Optimized Ecotoxicological Assessments: Integrating Machine Learning and Molecular Descriptors for Characterization Factor Prediction in Life Cycle Assessment of Inorganic Elements

Konstantinos M. Kritsotakis

School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechneiou st. Zografou Campus, 15780, Athens, Greece

Abstract

This study pioneers the integration of machine learning (ML) and molecular descriptors to predict Characterization Factors (CFs) for inorganic elements in Life Cycle Assessment (LCA), addressing a key gap in ecotoxicological research. CFs are essential metrics that connect chemical emissions to potential environmental impacts across various categories, including freshwater ecotoxicity (ECOTOX). Traditional methods for determining CFs rely heavily on extensive experimental data, which can be both resource-intensive and time-consuming. This research leverages advanced ML models to enhance the accuracy and efficiency of CF predictions. The methodology involves the computation of 1294 molecular descriptors for 51 elements, followed by normalization using the Yeo-Johnson transformation. Dimensionality reduction and feature selection are conducted through Spearman Rank Correlation and XGBoost feature importance. The study employs a Random Forest classifier and a Decision Tree regressor for classification and precise CF value predictions, respectively, with data augmentation achieved via the SMOTE algorithm and Gaussian copula techniques. The results reveal significant discrepancies between the predicted CFs and those in the EU Environmental Footprint v.3.1 database, for elements with low-quality scores, highlighting the uncertainties and limitations introduced by the existing methodologies. The developed model demonstrates superior accuracy in predicting CFs, validated against high-quality datapoints, thereby providing a robust and computationally efficient framework for environmental impact assessments. This integration of cheminformatics and ML offers a scalable and accurate method for CF prediction, thereby enhancing the precision of LCA practices and supporting informed environmental policy-making and resource management.

Funding

This research did not receive any specific funding.

Acknowledgements

The author extends his gratitude to his scientific collaborator, Mr. *Polykarpos Beikos*, for his invaluable assistance in surmounting various obstacles during this project. His support and inspiration were crucial in overcoming the challenges encountered throughout the development of this comprehensive model.

Keywords ecotoxicity · machine learning · molecular descriptors · life cycle assessment · characterization factors · inorganic elements

1 Introduction

In Life Cycle Assessment (LCA), Characterization Factors (CFs) are pivotal, providing a metric that relates the emissions and resource use of a chemical substance to its potential environmental impacts across various impact categories, such as global warming potential, acidification, eutrophication and others [43]. The mathematical model to calculate the impact I $(arbitary\ units)$ in a particular impact category involves multiplying the emission *E* of a substance by its respective *CF*:

$$I = E\ x\ CF \qquad \text{[Eq. 1]}$$

The *CF* itself is typically calculated by combining *Effect Factor (EfF)*, *Fate Factor (FF), Exposure Factor (XF) and Robust Factor (RF)*, with the RF specifically being defined solely within the context of the Environmental Footprint database:

$$CF = EfF\ x\ FF\ x\ XF\ x\ RF \qquad \text{[Eq. 2]}$$

,where *EfF* illustrates the potential harm a substance can cause upon exposure*, FF* quantifies the distribution of a substance in the environment, *XF* represents the degree to which the substance contacts receptors (e.g., humans or ecosystems) and RF is a numerical value representing the ratio of the highest plausible value to the lowest plausible value within the variability range of uncertain parameters associated with a substance. RF is typically set at 0.1 for non-essential metals and 0.01 for essential metals in the European Environmental Footprint (EF) database [1,10]. The values of FF, XF, and RF for a substance are linked to its physicochemical properties and are definitely well-known. In contrast, determining the EfF coefficient for a substance can be considerably challenging [14].

The calculation of CFs should be executed with meticulous attention to ensure LCA results' reliability and relevance, as detailed in several LCA methodologies and applications [43]. In the endeavor to determine reliable CFs for chemical substances, various databases and research initiatives have been developed. Notably, the USEtox database - continually refined by a global team of LCA experts - offers CFs across numerous impact categories, including human toxicity and freshwater ecotoxicity, gaining wide recognition in LCA studies and policy frameworks [2]. Similarly, the open EF database, with its updated version 3.1 featuring over 300,000 CFs across 25+ impact categories, also stands as a valuable data resource for the LCA practitioners, computing CFs through diverse data sources and modeling approaches, to uphold their relevance and reliability [11].

Leveraging Machine Learning (ML) to estimate or predict CFs, particularly when they are absent, for chemical substances in LCA has ushered in an era of more accurate and rapid assessment methodologies, especially when integrated with molecular descriptors or fingerprints. ML models, particularly those utilizing molecular descriptors, have demonstrated significant potential in accurately predicting characterization factors for chemicals [3]. These models exploit complex data patterns, resulting in more accurate and robust predictions. Ping et al (2020) demonstrated how machine learning models, specifically a random forest model, can be used to estimate ecotoxicity hazardous concentrations 50% (HC_{50}) in USEtox to calculate characterization factors for chemicals based on their physical-chemical properties [4]. In addition, methodologies, like Quantitative Structure-Activity Relationships (QSAR) models and also, hybrid models, amalgamating different types of descriptors and biological endpoints, have also been utilized for predicting the environmental impacts. Hamadache et al (2018) developed a QSAR model to predict the contact acute toxicity (LD_{50}) of 111 pesticides to bees, illustrating an alternate strategy for predicting the

toxicological impacts for chemicals [5]. Databases, such as USEtox and European EF v.3.1, derive CFs through rigorous experimental data and deterministic models, a method that is both resource and time-intensive. ML algorithms, when properly developed, ensure the robust and computationally efficient derivation of CFs, providing accurate estimations, that not only fill the existing gaps in the databases but also expedite their faster expansion.

Molecular descriptors, quantifiable properties of molecules encompassing their physicochemical, electronic and topological characteristics, can be strategically exploited to correlate with potential ecotoxicity impacts. Xingmei et al (2020) successfully developed a machine learning model to predict the acute toxicity of diverse chemicals on fathead minnows, using six molecular descriptors for validation [6]. In addition, molecular fingerprints, encoding the structural and topological information of molecules into binary strings or count vectors, emerge as a compelling methodology for estimating CFs. Weizhe et al (2022) developed predictive models using molecular dynamic simulation and multi-dimensional molecular fingerprints to accurately predict hERG cardiotoxicity of compounds using machine learning algorithms. Xie et al (2020) showed how the accurate prediction of physical properties and bioactivity of drug molecules in deep learning depends on how molecules are represented through fingerprints, thus highlighting the importance of molecular representation in machine learning models [8]. Also, Ucak et al (2023) highlighted that applying a more refined molecular representation, specifically atom-in-SMILES (AIS) tokenization, which tokenizes molecules into individual atoms, rather than groups of atoms, not only significantly enhances prediction quality in learning models but also surpasses other schemes in both accuracy and token's degeneration minimization, markedly boosting machine learning models in predicting chemical properties.

The scientific community has already agreed on the establishment of a new ecotoxicological endpoint HC_{20} (a concentration estimate where 20% of exposed organisms experience harmful effects, derived from Species Sensitivity Distributions (SSD) curves using chronic EC_{10} endpoints) for the ECOTOX impact category [11]. Notably, not all elements have this value, therefore the calculation of $CF_{HC_{20}}$ value for 27 elements in the EF v.3.1 database, for the ECOTOX category, was done by converting their HC_{50} value (which is provided by USEtox) to HC_{20}, by applying 0.34 as an extrapolation factor based on the proposal of Saouter et al 2018 [10, 14]. However, this calculation methodology led to an overestimation of CF values of these elements [11]. In this study, we developed an advanced machine learning model capable of predicting $CF_{HC_{20}}$ values (and by extension HC_{20} values, since $HC_{20} = 0.2 / EfF$) for inorganic elements. This was achieved using molecular descriptors from various cheminformatics toolkits to estimate the CFs of 51 elements (metals, metalloids, and non-metals) from the EF v.3.1 European database for the "Ecotoxicity Freshwater-inorganics" impact category. The model is applicable to inorganic elements, primarily metals and metalloids, and to any other elements exhibiting molecular descriptor distributions similar to those of metals and metalloids, details of which will be subsequently elucidated. This model is not limited to elements currently listed in the EF v.3.1 European database but also accommodates inorganic elements that may be integrated to it in the future. By eliminating the need for constructing chronic NOEC-equivalent SSDs to calculate HC_{20} values and providing essential CF data for elements lacking experimental data, this approach fills critical data gaps and greatly improves LCA study efficiency. The significance of this study is further emphasized by the findings from Bassi et al (2023) at the EC – Joint Research Centre, which highlight that, metals play a substantial role in toxicity impact categories, thereby emphasizing the critical importance of accurately predicting their CF as well as HC_{20} value. Finally, while the current model primarily addresses chemical elements, there is potential to expand it to include both inorganic molecules and organometallic compounds as identified in the EF v.3.1 database. This extension would enhance the predictive analysis of CFs for freshwater ecotoxicity and support the determination of HC_{20} values for

inorganics. The significant void in the literature concerning inorganic substances, as opposed to organics, underscores the urgent necessity and paramount importance of this research.

2 Materials and Methods

For the 51 elements listed in **Table 1.** of the EF v.3.1 database, a total of 1294 molecular descriptors were initially computed using toolkits based on Python or Java. Detailed computational data for these descriptors can be found in **Supplementary Material 1**.

Relying solely on a single toolkit for the extraction of molecular descriptors would complicate the identification of suitable features that could lead, through the predictive model, to the determination of either CFs or HC_{20} values for inorganics. This is particularly challenging given that existing cheminformatics platforms are primarily designed with a focus on organic compounds. For the generation of molecular descriptors using all this software, it is necessary to convert the 51 names of these elements into SMILES format in order to serve as inputs for these toolkits. This conversion was performed using PubChem & ChemSpider platforms [12,13]. In addition to the 1294 molecular descriptors, the FF, XF and RF were also used as general input data for the model. These coefficients were sourced from the database's "Ecotox Explorer". The "Ecotox Explorer" is an online interactive application developed by Sala et al (2022), that provides users with access to underlying ecotoxicological data and physico-chemical properties, enabling the comparison of chemicals, interpretation of results, and access to key research insights and external resources [15].

In the harmonization effort outlined by EF v.3.1 database, the CFs of aluminium, cadmium, cobalt, copper, lead, manganese, mercury, molybdenum, nickel, silver, tin, vanadium, zinc, magnesium and boron were derived using robust ecotoxicological endpoints (HC_{20}) and reliable EU data sources, ensuring their reliability. For elements lacking EU ecotoxicological data, EC – JRC's researchers derived $CF_{HC_{20}}$ and HC_{20} values using significant assumptions. For instance, they converted HC_{50} values from USEtox to HC_{20} values by applying an extrapolation factor of 0.34 (i.e., $HC_{50} = HC_{20} \times 0.34$), which consequently increases CFs by approximately 18%. This approach, based on elementary equations, provides a necessary yet inherently imprecise method for estimating ecotoxicity impacts, as evidenced by the fact that the reliable CF values of the aforementioned metals differ by up to 1173% from the estimated values derived using the 0.34 extrapolation factor. Additionally, CFs for certain elements were aligned with similar speciations or averaged across two forms, thereby enhancing the approximation involved in the ecotoxicity assessment of the inorganic elements [11].

Due to inherent uncertainties in determining CF and HC_{20} values, the assigned, by EF v.3.1 database, CF values are considered approximate rather than exact (with the exception of the aforementioned 15 elements). Nonetheless, they provide a useful framework for determining the order of magnitude of CFs for these elements. Consequently, the predictive model was divided into two parts to calculate the final CF values for the inorganic elements. Initially, a classification model was developed to categorize the elements into classes based on the magnitude of their estimated-by-the-EF-v.3.1-database CF value. The following boundaries for defining each class were meticulously established through iterative trials until they accurately predicted all unseen datapoints with respect to their class, a process which is explained in more detail below.

