RSC Sustainability

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: T. ding, G. Larrea-Gallegos, F. Busio, A. Marvuglia and T. Schaubroeck, *RSC Sustainability*, 2026, DOI: 10.1039/D6SU00023A.

ROYAL SOCIETY OF CHEMISTRY

View Article Online
DOI: 10.1039/D6SU00023A

Sustainability Spotlight Statement

This study advances sustainable chemical management by deriving models, based on machine and deep learning, to predict toxicity characterization factors for chemicals previously without in the EU Environmental Footprint methodology. This allows for animal-free toxicity assessment directly from molecular structure, addressing critical data gaps in Life Cycle Assessment (LCA) and Safe and Sustainable by Design (SSbD), directly aligning with potential future EU policy. The prediction framework supports more complete and precautionary impact accounting, reducing the risk of burden shifting and underestimated toxicity. The approach facilitates faster (high-throughput) screening and substitution of hazardous substances, contributing to safer product design and more transparent environmental decision-making. It directly aligns with the United Nations Sustainable Development Goals, particularly SDG 3 (Good Health and Well-Being) through improved chemical safety and SDG 12 (Responsible Consumption and Production) via informed material & product selection.

View Article Online
DOI: 10.1039/D6SU00023A

ARTICLE

Characterizing Chemical Toxicity for Life Cycle Assessment Using Machine Learning and Deep Learning Models Based on Environmental Footprint – Methodological Comparison & textile case study

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Tianran Ding,*[a] Gustavo Larrea-Gallegos [a], Federico Busio [a] , Antonino Marvuglia [a], & Thomas Schaubroeck *[a]

The rapid expansion of registered chemicals, coupled with persistent data gaps, poses a major challenge for toxicity assessment in Life Cycle Assessment (LCA) and Safe and Sustainable by Design (SSbD). This study proposes a data-driven framework to directly predict toxicity characterization factors (CFs) from molecular Simplified Molecular Input Line Entry System (SMILES), using the Environmental Footprint (EF) v3.1 database as the training benchmark. We evaluate five machine-learning and deep-learning approaches—random forest, XGBoost, Gaussian process, deep neural networks, and graph neural networks via message-passing neural networks (MPNN)—across three molecular representations: Mordred descriptors, molecular graphs, and large-scale pretrained molecular embeddings (GROVER). Predictive performance is strongly target-dependent, with ecotoxicity CFs showing consistently higher predictability (R^2 = 0.47–0.67) than human toxicity CFs (R^2 = 0.44–0.56). Mordred-based models, particularly XGBoost, shows robust better performance across multiple targets. Graph-based MPNNs achieved competitive performance, with graph-only multi-target MPNNs showing the clearest benefit over single-target training, especially for human toxicity targets. Adding Mordred descriptors to graph-based model generally improved human toxicity prediction, but can sometimes reduced performance for ecotoxicity prediction. GROVER embeddings provided advantages in specific clusters (e.g., highest mean R^2 over three run for one cluster is 0.70) and offer a promising alternative to handcrafted descriptors. The framework further integrates applicability domain analysis and chemical clustering to enable domain-consistent prediction. A textile-sector case study shows that incorporating predicted CFs for previously uncovered chemicals can lead to substantial underestimation of toxicity impacts—by up to about 60%.

Keywords: life cycle assessment (LCA); toxicity; characterization factors (CF), machine learning; deep learning; safe and sustainable by design; Environmental Footprint (EF)

Abbreviations: AD, applicability domain; CF, characterization factor; CNN, convolutional neural network; DL, deep learning; DNN, deep neural network; EC20, effect concentration affecting 20% of species; EC50, effect concentration affecting 50% of species; EF, Environmental Footprint; GNN, graph neural network; GP, Gaussian process; GROVER, Graph Representation frOm self-supervised mEssage passing tRansformer; HC50, hazard concentration affecting 50% of species; k-NN, k-nearest neighbors; LCA, life cycle assessment; LCIA, life cycle impact assessment; ML, machine learning; MLP, multilayer perceptron; MPNN, message passing neural network; PCA, principal component analysis; QSAR, quantitative structure–activity relationship; RF, random forest; RMSE, root mean squared error; SD, standard deviation; SI, Supporting Information; SMILES, Simplified Molecular Input Line Entry System; SSbD, Safe and Sustainable by Design; SVM, support vector machine; USEtox, UNEP-SETAC toxicity model; XGBoost, extreme gradient boosting.

Introduction

With an expanded list of chemicals identified in environmental media, and the rapid growth of newly registered substances, the issue of "chemicals of emerging concern" has remained critical since its initial recognition in the early 2000s. [1] While considerable efforts have been devoted to technologies for monitoring and removing hazardous contaminants, [2] efficient assessment of their toxicity is also urgently needed to support rapid chemical screening and the prioritization of safer

[a] *Environmental Sustainability Assessment and Circularity (SUSTAIN) unit, Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg*
[b] *† Footnotes relating to the title and/or authors should appear here.*
Supplementary Information available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/D6SU00037A

alternatives. This need is particularly pressing in regulatory and design-oriented context, such as Life Cycle Assessment (LCA) and Safe and Sustainable by Design (SSbD), where large chemical inventories need to be assessed to select non-hazaradous or less hazardous chemical alternatives despite substantial data gap. [3–5]

In this context, *in silico* data-driven approaches, especially machine learning (ML) and deep learning (DL), have emerged as powerful tools due to their capacity to approximate almost any kind of function with few apriori assumptions. [6] Such data-driven strategies have been used to accelerate computational chemistry and materials-property prediction. [7] Various researches have predicted toxicity related parameters based on well-known ML algorithms such as Random Forests (RF), [8] k-Nearest Neighbors (k-NN), [9] Gaussian Process (GP), [10] and other ensemble classifiers such as Extreme Gradient Boosting (XGBoost), [11] and DL models such as Multilayer Perceptron (MLP), [12] Deep Neural Network (DNN), [13] Convolutional Neural Network (CNN), [14] and Graph Neural Network (GNN). [15] By leveraging molecular descriptors derived from chemical structure (e.g., atom counts, topological indices), physicochemical properties (e.g., molecular weight, partition coefficients) and chemical graph information, these ML and DL models offer faster, cost-effective, and animal-free alternatives to traditional experimental testing, which is often slow, resource-intensive, and subject to increasing ethical restrictions. [16]

Recent research points out a growing role of DL architectures is observed in quantitative structure–activity relationship (QSAR) to predict both acute and chronic toxicity endpoints. [16] Crucially, while traditional ML approaches, e.g., RF, K-NN, and SVM, have been widely applied to toxicity prediction, their performance is often limited by a strong dependence on manually engineered features and a reduced capacity to capture the complex nonlinear relationships intrinsic to chemical data. In contrast, DL models offer a transformative potential by employing expressive architectures capable of automatically learning and integrating rich patterns from data. [16] For example, GNNs learn molecular representations directly from graph structures and has been successfully used for a wide range of chemical property and toxicity prediction tasks. [15] These GNN are most often constructed as a message-passing neural networks (MPNNs), which learn molecular embeddings by iteratively propagating and aggregating information across the molecular graph, enabling the model to capture local and global structural patterns without reliance on predefined descriptors. [17]

In addition, self-supervised pre-training strategies based on transformer-style architectures have been proposed mainly to address key limitations of GNNs, notably their reliance on large quantities of labelled data and their limited generalization to newly synthesized molecules. By learning transferable molecular representations from large unlabelled chemical corpora, these models, e.g., GROVER, [18] substantially improve data efficiency and predictive robustness, and have been shown to outperform existing state-of-the-art approaches across a range of molecular property and toxicity prediction tasks.

In the LCA field, these state-of-the-art advancement in ML and DL can be especially useful as LCA is data intensive, requiring accounting for the different environmental impacts through the whole life cycle of the related product that contains or emits vast amounts of chemical-related substances. When conducting LCA and characterizing chemical toxicity, the consensus method USEtox [19] provides around 3,000 characterization factors (CFs) for ecotoxicity and human toxicity categories. A recent update of the Environmental Footprint (EF) version 3.1, largely based on the USEtox methodology and recommended (2013/179/EU) [20] to be used as the standard method for environmental claims of all products on the EU market, has doubled the coverage to 6,459 chemicals for toxicity characterization by incorporating new data sources. However, this represents only a small fraction of the estimated 350,000 chemicals registered for large-scale use. [21] Missing CFs lead to uncharacterised elementary flows, which results in systematic underestimation of toxicity impacts in LCA studies. [22] Improving the coverage of toxicity CFs is therefore needed to enhance the completeness and interpretability of life cycle impact assessment results. [22]

To fill this gap in LCA, researchers have employed different ML models such as RF, [23–25] SVM and XGBoost, [24] and DNN [24,26] These works primarily follow a two-step parameter-centric paradigm, in which ML models are first trained to predict intermediate missing parameters or metrics used in the USEtox framework, [27] such as fate factors, intake fractions, [28] species sensitivity distribution, [29] ecotoxicological effect (e.g., HC50, hazard concentration of a chemical at which50% of the freshwater species face harmful chemical exposure), [23,24,30] and then feed them into the USEtox method to derive CFs. This two-step paradigm offers a mechanistically interpretable and modular pathway that aligns well with the USEtox framework. Especially as some intermediate parameters can be data-efficient and empirically grounded, predicting intermediate parameters allows training data with larger and more diverse training datasets, which can improve model robustness. However, this approach is ad hoc and cannot be easily automated since there is no guarantee that the missing physicochemical properties are available or predicted robustly. [31] In practice, missing or poorly predicted intermediate parameters can prevent CF estimation altogether, undermining automation and limiting applicability to large or emerging chemical inventories. Moreover, most existing studies rely on handcrafted physicochemical descriptors and conventional ML models, with limited exploration of modern molecular representation learning approaches, such as GNN and large-scale pretrained molecular embeddings. As a result, the potential of state-of-the-art ML and DL architectures to learn transferable, structure-aware representations for direct CF prediction remains largely unexplored.

