C(CCO)C(=O)N2C=CC=C(O)C2=N1
CAS RN™	**136210-30-5**
SMILES Huang and Zhang[16]	CCOC(=O)/C=C/C(=O)OCC
Molecular formula from SMILES	C8H12O4
Molecular formula from CAS RN™	C29H50N2O8
SMILES according to CAS RN™	CCOC(=O)CC(NC1CCC(CC2CCC(CC2)NC(CC(=O)OCC)C(=O)OCC)CC1)C(=O)OCC

Table S4: Examples of substances with data points with differing SMILES in the HUANG-CLASSIFICATION-DATASET. The data points belonging to the same substances were identified using the InChI™, which was based on the SMILES. The second example does not list CAS RN™, because the data points from the LUNGHINI-DATASET, which were added to the HUANG-CLASSIFICATION-DATASET by Huang and Zhang[16], did not have CAS RN™.

	CAS RN™	**SMILES**	**Label**
InChI=1S/C7H8O2S.Na/c1-6-2-4-7(5-3-6)10(8)9;/h2-5H,1H3,(H,8,9);/q;+1/p-1			
Entry 1	824-79-3	Cc1ccc(S(=O)[O-])cc1.[Na+]	NRB
Entry 2	824-79-3	Cc1ccc(S(=O)O[Na])cc1	RB
InChI=1S/C13H13N3/c14-13(15-11-7-3-1-4-8-11)16-12-9-5-2-6-10-12/h1-10H ,(H3,14,15,16)			
Entry 1	102-06-7	NC(=Nc1ccccc1)Nc1ccccc1	RB
Entry 2	NaN	N/C(=N\c1ccccc1)Nc1ccccc1	NRB
Entry 3	NaN	N=C(Nc1ccccc1)Nc1ccccc1	NRB
InChI=1S/C3H7NO2.Na/c1-4-2-3(5)6;/h4H,2H2,1H3,(H,5,6);/q;+1/p-1			
Entry 1	68411-97-2	CNCC(=O)[O-].[Na+]	RB
Entry 2	4316-73-8	CNCC(=O)O[Na]	RB

Table S5: Examples of substances in the HUANG-REGRESSION-DATASET with multiple study results that were strongly differing. All studies are ready-biodegradability tests carried out for 28 days.

CAS RN™	**Reliability**	**Guideline**	**Principle**	**Biodegrad. percent**
92128-65-9	2	OECD Guideline 301 D	Closed Bottle Test	21.0%
	1	OECD Guideline 301 B	CO2 Evolution	55.0%
	2	OECD Guideline 301 F	Closed Respirometer	70.0%
	1	OECD Guideline 301 F	Closed Respirometer	91.2%
3886-69-9	2	OECD Guideline 301 A	DOC Die Away	98.0%
	1	OECD Guideline 301 F	Closed Respirometer	5.0%
	2	OECD Guideline 301 F	Closed Respirometer	65.0%
13170-23-5	1	OECD Guideline 301 B	CO2 Evolution	41.5%
	1	OECD Guideline 301 C	Closed Respirometer	15.0%
	1	OECD Guideline 301 F	Closed Respirometer	79.5%

S6 Model performance

Table S6: Performance metrics reported by Huang and Zhang [16] and for the XGBClassifiers trained on the Huang-Classification-Dataset.

Datasets	Balanced accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score
Huang-Classification-Dataset reported				
No chemical speciation	85.1	89.0	80.9	86.2
With chemical speciation	87.6	87.8	87.4	87.9
Huang-Classification-Dataset replicated				
No chemical speciation	79.6 ± 1.1	74.1 ± 1.5	85.1 ± 1.2	0.73 ± 0.01
With chemical speciation	80.3 ± 1.1	75.1 ± 1.0	85.5 ± 1.2	0.74 ± 0.01

Table S7: Performance metrics for the XGBClassifiers trained on the Curated-Datasets. The classifiers were tested five times on fixed test sets from the $\text{Curated}_{\text{SCS}}$ and $\text{Curated}_{\text{BIOWIN}}$ datasets.

Test sets	Datasets	Balanced accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score
$\text{Curated}_{\text{SCS}}$	Huang-Classification-Dataset	80.6 ± 1.5	74.2 ± 2.7	87.1 ± 1.0	0.75 ± 0.02
	$\text{Curated}_{\text{SCS}}$	80.9 ± 1.7	74.9 ± 3.0	86.9 ± 1.0	0.75 ± 0.02
	$\text{Curated}_{\text{BIOWIN}}$	78.3 ± 0.9	71.0 ± 2.1	85.6 ± 0.6	0.72 ± 0.01
	$\text{Curated}_{\text{final}}$	79.4 ± 0.8	72.0 ± 1.9	86.9 ± 0.7	0.73 ± 0.01
$\text{Curated}_{\text{BIOWIN}}$	Huang-Classification-Dataset	88.0 ± 1.3	83.1 ± 2.6	92.9 ± 1.0	0.84 ± 0.02
	$\text{Curated}_{\text{SCS}}$	88.4 ± 2.1	84.1 ± 3.4	92.6 ± 1.0	0.84 ± 0.03
	$\text{Curated}_{\text{BIOWIN}}$	93.7 ± 1.0	90.9 ± 2.2	96.4 ± 0.4	0.92 ± 0.01
	$\text{Curated}_{\text{final}}$	94.2 ± 1.2	91.6 ± 2.8	96.9 ± 0.4	0.92 ± 0.01

S7 Analysis of the new datasets

S7.1 Data characteristics

To analyze if the label noise filtering led to the removal of difficult-to-predict data points, several characteristics of the CURATED$_{\text{SCS}}$, CURATED$_{\text{BIOWIN}}$, and CURATED$_{\text{FINAL}}$ datasets were analyzed. No substances had been removed from the CURATED$_{\text{SCS}}$ dataset based on the label, and, therefore, it served as the baseline. The results of the analysis are shown in the main article in **??**. The absolute values are shown in Figure S2, Figure S3, and Figure S4 below.

S7.1.1 Molecular Weight

Molecular weight bears significance in the context of biodegradation due to the established connection between increasing molecular size and reduced biodegradability[25]. Overall, a slightly higher proportion of data points in the lowest and highest molecular weight categories (0–250 Da and 1000–2000 Da) were removed in the CURATED$_{\text{BIOWIN}}$ and CURATED$_{\text{FINAL}}$ datasets. However, it should be noted that the lowest weight category also contained the highest number of substances, which means that the weight category 0–250 Da is still well represented in the datasets. In contrast, the lowest molecular weight category had very few substances, so removing even a small number of substances significantly impacts the percentage of removed substances.