- Elements with CF value between 0.1 and 9.9 are classified into Class 0
- Elements with CF value between 10 and 499 are classified into Class 1

- Elements with CF value between 500 and 1700 are classified into Class 2
- Elements with CF value between 1701 and 9999 are classified into Class 3
- Elements with CF value between 10000 and 99999 are classified into Class 4
- Elements with CF value between 100000 and 999999 are classified into Class 5

Following the initial classification, which leverages selected molecular descriptors as will be elaborated later, a regression-based machine learning model is developed separately for each class and then calculates the precise CF value for each element. This sophisticated two-tiered approach, despite initial data limitations, yields a robust assessment of the ecotoxicity impact of the elements. This is confirmed by Servien et al. (2022), who found that clustering followed by regression is superior to direct regression because it enhances predictive accuracy by addressing the inherent heterogeneity and non-linearity in the data, simplifying the regression task through pre-classification into groups, and improving the robustness and reliability of predictions. It is worth emphasizing again that the assignment of CF values to elements with significant uncertainties by EC-JRC researchers, while based on approximations, ultimately provided a reliable framework for determining the orders of magnitude of CFs for these elements.

As previously mentioned, our supervised machine learning project aimed to construct a classification and regression model to predict the CFs of various inorganic elements. However, a critical question arises here: which labeled datapoints from the EF v.3.1 database could be considered reliable to serve as the training set for the model? According to researchers at the EC-JRC [14], only data characterized by a high or average Quality Score should be deemed reliable in terms of their precise CF values. Consequently, only these datapoints could be selected for validating the classification or regression model, while the others can be regarded reliable only in terms of the order of magnitude of their CF values. To this end, we selected ten out of fifteen high-quality and reliable CF values, encompassing classes 0 through 5, to serve as the external validation datapoints [46], as illustrated in **Figure 1**. The accuracy with which the classification model categorizes these ten validation datapoints serves as a robust indicator of its efficacy. In other words, the successful classification of validation datapoints substantiates the model's proficiency in reliably determining the magnitude of the CF values for new, unseen inorganic elements, thereby validating its predictive robustness.

Table 1 Data derived from "Ecotox Explorer" for the 51 inorganic elements existing in the EF v.3.1 database

Name	SMILES	Database's CF*	Database's HC_{20}	FF	XF	RF	Quality Score**
bismuth	[Bi]	0,28	0,036428571	0,13	3,9	0,1	Only one test
molybdenum	[Mo]	0,98	0,031632653	0,97	16	0,01	High
magnesium	[Mg]	1,2	0,031666667	1	19	0,01	High
phosphorus	[P]	2	0,035	1	3,5	0,1	Only one test
silicon	[Si]	4,7	0,080851064	1	19	0,1	Only one test
titanium	[Ti]	4,8	0,00325	0,23	3,4	0,1	Low
boron	[B]	24	0,015833333	1	19	0,1	High

copper	[Cu]	46	0,000104348	0,49	4,8	0,01	Unavailable
copper (ii)	[Cu+2]	46	1,30435E-05	0,019	15	0,01	High
zirconium	[Zr]	48	0,007916667	1	19	0,1	Only one test
manganese	[Mn]	52	0,000730769	1	19	0,01	Unavailable
manganese (ii)	[Mn+2]	52	0,000730769	0,49	93	0,01	Average
lead	[Pb]	68	8,82353E-05	0,082	3,7	0,1	Unavailable
lead (ii)	[Pb+2]	68	2,64706E-05	0,0079	12	0,1	High
tungsten	[W]	81	0,004444444	1	18	0,1	Low
selenium	[Se]	86	0,000332558	0,95	15	0,01	Unavailable
selenium (iv)	[Se+4]	86	0,000569767	0,72	34	0,01	Unavailable
antimony (iii)	[Sb+3]	140	0,01023	0,93	77	0,1	Unavailable
lithium	[Li]	150	0,002533333	1	19	0,1	Low
iron (ii)	[Fe+2]	160	0,00135	0,9	120	0,01	Unavailable
tellurium	[Te]	500	0,00076	1	19	0,1	Low
tin	[Sn]	510	0,000705882	1	18	0,1	Unavailable
tin (ii)	[Sn+2]	510	0,000342745	0,46	19	0,1	Average
chromium (iii)	[Cr+3]	950	1,71368E-07	0,00074	11	0,1	Unavailable
beryllium (ii)	[Be+2]	1600	0,00000875	0,044	16	0,1	Unavailable
zinc	[Zn]	1700	2,23529E-05	1	19	0,01	Unavailable
zinc (ii)	[Zn+2]	1700	5,52941E-05	0,5	94	0,01	High
arsenic (iii)	[As+3]	1800	0,000791111	0,89	80	0,1	Unavailable
barium (ii)	[Ba+2]	3700	0,000741622	0,98	140	0,1	Unavailable
cerium(3^{+})	[Ce+3]	3900	9,23077E-05	1	18	0,1	Low
iron (iii)	[Fe+3]	4100	9,7561E-08	0,016	14	0,01	Unavailable
thallium (i)	[Tl+]	4100	0,000491707	0,84	120	0,1	Unavailable
aluminium	[Al]	4400	8,63636E-05	1	19	0,1	Unavailable
aluminium (iii)	[Al+3]	4400	9,81818E-05	0,18	120	0,1	High
cesium (i)	[Cs+]	4500	0,000616	0,99	140	0,1	Unavailable
arsenic (v)	[As+5]	4700	0,000302979	0,89	80	0,1	Unavailable
vanadium	[V]	4800	1,81667E-05	0,67	6,5	0,1	High
cobalt	[Co]	5500	9,81818E-07	0,53	5,1	0,01	Unavailable
cobalt (ii)	[Co+2]	5500	2,90909E-05	0,86	93	0,01	High
antimony	[Sb]	11000	1,98545E-05	0,91	12	0,1	Unavailable

chromium (vi)	[Cr+6]	12000	0,00012555	0,81	93	0,1	Unavailable
strontium (ii)	[Sr+2]	18000	0,000128	0,96	120	0,1	Unavailable
antimony (v)	[Sb+5]	22000	0,0000651	0,93	77	0,1	Unavailable
nickel	[Ni]	27000	9,85185E-06	0,95	14	0,1	Unavailable
nickel (ii)	[Ni+2]	27000	5,78519E-05	0,71	110	0,1	High
mercury	[Hg]	33000	1,46061E-06	0,28	8,6	0,1	Unavailable
mercury (ii)	[Hg+2]	33000	2,07273E-06	0,19	18	0,1	High
silver	[Ag]	180000	1,18889E-07	0,29	3,7	0,1	Unavailable
silver (i)	[Ag+]	180000	0,00000082	0,41	18	0,1	High
cadmium	[Cd]	670000	5,67164E-07	1	19	0,1	Unavailable
cadmium (ii)	[Cd+2]	670000	1,97015E-06	0,66	100	0,1	High

• CFs pertain to the impact category Freshwater Ecotoxicity (or ECOTOX).

•• Quality Score is a metric used to evaluate the reliability of SSD curves, where only those derived from data on multiple species and taxonomic groups, classified as ''average'' or ''high'', are deemed reliable. Additionally, data classified as "Unavailable" are excluded from the quality score evaluation due to the unavailability of reliable ecotoxicological data from both EU and other reputable sources [11,14,15].

Name	Quality Score	Labeled Class	Database's CF
magnesium	high	0	1,2
copper (ii)	high	1	46
manganese (ii)	average	1	52
lead (ii)	high	1	68
tin (ii)	average	2	510
zinc (ii)	high	2	1700
aluminum (iii)	high	3	4400
cobalt (ii)	high	3	5500
mercury (ii)	high	4	33000
cadmium (ii)	high	5	670000

Fig. 1 Reliable elements for classification model's validation: *Data assigned to classes based on CF values*

1. Goal & Scope

This study aims to enhance the accuracy of estimating ECOTOX CFs and HC_{20} values in LCA for all of the inorganic elements listed in the EF v.3.1 database. Utilizing a combination of machine learning and a huge amount of diverse molecular descriptors, the research focuses on refining CF predictions, particularly emphasizing the reliable calculation of HC_{20}, a crucial indicator of ecotoxicological impact. This effort addresses existing gaps in CF methodologies by applying sophisticated computational models that not only cater to current database elements but also anticipate future additions. Through this approach, the study seeks to integrate robust, scientifically validated data into LCA practices, thereby enhancing environmental policy-making and resource management by providing more precise and actionable environmental impact assessments.

2. Exploratory Analysis & Data Filtering

Data cleaning and preprocessing are indispensable steps in machine learning model development, primarily aimed at enhancing data quality to improve model performance and reliability. Key data cleaning tasks include handling missing values through techniques like imputation or deletion, removing duplicates to prevent overfitting, filtering out noise and outliers using statistical methods, and correcting structural errors such as inconsistencies in units [2]. In this study, the initial dataset of molecular descriptors was reduced from 1297 to 190 by primarily eliminating columns with non-varying values across the datapoints.

After the cleaning process, applying a normalization (or standardization) technique is essential to ensure that all model input data is numerical and consistently scaled. Feature engineering further refines the dataset by creating or modifying features to increase predictive capability, coupled with robust feature selection methods to streamline the model by retaining only the most relevant ones. Additionally, data partitioning into training, validation and testing sets aids in unbiased model evaluation and hyperparameter tuning. These preprocessing steps are essential not only for training accurate and efficient models but also for enhancing the models' ability to generalize from training data to real-world applications, thereby substantiating their practical efficacy in diverse settings [4,16,17].

Normalization technique

In the thorough examination of our initial dataset of all possible molecular descriptors (see Supplementary Material 1), we are confronted with multiple normalization challenges, vividly depicted in **Figures 2-4**. Our initial data not only exhibit pronounced skewness (Fig. 3) but also contain a mix of exceptionally high and low values across various features (Fig. 2), alongside the presence of zero and negative values (Figure 4). To illustrate these challenges, we randomly selected a subset of descriptors (ATSC0dv, FilterItLogS, SLogP, and ATS0m), effectively representing these normalization issues. These challenges significantly complicate the application of standard normalization techniques.