Building on these limitations, the aim of this work is threefold. Firstly, we propose a ML and DL-based framework for predicting toxicity CFs from molecular information, with the goal of enabling an automatable workflow for high-throughput chemical screening. Secondly, using the toxicity characterization space defined by EF3.1, which comprises around twice as many data entries as the previous studies that

relied on the USEtox dataset only we systematically benchmark a diverse set of ML models and molecular representations. These include conventional descriptor-based approaches using physicochemical descriptors, GNN that operate directly on chemical structures, and large-scale pretrained molecular embeddings (GROVER). Third, we then apply the trained model(s) to predict missing CFs in a textile-related case study, illustrating how directly estimated toxicity impacts can be integrated into life cycle assessment to support chemical prioritization and safer material selection in an industrially relevant context.

Materials and Methods

Data Sources and Processing

Environmental Footprint Characterization Factors. The training data were limited to 5899 organic data, which were collected from the EF v3.1 database (Figure 1). This database calculated CF with the same method as in USEtox, except using a hazard concentration affecting 20% of the species identified (i.e. EC20) rather than 50% (i.e. EC50). In total, 46 unique targets, i.e., combinations of category and (sub)compartment can be found, covering three categories of freshwater ecotoxicity and human toxicity (cancer and non-cancer) over multiple compartments (e.g., air) and sub-compartments (e.g., urban air close to ground). [32] In this work, we focused on 22 targets, excluding non-informative targets with fewer than 100 valid observations or exhibited near-zero variance to ensure the models have sufficient and informative data to learn from. For human toxicity, we chose to focus on the total human toxicity rather than differentiating between cancer and non-cancer effects due to less data availability and zero-inflation observed for cancer effects. [32] However, this aggregation represents a limitation, as it may obscure differences in underlying mechanisms and dose–response relationships between cancer and non-cancer toxicity pathways.

To explicitly assess this trade-off, we evaluated a modeling strategy that treats cancer and non-cancer effects as separate targets. Given the presence of zero-inflated and mixed discrete–continuous distributions, we implemented a data-driven hurdle modeling framework. Specifically, for targets exhibiting zero inflation, the modeling task was decomposed into two stages: (i) a binary classification model (random forest) to predict the occurrence of non-zero toxicity values, and (ii) a conditional regression model to predict the magnitude of toxicity for non-zero observations. For targets without a detectable hurdle structure, a single regression model was applied directly.

Using this procedure, we evaluated whether explicitly modelling cancer and non-cancer effects as separate targets improved predictive performance across multiple algorithms. However, no consistent improvement in predictive accuracy (R^2) was observed compared with modelling total human toxicity as a single endpoint throughout the target. Detailed descriptions of the hurdle-detection methodology and associated results are provided in the Supporting Information (SI). Accordingly, in this study we adopt total human toxicity as a pragmatic choice for systematic model evaluation and benchmarking different algorithms, while noting that future work may revisit separate modelling of cancer and non-cancer effects as larger and more balanced datasets become available or modelling separate effect is necessary.

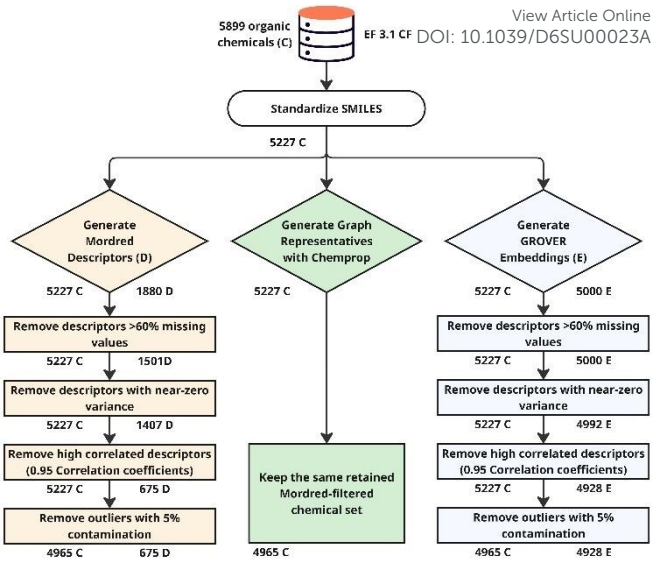

Figure 1 Data preprocessing and feature construction workflow for toxicity CF modelling (after each step, remaining chemical (C) amount, Mordred Descriptors (D) amount, and GROVER Embedding (E) amounts are followed.

In addition, since the CFs in the dataset exhibit a highly skewed distribution, the remaining 22 target variables were converted into log form, which are more symmetric and closer to a normal distribution. [23] This data transformation was done to facilitate the model to reduce the difference in the scale of the values and to capture the variance adequately.[33]

View Article Online
DOI: 10.1039/D6SU00023A

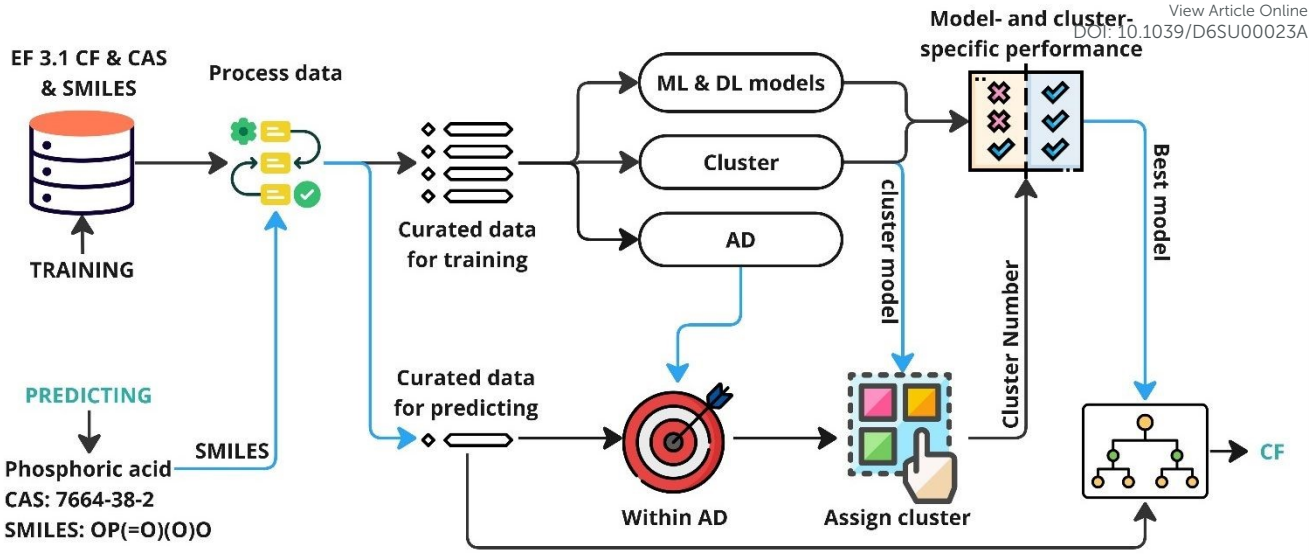

Figure 2 Illustration of machine learning workflow to predict characterization factors. SMILES (Simplified Molecular-Input Line-Entry System); AD (Applicability Domain); ML (Machine Learning); DL (Deep Learning); CF (Characterization Factors)

Chemical Feature Generation. Chemical features were generated from Simplified Molecular Input Line Entry System (SMILES) strings. SMILES were retrieved and standardized using the Python-based tool MoleculeResolver. [34] Salt counterions were removed following the protocol of Mansouri et al. [35] Stereoisomeric, tautomeric, and isotopic variants were not differentiated during SMILES generation. All structures were then converted to canonical SMILES, and duplicates were removed based on the final standardized representation. This procedure was applied before train/test splitting to reduce the risk of chemical overlap and data leakage caused by inconsistent SMILES representations across sources.