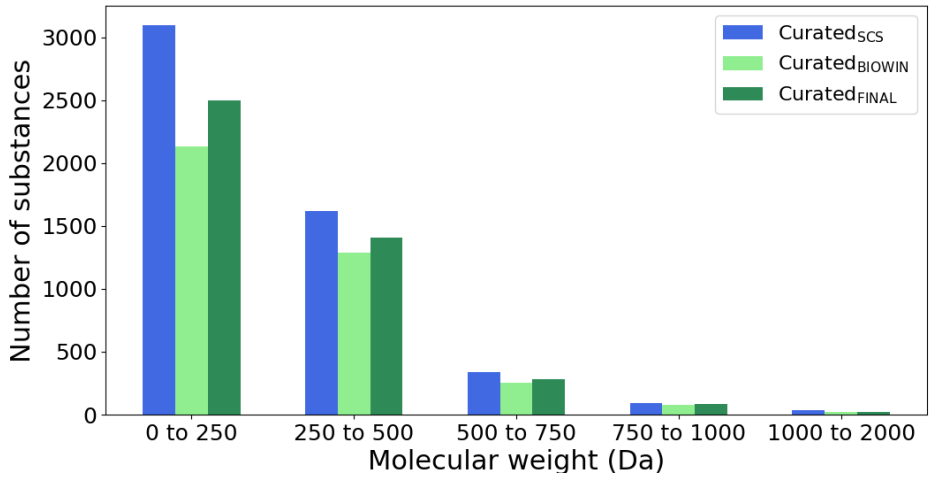


Figure S2: The distribution of data points associated with substances with a given molecular weight in the CURATED$_{SCS}$, CURATED$_{BIOWIN}$, and CURATED$_{FINAL}$ datasets.

S7.1.2 Halogens

The presence of halogen compounds influences biodegradation. Substances containing halogens such as fluorine, bromine, and chlorine generally have been reported to have decreasing biodegradability with increasing degree of halogenation[26–30]. The analysis shows that those substances with one of the three halogens were not removed more than others (cf. **??**).

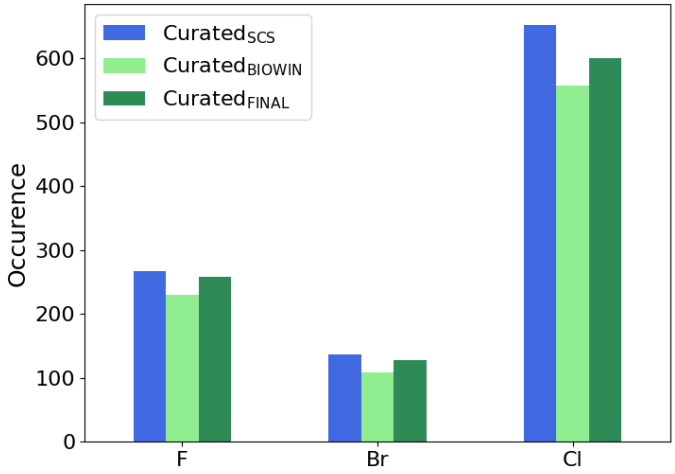


Figure S3: The occurrence of the halogens fluorine (F), bromide (Br), and chlorine (Cl) relative to the total number of data points in each of the classification datasets.

S7.1.3 Distribution of biodegradation labels

For both datasets ($\text{CURATED}_{\text{BIOWIN}}$ and $\text{CURATED}_{\text{FINAL}}$), the relative number of substances with a NRB label was $\approx 10\%$ higher than the number of substances with a RB label. This indicates that the label noise filtering identified more substances labeled as RB than NRB as potentially being incorrectly labeled.

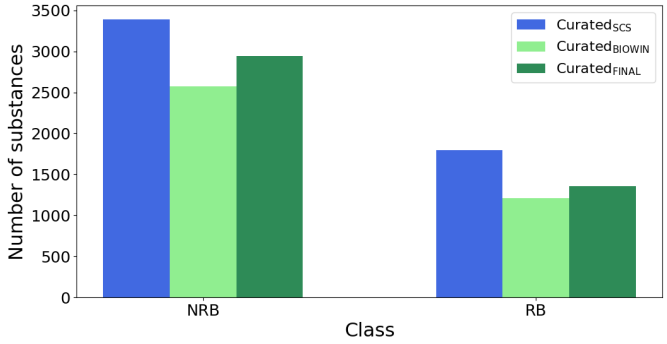


Figure S4: Distribution of biodegradation class labels in the classification datasets

S7.2 Analysing feature adequacy

To check if other features are more suitable to present all the important information about the chemicals than MACCS key, three other feature creation methods were tested. The four feature creation methods differ in the number of features they create and the underlying method used to create them.

S7.2.1 Morgan FPs

Morgan FPs, which are also known as circular FPs, are part of the Extended Connectivity Fingerprints (ECFP) family, which are based on the concept of circular substructures. Morgan fingerprints encode information about circular substructures in a molecule. The algorithm considers atom environments within a defined radius around each atom in the molecule, capturing circular connectivity patterns[31,32]. Given that the common range for the radius is 1 to 3, we opted for a radius value of 2[33]. Morgan FPs with 1 024, 2 048, and

4 096 binary bits can be created using RDKit, where each bit corresponds to the presence or absence of a specific circular substructure in the molecule[34]. Here, Morgan FPs with 1 024 bits were used.

S7.2.2 RDK FPs

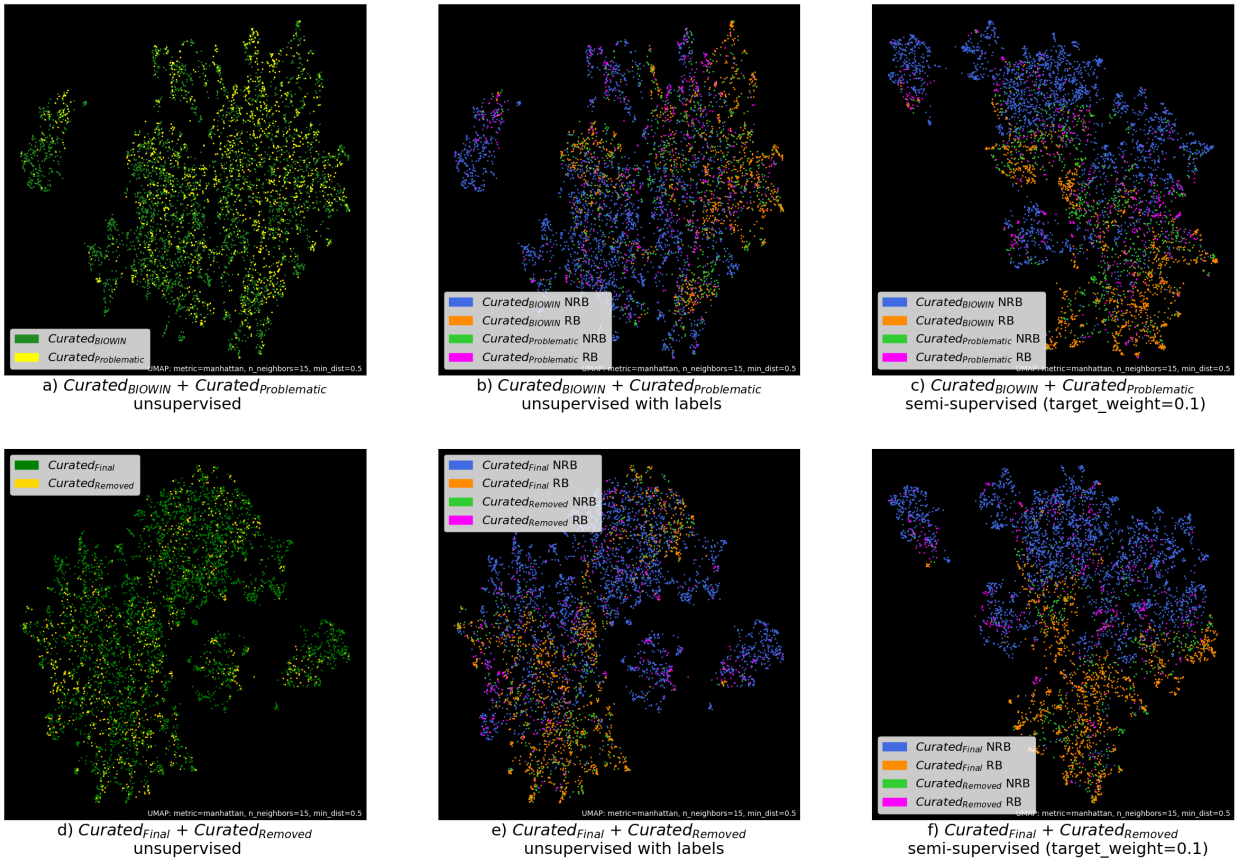
RDKit FPs are topological FPs similar to Daylight fingerprint, that are based on hashing molecular subgraphs[35,36]. To create RDKit FPs, all branched and linear molecular subgraphs of a chemical up to a specified size are hashed by combining information about atom types, atomic numbers, and aromaticity states, and bond types[37]. RDKit FPs generate a binary bit vector of size 2 048[35].