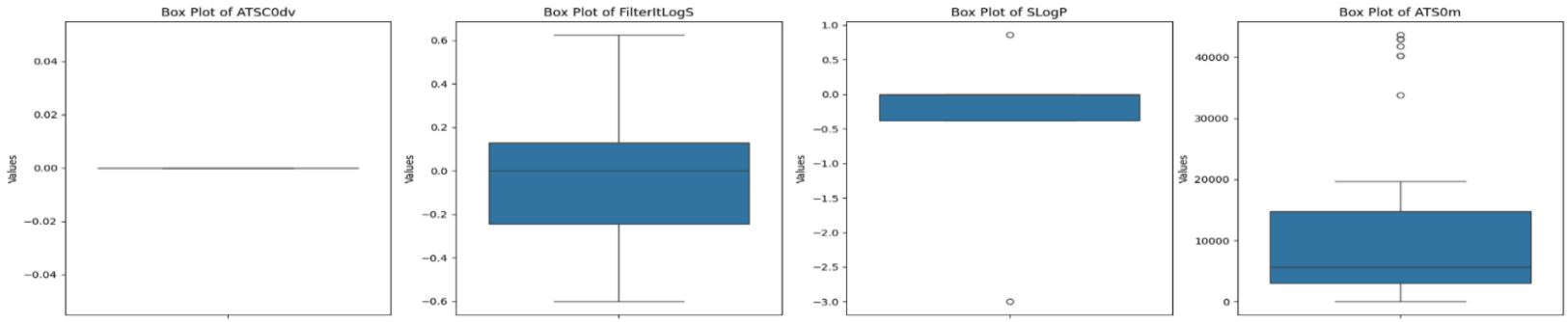

Fig.2 Box Plots of some selected molecular descriptors: *Highlights variability and outlier presence, underscoring the need for robust normalization to manage outliers effectively*

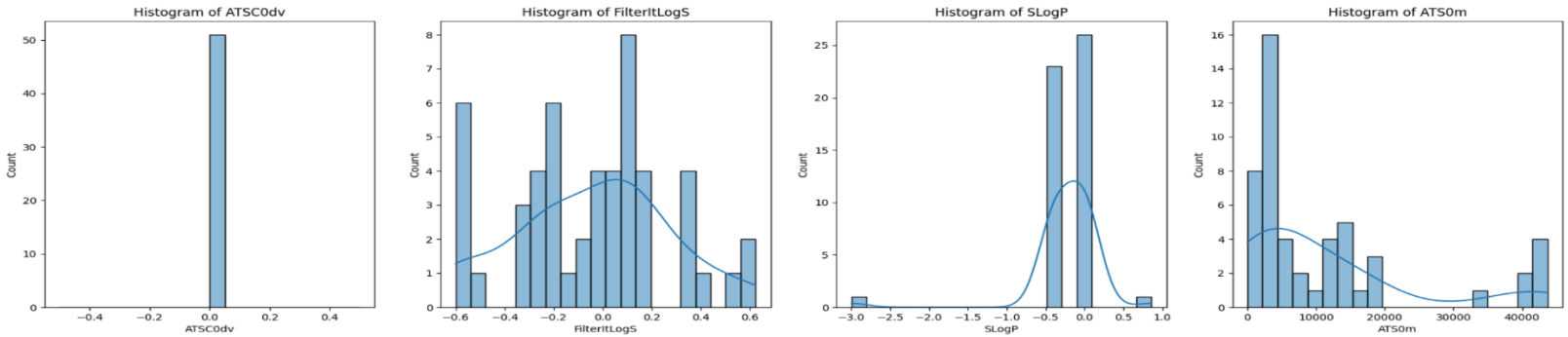

Fig.3 Frequency distribution of sample descriptors: *Showcases the diversity of distribution shapes and skewness, emphasizing the complexity in feature behavior*

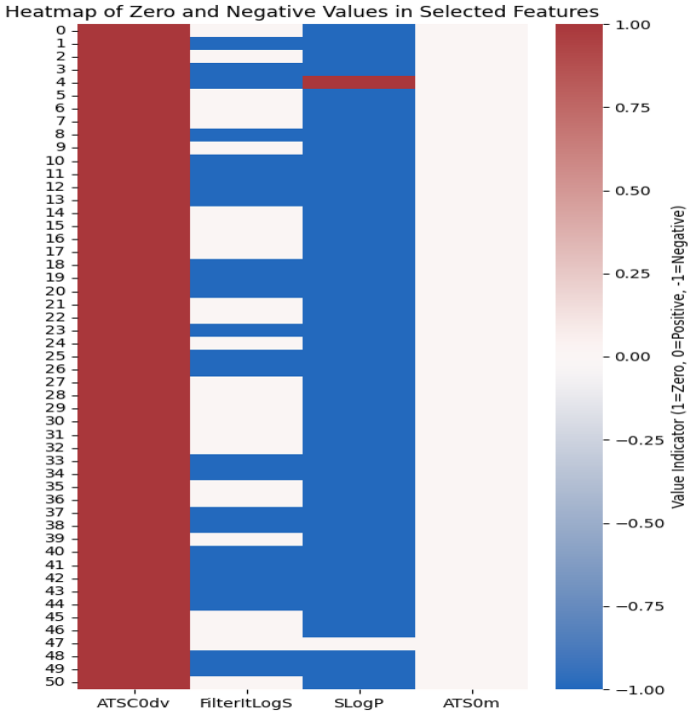

Fig.4 Heatmap of Zero & Negative values for the selected molecular descriptors: *Visualizes the prevalence of zero and negative values, illustrating the challenges in direct logarithmic or the Box-Cox transformation*

Z-score normalization, which standardizes data to achieve zero mean and unit variance, does not inherently accommodate skewness and demonstrates a pronounced sensitivity to outliers, as substantiated by the considerable variability and outlier visibility in Fig.2 [17]. This predisposition may lead to distortions in the representation of the underlying distribution. Additionally, the presence of skewness within the dataset is compellingly demonstrated by Fig. 3, where the distribution profiles of molecular descriptors such as SLogP and ATS0m exhibit marked asymmetry. Notably, the bimodal behavior observed in these descriptors underscores the complexity of the data, further challenging the efficacy of conventional normalization methods. Min-Max normalization, another traditional technique, by linearly mapping data to a bounded [0,1] range, can excessively compress the dynamic range of this dataset. This compression risks obfuscating critical patterns and nuances, particularly evident in the presence of extreme values as shown in Fig. 2. Such transformation may result in a loss of meaningful variability essential for accurate data interpretation. Furthermore, log normalization, often utilized to mitigate right skewness, is inapplicable to this dataset, due to the presence of zero and negative values. Additionally, despite its effectiveness in reducing outlier influence, the sigmoidal normalization with the

Hyperbolic Tangent (HT) function falls short in addressing skewness and accommodating diverse distribution patterns, leaving the asymmetrical and varied profiles of molecular descriptors inadequately normalized [18].

The Yeo-Johnson transformation, an effective alternative to all above, specifically engineered to address skewness and accommodate zero and negative values, excels in normalizing complex distributions, by dynamically adjusting its parameters to closely align with the underlying distributional shape. This adaptability ensures the preservation of intricate data characteristics, making it really effective for datasets with the diverse and challenging features depicted in Figures 2 through 4. Superior to the Box-Cox transformation, which only handles positive values, the Yeo-Johnson extends its functionality with power transformations—logarithmic, square, and inverse—that are tailored to the specific skewness and kurtosis of the data [19].

Correlation analysis

When dealing with a large dataset containing numerous features, it is crucial to employ effective techniques for feature reduction that prioritize both accuracy and interpretability. In the context of raw datasets, which exhibit challenging characteristics as previously discussed, identifying the most relevant features becomes even more crucial. **Figure 5** illustrates the heatmap of the remaining molecular descriptors after applying Spearman Rank Correlation for dimensionality reduction, which reduced the raw dataset to 15 essential descriptors.

Spearman Rank Correlation provides an effective approach for identifying uncorrelated features, especially when dealing with datasets that exhibit non-linear dependencies or asymmetrical distributions. This non-parametric method measures the strength and direction of association between ranked variables, making it highly suitable for datasets where the relationships are better described using a monotonic function rather than a linear one [20]. By concentrating on data rankings instead of absolute values, Spearman's correlation effectively captures complex relationships while remaining resilient to outliers, aligning well with the nuanced characteristics of our dataset.

Non-linear Principal Component Analysis (PCA) algorithms, such as kernel PCA, along with autoencoders, are effective tools for dimensionality reduction and excel at capturing non-linear relationships within, especially complex, datasets [44,48]. Nevertheless, despite their proven effectiveness for our dataset, these methods were excluded from our definitive model due to considerable limitations. Both techniques transform the original features into new constructs - principal components in the case of kernel PCA and encoded features in autoencoders - which lead to a critical loss of the physical meaning of the molecular descriptors. This challenge is exacerbated by the irreversible loss of information inherent in these transformations, further complicating any attempts to reconstruct the original features accurately [27,28]. Consequently, this transformation impedes applications where specific knowledge and transparent features are essential, as in our study which aims to identify measures that reveal tangible quantities with physical significance, such as CF or HC_{20}. Given these constraints, Spearman Rank Correlation is favored for dimensionality reduction in our analysis, as it effectively balances accuracy with the retention of interpretability in the original descriptors.

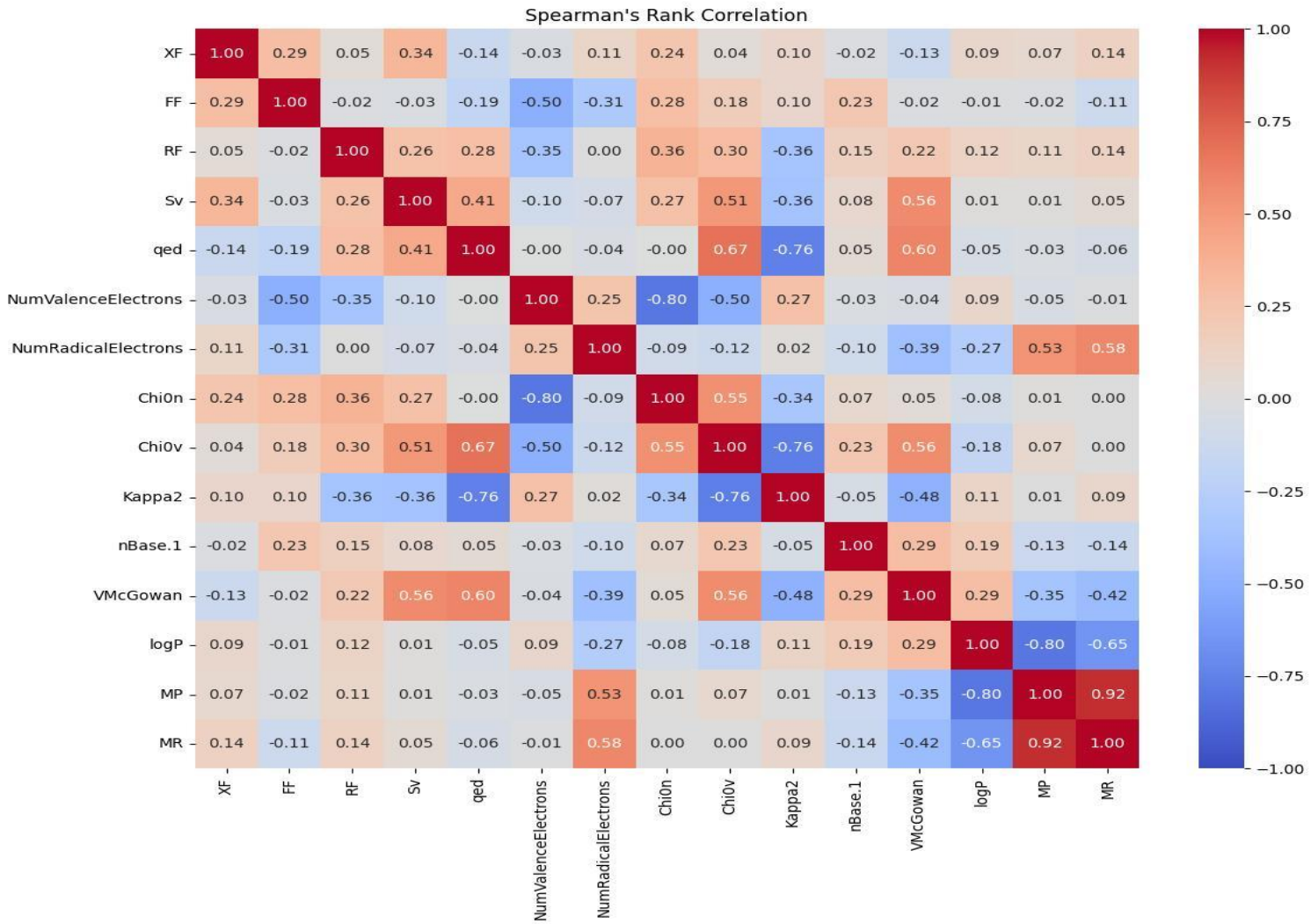

Fig.5 Feature Correlation Analysis: *Heatmap of retained independent molecular descriptors*

Feature selection technique

In the domain of supervised classification, selecting crucial features is essential for achieving robust predictive performance on unseen data. Among the array of feature selection techniques tested, XGBoost (eXtreme Gradient Boosting) has consistently emerged as a superior choice compared to Random Forest classifier, the Recursive Feature Elimination (RFE), Lasso Regression, and the Mutual Information. In our analysis, only the XGBoost selection algorithm succeeded in developing a model capable of accurately classifying the CF values of the ten unseen elements (Fig.1).