This work explored three types of feature data from valid SMILES (Figure 1). First, traditional physicochemical and structural properties were computed using the Mordred library, [36] which provides a comprehensive suite of over 1,800 2D and 3D descriptors, including constitutional, topological, and geometric features. The second type of data is chemical graph representations, in which atoms and bonds are represented as nodes and edges, respectively. Such graph data was only used for GNN and we used Chemprop (v2.0) framework, [37] which is a directed message-passing algorithm derived from the Gilmer et al., [38] to extract graph information from SMILES and perform GNN. The third type of representation is molecular embeddings generated with GROVER, [18] a large-scale self-supervised graph neural network pretrained on approximately 10 million unlabelled molecules. We obtained fixed-length (5000-dimensional) vector representations for each compound by applying the pretrained GROVER model, as distributed via the Ersilia Model Hub (identifier: eos7w6n). [39]

Pre-treatment of Chemical Descriptors. Before feeding into training algorithms, Mordred descriptors and GROVER embeddings were processed following the same procedure in the study of Liang et al., [40] (Figure 1). These molecular descriptors were curated through sequential filtering to remove features with excessive missingness and retaining only those descriptors with at least 40% non-missing values across the chemicals. Descriptors with near-zero variance were removed to eliminate features carrying negligible discriminatory information. Highly correlated descriptors were pruned based on Spearman correlation. Correlation coefficients over 0.95 were first grouped into correlation clusters and a single representative descriptor per cluster was retained based on minimal missingness and maximal variance.

Outliers Removal. As done in another similar study, [41] outliers were removed using an Isolation Forest [42] applied to a combined space of standardized molecular features (either Mordred descriptors or GROVER) and Morgan fingerprints, retaining only compounds classified as inliers under a fixed contamination rate (5%). Target-specific counts before and after outlier removal are provided in the Supplementary Information (SI). To assess the effect of this preprocessing step, the full modelling workflow was additionally repeated without outlier removal.

Model Construction and Evaluation

After data treatment above, the final dataset contains 4965 chemicals, and 22 targets as prediction variables. These processed data will feed into the training working flow (Figure 2) implemented for the prediction of CFs.

Note that the GROVER dataset after pre-processing still contains more than four thousand dimensions, which can substantially increase computational cost and exacerbate overfitting for non-DNN approach (RF, XGBoost, and GP). Therefore, before feeding GROVER data into these algorithms, a feature reduction was performed using a DNN approach. [43] First, the original high-dimensional feature vectors were standardized and then used to train a neural network autoencoder to reconstruct the input data while compressing it into a lower-dimensional space (512 dimensions). These compact latent representations were subsequently used as reduced feature inputs.

As done already in other similar studies [44] missing values in Mordred descriptors were imputed using the median value of each feature since median imputation provides a robust estimate of central tendency and reduces sensitivity to outliers commonly encountered in molecular descriptor calculations. GROVER embeddings do not contain missing values and therefore required no imputation. Continuous descriptors and embeddings were subsequently standardized using z-score normalization, defined as subtracting the mean and dividing by the standard deviation, with scaling parameters estimated from the training data only. Principal component analysis (PCA) retaining 95% of the cumulative variance was applied exclusively for GP models to reduce descriptor dimensionality and mitigate kernel sensitivity to correlated and high-dimensional inputs. PCA was not applied to tree-based or neural network models, as these algorithms are either scale-invariant or capable of learning latent representations directly from the original feature space.

Models Overview. Across three molecular representations—(i) Mordred descriptor, (ii) molecular graphs, and (iii) pretrained GROVER embeddings—we benchmarked a consistent panel of learners spanning classical ML and DL, namely RF, XGBoost, GP, and two neural architectures (DNN and MPNN). Additionally, as DNN and MPNN could be trained with multiple targets at the same time, results were compared employing these two DL models in single-target (and then looping through all targets) or multi-target settings (training for multiple targets simultaneously). For MPNN, which is based on a graph neural network architecture, two input configurations were implemented: a graph-only configuration using molecular graphs derived from SMILES, and a hybrid graph–tabular configuration integrating molecular graphs with curated Mordred descriptors. To ensure comparability with descriptor-based models, the graph-only MPNN was trained and evaluated on the same retained chemical set as the Mordred-based models (Figure 1), although its molecular representation was generated directly from SMILES rather than from Mordred descriptors. The DNN and MPNN were achieved through implementation of two DL frameworks: DeepMPT [45] and Chemprop. [37] Table 1 summarizes the applicability of each algorithm to different input data sources and indicates whether models were trained in single-target or multi-target settings. Additionally, in single target mode, chemicals with a missing target, were removed entirely to allow models to learn from features and target relationships. In multiple-target mode, chemicals are only removed if all targets are missing. If a chemical contains at least one target value, the chemical and its features are kept for training multiple targets. Model explanation and parameter settings are provided in SI.

Tree-based and kernel-based models (RF, XGBoost, GP) were trained using an 80/20 train–test split, as these methods do not require a separate validation set for early stopping or model selection. In contrast, DL models (DNN and MPNN) were trained using an 80/10/10 train–validation–test split. All the models were trained three times with different random splitting and compared with average predictions.

Table 1 Overview of model–representation compatibility and prediction scope across the evaluated machine learning frameworks (RF: Random Forest; XGB: XGBoost; GP: Gaussian Process; S: Single target; M: Multiple targets; GO: Graph only; GE: Graph and extra Mordred descriptors; R: Reduced dimension through DNN approach; F: Full dimensions without reduction).

	Mordred	Graph	GROVER
RF	S	-	S (R)
XGB	S	-	S (R)
GP	S	-	S (R)
DeepMPT	Both S and M	-	Both S and M (F)
MPNN_GO	-	Both S and M	-
MPNN_GE	Both S and M		-

Models Evaluation. Model performance was assessed using the coefficient of determination R^2 and Taylor diagram. R^2 measures how much of the variation in actual CF values from EF dataset can be explained by the model's predictions. Taylor diagram [46] provides a compact, multimeric comparison of model performance, jointly summarizing correlation, normalized standard deviation, and centered RMSE relative to observations.

Chemical Clustering

To characterize structural heterogeneity within the chemical dataset, unsupervised clustering was performed independently in two complementary representation spaces: (i) a hand-crafted physicochemical descriptor space (Mordred) and (ii) a learned molecular embedding space derived from GROVER. Clustering was conducted in the processed feature space using the *k*-means algorithm. The number of clusters was fixed at $k = 10$, selected based on maximizing the average silhouette score [47] across a range of candidate cluster numbers. Following model fitting, each molecule was assigned to the nearest cluster centroid in the corresponding transformed feature space.

Applicability Domain (AD)

We used Mahalanobis distance to determine whether a chemical falls within the feature space covered by the training data. [48] The applicability domain (AD) was defined separately for the Mordred descriptor space and the GROVER embedding space. For graph-only MPNN models, the AD was approximated using the GROVER embedding space, because the molecular graph input itself does not provide a fixed tabular feature matrix suitable for Mahalanobis-distance calculation. For MPNN models augmented with Mordred descriptors, the AD was calculated in the same curated Mordred descriptor space used for descriptor-based models. Chemicals outside this AD were considered less reliable for prediction. An empirical threshold was defined as the 95th percentile of the resulting distribution. New chemicals were projected into the same feature space, and their Mahalanobis distances to the training centroid were evaluated against this threshold. Compounds exceeding the threshold were flagged as out-of-domain, whereas in-domain chemicals were considered eligible for reliable prediction. Predictive errors with test dataset were calculated separately for in-domain and out-of-domain compounds.

View Article Online
DOI: 10.1039/D5SU00037A

Training and Predicting Framework

Finally, in this section, we present the general framework to predict CFs for missing chemicals based on EF3.1. In the training phase, EF 3.1 CF toxicity data linked to CAS numbers and SMILES representations are used to generate descriptors (in our case Mordred descriptors or GROVER embeddings). The curated representations serve as the basis for three parallel components: (i) training multiple ML and DL models, (ii) grouping similar chemicals into same clusters that capture structural and physicochemical similarity, and (iii) defining the AD that delineates the reliable chemical space of the training data. Model performance has been pre-evaluated in a cluster-specific manner, enabling identification of the best-performing model conditional on cluster identity.

During the prediction phase, a new chemical, provided as a SMILES string, is transformed using the same descriptor pipeline and projected into the same descriptor space. An AD check is first performed to determine whether the compound lies within the domain of reliable inference. The criterion that determines if the target chemical lies within or outside of the AD, is based on a predefined threshold value. If the new chemical is out of the AD, the pipeline terminates, reporting "No CF predicted". For in-domain chemicals, cluster membership is assigned using the pre-trained clustering model. The final CF prediction is then generated using this selected model, ensuring that predictions are both domain-consistent and locally optimized with respect to chemical similarity.

Note that the workflow including model selection is based on R^2 only, while within the selected models, prediction uncertainty was also evaluated in GP and MPNN. GP models inherently provide posterior predictive distributions, while evidential MPNN models yield distributional parameters that allow decomposition of predictive uncertainty into aleatoric and epistemic components. These uncertainty estimates provide complementary information on model confidence and reliability and are therefore reported alongside prediction results, as also advised in literature.[26] Depending on the research objective, the selection of the best suitable algorithms for prediction task could be adapted to be influenced by uncertainty.