S7.2.3 MolFormer

Recently, Ross et al.[38] presented MolFormer, a transformer-based language model trained on 1.1 billion unlabelled SMILES of molecules from the PubChem and ZINC datasets. MoLFormer learned molecular embeddings that outperformed existing baselines, including supervised and self-supervised graph neural networks and language models, across various downstream tasks on ten benchmark datasets[38]. Ross et al.[38] also demonstrated that MolFormer effectively learns spatial relationships between atoms within a molecule from chemical SMILES representations. Here, an available MolFormer checkpoint was used to generate embeddings for all substances in our datasets based on the SMILES. This resulted in feature vectors of length 768.

S7.3 UMAP

UMAP plots were created using the following parameters: n_neighbors=15, min_dist=0.5, target_weight=0.1 (for semi-supervised), metric="manhattan", random_state=42



a) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ unsupervised

b) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ unsupervised with labels

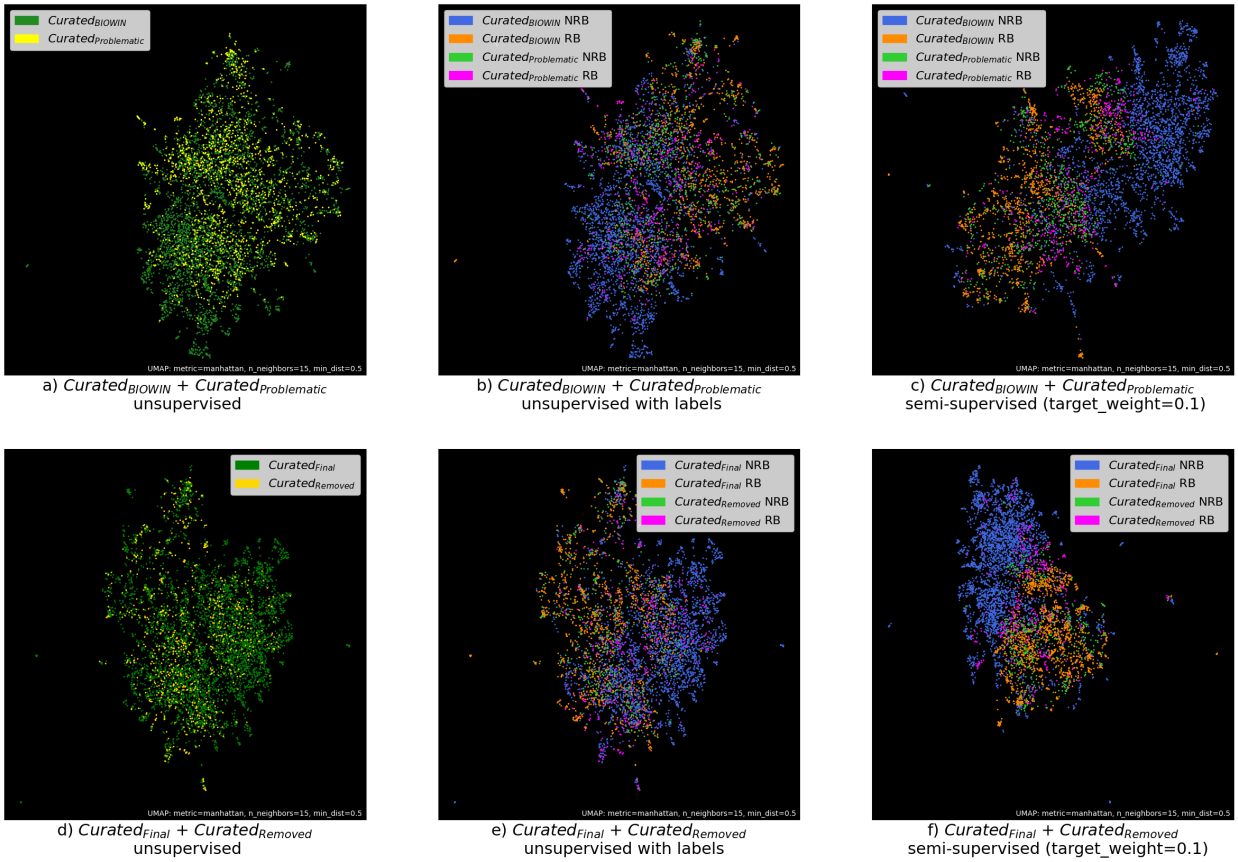
c) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ semi-supervised (target_weight=0.1)

d) Curated$_{Final}$ + Curated$_{Removed}$ unsupervised

e) Curated$_{Final}$ + Curated$_{Removed}$ unsupervised with labels

f) Curated$_{Final}$ + Curated$_{Removed}$ semi-supervised (target_weight=0.1)

Figure S5: UMAP plot for MACCS keys



a) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ unsupervised

b) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ unsupervised with labels

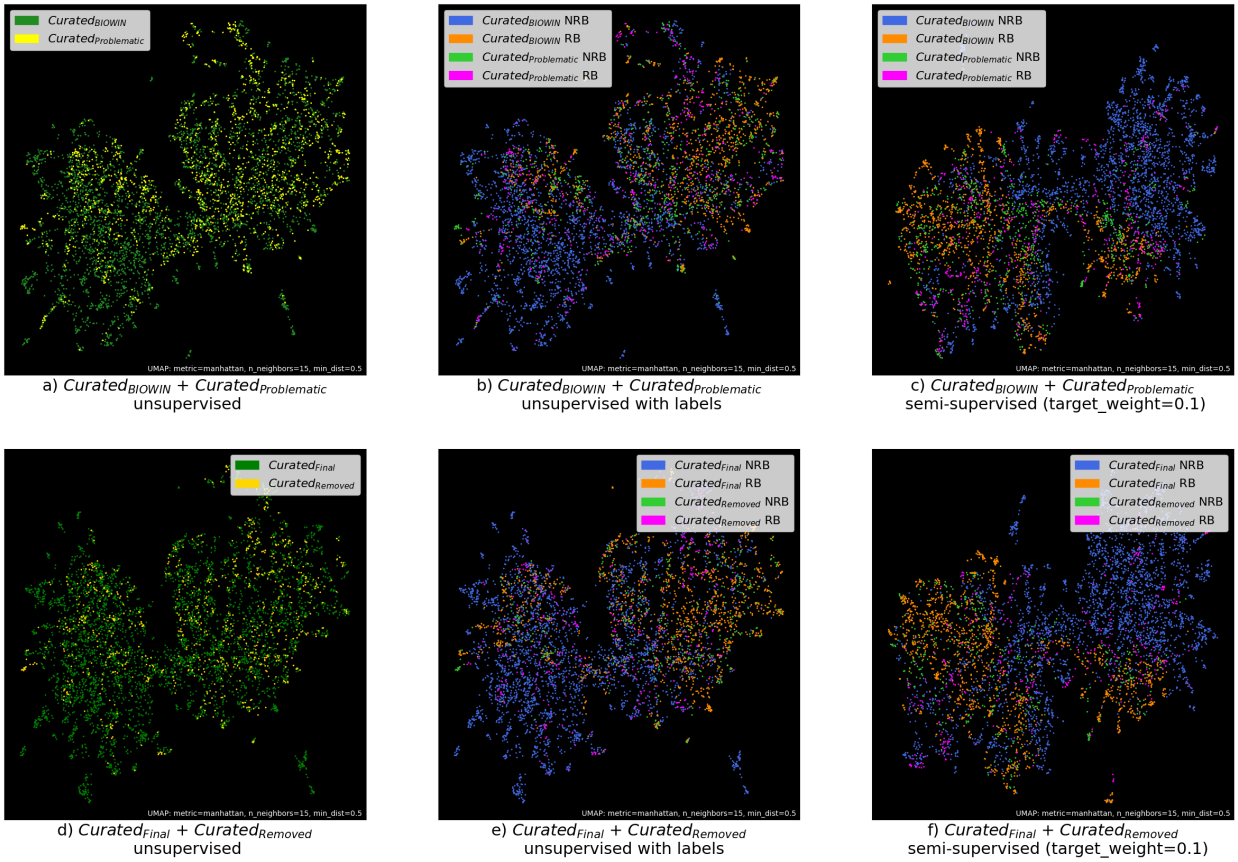
c) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ semi-supervised (target_weight=0.1)

d) $Curated_{Final}$ + $Curated_{Removed}$ unsupervised

e) $Curated_{Final}$ + $Curated_{Removed}$ unsupervised with labels

f) $Curated_{Final}$ + $Curated_{Removed}$ semi-supervised (target_weight=0.1)

Figure S6: UMAP plot for Morgan FPs



a) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ unsupervised

b) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ unsupervised with labels

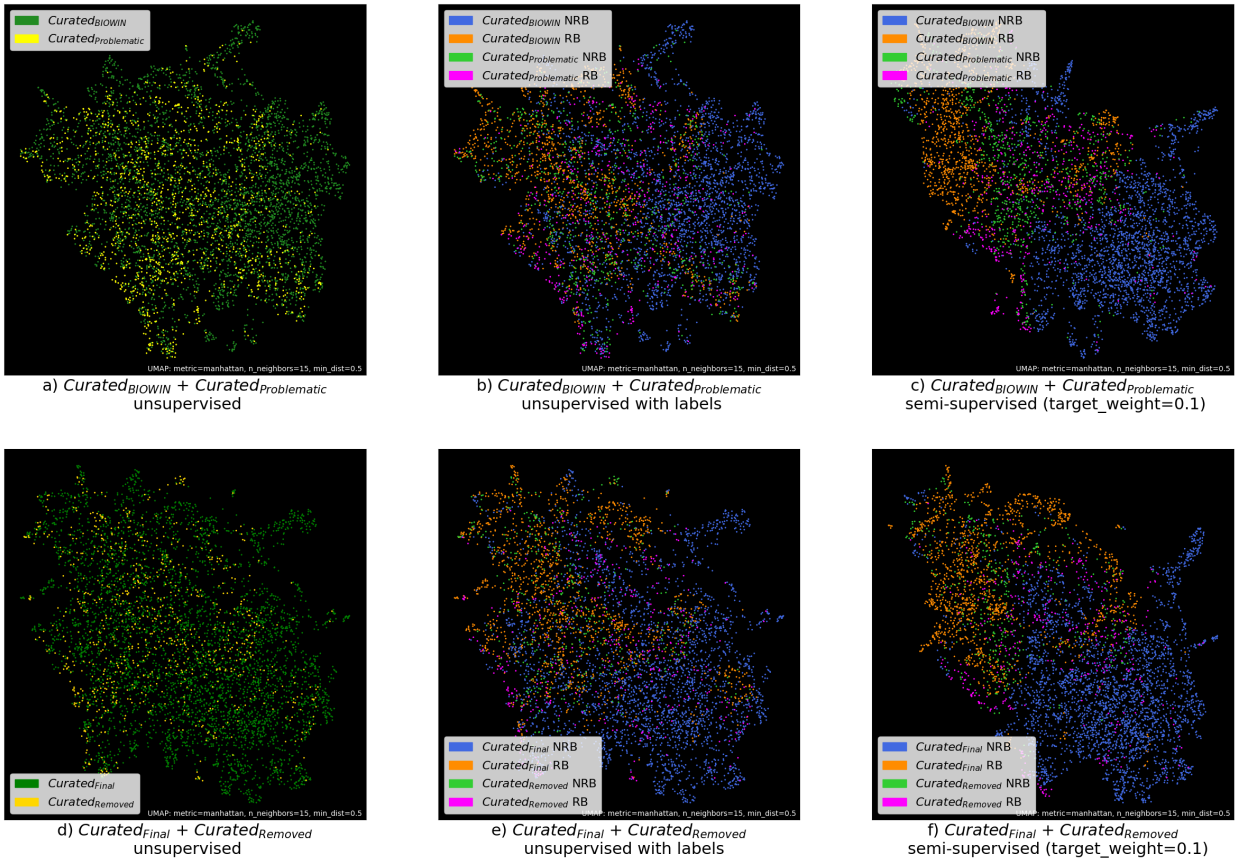
c) $Curated_{BIOWIN}$ + $Curated_{Problematic}$ semi-supervised (target_weight=0.1)

d) $Curated_{Final}$ + $Curated_{Removed}$ unsupervised

e) $Curated_{Final}$ + $Curated_{Removed}$ unsupervised with labels

f) $Curated_{Final}$ + $Curated_{Removed}$ semi-supervised (target_weight=0.1)

Figure S7: UMAP plot for RDKit FPs



a) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ unsupervised

b) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ unsupervised with labels

c) Curated$_{BIOWIN}$ + Curated$_{Problematic}$ semi-supervised (target_weight=0.1)

d) Curated$_{Final}$ + Curated$_{Removed}$ unsupervised

e) Curated$_{Final}$ + Curated$_{Removed}$ unsupervised with labels

f) Curated$_{Final}$ + Curated$_{Removed}$ semi-supervised (target_weight=0.1)

Figure S8: UMAP plot for features generated using MolFormer

S7.4 XGBClassifier performance with other features

Here XGBClassifier was run with default hyperparameters (as suggested by Huang and Zhang[16]) for the four different features.

Table S8: Balanced accuracy, sensitivity, specificity and F1 score for XGBClassifier using the default hyperparameters from Huang and Zhang [16].

Test datasets	Train datasets	Balanced accuracy (%)	Sensitivity (%)	Specificity (%)	F1
MACCS key					
CURATED$_{SCS}$	HUANG-CLASSIFICATION-DATASET	80.9 ± 1.4	74.4 ± 2.1	87.3 ± 0.8	0.75 ± 0.02
	CURATED$_{SCS}$	81.0 ± 1.4	74.7 ± 2.3	87.2 ± 1.0	0.75 ± 0.02
	CURATED$_{BIOWIN}$	78.3 ± 1.5	70.9 ± 2.8	85.6 ± 1.1	0.72 ± 0.02
	CURATED$_{FINAL}$	79.4 ± 1.6	72.1 ± 2.8	86.6 ± 1.6	0.73 ± 0.02
CURATED$_{BIOWIN}$	HUANG-CLASSIFICATION-DATASET	88.1 ± 0.7	83.0 ± 2.0	93.3 ± 1.3	0.84 ± 0.01
	CURATED$_{SCS}$	88.7 ± 1.0	85.1 ± 2.7	92.3 ± 1.2	0.84 ± 0.01
	CURATED$_{BIOWIN}$	94.0 ± 1.0	91.7 ± 1.4	96.2 ± 1.0	0.92 ± 0.02
	CURATED$_{FINAL}$	94.6 ± 0.3	92.2 ± 0.4	97.1 ± 0.6	0.93 ± 0.01
RDKit FPs					
CURATED$_{SCS}$	HUANG-CLASSIFICATION-DATASET	80.6 ± 1.0	73.9 ± 2.2	87.3 ± 1.2	0.75 ± 0.01
	CURATED$_{SCS}$	80.8 ± 1.8	74.3 ± 2.4	87.3 ± 1.7	0.75 ± 0.02
	CURATED$_{BIOWIN}$	78.7 ± 1.8	72.9 ± 3.0	84.5 ± 2.5	0.72 ± 0.02
	CURATED$_{FINAL}$	79.6 ± 1.3	72.7 ± 2.5	86.4 ± 1.6	0.73 ± 0.02
CURATED$_{BIOWIN}$					
	HUANG-CLASSIFICATION-DATASET	88.4 ± 1.3	83.2 ± 3.2	93.6 ± 1.3	0.85 ± 0.02
	CURATED$_{SCS}$	88.0 ± 1.3	83.3 ± 2.8	92.6 ± 1.1	0.84 ± 0.02
	CURATED$_{BIOWIN}$	93.8 ± 0.5	91.7 ± 1.0	95.8 ± 0.5	0.91 ± 0.01
	CURATED$_{FINAL}$	93.4 ± 1.5	90.6 ± 2.4	96.3 ± 0.8	0.91 ± 0.02
Morgan FPs					
CURATED$_{SCS}$	HUANG-CLASSIFICATION-DATASET	79.1 ± 1.3	72.0 ± 2.4	86.1 ± 1.7	0.73 ± 0.02
	CURATED$_{SCS}$	79.2 ± 2.2	72.6 ± 3.0	85.7 ± 2.2	0.73 ± 0.03
	CURATED$_{BIOWIN}$	77.3 ± 1.2	70.5 ± 2.5	84.0 ± 2.0	0.70 ± 0.02
	CURATED$_{FINAL}$	77.7 ± 1.9	69.9 ± 3.6	85.5 ± 1.4	0.71 ± 0.02
CURATED$_{BIOWIN}$	HUANG-CLASSIFICATION-DATASET	86.7 ± 1.0	81.4 ± 1.7	92.0 ± 1.7	0.82 ± 0.02
	CURATED$_{SCS}$	86.8 ± 1.9	82.5 ± 3.7	91.1 ± 2.1	0.82 ± 0.03
	CURATED$_{BIOWIN}$	91.8 ± 1.3	88.8 ± 2.2	94.7 ± 0.5	0.89 ± 0.02
	CURATED$_{FINAL}$	91.2 ± 1.4	87.0 ± 2.7	95.4 ± 0.6	0.88 ± 0.02

MolFormer					
CURATED$_{SCS}$	HUANG-CLASSIFICATION-DATASET	78.3 ± 1.3	73.3 ± 2.4	83.4 ± 1.3	0.72 ± 0.02
	CURATED$_{SCS}$	78.9 ± 1.3	74.4 ± 2.0	83.3 ± 1.5	0.72 ± 0.02
	CURATED$_{BIOWIN}$	77.9 ± 1.7	73.5 ± 3.6	82.2 ± 1.4	0.71 ± 0.02
	CURATED$_{FINAL}$	78.7 ± 1.0	73.4 ± 2.3	84.1 ± 1.5	0.72 ± 0.01
CURATED$_{BIOWIN}$	HUANG-CLASSIFICATION-DATASET	86.0 ± 1.7	80.5 ± 3.4	91.5 ± 0.7	0.81 ± 0.02
	CURATED$_{SCS}$	87.2 ± 1.4	84.1 ± 2.7	90.3 ± 1.7	0.82 ± 0.02
	CURATED$_{BIOWIN}$	92.0 ± 0.6	90.6 ± 1.2	93.4 ± 0.4	0.88 ± 0.01
	CURATED$_{FINAL}$	90.9 ± 0.7	87.4 ± 1.6	94.5 ± 1.2	0.88 ± 0.01

S7.5 LAZYPREDICT with other features

Table S9: LAZYPREDICT, all features created for the CURATED$_{FINAL}$ dataset, test set CURATED$_{SCS}$. ROC AUC stands for Receiver Operating Characteristic Area Under the Curve, and the Time Taken is in seconds.

Best models	**Accuracy**	**Balanced accuracy**	**ROC AUC**	**F1 Score**	**Time Taken**
MACCS key					
MLPClassifier	0.83	0.82	0.82	0.83	2.10
HistGradientBoostingClassifier	0.82	0.81	0.81	0.82	0.56
RandomForestClassifier	0.82	0.81	0.81	0.82	0.36
GradientBoostingClassifier	0.81	0.80	0.80	0.82	0.97
ExtraTreesClassifier	0.82	0.80	0.80	0.82	0.32
RDKit FPs					
LogisticRegressionCV	0.83	0.82	0.82	0.83	5.96
LogisticRegression	0.82	0.81	0.81	0.82	0.59
PassiveAggressiveClassifier	0.82	0.81	0.81	0.82	1.07
Perceptron	0.81	0.81	0.81	0.81	0.49
MLPClassifier	0.82	0.80	0.80	0.82	11.95

Morgan FPs					
ExtraTreesClassifier	0.81	0.79	0.79	0.81	0.95
HistGradientBoostingClassifier	0.81	0.78	0.78	0.80	2.04
MLPClassifier	0.80	0.78	0.78	0.80	3.78
XGBClassifier	0.80	0.78	0.78	0.80	0.43
LogisticRegressionCV	0.79	0.78	0.78	0.79	1.19
MolFormer					
SVC	0.82	0.81	0.81	0.82	1.57
NuSVC	0.79	0.80	0.80	0.80	2.92
MLPClassifier	0.81	0.80	0.80	0.81	2.96
HistGradientBoostingClassifier	0.81	0.80	0.80	0.81	2.95
RandomForestClassifier	0.80	0.79	0.79	0.80	6.33

Table S10: LazyPredict, all features created for the $\text{Curated}_{\text{FINAL}}$ dataset, test set $\text{Curated}_{\text{BIOWIN}}$ dataset. ROC AUC stands for Receiver Operating Characteristic Area Under the Curve, and the Time Taken is in seconds.

Best models	**Accuracy**	**Balanced accuracy**	**ROC AUC**	**F1 Score**	**Time Taken**
MACCS key					
RandomForestClassifier	0.95	0.94	0.94	0.95	0.36
XGBClassifier	0.95	0.94	0.94	0.95	0.18
HistGradientBoostingClassifier	0.95	0.94	0.94	0.95	0.78
ExtraTreesClassifier	0.94	0.93	0.93	0.94	0.35
GradientBoostingClassifier	0.94	0.93	0.93	0.94	1.11
RDKit FPs					
MLPClassifier	0.94	0.93	0.93	0.94	6.81
HistGradientBoostingClassifier	0.94	0.93	0.93	0.94	3.82
XGBClassifier	0.94	0.93	0.93	0.94	2.28
LogisticRegressionCV	0.93	0.93	0.93	0.93	5.17
SVC	0.92	0.92	0.92	0.92	4.41
Morgan FPs					
ExtraTreesClassifier	0.93	0.92	0.92	0.93	1.05
SVC	0.92	0.90	0.90	0.92	2.51
MLPClassifier	0.91	0.90	0.90	0.91	4.65
HistGradientBoostingClassifier	0.92	0.90	0.90	0.92	2.27
RandomForestClassifier	0.92	0.90	0.90	0.92	0.92
MolFormer					
MLPClassifier	0.94	0.92	0.92	0.94	3.25
SVC	0.93	0.92	0.92	0.93	1.58
XGBClassifier	0.93	0.92	0.92	0.93	1.72
PassiveAggressiveClassifier	0.91	0.91	0.91	0.92	0.41
RidgeClassifier	0.90	0.91	0.91	0.91	0.42

Table S11: After hyperparameter tuning, all features created for the CURATED$_{\text{FINAL}}$ dataset, test set CURATED$_{\text{SCS}}$

Best models	Best hyperparameters	Balanced accuracy	Sensitivity	Specificity	F1
MACCS key					
MLPClassifier	activation: relu alpha: 0.01443 early_stopping: True hidden_layer_sizes: 218 learning_rate_init: 0.03084 max_iter: 600 random_state: 42 solver: adam	79.4 ± 1.6	73.2 ± 3.3	85.6 ± 1.7	0.73 ± 0.02
HistGradient-BoostingClassifier	learning_rate: 0.07371 max_iter: 200 max_leaf_nodes: 40 min_samples_leaf: 2 random_state: 42	79.5 ± 1.8	72.3 ± 2.9	86.6 ± 1.6	0.73 ± 0.02
RandomForest-Classifier	criterion: gini max_features: sqrt min_samples_leaf: 1 min_samples_split: 2 n_estimators: 2208 random_state: 42	79.3 ± 1.1	71.9 ± 2.0	86.7 ± 1.0	0.73 ± 0.02
GradientBoosting-Classifier	criterion: squared_error learning_rate: 0.4 loss: exponential max_depth: 5 max_features: log2 n_estimators: 120 random_state: 42	78.4 ± 1.7	83.7 ± 2.5	73.2 ± 1.4	0.71 ± 0.02
ExtraTrees-Classifier	criterion: entropy max_depth: None max_features: sqrt min_samples_leaf: 1 min_samples_split: 2 n_estimators: 50 random_state: 42	78.2 ± 0.8	78.1 ± 1.2	78.3 ± 1.0	0.71 ± 0.01
RDKit FPs					
Logistic-RegressionCV	max_iter: 131 random_state: 42 solver: sag	80.4 ± 1.8	75.8 ± 2.7	85.0 ± 1.7	0.74 ± 0.02

Logistic-Regression	max_iter: 117 multi_class: ovr random_state: 42 solver: sag	80.6 ± 1.5	77.0 ± 2.6	84.1 ± 2.1	0.74 ± 0.02
PassiveAggressive-Classifier	early_stopping: True max_iter: 1091 random_state: 42	80.0 ± 1.7	76.5 ± 3.0	83.4 ± 1.4	0.74 ± 0.02
Perceptron	alpha: 0.00042 early_stopping: True max_iter: 1173 random_state: 42	79.5 ± 1.8	75.3 ± 2.1	83.7 ± 3.1	0.73 ± 0.02
MLPClassifier	activation: relu alpha: 0.01238 early_stopping: True hidden_layer_sizes: 120 learning_rate_init: 0.08337 max_iter: 559 random_state: 42 solver: lbfgs	79.8 ± 1.7	74.0 ± 2.4	85.6 ± 1.9	0.74 ± 0.02
Morgan FPs					
ExtraTrees-Classifier	criterion: gini max_depth: None max_features: log2 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 483 random_state: 42	79.0 ± 1.5	71.4 ± 2.8	86.7 ± 1.3	0.73 ± 0.02
HistGradient-BoostingClassifier	learning_rate: 0.09216 max_iter: 161 max_leaf_nodes: 40 min_samples_leaf: 2 random_state: 42	78.5 ± 1.3	70.0 ± 2.9	86.9 ± 1.1	0.72 ± 0.02
MLPClassifier	activation: relu alpha: 1e-06 early_stopping: True hidden_layer_sizes: 250 learning_rate_init: 0.06897 max_iter: 200 random_state: 42 solver: lbfgs	78.3 ± 1.5	71.5 ± 3.0	85.2 ± 0.6	0.72 ± 0.02
XGBClassifier	alpha: 2.23031 base_score: 0.54555	78.3 ± 1.4	68.5 ± 2.6	88.2 ± 1.1	0.72 ± 0.02

	booster: gbtree colsample_bylevel: 0.79474 colsample_bynode: 0.36806 colsample_bytree: 0.82424 gamma: 0.03574 lambda: 0.46975 learning_rate: 0.25053 max_delta_step: 3.29558 max_depth: 210 min_child_weight: 1.25367 n_estimators: 4907 num_parallel_tree: 21 scale_pos_weight: 0.69473 subsample: 0.86006 tree_method: exact validate_parameters: True				
Logistic-RegressionCV	max_iter: 118 random_state: 42 solver: saga	77.7 ± 1.0	71.7 ± 2.1	83.7 ± 1.0	0.71 ± 0.01
MolFormer					
SVC	degree: 3 gamma: scale kernel: poly random_state: 42	80.1 ± 1.2	78.4 ± 1.5	81.7 ± 2.1	0.74 ± 0.02
NuSVC	degree: 2 kernel: rbf random_state: 42	78.5 ± 0.9	82.3 ± 2.0	74.8 ± 1.5	0.72 ± 0.01
MLPClassifier	activation: relu alpha: 0.07908 early_stopping: True hidden_layer_sizes: 161 learning_rate_init: 0.1 max_iter: 285 random_state: 42 solver: lbfgs	79.3 ± 1.1	74.0 ± 2.2	84.6 ± 1.3	0.73 ± 0.01
HistGradient-BoostingClassifier	learning_rate: 0.30261 max_iter: 200 max_leaf_nodes: 15 min_samples_leaf: 25 random_state: 42	79.2 ± 1.5	73.3 ± 2.3	85.1 ± 1.8	0.73 ± 0.02

RandomForest-Classifier	criterion: entropy max_features: log2 min_samples_leaf: 1 min_samples_split: 4 n_estimators: 2024 random_state: 42	78.6 ± 1.4	73.2 ± 2.6	84.0 ± 1.5	0.72 ± 0.02

Table S12: After hyperparameter tuning, all features created for the CURATED$_{\text{FINAL}}$ dataset and tested on the CURATED$_{\text{BIOWIN}}$ dataset

Best models	**Best hyperparameters**	**Balanced accuracy**	**Sensitivity**	**Specificity**	**F1**
MACCS key					
RandomForest-Classifier	criterion: log_loss max_features: sqrt min_samples_leaf: 1 min_samples_split: 2 n_estimators: 1153 random_state: 42	94.3 ± 0.7	91.7 ± 1.1	96.8 ± 0.7	0.92 ± 0.01
XGBClassifier	alpha: 1.54138 base_score: 0.53733 booster: gbtree colsample_bylevel: 0.87768 colsample_bynode: 0.24143 colsample_bytree: 0.96502 gamma: 0.001 lambda: 0.55188 learning_rate: 0.71465 max_delta_step: 2.87329 max_depth: 211 min_child_weight: 2.84452 n_estimators: 5356 num_parallel_tree: 24 scale_pos_weight: 0.86930 subsample: 0.99 tree_method: exact validate_parameters: True	94.4 ± 0.6	91.6 ± 0.7	97.2 ± 0.6	0.93 ± 0.01
HistGradient-BoostingClassifier	learning_rate: 0.07890 max_iter: 200 max_leaf_nodes: 29 min_samples_leaf: 2 random_state: 42	94.8 ± 0.2	92.2 ± 0.6	97.4 ± 0.5	0.93 ± 0.00
ExtraTrees-Classifier	criterion: log_loss max_depth: None	94.4 ± 0.6	92.0 ± 0.8	96.8 ± 0.6	0.93 ± 0.01

	max_features: sqrt min_samples_leaf: 1 min_samples_split: 3 n_estimators: 1000 random_state: 42				
GradientBoosting-Classifier	criterion: friedman_mse learning_rate: 0.37841 loss: log_loss max_depth: 5 max_features: sqrt n_estimators: 119 random_state: 42	94.4 ± 0.8	92.2 ± 1.1	96.7 ± 1.0	0.93 ± 0.01
RDKit FPs					
MLPClassifier	activation: relu alpha: 0.00122 early_stopping: True hidden_layer_sizes: 249 learning_rate_init: 0.03344 max_iter: 389 random_state: 42 solver: adam	93.1 ± 1.5	89.8 ± 3.1	96.3 ± 0.7	0.91 ± 0.02
HistGradient-BoostingClassifier	learning_rate: 0.28092 max_iter: 60 max_leaf_nodes: 29 min_samples_leaf: 24 random_state: 42	93.7 ± 1.3	90.6 ± 2.3	96.7 ± 0.6	0.92 ± 0.02
XGBClassifier	alpha: 2.23031 base_score: 0.54555 booster: gbtree colsample_bylevel: 0.79474 colsample_bynode: 0.36806 colsample_bytree: 0.82424 gamma: 0.03574 lambda: 0.46975 learning_rate: 0.25053 max_delta_step: 3.29558 max_depth: 210 min_child_weight: 1.25367 n_estimators: 4907 num_parallel_tree: 21 scale_pos_weight: 0.69473 subsample: 0.86006 tree_method: exact validate_parameters: True	93.2 ± 1.6	89.3 ± 2.6	97.0 ± 0.7	0.91 ± 0.02

Logistic-RegressionCV	max_iter: 194 random_state: 42 solver: saga	93.4 ± 1.6	92.0 ± 2.9	94.8 ± 0.7	0.91 ± 0.02
SVC	degree: 2 gamma: scale kernel: linear random_state: 42	91.4 ± 1.2	89.5 ± 2.5	93.2 ± 0.3	0.88 ± 0.01
Morgan FPs					
ExtraTrees-Classifier	criterion: log_loss max_depth: None max_features: log2 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 1000 random_state: 42	92.1 ± 1.8	88.1 ± 2.9	96.1 ± 0.7	0.90 ± 0.02
SVC	degree: 2 gamma: scale kernel: rbf random_state: 42	92.4 ± 1.7	88.9 ± 2.6	96.0 ± 0.9	0.90 ± 0.02
MLPClassifier	activation: relu alpha: 0.1 early_stopping: True hidden_layer_sizes: 250 learning_rate_init: 0.02689 max_iter: 200 random_state: 42 solver: lbfgs	91.4 ± 1.4	88.1 ± 2.5	94.7 ± 0.9	0.88 ± 0.02
HistGradient-BoostingClassifier	learning_rate: 0.11409 max_iter: 148 max_leaf_nodes: 40 min_samples_leaf: 25 random_state: 42	92.1 ± 1.2	88.1 ± 2.3	96.2 ± 0.6	0.90 ± 0.01
RandomForest-Classifier	criterion: log_loss max_features: log2 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 1056 random_state: 42	91.1 ± 1.4	85.7 ± 2.5	96.5 ± 0.6	0.89 ± 0.02

MolFormer					
MLPClassifier	activation: tanh alpha: 0.1 early_stopping: True hidden_layer_sizes: 130 learning_rate_init: 0.06731 max_iter: 414 random_state: 42 solver: lbfgs	93.2 ± 0.8	90.9 ± 1.1	95.4 ± 1.2	0.91 ± 0.01
SVC	degree: 2 gamma: scale kernel: linear random_state: 42	89.4 ± 1.2	86.1 ± 2.5	92.7 ± 1.3	0.85 ± 0.02
XGBClassifier	alpha: 1.43587 base_score: 0.48343 booster: gbtree colsample_bylevel: 0.80991 colsample_bynode: 0.37732 colsample_bytree: 0.85652 gamma: 0.01834 lambda: 0.10876 learning_rate: 0.02336 max_delta_step: 3.83059 max_depth: 134 min_child_weight: 2.08288 n_estimators: 5364 num_parallel_tree: 23 scale_pos_weight: 0.50767 subsample: 0.83728 tree_method: exact validate_parameters: True	91.5 ± 1.0	86.9 ± 2.0	96.1 ± 0.4	0.89 ± 0.01
PassiveAggressive-Classifier	early_stopping: True max_iter: 1091 random_state: 42	88.7 ± 3.3	85.1 ± 9.3	92.4 ± 3.1	0.84 ± 0.03
RidgeClassifier	random_state: 42 solver: sparse_cg	89.4 ± 0.9	89.0 ± 1.6	89.7 ± 1.7	0.84 ± 0.01

S7.6 Comparing the AD

The similarity threshold here was set based on the share of data points in each similarity category. All datasets had <1% of substances with a similarity score below 0.5, and it was therefore decided to set the AD threshold to 0.5. The similarity threshold was not based on the expected accuracy because we didn't see a clear drop in the accuracy. However, the low number of substances with a similarity score below 0.5 meant that the determined accuracies of the similarity categories below 0.5 might not be robust.

Table S13: Results of defining the similarity threshold for the HUANG-CLASSIFICATION-DATASET. The dashed line indicates the determined similarity threshold.

	HUANG-CLASSIFICATION-DATASET reported		**HUANG-CLASSIFICATION-DATASET replicated**	
Similarity	**Expected accuracy**	**Share of datapoints**	**Expected accuracy**	**Share of datapoints**
0.9 to <1.0	88.9%	-	84.0 ± 0.8%	39.1%
0.8 to <0.9	87.1%	-	81.1 ± 2.5%	27.5%
0.7 to <0.8	86.3%	-	81.3 ± 2.2%	20.7%
0.6 to <0.7	85.6%	-	78.6 ± 2.0%	9.8%
0.5 to <0.6	85.1%	-	76.4 ± 6.3%	2.3%
0.4 to <0.5	Out of AD	-	76.0 ± 14.6%	0.5%
<0.4	Out of AD	-	50.0 ± 0.0%	0.1%

Table S14: Results of defining the similarity threshold for the CURATED$_{SCS}$, CURATED$_{BIOWIN}$, CURATED$_{FINAL}$ dataset. The dashed line indicates the determined similarity threshold.

	CURATED$_{SCS}$		**CURATED$_{BIOWIN}$**		**CURATED$_{FINAL}$**	
Similarity	**Expected accuracy**	**Share of datapoints**	**Expected accuracy**	**Share of datapoints**	**Expected accuracy**	**Share of datapoints**
0.9 to <1.0	84.9 ± 1.1%	37.0%	98.3 ± 0.8%	36.7%	98.2 ± 0.3%	36.8%
0.8 to <0.9	80.0 ± 2.5%	27.8%	94.0 ± 2.4%	25.8%	94.0 ± 1.3%	26.2%
0.7 to <0.8	79.3 ± 2.3%	21.7%	92.9 ± 2.2%	21.2%	92.5 ± 2.4%	21.6%
0.6 to <0.7	84.4 ± 3.0%	10.5%	92.5 ± 1.5%	12.5%	93.0 ± 3.4%	11.9%
0.5 to <0.6	81.0 ± 5.6%	2.4%	89.8 ± 7.4%	3.0%	91.9 ± 4.3%	2.9%
0.4 to <0.5	100.0 ± 0.0%	0.5%	66.7 ± 40.8%	0.7%	89.3 ± 15.3%	0.6%
<0.4	25.0 ± 35.4%	0.1%	66.7 ± 57.7%	0.1%	100.0 ± 0.0%	0.0%

Table S15: Share of substances in the DSSTox dataset that are in the AD of the classification model presented by Huang and Zhang[16] and the classifiers trained on the CURATED$_{SCS}$, CURATED$_{BIOWIN}$, CURATED$_{FINAL}$.

Dataset	**Percent in AD**
HUANG-CLASSIFICATION-DATASET (reported)	98.4%
CURATED$_{SCS}$	97.9%
CURATED$_{BIOWIN}$	97.3%
CURATED$_{FINAL}$	97.7%

References

(1) CAS CAS REGISTRY and CAS Registry Number FAQs. https://www.cas.org/support/documentation/chemical-substances/faqs, 2023; Accessed: 2023-05-20.

(2) National Center for Biotechnology Information PubChem Compound Summary for CID 441328, 7-Aminocephalosporanic acid. https://pubchem.ncbi.nlm.nih.gov/compound/441328#section=CAS, 2023; Accessed: 2023-09-01.

(3) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(4) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.

(5) Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *Journal of chemical information and computer sciences* **1990**, *30*, 237–243.

(6) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* **2020**, *12*, 1–22.

(7) O'Boyle, N. M. Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics* **2012**, *4*, 1–14.

(8) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI-the worldwide chemical structure identifier standard. *Journal of cheminformatics* **2013**, *5*, 1–9.

(9) Glüge, J.; McNeill, K.; Scheringer, M. Getting the SMILES right: identifying inconsis-

tent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard. *Environmental Science: Advances* **2023**, *2*, 612–621.

(10) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* **2015**, *7*, 1–34.

(11) Sree, K. S.; Karthik, J.; Niharika, C.; Srinivas, P.; Ravinder, N.; Prasad, C. Optimized conversion of categorical and numerical features in machine learning models. 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). 2021; pp 294–299.

(12) Domingos, P. A few useful things to know about machine learning. *Communications of the ACM* **2012**, *55*, 78–87.

(13) Kaufman, S.; Rosset, S.; Perlich, C. Leakage in data mining: Formulation, detection, and avoidance. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2011**, 556–563.

(14) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(15) Statology What is Balanced Accuracy? (Definition & Example). 2021; `https://www.statology.org/balanced-accuracy/`.

(16) Huang, K.; Zhang, H. Classification and Regression Machine Learning Models for Predicting Aerobic Ready and Inherent Biodegradation of Organic Chemicals in Water. *Environmental Science & Technology* **2022**, *56*, 12755–12764.

(17) Northcutt, C. G.; Athalye, A.; Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* **2021**,

(18) Frénay, B.; Verleysen, M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* **2013**, *25*, 845–869.

(19) Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems* **2015**, *145*, 22–29.

(20) Gramatica, P. Principles of QSAR modeling: comments and suggestions from personal experience. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2020**, *5*, 61–97.

(21) Kar, S.; Roy, K.; Leszczynski, J. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. *Computational toxicology: methods and protocols* **2018**, 141–169.

(22) Dias, J. R. pyADA. `https://github.com/jeffrichardchemistry/pyADA`, 2023; Accessed: 2023-09-01.

(23) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **2015**, *7*, 1–13.

(24) Liu, R.; Wallqvist, A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *Journal of Chemical Information and Modeling* **2018**, *59*, 181–189.

(25) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group contribution method for predicting probability and rate of aerobic biodegradation. *Environmental Science & Technology* **1994**, *28*, 459–465.

(26) Salkinoja-Salonen, M.; Uotila, J.; Jokela, J.; Laine, M.; Saski, E. Organic halogens in the environment: studies of environmental biodegradability and human exposure. *Environmental Health Perspectives* **1995**, *103*, 63–69.

(27) Pimviriyakul, P.; Wongnate, T.; Tinikul, R.; Chaiyen, P. Microbial degradation of halogenated aromatics: molecular mechanisms and enzymatic reactions. *Microbial Biotechnology* **2020**, *13*, 67–86.

(28) Mohn, W. W. *Biodegradation and bioremediation*; Springer, 2004; pp 125–148.

(29) Alexander, M. Biotechnology report. Nonbiodegradable and other recalcitrant molecules. *Biotechnol. Bioeng* **1973**, *15*.

(30) Loonen, H.; Lindgren, F.; Hansen, B.; Karcher, W.; Niemelä, J.; Hiromatsu, K.; Takatsuki, M.; Peijnenburg, W.; Rorije, E.; Struijś, J. Prediction of biodegradability from chemical structure: modeling of ready biodegradation test data. *Environmental Toxicology and Chemistry: An International Journal* **1999**, *18*, 1763–1768.

(31) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

(32) Ding, Y.; Chen, M.; Guo, C.; Zhang, P.; Wang, J. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. *Journal of Molecular Liquids* **2021**, *326*, 115212.

(33) Ying, C. Z.; Leong, C. H.; Alvin, L. W. X.; Yang, C. J. Improving the Workflow of Chemical Structure Elucidation with Morgan Fingerprints and the Tanimoto Coefficient. IRC-SET 2020: Proceedings of the 6th IRC Conference on Science, Engineering and Technology, July 2020, Singapore. 2021; pp 13–24.

(34) RDKit Morgan Fingerprints (Circular Fingerprints). `https://rdkit.readthedocs.io/en/latest/GettingStartedInPython.html`.

(35) RDKit: open-source cheminformatics. `http://www.rdkit.org`.

(36) Daylight Theory Manual. `https://www.daylight.com/dayhtml/doc/theory/index.pdf`.

(37) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics* **2013**, *5*, 26.

(38) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **2022**, *4*, 1256–1264.