XGBoost classifier excels in feature importance ranking through its unique boosting mechanism, where each new tree corrects errors from previous iterations, creating a highly efficient model capable of capturing nuanced patterns and highlighting features vital for classification accuracy. Unlike Random Forest, which utilizes bagging (bootstrap aggregating), XGBoost's sequential boosting can easily focus on challenging cases, refining feature importance [21,26]. This contrasts with RFE, which systematically removes the least important features but can be computationally intensive and struggle with complex interactions, while XGBoost dynamically adjusts feature weights to refine selection [22]. Lasso Regression, typically effective when the number of features exceeds the number of observations, may suppress important features in our dataset with 15 features (Fig.5) and 41 observations [23]. On the other hand, Mutual Information evaluates features independently and overlooks interdependencies, while XGBoost's decision tree-based architecture captures these interactions, offering a holistic view of feature importance and enhancing its suitability for complex classification tasks [24]. The robust boosting mechanism, sophisticated importance ranking, and capacity to handle complex interactions make XGBoost exceptional for such classification task, capturing nuanced patterns and emphasizing critical features, ensuring optimal performance on unseen datapoints.

In the task of predicting the class of inorganic elements according to the order of magnitude of their CF value, the XGBoost classifier's feature importance algorithm plays a crucial role in evaluating the impact of molecular descriptors on prediction accuracy. This methodological approach provides detailed insights into which descriptors most significantly influence the classification of elements into their respective classes, enhancing the precision of the classification model. The application of the XGBoost classifier's feature importance algorithm facilitates a detailed evaluation of feature efficacy, with the analysis performed on the refined set of fifteen features previously distilled through Spearman rank correlation test (Fig.5). As illustrated in **Table 2.**, the importance scores of each of these fifteen molecular descriptors are precisely computed to clarify their respective contributions to the final predictive capability of the model. This assessment is crucial for validating the model's predictive accuracy against, firstly, the ten validation datapoints (Fig.1) and, then, for verifying the model's capacity to accurately classify any real-world inorganic element. The analysis is executed through a ten-fold cross-validation process, ensuring that the importance scores are both robust and reliable, effectively minimizing variability and bias in the model's performance across different data subsets [25]. This structured validation framework not only enhances the statistical rigor of the feature importance assessment but also guarantees that the classification model is both accurate and robust, facilitating dependable predictions essential for a strategic classification of inorganic elements.

Table 2 Importance scores of the fifteen molecular descriptors via the XGBoost Classifier's Feature Importance algorithm

Molecular Descriptor	Importance Score
XF	0.0422
FF	0.0406
RF	0.0958
Sv	0.0774
qed	0.1201
NumValenceElectrons	0.0734
NumRadicalElectrons	0.0493
Chi0n	0.0324
Chi0v	0.1684
Kappa2	0.0882
nBase.1	0.0304
VMcGowan	0.0873
logP	0.0691
MP	0.0000
MR	0.0253

Data Augmentation

To construct any machine learning model for predictive analysis of the elements' ecotoxicity impacts, an expansive and comprehensive dataset is essential. In the feature selection phase, a robust and extensive dataset ensures that the molecular descriptors identified through the XGBoost feature selection algorithm truly represent broader data patterns, mitigating bias towards a limited subset of data. Similarly, in subsequent classification and regression analyses, this extensive dataset underpins the models' accuracy and enhances their generalizability, ensuring that predictions are reliably extrapolated across different scenarios. Given the constraints imposed by a dataset comprising only 41 elements in our study (51 elements in total, excluding the 10 validation datapoints), subdivided further by class, the likelihood of biased results increases if not addressed appropriately. To counteract potential overfitting and enhance the representativeness of each class within the dataset, data augmentation techniques are employed. To ensure that data augmentation techniques accurately reflect the distribution of features within each class, thereby generating new datapoints that follow the distribution of the original class-specific points, a detailed analysis of the distribution patterns is necessary. This preliminary step is crucial to verify that the augmented dataset maintains the integrity of the original distribution, allowing the enhanced dataset to support robust and reliable model training without introducing unintended biases or altering the fundamental characteristics of the data [29,30].

To optimize the development of a classification model, a variety of data augmentation techniques can be applied, including SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling Approach), random oversampling, Borderline-SMOTE, and k-means SMOTE. Each method is designed to address specific challenges in dataset dynamics, enhancing model robustness through tailored strategies. SMOTE is particularly acclaimed for its capability to generate synthetic samples through interpolation between similar instances, effectively preserving the multidimensional distribution of molecular descriptors. This method not only enriches the dataset diversity but also maintains the integrity of the original statistical properties, unlike ADASYN which adjusts synthetic sample generation based on local minority class density, potentially biasing the model towards complex boundary regions [53]. In comparison, random oversampling simply replicates existing datapoints, increasing the risk of overfitting and not contributing new informational diversity to the model [52]. Both Borderline-SMOTE and K-means SMOTE refine sample generation by targeting specific areas within the data—near decision boundaries and through cluster-based synthesis, respectively. However, these approaches may subtly shift the inherent class distribution, unlike SMOTE's uniform synthesis across the feature space [50,51]. This comprehensive application of SMOTE in our dataset preparation ensures a balanced enhancement of data diversity without skewing the original distribution, thereby supporting the development of robust, generalizable classification model. Given these considerations, SMOTE was selected for its demonstrated efficacy in producing a balanced dataset expansion that faithfully reflects the original molecular descriptor distributions, as detailed in the respective similarity tests in **Supplementary Material 2**. This strategic augmentation aligns with the goal of minimizing bias, particularly the types introduced by less sophisticated oversampling methods, making it an ideal choice for our classification task.

In regression tasks, where the target variable - CF values - is represented by real numbers instead of categorical data like class numbers, data augmentation techniques such as SMOTER and Gaussian Mixture Models (GMM) could be effectively employed to replicate the distribution of the original dataset. Each of these two methods brings a distinct advantage to the augmentation process, ensuring that the synthetic data generated is both representative and robust. GMM is particularly valued for its ability to model complex distributions by representing them as mixtures of multiple Gaussian distributions, effectively capturing the variability and subtleties within the dataset [54]. Conversely, SMOTER enhances the minority class representation by generating synthetic examples from challenging samples, thereby increasing the overall diversity within the training dataset. Each method contributes to a more comprehensive and nuanced augmentation strategy, which is crucial for developing regression models that are both accurate and generalizable. However, both techniques lack inherent support for extrapolating target values, a capability that is essential for ensuring comprehensive data coverage within predefined value ranges. This is particularly vital in scenarios like ours, where the training set for each class must encompass the entire spectrum of class-specific values. For instance, in constructing a regression model for Class 1, which spans values from 10 to 499, the original dataset does not adequately cover this range. Consequently, it is imperative to synthesize at least two new datapoints at the extremities of this interval to initiate an expansive data generation process that guarantees comprehensive coverage across the specified range. To meet the specific needs of our regression task, the Gaussian Copula Multivariate method was selected for its proficiency in closely adhering to the original molecular descriptor distributions and its exceptional capability in generating datapoints with extrapolated target values. This approach successfully upholds the statistical integrity of the original dataset, preserving its distributional and variance characteristics as validated by statistical tests detailed in Supplementary Material 2 [49]. This methodology facilitates the development of a robust and highly generalizable regression model designed to predict across a comprehensive range of target values, with an emphasis on preserving data integrity and broadening value ranges to enhance predictive accuracy and reliability in real-world environmental impact assessments.

3 Results and Discussion

Development and validation of the classification model

To further refine the molecular descriptors and isolate only those capable of accurately predicting the CF class, the XGB feature selection algorithm is applied to the SMOTE-augmented dataset, rather than the initial dataset, for the reasons previously discussed. Subsequently, through a rigorous 10-fold cross-validation process, a refined subset of key descriptors is identified from the prior set that was retained following the application of the Spearman Rank Correlation. This selection process performed by the XGBoost algorithm is contingent on the defined XGBoost importance threshold parameter, which typically ranges from 0.1 to 2. Higher values of this parameter result in fewer selected features, as the minimum importance threshold for a feature to be retained increases.

Once the important molecular descriptors are selected, the SMOTE-augmented dataset, incorporating only these selected features, is then employed as the training dataset. The classification machine learning models subsequently employed to train on this augmented dataset include the XGBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, K-Neighbors Classifier, and Decision Tree Classifier. The parameter tuning for these classification models, along with the explanation of these parameters (Pedregosa et al., 2011), is detailed in **Table 3.**

Table 3 Parameter tuning for the classification models. *This table provides detailed information on the critical parameters and their specific values used for tuning the various classification models in this study*

Classifier	Parameter	Value	Explanation
XGBoost	random_state*	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
	n_estimators	100	Number of boosting rounds. Typically ranges from 100 to 1000. More rounds increase accuracy but also computation time
	max_depth	6	Maximum depth of a tree. Typically ranges from 3 to 10. Controls model complexity
	learning_rate	0.1	Step size shrinkage to prevent overfitting. Typically ranges from 0.01 to 0.3. Lower values require more boosting rounds
	booster	"gbtree"	Type of booster used. Common options include "gbtree", "gblinear", and "dart"
	subsample	1	Fraction of samples used for each tree. Typically ranges from 0.5 to 1. Lower values help prevent overfitting
	colsample_bytree	1	Fraction of features used for each tree. Typically ranges from 0.5 to 1. Lower values help prevent overfitting

	gamma	0	Minimum loss reduction for partitioning. Typically ranges from 0 to 0.5. Higher values make the algorithm more conservative
	min_child_weight	1	Minimum sum of instance weight needed in a child. Typically ranges from 1 to 10. Higher values prevent overfitting
	objective	"binary:logistic"	Specifies the learning task and objective. Commonly used for binary classification.
	eval_metric	"logloss"	Evaluation metric for validation data. Common choices include "logloss" and "error"
Random Forest	random_state	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
	n_estimators	100	Number of trees in the forest. Typically ranges from 100 to 1000. More trees increase accuracy but also computation time
	criterion	"gini"	Function to measure quality of a split. Options include "gini" and "entropy"
	max_depth	None	Maximum depth of the tree. Typically ranges from 10 to None. Controls model complexity
	min_samples_split	2	Minimum samples required to split an internal node. Typically ranges from 2 to 10. Higher values prevent overfitting
	min_samples_leaf	1	Minimum samples required at a leaf node. Typically ranges from 1 to 4. Higher values prevent overfitting
	max_features	"auto"	Number of features for the best split. Common options include "auto", "sqrt", and "log2"
	bootstrap	True	Whether bootstrap samples are used. Typically set to True for better generalization
Gradient Boosting	random_state	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
	n_estimators	100	Number of boosting stages. Typically ranges from 100 to 1000. More stages increase accuracy but also computation time
	loss	"deviance"	Loss function to optimize. Common options include "deviance" and "exponential"
	learning_rate	0.1	Shrinks contribution of each tree. Typically ranges from 0.01 to 0.3. Lower values require more stages
	max_depth	3	Maximum depth of individual regression estimators. Typically ranges from 3 to 10. Controls model complexity

	subsample	1	Fraction of samples for fitting base learners. Typically ranges from 0.5 to 1. Lower values help prevent overfitting
	criterion	"friedman_mse"	Function to measure quality of a split. Options include "friedman_mse", "mse", and "mae"
AdaBoost	random_state	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
	n_estimators	50	Number of weak learners to train. Typically ranges from 50 to 500. More learners increase accuracy but also computation time
	learning_rate	1.0	Weight applied to each classifier at boosting iteration. Typically ranges from 0.01 to 1.0. Lower values improve robustness
	algorithm	"SAMME.R"	Algorithm type used. Options include "SAMME" and "SAMME.R".
K-Neighbors	n_neighbors	5	Number of neighbors to use. Typically ranges from 3 to 15. Lower values capture more noise, higher values smooth predictions
	weights	"uniform"	Weight function used in prediction. Options include "uniform" and "distance"
	algorithm	"auto"	Algorithm to compute nearest neighbors. Options include "auto", "ball_tree", "kd_tree", and "brute"
	leaf_size	30	Leaf size for BallTree or KDTree. Typically ranges from 20 to 40. Affects speed and memory
	p	2	Power parameter for Minkowski metric. 1 for Manhattan distance, 2 for Euclidean distance
Decision Tree	random_state	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
	criterion	"gini"	Function to measure quality of a split. Options include "gini" and "entropy"
	splitter	"best"	Strategy to choose split at each node. Options include "best" and "random"
	max_depth	None	Maximum depth of the tree. Typically ranges from 10 to None. Controls model complexity
	min_samples_split	2	Minimum samples required to split an internal node. Typically ranges from 2 to 10. Higher values prevent overfitting
	min_samples_leaf	1	Minimum samples required at a leaf node. Typically ranges from 1 to 4. Higher values prevent overfitting

	max_features	None	Number of features for the best split. Common options include "auto", "sqrt", "log2", and None

* In this case, it is set to a very large value ($> 10^6$) to emphasize reproducibility across multiple runs. The size of the random_state value does not inherently improve the model's performance but ensures that the results are consistent and not subject to random variations.