Application & LCA Case Study in Textile Sector

The developed models were employed to predict missing CFs for a case study in the textile sector. This case study originates from the CALIMERO Project ("Industry CAse studies anaLysis to IMprove EnviROnmental performance and sustainability of bio-based industrial processes"), which aims to create a common LCA methodological framework for certain bio-based industries' sectors, among which the textile sector is a significant contributor to global environmental pollution, involving widespread use of chemicals, dyes, and other processing agents. [49] However, the absence of CFs for many textile chemicals hinders a comprehensive toxicity impacts estimation in LCA studies. [50]

In this case study, we aimed to compare the ecotoxicity and human toxicity impacts of processing one pair of jeans washed with either pumice stone or synthetic stone, performed by the company EREKS in Turkey. Pumice stone is a non-renewable natural resource, which is used as an industry standard during the denim abrasion process. The pumices stone lasts maximum 2 washing cycles and degrades into sludge that needs to be disposed of. The alternative synthetic stone lasts much longer and remains nearly undegraded during the use stage of up to 3000h of washing cycle, thus avoiding generation of sludge and being ultimately suitable for recycling into a new stone. Detailed information including, among others, the process flow and life cycle inventory can be found in SI.

Note that for the case study, we obtained a list of chemicals that are added to the processes instead of the measured emissions from the processes. As the chemicals that are actually added to processes may be transformed and only a fraction of them (chemical and transformation compounds) may end up in the environment, the emitted amount and emission compartments need to be predefined. These assumptions were derived from textile-sector literature, especially the work of Roos et al.[49,50] for the textile sector. Following this procedure, five chemicals released into environment lack CFs in EF v3.1 (SI). By employing different ML and DL models, this work first predicted these missing CFs and then used them into the LCA study for jeans washing.

The SI provide a short procedure for the application of CF to chemical amounts used in processes, considering as well derivation of transformation and emission amounts, to better support implementation, especially in the context of SSbD.

Results and Discussion

Comparing Model Performance

Target-specific values of mean R^2 and RMSE and their corresponding standard deviation (SD) across three independent random splits were calculated. The performance of the algorithms with the highest R^2 are provided in Table 2. Figure 3 presents the mean predictive performance (R^2) of multiple learning algorithms across EF 3.1 ecotoxicity and human health CF targets, evaluated under different algorithms and data sources.

Table 2 Best-model performance (based on R^2). Values represent mean ± SD. The highest R2 is always obtained with Mordred descriptors.

Target	Best algorithm	R^2	RMSE
eco: agri soil	XGB	0.47 ± 0.07	1.34 ± 0.07
eco: air indoor	XGB	0.52 ± 0.03	1.28 ± 0.12
eco: air unspec.	XGB	0.52 ± 0.03	1.36 ± 0.04
eco: fresh water	MPNN_GE_S	0.5 ± 0.07	1.08 ± 0.05
eco: non-agri soil	XGB	0.49 ± 0.05	1.34 ± 0.01
eco: non-urban	XGB	0.53 ± 0.01	1.43 ± 0.03
eco: sea water	XGB	0.64 ± 0.01	2.39 ± 0.02
eco: soil unspec.	XGB	0.49 ± 0.04	1.33 ± 0.01

View Article Online
DOI: 10.1039/D5SU00234

eco: stratosphere	XGB	0.51 ± 0.04	1.4 ± 0.09
eco: urban air	XGB	0.54 ± 0.01	1.33 ± 0.02
eco: water unspec.	MPNN_GE_S	0.49 ± 0.08	1.23 ± 0.16
hh: agri soil	XGB	0.45 ± 0.02	1.17 ± 0.06
hh: air indoor	MPNN_GE_M	0.44 ± 0.04	0.89 ± 0.05
hh: air unspec.	MPNN_GE_M	0.46 ± 0.04	0.88 ± 0.06
hh: fresh water	XGB	0.52 ± 0.03	1.02 ± 0.04
hh: non-agri soil	XGB	0.46 ± 0.07	1.3 ± 0.09
hh: non-urban	MPNN_GE_M	0.49 ± 0.04	0.98 ± 0.05
hh: sea water	XGB	0.56 ± 0.01	1.34 ± 0.12
hh: soil unspec.	XGB	0.44 ± 0.02	1.18 ± 0.06
hh: stratosphere	MPNN_GE_M	0.5 ± 0.04	0.97 ± 0.06
hh: urban air	MPNN_GE_M	0.46 ± 0.05	0.87 ± 0.06
hh: water unspec.	XGB	0.52 ± 0.03	1.02 ± 0.06

Model Performance. Overall, model performance was moderate, with the best R^2 values across targets ranging from 0.42 to 0.64, and corresponding RMSE between 0.87 to 1.4 (Table 2). Across the tested algorithms, the highest R^2 values were generally obtained with models using Mordred descriptors, particularly XGBoost and MPNN-based models. In addition, performance is strongly target-dependent, with consistently higher R^2 values observed for ecotoxicity, ranging from 0.47 to 0.64 based on best algorithm for each target, compared to their human health equivalents, ranging from 0.44 to 0.56 from the best algorithm. These values of R^2 are within the range of the results obtained by other studies for toxicity prediction [23,24,30,51] and illustrate reasonable predictive performance across several environmental compartments.

Multi-Target Prediction Benefits. No consistent benefits with multi-target training are observed for DeepMTP comparing to single target training (Figure 3). In contrast, graph-only MPNNs display a more systematic and robust advantage under multi-target training (MPNN-GO-M) than single-target performance (MPNN-GO-S), particularly for human toxicity targets. This pattern suggests that shared message-passing representations can enable effective transfer of structural information across related toxicity targets, allowing the model to exploit common substructural motifs and physicochemical drivers encoded directly in the molecular graph. [15] The benefit of multi-target learning diminishes for ecotoxicity targets, where gains are smaller and occasionally negligible.

Mordred vs. GROVER Across the tested algorithms, the best-performing models were generally based on Mordred descriptors, particularly XGBoost and MPNN-based models (Table 2). This result suggests that for our task of predicting CF directly, 2D/3D molecular descriptors remain highly informative for these toxicological domains. This result is in line with another study comparing XGBoost, RF, and DNN. [24]
In most cases, replacing handcrafted descriptors with GROVER features in MLs of RF, XGB, and GP led to a systematic reduction in R^2 (Figure 3), indicating that the high-dimensional pretrained embeddings were not optimally exploited by tree-based or kernel-based learners. This trend was consistent across targets and algorithms, suggesting a general mismatch between GROVER's representations and conventional ML models.
In contrast, DL–based models showed the opposite behavior. For DeepMPT, GROVER embeddings achieved performance comparable to, and in several cases exceeding, that obtained with Mordred descriptors. This observation is in line with experiments by GROVER developers, [18] highlighting the ability of deep architectures to effectively leverage the rich, high-dimensional information encoded in pretrained graph representations. However, even with GROVER representation, DeepMPT generally underperformed both the MPNN models and the ML models explored in this study, indicating GROVER features do not guarantee improved predictive performance. Further work is needed to assess whether other DL architectures can exploit GROVER embeddings more effectively and achieve competitive or superior performance, enabling end-to-end learning of complex nonlinear feature interactions without reliance on handcrafted descriptor design.