The test size is consistently set to 0.25, and the unseen validation dataset comprises the 10 high-quality reliable elements (see Fig. 1), which serve as the model's external validation points. Each trained machine learning model is then evaluated using a series of metrics for classification tasks, including:

- **Accuracy**: Represents the proportion of true results (both true positives and true negatives) among the total number of cases examined. It indicates the overall effectiveness of the model in making correct predictions.
- **Precision**: The ratio of true positive results to all positive predictions made by the model. It indicates how many of the positively classified instances are actually relevant, thereby reflecting the accuracy of the positive predictions.
- **Recall** (Sensitivity): The ratio of true positive results to all actual positive instances. It measures the model's ability to correctly identify all relevant instances in the dataset.
- **F1 Score**: Measures the harmonic mean of precision and recall, balancing both metrics. It is particularly useful for imbalanced class distributions as it gives a single metric that accounts for both false positives and false negatives.
- **Matthews Correlation Coefficient** (MCC): Provides a balanced measure of the quality of classifications, considering true and false positives and negatives across all classes. It is particularly useful for imbalanced datasets as it takes into account all four confusion matrix categories (true positives, false negatives, true negatives, and false positives), offering a comprehensive evaluation.

These metrics are calculated using a 10-fold cross-validation procedure to ensure the model's robustness and generalizability, minimize overfitting, and provide a comprehensive evaluation across different data subsets. The final model of our classification task is the one that selects descriptors resulting in (a) high evaluation metrics for the model defined as those exceeding a threshold of 0.85, and (b) accurate prediction of the classes for all 10 high-quality validation datapoints.

It is important to clarify that the entire process, from the application of Spearman Rank Correlation to the 10-fold cross-validation for calculating classification metrics, is conducted within an iterative framework. In each iteration, only one machine learning classification model is activated, and the following two parameters are varied:

a) Spearman Rank Correlation threshold: This threshold determines the minimum correlation value above which correlated features are discarded. It is randomly selected in each iteration of the iterative process, with values ranging from 0.85 to 0.99.
b) XGBoost Feature Importance threshold: This threshold determines the minimum importance value required for descriptors to be retained, based on their importance values calculated by the XGBoost selection algorithm. In each iteration, the threshold is randomly selected from a range of 0.02 to 0.07.

The iterative process terminates when the model achieves high evaluation metrics and accurately predicts 100% of the external high-quality datapoints. Notably, in each iteration, approximately 100 new datapoints per class are generated using the SMOTE algorithm. Different points are created each time to enhance the model's robustness. **Figure 6** schematically presents the entire iterative process involved in refining the molecular descriptors and constructing the final predictive classification model.

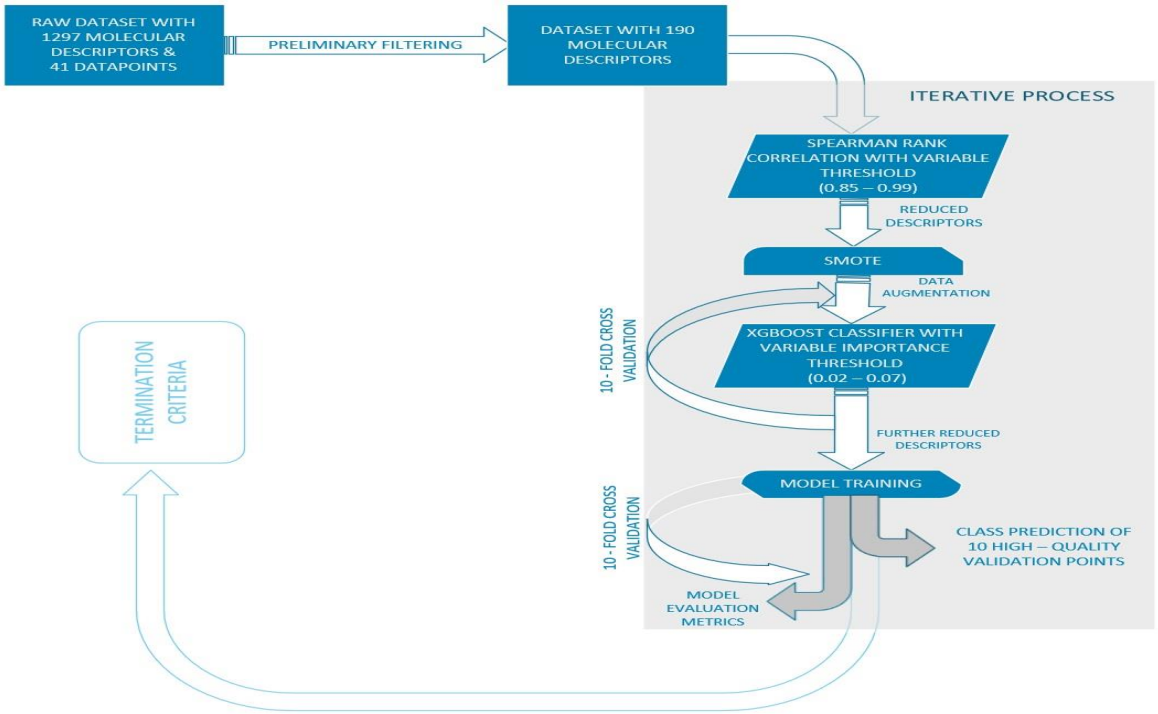

Fig.6 Schematic representation of the iterative process for refining molecular descriptors and constructing the final predictive classification model. *The process includes preliminary filtering, Spearman Rank Correlation with variable threshold, data augmentation using SMOTE, feature selection with XGBoost classifier, and model evaluation through 10-fold cross-validation. The termination criteria are based on achieving high evaluation metrics and accurate prediction of the classes for all 10 high-quality validation datapoints*

Based on the refined iterative process described, the final selected parameters were a Spearman Rank correlation threshold of 0.92 and an XGB feature importance threshold of 0.05. The correlation method reduced the features to 15 descriptors, which were previously detailed in Figure 5. The XGB feature selection technique ultimately selected the following eight features: RF, Sv, qed, NumValenceElectrons, Chi0v, Kappa2, VMcGowan, and logP. These descriptors

not only result in high metrics but also enable the accurate prediction of the 10 unseen validation points. The XGB importance scores of these descriptors were previously presented in Table 2. The evaluation metrics, averaged from the 10-fold cross-validation, are shown in **Table 4**, together with their standard deviations. The predicted classes of the validation datapoints, as previously noted, are [0, 1, 1, 1, 2, 2, 3, 3, 4, 5].

The classification model that ultimately satisfied the termination criteria was the Random Forest Classifier, with its parameters detailed in Table 3. The fact that Random Forest emerged as the best model is not coincidental. Tuulaikhuu et al. (2017) identified Random Forests as the most effective models for predicting chemical toxicity in freshwater ecosystems, due to their ability to manage complex interactions between variables such as chemical properties, environmental conditions, and biological characteristics of fish species. This finding is further corroborated by Hou Ping (2019), who emphasized that tree-based models, particularly Random Forest models, outperform neural networks and linear regression models in predicting ecotoxicity characterization factors because of their superior handling of data heterogeneity and computational efficiency. This superiority was also evident in the present study, where ANN and linear models were tested but not elaborated upon for brevity.

The augmented dataset used as the training dataset, generated via SMOTE and consisting of 559 datapoints, is presented in Supplementary Material 1, under the sheet named **'Classification Training Set.**' This dataset is crucial as it serves as the training set for predicting the CF ECOTOX classes of new unseen inorganic elements. Additionally, Supplementary Material 1 includes the dataset with the 51 elements and the 190 molecular descriptors, which is used as described below.

Table 4 Performance metrics and parameters of the final Random Forest classification model. *This table presents the performance metrics (Accuracy, Precision, Recall, F1 Score, and MCC) of the Random Forest classification model, along with the Spearman Rank correlation threshold, XGBoost feature importance threshold, and the total number of datapoints in the final training set, generated via SMOTE. These metrics demonstrate the model's precision and robustness as computed from the 10-fold cross-validation procedure*

Results	Value	Standard Deviation
Accuracy	0.9810	± 0.0178
Precision	0.9839	± 0.0147
Recall	0.9810	± 0.0178
F1 Score	0.9810	± 0.0177
MCC	0.9777	± 0.0208
Spearman Rank correlation threshold	0.92	-
XGBoost feature importance threshold	0.05	-
Datapoints in the final training set (via SMOTE)	559	-

To utilize the developed classification model for predicting the CF ECOTOX class of any new element, one must first obtain the corresponding SMILES notation for the element and calculate the eight descriptors. Specifically, Sv is calculated using PaDeL; qed, NumValenceElectrons, and Chi0v are calculated using RDKit;

and VMcGowan and logP are calculated using the Mordred and Pybel toolkits, respectively. Subsequently, the new datapoint should be added to the dataset with the 51 datapoints and the 190 molecular descriptors. This is necessary for applying, correctly, the Yeo-Johnson transformation, to normalize the values for the element. Following this, the prepared classification model from this study can be used to predict the CF ECOTOX class. The reason for adding the new element to the aforementioned dataset is that the Yeo-Johnson transformation, being a power transformation, requires at least two existing datapoints to transform the new datapoint accurately.

The selected descriptors are not arbitrary but have physical significance, indicating the combinatorial factors that influence the toxicity of inorganic elements in the aquatic ecosystems. The physical interpretation of these descriptors is detailed in **Table 5** [32-35].

Table 5 Final selected molecular descriptors and their physical interpretation related to the ecotoxicity of inorganic elements. *This table presents the final selected molecular descriptors along with their physical interpretations, highlighting their critical role in influencing the ecotoxicity impact of inorganic elements*

Final selected molecular descriptors	Physical interpretation in relation to inorganic elements' ecotoxicity
RF	Indicates the variability and uncertainty in parameter estimation, essential for assessing the range of potential ecotoxic impacts based on whether the element is non-essential or essential. RF is typically set at 0.1 for non-essential inorganic elements and at 0.01 for essential ones in the EF v.3.1 database.
Sv	Sv indicates the physical size and space an ion occupies, which affects how it interacts with biological membranes and environmental matrices. Larger elemental ions may have different transport and bioavailability profiles, impacting their potential to bioaccumulate and cause toxicity in aquatic ecosystems.
qed	Although originally for drug-likeness, qed provides insight into the bioavailability and reactivity of ions in ecological contexts. A higher qed value correlates with a higher biological activity, influencing the toxicity of these ions in environmental systems.
NumValenceElectrons	The number of valence electrons determines the chemical reactivity and bonding patterns of ions, which are critical for predicting their potential to form harmful compounds. Reactivity influences how these ions interact with biological molecules and environmental chemicals, affecting their ecotoxicity.
Chi0v	Reflects the structural complexity and connectivity of the elemental ions. Higher connectivity often leads to greater interaction with biological systems and environmental matrices, thereby increasing the toxicity of the inorganic ions.
Kappa2	Indicates the geometric configuration and symmetry of ions, affecting their spatial distribution and interaction in biological environments.
VMcGowan	Provides an estimate of the size of ions, which influences their transport, distribution, and persistence in the environment.
logP	Determines the lipophilicity of inorganic ions, impacting their ability to bioaccumulate in organisms and their potential toxicity in aquatic ecosystems.

The detailed procedure for calculating and predicting the CF ECOTOX class of any new, unseen inorganic element using the model developed in this study is provided in Supplementary Material 2 under the title '**Procedure for Predicting the CF ECOTOX Class of Inorganic Elements Using the Developed Classification Model**'. This guide comprehensively outlines all necessary steps to effectively utilize the classification model for predicting the CF ECOTOX class of any inorganic element.

Development of Regression model for precise CF prediction

To accurately determine the CF values for inorganic elements, it is essential to develop a regression model subsequent to the classification model. For each class, it is optimal to develop a separate regression model to achieve greater accuracy in CF value predictions. This approach entails using datapoints from each specific class to train the respective regression models, ensuring that the final constructed models are tailored to the unique characteristics of each class. This class-specific training approach enhances the models' ability to capture the unique variations within each class, ultimately leading to more reliable CF value predictions. However, we face two primary challenges:

- Challenge 1: Limited reliable datapoints

  To ensure precise CF estimation, it is crucial to train the machine learning models exclusively on datapoints with high-quality scores. However, the number of these reliable datapoints is insufficient. While only fifteen out of the 51 elements in the EF v.3.1 database have dependable CF values (see Table 1), a larger dataset is necessary to develop a model that is robust. Therefore, utilizing a data augmentation technique designed specifically for regression tasks is essential to address this limitation and enhance the dataset's comprehensiveness and reliability.

- Challenge 2: Incomplete coverage of the class's value range

  The reliable datapoints within each class fail to encompass the entire range of CF values for that class. This inadequate coverage presents a significant challenge, as the training set for each regression model must represent all potential CF values within the class to ensure comprehensive representation. To address this issue, the Gaussian copula technique was employed. This method leverages its extrapolation capabilities to generate additional target values from the existing high-quality datapoints, ensuring full coverage of the CF value range within each class. By applying this data augmentation technique, we effectively addressed the issues of limited high-quality datapoints and the insufficient coverage of the CF value range, resulting in a robust and generalizable model capable of providing accurate CF predictions for inorganic elements.

Extrapolation with Copula: Given the inadequacy of high-quality datapoints to cover the entire CF range within a class, the copula technique enables the extrapolation of existing points to generate at least two new points at the extremes of the CF range for each class. This ensures that the regression model has the necessary boundary conditions to predict accurately across the entire spectrum of CF values. This methodology is indispensable because conventional data augmentation techniques for regression tasks, such as SMOTER or GMM, inherently lack the capability to extrapolate values beyond the existing CF value range.

Data Augmentation: Once these boundary points are established, the copula technique is applied again to augment the limited datapoints. It is crucial that the distribution of the extrapolated points follows the same distribution as the original high-quality points, thus maintaining the statistical integrity of the dataset. This approach is validated in Supplementary Material 2, which demonstrates that the extrapolated data align with the distribution of the original reliable data within the class.

Overall, the Gaussian copula technique, through both data augmentation and its extrapolation capability, ensures that the six regression models developed are capable of providing reliable and accurate predictions across the full spectrum of CF values. Using this technique, six individual training sets for the six regression models were generated, each comprising 100 datapoints. These training sets, created through the Gaussian copula algorithm, are detailed in Supplementary Material 1 under the sheet named **'Regression full set.'** This structured approach ensured that the training sets developed are both comprehensive and robust, thereby enhancing the generalizability and predictive accuracy of the regression models.

To construct the most capable class-specific regression model for each of the six classes, several machine learning models were evaluated. The models used included XGB Regressor, Random Forest Regressor, Gradient Boosting Regressor, Decision Tree Regressor, and AdaBoost Regressor. The performance of these models was assessed using evaluation metrics appropriate for regression tasks: R-squared (Train), RMSE (Train), R-squared (Test), and RMSE (Test). The model demonstrating the best performance metrics was selected as the final training algorithm. For all six classes, the Decision Tree Regressor exhibited the best performance. The detailed parameter tuning for the Decision Tree Regressor, which is consistent across the six regression models, along with an explanation of these parameters (Pedregosa et al., 2011), is provided in **Table 6**. The training and testing performance metrics for all six regression models were recorded as follows: R-squared (Train) = 1.0, RMSE (Train) = 0.0, R-squared (Test) = 1.0, and RMSE (Test) = 0.0, confirming that the constructed models are highly accurate, as expected.

The comprehensive predictive model for determining $CF_{HC_{20}}$ values, and consequently HC_{20} values, for any inorganic element is detailed in Supplementary Material 2 under the title **'Consolidated code for predicting Element classes and corresponding CF values.'** By following the straightforward steps outlined in the code, users can precisely predict the $CF_{HC_{20}}$ value for any inorganic element of their interest.

Table 6 Detailed parameters of the Decision Tree Regressor model used for training the six predictive models

Parameter	Value	Explanation
criterion	mse	The function to measure the quality of a split. Mean Squared Error (MSE) is used, which measures variance reduction. Typically, options include "mse" and "friedman_mse"

splitter	best	Strategy used to choose the split at each node. The "best" strategy selects the best split. Typically, options include "best" or "random"
max_depth	None	The maximum depth of the tree. If None, nodes are expanded until all leaves are pure or contain fewer than min_samples_split samples. Typically ranges from None (unlimited) to an integer value specifying the maximum depth
min_samples_split	2	The minimum number of samples required to split an internal node. Ensures nodes are split only if they contain at least this many samples. Typically ranges from 2 to a higher integer
min_samples_leaf	1	The minimum number of samples required to be at a leaf node. Typically ranges from 1 to a higher integer
min_weight_fraction_leaf	0.0	The minimum weighted fraction of the input samples required to be at a leaf node. Ensures leaf nodes contain at least this fraction of the overall sample weight. Typically ranges from 0.0 to 0.5
max_features	None	The number of features to consider when looking for the best split. If None, all features are considered. Typically ranges from None to a fraction or integer value of total features
random_state	$> 10^6$	Ensures reproducibility of results, preventing random fluctuations in metrics. Typically set to 42 for consistency
max_leaf_nodes	None	Limits the number of leaf nodes in the tree. If None, leaf nodes are not limited. Typically ranges from None to an integer specifying the maximum number of leaf nodes
min_impurity_decrease	0.0	A node will be split if this split induces a decrease in impurity greater than or equal to this value. Typically ranges from 0.0 to a small positive number

ccp_alpha	0.0	Complexity parameter used for Minimal Cost-Complexity Pruning. Only subtrees with cost complexity smaller than this value are retained. Typically ranges from 0.0 to a small positive number

Analysis of CF and HC_{20} Predictions for 51 inorganic elements

The results regarding the EF v.3.1 database's 51 elements are presented in **Table 7**. This table illustrates the ECOTOX $CF_{HC_{20}}$ values for the 51 elements as predicted by the model developed in this study. For comparison, the ECOTOX $CF_{HC_{20}}$ values assigned by the EF v.3.1 database are also included. When the $CF_{HC_{20}}$ value for a chemical is known, the HC_{20} value can be immediately calculated if the XF and FF are known for that substance. Consequently, for all 51 elements, the predicted HC_{20} values have been calculated, and for comparison, the HC_{20} values assigned to these elements by the EF v.3.1 database are also presented.

To facilitate the interpretation of the model's CF predictions, the Quality Scores for all 51 elements are included. These scores indicate the reliability of the CF values assigned by the EF database, providing a benchmark for assessing the reasonableness of the corresponding predicted CF values, depending on the Quality Score of each element. It should be noted that elements classified as 'Unavailable' lack reliable ecotoxicological data from both EU and other reputable sources, which accounts for the discrepancies observed between the assigned and predicted CF values.

Table 7 Comparative analysis of Predicted and Database CF and HC_{20} values for the 51 Inorganic Elements. *This table illustrates the $CF_{HC_{20}}$ values predicted by the developed model alongside those assigned by the EF v.3.1 database. It also includes their corresponding HC_{20} (mg/L) values calculated from both the model and the EF database, as well as the Quality Scores to facilitate the evaluation of the predicted values' validity*

Name	Quality Score	Predicted CF	Database's CF	Predicted HC_{20}	Database's HC_{20}
bismuth	Only one test	0,9	0,28	1,14E-02	3,64E-02
molybdenum	High	0,98	0,98	3,16E-02	3,16E-02
magnesium	High	1,2	1,2	3,17E-02	3,17E-02
phosphorus	Only one test	6,9	2	1,02E-02	3,50E-02
silicon	Only one test	8,4	4,7	4,54E-02	8,09E-02
titanium	Low	1	4,8	1,59E-02	3,25E-03
boron	High	24	24	1,58E-02	1,58E-02
copper	Unavailable	32	46	1,50E-04	1,04E-04
copper (ii)	High	46	46	1,30E-05	1,30E-05
zirconium	Only one test	20	48	1,87E-02	7,92E-03
manganese	Unavailable	52	52	7,31E-04	7,31E-04
manganese (ii)	Average	52	52	7,31E-04	7,31E-04

lead	Unavailable	68	68	8,82E-05	8,82E-05
lead (ii)	High	68	68	2,65E-05	2,65E-05
tungsten	Low	20	81	1,77E-02	4,44E-03
selenium	Unavailable	20	86	1,41E-03	3,33E-04
selenium (iv)	Unavailable	42	86	1,16E-03	5,70E-04
antimony (iii)	Unavailable	92	140	1,57E-02	1,02E-02
lithium	Low	13,5	150	2,81E-02	2,53E-03
iron (ii)	Unavailable	12	160	1,82E-02	1,35E-03
tellurium	Low	606	500	6,27E-04	7,60E-04
tin	Unavailable	510	510	7,06E-04	7,06E-04
tin (ii)	Average	510	510	3,43E-04	3,43E-04
chromium (iii)	Unavailable	1350	950	1,21E-07	1,71E-07
beryllium (ii)	Unavailable	1350	1600	1,04E-05	8,75E-06
zinc	Unavailable	1700	1700	2,24E-05	2,24E-05
zinc (ii)	High	1700	1700	5,53E-05	5,53E-05
arsenic (iii)	Unavailable	9838	1800	1,45E-04	7,91E-04
barium (ii)	Unavailable	2058	3700	1,33E-03	7,42E-04
cerium(3^{+})	Low	2058	3900	1,75E-04	9,23E-05
iron (iii)	Unavailable	6554	4100	6,10E-08	9,76E-08
thallium (i)	Unavailable	6845	4100	2,95E-04	4,92E-04
aluminium	Unavailable	9838	4400	3,86E-05	8,64E-05
aluminium (iii)	High	4400	4400	9,82E-05	9,82E-05
cesium (i)	Unavailable	9838	4500	2,82E-04	6,16E-04
arsenic (v)	Unavailable	9838	4700	1,45E-04	3,03E-04
vanadium	High	4800	4800	1,82E-05	1,82E-05
cobalt	Unavailable	5500	5500	9,82E-07	9,82E-07
cobalt (ii)	High	5500	5500	2,91E-05	2,91E-05
antimony	Unavailable	55224	11000	3,95E-06	1,99E-05
chromium (vi)	Unavailable	30544	12000	4,93E-05	1,26E-04
strontium (ii)	Unavailable	30544	18000	7,54E-05	1,28E-04
antimony (v)	Unavailable	16038	22000	8,93E-05	6,51E-05
nickel	Unavailable	81828	27000	3,25E-06	9,85E-06
nickel (ii)	High	27000	27000	5,79E-05	5,79E-05

mercury	Unavailable	40039	33000	1,20E-06	1,46E-06
mercury (ii)	High	33000	33000	2,07E-06	2,07E-06
silver	Unavailable	180000	180000	1,19E-07	1,19E-07
silver (i)	High	180000	180000	8,20E-07	8,20E-07
cadmium	Unavailable	670000	670000	5,67E-07	5,67E-07
cadmium (ii)	High	670000	670000	1,97E-06	1,97E-06

Observing Table 7, even for elements with 'Low' Quality Scores, significant discrepancies are noted between the predicted CF values and those assigned by the EF v.3.1 database. For example, in the case of lithium, despite being classified with a 'Low' Quality Score rather than 'Unavailable,' there is a substantial difference of approximately 90% between the model's predicted CF value and the value provided by the EF v.3.1 database, with the relative difference for the HC_{20} value reaching nearly 1000%. Although Slapnik et al. (2015) point out that uncertainties in ecotoxicological EfFs and the corresponding HC_{20} values can span up to 8 orders of magnitude in some cases, this disparity particularly highlights the unreliability of the database's CF and HC_{20} values for elements with low-quality scores. Conversely, for elements with reliable Quality Scores, the predicted CF values from our model match the CF values assigned by the database, as expected; the same consistency is observed with the HC_{20} values.