View Article Online
DOI: 10.1039/D6SU00023A

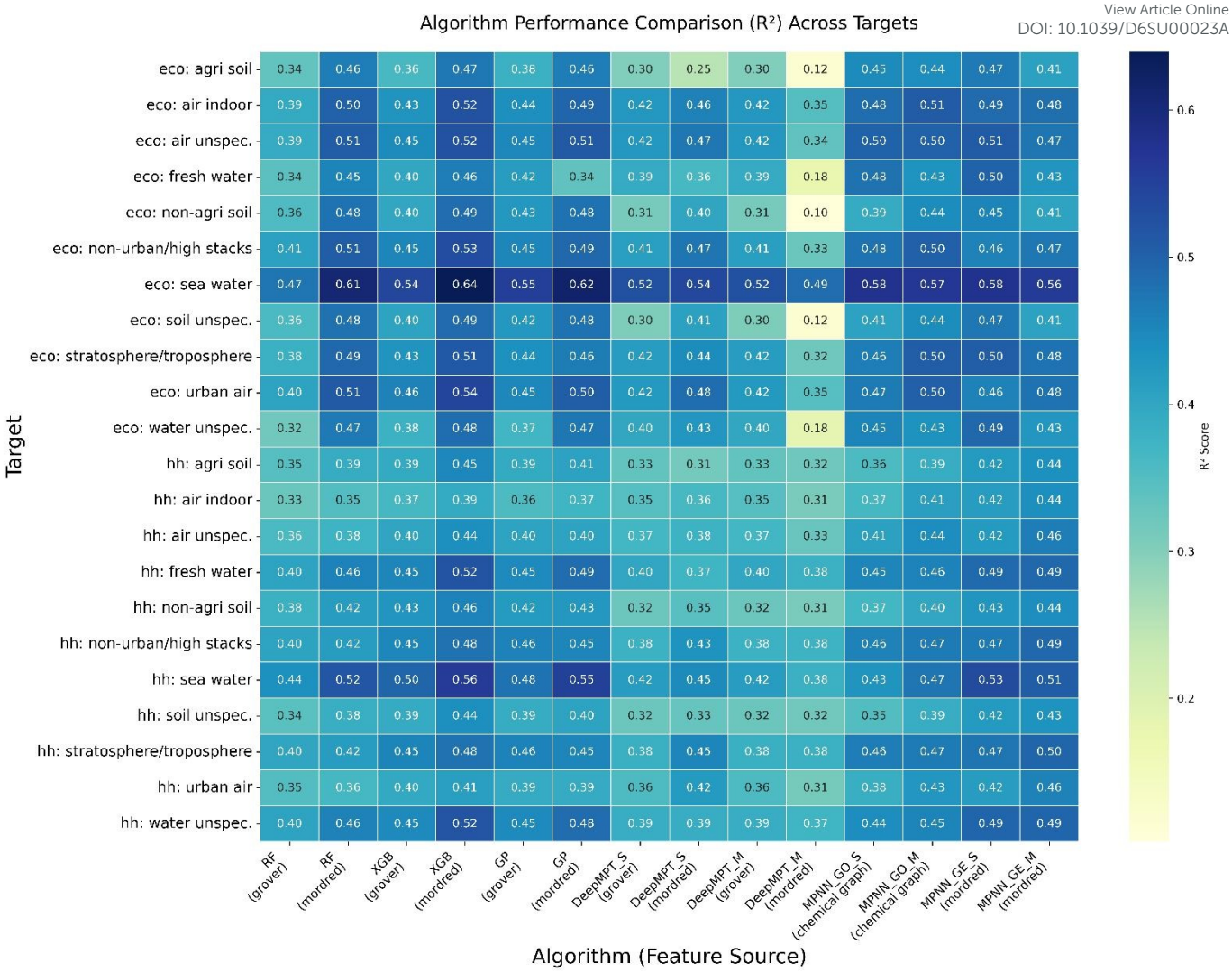

Figure 3 Heatmap of predictive performance (R^2) across multiple characterization factor (CF) targets for different learning algorithms and feature sources. Rows represent individual CF targets spanning ecotoxicity and human toxicity compartments, while columns correspond to model–feature combinations (e.g., DeepMPT, GP, MPNN, RF, and XGBoost using molecular descriptors or GROVER embeddings). Colour intensity and annotated values indicate R^2 scores, enabling a comparative assessment of algorithm robustness and feature representations across targets. (MPNN: Message-passing neural networks; GO: Graph only; GE: Graph and extra Mordred descriptors; M: Multi-target regression; S: Single-target regression; XGB: XGBoost; eco: ecotoxicity; hh: human health)

Combining Mordred descriptors. The effect of combining graph-based representations with additional Mordred descriptors (MPNN_GE_S and MPNN_GE_M comparing to MPNN_GO_S and MPNN_GO_M) is target-dependent (Figure 3). For ecotoxicity endpoints, adding Mordred descriptors to graph-based MPNN models resulted in only marginal improvements, and in some cases slightly reduced R^2 values. This suggests that the molecular graph representation alone already captured much of the structure–toxicity information relevant to ecotoxicological effects, particularly in the multi-target setting. In these cases, the additional Mordred descriptors may have been partly redundant or may have introduced noise and collinearity. For human toxicity endpoints, however, adding Mordred descriptors generally improved model performance. This indicates that hand-crafted descriptors contributed complementary information beyond the learned graph representation, possibly because human toxicity CFs depend more strongly on physicochemical or exposure-relevant molecular properties that are not fully represented by graph topology alone.

Comparing Model Performance over Clusters

The heatmap (Figure 4) summarizes the best algorithms and chemical representations in terms of R^2 scores across chemical clusters. The maximum R^2 values differ substantially across both chemical clusters and toxicity targets, indicating that model performance is not uniform across the chemical space. Although XGBoost and MPNN with Mordred data in general appear to be the best algorithm and data combination (Table 2), the identity of the best-performing algorithm for each cluster varies. Importantly, GROVER-based representations, despite showing lower overall performance in many conventional ML settings (Figure 3), emerge as the top-performing

View Article Online
DOI: 10.1039/D6SU00023A

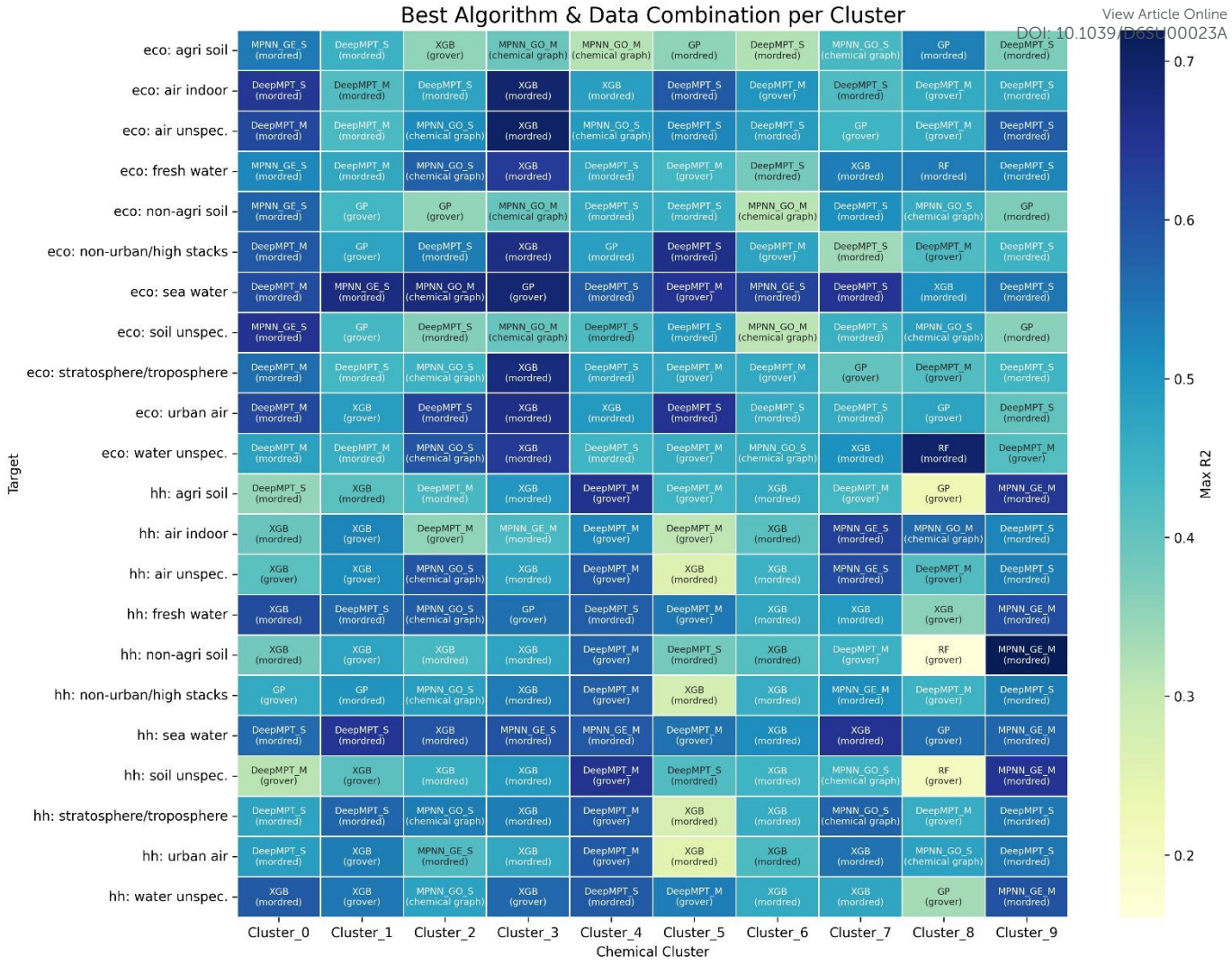

Figure 4 Comparative analysis of top-performing model architectures and molecular representations across chemical clusters for each target (eco: ecotoxicity; hh: human health)

representation in several specific clusters. This indicates that pretrained molecular embeddings may capture structural patterns that are particularly informative for certain regions of chemical space, even if their average performance across the full dataset is not consistently superior. Therefore, GROVER should not be dismissed solely based on global performance (Figure 3), but may be useful within a cluster-aware model-selection framework.

This observation also supports the usefulness of the proposed training and prediction pipeline, which explicitly couples global learning (i.e., the model trained with the whole training set from EF3.1 in this study) with local cluster-aware model selection. The global training enables broad coverage of heterogeneous chemical space, which is in line with recommendations in literature, [52] where it was highlighted that applying global models is preferred to developing a series of local models. In parallel, the selection of the best model and chemical representations for each cluster allows predictions to be refined.

Comparing Model Performance Taylor Diagram

In Taylor diagram results models positioned closer to the reference point (unit standard deviation, correlation =1) indicate superior agreement in both variability and pattern reproduction. Across all targets, the Taylor diagrams indicate that most algorithms achieve broadly comparable correlation coefficients and centred RMSE values (shows only freshwater ecotoxicity, other target diagram results can be found in SI). Model points are concentrated within a relatively narrow angular range and fall between similar centred RMSE contours, suggesting that differences in pattern agreement and error structure across algorithms are limited. In contrast, substantial variability is observed in the normalized standard deviations across models, indicating marked differences in how algorithms reproduce the amplitude of observed variability. RF constantly exhibit lowest standard deviation, far from the observed reference. DL models including MPNN and DeepMTP, generally achieve higher standard deviations, closer to the observed reference.

Sensitivity analysis with outliers

View Article Online
DOI: 10.1039/D6SU00037A

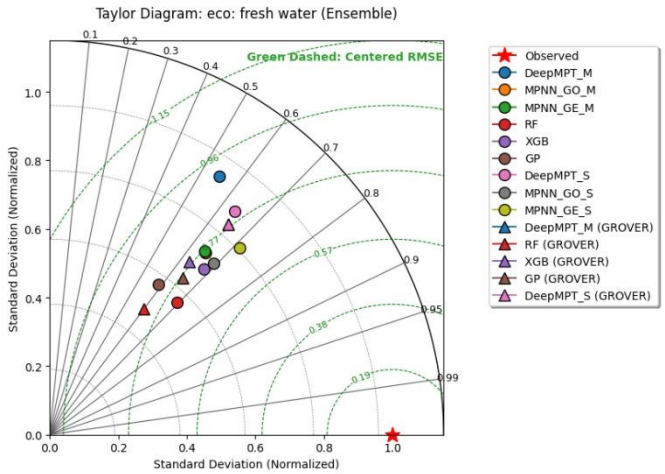

Figure 5 Taylor diagram comparing model performance for the target of ecotoxicity fresh water. The red star denotes the observed reference, while markers represent different algorithms and feature representations. Models closer to the reference point and with higher correlation and lower centered RMSE, indicate superior predictive performance.

Detailed target-specific sensitivity results for including outliers are provided in the SI. Overall, the results showed that outlier removal had no systematic effect on model performance: it slightly improved performance for some target–model combinations but reduced it for others. This result suggests that the identified outliers may harbor critical chemical or biological signals rather than mere stochastic noise. This inconsistency underscores the risk of information loss in high-dimensional datasets where extreme values may represent valid boundary conditions of the modeled phenomenon. Future work should therefore explore chemically informed and target-specific outlier-handling strategies with expert review, rather than relying solely on fixed-rate automatic removal.

In-domain and Out-of-domain Error

The AD-based error analysis (SI) showed that out-of-domain compounds generally have higher prediction errors than in-domain compounds, particularly for models using the Mordred descriptor representation. This indicates that the Mahalanobis-distance-based AD provided a useful indicator of prediction reliability in the curated descriptor space. However, this pattern was less consistent for GROVER-based models, suggesting that simple distance-to-centroid thresholds may be less suitable for learned molecular representations. Recent work [53] has similarly shown that measuring distance in an embedding space (GROVER embedding in this study), may not fully capture whether a molecule is reliable for prediction. Future work should therefore investigate alternative AD strategies for embedding-based models. For example, reconstruction-based metric was developed [53] for AD of embedding based chemical spaces, which may better capture molecular distribution shifts and prediction reliability.

Case Study Results

Illustrative Comparison of Model Performance with Literature

Figure compares freshwater ecotoxicity and total human toxicity impacts associated with washing one pair of jeans using pumice stone and synthetic stone, evaluated with and without the inclusion of predicted CFs. Across both assessment setups, the pumice-stone. process exhibits consistently higher toxicity impacts than the synthetic-stone alternative, highlighting the benefits of the use of synthetic stone process. Note that much less chemicals are used in the synthetic stone process.

When only existing EF v3.1 CFs are considered, toxicity impacts are systematically underestimated. The inclusion of predicted ecotoxicity CFs leads to a 39% and 59% increase for pumice and synthetic stone respectively. This increase is mainly due to the additional sodium hydroxide, hydrogen peroxide, and alcohols ethoxylated, highlighting that omitting these chemicals with missing CFs can lead to a substantial underrepresentation of toxicity burdens in LCA studies. For human toxicity, the increase is less considerable, with only 7% increase for synthetic stone and no observable difference for pumice stone.

One textile study [49] calculated CF for missing chemicals, and we found one common chemical, alcohols ethoxylated. In their study, [49] CFs were collected from COSMEDE database, which is developed for cosmetic and detergent substances, drawing primarily on experimental data from publicly available toxicological and physicochemical databases and peer-reviewed literature. By comparing our predicted values with the calculated values reported in the literature, the predicted freshwater ecotoxicity CF (3.69×10^{3}) is close to the literature value (2.74×10^{3}), while the predicted urban-air ecotoxicity CF (1.69×10^{1}) is lower than the literature value (8.70×10^{1}). For human toxicity, the predicted CFs are very small for freshwater (8.24×10^{-8}) and urban air (7.01×10^{-8}), whereas the literature values are both 0, reflecting the limited coverage of human toxicity in the USEtox-based databases, including COSMEDE. [50] All the predicted values and their uncertainties are provided in SI.

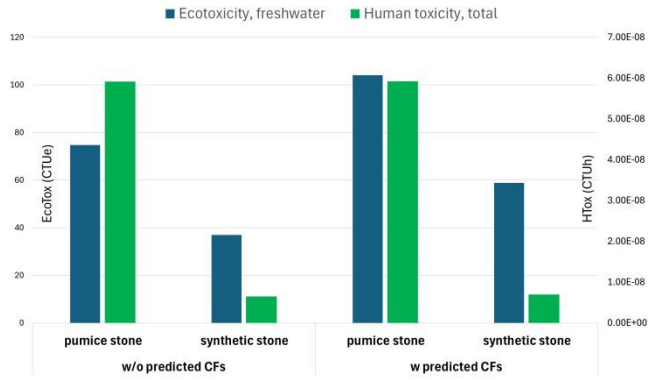

Figure 6 Comparison of freshwater ecotoxicity and total human toxicity impacts for washing one pair of jeans using pumices stone and synthetic stone, evaluated without (w/o) and with (w) machine learning predicted characterization factors

Conclusions

View Article Online
DOI: 10.1039/D6SU00023A

This study demonstrates the feasibility and value of a fully data-driven, automatable framework for predicting toxicity characterization factors (CFs) as a whole directly from molecular information, addressing a central bottleneck in Life Cycle Assessment (LCA) and Safe and Sustainable by Design (SSbD). By leveraging the expanded Environmental Footprint (EF) v3.1 database, we benchmark a broad spectrum of machine learning (ML) and deep learning (DL) models across multiple molecular representations.

Several model-related insights emerged from the benchmarking. First, predictive performance is strongly dependent on the target, with ecotoxicity CFs being systematically more predictable (R^2 ranging from 0.47 to 0.64) than human health CFs (R^2 ranging from 0.44 to 0.56), reflecting fundamental differences in mechanistic complexity and data availability. Second, no single model–representation combination dominates across all targets and chemical subspaces. Descriptor-based approaches coupled with XGBoost remain highly competitive for many CF targets, while DL models generally achieve higher and more realistic standard deviations. Third, graph-only MPNNs showed a more systematic advantage under multi-target training compared with single-target training, particularly for human toxicity targets, indicating that shared molecular graph representations can transfer useful structural information across related endpoints. For ecotoxicity targets, however, the benefit of multi-target learning was weaker and sometimes negligible. Fourth, the effect of combining molecular graph representations with Mordred descriptors was also target-dependent. For ecotoxicity endpoints, adding Mordred descriptors to graph-based MPNNs led only to marginal improvements and in some cases slightly reduced performance. In this case, graph-only information is sufficient and combining with Mordred descriptors can be redundant or even detrimental. In contrast, for human toxicity endpoints, adding Mordred descriptors generally improved performance, indicating that hand-crafted descriptors may provide complementary information beyond graph topology, possibly related to physicochemical or exposure-relevant molecular properties. Finally, although task-specific fine-tuning of the pre-trained GROVER model for toxicity CF prediction did not yield substantial performance gains compared, this approach nonetheless provides local advantages in certain chemical clusters. In addition, since GROVER-based representations eliminate the need for handcraft descriptor generation and manual feature engineering, it provides a promising alternative for CF toxicity prediction. Future work should explore whether other DL architectures can exploit pretrained molecular embeddings within cluster-specific applicability domains more effectively for CF prediction.

Another contribution of this work is the explicit integration of applicability domain analysis and chemical clustering into the prediction workflow. By coupling global models trained on the full EF 3.1 chemical space with cluster-aware model selection, the proposed framework reconciles generalizability with local interpretability and robustness. This strategy aligns with established QSAR best practices while extending them to enable domain-consistent predictions even in structurally heterogeneous chemical inventories.

The textile-sector case study illustrates the practical implications of this approach: predicted CFs for previously uncovered chemicals can be seamlessly integrated into LCA, enabling more comprehensive toxicity accounting and supporting informed substitution decisions in real industrial contexts. It should be noted that the results should be interpreted with caution, as model performance remains moderate and the case-study outcomes depend on assumptions to estimate emissions from chemical inputs, including release/degradation factors and emission compartments (SI). Nevertheless, this example illustrates the central concern highlighted in LCA [22] that omitting chemicals solely because of missing CFs may substantially underestimate toxicity impacts, by up to approximately 60% based on this study.

Overall, this work advances the state of the art by systematically positioning classical ML, graph-based DL, and pretrained molecular representations within a unified CF prediction workflow. This work moves beyond intermediate-parameter prediction for CF derivation, toward direct, structure-based CF estimation, offering a scalable pathway for high-throughput chemical screening, and a potential systematic application in LCA case studies, bridging data driven modelling and life cycle impact assessment practice. Notably, it can enrich the EF3.1 database with new CF for chemicals of interest. However, this work does not attempt to replace intermediate-parameter prediction paradigms which has clear benefits, as mentioned in the introduction. Instead, it rather serves as a benchmark for further studies and an alternative when fast screening of large amount of chemicals is needed while such data is lacking. In addition, this study focuses on exploring different ML and DL and data sources. Hence, feature importance is not discussed here, which can be further studied. This work provides a foundation for future research, which should focus on improving human toxicity modelling and combining data-driven predictions with mechanistic constraints and uncertainty to further enhance credibility for regulatory and SSbD applications.

Author contributions

Tianran Ding: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation Visualization Writing, original draft Writing; **Gustavo Larrea-Gallegos**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources Software, Validation Visualization Writing, original draft Writing; **Federico Busio**: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization Writing; **Antonino Marvuglia**: Conceptualization, Methodology, Resources, Supervision, Validation review & editing; **Thomas Schaubroeck**: Conceptualization, Investigation, Methodology, Resources, Project administration, Validation Visualization Writing, Supervision Validation review & editing

View Article Online
DOI: 10.1039/D6SU00023A

Conflicts of interest

There are no conflicts to declare

Data availability

This study was carried out using publicly available data from European Platform on LCA EPLCA at https://eplca.jrc.ec.europa.eu/LCDN/developerEF.html

Acknowledgements

This study is part of the CALIMERO project (the grant number: 101060546) DESIDERATA project (grant number: 101178011) and CompSafeNano project (grant number: 101008099) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. We thank EREKS for providing us information and data on their case study. Gratitude also goes towards other CALIMERO partners for their reflections and information on used chemicals.

Notes and references

(1) Diamond, J. M.; Latimer II, H. A.; Munkittrick, K. R.; Thornton, K. W.; Bartell, S. M.; Kidd, K. A. Prioritizing Contaminants of Emerging Concern for Ecological Screening Assessments. *Environmental Toxicology and Chemistry* **2011**, *30* (11), 2385–2394. https://doi.org/10.1002/etc.667.

(2) Zhu, X.; Zhang, M.; Zhang, Z.; Li, M.; Liu, M.; Jing, H.; Tan, J.; Jia, H.; Wang, J. Electrochemical Sensor with Redox-Active Poly (Acridine Orange) Imprinted Film for Ultrasensitive and Selective Detection of Perfluorooctanoic Acid: A Binding-Induced Electroactivity Suppression Strategy. *Journal of Hazardous Materials* **2026**, *505*, 141442. https://doi.org/10.1016/j.jhazmat.2026.141442.

(3) Abbate, E.; Ragas, A. M. J.; Caldeira, C.; Posthuma, L.; Garmendia Aguirre, I.; Devic, A. C.; Soeteman-Hernández, L. G.; Huijbregts, M. A. J.; Sala, S. Operationalization of the Safe and Sustainable by Design Framework for Chemicals and Materials: Challenges and Proposed Actions. *Integrated Environmental Assessment and Management* **2025**, *21* (2), 245–262. https://doi.org/10.1093/inteam/vjae031.

(4) Fantke, P. Safe and Sustainable-by-Design (SSbD): Calling for Efficient Metrics, Biophysical Benchmarks, and Broader Application. *Sustainable Chemistry and Pharmacy* **2025**, *45*, 101986. https://doi.org/10.1016/j.scp.2025.101986.

(5) Larrea-Gallegos, G. M.; Hofer, S.; Hofstätter, N.; Punz, B.; Watzek, N.; Lölsberg, W.; Wiench, K.; Wohlleben, W.; Aguirre, I. G.; Athanassios, N.; Sarimveis, H.; Costa, A.; Seitz, C.; Friedrichs, S.; Exner, T. E.; Hischier, R.; Marvuglia, A.; Himly, M. Integrate & Balance Aspects for Safe and Sustainable Innovation: Needs Analysis on SSbD Categories and Product Development Stage Requirements to Cover the Entire Life Cycle. *Computational and Structural Biotechnology Journal* **2025**, *29*, 201–221. https://doi.org/10.1016/j.csbj.2025.07.030.

(6) Wu, X.; Zhou, Q.; Mu, L.; Hu, X. Machine Learning in the Identification, Prediction and Exploration of Environmental Toxicology: Challenges and Perspectives. *Journal of Hazardous Materials* **2022**, *438*, 129487. https://doi.org/10.1016/j.jhazmat.2022.129487.

(7) Kurban, H.; Kurban, M.; Dalkilic, M. M. Rapidly Predicting Kohn–Sham Total Energy Using Data-Centric AI. *Sci Rep* **2022**, *12* (1), 14403. https://doi.org/10.1038/s41598-022-18366-7.

(8) Feng, H.; Zhang, L.; Li, S.; Liu, L.; Yang, T.; Yang, P.; Zhao, J.; Arkin, I. T.; Liu, H. Predicting the Reproductive Toxicity of Chemicals Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicology Letters* **2021**, *340*, 4–14. https://doi.org/10.1016/j.toxlet.2021.01.002.

(9) Siramshetty, V. B.; Chen, Q.; Devarakonda, P.; Preissner, R. The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data. *J. Chem. Inf. Model.* **2018**, *58* (6), 1224–1233. https://doi.org/10.1021/acs.jcim.8b00150.

(10) Obrezanova, O.; Segall, M. D. Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity. *J. Chem. Inf. Model.* **2010**, *50* (6), 1053–1061. https://doi.org/10.1021/ci900406x.

(11) Liu, M.; Zhang, L.; Li, S.; Yang, T.; Liu, L.; Zhao, J.; Liu, H. Prediction of hERG Potassium Channel Blockage Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicology Letters* **2020**, *332*, 88–96. https://doi.org/10.1016/j.toxlet.2020.07.003.

(12) Hamadache, M.; Benkortbi, O.; Hanini, S.; Amrane, A. Application of Multilayer Perceptron for Prediction of the Rat Acute Toxicity of Insecticides. *Energy Procedia* **2017**, *139*, 37–42. https://doi.org/10.1016/j.egypro.2017.11.169.

(13) Sharma, B.; Chenthamarakshan, V.; Dhurandhar, A.; Pereira, S.; Hendler, J. A.; Dordick, J. S.; Das, P. Accurate Clinical Toxicity Prediction Using Multi-Task Deep Neural Nets and Contrastive Molecular Explanations. *Sci Rep* **2023**, *13* (1), 4908. https://doi.org/10.1038/s41598-023-31169-8.

(14) Yuan, Q.; Wei, Z.; Guan, X.; Jiang, M.; Wang, S.; Zhang, S.; Li, Z. Toxicity Prediction Method Based on Multi-Channel Convolutional Neural Network. *Molecules* **2019**, *24* (18), 3383. https://doi.org/10.3390/molecules24183383.

(15) Wu, X.; Zhu, Y.; Hou, J.; Zhang, J.; Xu, J.; Lin, D. Deep Learning Advances High-Throughput Toxicity Screening of Chemicals at Multi-Biological Levels. *Journal of Environmental Sciences* **2025**. https://doi.org/10.1016/j.jes.2025.09.045.

(16) Guo, W.; Liu, J.; Dong, F.; Song, M.; Li, Z.; Khan, M. K. H.; Patterson, T. A.; Hong, H. Review of Machine Learning and Deep Learning Models for Toxicity Prediction. *Exp Biol Med (Maywood)* **2023**, *248* (21), 1952–1973. https://doi.org/10.1177/15353702231209421.

(17) Cremer, J.; Medrano Sandonas, L.; Tkatchenko, A.; Clevert, D.-A.; De Fabritiis, G. Equivariant Graph Neural Networks for Toxicity Prediction. *Chem. Res. Toxicol.* **2023**, *36* (10), 1561–1573. https://doi.org/10.1021/acs.chemrestox.3c00032.

(18) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. arXiv October 29, 2020. https://doi.org/10.48550/arXiv.2007.02835.

(19) Fantke, P.; Bijster, M.; Hauschild, M. Z.; Huijbregts, M.; Jolliet, O.; Kounina, A.; Magaud, V.; Margni, M.; McKone, T. E.; Rosenbaum, R. K.; Van De Meent, D.; Van Zelm, R. USEtox® 2.0 Documentation (Version 1.00). **2017**. https://doi.org/10.11581/DTU:00000011.

(20) EU. Commission Recommendation of 9 April 2013 on the Use of Common Methods to Measure and Communicate the Life Cycle Environmental Performance of Products and Organisations Text with EEA Relevance, 2013. https://doi.org/http://data.europa.eu/eli/reco/2013/179/oj.

(21) Wang, Z.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environ. Sci. Technol.* **2020**, *54* (5), 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

(22) Biganzoli, F.; Sanyé-Mengual, E.; Sala, S. Toxicity Impacts Evolution in the Environmental Footprint Methods: An Example-Based Interpretation of the Results. *Int J Life Cycle Assess* **2026**, *31* (1), 49. https://doi.org/10.1007/s11367-026-02609-0.

(23) Hou, P.; Jolliet, O.; Zhu, J.; Xu, M. Estimate Ecotoxicity Characterization Factors for Chemicals in Life Cycle Assessment Using Machine Learning Models. *Environment International* **2020**, *135*, 105393. https://doi.org/10.1016/j.envint.2019.105393.

(24) Li, D.; Qin, J.; Hong, J. Toward a Comprehensive Life Cycle Aquatic Ecotoxicity Assessment via Machine Learning: Application to Coal Power Generation in China. *Journal of Cleaner Production* **2024**, *445*, 141373. https://doi.org/10.1016/j.jclepro.2024.141373.

(25) Kvasnicka, J.; Aurisano, N.; von Borries, K.; Lu, E.-H.; Fantke, P.; Jolliet, O.; Wright, F. A.; Chiu, W. A. Two-Stage Machine Learning-Based Approach to Predict Points of Departure for Human Noncancer and Developmental/Reproductive Effects. *Environ. Sci. Technol.* **2024**. https://doi.org/10.1021/acs.est.4c00172.

(26) von Borries, K.; Beckwith, K. V.; Goodman, J. M.; Chiu, W. A.; Jolliet, O.; Fantke, P. Uncertainty-Aware Machine Learning to Predict Non-Cancer Human Toxicity for the Global Chemicals Market. *Nat Commun* **2026**. https://doi.org/10.1038/s41467-025-67374-4.

(27) Blanco, C. F.; Pauliks, N.; Donati, F.; Engberg, N.; Weber, J. Machine Learning to Support Prospective Life Cycle Assessment of Emerging Chemical Technologies. *Current Opinion in Green and Sustainable Chemistry* **2024**, *50*, 100979. https://doi.org/10.1016/j.cogsc.2024.100979.

(28) Marvuglia, A.; Kanevski, M.; Benetto, E. Machine Learning for Toxicity Characterization of Organic Chemical Emissions Using USEtox Database: Learning the Structure of the Input Space. *Environment International* **2015**, *83*, 72–85. https://doi.org/10.1016/j.envint.2015.05.011.

(29) Song, R.; Li, D.; Chang, A.; Tao, M.; Qin, Y.; Keller, A. A.; Suh, S. Accelerating the Pace of Ecotoxicological Assessment Using Artificial Intelligence. *Ambio* **2022**, *51* (3), 598–610. https://doi.org/10.1007/s13280-021-01598-8.

(30) Hou, P.; Zhao, B.; Jolliet, O.; Zhu, J.; Wang, P.; Xu, M. Rapid Prediction of Chemical Ecotoxicity

Through Genetic Algorithm Optimized Neural Network Models. *ACS Sustainable Chem. Eng.* **2020**, *8* (32), 12168–12176. https://doi.org/10.1021/acssuschemeng.0c03660.

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

(32) Sala, S.; Biganzoli, F.; Mengual, E. S.; Saouter, E. Toxicity Impacts in the Environmental Footprint Method: Calculation Principles. *Int J Life Cycle Assess* **2022**, *27* (4), 587–602. https://doi.org/10.1007/s11367-022-02033-0.

(33) Birkved, M.; Heijungs, R. Simplified Fate Modelling in Respect to Ecotoxicological and Human Toxicological Characterisation of Emissions of Chemical Compounds. *Int J Life Cycle Assess* **2011**, *16* (8), 739–747. https://doi.org/10.1007/s11367-011-0281-y.

(34) Müller, S. How to Crack a SMILES: Automatic Crosschecked Chemical Structure Resolution across Multiple Services Using MoleculeResolver. *J Cheminform* **2025**, *17* (1), 117. https://doi.org/10.1186/s13321-025-01064-7.

(35) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *Journal of Cheminformatics* **2018**, *10* (1), 10. https://doi.org/10.1186/s13321-018-0263-1.

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

(37) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17. https://doi.org/10.1021/acs.jcim.3c01250.

(38) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv June 12, 2017. https://doi.org/10.48550/arXiv.1704.01212.

(39) Turon, G.; Legese, A.; Arora, D.; Duran-Frigola, M. Ersilia Model Hub: A Repository of AI/ML Models for Infectious and Neglected Tropical Diseases, 2026. https://doi.org/10.5281/ZENODO.7274645.

(40) Liang, W.; Li, J.; Wang, X.; Giesy, J. P.; Zhao, X. Machine Learning-Driven Cross-Species Toxicity Prediction for Advancing Ecologically Relevant PFAS Water Quality Criteria. *Environ. Sci. Technol.* **2025**, *59* (48), 25688–25702. https://doi.org/10.1021/acs.est.5c12013.

(41) Chen, W.-R.; Yun, Y.-H.; Wen, M.; Lu, H.-M.; Zhang, Z.-M.; Liang, Y.-Z. Representative Subset Selection and Outlier Detection via Isolation Forest. *Anal. Methods* **2016**, *8* (39), 7225–7231. https://doi.org/10.1039/C6AY01574C.

(42) Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*; IEEE: Pisa, Italy, 2008; pp 413–422. https://doi.org/10.1109/ICDM.2008.17.

(43) Liang, Y.; Li, X.; Huang, X.; Zhang, Z.; Yao, Y. An Automated Data Mining Framework Using Autoencoders for Feature Extraction and Dimensionality Reduction. arXiv December 3, 2024. https://doi.org/10.48550/arXiv.2412.02211.

(44) Wen, T.; Cai, X.; Li, J. Graph Neural Networks vs. Traditional QSAR: A Comprehensive Comparison for Multi-Label Molecular Odor Prediction. *Molecules* **2025**, *30* (23). https://doi.org/10.3390/molecules30234605.

(45) Iliadis, D.; De Baets, B.; Waegeman, W. DeepMTP: A Python-Based Deep Learning Framework for Multi-Target Prediction. *SoftwareX* **2023**, *23*, 101516. https://doi.org/10.1016/j.softx.2023.101516.

(46) Taylor, K. E. Summarizing Multiple Aspects of Model Performance in a Single Diagram. *Journal of Geophysical Research: Atmospheres* **2001**, *106* (D7), 7183–7192. https://doi.org/10.1029/2000JD900719.

(47) Januzaj, Y.; Beqiri, E.; Luma, A. Determining the Optimal Number of Clusters Using Silhouette Score as a Data Mining Technique. *International Journal of Online and Biomedical Engineering (iJOE)* **2023**, *19*, 174–182. https://doi.org/10.3991/ijoe.v19i04.37059.

(48) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791–4810. https://doi.org/10.3390/molecules17054791.

View Article Online
DOI: 10.1039/D6SU00023A

(49) Roos, S.; Jönsson, C.; Posner, S.; Arvidsson, R.; Svanström, M. An Inventory Framework for Inclusion of Textile Chemicals in Life Cycle Assessment. *Int J Life Cycle Assess* **2019**, *24* (5), 838–847. https://doi.org/10.1007/s11367-018-1537-6.

(50) Roos, S.; Holmquist, H.; Jönsson, C.; Arvidsson, R. USEtox Characterisation Factors for Textile Chemicals Based on a Transparent Data Source Selection Strategy. *Int J Life Cycle Assess* **2018**, *23* (4), 890–903. https://doi.org/10.1007/s11367-017-1330-y.

(51) Kvasnicka, J.; Aurisano, N.; von Borries, K.; Lu, E.-H.; Fantke, P.; Jolliet, O.; Wright, F. A.; Chiu, W. A. Two-Stage Machine Learning-Based Approach to Predict Points of Departure for Human Noncancer and Developmental/Reproductive Effects. *Environ. Sci. Technol.* **2024**. https://doi.org/10.1021/acs.est.4c00172.

(52) Gramatica, P. Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. *International Journal of Quantitative Structure-Property Relationships* **2020**, *5* (3), 61–97. https://doi.org/10.4018/IJQSPR.20200701.oa1.

(53) van Tilborg, D.; Rossen, L.; Grisoni, F. Molecular Deep Learning at the Edge of Chemical Space. *Nat Mach Intell* **2026**, *8* (4), 575–587. https://doi.org/10.1038/s42256-026-01216-w.

View Article Online
DOI: 10.1039/D6SU00023A

Data available

This study was carried out using publicly available data from European Platform on LCA EPLCA at https://eplca.jrc.ec.europa.eu/LCDN/developerEF.html