The discrepancies observed in elements with low-quality scores or those classified as 'Unavailable' suggest that the CF values assigned by the EF v.3.1 database for these elements are subject to significant uncertainty. The rigorously validated model developed in this study offers a more reliable prediction framework, demonstrating its efficacy in providing dependable CF predictions for freshwater ecotoxicity and corresponding HC_{20} values. It should be reiterated that these HC_{20} values are typically derived from time-consuming and costly experiments, highlighting the utility of the developed model.

The following **Table 8** presents the toxicity (in freshwater) rankings of the 51 elements based on their HC_{20} values. Lower HC_{20} values indicate higher toxicity to aquatic life. This table compares the toxicity rankings of the 51 elements according to the HC_{20} values assigned by the EF v.3.1 database with the new toxicity rankings as determined by the model developed in this study.

Table 8 Comparative toxicity rankings of the 51 Elements. *This table presents the toxicity rankings of the 51 elements based on their HC_{20} values, comparing the rankings derived from the EF v.3.1 database with those determined by the model developed in the study. Elements are listed in order of increasing HC_{20} values, where a lower HC_{20} value indicates higher toxicity to aquatic life. The table demonstrates alignment between the model and the database for the most and least toxic elements, as well as notable consistency in the order of magnitude of HC_{20} values for highly and moderately toxic elements. Differences are more pronounced in the low toxicity elements, with some elements showing shifts in ranking of up to nine positions. This variation underscores the necessity for a more reliable method to calculate HC_{20} values, as demonstrated by the new model*

Rank	**The Model**		**Database EF v.3.1**	
	Name	*HC_{20}*	*Name*	*HC_{20}*
1	iron (iii)	6,10E-08	iron (iii)	9,76E-08
2	silver	1,19E-07	silver	1,19E-07

3	chromium (iii)	1,21E-07	chromium (iii)	1,71E-07
4	cadmium	5,67E-07	cadmium	5,67E-07
5	silver (i)	8,20E-07	silver (i)	8,20E-07
6	cobalt	9,82E-07	cobalt	9,82E-07
7	mercury	1,20E-06	mercury	1,46E-06
8	cadmium (ii)	1,97E-06	cadmium (ii)	1,97E-06
9	mercury (ii)	2,07E-06	mercury (ii)	2,07E-06
10	nickel	3,25E-06	beryllium (ii)	8,75E-06
11	antimony	3,95E-06	nickel	9,85E-06
12	beryllium (ii)	1,04E-05	copper (ii)	1,30E-05
13	copper (ii)	1,30E-05	vanadium	1,82E-05
14	vanadium	1,82E-05	antimony	1,99E-05
15	zinc	2,24E-05	zinc	2,24E-05
16	lead (ii)	2,65E-05	lead (ii)	2,65E-05
17	cobalt (ii)	2,91E-05	cobalt (ii)	2,91E-05
18	aluminium	3,86E-05	zinc (ii)	5,53E-05
19	chromium (vi)	4,93E-05	nickel (ii)	5,79E-05
20	zinc (ii)	5,53E-05	antimony (v)	6,51E-05
21	nickel (ii)	5,79E-05	aluminium	8,64E-05
22	strontium (ii)	7,54E-05	lead	8,82E-05
23	lead	8,82E-05	cerium(3+)	9,23E-05
24	antimony (v)	8,93E-05	aluminium (iii)	9,82E-05
25	aluminium (iii)	9,82E-05	copper	1,04E-04
26	arsenic (v)	1,45E-04	chromium (vi)	1,26E-04
27	arsenic (iii)	1,45E-04	strontium (ii)	1,28E-04
28	copper	1,50E-04	arsenic (v)	3,03E-04
29	cerium(3+)	1,75E-04	selenium	3,33E-04
30	cesium (i)	2,82E-04	tin (ii)	3,43E-04
31	thallium (i)	2,95E-04	thallium (i)	4,92E-04
32	tin (ii)	3,43E-04	selenium (iv)	5,70E-04
33	tellurium	6,27E-04	cesium (i)	6,16E-04
34	tin	7,06E-04	tin	7,06E-04
35	manganese	7,31E-04	manganese	7,31E-04

36	manganese (ii)	7,31E-04	manganese (ii)	7,31E-04
37	selenium (iv)	1,16E-03	barium (ii)	7,42E-04
38	barium (ii)	1,33E-03	tellurium	7,60E-04
39	selenium	1,41E-03	arsenic (iii)	7,91E-04
40	phosphorus	1,02E-02	iron (ii)	1,35E-03
41	bismuth	1,14E-02	lithium	2,53E-03
42	antimony (iii)	1,57E-02	titanium	3,25E-03
43	boron	1,58E-02	tungsten	4,44E-03
44	titanium	1,59E-02	zirconium	7,92E-03
45	tungsten	1,77E-02	antimony (iii)	1,02E-02
46	iron (ii)	1,82E-02	boron	1,58E-02
47	zirconium	1,87E-02	molybdenum	3,16E-02
48	lithium	2,81E-02	magnesium	3,17E-02
49	molybdenum	3,16E-02	phosphorus	3,50E-02
50	magnesium	3,17E-02	bismuth	3,64E-02
51	silicon	4,54E-02	silicon	8,09E-02

From the Table 8, we observe that in both cases - whether based on the rankings made by the model developed in this study or those from the EF v.3.1 database - the elements at the extremes of the toxicity rankings, corresponding to the highest and lowest toxicity, are the same. The order of magnitude remains consistent, with 10^{-8} for the most toxic element and 10^{-2} for the least toxic element in both taxonomic frameworks. This indicates alignment between the model and the database in identifying the most and least toxic elements.

The primary similarities between the model developed and the EF v.3.1 database lie in the rankings of highly toxic elements. Elements such as iron (III), silver, and chromium (III), despite being classified as “Unavailable” in terms of Quality Score, are consistently ranked among the top positions for toxicity in both taxonomic frameworks. This indicates a strong agreement in their toxicity assessments. Additionally, there is notable consistency in the order of magnitude of the HC_{20} values in both the database and the study’s model for elements with high and moderate toxicity.

The greatest differences are observed in the low toxicity elements, with rankings assigned by the database shifting by up to nine positions compared to the model developed in this study; for instance, phosphorus is one such element showing a significant shift. Additionally, the HC_{20} values for the low toxicity elements differ in order of magnitude, being 10^{-3} in the database and 10^{-2} in the model. When we refer to low toxicity elements, we mean those whose CF values are classified within a maximum of class 1. This significant variation underscores the need for a more reliable method to precisely calculate HC_{20} values, as demonstrated by the model developed. Despite the model's generalizability and accuracy in predicting both the HC_{20} and CF values of new elements, *continuous refinement with validated and unquestionable data* will further enhance its reliability.

To provide a visual representation of these comparisons, we include **Figure 7**, which depicts the relationship between the log-transformed HC_{20} values as predicted by our model and those reported in the EF v.3.1 database. This scatter plot serves to illustrate the strong concordance for elements with high and moderate toxicity, as well as the pronounced discrepancies for those with lower quality scores. The alignment observed for the most toxic elements reaffirms the reliability of our model's predictions. In contrast, the variations in the less toxic elements highlight the inherent challenges and limitations of the methodologies currently employed by the EF v.3.1 database. These discrepancies underscore the need for more advanced, and reliable approaches to accurately calculate HC_{20} values, as demonstrated by the developed model.

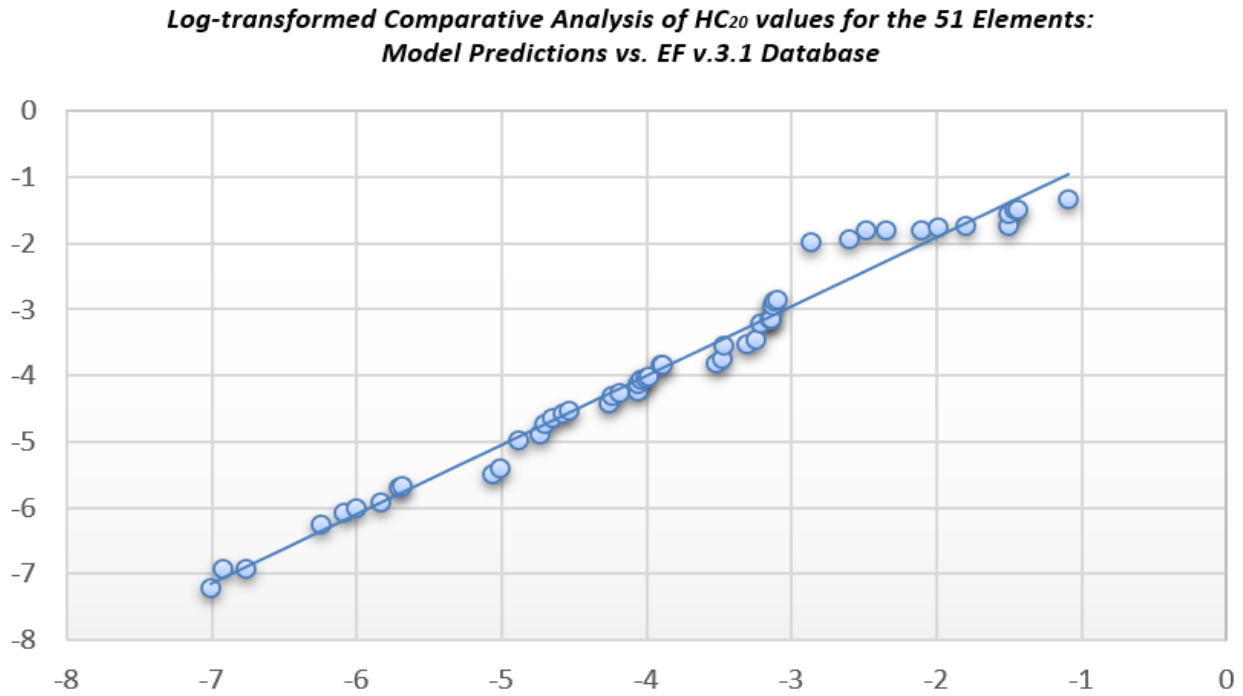

Fig.7 Log-transformed Comparative Analysis of HC_{20} values for the 51 Elements. *This figure presents a scatter plot comparing the log-transformed HC_{20} values for the 51 elements as predicted by the study's developed model against those reported in the EF v.3.1 database. Each point represents an element, with the x-axis indicating the HC_{20} values from the database and the y-axis showing the corresponding values predicted by the model. The plot highlights the alignment and discrepancies between the two sets of toxicity rankings, emphasizing the model's accuracy in predicting HC_{20} values for elements with high and moderate toxicity. Nevertheless, the analysis reveals significant discrepancies for elements with low-quality scores. These discrepancies highlight the unreliability of the database's HC_{20} values for such elements and underscore the need for a more robust method to calculate HC_{20} values, as provided by the developed model of this study*

4 Conclusions

This study represents a significant advancement in ecotoxicological assessment by integrating machine learning techniques with molecular descriptors to predict CFs for the freshwater ecotoxicity (ECOTOX) impact category and HC_{20} values for inorganic elements within LCA. By successfully bridging the gap between cheminformatics and environmental impact prediction, our research employs sophisticated computational models to enhance the precision and reliability of these estimations. This integration provides a more accurate and comprehensive framework for assessing the environmental impacts of inorganic elements, addressing the current limitations of existing methodologies that rely on unreliable data.

The methodology involves the computation of 1294 molecular descriptors, followed by normalization using the Yeo-Johnson transformation. Initial feature reduction was achieved through Spearman Rank Correlation, and further refinement was conducted using XGB feature importance. Data augmentation was performed with the SMOTE algorithm, leading to the application of a Random Forest classifier for classification and a Decision Tree regressor for precise CF value predictions. This multi-stage approach effectively addresses the complexities and variabilities inherent in ecotoxicological data, providing a robust framework for predictive modeling. Our results reveal significant discrepancies between the predicted CF values and those assigned by the EF v.3.1 database for elements with low-quality scores, highlighting the limitations of existing methodologies that depend on incomplete or uncertain experimental data.

The developed model has demonstrated superior accuracy in predicting CF values, validated against high-quality datapoints. This not only fills critical gaps in existing environmental databases but also offers a computationally efficient and scientifically grounded alternative for ecotoxicity assessment. Key molecular descriptors identified, such as Sv, NumValenceElectrons, logP, encapsulate essential physicochemical properties that significantly influence the environmental behavior and toxicological impacts of inorganic elements. The strategic incorporation of these descriptors into our predictive model enhances both the interpretability and reliability of the results, offering deeper insights into the mechanisms underlying elemental ecotoxicity.

The implications of this study are extensive and multifaceted. By integrating advanced machine learning techniques with cheminformatics, we have enhanced existing methodologies in the international literature, supporting broader environmental sustainability and resource management goals [36-41]. The developed model is not only applicable to existing database elements but also adaptable for future inclusions, promoting the continuous enhancement of environmental impact assessments. This adaptability ensures that our approach remains relevant in the dynamic field of ecotoxicological assessment, facilitating more informed and efficient decision-making processes. The ability to integrate new data seamlessly allows for ongoing improvement and expansion of ecotoxicological databases, ensuring they remain comprehensive and up-to-date, ultimately contributing to more effective environmental protection strategies.

Future research should prioritize expanding the dataset to include a broader spectrum of inorganic elements and refining the model with additional high-quality experimental data to enhance its robustness and applicability. Additionally, extending this methodology to encompass both inorganic molecules and organometallic compounds will significantly broaden the scope of predictive ecotoxicology. This study advances LCA methodologies by advocating for the use of advanced computational techniques to address pressing environmental challenges, ultimately contributing to sustainable environmental stewardship.

Statements and Declarations

Author information

Konstantinos M. Kritsotakis

Affiliations

School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechneiou st. Zografou Campus, 15780, Athens, Greece

Contributions

The author solely contributed to the conception, design, implementation, and writing of the manuscript.

Corresponding Author

Correspondence to: **Konstantinos M. Kritsotakis (kostaskritsos@gmail.com)**

References

1. Song Runsheng (2019) Machine learning for addressing data deficiencies in life cycle assessment, Dissertation, University of California Santa Barbara
2. Rosenbaum Ralph K. et al (2008) USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment, The International Journal of Life Cycle Assessment 13, 532-546
3. Servien Rémi et al (2022) Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater, Peer Community Journal, 2
4. Hou Ping et al (2020) Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models, Environment International, 135, 105393
5. Hamadache Mabrouk et al (2018) QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (Apis mellifera L.), Environmental Science and Pollution Research, 25, 896-907
6. Chen Xingmei et al (2020) Machine learning-based prediction of toxicity of organic compounds towards fathead minnow, RSC advances 10, 59, 36174-36180
7. Ding Weizhe et al (2022) Combining multi-dimensional molecular fingerprints to predict the hERG cardiotoxicity of compounds, Computers in Biology and Medicine, 144, 105390
8. Liangxu X. et al (2020) Improvement of prediction performance with conjoint molecular fingerprint in deep learning, Frontiers in pharmacology, 11, 606668
9. Ucak Umit V. et al (2023) Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization, Journal of Cheminformatics 15, 1, 55
10. Saouter Erwan et al (2018) Environmental Footprint: Update of Life Cycle Impact Assessment methods–Ecotoxicity freshwater, human toxicity cancer, and non-cancer, European Union, Luxembourg
11. Andreasi Bassi S. et al (2023) Updated characterisation and normalisation factors for the Environmental Footprint 3.1 method, Technical Report by European Joint Research Centre

12. National Center for Biotechnology Information (NCBI), PubChem Database (2024). https://pubchem.ncbi.nlm.nih.gov/
13. Royal Society of Chemistry, ChemSpider (2024). https://www.chemspider.com/
14. Sala Serenella et al (2022) Toxicity impacts in the environmental footprint method: calculation principles, The International Journal of Life Cycle Assessment 27, 4, 587-602
15. European Platform on LCA, EPLCA (2024). https://eplca.jrc.ec.europa.eu/ecotox.html
16. Kang Myeongsu and Jing Tian (2018) Machine Learning: Data Pre-processing, Prognostics and health management of electronics: fundamentals, machine learning, and the internet of things, 111-130
17. Sun Ye et al (2022) Improved machine learning models by data processing for predicting life-cycle environmental impacts of chemicals, Environmental Science & Technology 57, 8, 3434-3444
18. Singh Dalwinder and Birmohan Singh (2020) Investigating the impact of data normalization on classification performance, Applied Soft Computing 97, 105524
19. Pan Pengfei et al (2023) Predicting punching shear in RC interior flat slabs with steel and FRP reinforcements using Box-Cox and Yeo-Johnson transformations, Case Studies in Construction Materials 19, 2409
20. Xiao Chengwei et al (2016) Using Spearman's correlation coefficients for exploratory data analysis on big dataset, Concurrency and Computation: Practice and Experience 28, 14, 3866-3878
21. Ramraj Santhanam et al (2016) Experimenting XGBoost algorithm for prediction and classification of different datasets, International Journal of Control Theory and Applications 9, 40, 651-662
22. Darst Burcu F. et al (2018) Using recursive feature elimination in random forest to account for correlated variables in high dimensional data, BMC genetics 19, 1-6
23. Genovese Christopher R. et al (2012) A comparison of the lasso and marginal regression, The Journal of Machine Learning Research 13, 1, 2107-2143
24. Liu Huawen et al (2009) Feature selection with dynamic mutual information, Pattern Recognition 42, 7, 1330-1339
25. Chang Chih-Chi et al (2022) Melanoma detection using XGB classifier combined with feature extraction and K-means SMOTE techniques, Diagnostics 12, 7, 1747
26. V.F. Rodriguez-Galiano et al (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS journal of photogrammetry and remote sensing 67, 93-104
27. Cao L. J. et al (2003) A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, Neurocomputing 55, 1-2, 321-336
28. Wang Yasi et al (2016) Auto-encoder based dimensionality reduction, Neurocomputing 184, 232-242
29. Madan Spandan et al (2022) When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations, Nature Machine Intelligence 4, 2, 146-153
30. Gerassis Saki et al (2021) AI approaches to environmental impact assessments (EIAs) in the mining and metals sector using AutoML and Bayesian modeling, Applied Sciences 11, 17, 7914
31. Choulakian Vartan et al (1994) Cramér-von Mises statistics for discrete distributions, The Canadian Journal of Statistics 125-137
32. Landrum Greg (2013) Rdkit documentation, Release 1, 4, 1-79
33. Consonni Viviana and Roberto Todeschini (2010) Molecular descriptors, Recent advances in QSAR studies: methods and applications, 29-102
34. Yap Chun Wei (2011) PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints, Journal of computational chemistry 32, 7, 1466-1474
35. O'Boyle Noel M. et al (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, Chemistry Central Journal 2, 1-7
36. Kleinekorte Johanna et al (2019) Combining process short cuts and artificial neural networks for predictive life cycle assessment of chemicals, Foundations of Process Analytics and Machine learning (FOPAM 2019)

37. Galal M Abdella et al (2020) Sustainability assessment and modeling based on supervised machine learning techniques: The case for food consumption
38. José Oduque de Jesus et al (2021) Integration of artificial intelligence and life cycle assessment methods, IOP Conference Series: Materials Science and Engineering, 1196, 1999
39. Koyamparambath Anish et al (2022) Implementing artificial intelligence techniques to predict environmental impacts: case of construction products, Sustainability 14, 6, 3699
40. Tuulaikhuu Baigal-Amar et al (2017) Examining predictors of chemical toxicity in freshwater fish using the random forest technique, Environmental Science and Pollution Research 24, 10172-10181
41. Marvuglia Antonino et al (2015) Machine learning for toxicity characterization of organic chemical emissions using USEtox database: Learning the structure of the input space, Environment International 83, 72-85
42. Hou Ping (2019) Data-Driven Environmental System Analysis: Addressing Data Gaps in Life Cycle Assessment, Dissertation, The University of Michigan
43. Kritsotakis Konstantinos et al (2022) Life Cycle Assessment (LCA) upon the production chain of a powder containing modified olive leaves' extract, Biomass Conversion and Biorefinery 12, 10, 4503-4518
44. Zhu Xinzhe et al (2020) Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes, ACS Sustainable Chemistry & Engineering 8, 30, 11141-11151
45. Slapnik Matej et al (2015) Extending life cycle assessment normalization factors and use of machine learning - a Slovenian case study, Ecological indicators 50, 161-172
46. Li Fuxing et al (2017) In silico prediction of pesticide aquatic toxicity with chemical category approaches, Toxicology research 6, 6, 831-842
47. Pedregosa et al (2011) Scikit-learn: Machine learning in Python. Retrieved from https://scikit-learn.org/stable/
48. Wang Wei et al (2014) Generalized autoencoder: A neural network framework for dimensionality reduction, Proceedings of the IEEE conference on computer vision and pattern recognition workshops
49. Silva Diogo et al (2020) Copula-based data augmentation on a deep learning architecture for cardiac sensor fusion, IEEE Journal of Biomedical and Health Informatics 25, 7, 2521-2532
50. Douzas Georgios et al (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, Information sciences 465, 1-20
51. Lee Taejun et al (2020) Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI, 8th International Winter Conference on Brain-Computer Interface (BCI), IEEE
52. Wong Sebastien C. et al (2016) Understanding data augmentation for classification: when to warp? International conference on digital image computing: techniques and applications (DICTA), IEEE
53. Strelcenia Emilija and Simant Prakoonwit (2023) Improving classification performance in credit card fraud detection by using new data augmentation, AI 4, 1, 172-198
54. Arora Ashish et al (2021) Data augmentation using Gaussian mixture model on CSV files, Distributed Computing and Artificial Intelligence, 17th International Conference Springer International Publishing
55. Lanzante John R. (2021) Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests, International Journal of Climatology 41, 14, 6314-6323
56. Özmen Tamer (1993) A modified Anderson-Darling goodness-of-fit test for the gamma distribution with unknown scale and location parameter, Dissertation, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio