

Data-Driven Environmental System Analysis: Addressing Data Gaps in Life Cycle Assessment

by

Ping Hou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environment and Sustainability and Scientific Computing)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Ming Xu, Chair
Professor Olivier Jolliet
Associate Professor Shelie Miller
Professor Ji Zhu

Ping Hou

pinghou@umich.edu

ORCID ID: 0000-0003-1001-3107

© Ping Hou 2019

Dedication

In loving memory of my grandmother, Guiying Li (1929-2015)

Acknowledgements

Undertaking this Ph.D. has been a truly life-changing experience for me, which was not possible without the support and guidance that I received from many people.

First of all, I would like to express my sincere gratitude to my advisor Professor Ming Xu for his guidance, patience, and support. He is creative in tailoring projects for me to work on when I started my program. He also gave me valuable feedback and advice on projects that I grow interested in over time. Professor Xu taught me the fundamentals of scientific research and reviewed my work extensively. Under his supervision, I learned to pay attention to details and convey ideas clearly in writing and oral presentations.

I gratefully acknowledge the rest of my Ph.D. committee members for their priceless help and feedback on my research. I am indebted to Professor Oliver Jolliet. His enthusiasm for research and teaching motivated me to work hard and joyfully. I thank him for introducing me to his lab, where I received constructive feedback and enjoyed stimulating discussions. I appreciate Professor Shelie Miller for her kindness and support. She is a role model who inspires me to conduct research with positive societal impacts. I also thank Professor Ji Zhu. His expertise in statistical modeling and data insights greatly improved my work and enriched my final dissertation. I extend my appreciation to Dr. Dali Wang at Oak Ridge National Lab, who guided me on high-performance computing applied in my thesis. It was a privilege and pleasure to work with all of them. I also thank Gregor Wernet at the ecoinvent for providing data for this research.

I will always treasure the mentors and friends in Professor Ming Xu's lab, in our Ph.D. community, and at the Center for Sustainable Systems. Their mentorship and friendship are indispensable for my Ph.D. life. I have also received tremendous help from the English Language Institute, Sweetland Writing Center, and CSCAR (Consulting for Statistics, Computing & Analytics Research). Their classes and workshops greatly improved my writing and coding skills.

During my Ph.D. years, I am much blessed to have shared my life with many Christian brothers and sisters at the Olive Tree Campus Church. It is like a home away from home to me. Whenever I feel stressed and overwhelmed, the church family always gives me peace and strengthens my faith.

I deeply appreciate my family for their sacrificing love and support. I dedicate my dissertation to the memory of my grandmother, Guiying Li, whose role in my life was, and remains, immense. I want to thank my mom for her support and continuous prayers for me. To my dear friend and husband Jun Guo, I thank him for his love, encouragement, and company. We met and got married during this Ph.D., and he made them the best years of my life.

Last and above all, I give all thanks to my Lord Jesus Christ, in Him I found meaning and purpose of life. *You make known to me the path of life; you will fill me with joy in your presence, with eternal pleasures at your right hand.* – *Psalms 16:11*, Praise the Lord.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	viii
List of Figures	x
List of Appendices	xii
Abstract	xiii
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Literature Review	3
1.2.1 Estimate Unit Process Data in Life Cycle Inventory	4
1.2.2 Estimate Characterization Factors in Life Cycle Impact Assessment	5
1.3 Research Questions	7
1.4 Structure of the Dissertation	8
Chapter 2 Estimate Unit Process Data Using Similarity-based Link Prediction	10
2.1 Introduction	10
2.2 Data and Methods	10
2.2.1 Unit Process Data Structure	10
2.2.2 Link Prediction and Similarity-Based Link Prediction	12

2.2.3 Steps to Develop the Link Prediction Model	13
2.2.4 Data	16
2.3 Results and Discussion	18
2.3.1 Similarity of Processes	18
2.3.2 Estimation Performance	21
2.3.3 Computational Time	26
2.3.4 Theoretical Grounds	27
2.3.5 Case Study	27
2.3.6 Implications for LCA	28
2.3.7 Future Work	29
2.4 Summary	30
Chapter 3 Estimate Ecotoxicity Characterization Factors for Chemicals Using Neural Network	
Models	31
3.1 Introduction	31
3.2 Data and Methods	32
3.2.1 Data Filtering and Exploratory Analysis	34
3.2.2 Neural Network and Generic Algorithm	35
3.2.3 Steps to Develop the Neural Network Model Using a Genetic Algorithm	38
3.2.4 Model Performance Comparison	41
3.2.5 Variable Importance	42
3.3 Results and Discussion	43
3.3.1 Data Filtering and Exploratory Analysis	43
3.3.2 Model Performance	48
3.3.3 Evolution of the Parameters	51
3.3.4 Performance Comparison with Grid Search	53
3.3.5 Performance Comparison with the ECOSAR Model and Linear Regression Models	55
3.3.6 Variable Importance via Shuffling Procedure	58
3.3.7 Computational Time	61
3.3.8 Drawbacks and Advantages of Neural Networks	62
3.4 Summary	65

Chapter 4 Estimate Ecotoxicity Characterization Factors for Chemicals Using Random Forest

Models	66
4.1 Introduction	66
4.2 Data and Methods	68
4.2.1 Data	68
4.2.2 Random Forests	68
4.2.3 Steps to Develop the Random Forest Models	70
4.2.4 Variable Importance	71
4.2.5 Uncertainty and Application Domain of Estimated HC ₅₀	71
4.3 Results and Discussion	72
4.3.1 Model Selection Results	72
4.3.2 Performance of Random Forest Models	74
4.3.3 Variable Importance via Random Forests	76
4.3.4 Computational Time	80
4.3.5 Drawbacks and Advantages of Random Forests	81
4.3.6 Estimation of Missing HC ₅₀ and CF _{eco} in USEtox	82
4.3.7 Implications for LCA	83
4.3.8 Future Work	84
4.4 Summary	85
Chapter 5 Conclusions	86
Appendices	89
References	127

List of Tables

Table 2-1. The ten most similar processes for “machine operation, diesel, <18.64 kW, underground mining” (the index number is 2,429 in Figure 2-3).....	20
Table 2-2. MPEs with different percentages of data missing.	25
Table 2-3. Computational time required for completing the link prediction estimation.	26
Table 3-1. Physical-chemical properties in USEtox.	34
Table 3-2. Commonly used activation functions.	37
Table 3-3. Parameters and options in genomes.	38
Table 3-4. Correlation coefficients of input and output variables in dataset1.	46
Table 3-5. Best genomes selected by genetic algorithm for dataset1 and their performance.	50
Table 3-6. Best genomes selected by genetic algorithm for dataset2 and their performance.	50
Table 3-7. Best genomes selected by genetic algorithm and grid search for dataset1.....	53
Table 3-8. Best genomes selected by genetic algorithm and grid search for dataset2.....	54
Table 3-9. Remove multicollinearity by calculating variance inflation factor (VIF).	56
Table 3-10. Performance comparison of neural network models and linear regression models. .	57
Table 4-1. Performance comparison of random forest models and neural network models.	75
Table 4-2. Important features in dataset2 identified by random forest models.	78
Table 4-3. Computational time comparison of random forest models and neural network models.	80

Table A-1. Average Mean Percentage Error (MPE) when missing 1% data calculated using different distance functions91

Table B-1. Average Mean Percentage Error (MPE) using three different normalization strategies.....95

Table C-1. Estimation of missing HC₅₀.....96

Table C-2. Ecotoxicity characterization factors (Midpoint, [PAF.m³.day/kg emitted]) calculated based on the estimated HC₅₀. 111

List of Figures

Figure 1-1. Data in LCA.	2
Figure 2-1. Data structure of a unit process database.	12
Figure 2-2. Methodological framework of similarity-based link prediction applied to estimating unit process data.....	14
Figure 2-3. Heat map of the similarity matrix for the ecoinvent 3.1 UPR dataset.	20
Figure 2-4. MPEs with respect to percentage of data missing, k (the number of most similar processes), and q (the parameter in the distance function).	23
Figure 2-5. The distribution of MPEs with respect to the percentage of data missing and value of q (the parameter in the distance function).....	24
Figure 2-6. Histograms of MPEs when different percentages of data are missing with the best q	25
Figure 3-1. A neural network model with one hidden layer.	36
Figure 3-2. Steps of developing the neural network model using a genetic algorithm.	39
Figure 3-3. Data filtering.	44
Figure 3-4. Pair plot of log-transformed input and output variables in dataset1.	45
Figure 3-5. Correlation of variables in dataset2.....	47
Figure 3-6. Empirical cumulative distribution function for dataset1 and dataset2.	48
Figure 3-7. Model performance of the best genomes selected by genetic algorithm.	49
Figure 3-8. Evolution of the four parameters in neural network models.....	52

Figure 3-9. Fitness (i.e., validation MSE) of the best model along eight generations.....	53
Figure 3-10. Performance of the linear regression models.....	57
Figure 3-11. Performance of the neural network models compared with other models.....	58
Figure 3-12. Test MSE when each variable is shuffled.....	59
Figure 3-13. Regularization parameters for preventing overfitting.....	64
Figure 4-1. Procedure of constructing a tree in random forests.....	69
Figure 4-2. Average OOB error rate with different <i>n_estimators</i> and <i>max_features</i>	73
Figure 4-3. Performance of the random forest models on the two datasets.....	75
Figure 4-4. Performance of the random forest models compared with other models.....	76
Figure 4-5. Feature importance by the random forest model for dataset1.....	77
Figure 4-6. Feature importance by the random forest model for dataset2.....	78
Figure 4-7. Visualization of the application domain of the models developed based on dataset1.	83

List of Appendices

Appendix A. Distance measuring methods tested	89
Appendix B. Normalization of the unit process database	93
Appendix C. Estimation of missing HC_{50} and CF_{eco}	96

Abstract

Life Cycle Assessment (LCA) is a widely used analytical tool in environmental system analysis. LCA examines the environmental impacts of a product along its whole life cycle, including raw materials extraction, manufacturing, transport, use, and disposal. LCA studies are data-intensive, requiring two types of data. Unit process data are first used to calculate life cycle consumptions and emissions of a product system. And then characterization factors are used to convert the consumptions and emissions to their potential damage on the ecosystem and human health. Traditional ways to collect the two types of data involve on-site investigation of manufacturing processes and laboratory tests, which are time-consuming and expensive. Therefore, many data in LCA are missing, which generate data gaps and make LCA unable to support decision making effectively.

In this research, taking advantage of existing already collected empirical data, I propose three data-driven frameworks to estimate the missing data in LCA. For the unit process data, I develop a link prediction method based on the ecoinvent database. The results show that on average missing data can be accurately estimated when less than 5% data are missing in one process. For the characterization factors, I first develop neural network models based on existing data in USEtox. The results show that the neural network models outperform a traditional quantitative structure-activity relationship (QSAR) model and linear regression models. Also based on USEtox data, I develop random forest models. The results show random forest models outperform neural network models both in prediction accuracy and computational time. Using

the validated random forest model, I provide estimated missing ecotoxicity characterization factors for LCA practitioners to use.

In summary, I use data-driven approaches to explore the underlying patterns of LCA data and reveal the interrelationship between manufacturing processes and the environment and between properties of contaminants and their hazard impacts. Correctly extracting the patterns behind LCA data helps estimate the missing data without relying on the time-consuming, expensive empirical data collection. The developed data-driven computational approaches will significantly reduce the cost of and save time for LCA studies, therefore help broaden the applications of LCA for sustainability decision making.

Chapter 1 Introduction

1.1 Overview

Life cycle assessment (LCA) measures the environmental impacts of a product in its whole life cycle, including resource extraction, raw materials processing, manufacturing, transport, use, and disposal.¹ By covering all stages of a product life cycle and a wide range of environmental impacts, LCA can help guide policy and technology development to avoid environmental burdens shifting among different life cycle stages. For example, with the rapid economic growth, the world's increasing demand for electronics has dramatically increased the production of electronic products, which causes considerable environmental pressures threatening public health and social development. However, at the end of the life cycle, a large amount of e-waste has also posed environmental and health risks by releasing toxic chemicals, poisoning people, land, air, and water. Such hidden environmental burdens in products can be identified through LCA studies by investigating all stages of a product's life cycle and a wide range of environmental impacts. LCA-guided policy and technology can therefore prevent the hidden burdens of the products. For instance, Apple has been regularly conducting LCA studies for its products, which have identified greenhouse gas emissions from manufacturing as the major contributor to climate change impact. These studies have directed Apple to develop strategies to help their suppliers reduce those emissions. Increasingly, LCA has become an important tool in environmental policy and voluntary actions around the world, supporting decision-making towards sustainability.^{2, 3}

In an LCA study, life cycle inventory (LCI) and life cycle impact assessment (LCIA) are two important quantitative steps (**Figure 1-1**). An LCA model of a product is an ensemble of interconnected unit processes and system processes, respectively represented by foreground and background data. Foreground data quantify intermediate flows (i.e., materials/energy transmitted between unit processes) and elementary flows (i.e., resource from the environment and emission/waste released to the environment) associated with each unit process. Background data are aggregated life cycle inventory (LCI), normally provided by LCI database as system processes, which only contain elementary flows since all intermediate flows are traced back to resource extraction. Based on the product's LCA model, given a functional unit (e.g., produce 1 kg particular product), we can calculate the aggregated LCI of the product. The LCI results multiplying with the corresponding characterization factors (i.e., the relative impact of LCI) are the characterized LCA results, which are the final results of an LCA study.

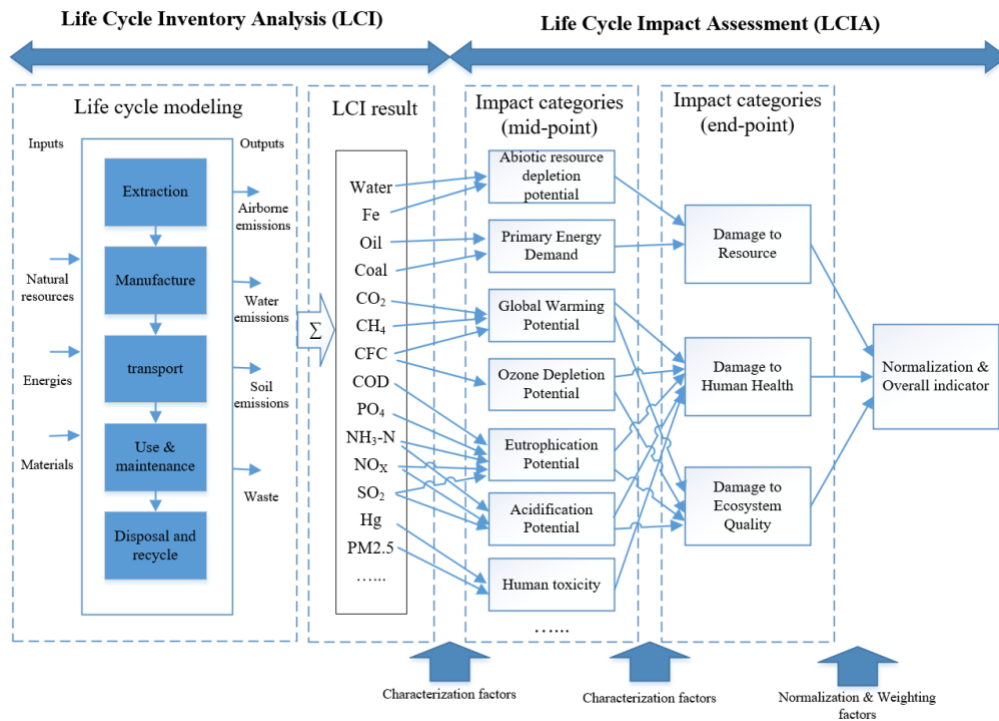


Figure 1-1. Data in LCA.

A good LCA study relies on the availability, quality, and completeness of unit process data in LCI and characterization factors in LCIA. The traditional ways to compile a high-quality unit process data involve collecting empirical data. However, this approach can be time-consuming and expensive because it often requires onsite investigation of manufacturing processes. Additionally, companies often treat their manufacturing data as confidential business information and are unwilling to make it public. In the case of emerging technologies, the manufacturing data are still unavailable at the early design stage, even though evaluating their environmental impacts beforehand is critical to prevent unexpected consequences in the future. All the above reasons cause data gaps in the LCI phase. LCIA phase are also affected by the missing data problem. Laboratory tests are often required to measure impacts of specific wastes or emissions on the ecosystem and human health. However, this approach is not only time-consuming and expensive, but is also impossible to cover all contaminants, especially since new chemicals are developed every day. This means that critical information is missing to understand the potential harm of the chemicals, highlighting the significance of developing a cost-effective and fast way to evaluate their impacts, even before they are produced and contaminate the environment. When the LCI and LCIA data are incomplete or inaccurate, LCA results are embedded with more uncertainty. Therefore, LCA studies based on such incomplete data may lack credibility and may lead to incorrect sustainable decision-making, highlighting the significance of developing a cost-effective and fast way to compile LCI and LCIA data.

1.2 Literature Review

There are increasing efforts to use computational models to estimate LCA data, including unit process data, characterization factors, and characterized LCA results. Most of the studies focus on estimating characterized LCA results for various products, such as chemicals,^{4,5}

consumer products,⁶⁻¹¹ buildings,¹² and others.¹³⁻¹⁵ For example, Wernet *et al.* developed molecular-structure-based models that use molecular features of chemicals as input to estimate chemicals' characterized LCA results using neural networks.^{4, 16} Song *et al.* also used neural networks to estimate LCA results for chemicals, but with more layers in the neural networks.⁵ In addition, given that electronic products are often designed as a bill of materials (BOM) to record the use of raw materials, components, and assemblies, some have proposed automating processes to assign impact factors to inventory components in the BOM to help the sustainable design of electronic products.¹⁷⁻¹⁹ Similarly, for buildings, studies have characterized the relationship between material uses and their environmental impacts to create new designs and optimize the energy use of buildings.^{20, 21} However, characterized LCA results are integrated results from unit process data and characterization factors; thus, they have a higher degree of freedom and involve more uncertainty. Therefore, directly estimating unit process data and characterization factors can be more effective.

1.2.1 Estimate Unit Process Data in Life Cycle Inventory

Collecting unit process data is fundamental to LCA study, but it is time-consuming, and needs large amounts of data, such as raw material inputs, energy use, the ratio of the main products to by-products, production rates, and releases of emissions and waste. LCA practitioners often need to collect foreground and background data from a variety of sources, including direct reports from operations (e.g., meter readings, operation logs/journals), publications, government statistics, and LCA databases. In particular, LCA databases that provide LCA data for common processes are often the major sources for LCA data.²² While convenient for LCA practitioners to use, LCA databases still largely rely on collecting empirical data from various sources, requiring a significant investment of human and capital resources.^{2, 3}

In response to the difficulties facing collecting unit process data, there have been efforts to computationally estimate unit process data instead of relying on empirical data. Among them, neural network models are the mostly used methods. For example, Piao *et al.* use a neural network model to estimate electric power consumptions of wastewater plants.²³ Chiang *et al.* use one to estimate the amount of hazardous chemicals produced in electronic products' life cycle.²⁴ Yin *et al.* estimate CO₂ emissions from power generation also use a neural network model.²⁵ These studies focus on some particular consumptions or emissions in certain industries, lack of broad applicability. Suh and Huppel developed the Missing Inventory Estimation Tool (MIET) using extended input-output analysis. MIET is based on the national accounting system, thus includes the entire national economy; but the application is limited due to the coarse resolution of the input-output tables and product prices are required to convert monetary units to physical units.²⁷

All the previous studies still rely on collecting large amounts of information, such as chemical reaction equations, process characteristics of chemicals, and design details of electronic products and buildings. Therefore, these works only apply to specific products and rely on extensive domain knowledge. There is still a need to develop a convenient, computational approach to estimate unit process data with broader applications to a wide range of products.

1.2.2 Estimate Characterization Factors in Life Cycle Impact Assessment

In LCA, the step of quantifying the impacts of chemicals and other pollutants is called life cycle impact assessment (LCIA). Current practice of LCIA is significantly constrained by limited information on environmental impacts of chemicals. For example, USEtox is the most widely used LCIA method for toxicity, which provides characterization factors (i.e., impact of per unit chemical) for human and ecotoxicology impacts of chemicals.^{28, 29} Despite its wide

applications in LCA, USEtox only offers characterization factors for approximately 3,000 chemicals. As a comparison, the U.S. Environmental Protection Agency (EPA) has more than 85,000 chemicals listed under the Toxic Substances Control Act (TSCA).³⁰ Even for the limited number of chemicals covered in USEtox, 19% and 67% of them miss ecotoxicity characterization factors and human toxicity characterization factors, respectively.

The lack of ecotoxicity characterization factors is essentially due to missing effect factors (EFs). EFs are an integral part of characterization factors that express the ability of chemicals to cause toxic effects to the exposed species in the ecosystems, which are generally obtained from laboratory tests. Since laboratory tests are time-consuming and expensive, EFs of many chemicals are currently unknown, thus their characterization factors.

There are many efforts to estimate toxicity information of chemicals when laboratory experimental data are not available, such as widely used quantitative structure-activity relationship (QSAR) models. QSAR refers to a broad area of inquiry on the relationship between chemical structures and biological activities of chemicals.³¹ It relates a set of predictor variables (e.g., number of carbon atoms in the molecule) to the response variable (e.g., boiling point). The discovered relationship can then be used to predict the activities of new chemicals. QSAR tools have been developed to predict aquatic ecotoxicity of chemicals, such as Ecological Structure Activity Relationships (ECOSAR),³² Kashinhou Tool for Ecotoxicity (KATE),³³ Toxicity Estimation Software Tool (TEST),³⁴ ADMET (absorption, distribution, metabolism, excretion, and toxicity),³⁵ and Computer-Aided Discovery and REdesign for Aquatic Toxicity (CADRE-AT).³⁶⁻³⁸ A recent review compared these tools and concluded that ECOSAR outperforms other tools except CADRE-AT, which however is not currently available to the public.³⁹

Traditional QSAR models, including ECOSAR, are mostly based on linear models (i.e., ordinary least squares regression), or advanced linear models, such as partial least square regression,⁴⁰ principal component regression.⁴¹ Recent focus has been shifted towards more complex and nonlinear approaches, such as machine learning models.⁴² However, applications of machine learning in environmental toxicology are still limited.⁴³ A few studies developed machine learning models for predicting ecotoxicity of chemicals, such as decision tree,⁴⁴ discriminate analysis.^{36, 38} Li *et al.* (2017) compared six machine learning methods and results show support vector machine and neural network gave best results in acute toxicity prediction.⁴⁵ Most of the studies using these models to classify chemicals into different level of concerns for screening and prioritization of chemicals for regulation. However, in life cycle assessment, exact toxicity values are required to quantitatively evaluate the risk of chemicals along products' life cycle. A few studies use machine learning models including neural networks to estimate parameters for calculating characterization factors, such as removal rates,⁴⁶ fate factors, and intake fractions.⁴⁷⁻⁵⁰ Nevertheless, these parameters can be calculated by the multimedia model in USEtox. The missing of CF_{eco} data in USEtox is actually due to the missing of effect factors (EFs). To the best of my knowledge, no study has been done using machine learning models to estimate EFs of chemicals, which can then be used to calculate ecotoxicity characterization factors for LCIA.

1.3 Research Questions

My overall research question is: how can we transform the current time-consuming, expensive practice of LCA data collection into a faster, less expensive process that still generates reliable LCA data?

Regarding LCI data, I focus on estimating unit process data (Chapter 2). Unit process data mean material/energy used and waste/emission generated in specific industrial processes, which are used to develop a life cycle inventory for a product system. Specifically, I address the following questions:

- (1) How can we estimate the missing unit process data in an LCI database?
- (2) What are the implications for LCA?

Regarding LCIA data, I focus on estimating effect factors (Chapters 3-4), which are an integral part of characterization factors that reflect the relative contributions of life cycle resource consumption and emissions to various environmental impacts. Here, I focus on ecotoxicity impact caused by chemicals used along products' life cycles. The specific research questions are:

- (1) How can we estimate the missing effect factors based on existing data, and then calculate ecotoxicity characterization factors?
- (2) How can we evaluate the uncertainty of the estimated data?
- (3) What are the implications for LCA?

1.4 Structure of the Dissertation

The remainder of the dissertation is organized as follows. Chapter 2 proposes a data-driven method to estimate LCI data focusing on unit process data. Chapters 3 and 4 present two data-driven methods to estimate LCIA data focusing on ecotoxicity characterization factors. The last chapter concludes and discusses the significance of the work.

In Chapter 2, I develop a computational approach to estimate missing unit process data solely relying on limited known data based on a similarity-based link prediction method. The basic idea is that similar processes in the industrial system tend to have similar material/energy

inputs and waste/emission outputs. The results have been published in the journal *Environmental Science & Technology* (Vol.52, No.9, p.5259-5267).⁵¹

In Chapter 3, I develop neural network models to estimate ecotoxicity characterization factors for chemicals based on their physical-chemical properties and chemical descriptors. I use a genetic algorithm to optimize the structure and design of the neural network. The discovered relationship can be used to predict the toxic effects of new chemicals in both LCA and, more broadly, environmental impact assessment of chemicals. A manuscript based on this work is currently under review in the journal *Environment International*.

In Chapter 4, based on the same datasets in chapter 3, I develop random forest models to estimate ecotoxicity characterization factors for chemicals. A random forest model builds multiple decision trees and merges them together to get a more accurate and stable prediction. A manuscript based on this work is in preparation.

Chapter 2 Estimate Unit Process Data Using Similarity-based Link Prediction

2.1 Introduction

In life cycle assessment (LCA), collecting unit process data from the empirical sources (i.e., meter readings, operation logs/journals) is often costly and time-consuming. In this chapter, I develop a similarity-based computational framework for estimating missing unit process data solely based on limited known data, without relying on additional empirical data. In particular, I use similarity-based link prediction. The intuition is that similar processes in a unit process network tend to have similar material/energy inputs and waste/emission outputs. I use the similarity of unit processes in a unit process database to characterize the structure of the unit process network and develop the computational model to estimate missing unit process data. I use the ecoinvent 3.1 unit process datasets (UPR) to test this method. I first use this method to estimate a subset of the UPR data using the other subset as the training data. I then compare the estimated data with the original UPR data to evaluate the performance of the method.

2.2 Data and Methods

2.2.1 Unit Process Data Structure

The intuition of our method is based on the fact that unit process data essentially represent the interrelationship of unit processes (by intermediate flows) and the interrelationship between unit processes and the environment (by elementary flows). The ensemble of such interrelationship characterizes the structure of the underlying technology network (or unit

process network). If sufficient, observed unit process data, although not complete, can potentially be used to extract structural features of the underlying technology network. Such structural features, in turn, can be used to predict the structure of the unknown area of the technology network, which is equivalent to estimating the unknown data in the unit process database.

Unit process data are commonly represented as matrices when used in matrix-based LCA models. As illustrated in **Figure 2-1**, a unit process database is a matrix with columns representing unit processes (e.g., production of 1 kWh electricity) and rows representing either intermediate flows (i.e., inputs required by each unit process from other unit processes) or elementary flows (e.g., water consumption, CO₂ emissions). Each element of the matrix indicates the amount of a particular type of intermediate or elementary flow (row) associated with the unitary output of a particular unit process (column), e.g., 1.07 kg CO₂ emissions per kWh electricity production in hard coal power plants. Therefore, this matrix is a combination of technology matrix (A matrix) and emission matrix (B matrix).⁵²

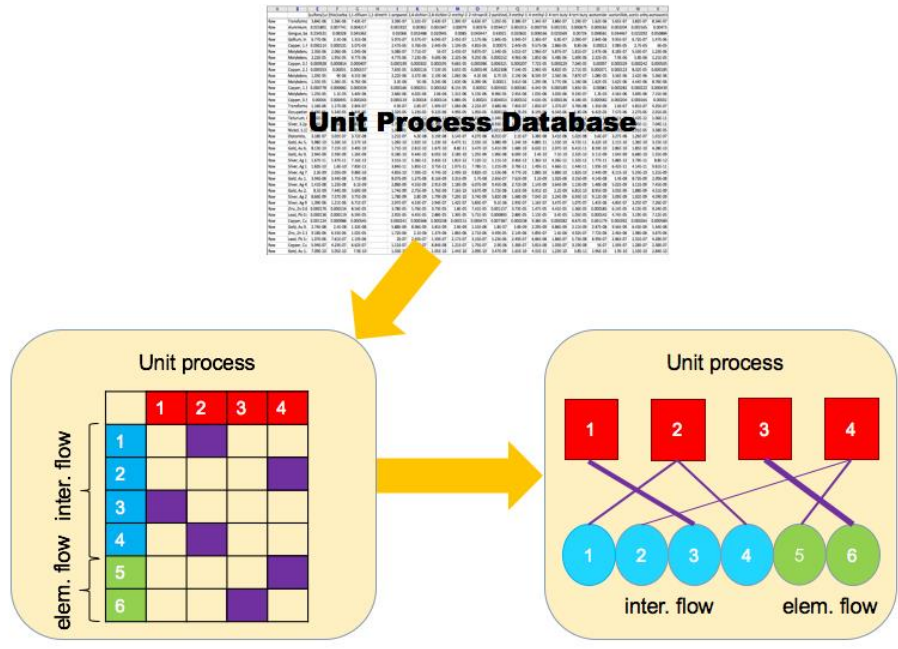


Figure 2-1. Data structure of a unit process database.

A unit process database can also be represented as a network, using the unit process matrix as the adjacency matrix (**Figure 2-1**). In particular, there are two types of nodes (or vertices) in a unit process network respectively representing unit processes and intermediate and elementary flows. Unit process nodes are connected with intermediate or elementary flow nodes by links (or edges) indicating how much and what type of flow each unit process is associated with. This network is a weighted, bipartite network,⁵³ as its links have strengths (the amount of flow) and nodes are divided into two disjoint sets. The network representation allows studying structural features of complex systems.^{54, 55} Representing a unit process database as a network allows identifying the features of its structure which is then used to estimate missing unit process data.

2.2.2 Link Prediction and Similarity-Based Link Prediction

Link prediction is a branch of the emerging network science to predict missing links in a

network based on limited observations.⁵⁶ Link prediction has mostly been applied in recommendation systems,⁵⁷⁻⁶⁰ such as in e-commerce sites to recommend products/services to likely customers.⁶¹ Link prediction has also been applied in analyzing social networks, such as characterizing the structure of literature citation networks,⁶² predicting collaborations in co-authorship networks,⁵⁶ and detecting relationships among terrorists.⁶³ Viewing the unit process data as a network allows us to use link prediction to explore the interrelationship between unit processes.

Many link prediction applications use similarity-based methods.⁶⁴ As the name suggests, similarity-based methods first measure the similarity (or proximity) between each pair of nodes in the network. For bipartite networks, the similarity is measured for the same type of nodes. Two nodes that are similar tend to have similar patterns of linkages with other nodes in the network. Based on appropriate measures of similarity, we can then evaluate the likelihood of unknown links that exist for a node by comparing it with other similar nodes. I apply the principle of similarity-based link prediction in this work to estimate missing unit process data. Note that in unit process networks, only predicting the existence of links between processes is not enough. We also need to predict the strength of particular links. Although this is different from simply applying existing link prediction methods that are mostly developed for unweighted networks, the same principles still apply.^{65, 66}

2.2.3 Steps to Develop the Link Prediction Model

As shown in **Figure 2-2**, I use a reputable unit process database as the complete, observed dataset, which is an $m \times n$ matrix including m types of intermediate and elementary flows and n types of unit processes. For each process j (column $j \in [1, n]$ in the matrix), when I use it as the test set, the rest of the matrix becomes the training set. I then randomly select p

($1 \leq p < m$) number of data from the test set (column j) and assume they are missing. I use the training set to estimate those missing data based on the similarities of the remaining data in the process j with the k ($1 \leq k \leq n-1$) most similar processes (details described below). Finally, I compare the estimated data with the original data to evaluate the performance of the method. To measure the effectiveness of our method, this procedure is repeated for each process being used as the test set, the same as in leave-one-out-cross-validation (LOOCV). The overall performance of the method with respect to the selected unit process database can then be evaluated by averaging the performance metrics obtained each time (details see equation (2-4)).

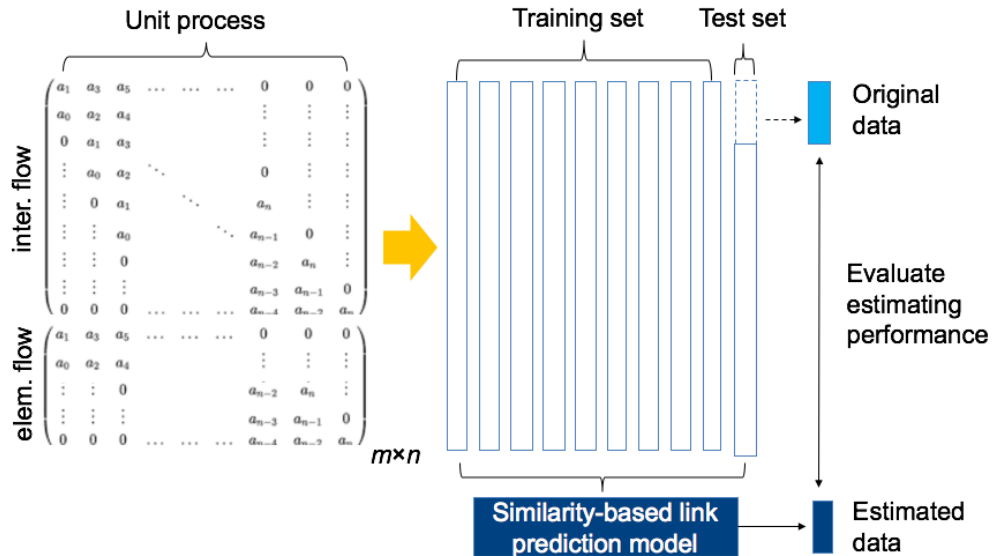


Figure 2-2. Methodological framework of similarity-based link prediction applied to estimating unit process data.

For each test set with certain numbers of missing data, I use the following three steps to complete the estimation and evaluate the performance of the estimation:

(1) Similarity calculation. I compute the similarities of test set (process j) which has missing data with other processes in the training set by comparing the remaining portion of

process j and the corresponding portion of each process in the training set. While many methods are available to measure the difference between two vectors, I choose the Minkowski distance based on comparison of these available methods (**Appendix A**). In particular, the Minkowski distance between the remaining portion of the test set and the corresponding portion of each process in the training set is calculated as:

$$d_{ij} = \left(\sum_{t=1}^{m-p} |a_{ti} - a_{tj}|^q \right)^{1/q} \quad (2-1)$$

where t indexes intermediate and elementary flows, $m-p$ is the total number of flows minus the number of the missing data (i.e., number of known data in process j), a_{ti} is the flow t in the training process i , a_{tj} is the flow t in the test process j , and q is the parameter in the definition of Minkowski distance. One can get different measurements of the distance by adjusting q . During the training, I find the best q which achieves the lowest estimation errors for each training dataset. The similarity of the known portion of process j and the corresponding portion of process i , s_{ij} , is then calculated based on their distance d_{ij} :

$$s_{ij} = \frac{1}{d_{ij}+1} \quad (2-2)$$

The larger the s_{ij} is, the more similar the two processes are. If one process in the training set is identical with the process in the test set, then their distance d_{ij} is 0 and their similarity s_{ij} becomes 1.

(2) Missing data estimation. Each missing data point e_{tj} in the test set process j is estimated by averaging the corresponding data in the k most similar processes weighted by their similarities, which are calculated by equation (2-1) and (2-2):

$$e_{tj} = \frac{\sum_{i=1}^k a_{ti} s_{ij}}{\sum_{i=1}^k s_{ij}} \quad (2-3)$$

where k ($1 \leq k \leq m-1$) represents the number of most similar processes in the training set used to estimate the missing data in the test set and a_{ti} is the corresponding flow t of the i -th similar process when the training processes are ranked in descending order of similarity. For every set of missing data, there are $m-1$ different estimations with k ranging between 1 and $m-1$.

(3) Performance measurement. I evaluate the performance of the model by comparing the estimated data e_{tj} with the original data a_{tj} using mean percentage error (MPE):

$$MPE = \sqrt{\frac{\sum_1^p (a_{tj} - e_{tj})^2}{\sum_1^p a_{tj}^2}} \quad (2-4)$$

where p is the number of the missing data. Lower MPE indicates more accurate estimation of the missing data.

One important assumption of our method is that the observed unit process data we use to estimate the missing data should be complete. In other words, the applicability of our method largely depends on the completeness and quality of the observed unit process database. Our method does not intend to replace primary data collection for unit processes, but as a complementary approach when primary data are not available.

2.2.4 Data

I use ecoinvent 3.1 database⁶⁷ as a reputable database to test this method. The ecoinvent database is perhaps the most widely used LCA database, which comprises data for thousands of common unit processes. There are three models in ecoinvent including default model, cut-off model, and consequential model. The three models have the same matrix structure, except they use different methods to deal with co-products and wastes. For each model, ecoinvent provides three datasets including unit process datasets (UPR), aggregated life cycle inventories (LCI), and calculated impact assessment results (LCIA). UPR records the data for energy/resource inputs

and emission outputs of a process. Aggregated LCI converts all upstream UPR data into the life cycle inputs and outputs of a process. LCIA is the characterized LCA results by categorizing energy and resource uses and emissions into various categories of environmental impacts. In this study, I use the UPR data in the default model to represent the underlying technology network of the ecoinvent database, as the aggregated LCI and LCIA data are essentially calculated based on UPR for users' convenience.

Ecoinvent 3.1 UPR database is a 13,201 by 11,332 matrix, corresponding to 13,201 types of flows (including 11,332 types of intermediate flows and 1,869 types of elementary flows) and 11,332 unit processes. Because UPR only includes onsite energy and resource use and emission data for each process, most entries in the UPR matrix are zeros. Thus, the ecoinvent UPR database is a sparse matrix, with only 9.8% of entries are nonzero.

The ecoinvent database implements inheritance for geography, which means a local process can be created as a child of the global parent process. The child process inherits all flows from the parent unless otherwise specified to ensure consistency of processes for the same activity in different regions.⁶⁸ Some local processes are generated as an exact copy of the global process with uncertainty adjusted. I kept only the parent process by removing the child processes specific for different locations. Empty rows and columns in the UPR matrix are also removed. As a result, the processed UPR matrix has 7,029 intermediate and elementary flows (row) and 2,546 processes (column).

Data in the ecoinvent database have very different orders-of-magnitude due to the nature of intermediate and elementary flows and the choice of units. For example, CO₂ emissions to the air can be in the order of 10⁻⁵-10¹ kg for the unitary output of a process, while lead discharges to the air for the same unitary output of the same process can only be in the order of 10⁻¹⁵-10⁻¹⁸ kg.

If one of the flows has a relatively high order of magnitude, the similarity of unit processes will be dominated by this particular flow. In addition, the choice of units of processes also affects the order of magnitude. For instance, the dataset of 1 km passenger car transportation and the dataset of 1 metric ton-km freight train transportation have very different orders of magnitude since the later converted the data to per metric ton freight being transported. Normalization sometimes is needed to represent data in similar orders of magnitude. In this study, I define a specific procedure of matrix normalization (**Appendix B**) and compare three different strategies: 1) normalization based on the complete UPR matrix; 2) normalization based on the training set to avoid introducing future information in the test set; and 3) without normalization. Error! Reference source not found. compares the estimation results of using these three strategies. Overall, estimation without normalization offers the best results. This is because normalization, while making the data more regular, can actually lose important information from the raw dataset. Such information can be useful to improve the estimation accuracy.

2.3 Results and Discussion

2.3.1 Similarity of Processes

I calculate the similarities of each pair of processes in the ecoinvent 3.1 UPR dataset. The resulting similarity matrix is a symmetric, square matrix with both rows and columns representing unit processes and elements standing for the similarities (s_{ij}) between pairs of processes. Since the processed matrix has 2,546 processes, the similarity matrix is a square matrix of 2,546 by 2,546. It shows the similarities of each process with other 2,545 processes in the matrix. Each cell's value s_{ij} is calculated by equations (1) and (2), showing the similarity of process i with process j . The smaller the s_{ij} is, the less similar process i is to process j . **Figure**

2-3 shows the heat map of the similarity matrix to demonstrate the disparity of similarities of each pair of processes. The grid-like pattern indicates that the process pairs have significantly different levels of similarity, which provides valuable information to extract the underlying structure of the dataset. The processes in **Figure 2-3** are ordered by International Standard Industrial Classification of All Economic Activities (ISIC) in which processes in the same industry are next to each other. We observe irregular distribution of similarities in **Figure 2-3**. Specifically, there are no light squares around the diagonal, meaning processes from the same industry do not necessarily have similar intermediate and elementary flows. Therefore, using processes from the same industry to update missing intermediate and elementary flows as commonly done in LCI is in fact not always appropriate. Note that the similarities shown here are not the similarities we use to estimate the missing data. The similarities here are based on the complete dataset. When we estimate missing data, similarities are calculated each time only based on the remaining data after the missing data are removed.

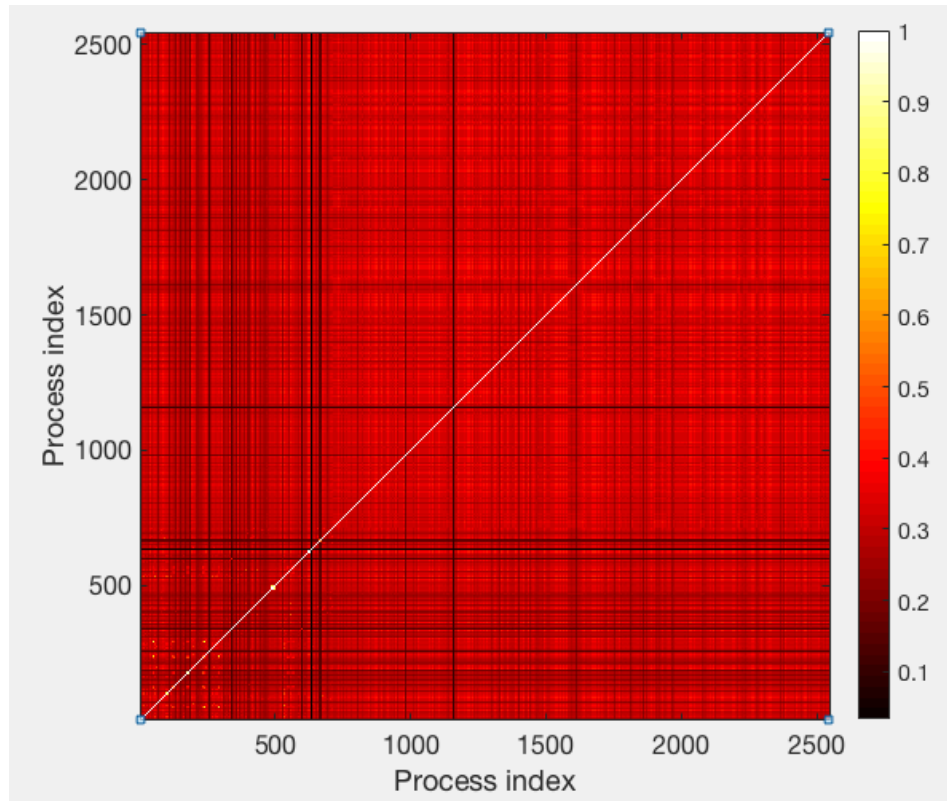


Figure 2-3. Heat map of the similarity matrix for the ecoinvent 3.1 UPR dataset.

As an example, I find the ten most similar processes for the process “machine operation, diesel, <18.64 kW, underground mining” (**Table 2-1**). The four most similar processes are all basically the same as the selected process, just with different conditions. The fifth to ninth most similar processes all involve the combustion of diesel as the selected process does. The tenth is a waste incineration process which also involves combustion.

Table 2-1. The ten most similar processes for “machine operation, diesel, <18.64 kW, underground mining” (the index number is 2,429 in **Figure 2-3**).

	Processes	Similarity	Index number in Figure 2-3
1	'machine operation, diesel, < 18.64 kW, steady-state'	0.9083	2,428
2	'machine operation, diesel, < 18.64 kW, low load factor'	0.8925	2,427
3	'machine operation, diesel, < 18.64 kW, high load factor'	0.8498	2,426
4	'machine operation, diesel, < 18.64 kW, underground mining'	0.7843	2,429

5	'excavation, skid-steer loader'	0.7098	294
6	'excavation, hydraulic digger'	0.7095	293
7	'diesel, burned in building machine'	0.7062	688
8	'bale loading'	0.7055	119
9	'baling'	0.7053	120
10	'waste vapour barrier, flame-retarded'	0.7052	1,820

In this method, I solely use unit process data to measure the similarity between any pair of processes, while ecoinvent also categorizes unit processes based on their industrial classification. Intuitively, one would assume that processes in the same category would be more similar to each other. I calculate the similarity of each pair of processes and found the most similar process for each process. However, the results show that only approximately 32% unit processes are in the same category with their most similar process. For example, cement and clinker both belong to the construction industry. However, because clinker production is high energy intensive but cement production is just to mix raw materials including clinker, clinker and cement production processes are very different despite same industry. Our results prove that their similarity is low (0.2308). Therefore, using processes in the same industry to update missing flows, as commonly done in LCA practice, is not always appropriate.

2.3.2 Estimation Performance

I use mean percentage error (MPE) to evaluate the accuracy of estimating different numbers of missing data. Specifically, I test missing 1%, 5%, 10%, and 20% of the total number of intermediate and elementary flows (7,029). **Figure 2-4** shows the MPEs when different percentages of data are missing when k (the number of most similar processes) ranges from 1 to 2,545 and q (the parameter in the distance function) ranges from 0.01 to 0.1. The MPEs are the average of the MEPs by making every process in the dataset as testing data one by one. When

less data are missing, the estimation MPEs are distributed in relatively narrow ranges (e.g., 1% and 5% data missing in **Figure 2-4**). This implies that most processes can be estimated relatively well except for a few outliers. When more data are missing (e.g., 10% data missing), the distribution of MPEs becomes much broader. This indicates the number of processes that are difficult to estimate becomes larger when more data become missing. However, when even more data are missing (e.g., 20% data missing), the MPEs are distributed again in a narrow range but with large values, which means when missing data exceeds a certain level, the missing data are generally hard to estimate. This is because, when more data are missing, the less information we can use to estimate those missing data and the similarity measures are getting less reliable in finding similar unit processes. MPE is the lowest when a few most similar processes are used for the estimation. However, when more processes are included, MPE values actually increase, because newly added processes are less similar and introduce more noises. Therefore, using more processes for the estimation doesn't necessarily mean lower MPE.

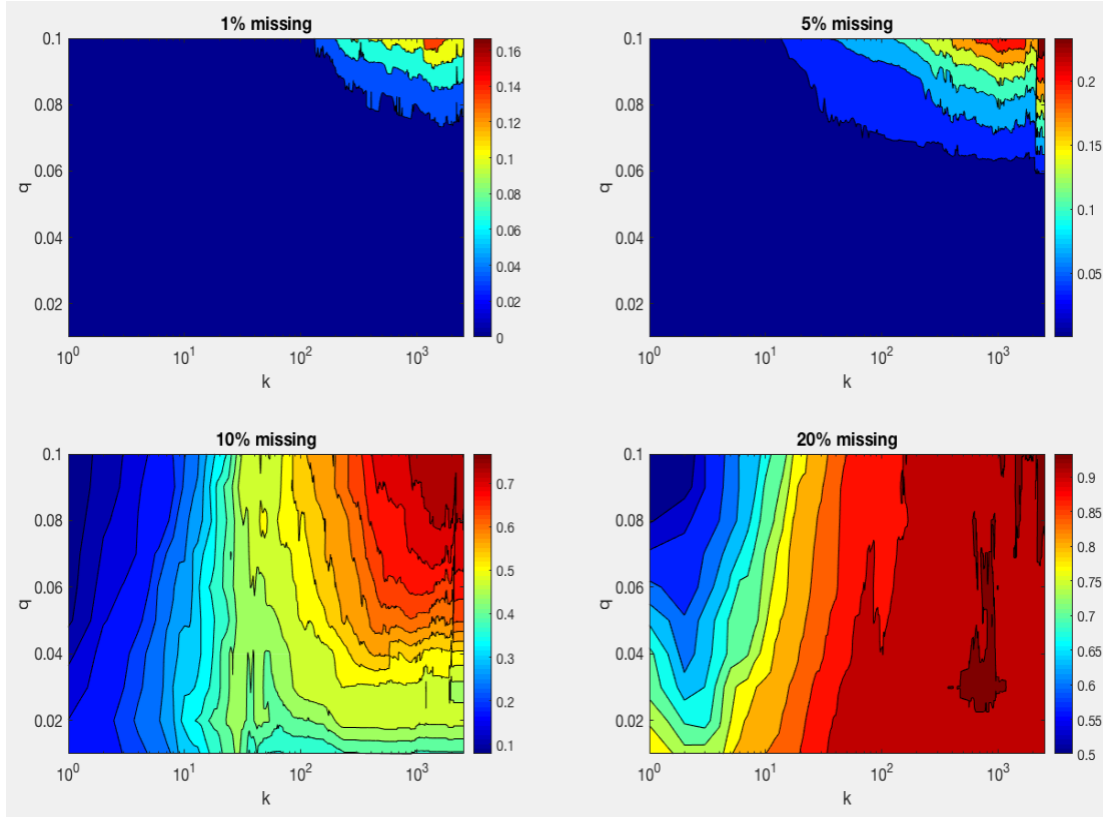


Figure 2-4. MPEs with respect to percentage of data missing, k (the number of most similar processes), and q (the parameter in the distance function).

Figure 2-5 shows the distribution of the MPEs for all processes with respect to different percentages of data missing. Based on these distributions, we can choose the value of parameter q that has the best estimation performance. The parameter q in the Minkowski distance is essentially a general representation of distance function. When $q = 1$, it is Manhattan distance; when $q = 2$, it is Euclidean distance. Larger values of q place more emphasis on large differences in intermediate and elementary flows, because all differences are raised to the power of q . Consequently, the distance with higher q is strongly influenced by a single large difference in one flow. From **Figure 2-5**, we can see that smaller q generally tends to correspond to lower

MPEs. Because we have in total 7,029 flows, placing more emphasis on small number of large differences is not helpful on the estimation.

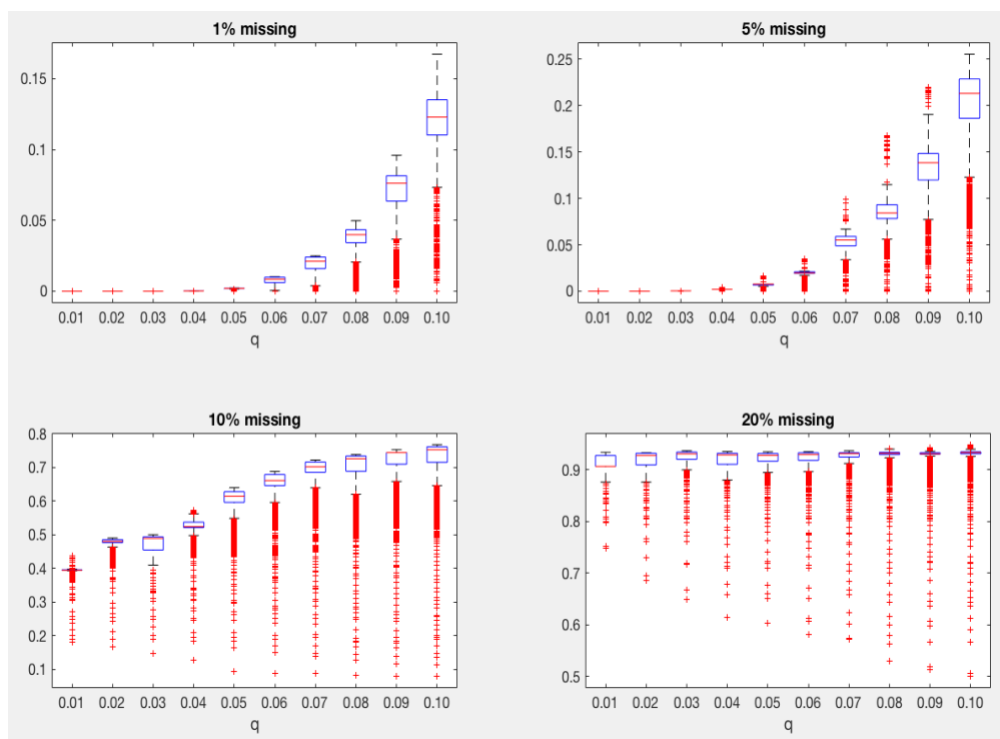


Figure 2-5. The distribution of MPEs with respect to the percentage of data missing and value of q (the parameter in the distance function).

Table 2-2 and **Figure 2-6** show the distribution of the MPEs for all the processes with respect to the best q , which is 0.01. MPEs are distributed around the average value with relatively small standard deviations. These results show that the accuracy of estimation increases (i.e., average MPE decreases) when less data are missing. When 1% and 5% data are missing, we can estimate those missing data with a very high accuracy (i.e., the average MPE are almost near zero). Given that the dataset includes many entries as zero, missing less than 5% data often means a few non-zero data points are missing. As a result, one or two processes in the training set that are most similar to the process in the test set can effectively dominate the estimation and make the estimation very close to the original values. When 10% data are missing, the average

MPE becomes 39.32%. The average MPE exceeds 90% when more than 20% data are missing in one process. I did not test more scenarios of data missing since the result for 20% missing is already beyond acceptance and more data missing will only worsen the estimation results.

Table 2-2. MPEs with different percentages of data missing.

MPE	Average	25% quantile	Median	75% quantile	Standard deviation
1% missing	$2.09 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$1.61 \times 10^{-14}\%$
5% missing	$2.85 \times 10^{-12}\%$	$2.54 \times 10^{-12}\%$	$2.54 \times 10^{-12}\%$	$3.35 \times 10^{-12}\%$	$4.39 \times 10^{-13}\%$
10% missing	39.32%	39.45%	39.46%	39.58%	1.26%
20% missing	91.39%	90.61%	90.61%	92.71%	1.37%

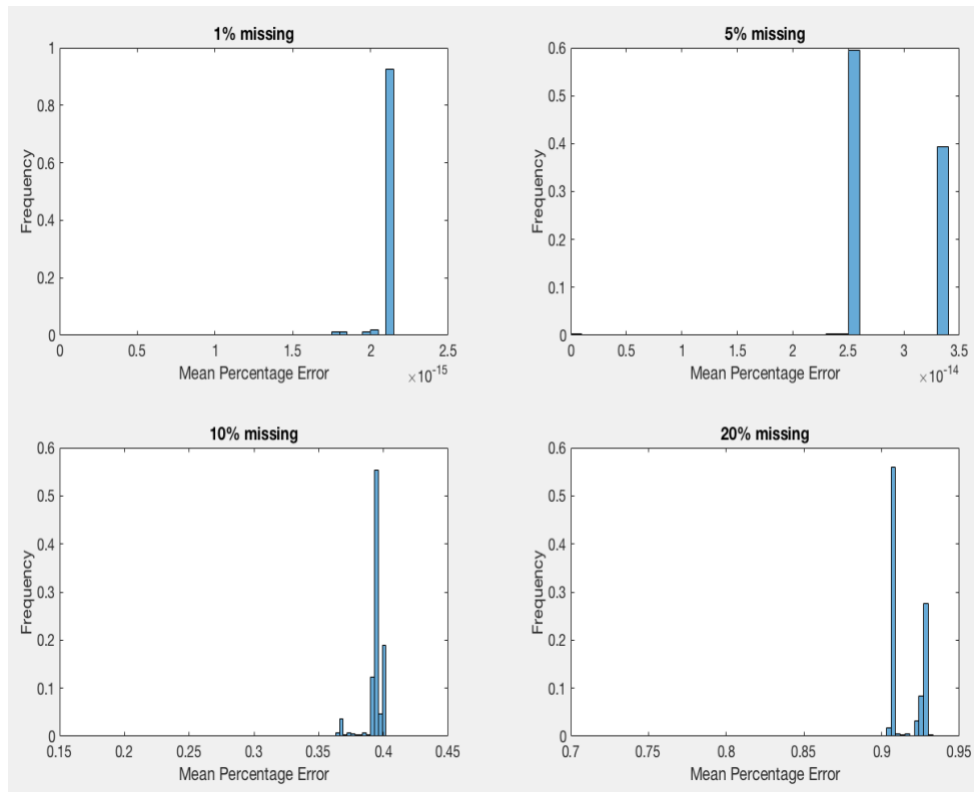


Figure 2-6. Histograms of MPEs when different percentages of data are missing with the best q .

2.3.3 Computational Time

I examine the computational time required to implement this method. I use Flux, the high-performance computer available at the University of Michigan, specifically, the Nehalem architecture compute nodes, each configured with 12 cores (two 6-core 2.67 GHz Intel Xeon X5650 processors) and 48 GB RAM. The code is programmed in MATLAB. In the case of missing 1% data with one processor, it takes approximately 25 minutes to estimate the missing data for all processes. The estimation itself takes the greatest computational time (94.1%), while other procedures including the similarity calculation and data preparation only need approximately 5.9% of the computational time. The estimation step dominates the computational time because it involves many matrix multiplications, i.e., 2,545 times of matrix multiplication for estimating missing data in each process, in total 6.5 million times of matrix multiplication. When more data are missing, the time spent on similarity calculation would decrease since the dataset for calculating similarity is smaller. However, the time for estimation would increase significantly since the dataset for estimation becomes larger. In order to improve the computational efficiency, I use 10 processors regarding to 10 different q to calculate simultaneously. **Table 2-3** shows the computational time required for calculating the whole database. Overall the computational resource needed for implementing our method in ecoinvent 3.1 is manageable.

Table 2-3. Computational time required for completing the link prediction estimation.

Missing data scenarios	Time required
1% missing	1.25 hours
5% missing	5.62 hours
10% missing	11.94 hours
20% missing	57.22 hours

2.3.4 Theoretical Grounds

While the computational results are promising, I further explore theoretical grounds that can explain the results. Essentially, the ecoinvent UPR dataset is high dimensional with 7,029 variables (intermediate and elementary flows) but only 2,546 samples (processes). Such high dimensional data often lie close to low dimensional subspaces. The intrinsic dimension of the matrix can be measured with the effective rank:⁶⁹

$$r(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|} \quad (2-5)$$

where $\text{tr}(\Sigma)$ is the trace of a matrix, which is the sum of all the eigenvalues of Σ , and $\|\Sigma\|$ denotes the largest singular value of Σ .

Based on equation (2-5), the effective rank of the processed dataset is approximately 2. This means that the first two components explain 89.3% of the total variance. This is reasonable since some intermediate flows and elementary flows are highly related with each other. For example, processes with high energy consumption normally have high levels of greenhouse gas emissions. This low rank characteristic of the ecoinvent 3.1 UPR dataset allows us to estimate limited number of missing data with the underlying pattern of the dataset and explains why our method works very well when missing 1% -5% data. I also recognize the diversity among the types of flows and the categories of unit processes. Such diversity could be the reason why the estimated results are not satisfactory when more data are missing.

2.3.5 Case Study

To demonstrate the application of the link prediction model, I choose one process in the US LCI database,⁷⁰ “Diesel, combusted in industrial boiler”, to identify its possible missing flows. I calculate the similarities between this process and all the unit processes in ecoinvent.

The most similar process in ecoinvent is “machine operation, diesel, < 18.64 kW, low load factor”. The descriptions of these two processes suggest strong similarity as well. The US LCI process and its most similar process in ecoinvent have 17 flows in common. The MPE of estimating the 17 flows of the US LCI process using its similarities with ecoinvent processes is 7.05%. Comparing the two processes, the ecoinvent process has 36 additional intermediate flows (e.g., lubricating oil) and elementary flows (e.g., ammonia and benzene emissions). This suggests the US LCI process potentially misses data for these additional flows. Therefore, ecoinvent data can be used to estimate data for these additional flows for the US LCI process.

2.3.6 Implications for LCA

I envision three major implications for LCA research and practice. First, empirically unit process data collection is expensive and time-consuming. Based on ecoinvent database, we show that the similarity-based link prediction approach can effectively estimate missing data with high levels of accuracy. The adoption of this approach will provide reasonably accurate unit process data when primary data are not available, with only a fraction of cost and time for collecting primary data.

Second, unit process data are collected from various sources with different quality. We can use a portion of the unit process dataset that is trustworthy to estimate the less trustworthy data. By comparing the estimated results with the collected data, we can identify the data potentially with low quality. This result can help guide directions for improving data quality.⁷¹ In ecoinvent, there are many missing intermediate and elementary flows that are represented as zeros. To investigate which of the current zero entries should in fact not be zeros, I apply this method on the zero entries when 1% data missing. I find that the estimated data for the zero entries are also zeros for 90% of the processes. The rest 10% unit processes, for which estimated

data are non-zeros, are generally non-exceptionally market processes, newly introduced in ecoinvent v3. The distinctive feature of these processes is that there is no transformation of materials happening, simply adding transport activities, wholesale and retail activities, and product losses in trade and transport. Missing data for market processes in ecoinvent are substituted by a simple market dataset.⁶⁸ This is the reason why this method estimated the zero entries as non-zeros. In other words, this method can be used to identify those missing or low-quality data and direct future data collection efforts.

Lastly, LCA databases are constantly expanding due to the addition of new unit processes from new technologies. It is often the case that the unit process data for new, emerging processes are incomplete. Our method can be used to estimate the incomplete data for a new process based on its similarities with other processes in a known LCA database.⁷² Note that this method does not apply if there is no data at all for a new process, because we cannot compute the similarity and find the relationship between the new process and other processes. However, it is also very rare that we do not know anything about the new process except what it produces. At least, one should be able to know energy and key material uses for producing unitary product from this process, which can potentially be used to estimate other intermediate and elementary flows.

2.3.7 Future Work

The results and conclusions only apply to the ecoinvent 3.1 UPR dataset. In addition, I need to test this method in other commonly used LCA databases, such as the Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation (GREET) model⁷³ and the US LCI Database⁷⁰. In particular, the ecoinvent database is one of the proprietary databases with comprehensive coverage; GREET has been developed for a particular sector (transportation); and the US LCI Database is a national reference LCA database that provides industrial-

representative LCA data for a particular country. Testing this method using these representative LCA databases will help understand the applicability of the method and its limitations.

In addition to the similarity-based link prediction method used in this study, other methods we have seen in the recent rapid development of data science can potentially also be used to estimate missing unit process data. My future work will explore the potential applications of these methods in LCA data estimation.

2.4 Summary

In this chapter, I propose a new computational approach to estimate missing unit process data solely relying on limited known data based on a similarity-based link prediction method. I use the ecoinvent 3.1 unit process datasets to test our method in four steps: 1) dividing the datasets into a training set and a test set; 2) randomly removing certain numbers of data in the test set indicated as missing; 3) using similarity-weighted means of various numbers of most similar processes in the training set to estimate the missing data in the test set; and 4) comparing estimated data with the original values to determine the performance of the estimation. The results show that missing data can be accurately estimated when less than 10% data are missing in one process. The estimation performance decreases as the percentage of missing data increases. This study provides a new approach to compile unit process data and demonstrates a promising potential of using computational approaches for LCA data compilation.

Chapter 3 Estimate Ecotoxicity Characterization Factors for Chemicals Using Neural Network Models

3.1 Introduction

A wide range of environmental impacts in product life cycles are associated with the usage and release of chemicals. For example, the life cycle of food generally includes production, harvesting, processing, packing, transport, marketing, consumption, and waste treatment and disposal. In every step of this life cycle, chemicals are used for food processing and preservation. The release of and exposure to those chemicals can impact ecosystem and human health. Quantifying the potential environmental impacts of chemicals is thus critical for LCA. Broadly, understanding environmental impacts of chemicals is also an important step towards creating effective policy and regulations to reduce harmful impacts to human health from chemical exposure. Witnessing the rapid development of neural networks (a.k.a., artificial neural networks) and the promising of advanced neural networks (e.g., deep learning),⁷⁴ here I use neural networks to predict the ecotoxicity values of chemicals for calculating their characterization factors in LCA.

The success of neural network models depends on designing an optimal architecture and hyper parameters to fit the task. In the traditional neural network training, the architectures are normally fixed and the training parameters are previously assigned. Therefore, finding a satisfactory architecture and hyper parameters involves many trials and errors. Within a given training period, one may not find neural networks with satisfactory performance. Indeed, if there

were infinite time and infinite computing resources, one could brute force the problem and compare all parameter combinations by grid search. However, in most real-world applications of neural networks, we have to balance the time and cost to generate acceptable networks rapidly. In this context, one of the global stochastic optimization algorithms, the genetic algorithm, can be used to find the best neural network rapidly.

Genetic algorithm is a directed random search technique that simulates the natural selection and evolution process.⁷⁵ Because it can be directly integrated to existing simulations and models, genetic algorithm has been widely used for many optimization problems which have a large number of parameters and their analytical solutions are hard to derive.⁷⁶ Rationally, genetic algorithm has also been used to optimize neural networks,^{77, 78} but rarely used for predicting chemical toxicity.^{79, 80} In these studies, the neural network architectures are predefined and the genetic algorithm is only used for selecting input variables and optimizing other parameters. In addition, these studies are specific for certain types of chemicals and the predicted toxicity is for certain species, which limits the applications of the developed models.

In this chapter, I aim to provide missing EFs and characterization factors for chemicals in USEtox. I build a neural network model to estimate EFs based on the USEtox data. I use genetic algorithms to rapidly find optimal architecture and hyper parameters for the neural network model. To evaluate the performance of the model, I compare its performance with those of traditional QSAR models, including ECOSAR model, ordinary least squares regression (OLS), partial least squares regression (PLS), and principal component regression (PCR).

3.2 Data and Methods

In USEtox, ecotoxicity characterization factor of a chemical (CF_{eco} [PAF·m³·d·kg⁻¹]) is calculated by:

$$CF_{eco} = FF \times XF \times EF \quad (3-1)$$

where FF [d] is the fate factor calculated by the multimedia transport and transformation model in USEtox, which characterizes the distribution of emitted contaminants among different compartments (e.g., urban air, agricultural soil, freshwater). XP [-] is the ecotoxicity exposure factor, calculated as the fraction of a chemical dissolved in freshwater. EF [PAF*m³.kg⁻¹] is the effect factor that relates ecosystem exposures and dissolved masses in the freshwater ecosystem to a measure of the potentially affected fraction of exposed species.

The missing of CF_{eco} are generally due to missing EFs, which are calculated as:

$$EF = \frac{0.5}{HC_{50}} \quad (3-2)$$

where HC₅₀ [kg·m⁻³] is defined as the hazardous concentration of a chemical at which 50% of the freshwater species are exposed above their EC₅₀. The EC₅₀ is the effective concentration at which 50% of a population displays an effect (e.g. mortality) in a laboratory test. Besides EC₅₀, laboratory tests are results in other concentration values, such as LC₅₀ (lethal concentration 50%), NOEC (no observed effect concentration), and LOEC (lowest observed effect concentration). Given that EC₅₀ is less fluctuant with the test conditions and is the endpoint with the lowest uncertainty,⁸¹ USEtox model uses EC₅₀ to determine the relative aquatic ecotoxicity of chemicals and calculates the geometric mean of the EC₅₀ across different species, which is called hazard concentration (HC₅₀), indicating the average concentration affecting 50% of the species at a level above their EC₅₀.²⁹ In USEtox, HC₅₀ are primarily derived from chronic EC₅₀ because LCA mainly deals with chronic exposure. Wherever chronic EC₅₀ data are unavailable, best-estimate acute to chronic ratios are developed to extrapolate acute EC₅₀ to chronic EC₅₀.^{29, 82, 83}

3.2.1 Data Filtering and Exploratory Analysis

(1) **HC₅₀ data in USEtox.** USEtox version 2.1 contains 3,077 organic chemicals, in which 2,499 chemicals have HC₅₀ data based on laboratory tests. However, the toxicity data from laboratory experiments often have different levels of uncertainties.⁸⁴ I remove high uncertainty data using two steps. First, remove chemicals for which HC₅₀ are more than one order of magnitude above their baseline toxicity. The baseline toxicity is derived by a simple linear relationship as a function of the octanol-water partitioning coefficient (Kow), which increases with increasing hydrophobicity. EC₅₀ data that are much higher than the baseline EC₅₀ derived from Kow are likely to be unreliable or incorrect. Second, remove chemicals whose Kow value is larger than the solubility cut-off, a threshold above which a chemical is no longer soluble enough to result in toxicity. For chronic effects, chemicals with log Kow > 8.0 are expected to have no effects at saturation. Both baseline toxicity and solubility cut-off data are acquired from the ECOSAR model.⁸⁵

(2) **Physical-chemical properties in USEtox.** USEtox also provides physical-chemical characteristics for these chemicals. USEtox has complete data for the 2,499 chemicals on 11 physical-chemical properties (**Table 3-1**).

Table 3-1. Physical-chemical properties in USEtox.

Physical-chemical properties	Description
MW (g·mol ⁻¹)	Molecular weight
Kow (L·L ⁻¹)	Octanol-water partitioning coefficient
KH25C (Pa·m ³ ·mol ⁻¹)	Henry coefficient
Pvap25 (Pa)	Vapor pressure
Sol25 (mg·L ⁻¹)	Solubility
kdegA (s ⁻¹)	Degradation in air
kdegW (s ⁻¹)	Degradation in water
kdegSd (s ⁻¹)	Degradation in sediment

kdegSI (s ⁻¹)	Degradation in soil
kdissP (s ⁻¹)	Dissipation rates in above-ground
BAF _{fish} (L·kg _{fish} ⁻¹)	Bioaccumulation factor in fish

(3) Mode of action (MoA). The mode of action (MoA) is recognized as an important determinant of chemical toxicity.⁸⁶ To incorporate MoA information in the model, I use ToxTree⁸⁷ to assign MoA of chemicals by Verhaar scheme,⁸⁸ which classifies chemicals into five categories based on the presence and absence of certain chemical structures and elements.

(4) Theoretical molecular descriptors. I collect additional theoretical molecular descriptors, which are derived from symbolic representations of molecules by logic and mathematical procedures.⁸⁹ I use Toxicity Estimation Software Tool (T.E.S.T)³⁴ developed by the US Environmental Protection Agency (EPA) and QikProp⁹⁰ developed by Schrodinger to calculate theoretical molecular descriptors for the chemicals in USEtox. T.E.S.T. calculates 797 descriptors and QikProp calculates 51 physically and pharmaceutically significant properties based on the full 3D molecular structure.

3.2.2 Neural Network and Generic Algorithm

A neural network model simulates the way biological nervous systems (e.g., human brain) process information.⁹¹ A feedforward neural network composes multiple layers of neurons. The first layer contains predictors (or inputs), and the last layer contains responses (or outputs). Between the input layer and the output layer are one or more hidden layers interconnected with each other by hidden neurons. **Figure 3-1** shows a one hidden layer neural network model.

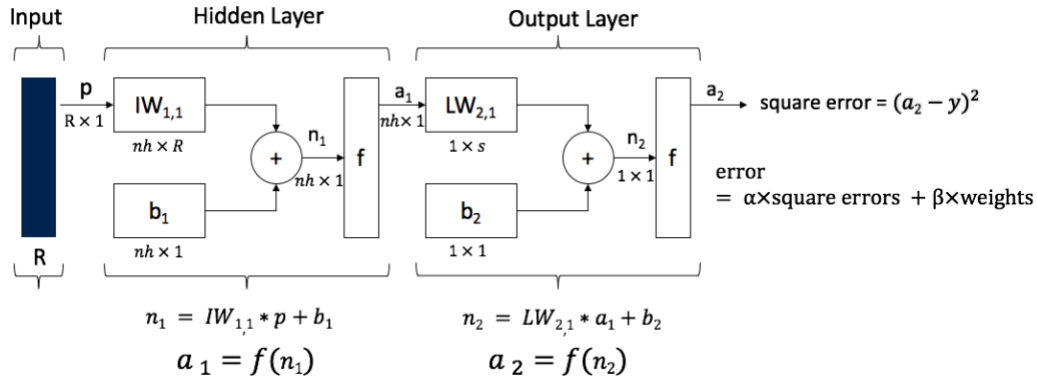
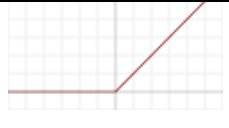
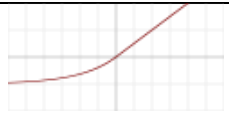
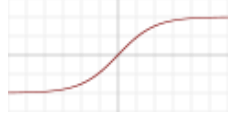



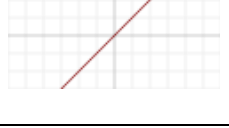


Figure 3-1. A neural network model with one hidden layer.

The solid vertical bar on the left represents the input vector p . The dimensions of p are $R \times 1$, indicating that the input is a single vector of R elements. The input vector is multiplied with the weight matrix IW ($nh \times R$), which represents the connection weights of nh neurons with R inputs. The product $IW \times p$ plus the bias vector b_1 is a $nh \times 1$ vector n_1 , which is the input of the activation function f (Table 3-2 shows commonly used activation functions). Activation function decides whether or how much outside connections should consider this neuron, i.e., whether and how much this neuron should be “activated”. The activation function is where the nonlinearity is introduced in the network, which enables the network to capture the nonlinear relationship in the data set. The output of f is a_1 , which is then used as the input of the output layer. Using similar procedures and a_1 as the input, the output layer derives its output a_2 which is the estimation results. By comparing a_2 and y , the desired output, we get the error of the model estimation. This step is called forward-propagation because the calculation flow is going towards the forward direction. Back-propagation is then required to update the parameters (i.e., weights and biases) to reduce the error of the model.⁹²⁻⁹⁴ This general process is called back-propagation, or gradient descent, which is an iterative process. Therefore, the dataset is often passed multiple times to the network during the learning process. The number of times the

dataset is shown to the network is called epochs. In each epoch, the dataset is often divided into small batches to update the parameters. The number of samples in each batch is called batch size.

Table 3-2. Commonly used activation functions.

Name	Input output relation	Graphic representation
ReLU (Rectified Linear Unit)	$f(x) = \alpha * (x - threshold) \quad x < threshold$ $f(x) = x \quad threshold \leq x \leq max_value$ $f(x) = max_value \quad x \geq max_value$	
ELU (Exponential linear unit)	$f(x) = \alpha * (\exp(x) - 1) \quad x < 0$ $f(x) = x \quad x \geq 0$	
TanH (Hyperbolic tangent)	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$	
Hard_sigmoid	$f(x) = 0 \quad x < -2.5$ $f(x) = 0.2 * x + 0.5 \quad -2.5 \leq x \leq 2.5$ $f(x) = 1 \quad x \geq 2.5$	
SoftPlus	$f(x) = \log(\exp(x) + 1)$	
Linear	$f(x) = x$	

Traditionally, the backpropagation (BP) algorithm⁹⁵ and its variations are used to optimize neural networks, but they often get trapped in local optima and cannot converge to the global minima.⁹⁶ Therefore, more advanced algorithms are proposed, such as RMSprop, Adam,⁹⁷ stochastic gradient descent (SGD), Adagrad,⁹⁸ Adadelta,⁹⁹ AdaMax, and Nadam¹⁰⁰. These

algorithms leverage information obtained during the training and adapt their learning rates accordingly. They are often called as optimizers.

Genetic algorithms have been increasingly used to evolve neural networks for their good global search capabilities.¹⁰¹⁻¹⁰⁵ Genetic algorithms search for the global optimal solution by continually transforming a population of individual solutions. At each generation, the genetic algorithm selects optimal individuals from the current population and uses them as parents to generate children for the next generation. After several generations, the population "evolves" towards a global optimal solution. Genetic algorithms are mainly used in three aspects in evolving neural networks: optimizing weights,^{101, 106, 107} optimizing network architectures,¹⁰⁸⁻¹¹² and optimizing optimizers.^{113, 114} In addition, genetic algorithms are also used to select proper input variables for neural networks from a high-dimensional space of raw data.¹¹⁵

3.2.3 Steps to Develop the Neural Network Model Using a Genetic Algorithm

I use a genetic algorithm to evolve the architecture and hyper parameters of neural networks for predicting ecotoxicity of chemicals. My objective is to find the best combination of four parameters: number of hidden layers, number of neurons per hidden layer, activation function, and network optimizer (**Table 3-3**). A combination of the four parameters is defined as a genome. Other parameters, such as number of epochs and batch size are determined based on trial and error.

Table 3-3. Parameters and options in genomes.

Parameters	Options
Number of hidden layers	1, 2, 3, 4, 5
Number of neurons per hidden layer	32, 64, 128, 256, 512
Activation function	ReLU, ELU, TanH, Sigmoid, Hard_Sigmoid, SoftPlus, Linear
Network optimizer	RMSprop, Adam, SGD, Adagrad, Adadelata, AdaMax, Nadam

I use mean squared error (MSE) and correlation of determination (R^2) between observed data in experiments and predicted values to evaluate the model performance. MSE measures the average of the squares of the error (i.e., difference between the estimated ecotoxicity and the observed ecotoxicity). MSE is always non-negative and values closer to zero the better. Ranging from 0 to 1, R^2 measures how much of the variability in the observed ecotoxicity can be explained by the predicted ecotoxicity. The higher R^2 , the better the model is at predicting the ecotoxicity. During the evolving process, I calculate validation MSE as the criteria to select the parents of next generation genomes. After the evolving process, when the best genome has been selected, I use test MSE and test R^2 to measure its performance on the “unknown” data (test set).

As shown in **Figure 3-2**, I build the neural network model in the following three steps:

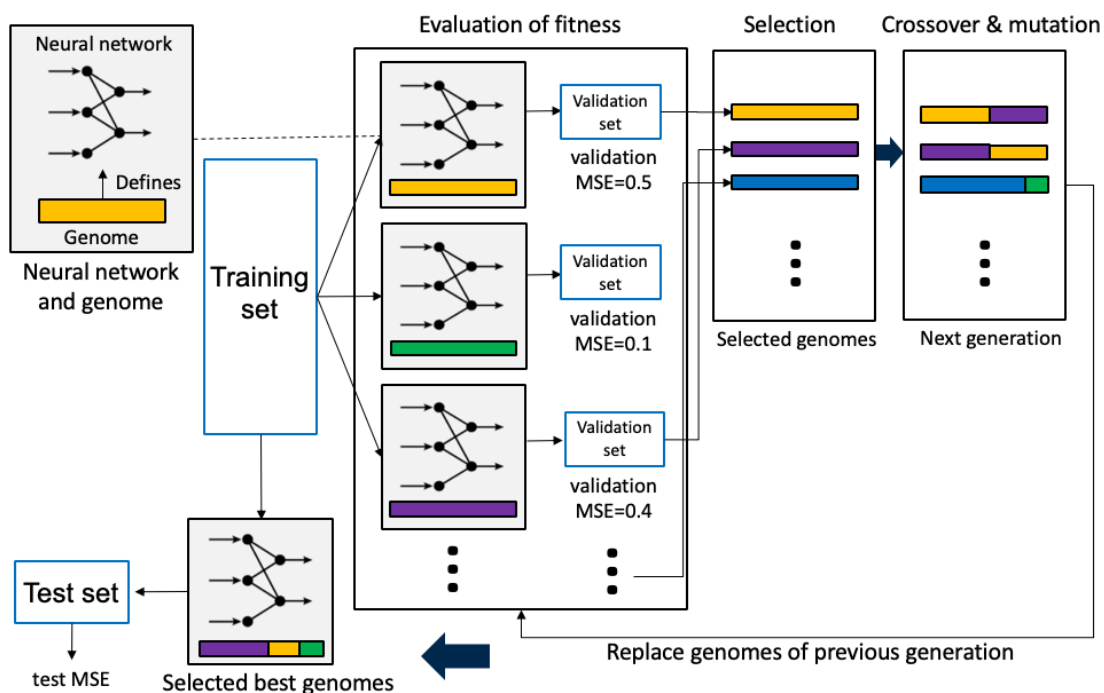


Figure 3-2. Steps of developing the neural network model using a genetic algorithm.

(1) Data splitting. Given that my goal is to build prediction models for new chemicals, I need to evaluate the model performance on new data. I randomly split the data set into a training set

(63%), a validation set (7%), and a test set (30%). The training set is used to train the model, the validation set is used to select the best genomes in the evolving process, and the test set is only used to test the selected model after the evolutionary process.

(2) Model selection using genetic algorithm. I use genetic algorithm to select best neural network configuration in the following steps:

- 1) Data preprocessing: I normalize the “known” data (training set and validation set) in the range of $[-1, 1]$ via min-max normalization. I also normalize the “unknown” data (test set) based on the minimum and maximum values of the “known” data for each variable. This is important for neural networks because unscaled input variables can result in slow or unstable learning process and unscaled output variables can cause the learning process to fail.¹¹⁶
- 2) Initialization: I create 30 neural networks with randomly generated genomes to be the population of the first generation.
- 3) Fitness evaluation: I train each network in the population and evaluate its performance on the validation set using validation MSE.
- 4) Selection: I rank all networks in the population by validated MSE and keep 20% of the top-ranked networks (6 networks) to become part of the next generation to breed children. Also, I randomly keep 10% (3 networks) of the rest of the networks. This helps find potentially successful combinations between worse-performers and top-performers, and also helps avoiding stuck in a local maximum.
- 5) Crossover: Crossover is the combination process from two members of a population to generate one or more children. Besides the top 20% networks and the randomly kept 10%

non-top networks, to keep our population of 30 networks, 21 children are generated for breeding in each generation.

- 6) Mutation: I randomly mutate some of the parameters on some of the kept networks.
- 7) Genome replacement: Genomes of the previous generation are replaced using the genomes after crossover and mutation.
- 8) Repeat: Step 3 to Step 7 are repeated for multiple generations until the model performance converges, i.e., the validation MSE will not get any better. The best performed genome in the final generation is the selected best neural network model.

(3) Model testing. The neural network model is trained again with the whole training and validation sets and the best nh identified from the previous step. The trained model is applied to the test set. The test MSE and test R^2 then evaluates the performance of the model. The lower the test MSE, the higher test R^2 , the better the model is at predicting the ecotoxicity for new chemicals.

In this process, each child is a combination of a random assortment of parameters from its parents. For instance, one child might have the same number of layers as its mother and the rest of its parameters from its father. A second child of the same parents may have the opposite. This is how genetic algorithms mirrors real-world biology and how it can lead to an optimized network quickly. In order to evaluate the average performance of the model, I repeat the above steps ten times with different splits of data to calculate the average test MSE and test R^2 , which also generate ten best models.

3.2.4 Model Performance Comparison

I first compare the performance of the neural network model with the performance of the ECOSAR model. Besides baseline toxicity, ECOSAR also calculates predicted toxicity. The

predicted toxicity is either calculated by multiplying the baseline toxicity by a toxicity reduction factor or by simple linear models with only one predictor variable (e.g., number of carbons in the molecular formula), depending on the chemical class. The performance of ECOSAR is measured by comparing predicted HC₅₀ from ECOSAR with the HC₅₀ based on experimental data in USEtox.

I also compare the neural network model with linear regression models: including ordinary least squares (OLS) regression, partial least squares (PLS) regression, and principle component regression (PCR) that I build based on the same data set. OLS chooses the parameters in a linear function by minimizing the sum of the squares of the difference between the observed and the predicted response variable by the linear function. PLS is a linear regression model by projecting the predictor variables and the response variable to a new space.⁴¹ PCR first reduces the original predictor variables to a small number of variables by principal component analysis and then uses the reduced variables in an OLS regression fit.¹¹⁷

For the OLS model, since no hyper parameters need to tune in the model, I use 10-fold cross-validation to build the model and test their average performance on the 10 folds. For PLS and PCR models, the number of components needs to be identified. I use 10-fold cross-validation on the training set and choose the best number of components, then use the test set to evaluate the model performance. In order to do a fair comparison, all the models are fitted on the same ten splits of data as for the neural network models.

3.2.5 Variable Importance

Since not all the input variables are equally important in the model, I evaluate the relative importance of the variables and identify the important input variables for the neural network model. One way to test the importance of a variable is to shuffle or permutate the variable and

see its impact on the model performance¹¹⁸. The procedure is first to get a benchmark test MSE by training the model once and then predict multiple times while randomizing each variable in the test set. The difference between the benchmark test MSE and the test MSE after permuting one variable, i.e., including and excluding this variable, is used as an importance measure, called permutation importance. If the test MSE after randomizing a variable is lower than the benchmark test MSE, it is an important variable. On the other hand, if nothing changes or the test MSE is higher than the benchmark, it is a useless variable.

In this study, I use shuffling procedure to find the most salient variables that can predict the ecotoxicity of chemicals. I randomize 500 times and get an average test MSE for each variable and compare with the benchmark test MSE.

3.3 Results and Discussion

3.3.1 Data Filtering and Exploratory Analysis

After data filtering, 2,308 out of 2,499 chemicals (92.4%) remain for building the neural network model (**Figure 3-3**). However, T.E.S.T and QikProp do not have molecular descriptors for all these chemicals. Therefore, I compile two datasets, one with more chemical samples but less variables, the other with more variables but less chemicals. Variables with constant values and duplicated variables are removed.

- Dataset1: 2,308 chemical samples with 12 variables including 11 physical-chemical properties in USEtox and the mode of action (MoA);
- Dataset2: 1,869 chemical samples with 695 variables with additional descriptors acquired from T.E.S.T. and QikProp.

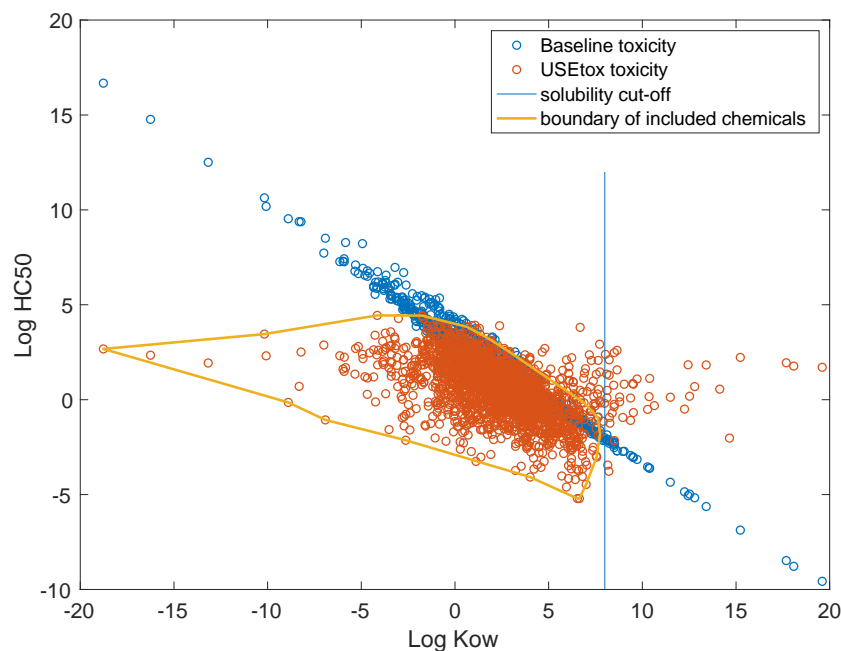


Figure 3-3. Data filtering.

(1) Exploratory analysis of dataset1. The exploratory analysis of dataset1 shows that the original data of the variables are highly skewed. Log-transforming the data can account for such skewed distributions and help making the data patterns more interpretable. The diagonal plots in **Figure 3-4** show the distribution of log-transformed variables. After the log-transformation, most of the variables show approximately normal distribution. Therefore, all raw data are log-transformed before building the neural network model and the linear regression models. The off-diagonal plots in **Figure 3-4** show the correlation between any two variables.

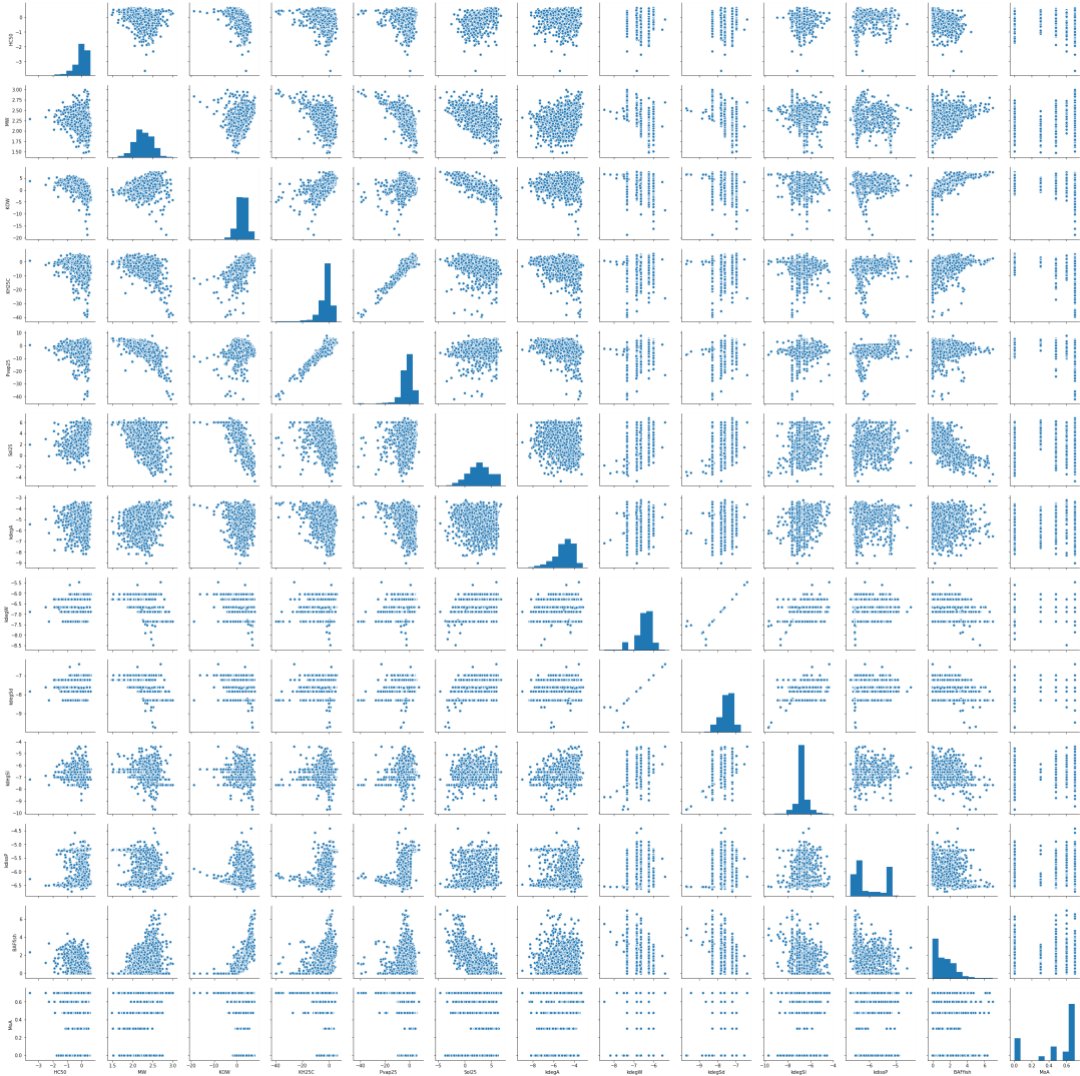


Figure 3-4. Pair plot of log-transformed input and output variables in dataset1.

Table 3-4 lists the calculated Pearson correlation coefficient (ρ) between any two variables. Most of the variables do not have strong correlation, except for kdegW and kdegSd ($\rho=0.991$) and Pvp25 and KH25C ($\rho=0.887$). The Pearson correlation coefficients between HC₅₀ and each variable in descending order of their absolute values are: Sol25 ($\rho=0.509$), BAFfish ($\rho=-0.489$), Kow ($\rho=-0.445$), MW ($\rho=-0.305$), kdegW ($\rho=0.269$), kdegSd ($\rho=0.269$), kdissP ($\rho=0.172$), KH25C ($\rho=-0.115$), MoA ($\rho=-0.068$), kdegSl ($\rho=0.066$), Pvp25 ($\rho=0.063$), kdegA ($\rho=-0.047$).

Table 3-4. Correlation coefficients of input and output variables in dataset1.

	HC50	MW	Kow	KH25C	Pvap25	Sol25	kdegA	kdegW	kdegSd	kdegSl	kdissP	BAFfish	MoA
HC50	1.000	-0.305	-0.445	-0.115	0.063	0.509	-0.047	0.269	0.269	0.066	0.172	-0.489	-0.068
MW	-0.305	1.000	0.361	-0.408	-0.706	-0.619	0.203	-0.625	-0.621	-0.189	-0.488	0.458	0.255
Kow	-0.445	0.361	1.000	0.495	0.117	-0.815	0.005	-0.395	-0.396	-0.170	-0.105	0.794	-0.177
KH25C	-0.115	-0.408	0.495	1.000	0.887	-0.228	-0.206	0.103	0.101	-0.012	0.387	0.319	-0.391
Pvap25	0.063	-0.706	0.117	0.887	1.000	0.232	-0.233	0.359	0.358	0.091	0.519	-0.056	-0.362
Sol25	0.509	-0.619	-0.815	-0.228	0.232	1.000	-0.063	0.548	0.550	0.224	0.268	-0.803	0.063
kdegA	-0.047	0.203	0.005	-0.206	-0.233	-0.063	1.000	0.095	0.096	0.165	-0.092	-0.008	0.157
kdegW	0.269	-0.625	-0.395	0.103	0.359	0.548	0.095	1.000	0.991	0.468	0.256	-0.467	-0.128
kdegSd	0.269	-0.621	-0.396	0.101	0.358	0.550	0.096	0.991	1.000	0.482	0.256	-0.468	-0.122
kdegSl	0.066	-0.189	-0.170	-0.012	0.091	0.224	0.165	0.468	0.482	1.000	0.144	-0.269	0.033
kdissP	0.172	-0.488	-0.105	0.387	0.519	0.268	-0.092	0.256	0.256	0.144	1.000	-0.179	-0.324
BAFfish	-0.489	0.458	0.794	0.319	-0.056	-0.803	-0.008	-0.467	-0.468	-0.269	-0.179	1.000	-0.192
MoA	-0.068	0.255	-0.177	-0.391	-0.362	0.063	0.157	-0.128	-0.122	0.033	-0.324	-0.192	1.000

(2) Exploratory analysis of dataset2. Due to the large number of variables in dataset2, I do not explore the distribution of each variable, only look at their correlations between any of the two variables. **Figure 3-5** shows the correlations of 695 variables in dataset2. A large portion of variables are positively correlated; a fair amount of them are independent with each other; and a small part of them are negatively correlated.

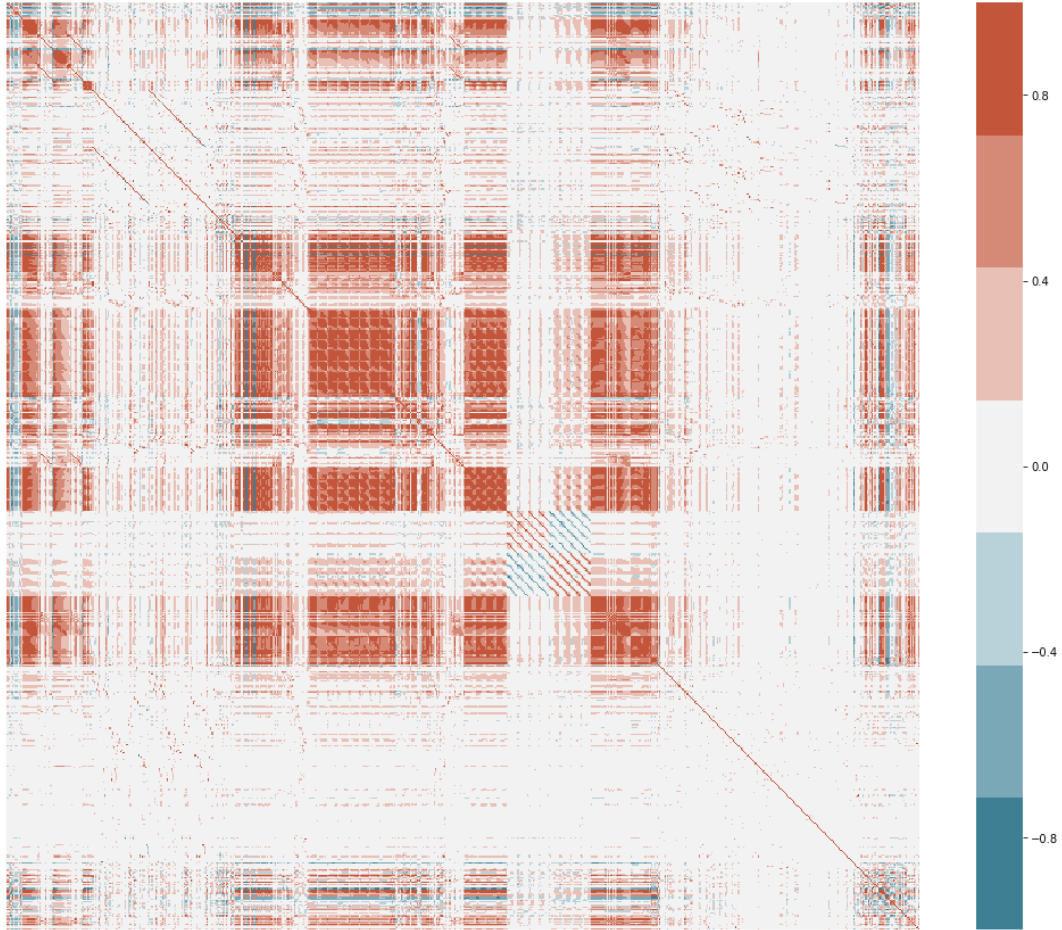


Figure 3-5. Correlation of variables in dataset2.

Figure 3-6 shows the empirical cumulative distribution function for dataset1 and dataset2. Each color represents the distribution of one variable. Most variables in dataset1 are distributed in a relatively narrow range; while the variables in dataset2 are distributed in a much broad range, differ by several orders of magnitude. Since the benchmark linear regression models are sensitive to the variable ranges, I transform the input variables in the range [0, 1] by sigmoid function in equation (3-3).

$$f(x) = \frac{1}{1+e^{-x}} \quad (3-3)$$

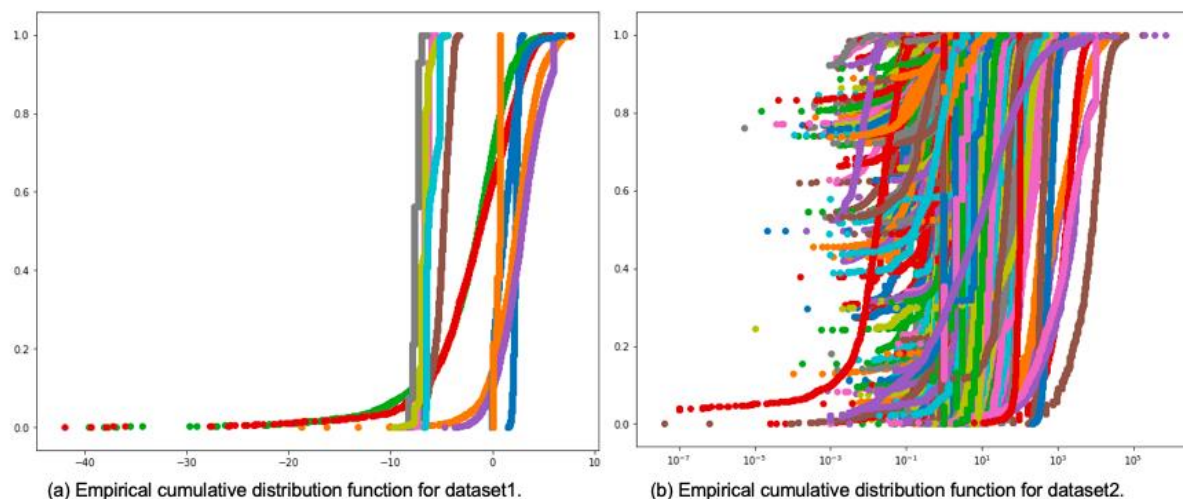


Figure 3-6. Empirical cumulative distribution function for dataset1 and dataset2.

3.3.2 Model Performance

Figure 3-7 shows the performance of the models that selected by genetic algorithm after eight generations on ten different splits of data. Neural networks for dataset2 have better performance (i.e., lower test MSE and higher test R^2) than networks for dataset1 because the additional chemical descriptor provides additional information to improve the model performance, although fewer chemical samples are used. In each subfigure, the models perform relatively stable on different splits of data. The training MSE is unsurprisingly lower than the validation MSE and test MSE. The validation MSE is the criteria the model is selected on validation data. The test MSE evaluates the model performance on the “unknown” data. The average test MSE of the ten splits of dataset1 is 0.037 with a standard deviation of 0.003. The average test MSE of the ten splits of dataset2 is 0.031 with a standard deviation of 0.005. Overall, the difference between the training, validation, and test MSE is trivial, which indicates no overfitting or underfitting occurs. Similar patterns can be observed for R^2 of the models. The

average test R^2 is 0.546 with a standard deviation of 0.03 for dataset1 and 0.628 with a standard deviation of 0.04 for dataset2.

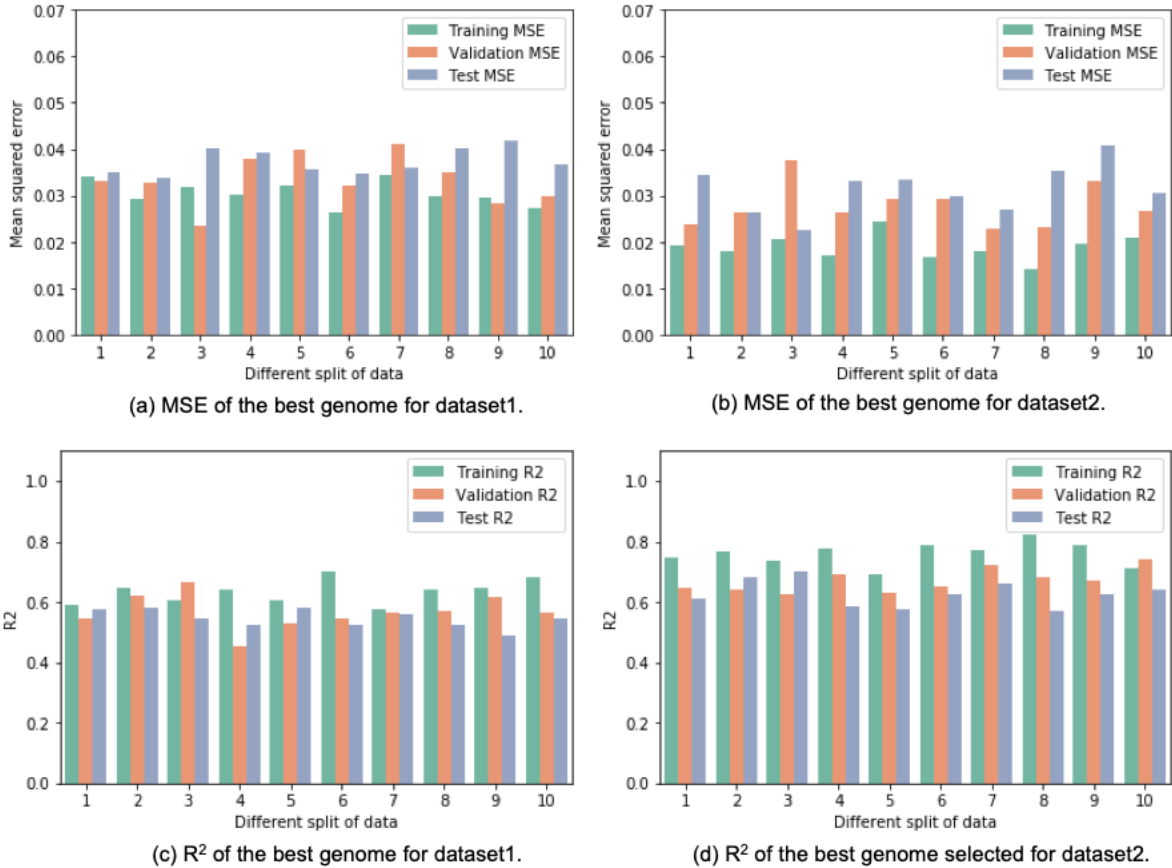


Figure 3-7. Model performance of the best genomes selected by genetic algorithm.

Table 3-5 and **Table 3-6** show the best genomes selected after eight generations by the genetic algorithm for 10 different splits of data and their performance on dataset1 and dataset2, respectively. For dataset1, the activation function selected by genetic algorithm is exclusively ReLU, which was found to achieve better results than other functions.^{119, 120} Almost all optimizers can achieve good performance on dataset1. This might be because dataset1 is a relatively small dataset, all kinds of optimizers can achieve good performance. For dataset2, the activation function selected are mostly Softplus and Sigmoid, which are the most widely used activation

functions. As a variation of Sigmoid, Hard_sigmoid on the third split has a substantial high test R^2 . This might be attributed to many factors (e.g., the random initialized weights and bias) besides activation function. The optimizer selected are mostly Adam and its variants, which have been proven effective for large dataset and high-dimensional parameters spaces.⁹⁷ On the other hand, the architecture selected are quite different. The reason I believe is that the architecture is not as important in the neural network design to some extent. Because no matter how the neural network is constructed, as long as the architecture is reasonable, the extra weights will be compressed toward zero by regularization techniques, and optimizers will always adjust the weights and the bias among the neurons to match the input and the output.

Table 3-5. Best genomes selected by genetic algorithm for dataset1 and their performance.

Splits	Best genomes selected by genetic algorithm	Training MSE	Training R^2	Validation MSE	Validation R^2	Test MSE	Test R^2
1	1, 64, ReLU, Adagrad	0.034	0.592	0.033	0.548	0.035	0.578
2	1, 256, ReLU, AdaMax	0.029	0.648	0.033	0.623	0.034	0.580
3	2, 32, ReLU, Adadelata	0.032	0.608	0.023	0.669	0.040	0.548
4	3, 64, ReLU, Adagrad	0.030	0.640	0.038	0.452	0.039	0.526
5	1, 32, ReLU, Nadam	0.032	0.605	0.040	0.529	0.036	0.581
6	4, 512, ReLU, Adagrad	0.026	0.703	0.032	0.547	0.035	0.525
7	1, 128, ReLU, SGD	0.034	0.577	0.041	0.566	0.036	0.562
8	5, 64, ReLU, Adam	0.030	0.640	0.035	0.571	0.040	0.526
9	2, 64, ReLU, Adam	0.029	0.648	0.028	0.617	0.042	0.490
10	2, 256, ReLU, Adagrad	0.027	0.681	0.030	0.565	0.037	0.543

Table 3-6. Best genomes selected by genetic algorithm for dataset2 and their performance.

Splits	Best genomes selected by genetic algorithm	Training MSE	Training R^2	Validation MSE	Validation R^2	Test MSE	Test R^2
1	2, 256, Sigmoid, Adam	0.019	0.746	0.024	0.647	0.034	0.610
2	1, 512, Softplus, Adam	0.018	0.770	0.026	0.643	0.026	0.682
3	1, 128, Hard_sigmoid, Nadam	0.021	0.736	0.038	0.625	0.023	0.704

4	1, 512, Softplus, Adam	0.017	0.778	0.026	0.692	0.033	0.588
5	2, 256, ELU, Adamax	0.024	0.693	0.029	0.631	0.033	0.573
6	1, 128, ReLU, Adadelta	0.017	0.787	0.029	0.654	0.030	0.628
7	1, 512, Sigmoid, Adam	0.018	0.773	0.023	0.720	0.027	0.659
8	1, 256, Sigmoid, RMSprop	0.014	0.821	0.023	0.682	0.035	0.571
9	1, 256, Softplus, Adam	0.019	0.789	0.033	0.670	0.041	0.628
10	1, 32, Softplus, RMSprop	0.021	0.715	0.027	0.742	0.031	0.641

3.3.3 Evolution of the Parameters

I analyze the evolutionary process of the four parameters along the eight generations for the first split of dataset2. From **Figure 3-8**, we can see that, at the first generation, different options are almost evenly selected. Along the evolutionary process, some options gradually show their advantage. This is particularly true for the evolution of the number of hidden layers and the number of hidden neurons. Two layers increasingly becomes favorable, even dominant in the selected number of hidden layers after the second generation. 128 is increasingly become the most selected number of hidden neurons. ReLU and Sigmoid are the most favorable activation functions. For optimizer, Adam is the most favorable option.

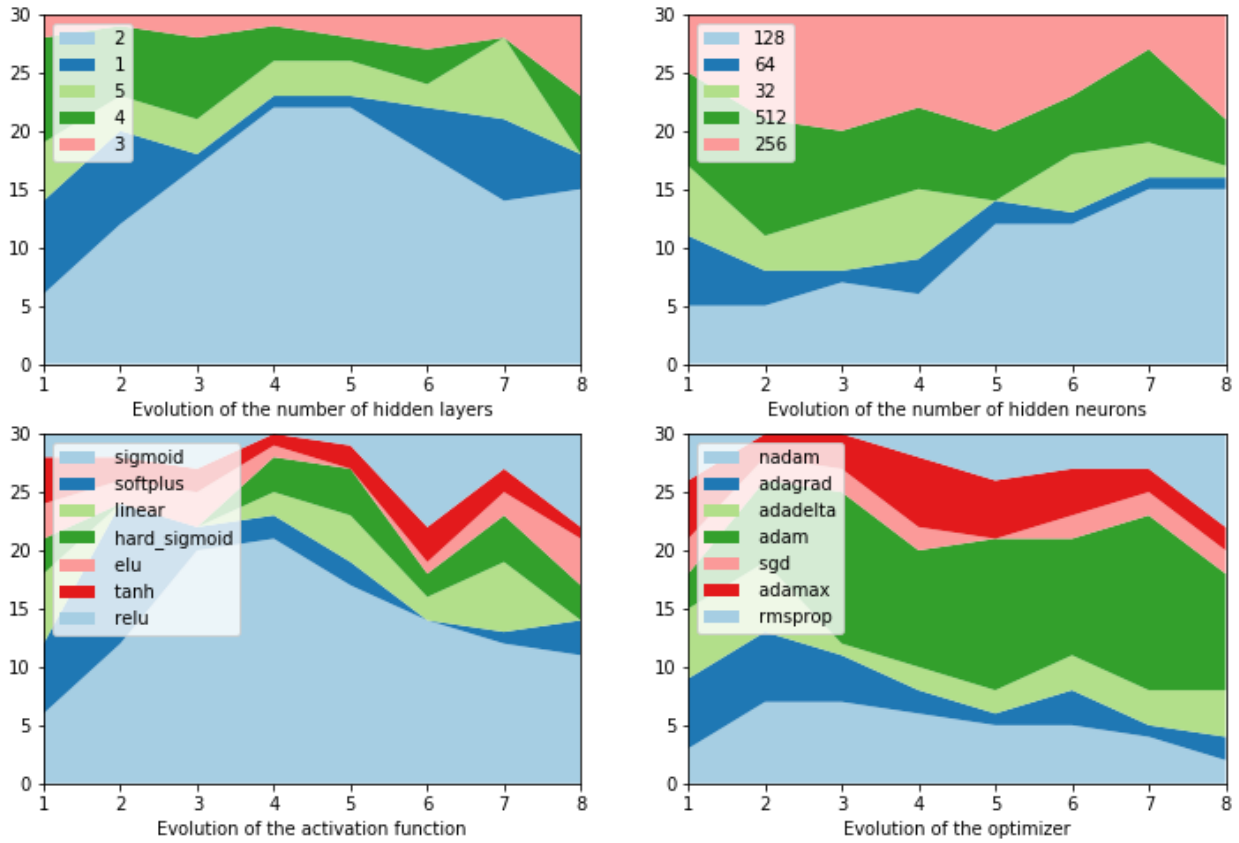
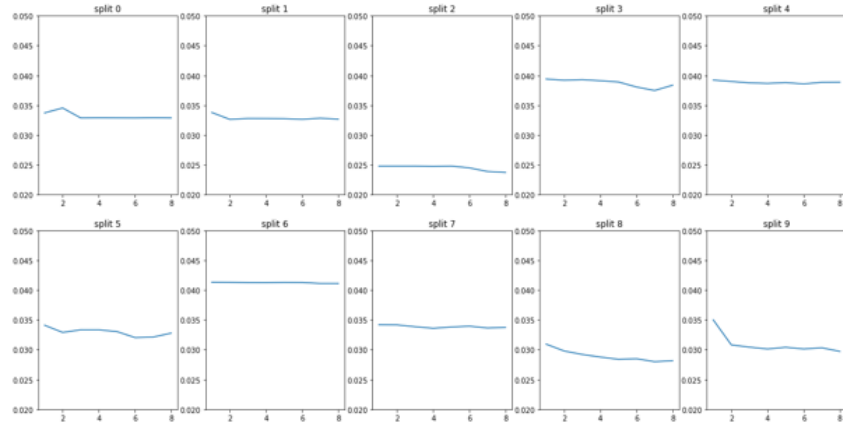
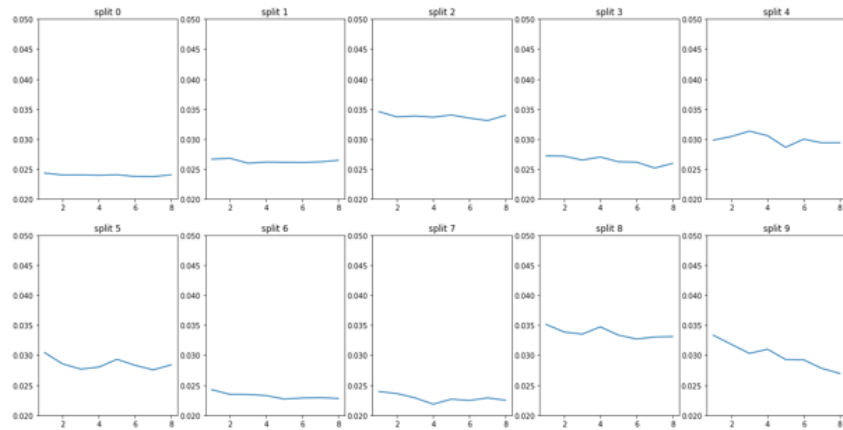


Figure 3-8. Evolution of the four parameters in neural network models.

The number of generations is set for eight because experiments show that the results generally converge after eight generations (**Figure 3-9**).



(a) Fitness of the best model along eight generations of dataset1.



(b) Fitness of the best model along eight generations of dataset2.

Figure 3-9. Fitness (i.e., validation MSE) of the best model along eight generations.

3.3.4 Performance Comparison with Grid Search

To show the advantage of the genetic algorithm, I implement grid search method as a benchmark to compare with the genetic algorithm. Grid search tests the performance of all the possible combinations of network parameters to identify the best one. **Table 3-7** and **Table 3-8** show the best genomes selected by genetic algorithm and grid search for the 10 different splits of dataset1 and dataset2 and their validation MSE (i.e., the criteria of selection).

Table 3-7. Best genomes selected by genetic algorithm and grid search for dataset1.

Splits	Best genomes selected by genetic algorithm	Validation MSE	Best genomes selected by grid search	Validation MSE
--------	--	----------------	--------------------------------------	----------------

1	1, 64, ReLU, Adagrad	0.033	1, 32, ReLU, Nadam	0.032
2	1, 256, ReLU, AdaMax	0.033	5, 512, ReLU, AdaMax	0.033
3	2, 32, ReLU, Adadelata	0.023	1, 512, ReLU, Adagrad	0.024
4	3, 64, ReLU, Adagrad	0.038	2, 256, ReLU, Adadelata	0.038
5	1, 32, ReLU, Nadam	0.040	3, 512, ReLU, Adagrad	0.038
6	4, 512, ReLU, Adagrad	0.032	4, 128, ReLU, Nadam	0.032
7	1, 128, ReLU, SGD	0.041	1, 128, ReLU, Adagrad	0.040
8	5, 64, ReLU, Adam	0.035	1, 64, ReLU, AdaMax	0.033
9	2, 64, ReLU, Adam	0.028	1, 256, ReLU, Adagrad	0.029
10	2, 256, ReLU, Adagrad	0.030	2, 64, ReLU, Adam	0.030

Table 3-8. Best genomes selected by genetic algorithm and grid search for dataset2.

Splits	Best genomes selected by genetic algorithm	Validation MSE	Best genomes selected by grid search	Validation MSE
1	2, 256, Sigmoid, Adam	0.024	1, 32, Hard_sigmoid, Adadelata	0.023
2	1, 512, Softplus, Adam	0.026	2, 128, Softplus, Nadam	0.025
3	1, 128, Hard_sigmoid, Nadam	0.038	1, 128, Sigmoid, RMSprop	0.034
4	1, 512, Softplus, Adam	0.026	1, 256, ReLU, Adam	0.025
5	2, 256, ELU, Adamax	0.029	1, 512, ELU, Adamax	0.029
6	1, 128, ReLU, Adadelata	0.029	1, 512, Tanh, Adagrad	0.027
7	1, 512, Sigmoid, Adam	0.023	1, 256, Tanh, Adam	0.023
8	1, 256, Sigmoid, RMSprop	0.023	1, 128, Softplus, Adadelata	0.022
9	1, 256, Softplus, Adam	0.033	1, 128, Linear, Adam	0.032
10	1, 32, Softplus, RMSprop	0.027	1, 512, Sigmoid, Adagrad	0.029

Results show the genomes selected by genetic algorithm and grid search are quite different. This is because the optimization of neural network is a nonconvex problem.¹²¹ There are more than one best solution for this problem; several different options of network architectures, activation functions and optimizers are able to achieve good performance on the two datasets. Although the genomes selected are not exactly the same, the validation MSE are very close

between the two methods. This proves genetic algorithm is able to find one of the best models through the evolutionary process.

Noted that the genetic algorithm selected models can sometimes be better than the one selected by grid search. This is due to the randomness in the training process. Neural networks and genetic algorithm are both stochastic, which means they make use of randomness (e.g., random weights initialized in neural network, population random generated in genetic algorithm) and therefore each time can produce different results. The traditional and practical way to address this problem is to run network several times and use statistics to summarize the performance of the model. Here, I repeat the training 10 times to get an average evaluation metric of the models.

3.3.5 Performance Comparison with the ECOSAR Model and Linear Regression Models

(1) ECOSAR model. ECOSAR predicts toxicity for different species for each chemical. I aggregate them into HC_{50} as required in LCA. The calculated test MSE of the ECOSAR model is 0.128. The calculated test R^2 is 0.218, which is close to 0.13 calculated by Melnikov *et al.* (2016).³⁹ Here, the HC_{50} are preprocessed in to range [-1, 1] using the same procedure for neural network for a fair comparison. Neural network models have a much better performance (i.e., test MSE is 0.031 and test R^2 is 0.628 for dataset2) than ECOSAR model.

(2) Linear regression models. The exploratory analysis for dataset1 shows that Pvp25 and KH25C and kdegW and kdegSd are highly correlated with each other, respectively. Such multicollinearity makes the coefficients in OLS models unstable and difficult to interpret. I calculate the variance inflation factor (VIF) of each variable, which indicates the extent to which multicollinearity is present in a regression analysis. I remove the variable with the highest VIF each time until the VIF of the remaining variables are down to an acceptable range (a rule of

thumb commonly used is when VIF is less than 10). As a result, Pvp25 and kdegSd are removed from this process (**Table 3-9**).

Table 3-9. Remove multicollinearity by calculating variance inflation factor (VIF).

Variables	VIF in step 1	VIF in step 2	VIF in step 3
MW	4.45	4.16	4.16
Kow	4.93	4.92	4.92
KH25C	73.15	3.16	3.16
Pvp25	77.48 (remove)		
Sol25	18.4	5.47	5.44
kdegA	1.18	1.18	1.18
kdegW	55.43	55.17	2.35
kdegSd	56.06	55.83 (remove)	
kdegSl	1.41	1.41	1.37
kdissP	1.54	1.52	1.52
BAFfish	3.64	3.64	3.63
MoA	1.33	1.33	1.33

I use the remaining ten variables to build the OLS model. I include all the twelve variables in the PLS and PCR models because they resolve multicollinearity issue by constructing latent independent variables underlying the collinear variables. Similarly, neural network models are also unsusceptible to the multicollinearity.¹²² Due to the high multicollinearity in dataset2, I only use PLS and PCR models to fit dataset2.

Figure 3-10 shows performance of the linear models for dataset1 and dataset2. Similar with neural network models, the linear models for dataset2 perform better (i.e., lower test MSE and higher test R^2) than models for dataset1 due to the additional variables provided by dataset2. In each subfigure, the green boxplots and yellow boxplots respectively show the distribution of training MSE (or test R^2) and test MSE (or test R^2) on the 10 different splits of data. They all on the same level, meaning no overfitting or underfitting occurs.

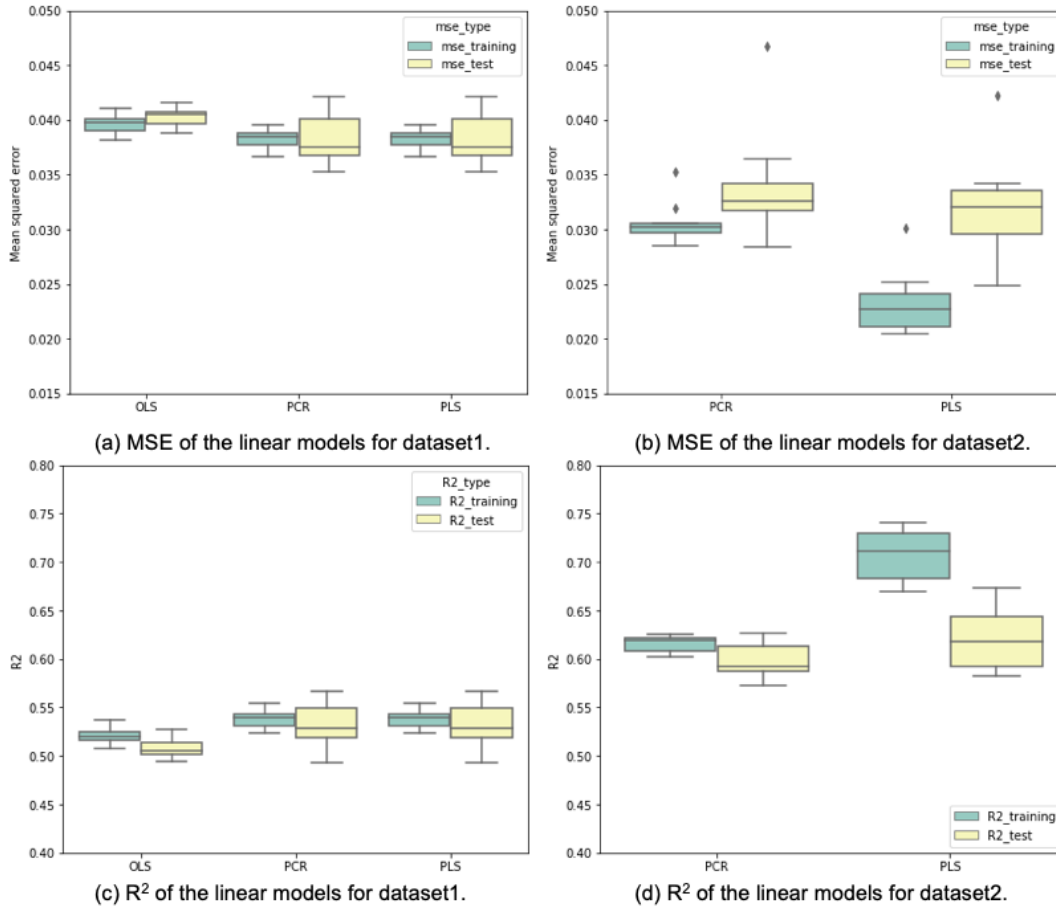


Figure 3-10. Performance of the linear regression models.

Table 3-10 lists the test MSE and test R^2 of the linear models and neural network models. **Figure 3-11** visualizes performance comparison of the models. PCR and PLS models perform same on dataset1, both better than OLS model. For dataset2, PLS performs better than PCR model. The best test MSE on dataset1 is 0.038 and best test R^2 is 0.532. The best test MSE on dataset2 is 0.032 and best test R^2 is 0.619. Overall, they all perform not as good as the neural network models.

Table 3-10. Performance comparison of neural network models and linear regression models.

Regression models	Dataset1	Dataset2
-------------------	----------	----------

	Average test MSE	Average test R^2	Average test MSE	Average test R^2
Ordinary least square	0.040 (0.001)	0.507 (0.01)	NA	NA
Principle component regression (PCR)	0.038 (0.002)	0.532 (0.02)	0.034 (0.005)	0.598 (0.02)
Partial least squares (PLS)	0.038 (0.002)	0.532 (0.02)	0.032 (0.005)	0.619 (0.03)
Neural networks	0.037 (0.003)	0.546 (0.03)	0.031 (0.005)	0.628 (0.04)

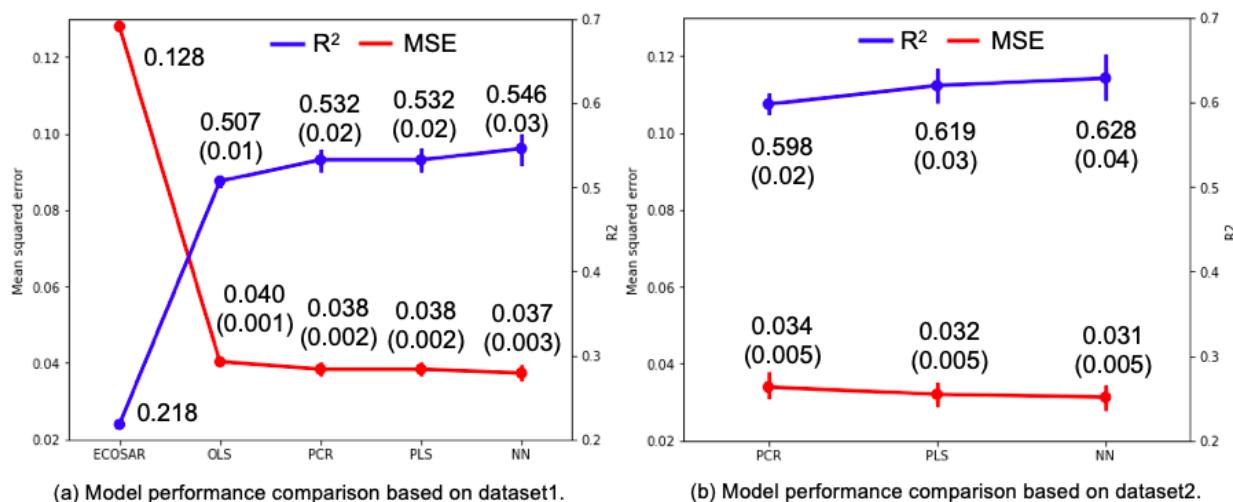


Figure 3-11. Performance of the neural network models compared with other models.

One the other hand, if we look at the standard deviation of the results (as shown in the parenthesis in **Table 3-10** and **Figure 3-11**), neural network models have the largest standard deviation. The reason is that neural networks are generally overparameterized and the optimization process is nonconvex and unstable.¹²¹ There are a lot randomness involved in the training process. Even with exactly the same parameters, each time neural network models can give different results.

3.3.6 Variable Importance via Shuffling Procedure

The results show that the neural network model has better prediction performance than ECOSAR and linear regression models for estimating ecotoxicity of chemicals using USEtox

data. However, the neural network model is often regarded as a “black box” and the weights attached to variables are hard to interpret due to its nonlinear functions.¹²³ Here I try to find reasonable explanations for the relative importance of variables in predicting ecotoxicity of chemicals based on the shuffling procedure. I mentioned before that neural network is unsusceptible to the multicollinearity. This is true for prediction purpose, but when it comes to variable selection, highly correlated variables will affect each other and lead to incorrect conclusion. Since many variables in dataset2 are highly correlated, here I only analyze dataset1 and remove the two variables with high VIF (i.e., Pvp25 and kdegSd) in order to perform the shuffling procedure.

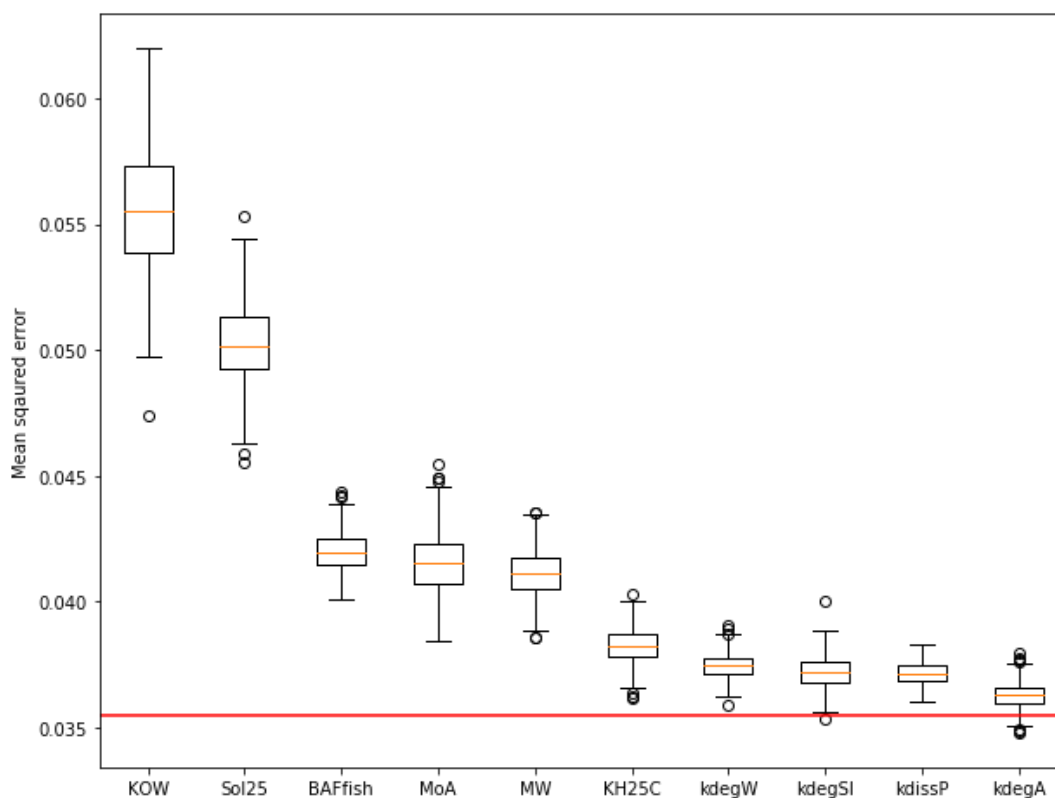


Figure 3-12. Test MSE when each variable is shuffled.

The permutation importance of variables is shown in **Figure 3-12**. The red horizontal line indicates the benchmark test MSE (0.036) when no variable is shuffled. Each box indicates the

distribution of the test MSE of each variable being shuffled 500 times. Overall, the average test MSE of all the variables are less than the benchmark test MSE, which means all the variables contribute to the model to a certain degree. But some are more important than others. The order of the importance is Kow, Sol25, BAFfish, MoA, MW, KH25C, kdegW, kdegSl, kdissP and kdegA.

The two most important variables are Kow and Sol25. Kow (octanol-water partitioning coefficient) is the ratio of a chemical's concentration in octanol (representing lipid "fat" in general) compared to water, measuring a chemical's affinity for the lipid portions of an organism's tissues. Kow has a relatively high negative correlation with HC_{50} ($\rho=-0.493$). Higher Kow indicates higher hydrophobicity, thus lower toxicity. Chemicals with Kow larger than the cut-off value are removed in this study. Sol25 (solubility at 25°C) has the highest correlation coefficient ($\rho=0.548$) with HC_{50} . Water solubility of a chemical influences its fate and transport in all environmental media and is especially relevant to exposure via aquatic pathways. Soluble chemicals are more available for traveling with water and for chemical and biological transformations.¹²⁴ This explains why higher Sol25 is generally associated with higher ecotoxicity.

BAFfish, MoA, and MW are the next three relatively important variables. BAFfish is the bioaccumulation factor of continental and global fresh and sea water fish. BAFfish also has a relatively high negative correlation coefficient with HC_{50} ($\rho=-0.582$). This is probably because BAFfish is calculated as the ratio of a chemical's concentration in fish to its concentration in the water, thus is relatively highly correlated with Kow ($\rho=0.746$). Since HC_{50} is also correlated with Kow ($\rho=-0.578$), BAFfish and HC_{50} are correlated. MoA is a categorical variable calculated by Verhaar scheme, which classify the chemicals into five categories: class 1 (inert chemicals),

class 2 (less inert chemicals), class 3 (reactive chemicals), class 4 (chemicals acting by a specific mechanism), and class 5 (unclassifiable chemicals). From class 1 to class 4, the toxicity of chemicals is increasing, thus an important variable for toxicity estimation. MW (molecular weight) is negatively associated with Sol25 ($\rho=-0.601$), which means the higher the MW, the lower the Sol25, and the lower the toxicity.

KH25C, kdegW, kdegSI, kdissP, and kdegA show relative low importance in predicting the ecotoxicity of chemicals. KH25C is weakly correlated with HC₅₀ ($\rho=-0.134$), indicating the transfer of chemicals from soil and water to air through volatilization. kdegSd is excluded from the OLS model due to the high collinearity with kdegW. kdegW, kdegSI, kdegA, and kdegSd represent the rates of degradation for chemicals in water, soil, air, and sediment, respectively. The higher the degradation rate, the shorter the persistence of the chemical. These parameters are more relevant to the persistence of chemicals in the environment than to the ecotoxicological effect on the environment. kdissP is the dissipation rate in above-ground plant tissues which can be regarded as the degradation of chemicals in plants. Since ecotoxicity in USEtox is currently based on freshwater toxicity only without consideration of terrestrial toxicity, kdissP is not significant either.

3.3.7 Computational Time

The code is programmed in Python and run on Flux. The computing for different splits of data is independent and can run simultaneously. The neural networks in each generation are also independent and therefore can parallel the computing. I use 30 processors to train the 30 neural network models in each generation. For ten splits of data, I use 300 processors at the same time to run the grid search code and genetic algorithm code, respectively.

For dataset1, for each split of the data, the grid search takes 2 hours 38 minutes 12 seconds to run all 1,225 parameter combinations, which correspond to training 1,225 neural networks. Conversely, the genetic algorithm only takes 26 minutes 5 seconds, because it only trains 240 networks (30 networks for each of the eight generations). For dataset2, the grid search takes 3 hours and 7 minutes 6 seconds, and the genetic algorithm only takes 29 minutes 2 seconds. The genetic algorithm gives similar results in approximately 16% time, providing exponential speed benefit.

3.3.8 Drawbacks and Advantages of Neural Networks

Neural networks have some drawbacks: (1) They do not have an explicit mathematical function; therefore, it is hard to interpret the reasons that some chemicals have high predicted toxicity and others not. When applied in LCA studies, the predicted characterization factors can first be used to calculate preliminary results, with which the chemicals with high impact in a product's life cycle can be identified. LCA practitioner can then look for more accurate data or conduct tests for the identified chemicals to improve the data quality of the LCA studies.

(2) Their structure is very flexible (i.e., it can have any number of hidden layers and any number of neurons in each layer) and has many variations (e.g., activation function). It could take a long time to try different options or require other algorithms to optimize the structure (e.g., genetic algorithm). Here I use a genetic algorithm to optimize the structure, which generate comparable networks derived by the grid search method, but only uses 16% time. The results demonstrate the ability of the genetic algorithm to efficiently design a neural network that generates desired performance. The genetic algorithm is inherent parallelized by evaluating multiple choices simultaneously. Through parallel computing, it can solve problems even more efficiently. This is especially valuable when the size of the training data is large. Although

convenient to use, genetic algorithm itself requires certain number of parameters to configure, such as population size, selection rate, crossover rate, and mutation rate. A small sized population will not give enough solution space to generate accurate results. An out of ranged selection rate, crossover rate, or mutation rate will disrupt the selection process. However, research shows that, as long as these parameters are in a reasonable scope, they have limited influence on the outcome.¹²⁵⁻¹²⁷ In this study, the population size is 30, selection rate is 0.3 (i.e., select 20% top genomes and 10% non-top genomes), the crossover rate is 0.7 (i.e., breed 70% children in each generation from the crossover), and the mutation rate is 0.3 (i.e., 30% of the population are randomly mutated after selection and crossover). Another key parameter of genetic algorithms is the fitness (evaluation) function. A wrong choice of the fitness function may lead to problems such as unable to find the solution or returning incorrect results. Here, the fitness function is validation MSE, which evaluate the performance of the model on validation set. Results show this configuration is good enough to outperform grid search method and it is unnecessary to fine tune the parameters in the genetic algorithm furthermore.

(3) Overfitting. A common problem of neural network training is overfitting. When overfitting occurs, the error on the training set is driven to be very small, but the error on the test set is relatively large. There are several ways to avoid overfitting and improve generalization. First is early stopping, which separates out a validation data set to monitor the training process, and stops the training when validation error increases for a specified number of iterations. My experiments show early stopping may stop too early and underfit the data. Another method for improving generalization is regularization. One often used approach is weight regularization, which introduce a penalty to the loss function when training a neural network to encourage the network to use smaller weights. Smaller weights in a neural network can result in a model that is

more stable and less likely to overfit the training data, in turn have better performance when predicting on new data. Unlike weight regularization encourage small weights, weight constraint forces weights to be small and can improve generalization when used in conjunction with other regularization methods like dropout.¹²⁸ Dropout means temporality removing some nodes from the network, forcing each node within a layer to take on more or less responsibility for the inputs. In this study, I use both weight constraint and dropout to prevent overfitting. **Figure 3-13** shows the performance of the model when use different dropout rate and different weight constraint. As a result, I use weights constraint as 1, dropout rate as 0.1 for networks training on dataset1; and weights constraint as 2, dropout rate as 0.4 for networks training on dataset2.

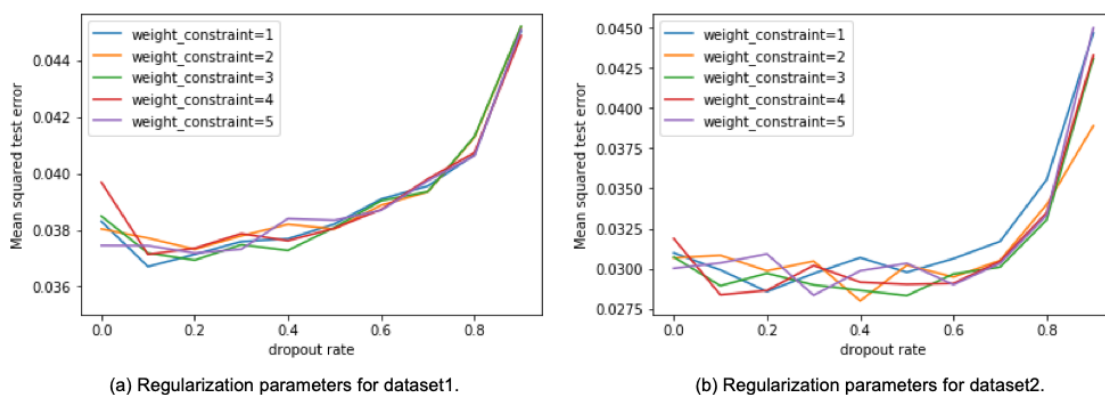


Figure 3-13. Regularization parameters for preventing overfitting.

On the other hand, neural networks have the following advantages: (1) Fast and less costly. Once trained, the prediction is pretty fast, so we can calculate characterization factors for a larger number of chemicals. (2) Good performance. As shown by the results, the neural network model outperforms other QSAR models. The performance can be further improved with more data, either toxicity data or relevant predictor variables. (3) Wide applicability. Neural networks can also be used to predict characterization factors for other environmental impacts in LCA given the existing data and relevant predictor variables.

3.4 Summary

In this chapter, I develop a neural network model to predict chemical ecotoxicity. The model is trained using experiment-based ecotoxicity data of approximately 2,000 chemicals from USEtox and chemical descriptors from different sources. I use a genetic algorithm to significantly reduce the computational time required to reach the optimal configuration for the neural network model. By using genetic algorithms, neural networks with comparable performance can be found in 16% time compared with the widely used grid search method, which finds the best parameter combination by calculating and comparing all possible ones. The developed neural network models outperform widely used ECOSAR model and also better than the linear regression models. The generated neural network model can be used to fill the data gap in LCA for toxicity evaluation of any product's life cycle.

More broadly, the neural network model can be used in chemical risk assessment to predict the ecotoxicity of chemicals for which experimental data are not available. Genetic algorithms can help rapidly predict the ecotoxicity of chemicals to help understand the potential risk of chemicals and develop strategies for chemical risk management.

Chapter 4 Estimate Ecotoxicity Characterization Factors for Chemicals Using Random Forest Models

4.1 Introduction

In chapter 3, I use neural networks to estimate HC_{50} and CF_{eco} for chemicals in USEtox. The results show on average neural networks performs better than ECOSAR and linear regression models. However, the standard deviation of the results is larger than the linear models. The main reason I believe is that I only use datasets with approximately 2,000 data samples and use a basic fully connected feedforward neural network. There have been many advanced neural networks, such as associative neural networks (ASNN), multitask deep neural network (MT-DNN), which have achieved top positions in the toxicity prediction challenge issued by National Institutes of Health Tox21 program¹²⁹ and US Environment Protection Agency ToxCast.¹³⁰ The success of these advanced neural networks is not only due to the use of novel algorithms and architectures, but also to the availability of high-performance computers (e.g., GPU) and large datasets. For example, a study proves that multitask networks can obtain predictive accuracy significantly better than single-task networks by using 40 million data points.¹³¹ However, in the current LCIA methods, we do not have that many data, which limits the application of neural networks in LCIA.

Another reason is that neural networks in nature are not good at handling “mixed” types of data.¹²¹ In fact, the chemicals’ descriptors are usually messy: the input variables tend to be mixtures of numerical and categorical variables, and measured on very different scales as shown

in **Figure 3-6**. While tree-based models are more robust when dealing with such kind of data. One of the tree-based machine learning methods that actively used in toxicity prediction is random forests,¹³² which are currently considered as one of the best approaches to build predictive models. The main advantages of random forests are their high predictive performance, high computational efficiency, and the ease of use because only a few model settings need to be configured.¹³³ Random forests generate an ensemble of decision trees. Each tree is a predictive model that uses input variables as predictors of outcome values, using a random subset of the training data. By limiting the number of variables used in each tree, random forests average many noisy but approximately unbiased trees, and hence reduce the prediction variance and get a more accurate and stable prediction.¹³⁴ Svetnik *et al.* pointed out that random forests are uniquely suited for modeling in cheminformatics because they can deal with a large number of (and all kinds of) descriptors simultaneously, handle redundant descriptors, and incorporate interactions and multiple mechanisms of actions.¹³² Polishchuk *et al.* applied random forests in predicting aquatic toxicity and have a better performance than corresponding partial least squares and k nearest neighbor models.¹³⁵ Recently, Li *et al.* used random forest to predict carcinogenicity of polycyclic aromatic hydrocarbons and found random forests outperform partial least squares and neural networks.¹³⁶

In this chapter, I aim to provide missing HC_{50} and CF_{eco} for chemicals in USEtox by random forest models. To evaluate the performance of the model, I compare its performance with those of neural network models in chapter 3.

4.2 Data and Methods

4.2.1 Data

I use the same two datasets used for random forest models in chapter 3 to develop the neural network models.

- Dataset1: 2,308 chemical samples with 12 variables including 11 physical-chemical properties in USEtox and the mode of action (MoA);
- Dataset2: 1,869 chemical samples with 695 variables with additional descriptors acquired from T.E.S.T. and QikProp.

4.2.2 Random Forests

Random forests are developed based on decision trees. Decision trees are used to model complex relationships. Although accurate, they often overfit the noise in the data and show a large degree of variability among different data samples from the same dataset. As a result, decision trees are known for giving high variance and low bias results. The objective of random forests is to take a group of high-variance, low-bias decision trees and transform them into a model that has both low variance and low bias. To achieve this, random forests introduce randomness to each generated decision tree at two levels. First, a certain proportion of training samples are randomly selected from the total sample set for each of the decision trees. Second, when the node of the decision tree is bifurcated, a feature subset is selected randomly from the feature set, and then the optimal feature is selected from the feature subset for bifurcation. The two levels of randomness increase the diversity of decision trees. As a result, by averaging the outputs of individual trees, random forests reduce the variance and have better predictive performance.

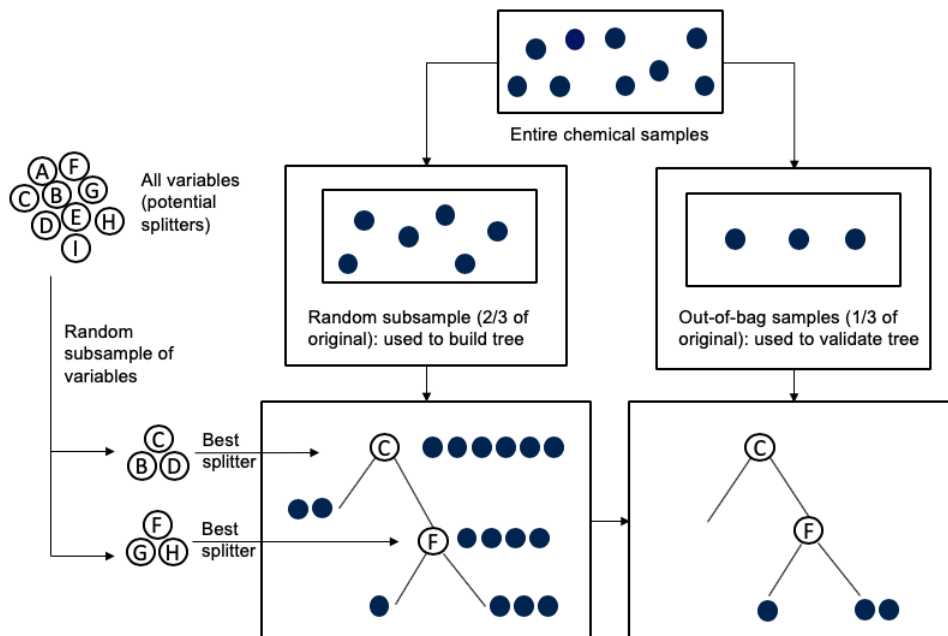


Figure 4-1. Procedure of constructing a tree in random forests.

In a random forest, each tree is constructed using the following procedure (**Figure 4-1**):

- (1) Choose a training set for this tree with replacement from all available training samples (i.e., take a bootstrap sample);
- (2) From the top of the tree, for each node, randomly choose a certain number of variables from all available variables. Pick the best variable among the selected variables to split the node into two nodes and continue to grow until it meets the criteria to stop;
- (3) Use the rest of the samples (i.e., out-of-bag samples) to estimate the error of the tree, which is called OOB error.

When making a prediction, a new sample is pushed down the tree. It is assigned the average outcome values of the training samples in the final node it ends up in. This procedure is iterated over all trees in the ensemble, and the average prediction of all trees is reported as random forest prediction. Note that this is in the case of regression since predicting toxicity

values is a regression problem. Random forests are also able to solve classification problems by taking the majority vote among the trees.

The two most important hyperparameters in random forest models are *n_estimators* and *max_features*. *n_estimators* is the number of trees in the forest. A higher number of trees generally improves the performance and produces stable predictions, but it also increases the computational time. *max_features* is the maximum number of features random forest considers to split a node. The other parameters include the *max_depth* (the maximum depth of the tree), *min_samples_leaf* (the minimum number of samples required to be at a leaf node), and *min_samples_split* (the minimum number of samples required to split an internal node). These parameters are used to control the size of the tree. If set *max_depth* = None, *min_samples_leaf* = 1, and *min_samples_split* = 2, the nodes are expanded until all leaves contain one sample, which allows the trees to fully grow. Setting a larger number of the three parameters avoids overfitting and speed up the computation.

4.2.3 Steps to Develop the Random Forest Models

Similar as in chapter 3, I build the random forest models in the following three steps:

- (1) Data splitting. I randomly split the data into 70% for training and validation, and 30% for testing.
- (2) Model selection. To select the best number of *n_estimators* and *max_features*, I use the out-of-bag (OOB) error, which is an important feature of random forests. An OOB error is similar to the validation error obtained by cross-validation. Unlike many other nonlinear machine learning models, random forests can perform cross-validation error during the training process. In this way, the best parameter can be chosen based on OOB error. Once the OOB error stabilizes, the training can be stopped. To control the size of

the tree, I use $min_samples_leaf = 5$ for regression problems, following the recommendation of the inventors.¹²¹

- (3) Model test. The random forest model is trained again with the training set and the best parameters identified from the previous step. The trained model is applied to the test set. The test MSE and test R^2 then evaluate the performance of the model. The lower test MSE, the higher test R^2 , the better the model is at predicting the ecotoxicity for new chemicals. In order to ensure a stable performance of the model, I repeat step 1 to 3 ten times with different splits of data to calculate the average test MSE and test R^2 , which also generate ten groups of best $n_estimators$ and $max_features$.

4.2.4 Variable Importance

Random forests have another great quality that they can easily measure the relative importance of each feature on the prediction. At each split in each tree, the improvement on split-criterion is attributed to the splitting variable as its importance and is accumulated over all the trees in the forest respectively for each variable. In our case, the outcome is numeric values (i.e., regression problem), the split-criterion is the decrease of errors, which is calculated by the reduction in the sum of squared errors whenever a variable is chosen to split. Feature importance is then calculated as the decrease of errors weighted by the probability of reaching that node, which is the number of samples reaching the node divided by the total number of samples.

4.2.5 Uncertainty and Application Domain of Estimated HC_{50}

The outcome by the trained model is a point prediction, which contains some uncertainty that comes from the errors in the model and noise in the input data. I use two ways to evaluate the uncertainty of the estimated values. The first way is to give a confidence interval of the estimation. A confidence interval provides a range of model results and a probability that the

model results will fall between the ranges when making predictions on new data. A robust way to determine confidence intervals for machine learning models is to use bootstrap, a common technique that can be used to derive empirical confidence intervals.¹³⁷ The basic idea is to resample the original data many times and use them to train the model respectively, and then we can have many predictions that can produce reasonable approximate confidence intervals of the predicted values. I train the model 100 times based on resampled training data and provide bootstrap confidence intervals of the predicted values for 578 chemicals.

The second way is to define the application domain of the developed model. An application domain is required in a QSAR study to express the scope and limitations of a model to specify the range of chemical properties for which the model is applicable.¹³⁸ I use a distance-based method to define the application domain of our model. I first calculate a centroid of all the chemicals in the available data set based on their input variables. Then I calculate the distance from each chemical to the centroid. Finally, I identify a distance in which 90% of the chemicals are enclosed. This distance is defined as the application domain of the model. When applying this model on a new chemical, one first calculates the distance of the new chemical to the centroid and determines whether it is in or outside of the application domain. This also gives an estimation for the confidence of the predicted results.

4.3 Results and Discussion

4.3.1 Model Selection Results

I test different number of trees ($n_{estimators}$) from 100 to 1,000 and three different options of $max_features$: *None* means using all the features to split a node; *sqrt* means the square root of the total number of features, which is 3 for dataset1 and 26 for dataset2; $p/3$ means the one third of the total number of features, which is 4 for dataset1 and 231 for dataset2. **Figure 4-2**

shows the average OOB error rate from 10 splits of data with different options of $n_estimators$ and $max_features$. $max_features = p/3$ shows the best performance for both datasets, which is in line with the recommendation of the inventors for regression problems.¹²¹ For the number of trees, $n_estimators = 1000$ shows the best performance for both datasets, and the OOB error rate is converged, indicating further increasing the number of trees would not improve the performance too much. The OOB error rate is calculated by one minus OOB R^2 , which is the only metric provided by Python sklearn package to evaluate the performance of random forests on out-of-bag samples. Note that OOB error rate is not equal to OOB MSE, here I use OOB error rate only for model selection.

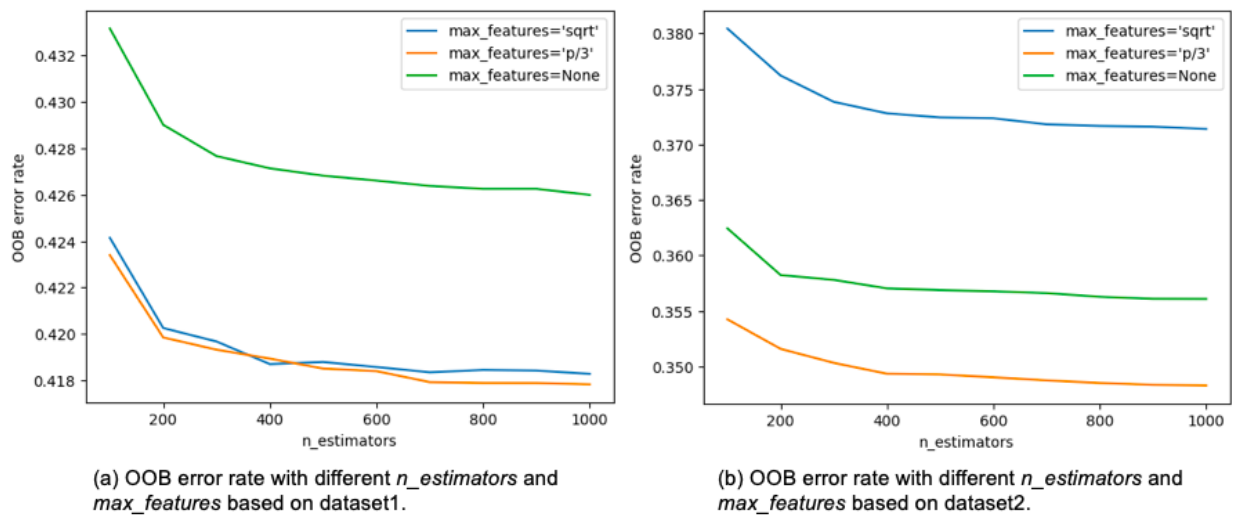


Figure 4-2. Average OOB error rate with different $n_estimators$ and $max_features$.

The performance order of $max_features = sqrt$ and $None$, however, are different for the two datasets. For dataset1, $max_features = sqrt$ has a better performance than using all features, which makes sense for random forest models because the essence of random forests is to use random subsets of features to decrease the variance. However, for dataset2, $max_features = None$ has a better performance, which means the model has a better performance when using

all the features than using the square root of the total number of features. This is because the square of 695 is 26, selecting 26 variables from 695 variables is highly likely to choose irrelevant variables that cannot predict HC_{50} ; therefore, using all features have a better performance.

4.3.2 Performance of Random Forest Models

Figure 4-3 shows the performance of the random forest models on the two datasets. On both datasets, random forest models perform much better on training set compared with the test set. Even though, the models still have a good performance on the test set. Again, the MSE for OOB set is not reported here because the package in Python only provides OOB R^2 as the score on out-of-bag samples. The average test MSE of the ten splits of dataset1 is 0.034, with a standard deviation of 0.002. The average test MSE of the ten splits of dataset2 is 0.029, with a standard deviation of 0.004. Similar patterns can be observed for R^2 of the models. The average test R^2 is 0.590, with a standard deviation of 0.01 for dataset1 and 0.657, with a standard deviation of 0.03 for dataset2.

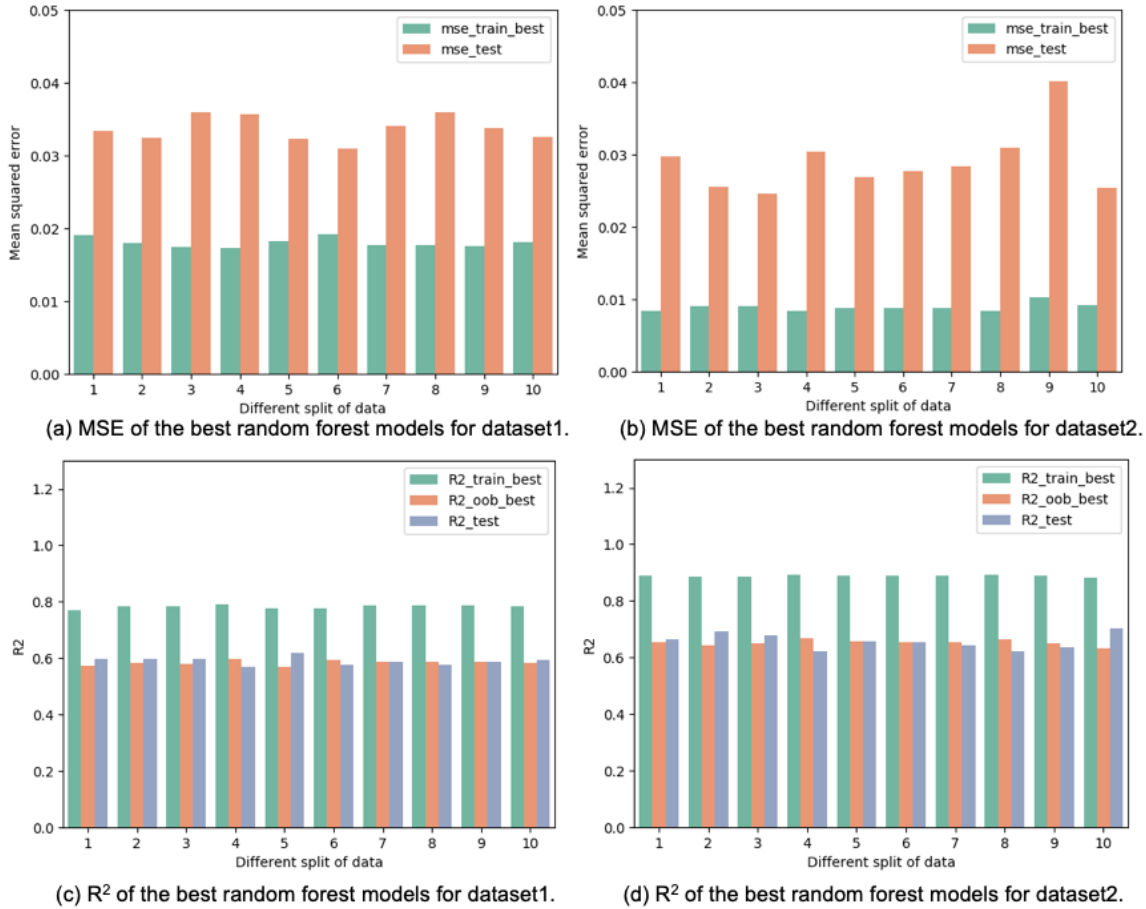


Figure 4-3. Performance of the random forest models on the two datasets.

Table 4-1 compares the performance of the random forest models developed in this chapter and the neural network models in chapter 3. Results show the random forest models perform better than neural network model. Moreover, random forests have a smaller standard deviation, meaning they are more stable compared with neural networks. **Figure 4-4** visualizes performance comparison between random forest models and other models.

Table 4-1. Performance comparison of random forest models and neural network models.

Models	Dataset1		Dataset2	
	Average test MSE	Average test R ²	Average test MSE	Average test R ²
Random forest models	0.034 (0.002)	0.590 (0.01)	0.029 (0.004)	0.657 (0.03)

Neural network models	0.037 (0.003)	0.549 (0.02)	0.031 (0.005)	0.628 (0.04)
-----------------------	---------------	--------------	---------------	--------------

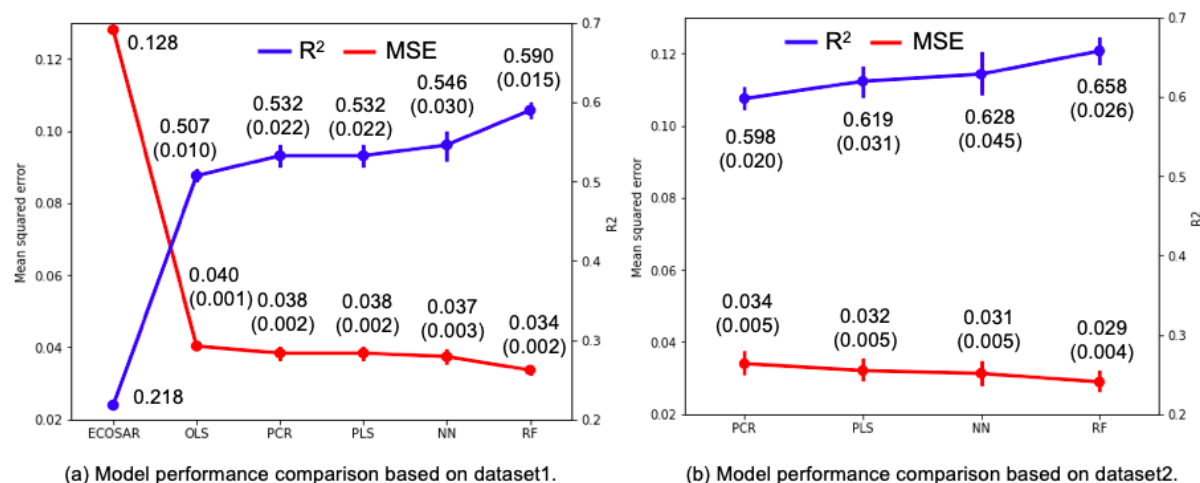


Figure 4-4. Performance of the random forest models compared with other models.

The main reason I believe is that neural networks are better when dealing with the same types of data than with “mixed” types of data. For example, many neural network applications involve images (each feature is a pixel) or speech signals (each feature is an amplitude sample), which all have the same kinds of values. While random forests, or tree-based models in general, are naturally incorporate both numeric and categorical variables, therefore, they have a better performance than neural networks on the two datasets.

4.3.3 Variable Importance via Random Forests

Figure 4-5 shows the variable importance calculated by the random forest model on dataset1. The results show a good consensus with the important features identified by neural network models. The two most important variables are Sol25 and Kow. Relatively important variables are BAFfish, MW, Pvp25, and KH25C. kdissP, kdegW, kdegA, and kdegSl show relatively low importance in predicting the ecotoxicity of chemicals. Different from neural network model, random forest model assigns less importance to MoA. This is because MoA is a

categorical variable, where the sparsity of the values makes it assigned lower feature importance.

It has been observed that random forest models are biased in such a way that categorical variables with a smaller number of categories are less preferred.^{139, 140}

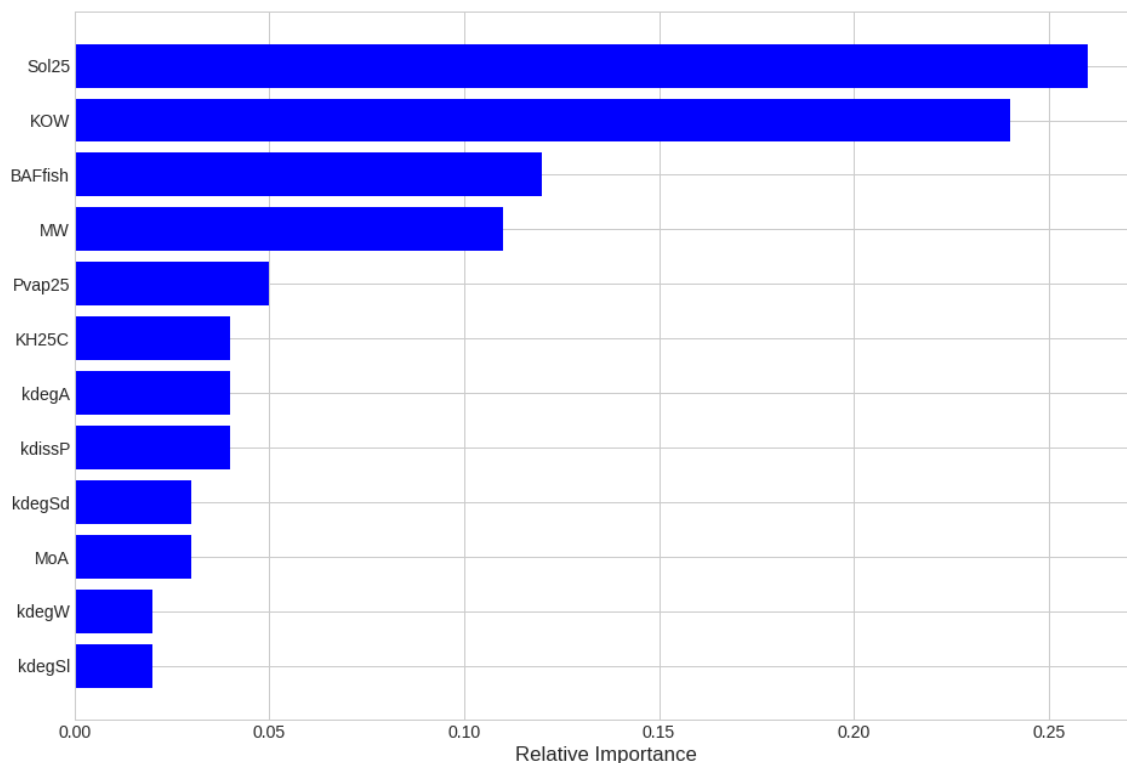


Figure 4-5. Feature importance by the random forest model for dataset1.

Figure 4-6 shows the variable importance calculated by the random forest model on dataset2. Most of the variables are not important in the random forest model, only 27 out of 695 variables have importance larger than zero. **Table 4-2** lists the feature importance of the 27 variables and their descriptions. Sol25 and Kow are still the two most important variables. Among other variables, some are also related with solubility, e.g., CIQPlogS and QPlogS; some are related with Kow, e.g., ALOGP and QPlogPo/w. BEHm1 to BEHm5 are burden eigenvalue descriptors. ATS1p to ATS4p are 2D autocorrelation descriptors. They reflect the topology of a molecule.⁸⁹ Mv and Mp are constitutional descriptors. EA(eV) (i.e., electron affinity) and dE

(i.e., the gap between the two frontier orbital energies HOMO and LUMO) are both electronic descriptors, which have been previously identified as important features in toxicity estimation.³⁸

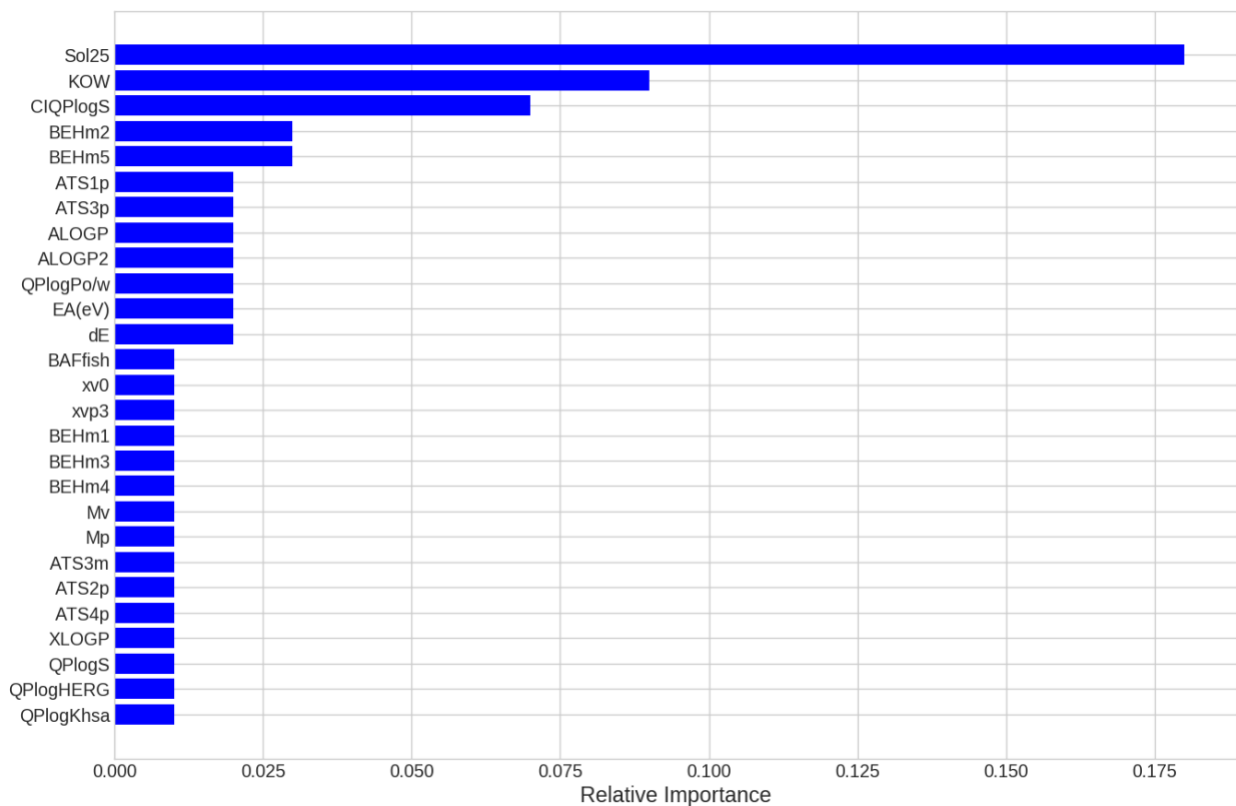


Figure 4-6. Feature importance by the random forest model for dataset2.

Table 4-2. Important features in dataset2 identified by random forest models.

	Property or descriptor	Feature importance	Description	Data source
1	Sol25	0.18	Solubility	USEtox
2	Kow	0.09	Octanol-water partitioning coefficient	USEtox
3	CIQPlogS	0.07	Conformation-independent predicted aqueous solubility, log S. S in mol dm ⁻³ is the concentration of the solute in a saturated solution that is in equilibrium with the crystalline solid.	QikProp
4	BEHm2	0.03	Highest eigenvalue n.2 of Burden matrix / weighted by atomic masses	T.E.S.T.
5	BEHm5	0.03	Highest eigenvalue n.4 of Burden matrix / weighted by atomic masses	T.E.S.T.
6	ATS1p	0.02	Broto-Moreau autocorrelation of a topological structure – lag 1 / weighted by atomic polarizabilities	T.E.S.T.
7	ATS3p	0.02	Broto-Moreau autocorrelation of a topological structure – lag 3 / weighted by atomic polarizabilities	T.E.S.T.

8	ALOGP	0.02	Ghose-Crippen octanol water coefficient	T.E.S.T.
9	ALOGP2	0.02	Ghose-Crippen octanol water coefficient squared	T.E.S.T.
10	QPlogPo/w	0.02	Predicted octanol/water partition coefficient	QikProp
11	EA(eV)	0.02	PM3 calculated electron affinity	QikProp
12	dE	0.02	Frontier orbital energies, HOMO–LUMO gap	QikProp
13	BAFfish	0.01	Bioaccumulation factor in fish	USEtox
14	xv0	0.01	Valence zero order chi index	T.E.S.T.
15	xvp3	0.01	Valence 3rd order path chi index	T.E.S.T.
16	BEHm1	0.01	Highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses	T.E.S.T.
17	BEHm3	0.01	Highest eigenvalue n. 3 of Burden matrix / weighted by atomic masses	T.E.S.T.
18	BEHm4	0.01	Highest eigenvalue n. 4 of Burden matrix / weighted by atomic masses	T.E.S.T.
19	Mv	0.01	Mean atomic van der Waals volume (scaled on Carbon atom)	T.E.S.T.
20	Mp	0.01	Mean atomic polarizability (scaled on Carbon atom)	T.E.S.T.
21	ATS3m	0.01	Broto-Moreau autocorrelation of a topological structure - lag 3 / weighted by atomic masses	T.E.S.T.
22	ATS2p	0.01	Broto-Moreau autocorrelation of a topological structure - lag 2 / weighted by atomic polarizabilities	T.E.S.T.
23	ATS4p	0.01	Broto-Moreau autocorrelation of a topological structure - lag 4 / weighted by atomic polarizabilities	T.E.S.T.
24	XLOGP	0.01	Wang octanol water partition coefficient	T.E.S.T.
25	QPlogS	0.01	Predicted aqueous solubility, log S. S in mol dm ⁻³ is the concentration of the solute in a saturated solution that is in equilibrium with the crystalline solid.	QikProp
26	QPlogHERG	0.01	Predicted IC50 value for blockage of HERG K ⁺ channels.	QikProp
27	QPlogKhsa	0.01	Prediction of binding to human serum albumin.	QikProp

Based on the identified 27 important variables, I rebuild a random forest model. The average test MSE of the model is 0.030 with a standard deviation of 0.004, and the average test R^2 is 0.642 with a standard deviation of 0.024, which are very close to the results of the model built based on all 695 variables. This is very helpful when applying the model on a much broader range of chemicals because we can obtain the same quality of results while having a much lower number of input data to collect.

4.3.4 Computational Time

The code is programmed in Python and run on Flux. For each split of data, 30 models (10 different number of trees and 3 different maximum number of features) are trained to calculate performance on training set and out-of-bag samples, one additional model is trained to calculate test performance. In total, 310 random forest models are trained for 10 different splits of data.

When using 10 processors regarding to 10 different splits to calculate simultaneously, the computational time for model selection and testing is 8 minutes 51 seconds for dataset1 and is 1 hour 8 minutes 4 seconds for dataset2. Once the model is developed, using 10 processors, it takes 59 seconds for predicting the missing HC_{50} (including calculating their bootstrap confidence interval) for dataset1 and 17 minutes 57 seconds for dataset2.

Compared with random forests, neural networks require more computational resource in terms of model selection and testing (**Table 4-3**). This is because neural network models have many parameters and each parameter has many options. 240 models were trained in chapter 3 to find an appropriate model. While random forests only need to find the best number of trees and the best number of maximum features for splitting the nodes, only 30 models were trained in model selection. Therefore, random forest models take less time for selecting and testing the model. Once the model is trained, however, neural networks require less time for prediction on a larger dataset (i.e., dataset2). This is very useful for real time prediction, which is one of the reasons neural networks have been extensively applied in complex tasks, such as computer vision, speech recognition, and natural language processing. In ecotoxicity estimation, for now, we care more about prediction accuracy than computational time. Therefore, I choose random forest models to estimate the missing HC_{50} and CF_{eco} in USEtox.

Table 4-3. Computational time comparison of random forest models and neural network models.

Dataset	Random forest models		Neural network models	
	Model selection and testing (10 cores)	Prediction (10 cores)	Model selection and testing (300 cores)	Prediction (10 cores)
Dataset1	00:08:51	00:01:05	00:26:05	00:03:29
Dataset2	01:08:04	00:15:52	00:29:02	00:05:15

4.3.5 Drawbacks and Advantages of Random Forests

Random forests have some drawbacks: (1) They are ensemble models consist of a large number of decision trees, therefore less interpretable than an individual decision tree. (2) The variable importance from random forests is not reliable for data containing categorical variables, in which case, random forests are in favor of numerical variables and categorical variables with more levels.¹⁴¹ For example, in our case, MoA is scored with less importance compared with neural network models. (3) Training a large number of deep trees can have high computational costs (but can be easily paralleled) and use a lot of memory.

On the other hand, random forests have the following advantages: (1) The predictive performance can compete with the best supervised learning algorithms. For many datasets, they produce highly accurate results; (2) Random forests can operate on both continuous and categorical variables directly. They do not require feature engineering, such as scaling or normalization. They can also cope with missing values and maintains accuracy when a large proportion of the data are missing;¹⁴² (3) They offer efficient estimates of the test error (OOB error) without incurring the cost of repeated model training associated with cross-validation; (4) The performance is not sensitive to parameters. Typically, two parameters need to be tuned, the number of trees and the number of features to be selected at each node.

4.3.6 Estimation of Missing HC_{50} and CF_{eco} in USEtox

Since random forests have a better performance than neural networks, I estimate the HC_{50} and CF_{eco} for the 578 chemicals that do not have HC_{50} and CF_{eco} values in USEtox version 2.1. I first use the best performed random forest model (1,000 trees and maximum number of features is 231) developed with dataset2 to estimate HC_{50} and CF_{eco} for 446 chemicals. For the rest 132 chemicals, for which some descriptors are not available in dataset2, I use the random forest model (1,000 trees and maximum number of features is 4) developed with dataset1 to estimate their HC_{50} and CF_{eco} . I also calculated their confidence interval and their application domains to indicate the uncertainty of the results. The predicted HC_{50} values with their confidence intervals and the distance to the centroid of these chemicals are listed in **Appendix C**.

To visualize the application domain of the model, take the dataset1 as an example, I project the 12 properties of the chemicals on two main principal components by principal component analysis¹⁴³ (**Figure 4-7**). The blue points represent chemicals I use to build the model. The red star in the middle indicates the centroid of these chemicals. The large red circle represents the range of the distance that 90% of the blue points are enclosed, which is defined as the application domain of this model. The green crosses represent the chemicals for which the HC_{50} values are missing in USEtox and predicted by this model. The green crosses located within the red circle are more reliable and recommended to use with confidence. Those located outside of the red circle should be used with caution.

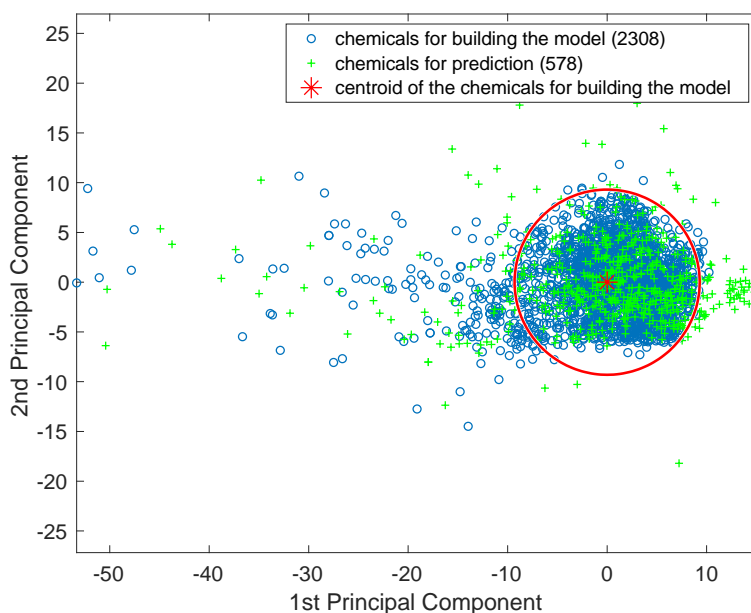


Figure 4-7. Visualization of the application domain of the models developed based on dataset1.

4.3.7 Implications for LCA

I envision three major implications for LCA research and practice. First, LCA practitioners can directly use the estimated CF_{eco} values in LCA case studies. I provide CF_{eco} values for 578 chemicals and their confidence interval and application domain as references for users. These estimates do not replace the need for chemical-specific laboratory tests to obtain accurate HC_{50} values; but they serve as a useful reference when laboratory test data are not available.

Second, when some chemicals that are not listed here missing their CF_{eco} , one can calculate their CF_{eco} by the developed model. First, collect or calculate the variables in dataset1 or dataset2 for these chemicals, or at least for those important variables. Some variables in dataset1, such as solubility and Kow , can be calculated by EPI suite¹⁴⁴. Variables in dataset2 can be calculated by T.E.S.T. and QikProp. Then, train the model with a subset of dataset1 or

dataset2, depending on which variables can be collected. Finally, the missing CF_{eco} can be predicted by the developed model.

Last but not least, for other impact categories, such as human toxicity, one can develop machine learning models to estimate their characterization factors. The first and most important thing is to find relevant variables related to the interested impact. Then, one can develop new machine learning models following the same procedure: data splitting, model selection, model testing, and prediction.

Beyond LCA, the developed neural network and random forest models can be used in chemical risk management to predict ecotoxicity of new chemicals or as a screening process to identify chemicals with high predicted ecotoxicity potential to further test in priority.

4.3.8 Future Work

Besides ecotoxicity, USEtox also developed characterization factors for human toxicity. Due to the limited availability of chronic data for estimating dose-response and disease incidences, 67.4% of chemicals in USEtox do not have human toxicity characterization factors. Estimating human toxicity for chemicals represents an interesting research direction for the future.

Besides neural networks and random forests, other machine learning models also show high prediction accuracy on many datasets. For example, boosting models, such as Adaptive boosting (AdaBoost) and Gradient boosting machine (GBM), offer systematic solutions to combine the predictive power of multiple models.¹⁴⁵ The result is a single model which gives the aggregated output from several models. Different from random forest, where each tree grows independently, AdaBoost and GBM trains many models in a gradual, additive and sequential way. By analyzing and correcting errors in previous models, the final model can converge to a

strong model. These machine learning models might give even better performance on toxicity estimation.

4.4 Summary

In this chapter, I develop random forest models to estimate HC_{50} and CF_{eco} for chemicals in USEtox. A random forest builds each decision tree with a randomly selected subset of training data. Each node of each decision tree is split using randomly chosen variables. The two levels of randomness ensure that a random forest generates trees that are uncorrelated with one another. As a result, prediction errors are dispersed throughout the model and are canceled out by averaging. Results show that random forest models quantitatively perform better and have more stable performance than neural network models. Although the model structure is not easy to interpret, the random forest models provide an efficient way to quickly predict the toxicity and characterization factors of chemicals for LCA and broader applications.

Chapter 5 Conclusions

Drawing from the rapid development in data science, this research applies data-driven methods to provide efficient and effective solutions to support decision making in sustainability. Specifically, I focus on LCA which evaluates environmental impacts along products' whole life cycle to avoid environmental impacts shifting among different life cycle stages. To improve the feasibility of LCA and reduce data collection efforts for LCA studies, I propose two computational frameworks to respectively estimate unit process data and characterization factors in two steps of LCA.

First, at the LCI step, unit process data characterize the resource/energy consumed and emission/waste generated in a particular process within a product's life cycle. Viewing a unit process database as a network provides a new perspective to understand unit process data. Using link prediction techniques, I develop a similarity-based method to estimate missing unit process data in the ecoinvent database. Results show that on average missing data can be accurately estimated when less than 5% data are missing in one process. However, when more data are missing (e.g., over 20%), the estimation accuracy becomes lower. This is because, when the number of known data is limited, the similarity calculated based on such limited data is no longer accurate and reliable. Therefore, this method is more applicable in the situation that most data are known and only a few data points are missing. In addition, this method assumes that the observed unit process data used to estimate the missing data are complete. In practice, it is impossible to have an observed database that is complete. In fact, as the most comprehensive and

widely used LCA database, ecoinvent still has many missing data that are simply filled with zeros.

Despite all these limitations, this study provides a new direction for estimating unit process data for LCA. In future research, how to find the optimal choice of similarity measurements and the number of most similar processes used for estimation needs to be carefully examined.

Second, at the LCIA step, many ecotoxicity characterization factors are missing for many chemicals because evaluating potential hazardous effects of chemicals has been traditionally done by laboratory experiments. Experiment-based ecotoxicity test results are only available for a small set of chemicals due to the high cost associated with laboratory experiments. Given the enormous amount and ever-increasing number of chemicals that are used in production and incorporated in products, characterizing chemical ecotoxicity with a lower cost has become critical for guiding technology and policy development for chemical risk management and LCA studies. Neural networks and random forests can effectively estimate the missing HC_{50} to derive ecotoxicity characterization factors. The results show they perform better than the ECOSAR model and linear regression models.

However, the proposed model relies on HC_{50} values in USEtox, which are highly aggregated across species, conditions, and modes of actions. Since the data underlying the HC_{50} across USEtox chemicals are based on different datasets for each chemical, they may introduce different levels of uncertainties. Consequently, the results here could inherit these uncertainties. Moreover, in USEtox, ecotoxicity of chemicals is currently based on freshwater toxicity without consideration of terrestrial or marine toxicity. The results thus are also confined in the scope of freshwater toxicity.

Despite these limitations, the proposed models perform better than ECOSAR and linear regression models. Estimating ecotoxicity of chemicals is a difficult task because of the complex physical, chemical, and biological processes how chemicals transform and interact in environmental media. The neural network and random forest models explore these complex processes through the pattern revealed from observed data. Although the model structure is not easy to interpret, they provide an efficient way to quickly predict the ecotoxicity and characterization factors of chemicals for LCA and broader applications.

This research advances computational modeling in the LCA field in various ways. First, estimating missing LCA data without on-site investigation and laboratory tests will significantly reduce the cost of and save time for LCA studies. Second, filling in the missing data and compiling complete data will enhance the credibility of LCA studies. Third, data used in an LCA often come from various sources with different quality and accuracy. By comparing the predicted results with the observed data, one can evaluate the quality of those observed data, identify inaccurate data, and guide future improvements. Lastly, the industrial system is constantly evolving in the way that new processes and products are invented all the time. Predicting emerging patterns between the new process and environmental interventions can help reasonably estimate LCA data for emerging technologies for which empirical LCA data are less available.

Furthermore, efficiently and cost-effectively conducting an LCA study will allow easier implementation of LCA, broadly promote LCA applications in various areas, such as policy making and corporate social responsibility, and enable businesses, governments, and citizens to make better decisions towards sustainability based on complete life cycle information.

Appendices

Appendix A. Distance measuring methods tested

A distance metric is a function that defines a distance between two observations. Given an $m \times n$ data matrix X , which is treated as m (1-by- n) row vectors x_1, x_2, \dots, x_m , and an $n \times m$ data matrix Y , which is treated as n (1-by- m) row vectors y_1, y_2, \dots, y_n , the various distances between the vector x_s and y_t are defined as follows:

1. Euclidean distance

The Euclidean distance is the straight-line distance between two points in Euclidean space. The Euclidean distance is a special case of the Minkowski distance, where $q=2$.

$$d_{st} = (x_s - y_t)(x_s - y_t)^T = \left(\sum_{i=1}^n |x_{si} - y_{ti}|^2 \right)^{1/2}$$

2. Standardized Euclidean distance

$$d_{st} = (x_s - y_t)V^{-1}(x_s - y_t)^T$$

where V is the $n \times n$ diagonal matrix whose j th diagonal element is $(S(j)^2)$, where S is a vector of scaling factors for each dimension.

3. City block distance (Manhattan distance)

The city block distance between two points is the sum of the absolute differences of their coordinates. The city block distance is a special case of the Minkowski distance, where $q=1$.

$$d_{st} = \sum_{i=1}^n |x_{si} - y_{ti}|$$

4. Chebychev distance

The Chebychev distance is a special case of the Minkowski distance, where $q=\infty$.

$$d_{st} = \max_i \{|x_{si} - y_{ti}|\}$$

5. Cosine distance

$$d_{st} = \left(1 - \frac{x_s y_t^T}{\sqrt{(x_s x_s^T)(y_t y_t^T)}}\right)$$

6. Correlation distance

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)^T}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)^T} \sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)^T}}$$

where

$$\bar{x}_s = \frac{1}{n} \sum_i x_{si}$$

$$\bar{y}_t = \frac{1}{n} \sum_i y_{ti}$$

7. Hamming distance

$$d_{st} = (\#(x_{si} \neq y_{ti})/n)$$

8. Jaccard distance

$$d_{st} = \frac{\#[(x_{si} \neq y_{ti}) \cap (x_{si} \neq 0) \cap (y_{ti} \neq 0)]}{\#[(x_{si} \neq 0) \cup (y_{ti} \neq 0)]}$$

9. Spearman distance

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)^T}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)^T} \sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)^T}}$$

where

r_{si} is the rank of x_{si} taken over $x_{1i}, x_{2i}, \dots, x_{mx,i}$

r_{ti} is the rank of y_{ti} taken over $y_{1i}, y_{2i}, \dots, y_{my,i}$

r_s and r_t are the coordinate-wise rank vectors of x_s and y_t , i.e., $r_s = (r_{s1}, r_{s2}, \dots, r_{sn})$ and $r_t =$

$(r_{t1}, r_{t2}, \dots, r_{tn})$

$$\bar{r}_s = \frac{1}{n} \sum_i r_{si} = \frac{(n+1)}{2}$$

$$\bar{r}_t = \frac{1}{n} \sum_i r_{ti} = \frac{(n+1)}{2}$$

10. Minkowski distance

$$d_{st} = \left(\sum_{i=1}^n |x_{si} - y_{ti}|^q \right)^{1/q}$$

For the special case of $q=1$, the Minkowski distance gives the city block distance. For the special case of $q=2$, the Minkowski distance gives the Euclidean distance. For the special case of $q=\infty$, the Minkowski distance gives the Chebychev distance.

Table A-1. Average Mean Percentage Error (MPE) when missing 1% data calculated using different distance functions.

Method	Definition	MPE
Euclidean distance	The straight-line distance between two points in Euclidean space. equivalent to Minkowsk distance when $q=2$	80.99%
Standardized Euclidean distance	Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation	105.96%
City block distance	Also called Manhattan distance, the sum of the absolute differences of their coordinates, equivalent to Minkowski distance when $q=1$	80.90%
Chebychev distance	Maximum coordinate difference, equivalent to Minkowski distance when $q=\infty$	86.08%
Cosine distance	One minus the cosine of the included angle between points	1393.67%
Correlation distance	One minus the correlation between points	1208.16%
Hamming distance	Percentage of coordinates that differ	100.00%
Jaccard distance	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ	24972.57%

Spearman	One minus the sample Spearman's rank correlation between observations	1194.69%
----------	---	----------

Note: Minkowski distance is a metric which can be considered as a generalization of the Euclidean distance ($q=2$), the City block distance ($q=1$), and the Chebychev distance ($q=\infty$). When q is smaller the MPE is getting smaller. Therefore, through adjusting the q value in Minkowski distance, we can find the best q value that can have the best estimation performance.

Appendix B. Normalization of the unit process database

Normalization sometimes is needed to represent data in similar order of magnitudes.

There are different ways of normalization, such as Z-score and min-max. In this paper, we define another way of matrix normalization. We first pre-multiplied original matrix A by a diagonal matrix, L , in which the diagonal elements are the inverse of the maximum values in each row of matrix A .

$$L = \begin{bmatrix} \frac{1}{\max a_{1,:}} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\max a_{n,:}} \end{bmatrix}$$

then,

$$B = LA = \begin{bmatrix} \frac{1}{\max a_{1,:}} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\max a_{n,:}} \end{bmatrix} \begin{bmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \frac{a_{11}}{\max a_{1,:}} & \dots & \frac{a_{n1}}{\max a_{1,:}} \\ \dots & \dots & \dots \\ \frac{a_{m1}}{\max a_{n,:}} & \dots & \frac{a_{mn}}{\max a_{n,:}} \end{bmatrix}$$

We then post-multiplied B by R , a diagonal matrix in which the diagonal elements are the inverse of the maximum values in each column of matrix B .

$$R = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{\max b_{:,1}} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\max b_{:,m}} \end{bmatrix}$$

then,

$$C = BR = \begin{bmatrix} b_{11} & \dots & b_{n1} \\ \dots & \dots & \dots \\ b_{m1} & \dots & b_{mn} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{\max b_{:,1}} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\max b_{:,m}} \end{bmatrix} = \begin{bmatrix} \frac{b_{11}}{\max b_{:,1}} & \dots & \frac{b_{1n}}{\max b_{:,m}} \\ \dots & \dots & \dots \\ \frac{b_{m1}}{\max b_{:,1}} & \dots & \frac{b_{mn}}{\max b_{:,m}} \end{bmatrix}$$

After the matrix normalization, the resulting matrix C has the maximum value as 1 and the minimum value as 0 for all the rows and columns. By doing so, the values in the whole matrix are normalized in the similar orders of magnitude. Another benefit of such transformation is that MPEs calculated based on the matrix C are the same as those based on the original matrix A since this transformation is basically the multiplication of constant values, which can be recorded for denormalizing the data after estimation. The row normalization is to reduce the order of magnitude difference in intermediate and elementary flows, which is equivalent to converting the units of the flows. While the column normalization is to remove the scale difference of the process units, which is equivalent to converting the functional units.

Based on this definition, we tried three different normalization strategies: 1) normalization based on the complete matrix; 2) normalization based on the training set to avoid introducing future information in the test set; and 3) without normalization. Table S2 shows the comparison of the average MPE using the three different strategies.

In general, estimation with normalization based on the training set has the highest average MPE. This is because, in ecoinvent, the order of the magnitude of the test data can be very different from that of the training data. If there are magnitude of difference between the test

data and the training data, test data after normalization can be extremely small or large, while the maximum of the normalized training data is always 1. Therefore, normalizing the test data based on the information from the training data can be problematic and skew the test data.

Much to my surprise, estimation without normalization has the lowest average MPE. I believe this is because normalization, while making the data more regular, can actually lose information that might be useful in the estimation.

Therefore, based on the above comparison and analysis, we did the estimation without normalization.

Table B-1. Average Mean Percentage Error (MPE) using three different normalization strategies.

Average MPE	Normalization based on the whole matrix	Normalization based on the training set	Without normalization
1% missing	$3.14 \times 10^{-14}\%$	17.95%	$2.09 \times 10^{-13}\%$
5% missing	13.43%	100%	$2.85 \times 10^{-12}\%$
10% missing	59.45%	100%	39.32%
20% missing	97.55%	100%	91.39%

Appendix C. Estimation of missing HC₅₀ and CF_{eco}

Table C-1 provides the estimated logHC₅₀, their 95% confidence interval, and the relative distance to the application domain (AD). The distance less than one means in the AD, larger than one means outside of the AD. The HC₅₀ values are first estimated by model2 (i.e., random forest model with 1000 trees and maximum features used for splitting a node is 231, trained by dataset2). When features are unavailable for some chemicals to use model 2, model1 (i.e., random forest model with 1000 trees and maximum features used for splitting a node is 4, trained by dataset1) is applied to estimate the missing HC₅₀.

Table C-2 provides midpoint ecotoxicity characterization factors (PAF.m3.day/kg emitted) calculated based on the HC₅₀ values in **Table C-1** and exposure factors (XF) and fate factors (FF) in USEtox. The endpoint ecotoxicity characterization factors (PDF.m3.day/kg emitted) can be easily calculated by dividing midpoint ecotoxicity characterization factors with two as defined in USEtox.

Table C-1. Estimation of missing HC₅₀.

CAS	prediction	95% lower	95% upper	Distance to AD	model
100-40-3	6.41E-01	4.81E-01	8.01E-01	6.78E-01	model2
100-75-4	2.08E+00	1.76E+00	2.40E+00	5.17E-01	model2
10034-93-2	1.78E+00	1.30E+00	2.26E+00	1.00E+00	model1
10048-13-2	-2.73E-01	-5.93E-01	4.77E-02	2.13E+00	model2
101-14-4	-1.69E-01	-3.64E-01	2.67E-02	4.49E-01	model2
101-61-1	-5.06E-01	-7.68E-01	-2.44E-01	6.81E-01	model2
101-79-1	8.64E-02	-3.76E-02	2.11E-01	1.35E-01	model2
101-80-4	8.24E-01	5.31E-01	1.12E+00	3.25E-01	model2
101-90-6	8.70E-01	5.69E-01	1.17E+00	1.40E+00	model2

10191-41-0	-1.88E+00	-2.60E+00	-1.16E+00	3.05E+00	model2
102-09-0	2.77E-01	1.43E-01	4.12E-01	1.83E-01	model2
102-50-1	1.15E+00	9.84E-01	1.32E+00	2.97E-01	model2
102-77-2	5.49E-01	2.34E-01	8.64E-01	5.24E-01	model2
103-03-7	1.62E+00	1.35E+00	1.89E+00	4.96E-01	model2
103-16-2	3.46E-01	2.60E-01	4.31E-01	1.62E-01	model2
104-46-1	4.02E-01	2.76E-01	5.29E-01	5.67E-01	model2
10473-70-8	-4.32E-01	-7.42E-01	-1.21E-01	8.80E-01	model2
105-11-3	1.21E+00	8.40E-01	1.58E+00	4.75E-01	model2
10589-74-9	1.61E+00	1.35E+00	1.86E+00	5.59E-01	model2
10595-95-6	2.16E+00	1.90E+00	2.42E+00	6.87E-01	model2
106-87-6	1.89E+00	1.60E+00	2.18E+00	6.23E-01	model2
106-99-0	1.26E+00	8.47E-01	1.68E+00	8.82E-01	model2
1068-57-1	2.18E+00	1.70E+00	2.67E+00	7.67E-01	model2
107-30-2	1.60E+00	1.32E+00	1.88E+00	1.05E+00	model2
107-35-7	1.93E+00	1.59E+00	2.26E+00	6.36E-01	model2
1078-38-2	1.45E+00	1.19E+00	1.71E+00	3.85E-01	model2
108-30-5	1.88E+00	1.59E+00	2.18E+00	6.01E-01	model2
108-60-1	1.19E+00	1.00E+00	1.39E+00	8.31E-01	model2
1083-57-4	7.85E-01	6.14E-01	9.55E-01	2.58E-01	model2
109-84-2	2.19E+00	1.86E+00	2.53E+00	8.05E-01	model2
1095-90-5	6.73E-01	5.07E-01	8.40E-01	3.05E-01	model1
110-85-0	2.28E+00	1.94E+00	2.62E+00	7.04E-01	model2
110-89-4	1.72E+00	1.35E+00	2.10E+00	6.66E-01	model2
1116-54-7	2.57E+00	2.09E+00	3.04E+00	6.46E-01	model2
1119-68-2	1.76E+00	1.33E+00	2.18E+00	3.29E-01	model1
112-63-0	-1.42E+00	-1.82E+00	-1.01E+00	8.80E-01	model2
1120-71-4	1.89E+00	1.53E+00	2.24E+00	5.07E-01	model2
1121-92-2	1.41E+00	1.28E+00	1.55E+00	6.50E-01	model2
1133-64-8	1.06E+00	8.30E-01	1.30E+00	2.67E-01	model2
114-83-0	1.61E+00	1.43E+00	1.79E+00	4.31E-01	model2
115-07-1	1.25E+00	9.43E-01	1.55E+00	8.94E-01	model2
115-09-3	8.95E-01	5.35E-01	1.26E+00	3.32E-01	model1
115-11-7	1.29E+00	9.84E-01	1.59E+00	8.69E-01	model2
115-28-6	6.03E-01	5.71E-02	1.15E+00	2.54E+00	model2
1156-19-0	2.96E-01	6.38E-02	5.27E-01	7.97E-01	model2
116-14-3	1.30E+00	1.10E+00	1.49E+00	3.55E+00	model1
1163-19-5	-1.83E+00	-2.63E+00	-1.04E+00	1.69E+00	model2
117-39-5	5.88E-01	2.57E-01	9.19E-01	1.44E+00	model2

117-79-3	-6.32E-02	-4.32E-01	3.06E-01	9.97E-01	model2
118-92-3	1.44E+00	1.27E+00	1.60E+00	4.37E-01	model2
119-47-1	-1.55E+00	-2.07E+00	-1.02E+00	1.76E+00	model2
119-53-9	7.08E-01	5.19E-01	8.97E-01	3.16E-01	model2
119-84-6	1.30E+00	1.16E+00	1.44E+00	2.40E-01	model2
1192-28-5	2.21E+00	1.93E+00	2.49E+00	6.05E-01	model2
120-32-1	-4.24E-02	-1.41E-01	5.59E-02	3.92E-01	model2
120-58-1	3.33E-01	1.35E-01	5.30E-01	4.97E-01	model2
120-71-8	1.25E+00	1.09E+00	1.40E+00	3.15E-01	model2
120-78-5	-1.02E+00	-1.38E+00	-6.67E-01	1.06E+00	model2
121-19-7	1.40E+00	1.17E+00	1.64E+00	3.28E-01	model1
121-59-5	1.38E+00	1.05E+00	1.71E+00	3.32E-01	model1
121-66-4	1.09E+00	8.66E-01	1.32E+00	5.06E-01	model2
121-88-0	1.16E+00	1.03E+00	1.30E+00	3.97E-01	model2
1212-29-9	1.36E-01	-1.56E-01	4.28E-01	1.62E-01	model2
122-20-3	2.37E+00	2.12E+00	2.61E+00	4.64E-01	model2
123-73-9	9.78E-01	2.50E-01	1.71E+00	7.62E-01	model2
124-58-3	2.05E+00	1.65E+00	2.45E+00	3.35E-01	model1
124-64-1	2.27E+00	1.89E+00	2.65E+00	4.24E-01	model1
1248-18-6	6.20E-01	8.41E-02	1.16E+00	3.37E-01	model1
125-33-7	1.30E+00	9.57E-01	1.65E+00	8.31E-01	model2
126-07-8	-1.62E-01	-3.73E-01	4.84E-02	2.05E+00	model2
126-13-6	-1.48E+00	-2.00E+00	-9.58E-01	1.34E+01	model2
126-98-7	1.14E+00	6.91E-01	1.59E+00	7.23E-01	model2
12663-46-6	6.02E-01	1.39E-01	1.07E+00	4.16E+00	model2
127-47-9	-1.71E+00	-2.25E+00	-1.16E+00	1.21E+00	model2
127-48-0	1.94E+00	1.56E+00	2.32E+00	3.51E-01	model2
127-69-5	5.19E-01	2.09E-01	8.29E-01	7.06E-01	model2
1271-19-8	-6.33E-01	-8.26E-01	-4.40E-01	2.95E-01	model1
128-66-5	-1.47E+00	-2.17E+00	-7.82E-01	2.88E+00	model2
129-15-7	-2.78E-01	-4.99E-01	-5.76E-02	1.39E+00	model2
129-43-1	-3.47E-02	-3.38E-01	2.69E-01	1.04E+00	model2
13010-07-6	1.43E+00	1.15E+00	1.71E+00	4.90E-01	model2
13073-35-3	1.42E+00	1.10E+00	1.74E+00	6.10E-01	model2
131-01-1	-5.25E-01	-9.29E-01	-1.21E-01	6.95E+00	model2
13256-06-9	2.72E-01	1.00E-01	4.44E-01	4.88E-01	model2
13256-11-6	1.24E+00	1.08E+00	1.41E+00	3.74E-01	model2
13292-46-1	9.33E-02	-3.31E-01	5.18E-01	3.20E-01	model1
134-03-2	2.03E+00	1.70E+00	2.35E+00	6.17E-01	model1

134-29-2	1.86E+00	1.39E+00	2.34E+00	8.40E-01	model1
135-20-6	2.01E+00	1.66E+00	2.37E+00	1.00E+00	model1
135-23-9	8.68E-01	4.42E-01	1.30E+00	3.05E-01	model1
13552-44-8	8.16E-01	5.24E-01	1.11E+00	3.37E-01	model1
136-23-2	-1.88E+00	-2.24E+00	-1.52E+00	1.17E+01	model1
136-40-3	1.31E+00	1.17E+00	1.46E+00	3.30E-01	model1
136-77-6	3.40E-01	2.09E-01	4.71E-01	2.86E-01	model2
137-09-7	1.90E+00	1.56E+00	2.23E+00	1.00E+00	model1
137-17-7	1.10E+00	8.97E-01	1.31E+00	2.51E-01	model2
13743-07-2	1.94E+00	1.82E+00	2.06E+00	6.06E-01	model2
13752-51-7	6.04E-01	1.62E-01	1.05E+00	5.83E-01	model2
13838-16-9	1.33E+00	1.13E+00	1.52E+00	3.32E-01	model1
139-65-1	7.31E-01	5.38E-01	9.24E-01	3.16E-01	model2
139-94-6	6.92E-01	4.54E-01	9.31E-01	3.68E-01	model2
13927-77-0	-1.40E+00	-1.65E+00	-1.15E+00	1.51E-01	model1
140-67-0	5.01E-01	4.07E-01	5.95E-01	5.67E-01	model2
1402-68-2	4.74E-01	-1.74E-03	9.50E-01	3.36E-01	model1
14026-03-0	1.78E+00	1.53E+00	2.03E+00	3.90E-01	model2
142-04-1	1.82E+00	1.36E+00	2.27E+00	3.10E-01	model1
142-46-1	1.10E+00	7.05E-01	1.49E+00	6.73E-01	model2
142-83-6	1.24E+00	9.15E-01	1.56E+00	7.24E-01	model2
14239-68-0	6.06E-01	3.58E-01	8.53E-01	3.37E-01	model1
143-19-1	-3.70E-01	-6.81E-01	-5.86E-02	3.29E-01	model1
14371-10-9	1.30E+00	1.20E+00	1.40E+00	4.85E-01	model2
144-02-5	1.77E+00	1.57E+00	1.97E+00	7.42E-01	model1
144-34-3	8.77E-01	5.21E-01	1.23E+00	3.37E-01	model1
144-48-9	1.04E+00	8.55E-01	1.22E+00	7.48E-01	model2
1456-28-6	2.09E+00	1.80E+00	2.38E+00	3.45E-01	model2
148-82-3	5.00E-01	2.17E-01	7.82E-01	5.05E-01	model2
149-29-1	1.30E+00	1.12E+00	1.49E+00	3.59E-01	model2
149-91-7	1.43E+00	1.28E+00	1.57E+00	3.90E-01	model2
150-38-9	1.69E+00	1.07E+00	2.31E+00	1.00E+00	model1
150-69-6	1.28E+00	1.14E+00	1.42E+00	3.79E-01	model2
153-18-4	5.30E-01	-3.03E-02	1.09E+00	7.16E+00	model2
15318-45-3	8.77E-01	3.52E-01	1.40E+00	8.24E-01	model2
15356-70-4	1.06E+00	9.67E-01	1.15E+00	2.57E-01	model2
155-04-4	-1.17E+00	-1.53E+00	-8.05E-01	2.40E-01	model1
156-10-5	2.90E-01	1.10E-01	4.70E-01	2.73E-01	model2
156-51-4	1.22E+00	1.00E+00	1.44E+00	1.00E+00	model1

156-62-7	1.59E+00	1.09E+00	2.09E+00	3.13E-01	model1
15879-93-3	7.75E-01	5.00E-01	1.05E+00	7.47E-01	model2
16071-86-6	1.82E+00	1.32E+00	2.32E+00	1.00E+00	model1
16301-26-1	1.69E+00	1.40E+00	1.97E+00	6.39E-01	model2
16338-97-9	1.29E+00	1.04E+00	1.53E+00	5.83E-01	model2
1643-20-5	2.39E-01	-1.49E-01	6.27E-01	3.63E-01	model2
16568-02-8	1.73E+00	1.46E+00	1.99E+00	6.96E-01	model2
16699-10-8	-5.40E-01	-7.51E-01	-3.29E-01	3.23E-01	model1
16813-36-8	1.67E+00	1.47E+00	1.87E+00	4.33E-01	model2
16846-24-5	4.38E-02	-5.17E-01	6.05E-01	1.53E+01	model2
169590-42-5	-2.65E-01	-4.48E-01	-8.14E-02	3.34E-01	model1
17026-81-2	1.11E+00	8.86E-01	1.33E+00	3.07E-01	model2
1717-00-6	1.20E+00	1.02E+00	1.38E+00	3.37E-01	model1
17608-59-2	1.11E+00	9.08E-01	1.31E+00	2.56E-01	model2
1777-84-0	3.99E-01	2.04E-01	5.93E-01	3.61E-01	model2
17924-92-4	2.72E-01	-6.66E-02	6.10E-01	9.42E-01	model2
18523-69-8	-7.30E-03	-1.56E-01	1.41E-01	4.94E-01	model2
18559-94-9	8.99E-01	5.79E-01	1.22E+00	4.58E-01	model2
18662-53-8	1.88E+00	1.47E+00	2.28E+00	1.00E+00	model1
18699-02-0	1.50E+00	1.30E+00	1.69E+00	3.70E-01	model2
19010-66-3	4.63E-01	1.49E-01	7.76E-01	3.36E-01	model1
191-24-2	-1.89E+00	-2.80E+00	-9.91E-01	2.43E+00	model2
193-39-5	-1.98E+00	-2.92E+00	-1.03E+00	2.28E+00	model2
1936-15-8	1.65E+00	1.32E+00	1.98E+00	3.10E-01	model1
1937-37-7	-8.86E-01	-1.40E+00	-3.69E-01	2.62E-01	model1
1955-45-9	2.30E+00	1.90E+00	2.69E+00	3.06E-01	model2
20265-96-7	1.81E+00	1.48E+00	2.15E+00	1.00E+00	model1
20325-40-0	8.40E-01	4.45E-01	1.24E+00	3.36E-01	model1
205-99-2	-1.80E+00	-2.63E+00	-9.69E-01	1.66E+00	model2
207-08-9	-1.93E+00	-2.96E+00	-8.97E-01	1.60E+00	model2
208-96-8	-3.97E-02	-1.37E-01	5.76E-02	6.19E-01	model2
20917-49-1	1.47E+00	1.27E+00	1.68E+00	5.00E-01	model2
20941-65-5	-8.57E-01	-1.40E+00	-3.17E-01	3.13E-01	model1
2122-86-3	1.01E+00	7.68E-01	1.26E+00	3.80E-01	model2
21260-46-8	1.01E+00	6.48E-01	1.38E+00	3.37E-01	model1
21436-96-4	1.31E+00	1.13E+00	1.50E+00	3.35E-01	model1
21436-97-5	1.09E+00	8.97E-01	1.27E+00	3.36E-01	model1
21626-89-1	5.48E-01	2.91E-01	8.05E-01	1.40E+00	model2
21638-36-8	6.34E-01	3.75E-01	8.93E-01	4.58E-01	model2

218-01-9	-1.85E+00	-2.67E+00	-1.02E+00	1.15E+00	model2
2185-92-4	6.79E-01	4.57E-01	9.01E-01	3.36E-01	model1
21884-44-6	9.66E-02	-4.12E-01	6.06E-01	8.42E+00	model2
22071-15-4	5.70E-01	2.06E-01	9.33E-01	6.44E-01	model2
22131-79-9	9.08E-01	7.33E-01	1.08E+00	3.28E-01	model2
2244-16-8	9.47E-01	7.82E-01	1.11E+00	2.24E-01	model2
22494-47-9	-1.74E-01	-4.81E-01	1.33E-01	9.11E-01	model2
22760-18-5	-8.85E-02	-5.17E-01	3.40E-01	1.15E+00	model2
22839-47-0	1.02E+00	6.59E-01	1.38E+00	6.49E-01	model2
22966-79-6	-2.29E+00	-3.09E+00	-1.49E+00	1.57E+01	model2
23031-25-6	9.25E-01	6.49E-01	1.20E+00	4.07E-01	model2
23282-20-4	9.62E-01	4.75E-01	1.45E+00	3.41E+00	model2
2353-45-9	1.68E+00	1.06E+00	2.30E+00	3.37E-01	model1
23746-34-1	9.19E-01	5.83E-01	1.26E+00	1.00E+00	model1
2409-55-4	2.95E-01	1.67E-01	4.24E-01	2.49E-01	model2
2425-85-6	-1.01E+00	-1.38E+00	-6.39E-01	1.30E+00	model2
2429-74-5	6.96E-01	1.58E-01	1.23E+00	3.37E-01	model1
2432-99-7	1.37E+00	1.08E+00	1.65E+00	5.98E-01	model2
2438-88-2	-7.09E-01	-8.54E-01	-5.65E-01	7.17E-01	model2
24382-04-5	1.52E+00	1.05E+00	1.99E+00	1.00E+00	model1
244-63-3	2.43E-01	5.70E-02	4.28E-01	3.90E-01	model2
24448-94-0	1.93E+00	1.49E+00	2.36E+00	4.43E-01	model2
24554-26-5	5.43E-01	2.73E-01	8.14E-01	4.09E-01	model2
24589-77-3	1.19E+00	7.37E-01	1.64E+00	3.29E-01	model1
2475-45-8	2.21E-01	-1.67E-01	6.09E-01	1.54E+00	model2
25013-15-4	5.51E-01	3.68E-01	7.35E-01	3.34E-01	model1
25265-71-8	2.35E+00	2.10E+00	2.59E+00	1.00E+00	model1
25321-14-6	5.87E-01	2.21E-01	9.52E-01	3.37E-01	model1
25451-15-4	1.10E+00	8.64E-01	1.33E+00	3.81E-01	model2
2578-75-8	5.51E-01	3.06E-01	7.95E-01	4.56E-01	model2
25812-30-0	-7.95E-02	-5.68E-01	4.08E-01	5.18E-01	model2
25843-45-2	1.76E+00	1.56E+00	1.96E+00	7.14E-01	model2
2602-46-2	1.57E+00	1.04E+00	2.10E+00	3.37E-01	model1
26049-68-3	7.55E-01	5.41E-01	9.68E-01	3.95E-01	model2
26049-69-4	4.05E-01	1.55E-01	6.55E-01	4.55E-01	model2
26049-70-7	4.07E-01	1.96E-01	6.19E-01	4.31E-01	model2
26049-71-8	9.54E-01	7.71E-01	1.14E+00	3.65E-01	model2
2611-82-7	1.14E+00	9.34E-01	1.34E+00	3.10E-01	model1
26148-68-5	3.55E-01	1.50E-01	5.61E-01	5.48E-01	model2

262-12-4	-2.73E-02	-3.13E-01	2.58E-01	6.34E-01	model2
26541-51-5	1.34E+00	9.81E-01	1.69E+00	5.04E-01	model2
26675-46-7	1.30E+00	1.08E+00	1.52E+00	3.33E-01	model1
2757-90-6	9.13E-01	5.17E-01	1.31E+00	5.24E-01	model2
27753-52-2	-2.50E+00	-3.00E+00	-2.00E+00	6.45E+05	model1
2783-94-0	1.63E+00	1.44E+00	1.82E+00	3.12E-01	model1
2784-94-3	9.36E-01	5.54E-01	1.32E+00	4.74E-01	model2
2832-40-8	-1.30E-01	-3.49E-01	8.91E-02	6.85E-01	model2
2835-39-4	1.15E+00	9.67E-01	1.34E+00	5.25E-01	model2
2837-89-0	1.52E+00	1.33E+00	1.70E+00	5.04E-01	model1
29069-24-7	-2.20E+00	-2.93E+00	-1.47E+00	9.18E+00	model2
29082-74-4	-1.65E+00	-2.39E+00	-9.19E-01	8.88E-01	model2
2955-38-6	-4.74E-01	-1.00E+00	5.41E-02	1.30E+00	model2
298-18-0	2.34E+00	1.94E+00	2.74E+00	6.04E-01	model2
298-81-7	1.86E-01	-4.33E-02	4.14E-01	7.27E-01	model2
29975-16-4	-3.34E-01	-6.15E-01	-5.39E-02	1.07E+00	model2
3012-37-1	1.09E+00	8.83E-01	1.30E+00	4.52E-01	model2
3012-65-5	1.78E+00	1.47E+00	2.09E+00	5.07E-01	model1
302-22-7	-6.57E-01	-1.05E+00	-2.65E-01	3.29E+00	model2
303-34-4	3.30E-01	-3.60E-02	6.95E-01	2.44E+00	model2
303-47-9	-4.23E-01	-8.50E-01	5.19E-03	2.16E+00	model2
3031-51-4	5.15E-01	1.71E-01	8.59E-01	3.37E-01	model1
305-03-3	-6.04E-02	-3.71E-01	2.50E-01	4.66E-01	model2
30516-87-1	6.41E-01	3.63E-01	9.19E-01	7.41E-01	model2
3054-95-3	1.78E+00	1.58E+00	1.98E+00	7.37E-01	model2
306-37-6	1.94E+00	1.51E+00	2.38E+00	1.00E+00	model1
3068-88-0	2.00E+00	1.72E+00	2.29E+00	4.19E-01	model2
3096-50-2	2.93E-01	2.80E-02	5.58E-01	8.67E-01	model2
315-22-0	7.84E-01	3.34E-01	1.23E+00	1.72E+00	model2
3165-93-3	1.05E+00	9.38E-01	1.16E+00	3.37E-01	model1
324-93-6	3.26E-01	1.72E-01	4.80E-01	3.36E-01	model1
3276-41-3	1.64E+00	1.40E+00	1.87E+00	5.48E-01	model2
32852-21-4	1.47E+00	1.23E+00	1.70E+00	5.21E-01	model2
3296-90-0	1.29E+00	9.34E-01	1.64E+00	3.71E-01	model2
331-39-5	1.20E+00	9.93E-01	1.41E+00	3.77E-01	model2
33229-34-4	9.31E-01	4.92E-01	1.37E+00	5.95E-01	model2
33857-26-0	-1.18E+00	-1.74E+00	-6.19E-01	9.00E-01	model2
34627-78-6	2.21E-01	5.78E-02	3.84E-01	2.94E-01	model2
34661-75-1	4.20E-01	8.99E-02	7.49E-01	2.33E+00	model2

3544-23-8	9.19E-02	-8.33E-02	2.67E-01	3.50E-01	model2
3570-75-0	6.95E-01	4.99E-01	8.91E-01	4.45E-01	model2
36133-88-7	8.30E-01	5.63E-01	1.10E+00	5.08E-01	model2
363-17-7	-9.26E-01	-1.18E+00	-6.78E-01	3.23E-01	model1
36322-90-4	2.85E-01	-1.15E-01	6.84E-01	1.64E+00	model2
36702-44-0	1.75E+00	1.45E+00	2.04E+00	3.92E-01	model2
3688-53-7	6.24E-01	3.45E-01	9.03E-01	5.04E-01	model2
3693-22-9	2.08E-01	4.88E-02	3.66E-01	4.95E-01	model2
37087-94-8	-1.24E-01	-4.39E-01	1.92E-01	1.14E+00	model2
3761-53-3	1.39E+00	1.13E+00	1.64E+00	3.10E-01	model1
3771-19-5	-8.00E-01	-1.08E+00	-5.15E-01	1.38E+00	model2
3775-55-1	1.18E+00	9.85E-01	1.37E+00	3.80E-01	model2
38434-77-4	1.69E+00	1.49E+00	1.89E+00	6.92E-01	model2
38514-71-5	7.59E-01	5.82E-01	9.36E-01	3.69E-01	model2
389-08-2	4.05E-01	-7.56E-02	8.85E-01	7.79E-01	model2
39156-41-7	1.76E+00	1.43E+00	2.09E+00	1.00E+00	model1
396-01-0	9.69E-01	7.87E-01	1.15E+00	9.49E-01	model2
398-32-3	2.41E-01	-6.07E-02	5.43E-01	3.36E-01	model1
39801-14-4	-1.62E+00	-2.42E+00	-8.22E-01	1.09E+01	model2
40548-68-3	2.00E+00	1.82E+00	2.18E+00	5.43E-01	model2
40580-89-0	5.46E-02	-2.17E-01	3.26E-01	3.76E-01	model2
4075-79-0	4.59E-01	2.85E-01	6.32E-01	2.40E-01	model2
4106-66-5	2.06E-01	4.06E-02	3.72E-01	4.94E-01	model2
41340-25-4	-1.83E-01	-6.39E-01	2.74E-01	1.40E+00	model2
4164-28-7	2.12E+00	1.92E+00	2.31E+00	6.72E-01	model2
4180-23-8	4.02E-01	2.75E-01	5.29E-01	5.67E-01	model2
42011-48-3	-6.45E-02	-2.06E-01	7.69E-02	3.36E-01	model1
42579-28-2	1.68E+00	1.45E+00	1.91E+00	4.91E-01	model2
434-07-1	-5.49E-01	-1.13E+00	2.92E-02	2.66E+00	model2
434-13-9	-1.25E+00	-1.79E+00	-7.00E-01	2.69E+00	model2
439-14-5	2.51E-01	2.62E-02	4.76E-01	8.46E-01	model2
443-72-1	1.48E+00	1.29E+00	1.67E+00	3.38E-01	model2
446-86-6	3.68E-01	2.55E-02	7.10E-01	8.35E-01	model2
4548-53-2	1.29E+00	1.12E+00	1.45E+00	3.36E-01	model1
471-29-4	2.03E+00	1.64E+00	2.42E+00	7.68E-01	model2
471-53-4	-1.50E+00	-2.06E+00	-9.42E-01	5.38E+00	model2
474-25-9	-4.66E-01	-1.17E+00	2.37E-01	2.91E+00	model2
476-66-4	6.57E-01	1.86E-01	1.13E+00	2.08E+00	model2
480-54-6	5.94E-01	2.95E-01	8.93E-01	1.61E+00	model2

4812-22-0	1.17E+00	9.33E-01	1.40E+00	5.45E-01	model2
493-78-7	3.30E-01	4.75E-02	6.12E-01	3.05E-01	model2
50-02-2	5.09E-01	7.33E-02	9.44E-01	3.37E-01	model1
50-14-6	-1.66E+00	-2.25E+00	-1.06E+00	2.46E+00	model2
50-18-0	8.58E-01	5.69E-01	1.15E+00	4.88E-01	model2
50-23-7	5.86E-01	1.25E-02	1.16E+00	3.04E+00	model2
50-24-8	5.09E-01	-7.14E-02	1.09E+00	3.04E+00	model2
50-33-9	2.07E-01	-1.28E-01	5.42E-01	1.19E+00	model2
50-55-5	-3.82E-01	-8.77E-01	1.14E-01	7.90E+00	model2
50-78-2	1.36E+00	1.25E+00	1.48E+00	3.62E-01	model2
50-81-7	1.61E+00	1.32E+00	1.91E+00	3.89E-01	model2
501-30-4	1.52E+00	1.29E+00	1.74E+00	4.30E-01	model2
50264-69-2	-4.37E-01	-7.77E-01	-9.60E-02	1.06E+00	model2
50594-66-6	-1.26E-01	-3.67E-01	1.16E-01	3.32E-01	model1
509-14-8	1.29E+00	9.48E-01	1.64E+00	5.64E-01	model2
513-37-1	1.05E+00	7.33E-01	1.36E+00	9.45E-01	model2
5131-60-2	9.15E-01	7.24E-01	1.11E+00	4.17E-01	model2
51325-35-0	6.57E-01	2.64E-01	1.05E+00	8.96E-01	model2
51333-22-3	1.31E-01	-4.37E-01	6.99E-01	4.05E+00	model2
51410-44-7	1.13E+00	9.86E-01	1.27E+00	2.20E-01	model2
51481-10-8	1.00E+00	5.39E-01	1.46E+00	3.12E+00	model2
51481-61-9	5.35E-01	2.28E-01	8.41E-01	4.09E-01	model2
51542-33-7	8.37E-01	6.68E-01	1.01E+00	4.10E-01	model2
517-28-2	7.84E-01	4.46E-01	1.12E+00	1.93E+00	model2
51786-53-9	1.28E+00	1.07E+00	1.49E+00	3.35E-01	model1
518-82-1	-4.80E-03	-3.16E-01	3.06E-01	1.46E+00	model2
52-76-6	-1.19E+00	-1.80E+00	-5.84E-01	1.96E+00	model2
520-18-3	7.15E-01	3.58E-01	1.07E+00	1.28E+00	model2
520-45-6	1.12E+00	8.41E-01	1.39E+00	3.64E-01	model2
5208-87-7	8.98E-01	7.04E-01	1.09E+00	1.86E-01	model2
52214-84-3	-1.25E-01	-6.10E-01	3.60E-01	9.99E-01	model2
53-03-2	5.92E-01	1.40E-02	1.17E+00	3.04E+00	model2
53-19-0	-1.65E+00	-2.04E+00	-1.26E+00	8.25E-01	model2
53-43-0	-1.34E-01	-8.81E-01	6.14E-01	2.05E+00	model2
53-86-1	-5.03E-01	-1.02E+00	1.88E-02	1.61E+00	model2
53-95-2	3.70E-01	1.35E-01	6.05E-01	8.15E-01	model2
5307-14-2	1.28E+00	1.14E+00	1.42E+00	3.94E-01	model2
531-18-0	9.38E-01	4.91E-01	1.38E+00	7.19E-01	model2
531-82-8	4.31E-01	1.93E-01	6.68E-01	4.51E-01	model2

531-85-1	6.84E-01	3.96E-01	9.71E-01	3.37E-01	model1
533-31-3	1.15E+00	1.03E+00	1.26E+00	2.40E-01	model2
536-33-4	1.25E+00	1.01E+00	1.50E+00	2.80E-01	model2
53609-64-6	2.26E+00	1.92E+00	2.59E+00	5.31E-01	model2
538-23-8	-1.32E+00	-1.73E+00	-9.13E-01	3.32E+00	model2
538-41-0	3.05E-01	-4.14E-02	6.51E-01	3.60E-01	model2
54-12-6	1.28E+00	1.01E+00	1.54E+00	4.27E-01	model2
54-31-9	3.40E-01	1.08E-02	6.69E-01	9.71E-01	model2
54-80-8	6.80E-01	5.03E-01	8.56E-01	4.52E-01	model2
540-23-8	1.24E+00	9.15E-01	1.56E+00	3.34E-01	model1
540-51-2	1.23E+00	1.02E+00	1.44E+00	7.45E-01	model2
541-69-5	1.43E+00	1.20E+00	1.66E+00	3.27E-01	model1
542-56-3	1.50E+00	1.09E+00	1.92E+00	7.01E-01	model2
542-88-1	1.39E+00	1.12E+00	1.66E+00	9.79E-01	model2
5456-28-0	-7.30E-02	-4.04E-01	2.58E-01	3.35E-01	model1
55-80-1	-6.10E-01	-7.27E-01	-4.92E-01	4.36E-01	model2
55-98-1	1.18E+00	8.33E-01	1.52E+00	3.84E-01	model2
55090-44-3	-5.86E-01	-9.47E-01	-2.25E-01	4.99E-01	model2
551-92-8	1.83E+00	1.66E+00	2.00E+00	3.84E-01	model2
5522-43-0	-1.16E+00	-1.46E+00	-8.62E-01	1.41E+00	model2
553-53-7	1.76E+00	1.18E+00	2.34E+00	4.75E-01	model2
55380-34-2	1.80E+00	1.52E+00	2.07E+00	3.42E-01	model2
555-84-0	2.72E-01	-2.20E-01	7.63E-01	3.95E-01	model2
55556-92-8	1.59E+00	1.31E+00	1.88E+00	5.24E-01	model2
55557-00-1	1.98E+00	1.81E+00	2.14E+00	4.98E-01	model2
55566-30-8	1.86E+00	1.49E+00	2.22E+00	1.00E+00	model1
55738-54-0	3.75E-01	6.31E-02	6.86E-01	6.62E-01	model2
56-04-2	1.04E+00	8.05E-01	1.28E+00	3.98E-01	model2
56-40-6	2.05E+00	1.74E+00	2.36E+00	7.79E-01	model2
56-86-0	1.97E+00	1.80E+00	2.13E+00	5.84E-01	model2
56038-13-2	9.61E-01	4.42E-01	1.48E+00	1.49E+00	model2
5632-47-3	2.11E+00	1.86E+00	2.36E+00	5.85E-01	model2
5634-39-9	7.87E-01	5.84E-01	9.90E-01	3.34E-01	model2
56654-52-5	6.03E-01	4.00E-01	8.06E-01	4.49E-01	model2
56795-65-4	1.83E+00	1.58E+00	2.08E+00	3.18E-01	model1
569-57-3	-1.66E+00	-2.15E+00	-1.18E+00	1.85E+00	model2
569-61-9	8.08E-01	4.59E-01	1.16E+00	3.36E-01	model1
56980-93-9	5.82E-01	3.03E-01	8.60E-01	1.67E+00	model2
57-30-7	1.72E+00	1.48E+00	1.96E+00	3.07E-01	model1

57-39-6	8.62E-01	4.85E-01	1.24E+00	1.32E+00	model2
57-41-0	5.73E-01	2.41E-01	9.04E-01	1.07E+00	model2
57-50-1	1.45E+00	8.07E-01	2.10E+00	1.49E+00	model2
57-57-8	1.86E+00	1.48E+00	2.25E+00	6.31E-01	model2
57-66-9	1.55E-01	-9.89E-02	4.09E-01	6.85E-01	model2
57-68-1	7.64E-01	4.42E-01	1.09E+00	7.63E-01	model2
57-97-6	-1.55E+00	-2.07E+00	-1.02E+00	1.52E+00	model2
57590-20-2	1.32E+00	1.12E+00	1.51E+00	6.26E-01	model2
57817-89-7	1.58E+00	1.15E+00	2.01E+00	3.37E-01	model1
58-15-1	8.68E-01	5.25E-01	1.21E+00	6.72E-01	model2
58-55-9	8.99E-01	5.82E-01	1.22E+00	5.76E-01	model2
58-93-5	7.33E-01	3.33E-01	1.13E+00	1.05E+00	model2
5834-17-3	1.46E-01	-5.25E-02	3.45E-01	6.83E-01	model2
59-05-2	8.02E-01	3.68E-01	1.24E+00	3.59E+00	model2
59-33-6	4.94E-01	-7.30E-02	1.06E+00	3.37E-01	model1
59-51-8	1.48E+00	1.14E+00	1.81E+00	6.30E-01	model2
59-88-1	1.50E+00	8.33E-01	2.16E+00	1.00E+00	model1
59-89-2	2.24E+00	2.10E+00	2.38E+00	8.21E-01	model2
590-21-6	1.08E+00	7.58E-01	1.40E+00	9.81E-01	model2
59122-46-2	-3.17E-01	-8.93E-01	2.59E-01	1.74E+00	model2
592-31-4	2.40E+00	2.23E+00	2.56E+00	7.10E-01	model2
593-60-2	1.18E+00	7.32E-01	1.63E+00	9.89E-01	model2
593-70-4	1.76E+00	1.35E+00	2.18E+00	4.55E-01	model1
597-25-1	1.35E+00	9.36E-01	1.76E+00	3.27E+00	model2
5979-28-2	-2.16E+00	-3.17E+00	-1.15E+00	1.15E+01	model2
598-55-0	2.19E+00	1.82E+00	2.55E+00	7.65E-01	model2
598-57-2	1.82E+00	1.62E+00	2.03E+00	7.78E-01	model2
599-79-1	-1.90E-01	-4.78E-01	9.88E-02	2.26E+00	model2
60-11-7	-3.41E-01	-5.52E-01	-1.31E-01	3.33E-01	model2
60-32-2	2.26E+00	2.02E+00	2.50E+00	7.00E-01	model2
60-56-0	1.60E+00	1.25E+00	1.95E+00	5.06E-01	model2
600-24-8	1.68E+00	1.36E+00	2.00E+00	6.18E-01	model2
60102-37-6	6.26E-01	2.46E-01	1.01E+00	2.14E+00	model2
60142-96-3	1.91E+00	1.61E+00	2.21E+00	3.81E-01	model2
602-87-9	-8.31E-02	-3.24E-01	1.58E-01	7.28E-01	model2
604-75-1	5.51E-01	3.31E-01	7.72E-01	8.54E-01	model2
60599-38-4	1.72E+00	1.47E+00	1.97E+00	5.31E-01	model2
607-35-2	1.11E+00	7.63E-01	1.45E+00	3.66E-01	model2
609-20-1	7.60E-01	5.57E-01	9.63E-01	2.97E-01	model2

61-76-7	1.07E+00	7.65E-01	1.37E+00	1.00E+00	model1
61-94-9	1.65E+00	1.35E+00	1.94E+00	1.00E+00	model1
6109-97-3	1.29E-02	-2.94E-01	3.20E-01	3.35E-01	model1
611-23-4	8.91E-01	6.88E-01	1.09E+00	4.01E-01	model2
611-32-5	8.89E-01	8.10E-01	9.68E-01	3.55E-01	model2
612-82-8	5.48E-01	1.84E-01	9.12E-01	3.36E-01	model1
613-94-5	1.82E+00	1.51E+00	2.12E+00	4.63E-01	model2
614-00-6	1.24E+00	1.08E+00	1.41E+00	3.88E-01	model2
614-95-9	1.62E+00	1.36E+00	1.88E+00	5.30E-01	model2
615-28-1	1.06E+00	5.29E-01	1.59E+00	3.21E-01	model1
615-53-2	1.82E+00	1.57E+00	2.07E+00	5.85E-01	model2
616-91-1	1.51E+00	1.29E+00	1.72E+00	5.50E-01	model2
617-84-5	2.37E+00	2.17E+00	2.58E+00	6.56E-01	model2
619-17-0	1.13E+00	1.03E+00	1.24E+00	3.79E-01	model2
62-23-7	1.02E+00	8.00E-01	1.24E+00	3.89E-01	model2
62-44-2	1.03E+00	8.33E-01	1.23E+00	2.46E-01	model2
62-54-4	2.56E+00	2.03E+00	3.09E+00	1.00E+00	model1
621-64-7	1.60E+00	1.33E+00	1.88E+00	5.84E-01	model2
622-51-5	1.62E+00	1.45E+00	1.80E+00	4.36E-01	model2
622-97-9	5.62E-01	4.19E-01	7.05E-01	6.17E-01	model2
624-18-0	9.98E-01	4.22E-01	1.57E+00	3.22E-01	model1
624-84-0	2.06E+00	1.74E+00	2.37E+00	8.17E-01	model2
625-89-8	6.63E-01	3.92E-01	9.35E-01	3.37E-01	model1
627-05-4	1.67E+00	1.45E+00	1.89E+00	7.04E-01	model2
628-02-4	2.18E+00	2.05E+00	2.31E+00	7.00E-01	model2
628-36-4	1.71E+00	1.42E+00	2.01E+00	7.86E-01	model2
628-94-4	2.08E+00	1.92E+00	2.24E+00	6.49E-01	model2
6294-89-9	2.30E+00	2.02E+00	2.57E+00	7.37E-01	model2
63412-06-6	1.15E+00	9.53E-01	1.34E+00	3.35E-01	model2
6358-85-6	-2.05E+00	-3.05E+00	-1.06E+00	8.96E+00	model2
636-21-5	1.26E+00	9.04E-01	1.62E+00	3.30E-01	model1
636-23-7	1.47E+00	1.01E+00	1.93E+00	3.11E-01	model1
6369-59-1	1.65E+00	1.32E+00	1.98E+00	1.00E+00	model1
637-07-0	1.47E-01	-5.07E-02	3.45E-01	4.91E-01	model2
6373-74-6	5.46E-01	2.42E-01	8.49E-01	3.37E-01	model1
638-03-9	1.12E+00	7.03E-01	1.54E+00	3.31E-01	model1
6381-77-7	2.03E+00	1.70E+00	2.35E+00	6.17E-01	model1
63885-23-4	1.60E+00	1.30E+00	1.90E+00	4.90E-01	model2
63886-77-1	1.23E+00	9.11E-01	1.55E+00	3.27E-01	model1

64-77-7	2.41E-01	1.43E-02	4.67E-01	4.73E-01	model2
64049-29-2	-4.01E-01	-7.42E-01	-6.00E-02	3.29E-01	model1
64091-91-4	8.95E-01	6.62E-01	1.13E+00	3.46E-01	model2
6459-94-5	1.17E+00	7.46E-01	1.59E+00	3.37E-01	model1
6471-49-4	-1.66E+00	-2.41E+00	-9.03E-01	4.30E+00	model2
6485-34-3	1.48E+00	1.16E+00	1.79E+00	3.32E-01	model1
66-22-8	1.42E+00	1.13E+00	1.70E+00	5.44E-01	model2
66-27-3	1.82E+00	1.36E+00	2.28E+00	6.08E-01	model2
6673-35-4	8.38E-01	5.10E-01	1.17E+00	5.09E-01	model2
67-20-9	3.53E-01	-4.69E-02	7.53E-01	4.62E-01	model2
67-21-0	1.42E+00	1.10E+00	1.74E+00	6.10E-01	model2
67-52-7	1.31E+00	1.00E+00	1.62E+00	4.79E-01	model2
6731-36-8	-1.05E+00	-1.27E+00	-8.33E-01	1.09E+00	model2
67730-10-3	5.83E-01	2.53E-01	9.14E-01	5.44E-01	model2
67730-11-4	3.11E-01	1.34E-01	4.88E-01	6.67E-01	model2
68-23-5	-6.56E-01	-1.27E+00	-4.26E-02	2.10E+00	model2
68-89-3	1.58E+00	1.28E+00	1.88E+00	1.00E+00	model1
68844-77-9	-1.80E+00	-2.33E+00	-1.27E+00	2.65E+00	model1
69-65-8	2.75E+00	2.43E+00	3.07E+00	4.28E-01	model2
695-53-4	2.11E+00	1.68E+00	2.55E+00	3.81E-01	model2
6959-48-4	1.83E+00	1.58E+00	2.07E+00	3.10E-01	model1
70-25-7	1.64E+00	1.35E+00	1.94E+00	5.59E-01	model2
7003-89-6	2.12E+00	1.76E+00	2.48E+00	1.00E+00	model1
71125-38-7	-2.00E-01	-5.12E-01	1.12E-01	1.66E+00	model2
712-68-5	9.72E-01	7.49E-01	1.19E+00	3.77E-01	model2
7177-48-2	9.87E-01	5.72E-01	1.40E+00	3.34E-01	model1
72-33-3	-1.22E+00	-1.80E+00	-6.38E-01	2.17E+00	model2
720-69-4	9.49E-01	6.66E-01	1.23E+00	4.71E-01	model2
72254-58-1	1.86E-01	-2.64E-02	3.98E-01	3.37E-01	model1
7227-91-0	8.44E-01	5.73E-01	1.11E+00	4.13E-01	model2
7235-40-7	-1.76E+00	-2.18E+00	-1.33E+00	7.54E+00	model2
73-22-3	1.28E+00	1.01E+00	1.54E+00	4.27E-01	model2
7336-20-1	1.39E+00	9.99E-01	1.78E+00	1.00E+00	model1
7347-49-1	-8.22E-02	-2.83E-01	1.18E-01	6.32E-01	model2
73590-58-6	3.27E-01	1.84E-03	6.53E-01	1.46E+00	model2
74-31-7	-4.37E-01	-6.49E-01	-2.26E-01	6.60E-01	model2
74-96-4	1.14E+00	6.40E-01	1.63E+00	9.90E-01	model2
7411-49-6	7.68E-01	3.54E-01	1.18E+00	3.24E-01	model1
7422-80-2	1.36E+00	1.21E+00	1.51E+00	6.50E-01	model2

75-00-3	1.60E+00	1.25E+00	1.95E+00	9.90E-01	model2
75-01-4	1.38E+00	9.75E-01	1.78E+00	9.90E-01	model2
75-02-5	1.73E+00	1.28E+00	2.18E+00	2.87E+00	model1
75-34-3	1.69E+00	1.52E+00	1.86E+00	9.68E-01	model2
75-38-7	1.35E+00	1.15E+00	1.55E+00	4.35E+00	model1
75-45-6	1.64E+00	1.33E+00	1.95E+00	1.07E+00	model1
75-69-4	1.29E+00	1.11E+00	1.46E+00	3.41E-01	model1
75-71-8	1.25E+00	1.06E+00	1.44E+00	7.44E-01	model1
75-88-7	1.67E+00	1.48E+00	1.85E+00	3.73E-01	model1
75104-43-7	4.66E-01	2.21E-01	7.10E-01	3.37E-01	model1
75330-75-5	-7.80E-01	-1.40E+00	-1.57E-01	2.25E+00	model2
756-79-6	2.14E+00	1.63E+00	2.65E+00	1.36E+00	model2
7572-29-4	1.61E+00	1.40E+00	1.82E+00	9.81E-01	model2
758-17-8	2.25E+00	1.93E+00	2.57E+00	7.64E-01	model2
759-73-9	1.96E+00	1.77E+00	2.16E+00	6.26E-01	model2
76-13-1	6.40E-01	4.52E-01	8.27E-01	3.38E-01	model1
76-25-5	4.70E-01	7.44E-02	8.66E-01	3.37E-01	model1
76-57-3	6.57E-01	3.87E-01	9.26E-01	2.57E+00	model2
760-60-1	1.91E+00	1.63E+00	2.18E+00	5.48E-01	model2
76180-96-6	4.92E-01	1.11E-01	8.73E-01	3.37E-01	model1
764-41-0	7.81E-01	5.72E-01	9.90E-01	9.53E-01	model2
765-34-4	1.65E+00	1.34E+00	1.95E+00	8.88E-01	model2
7681-93-8	1.43E+00	8.62E-01	2.00E+00	3.35E-01	model1
77-06-5	8.41E-01	4.12E-01	1.27E+00	3.68E+00	model2
77-09-8	2.15E-01	-7.02E-02	5.00E-01	1.86E+00	model2
77-46-3	6.10E-01	2.46E-01	9.75E-01	1.21E+00	model2
77-65-6	7.63E-01	4.98E-01	1.03E+00	3.60E-01	model2
77-79-2	1.65E+00	1.32E+00	1.99E+00	4.84E-01	model2
77-83-8	4.02E-01	2.63E-01	5.41E-01	6.99E-01	model2
785-30-8	9.26E-01	7.17E-01	1.14E+00	4.45E-01	model2
79-24-3	1.68E+00	1.53E+00	1.82E+00	7.45E-01	model2
79-40-3	1.03E+00	6.74E-01	1.39E+00	6.62E-01	model2
79-44-7	1.73E+00	1.47E+00	1.98E+00	6.19E-01	model2
80-07-9	-3.13E-01	-4.88E-01	-1.39E-01	6.68E-01	model2
80-08-0	8.26E-01	5.14E-01	1.14E+00	6.84E-01	model2
8015-30-3	-7.89E-02	-3.34E-01	1.76E-01	3.36E-01	model1
81-15-2	-6.64E-01	-9.44E-01	-3.83E-01	1.32E+00	model2
81-16-3	1.34E+00	9.55E-01	1.73E+00	7.10E-01	model2
81-21-0	1.61E+00	1.31E+00	1.91E+00	1.78E+00	model2

81-49-2	-1.00E+00	-1.25E+00	-7.47E-01	1.34E+00	model2
811-97-2	1.57E+00	1.40E+00	1.74E+00	7.62E-01	model1
816-57-9	1.88E+00	1.60E+00	2.15E+00	6.01E-01	model2
82-28-0	-9.13E-02	-3.32E-01	1.49E-01	1.18E+00	model2
838-88-0	2.21E-01	7.46E-02	3.68E-01	4.86E-01	model2
842-00-2	2.67E-01	-1.59E-01	6.93E-01	1.14E+00	model2
842-07-9	-6.37E-01	-8.11E-01	-4.63E-01	7.01E-01	model2
846-50-4	4.72E-01	2.87E-01	6.57E-01	1.02E+00	model2
853-23-6	-7.95E-01	-1.55E+00	-4.06E-02	2.29E+00	model2
86-29-3	4.80E-01	2.55E-01	7.04E-01	2.19E-01	model2
86-86-2	6.82E-01	4.15E-01	9.49E-01	3.67E-01	model2
86-88-4	8.70E-01	6.77E-01	1.06E+00	2.87E-01	model2
860-22-0	1.55E+00	1.33E+00	1.76E+00	3.33E-01	model1
86315-52-8	7.83E-01	4.92E-01	1.07E+00	3.37E-01	model1
869-01-2	1.83E+00	1.58E+00	2.08E+00	5.80E-01	model2
87-29-6	-1.20E-01	-3.70E-01	1.30E-01	5.28E-01	model2
88-19-7	1.44E+00	1.22E+00	1.65E+00	3.54E-01	model2
88-96-0	1.47E+00	1.20E+00	1.73E+00	3.75E-01	model2
88107-10-2	-9.75E-02	-3.26E-01	1.31E-01	1.09E+00	model2
89-25-8	8.90E-01	5.87E-01	1.19E+00	2.16E-01	model2
90-49-3	1.22E+00	9.95E-01	1.44E+00	3.38E-01	model2
90-94-8	1.04E-01	-9.05E-02	2.98E-01	7.05E-01	model2
91-59-8	6.96E-01	4.51E-01	9.42E-01	2.34E-01	model2
91-62-3	9.58E-01	8.87E-01	1.03E+00	2.54E-01	model2
91-76-9	1.02E+00	7.50E-01	1.28E+00	3.70E-01	model2
91-79-2	5.03E-01	2.69E-01	7.37E-01	3.06E-01	model2
91-93-0	-5.55E-01	-7.69E-01	-3.41E-01	9.64E-01	model2
915-67-3	1.75E+00	1.33E+00	2.17E+00	3.15E-01	model1
92-55-7	6.27E-01	2.56E-01	9.97E-01	3.83E-01	model2
92-67-1	4.97E-01	3.32E-01	6.63E-01	1.94E-01	model2
92-84-2	-4.19E-01	-7.65E-01	-7.28E-02	5.33E-01	model2
924-16-3	1.02E+00	8.90E-01	1.15E+00	5.37E-01	model2
924-42-5	1.71E+00	1.42E+00	2.00E+00	6.98E-01	model2
93-46-9	-1.85E+00	-2.68E+00	-1.02E+00	2.55E+00	model2
930-55-2	2.19E+00	1.74E+00	2.64E+00	5.92E-01	model2
932-83-2	1.66E+00	1.39E+00	1.93E+00	5.42E-01	model2
934-00-9	1.22E+00	1.04E+00	1.41E+00	3.62E-01	model2
938-73-8	1.29E+00	1.10E+00	1.49E+00	3.26E-01	model2
93957-54-1	-9.12E-01	-1.33E+00	-4.96E-01	2.70E-01	model1

94-20-2	3.79E-01	8.96E-02	6.68E-01	4.11E-01	model2
94-26-8	3.84E-01	2.37E-01	5.31E-01	3.09E-01	model2
94-36-0	2.71E-01	1.16E-01	4.25E-01	4.34E-01	model2
94-58-6	3.66E-01	2.03E-01	5.28E-01	4.97E-01	model2
94-59-7	3.59E-01	1.47E-01	5.71E-01	4.97E-01	model2
95-71-6	8.14E-01	4.66E-01	1.16E+00	4.15E-01	model2
95-79-4	8.29E-01	6.51E-01	1.01E+00	3.23E-01	model2
95-83-0	9.28E-01	7.14E-01	1.14E+00	4.10E-01	model2
959-24-0	3.61E-01	-6.37E-02	7.87E-01	3.05E-01	model1
96-69-5	-1.71E+00	-2.35E+00	-1.07E+00	1.72E+00	model2
968-81-0	4.58E-01	2.64E-02	8.89E-01	1.01E+00	model2
97-16-5	-3.07E-01	-5.20E-01	-9.34E-02	5.99E-01	model2
97-18-7	-1.10E+00	-1.52E+00	-6.74E-01	8.73E-01	model2
97-56-3	-2.37E-01	-4.46E-01	-2.86E-02	3.74E-01	model2
97-59-6	1.68E+00	1.38E+00	1.97E+00	4.35E-01	model2
971-15-3	-1.08E+00	-1.36E+00	-8.02E-01	1.05E+00	model2
98-85-1	1.57E+00	1.38E+00	1.77E+00	4.00E-01	model2
98-96-4	1.64E+00	1.41E+00	1.88E+00	5.11E-01	model2
98319-26-7	-2.63E-01	-6.84E-01	1.57E-01	2.79E+00	model2
989-38-8	1.83E-02	-1.90E-01	2.27E-01	3.18E-01	model1
99-50-3	1.52E+00	1.39E+00	1.64E+00	4.02E-01	model2
99-57-0	1.15E+00	1.00E+00	1.30E+00	3.97E-01	model2
99-59-2	8.30E-01	6.12E-01	1.05E+00	3.64E-01	model2

Table C-2. Ecotoxicity characterization factors (Midpoint, [PAF.m3.day/kg emitted]) calculated based on the estimated HC₅₀.

CAS	Em.hom.air I	Em.ind.air I	Em.airU	Em.airC	Em.fr.water C	Em.sea waterC	Em.nat.soil C	Em.agr.soil C
100-40-3	5.21E-03	7.66E-03	9.29E-03	1.14E-03	4.60E+02	1.03E-04	5.43E-01	5.43E-01
100-75-4	3.64E+00	3.77E+00	3.86E+00	3.42E+00	1.46E+02	1.17E-02	4.48E+01	4.48E+01
10034-93-2	3.20E+01	3.22E+01	3.24E+01	3.16E+01	1.56E+02	2.02E-21	5.05E+01	3.59E+01
10048-13-2	3.77E+02	5.11E+02	6.00E+02	1.54E+02	3.55E+04	2.30E-05	4.50E+02	4.50E+02
101-14-4	4.15E+02	5.46E+02	6.32E+02	1.99E+02	3.79E+04	9.19E-02	8.30E+02	8.21E+02
101-61-1	6.54E+02	9.30E+02	1.11E+03	1.94E+02	7.92E+04	2.43E-01	1.53E+03	1.20E+03
101-79-1	1.13E+02	1.54E+02	1.80E+02	4.64E+01	1.51E+04	5.34E-02	2.39E+02	2.29E+02
101-80-4	4.54E+01	5.55E+01	6.23E+01	2.84E+01	2.92E+03	2.15E-05	2.46E+02	2.81E+02
101-90-6	4.51E+01	4.95E+01	5.25E+01	3.77E+01	2.62E+03	9.62E-03	8.09E+02	8.09E+02
10191-41-0	1.04E+01	1.57E+01	1.92E+01	1.52E+00	2.89E+03	1.78E-02	1.27E+00	1.27E+00
102-09-0	1.27E+02	1.39E+02	1.47E+02	1.07E+02	4.73E+03	1.10E-01	6.47E+01	6.47E+01

102-50-1	8.50E+00	1.10E+01	1.26E+01	4.38E+00	1.28E+03	1.04E-02	6.12E+01	1.21E+02
102-77-2	2.11E+02	2.16E+02	2.19E+02	2.02E+02	5.54E+03	2.32E-04	1.97E+03	1.97E+03
103-03-7	9.50E+00	1.00E+01	1.04E+01	8.65E+00	2.25E+02	2.26E-05	4.63E+01	4.63E+01
103-16-2	6.24E+01	7.83E+01	8.89E+01	3.60E+01	4.17E+03	7.61E-04	4.16E+01	4.16E+01
104-46-1	7.74E-01	8.09E-01	8.32E-01	7.16E-01	1.12E+03	4.25E-02	1.04E+01	1.04E+01
10473-70-8	3.31E+02	4.21E+02	4.80E+02	1.82E+02	3.74E+04	6.25E-01	5.48E+01	5.48E+01
105-11-3	5.78E+00	7.50E+00	8.65E+00	2.92E+00	4.80E+02	7.79E-10	2.14E-01	2.14E-01
10589-74-9	1.90E+01	1.95E+01	1.99E+01	1.81E+01	2.32E+02	2.63E-03	4.23E+01	4.23E+01
10595-95-6	6.43E+00	6.56E+00	6.64E+00	6.23E+00	1.22E+02	2.22E-02	3.89E+01	3.89E+01
106-87-6	5.79E+00	5.94E+00	6.04E+00	5.54E+00	1.16E+02	6.61E-03	2.89E+01	2.89E+01
106-99-0	3.77E-04	4.66E-04	5.25E-04	2.29E-04	9.64E+01	2.44E-05	5.08E-01	5.08E-01
1068-57-1	1.02E+01	1.03E+01	1.04E+01	1.01E+01	6.13E+01	5.54E-04	1.89E+01	1.91E+01
107-30-2	1.87E+00	1.87E+00	1.87E+00	1.87E+00	8.57E+01	8.77E-02	1.66E+01	1.66E+01
107-35-7	8.08E+00	8.22E+00	8.32E+00	7.84E+00	1.11E+02	4.52E-09	3.72E+01	3.71E+01
1078-38-2	1.04E+02	1.05E+02	1.05E+02	1.03E+02	6.98E+02	8.08E-07	2.62E+02	2.62E+02
108-30-5	2.58E+01	2.60E+01	2.62E+01	2.54E+01	1.23E+02	3.91E-04	3.33E+01	3.33E+01
108-60-1	3.62E+00	3.62E+00	3.62E+00	3.62E+00	2.63E+02	3.81E-01	3.56E+01	3.56E+01
1083-57-4	4.66E+01	5.13E+01	5.44E+01	3.89E+01	1.54E+03	4.07E-08	1.63E+02	1.63E+02
109-84-2	1.13E+00	1.34E+00	1.47E+00	7.89E-01	6.03E+01	2.38E-06	3.93E+00	4.42E+00
1095-90-5	1.47E+03	1.48E+03	1.48E+03	1.46E+03	5.71E+03	1.24E-10	2.05E+03	2.05E+03
110-85-0	5.55E-01	7.58E-01	8.94E-01	2.17E-01	4.89E+01	7.47E-08	6.78E-01	6.79E-01
110-89-4	1.41E+00	1.90E+00	2.23E+00	5.86E-01	1.22E+02	6.33E-06	2.70E-02	2.70E-02
1116-54-7	1.98E+00	2.01E+00	2.03E+00	1.92E+00	2.54E+01	2.65E-09	8.54E+00	8.54E+00
1119-68-2	9.53E-01	1.29E+00	1.51E+00	3.92E-01	9.02E+01	3.52E-03	5.81E-02	6.27E-02
112-63-0	6.54E+01	9.78E+01	1.19E+02	1.15E+01	2.58E+04	2.68E-01	1.45E+00	1.45E+00
1120-71-4	2.04E+01	2.06E+01	2.07E+01	2.01E+01	1.21E+02	7.77E-03	3.68E+01	3.68E+01
1121-92-2	2.68E+00	3.64E+00	4.28E+00	1.08E+00	2.37E+02	2.41E-05	4.52E-02	4.52E-02
1133-64-8	7.84E+01	8.51E+01	8.97E+01	6.70E+01	2.32E+03	3.01E-05	1.67E+02	4.18E+02
114-83-0	7.39E+00	8.03E+00	8.45E+00	6.32E+00	2.30E+02	4.98E-06	3.27E+01	3.27E+01
115-07-1	3.33E-04	4.01E-04	4.47E-04	2.18E-04	9.52E+01	2.45E-05	2.93E-01	2.93E-01
115-09-3	2.26E+02	2.29E+02	2.32E+02	2.20E+02	2.41E+03	2.54E-01	8.14E+02	8.14E+02
115-11-7	2.21E-04	2.97E-04	3.48E-04	9.33E-05	9.12E+01	9.82E-06	2.03E-01	2.03E-01
115-28-6	2.07E+03	2.08E+03	2.09E+03	2.05E+03	1.15E+04	1.49E-09	3.98E+03	3.98E+03
1156-19-0	3.48E+02	3.78E+02	3.98E+02	2.97E+02	9.89E+03	4.92E-06	1.09E+03	1.09E+03
116-14-3	2.94E-03	2.99E-03	3.02E-03	2.86E-03	1.01E+02	2.69E-04	1.72E-01	1.72E-01
1163-19-5	4.29E-01	4.89E-01	5.29E-01	3.28E-01	1.59E+01	1.85E-08	6.80E-03	6.80E-03
117-39-5	1.84E+02	1.92E+02	1.97E+02	1.72E+02	2.42E+03	3.88E-12	1.87E+02	1.87E+02
117-79-3	3.94E+02	4.75E+02	5.29E+02	2.59E+02	2.22E+04	7.83E-05	5.02E+02	5.01E+02
118-92-3	1.07E+01	1.17E+01	1.24E+01	8.95E+00	3.44E+02	6.75E-08	3.95E+01	3.88E+01

119-47-1	5.36E+03	6.85E+03	7.84E+03	2.89E+03	4.06E+05	5.94E-01	4.81E+02	4.81E+02
119-53-9	9.32E+01	9.93E+01	1.03E+02	8.30E+01	1.83E+03	1.40E-03	1.61E+02	1.61E+02
119-84-6	3.41E+01	3.49E+01	3.54E+01	3.29E+01	4.58E+02	2.83E-02	1.03E+02	1.03E+02
1192-28-5	6.88E+00	7.02E+00	7.11E+00	6.65E+00	5.69E+01	1.30E-03	1.08E+01	1.08E+01
120-32-1	4.32E+02	5.12E+02	5.66E+02	2.98E+02	2.05E+04	2.65E-03	1.74E+02	1.74E+02
120-58-1	2.19E+00	2.22E+00	2.25E+00	2.12E+00	2.07E+03	2.00E-01	4.68E+01	4.68E+01
120-71-8	1.93E+00	2.46E+00	2.82E+00	1.03E+00	8.88E+02	7.77E-03	5.12E+01	5.84E+01
120-78-5	2.85E+03	3.47E+03	3.88E+03	1.81E+03	1.79E+05	1.41E-04	4.89E+02	4.89E+02
121-19-7	1.17E+02	1.19E+02	1.20E+02	1.15E+02	7.72E+02	4.89E-12	1.88E+02	1.88E+02
121-59-5	4.33E+01	4.47E+01	4.57E+01	4.09E+01	8.21E+02	3.33E-12	2.26E+02	2.26E+02
121-66-4	3.33E+02	3.36E+02	3.38E+02	3.28E+02	1.59E+03	1.89E-03	4.17E+02	4.17E+02
121-88-0	1.54E+02	1.57E+02	1.59E+02	1.50E+02	1.34E+03	2.96E-05	3.28E+02	3.28E+02
1212-29-9	1.34E+02	1.87E+02	2.23E+02	4.43E+01	1.39E+04	5.18E-03	2.30E+02	2.30E+02
122-20-3	4.86E-01	6.45E-01	7.52E-01	2.20E-01	4.00E+01	9.60E-10	7.95E-01	8.70E-01
123-73-9	2.04E+00	2.02E+00	2.01E+00	2.07E+00	4.76E+02	5.95E-02	9.17E+01	9.17E+01
124-58-3	1.19E+01	1.21E+01	1.23E+01	1.15E+01	8.37E+01	4.16E-08	1.37E+01	1.37E+01
124-64-1	3.23E+00	3.28E+00	3.32E+00	3.13E+00	3.10E+01	1.95E-10	8.45E+00	8.45E+00
1248-18-6	8.80E+02	8.88E+02	8.94E+02	8.65E+02	4.70E+03	8.74E-17	1.17E+03	1.17E+03
125-33-7	6.40E+01	6.60E+01	6.73E+01	6.08E+01	9.73E+02	3.83E-08	2.22E+02	2.22E+02
126-07-8	7.96E+02	8.81E+02	9.39E+02	6.52E+02	3.90E+04	6.18E-04	8.59E+03	8.59E+03
126-13-6	1.69E+04	1.92E+04	2.07E+04	1.31E+04	6.02E+05	1.66E-06	1.36E+03	1.36E+03
126-98-7	5.49E-01	5.47E-01	5.45E-01	5.53E-01	1.55E+02	4.69E-02	1.94E+01	1.94E+01
12663-46-6	3.78E+03	3.78E+03	3.78E+03	3.78E+03	1.15E+04	1.04E-26	5.41E+03	5.41E+03
127-47-9	5.46E+00	8.71E+00	1.09E+01	4.15E-02	4.37E+03	5.83E-04	3.91E-01	3.91E-01
127-48-0	5.02E+01	5.03E+01	5.04E+01	4.99E+01	2.26E+02	2.39E-04	8.78E+01	8.78E+01
127-69-5	3.28E+02	3.40E+02	3.49E+02	3.07E+02	5.93E+03	2.85E-05	1.38E+03	1.38E+03
1271-19-8	1.52E+02	2.23E+02	2.71E+02	3.24E+01	6.41E+04	1.63E-01	1.79E+02	1.79E+02
128-66-5	3.59E+03	4.10E+03	4.44E+03	2.75E+03	1.33E+05	1.10E-04	2.91E+01	2.91E+01
129-15-7	1.55E+03	1.71E+03	1.82E+03	1.29E+03	3.61E+04	1.01E-02	5.48E+02	5.48E+02
129-43-1	5.42E+02	6.23E+02	6.77E+02	4.07E+02	2.07E+04	9.67E-03	3.87E+02	3.87E+02
13010-07-6	1.31E+01	1.54E+01	1.70E+01	9.10E+00	5.80E+02	2.34E-08	8.61E-01	8.62E-01
13073-35-3	1.82E+01	1.87E+01	1.90E+01	1.74E+01	3.59E+02	1.05E-07	1.16E+02	1.15E+02
131-01-1	1.68E+04	1.72E+04	1.75E+04	1.61E+04	1.52E+05	3.60E-07	2.00E+04	2.00E+04
13256-06-9	1.13E+01	1.22E+01	1.28E+01	9.76E+00	2.79E+03	2.03E-01	1.29E+01	1.29E+01
13256-11-6	9.89E+01	1.00E+02	1.02E+02	9.61E+01	1.11E+03	2.19E-02	3.38E+02	3.38E+02
13292-46-1	1.50E+03	1.62E+03	1.71E+03	1.30E+03	3.43E+04	2.36E-30	9.20E+02	9.20E+02
134-03-2	1.10E+01	1.10E+01	1.11E+01	1.08E+01	5.40E+01	9.20E-21	1.47E+01	1.47E+01
134-29-2	2.37E+01	2.38E+01	2.39E+01	2.34E+01	1.29E+02	5.85E-09	4.32E+01	4.32E+01
135-20-6	1.18E+01	1.20E+01	1.20E+01	1.17E+01	9.09E+01	1.81E-10	3.07E+01	3.07E+01

135-23-9	1.44E+02	1.44E+02	1.44E+02	1.43E+02	3.65E+03	3.85E-06	1.60E+03	1.60E+03
13552-44-8	7.92E+01	8.40E+01	8.72E+01	7.12E+01	2.99E+03	1.32E-04	8.27E+02	8.27E+02
136-23-2	7.62E+00	1.22E+01	1.52E+01	5.90E-02	3.80E+04	1.05E-03	9.97E-01	9.97E-01
136-40-3	4.38E+02	4.38E+02	4.38E+02	4.38E+02	1.31E+03	1.53E-09	5.78E+02	5.78E+02
136-77-6	4.36E+01	6.15E+01	7.35E+01	1.38E+01	4.22E+03	1.23E-04	3.51E+01	3.51E+01
137-09-7	3.00E+01	3.01E+01	3.02E+01	2.98E+01	1.19E+02	1.68E-15	4.01E+01	4.01E+01
137-17-7	3.15E+00	4.27E+00	5.02E+00	1.27E+00	1.19E+03	1.17E-02	3.26E+01	5.36E+01
13743-07-2	1.55E+01	1.57E+01	1.58E+01	1.53E+01	1.07E+02	7.57E-06	3.58E+01	3.58E+01
13752-51-7	1.66E+02	1.67E+02	1.67E+02	1.65E+02	4.87E+03	2.17E-05	2.03E+03	2.03E+03
13838-16-9	1.14E+00	1.14E+00	1.14E+00	1.13E+00	1.22E+02	2.00E-01	4.27E+00	4.27E+00
139-65-1	5.09E+01	6.41E+01	7.30E+01	2.88E+01	3.61E+03	6.06E-04	1.97E+02	1.96E+02
139-94-6	4.84E+02	4.94E+02	5.01E+02	4.68E+02	3.97E+03	8.38E-05	7.92E+02	7.92E+02
13927-77-0	1.40E+01	2.13E+01	2.61E+01	1.94E+00	5.18E+04	6.88E-02	8.13E+00	8.13E+00
140-67-0	6.89E-01	7.01E-01	7.09E-01	6.70E-01	9.74E+02	9.45E-02	1.68E+01	1.68E+01
1402-68-2	3.90E+02	3.93E+02	3.96E+02	3.84E+02	6.58E+03	8.95E-09	2.58E+03	2.58E+03
14026-03-0	8.06E+00	8.43E+00	8.67E+00	7.44E+00	3.06E+02	1.35E-02	9.43E+01	9.43E+01
142-04-1	3.63E+01	3.64E+01	3.66E+01	3.60E+01	1.43E+02	4.62E-08	4.81E+01	4.81E+01
142-46-1	2.01E+01	2.12E+01	2.19E+01	1.83E+01	7.47E+02	6.09E-04	2.36E+02	2.36E+02
142-83-6	6.76E-01	6.91E-01	7.00E-01	6.52E-01	2.46E+02	5.91E-03	2.65E+01	2.65E+01
14239-68-0	1.12E+02	1.24E+02	1.32E+02	9.29E+01	4.85E+03	8.88E-06	9.58E+02	9.58E+02
143-19-1	6.22E+02	7.02E+02	7.55E+02	4.88E+02	2.11E+04	5.96E-08	8.90E+01	8.90E+01
14371-10-9	1.87E+00	2.00E+00	2.09E+00	1.64E+00	3.93E+02	9.16E-03	3.90E+01	3.90E+01
144-02-5	9.87E+01	9.89E+01	9.90E+01	9.85E+01	3.33E+02	2.76E-17	1.40E+02	1.40E+02
144-34-3	6.73E+02	6.74E+02	6.74E+02	6.73E+02	3.58E+03	2.18E-10	1.59E+03	1.59E+03
144-48-9	4.04E+02	4.06E+02	4.07E+02	4.01E+02	1.78E+03	1.13E-02	6.14E+02	6.14E+02
1456-28-6	8.83E+00	8.98E+00	9.08E+00	8.59E+00	1.57E+02	1.19E-03	6.03E+01	6.03E+01
148-82-3	2.60E+02	2.66E+02	2.70E+02	2.49E+02	8.51E+03	2.90E-06	3.08E+03	3.08E+03
149-29-1	1.62E+01	1.67E+01	1.70E+01	1.54E+01	4.67E+02	3.40E-08	1.57E+02	1.57E+02
149-91-7	7.37E+00	8.43E+00	9.14E+00	5.60E+00	3.51E+02	2.97E-11	4.32E+01	4.32E+01
150-38-9	3.27E+00	3.47E+00	3.60E+00	2.94E+00	1.18E+02	7.70E-20	3.21E+01	3.21E+01
150-69-6	2.83E+01	3.14E+01	3.36E+01	2.29E+01	1.02E+03	1.56E-04	1.04E+02	1.04E+02
153-18-4	6.75E+02	6.78E+02	6.80E+02	6.70E+02	2.77E+03	2.71E-23	9.34E+02	9.34E+02
15318-45-3	6.59E+02	6.62E+02	6.63E+02	6.55E+02	2.60E+03	6.49E-12	9.13E+02	9.13E+02
15356-70-4	1.68E+00	1.72E+00	1.74E+00	1.61E+00	3.97E+02	4.54E-02	5.72E+00	5.72E+00
155-04-4	3.04E+03	3.83E+03	4.36E+03	1.72E+03	2.22E+05	2.19E-03	3.21E+02	3.21E+02
156-10-5	3.91E+01	5.41E+01	6.41E+01	1.41E+01	9.36E+03	2.67E-02	3.53E+02	3.53E+02
156-51-4	3.32E+02	3.32E+02	3.32E+02	3.31E+02	1.17E+03	2.12E-14	4.92E+02	4.92E+02
156-62-7	5.68E+01	5.71E+01	5.73E+01	5.63E+01	2.41E+02	2.81E-20	7.83E+01	7.83E+01
15879-93-3	3.93E+02	4.00E+02	4.04E+02	3.81E+02	4.52E+03	1.71E-07	1.23E+03	1.23E+03

16071-86-6	2.34E+02	2.34E+02	2.34E+02	2.34E+02	6.96E+02	2.65E-41	3.35E+02	3.35E+02
16301-26-1	1.54E+01	1.59E+01	1.63E+01	1.45E+01	1.93E+02	1.69E-04	2.99E+01	2.99E+01
16338-97-9	6.29E+00	6.77E+00	7.09E+00	5.48E+00	8.59E+02	2.94E-02	2.28E+02	2.28E+02
1643-20-5	1.34E+02	1.53E+02	1.65E+02	1.03E+02	4.89E+03	3.58E-08	5.27E+00	5.27E+00
16568-02-8	4.58E+00	4.89E+00	5.11E+00	4.05E+00	1.70E+02	3.86E-03	2.16E+01	2.26E+01
16699-10-8	1.25E+03	1.58E+03	1.80E+03	7.07E+02	8.93E+04	2.09E-03	2.27E+03	1.59E+03
16813-36-8	1.44E+01	1.48E+01	1.52E+01	1.36E+01	1.98E+02	1.80E-06	3.90E+01	3.90E+01
16846-24-5	2.19E+03	2.32E+03	2.42E+03	1.95E+03	4.03E+04	2.09E-18	1.17E+03	2.30E+03
169590-42-5	1.49E+03	1.67E+03	1.79E+03	1.18E+03	4.84E+04	6.82E-04	1.50E+03	1.50E+03
17026-81-2	3.92E+01	4.18E+01	4.36E+01	3.48E+01	1.53E+03	3.33E-06	3.90E+02	4.15E+02
1717-00-6	1.35E+00	1.35E+00	1.35E+00	1.35E+00	1.47E+02	2.62E-01	3.21E+00	3.21E+00
17608-59-2	8.51E+01	8.80E+01	8.99E+01	8.03E+01	1.53E+03	8.33E-06	3.95E+02	3.95E+02
1777-84-0	4.51E+02	4.78E+02	4.96E+02	4.05E+02	7.78E+03	8.27E-04	6.25E+02	6.25E+02
17924-92-4	3.65E+02	4.03E+02	4.28E+02	3.02E+02	1.02E+04	1.24E-07	1.80E+02	1.80E+02
18523-69-8	6.95E+02	7.69E+02	8.18E+02	5.72E+02	1.97E+04	9.26E-03	8.56E+02	8.36E+02
18559-94-9	1.30E+01	1.60E+01	1.81E+01	7.81E+00	8.74E+02	5.58E-11	2.35E-01	2.36E-01
18662-53-8	3.01E+00	3.14E+00	3.23E+00	2.78E+00	7.64E+01	6.04E-13	2.08E+01	2.08E+01
18699-02-0	1.76E+01	1.85E+01	1.91E+01	1.61E+01	2.99E+02	6.21E-09	3.82E+01	3.82E+01
19010-66-3	3.70E+02	3.74E+02	3.76E+02	3.63E+02	9.27E+03	3.36E-04	3.82E+03	3.82E+03
191-24-2	3.64E+03	5.08E+03	6.04E+03	1.24E+03	5.05E+05	6.80E+00	1.30E+03	1.30E+03
193-39-5	5.36E+03	7.31E+03	8.62E+03	2.09E+03	6.88E+05	1.46E+01	1.99E+03	1.99E+03
1936-15-8	1.29E+02	1.30E+02	1.30E+02	1.29E+02	4.36E+02	6.48E-20	1.83E+02	1.83E+02
1937-37-7	7.39E+04	7.45E+04	7.48E+04	7.30E+04	3.53E+05	1.05E-27	9.51E+04	1.06E+05
1955-45-9	4.01E+00	4.05E+00	4.08E+00	3.95E+00	4.38E+01	9.43E-03	1.15E+01	1.15E+01
20265-96-7	9.41E+01	9.42E+01	9.43E+01	9.40E+01	3.03E+02	6.73E-08	1.27E+02	1.27E+02
20325-40-0	7.85E+02	7.87E+02	7.89E+02	7.82E+02	2.83E+03	6.32E-14	1.10E+03	1.10E+03
205-99-2	1.57E+04	1.93E+04	2.16E+04	9.74E+03	1.21E+06	1.08E+02	5.61E+03	5.61E+03
207-08-9	1.60E+04	2.12E+04	2.46E+04	7.45E+03	1.90E+06	1.25E+02	2.54E+04	2.54E+04
208-96-8	8.26E-01	8.92E-01	9.36E-01	7.17E-01	4.13E+03	1.06E-01	5.28E+01	5.28E+01
20917-49-1	1.29E+01	1.36E+01	1.41E+01	1.18E+01	6.03E+02	3.25E-02	1.45E+02	1.45E+02
20941-65-5	1.19E+04	1.29E+04	1.37E+04	1.01E+04	2.98E+05	8.79E-08	6.16E+03	6.16E+03
2122-86-3	1.83E+02	1.87E+02	1.89E+02	1.76E+02	1.90E+03	2.62E-11	4.53E+02	4.53E+02
21260-46-8	9.47E+01	9.45E+01	9.44E+01	9.51E+01	2.62E+03	6.94E-06	1.17E+03	1.17E+03
21436-96-4	8.15E-01	8.63E-01	8.94E-01	7.36E-01	5.61E+02	1.54E-02	1.08E+02	1.08E+02
21436-97-5	2.24E+00	2.66E+00	2.93E+00	1.55E+00	1.23E+03	1.42E-02	1.50E+02	1.50E+02
21626-89-1	3.51E+02	3.64E+02	3.73E+02	3.28E+02	5.54E+03	2.58E-06	9.66E+02	9.66E+02
21638-36-8	3.16E+02	3.25E+02	3.31E+02	3.02E+02	4.55E+03	2.39E-05	1.18E+03	1.18E+03
218-01-9	6.71E+03	8.66E+03	9.95E+03	3.47E+03	1.25E+06	2.45E+02	9.45E+03	9.45E+03
2185-92-4	1.82E+01	2.22E+01	2.48E+01	1.15E+01	3.61E+03	4.61E-02	2.19E+02	2.19E+02

21884-44-6	1.71E+03	1.77E+03	1.82E+03	1.60E+03	2.14E+04	5.18E-24	1.77E+03	1.77E+03
22071-15-4	1.36E+02	1.46E+02	1.52E+02	1.21E+02	2.52E+03	8.56E-05	1.55E+02	1.55E+02
22131-79-9	9.11E+01	9.75E+01	1.02E+02	8.03E+01	2.42E+03	3.07E-04	3.83E+02	3.83E+02
2244-16-8	6.34E-01	6.75E-01	7.03E-01	5.65E-01	6.14E+02	1.10E-02	1.92E+01	1.92E+01
22494-47-9	1.77E+03	1.87E+03	1.93E+03	1.62E+03	2.91E+04	1.78E-04	3.34E+03	3.34E+03
22760-18-5	3.80E+02	4.67E+02	5.26E+02	2.33E+02	2.37E+04	2.13E-03	6.95E+02	6.72E+02
22839-47-0	4.75E+01	4.92E+01	5.04E+01	4.45E+01	8.95E+02	1.11E-09	2.02E+02	2.04E+02
22966-79-6	5.82E+03	6.64E+03	7.19E+03	4.46E+03	2.16E+05	3.21E-04	1.10E+02	1.10E+02
23031-25-6	1.29E+01	1.59E+01	1.79E+01	7.88E+00	8.60E+02	2.36E-10	2.69E-01	2.73E-01
23282-20-4	6.13E+02	6.14E+02	6.15E+02	6.12E+02	2.14E+03	5.99E-13	8.97E+02	8.97E+02
2353-45-9	3.24E+02	3.24E+02	3.24E+02	3.24E+02	9.67E+02	3.34E-39	4.65E+02	4.65E+02
23746-34-1	2.76E+02	2.77E+02	2.78E+02	2.73E+02	1.13E+03	1.34E-16	3.81E+02	3.81E+02
2409-55-4	3.85E+00	4.02E+00	4.14E+00	3.57E+00	2.30E+03	3.28E-01	2.21E+01	2.21E+01
2425-85-6	2.63E+03	3.00E+03	3.25E+03	2.00E+03	9.36E+04	1.57E-02	8.94E+01	8.94E+01
2429-74-5	2.62E+03	2.62E+03	2.62E+03	2.61E+03	9.26E+03	8.27E-37	3.69E+03	3.68E+03
2432-99-7	7.85E+00	9.29E+00	1.03E+01	5.44E+00	4.03E+02	3.31E-09	8.49E-01	3.40E+01
2438-88-2	1.51E+03	1.53E+03	1.55E+03	1.48E+03	5.56E+04	1.91E+02	1.08E+03	1.08E+03
24382-04-5	1.01E+01	1.04E+01	1.06E+01	9.63E+00	2.85E+02	1.21E-10	9.62E+01	9.62E+01
244-63-3	7.67E+01	9.62E+01	1.09E+02	4.42E+01	5.28E+03	3.89E-03	5.96E+01	5.50E+01
24448-94-0	6.19E+01	6.21E+01	6.22E+01	6.16E+01	2.32E+02	2.26E-10	8.34E+01	8.34E+01
24554-26-5	3.94E+02	4.11E+02	4.21E+02	3.68E+02	5.59E+03	2.02E-05	8.15E+02	8.15E+02
24589-77-3	5.89E+01	6.00E+01	6.06E+01	5.72E+01	6.06E+02	8.89E-05	1.72E+02	1.72E+02
2475-45-8	1.22E+02	1.64E+02	1.92E+02	5.24E+01	1.55E+04	5.15E-02	6.83E+02	6.83E+02
25013-15-4	5.42E-02	6.22E-02	6.76E-02	4.08E-02	5.87E+02	3.56E-03	2.95E+00	2.95E+00
25265-71-8	3.10E+00	3.16E+00	3.20E+00	3.01E+00	4.23E+01	2.84E-05	1.40E+01	1.40E+01
25321-14-6	2.84E+02	2.87E+02	2.90E+02	2.78E+02	4.23E+03	1.80E+00	6.23E+02	6.23E+02
25451-15-4	1.08E+02	1.11E+02	1.13E+02	1.03E+02	1.56E+03	2.69E-04	3.86E+02	3.86E+02
2578-75-8	7.43E+02	7.51E+02	7.57E+02	7.29E+02	5.51E+03	2.16E-06	1.66E+03	1.66E+03
25812-30-0	2.48E+02	3.19E+02	3.66E+02	1.30E+02	2.29E+04	5.92E-02	9.01E+02	9.01E+02
25843-45-2	3.35E+01	3.38E+01	3.40E+01	3.30E+01	1.64E+02	2.65E-04	4.65E+01	4.65E+01
2602-46-2	4.13E+02	4.13E+02	4.13E+02	4.14E+02	1.23E+03	1.04E-30	5.92E+02	5.92E+02
26049-68-3	1.74E+02	1.86E+02	1.94E+02	1.55E+02	3.43E+03	1.94E-04	2.98E+02	3.01E+02
26049-69-4	3.20E+02	3.46E+02	3.64E+02	2.76E+02	7.67E+03	7.10E-04	5.58E+02	5.59E+02
26049-70-7	4.58E+02	4.88E+02	5.07E+02	4.09E+02	7.61E+03	5.23E-04	4.30E+02	4.28E+02
26049-71-8	3.70E+01	4.46E+01	4.97E+01	2.42E+01	2.17E+03	4.60E-06	6.40E+01	1.92E+02
2611-82-7	4.62E+02	4.64E+02	4.65E+02	4.58E+02	1.95E+03	3.75E-27	6.37E+02	6.37E+02
26148-68-5	9.79E+01	1.33E+02	1.57E+02	3.88E+01	8.53E+03	2.85E-04	1.07E+02	2.74E+02
262-12-4	1.42E+01	1.46E+01	1.49E+01	1.36E+01	3.85E+03	1.61E+00	1.93E+01	1.93E+01
26541-51-5	5.30E+01	5.39E+01	5.45E+01	5.14E+01	8.97E+02	1.29E-02	3.34E+02	3.34E+02

26675-46-7	6.29E-01	6.29E-01	6.29E-01	6.29E-01	1.33E+02	1.58E-01	3.48E+00	3.48E+00
2757-90-6	6.65E+01	6.77E+01	6.86E+01	6.44E+01	1.15E+03	2.35E-12	3.78E+02	3.78E+02
27753-52-2	2.57E+00	3.00E+00	3.29E+00	1.86E+00	1.40E+02	2.35E-04	6.22E-02	6.22E-02
2783-94-0	1.37E+02	1.37E+02	1.37E+02	1.36E+02	4.62E+02	1.60E-22	1.93E+02	1.93E+02
2784-94-3	1.21E+02	1.27E+02	1.30E+02	1.11E+02	2.26E+03	1.06E-08	3.33E+02	3.33E+02
2832-40-8	1.42E+03	1.55E+03	1.64E+03	1.20E+03	3.52E+04	8.38E-05	8.69E+02	8.69E+02
2835-39-4	3.58E-02	3.67E-02	3.73E-02	3.44E-02	1.55E+02	2.81E-03	2.96E+00	2.96E+00
2837-89-0	2.20E-01	2.20E-01	2.20E-01	2.20E-01	7.55E+01	5.84E-02	1.17E+00	1.17E+00
29069-24-7	8.93E+04	9.98E+04	1.07E+05	7.17E+04	2.83E+06	5.71E-08	2.22E+04	2.21E+04
29082-74-4	5.03E+01	6.35E+01	7.22E+01	2.84E+01	3.51E+04	3.51E+00	1.91E+01	1.91E+01
2955-38-6	2.13E+03	2.43E+03	2.63E+03	1.62E+03	7.77E+04	4.72E-03	1.79E+03	1.79E+03
298-18-0	1.86E+00	1.87E+00	1.88E+00	1.84E+00	3.34E+01	1.45E-02	7.89E+00	7.89E+00
298-81-7	1.12E+02	1.32E+02	1.44E+02	8.02E+01	6.11E+03	1.47E-03	6.37E+02	6.37E+02
29975-16-4	1.40E+03	1.57E+03	1.68E+03	1.13E+03	4.15E+04	8.73E-03	9.37E+02	9.28E+02
3012-37-1	6.88E+00	7.02E+00	7.12E+00	6.65E+00	5.81E+02	5.67E-02	5.21E+01	5.21E+01
3012-65-5	3.80E+01	3.82E+01	3.83E+01	3.77E+01	1.56E+02	1.05E-18	5.26E+01	5.26E+01
302-22-7	3.23E+03	3.95E+03	4.43E+03	2.02E+03	1.99E+05	2.69E-02	8.78E+03	8.78E+03
303-34-4	3.20E+02	3.64E+02	3.93E+02	2.47E+02	1.21E+04	8.87E-08	1.10E+02	1.90E+02
303-47-9	4.09E+03	4.25E+03	4.36E+03	3.82E+03	5.16E+04	7.57E-10	4.22E+03	4.22E+03
3031-51-4	1.76E+02	1.97E+02	2.11E+02	1.41E+02	8.21E+03	1.03E-06	1.03E+03	1.93E+03
305-03-3	4.33E+02	5.42E+02	6.14E+02	2.52E+02	3.06E+04	1.03E-02	2.40E+03	2.40E+03
30516-87-1	4.98E+02	5.00E+02	5.02E+02	4.93E+02	2.15E+03	1.05E-16	6.84E+02	6.84E+02
3054-95-3	1.05E-02	1.03E-02	1.02E-02	1.07E-02	3.72E+01	8.61E-04	2.44E+00	2.44E+00
306-37-6	2.36E+01	2.37E+01	2.38E+01	2.34E+01	1.07E+02	1.64E-11	3.60E+01	3.60E+01
3068-88-0	2.21E+00	2.21E+00	2.22E+00	2.20E+00	5.56E+01	4.17E-02	1.31E+01	1.31E+01
3096-50-2	4.68E+02	5.06E+02	5.31E+02	4.05E+02	9.89E+03	2.93E-04	4.95E+02	4.95E+02
315-22-0	5.15E+02	5.27E+02	5.35E+02	4.96E+02	4.41E+03	7.26E-13	4.57E+02	7.81E+02
3165-93-3	4.79E+00	4.95E+00	5.06E+00	4.52E+00	1.03E+03	9.68E-02	1.21E+02	1.21E+02
324-93-6	1.30E+02	1.64E+02	1.87E+02	7.29E+01	1.21E+04	8.85E-02	5.41E+02	5.10E+02
3276-41-3	5.27E+00	5.62E+00	5.85E+00	4.69E+00	2.14E+02	1.62E-03	5.88E+01	5.88E+01
32852-21-4	2.60E+01	2.69E+01	2.74E+01	2.46E+01	3.21E+02	6.17E-06	5.64E+01	5.67E+01
3296-90-0	3.21E+01	3.37E+01	3.47E+01	2.95E+01	4.85E+02	2.30E-06	5.40E+01	5.40E+01
331-39-5	8.91E+00	1.01E+01	1.09E+01	6.90E+00	3.62E+02	4.87E-11	2.62E+01	2.62E+01
33229-34-4	3.33E+02	3.37E+02	3.40E+02	3.26E+02	2.30E+03	3.72E-10	5.34E+02	5.34E+02
33857-26-0	6.65E+02	7.24E+02	7.64E+02	5.65E+02	6.24E+04	6.53E+01	7.50E+01	7.50E+01
34627-78-6	5.48E+01	6.77E+01	7.63E+01	3.34E+01	1.05E+04	1.15E-01	5.68E+02	5.68E+02
34661-75-1	3.11E+02	3.47E+02	3.70E+02	2.52E+02	1.00E+04	6.37E-08	1.17E+02	3.03E+02
3544-23-8	3.46E+02	4.16E+02	4.63E+02	2.29E+02	2.10E+04	1.20E-01	7.57E+02	7.56E+02
3570-75-0	3.47E+02	3.55E+02	3.60E+02	3.34E+02	3.95E+03	6.04E-07	9.59E+02	9.59E+02

36133-88-7	2.80E+02	2.83E+02	2.85E+02	2.74E+02	2.90E+03	5.59E-06	1.03E+03	1.03E+03
363-17-7	3.21E+03	3.94E+03	4.43E+03	1.99E+03	2.11E+05	1.53E+00	2.86E+03	2.86E+03
36322-90-4	5.54E+02	5.88E+02	6.11E+02	4.96E+02	1.01E+04	4.76E-10	4.44E+02	4.44E+02
36702-44-0	8.69E+00	9.09E+00	9.35E+00	8.02E+00	3.30E+02	1.46E-02	1.02E+02	1.02E+02
3688-53-7	2.44E+02	2.47E+02	2.49E+02	2.40E+02	4.66E+03	4.80E-05	1.89E+03	1.89E+03
3693-22-9	4.47E+01	6.24E+01	7.43E+01	1.51E+01	1.11E+04	3.51E-02	2.95E+02	2.78E+02
37087-94-8	1.13E+03	1.20E+03	1.25E+03	1.01E+03	2.60E+04	1.92E-07	3.97E+03	3.97E+03
3761-53-3	3.43E+02	3.43E+02	3.44E+02	3.43E+02	1.11E+03	3.03E-22	4.88E+02	4.88E+02
3771-19-5	2.87E+03	3.28E+03	3.56E+03	2.18E+03	1.21E+05	1.07E-03	8.12E+03	8.12E+03
3775-55-1	1.98E+02	1.98E+02	1.99E+02	1.96E+02	1.30E+03	4.76E-06	5.37E+02	5.37E+02
38434-77-4	2.78E+01	2.81E+01	2.82E+01	2.74E+01	1.91E+02	9.86E-04	6.18E+01	6.18E+01
38514-71-5	1.43E+02	1.55E+02	1.63E+02	1.23E+02	3.39E+03	4.13E-04	1.70E+02	2.44E+02
389-08-2	4.64E+02	4.83E+02	4.96E+02	4.32E+02	7.70E+03	1.02E-04	1.44E+03	1.44E+03
39156-41-7	1.03E+02	1.04E+02	1.04E+02	1.03E+02	3.40E+02	1.10E-15	1.42E+02	1.42E+02
396-01-0	1.13E+02	1.20E+02	1.25E+02	1.01E+02	2.09E+03	3.50E-10	3.24E+01	1.58E+02
398-32-3	6.30E+02	6.87E+02	7.25E+02	5.35E+02	1.52E+04	4.38E-03	7.49E+02	7.49E+02
39801-14-4	1.15E+02	1.23E+02	1.28E+02	1.02E+02	6.90E+04	1.50E+01	5.69E+01	5.69E+01
40548-68-3	5.33E+00	5.48E+00	5.58E+00	5.08E+00	9.11E+01	3.09E-03	2.37E+01	2.37E+01
40580-89-0	5.92E+01	7.00E+01	7.72E+01	4.12E+01	1.12E+04	5.94E-01	7.38E+01	7.38E+01
4075-79-0	2.33E+02	2.58E+02	2.75E+02	1.90E+02	6.74E+03	4.73E-04	2.70E+02	2.70E+02
4106-66-5	4.37E+01	6.12E+01	7.29E+01	1.46E+01	1.11E+04	3.39E-02	2.92E+02	2.79E+02
41340-25-4	4.35E+02	5.42E+02	6.14E+02	2.56E+02	2.97E+04	2.49E-05	2.38E+03	2.38E+03
4164-28-7	1.22E+01	1.23E+01	1.24E+01	1.20E+01	7.09E+01	2.85E-03	2.23E+01	2.23E+01
4180-23-8	7.75E-01	8.10E-01	8.33E-01	7.16E-01	1.12E+03	4.26E-02	1.04E+01	1.04E+01
42011-48-3	1.11E+03	1.23E+03	1.31E+03	9.10E+02	3.08E+04	1.53E-03	7.19E+02	1.42E+03
42579-28-2	2.94E+01	2.99E+01	3.02E+01	2.86E+01	1.98E+02	4.95E-06	4.23E+01	4.23E+01
434-07-1	3.70E+03	4.04E+03	4.27E+03	3.14E+03	9.26E+04	1.43E-05	2.44E+03	2.44E+03
434-13-9	7.91E+03	8.96E+03	9.66E+03	6.17E+03	2.67E+05	6.11E-05	1.51E+03	1.51E+03
439-14-5	5.23E+02	5.62E+02	5.88E+02	4.58E+02	1.09E+04	1.07E-03	7.52E+02	7.52E+02
443-72-1	1.32E+01	1.49E+01	1.60E+01	1.04E+01	6.45E+02	3.63E-05	9.21E+01	1.33E+02
446-86-6	6.45E+02	6.49E+02	6.51E+02	6.38E+02	8.40E+03	6.33E-09	3.41E+03	3.41E+03
4548-53-2	4.29E+02	4.29E+02	4.29E+02	4.28E+02	1.39E+03	1.92E-20	6.09E+02	6.09E+02
471-29-4	9.67E-01	1.29E+00	1.51E+00	4.21E-01	8.19E+01	2.90E-07	1.21E-01	1.21E-01
471-53-4	2.46E+04	2.71E+04	2.87E+04	2.05E+04	6.75E+05	1.51E-06	1.03E+04	1.03E+04
474-25-9	3.37E+03	3.57E+03	3.70E+03	3.04E+03	5.68E+04	2.19E-08	2.83E+03	2.83E+03
476-66-4	2.63E+02	2.69E+02	2.72E+02	2.55E+02	2.06E+03	3.25E-22	3.24E+02	3.24E+02
480-54-6	1.95E+02	2.13E+02	2.25E+02	1.65E+02	4.91E+03	3.65E-14	7.54E+01	1.25E+02
4812-22-0	1.81E-01	1.80E-01	1.80E-01	1.81E-01	1.61E+02	1.37E-02	6.46E+00	6.46E+00
493-78-7	1.06E+02	1.52E+02	1.83E+02	2.80E+01	1.04E+04	1.93E-03	3.23E+01	3.31E+01

50-02-2	1.09E+03	1.11E+03	1.12E+03	1.05E+03	8.32E+03	1.48E-09	1.35E+03	1.35E+03
50-14-6	3.86E+00	4.40E+00	4.77E+00	2.95E+00	1.44E+02	2.15E-10	1.29E-02	1.29E-02
50-18-0	7.40E+01	8.15E+01	8.65E+01	6.15E+01	2.71E+03	7.28E-05	4.06E+02	4.06E+02
50-23-7	1.66E+03	1.67E+03	1.67E+03	1.65E+03	6.97E+03	1.27E-09	2.30E+03	2.30E+03
50-24-8	1.98E+03	1.99E+03	1.99E+03	1.96E+03	8.33E+03	1.22E-09	2.73E+03	2.73E+03
50-33-9	1.62E+02	1.81E+02	1.94E+02	1.29E+02	5.82E+03	3.48E-08	4.47E+02	4.47E+02
50-55-5	3.37E+03	3.73E+03	3.96E+03	2.79E+03	9.58E+04	1.49E-10	1.24E+03	1.34E+03
50-78-2	4.19E+01	4.33E+01	4.42E+01	3.96E+01	4.06E+02	8.46E-07	4.59E+01	4.59E+01
50-81-7	3.90E+00	4.31E+00	4.58E+00	3.21E+00	1.40E+02	6.90E-15	1.69E+01	1.69E+01
501-30-4	1.21E+01	1.26E+01	1.30E+01	1.13E+01	2.85E+02	2.86E-09	7.72E+01	7.72E+01
50264-69-2	2.57E+03	2.83E+03	3.00E+03	2.14E+03	7.27E+04	7.15E-04	4.20E+03	4.20E+03
50594-66-6	6.52E+03	6.72E+03	6.86E+03	6.17E+03	6.12E+04	4.25E-06	7.01E+03	7.01E+03
509-14-8	2.26E+00	2.26E+00	2.26E+00	2.25E+00	1.61E+02	3.20E-01	3.66E+01	3.66E+01
513-37-1	5.65E-03	6.14E-03	6.47E-03	4.83E-03	1.75E+02	4.56E-04	1.23E+00	1.23E+00
5131-60-2	3.60E+01	4.32E+01	4.81E+01	2.40E+01	2.37E+03	2.32E-03	2.83E+02	3.09E+02
51325-35-0	1.50E+03	1.51E+03	1.51E+03	1.49E+03	5.92E+03	1.95E-09	1.93E+03	1.93E+03
51333-22-3	3.34E+03	3.38E+03	3.41E+03	3.27E+03	1.99E+04	2.44E-09	4.38E+03	4.38E+03
51410-44-7	1.52E+01	1.72E+01	1.86E+01	1.19E+01	6.98E+02	1.05E-03	7.85E+01	7.85E+01
51481-10-8	1.74E+02	1.75E+02	1.76E+02	1.72E+02	1.95E+03	2.64E-10	7.37E+02	7.37E+02
51481-61-9	7.33E+01	9.39E+01	1.08E+02	3.89E+01	5.68E+03	2.35E-08	7.92E+01	3.61E+02
51542-33-7	5.91E+01	6.71E+01	7.24E+01	4.58E+01	2.85E+03	1.01E-06	4.30E+02	4.30E+02
517-28-2	5.92E+02	5.98E+02	6.02E+02	5.82E+02	3.22E+03	2.15E-10	8.16E+02	8.16E+02
51786-53-9	9.42E-01	1.04E+00	1.10E+00	7.85E-01	6.74E+02	1.24E-02	1.19E+02	1.19E+02
518-82-1	5.53E+02	6.24E+02	6.72E+02	4.34E+02	1.91E+04	1.52E-06	2.04E+02	2.04E+02
52-76-6	3.65E+03	5.05E+03	5.98E+03	1.32E+03	3.46E+05	3.60E-02	1.28E+03	1.28E+03
520-18-3	1.03E+02	1.09E+02	1.14E+02	9.31E+01	1.80E+03	3.87E-12	8.61E+01	8.61E+01
520-45-6	9.41E+00	1.04E+01	1.11E+01	7.73E+00	7.04E+02	4.77E-03	1.64E+02	1.64E+02
5208-87-7	2.67E+01	3.03E+01	3.27E+01	2.07E+01	1.18E+03	1.37E-04	1.39E+02	1.39E+02
52214-84-3	1.74E+03	1.83E+03	1.89E+03	1.58E+03	3.58E+04	1.51E-02	6.09E+03	6.09E+03
53-03-2	1.75E+03	1.75E+03	1.76E+03	1.73E+03	6.88E+03	4.30E-09	2.42E+03	2.42E+03
53-19-0	7.06E+03	7.79E+03	8.27E+03	5.86E+03	4.29E+05	3.12E+02	9.29E+02	9.29E+02
53-43-0	3.48E+02	4.48E+02	5.14E+02	1.82E+02	2.63E+04	8.26E-05	9.26E+02	9.26E+02
53-86-1	8.05E+02	1.04E+03	1.20E+03	4.05E+02	6.18E+04	2.95E-05	3.08E+03	3.08E+03
53-95-2	2.60E+02	2.87E+02	3.06E+02	2.13E+02	8.31E+03	8.56E-06	5.79E+02	5.79E+02
5307-14-2	5.45E+01	5.73E+01	5.92E+01	4.99E+01	1.03E+03	4.81E-05	1.62E+02	1.63E+02
531-18-0	3.16E+02	3.17E+02	3.17E+02	3.15E+02	2.26E+03	3.20E-15	9.32E+02	9.39E+02
531-82-8	5.63E+02	5.84E+02	5.98E+02	5.29E+02	7.25E+03	1.83E-05	1.05E+03	1.05E+03
531-85-1	4.15E+01	5.76E+01	6.83E+01	1.46E+01	3.84E+03	3.46E-05	2.65E+01	2.65E+01
533-31-3	6.53E+00	8.56E+00	9.92E+00	3.14E+00	6.66E+02	4.44E-04	4.57E+01	4.57E+01

536-33-4	3.52E+01	3.89E+01	4.13E+01	2.91E+01	1.09E+03	9.54E-05	6.62E+01	9.38E+01
53609-64-6	3.40E+00	3.47E+00	3.51E+00	3.29E+00	5.22E+01	3.02E-08	1.72E+01	1.72E+01
538-23-8	6.76E-01	1.08E+00	1.35E+00	1.61E-03	4.52E+02	2.82E-06	9.15E-03	9.15E-03
538-41-0	2.44E+02	2.89E+02	3.19E+02	1.69E+02	1.32E+04	9.70E-05	1.02E+03	1.00E+03
54-12-6	1.32E+01	1.39E+01	1.44E+01	1.21E+01	4.97E+02	6.78E-10	1.49E+02	1.49E+02
54-31-9	4.09E+02	4.32E+02	4.47E+02	3.70E+02	1.23E+04	1.10E-09	3.02E+03	3.02E+03
54-80-8	1.81E+01	2.38E+01	2.76E+01	8.59E+00	1.60E+03	4.15E-08	6.58E-01	6.60E-01
540-23-8	1.34E+00	1.62E+00	1.80E+00	8.89E-01	4.83E+02	3.11E-03	3.30E+01	3.30E+01
540-51-2	9.07E+01	9.16E+01	9.22E+01	8.92E+01	5.44E+02	3.38E-02	1.59E+02	1.59E+02
541-69-5	2.51E+01	2.53E+01	2.54E+01	2.48E+01	7.28E+02	1.30E-04	3.00E+02	3.00E+02
542-56-3	8.73E-03	8.88E-03	8.98E-03	8.47E-03	6.30E+01	7.80E-04	6.52E-01	6.52E-01
542-88-1	1.28E-01	1.28E-01	1.28E-01	1.28E-01	9.48E+01	2.43E-02	7.69E+00	7.69E+00
5456-28-0	5.09E+03	5.24E+03	5.35E+03	4.82E+03	5.36E+04	2.49E-09	5.71E+03	5.71E+03
55-80-1	7.31E+01	9.26E+01	1.06E+02	4.08E+01	4.92E+04	1.48E+00	4.72E+02	4.61E+02
55-98-1	3.02E+02	3.02E+02	3.02E+02	3.01E+02	1.30E+03	4.17E-07	5.41E+02	5.41E+02
55090-44-3	1.57E+02	1.81E+02	1.97E+02	1.16E+02	3.03E+04	4.46E+00	7.20E+01	7.20E+01
551-92-8	3.42E+01	3.50E+01	3.55E+01	3.29E+01	2.88E+02	2.21E-04	5.10E+01	5.10E+01
5522-43-0	6.00E+03	6.66E+03	7.11E+03	4.89E+03	3.42E+05	1.15E+02	2.30E+03	2.30E+03
553-53-7	6.82E+01	6.85E+01	6.87E+01	6.78E+01	3.40E+02	2.21E-05	1.29E+02	1.32E+02
55380-34-2	2.57E+01	2.59E+01	2.60E+01	2.54E+01	4.30E+02	3.59E-05	1.88E+02	1.88E+02
555-84-0	9.08E+02	9.24E+02	9.34E+02	8.82E+02	1.05E+04	2.47E-04	3.22E+03	3.22E+03
55556-92-8	2.66E+00	2.88E+00	3.02E+00	2.30E+00	4.16E+02	1.39E-02	1.06E+02	1.06E+02
55557-00-1	2.17E+01	2.18E+01	2.19E+01	2.15E+01	2.84E+02	1.05E-05	1.26E+02	1.26E+02
55566-30-8	4.83E-01	5.00E-01	5.11E-01	4.55E-01	7.25E+01	1.25E-03	1.52E+01	1.52E+01
55738-54-0	1.15E+02	1.60E+02	1.90E+02	4.10E+01	1.08E+04	1.10E-06	9.31E+01	1.40E+02
56-04-2	1.06E+01	1.39E+01	1.61E+01	5.19E+00	8.33E+02	3.80E-06	2.12E+00	4.21E+00
56-40-6	3.49E+00	3.58E+00	3.64E+00	3.34E+00	5.11E+01	1.27E-10	1.37E+01	1.37E+01
56-86-0	3.37E+00	3.48E+00	3.55E+00	3.19E+00	6.22E+01	1.16E-09	1.68E+01	1.68E+01
56038-13-2	5.95E+02	5.96E+02	5.97E+02	5.92E+02	2.14E+03	1.21E-12	8.33E+02	8.33E+02
5632-47-3	1.58E+00	2.14E+00	2.52E+00	6.34E-01	1.39E+02	2.81E-06	5.87E-01	6.16E-01
5634-39-9	9.18E+01	9.50E+01	9.72E+01	8.64E+01	1.53E+03	1.15E-03	3.51E+02	3.51E+02
56654-52-5	2.58E+01	3.07E+01	3.39E+01	1.77E+01	1.41E+03	3.56E-03	1.33E+01	1.33E+01
56795-65-4	9.57E-02	9.41E-02	9.31E-02	9.83E-02	5.14E+01	1.69E-03	8.07E+00	8.07E+00
569-57-3	1.15E+03	1.63E+03	1.95E+03	3.53E+02	1.96E+05	3.68E-01	7.36E+01	7.36E+01
569-61-9	1.27E+03	1.27E+03	1.28E+03	1.27E+03	4.19E+03	2.22E-11	1.85E+03	1.85E+03
56980-93-9	1.14E+02	1.30E+02	1.41E+02	8.80E+01	4.15E+03	8.56E-10	6.43E+00	6.51E+00
57-30-7	1.11E+02	1.11E+02	1.11E+02	1.10E+02	3.74E+02	2.67E-18	1.57E+02	1.57E+02
57-39-6	2.48E+02	2.51E+02	2.53E+02	2.43E+02	2.68E+03	5.39E-02	9.67E+02	9.67E+02
57-41-0	3.67E+02	3.85E+02	3.97E+02	3.37E+02	5.21E+03	1.51E-07	3.20E+02	3.20E+02

57-50-1	4.07E+01	4.10E+01	4.12E+01	4.01E+01	2.02E+02	2.20E-18	5.48E+01	5.48E+01
57-57-8	1.76E+01	1.78E+01	1.79E+01	1.73E+01	1.24E+02	2.06E-02	3.63E+01	3.63E+01
57-66-9	6.64E+02	7.01E+02	7.26E+02	6.01E+02	1.37E+04	3.28E-06	2.11E+03	2.11E+03
57-68-1	4.18E+01	5.46E+01	6.31E+01	2.06E+01	3.30E+03	9.55E-08	8.21E+00	1.57E+01
57-97-6	6.59E+02	9.50E+02	1.14E+03	1.74E+02	2.10E+05	5.18E+00	1.75E+02	1.75E+02
57590-20-2	5.56E+00	6.78E+00	7.58E+00	3.54E+00	3.06E+02	3.05E-06	6.42E-02	6.42E-02
57817-89-7	1.51E+02	1.52E+02	1.52E+02	1.51E+02	5.12E+02	7.82E-30	2.14E+02	2.14E+02
58-15-1	5.45E+01	6.28E+01	6.83E+01	4.06E+01	2.65E+03	2.44E-06	2.62E+02	3.20E+02
58-55-9	6.72E+01	7.07E+01	7.31E+01	6.12E+01	1.18E+03	1.67E-09	1.30E+02	1.30E+02
58-93-5	2.50E+02	2.54E+02	2.57E+02	2.43E+02	4.98E+03	1.06E-07	1.91E+03	1.91E+03
5834-17-3	9.24E+01	1.29E+02	1.54E+02	3.05E+01	1.35E+04	2.10E-02	3.31E+02	3.24E+02
59-05-2	6.60E+02	6.64E+02	6.67E+02	6.52E+02	3.09E+03	1.06E-18	9.32E+02	8.91E+02
59-33-6	2.02E+03	2.03E+03	2.04E+03	2.00E+03	8.63E+03	2.73E-10	2.79E+03	2.79E+03
59-51-8	8.74E+00	9.11E+00	9.35E+00	8.13E+00	1.92E+02	2.40E-08	4.90E+01	4.84E+01
59-88-1	1.88E+01	1.91E+01	1.94E+01	1.81E+01	2.99E+02	3.67E-10	1.01E+02	1.01E+02
59-89-2	5.33E+00	5.42E+00	5.48E+00	5.19E+00	1.12E+02	5.20E-04	4.48E+01	4.48E+01
590-21-6	6.24E-03	6.51E-03	6.68E-03	5.80E-03	1.58E+02	5.69E-04	1.64E+00	1.64E+00
59122-46-2	1.05E+03	1.19E+03	1.28E+03	8.08E+02	3.67E+04	8.55E-06	1.41E+02	1.41E+02
592-31-4	2.04E+00	2.16E+00	2.24E+00	1.83E+00	3.75E+01	1.15E-04	3.15E+00	3.15E+00
593-60-2	5.52E-03	5.65E-03	5.73E-03	5.30E-03	1.33E+02	4.83E-04	1.81E+00	1.81E+00
593-70-4	2.26E-01	2.26E-01	2.26E-01	2.26E-01	3.22E+01	2.30E-02	1.50E+00	1.50E+00
597-25-1	1.56E+01	1.62E+01	1.65E+01	1.48E+01	4.21E+02	1.01E-04	1.40E+02	1.40E+02
5979-28-2	1.66E+05	1.85E+05	1.97E+05	1.36E+05	5.00E+06	7.07E-10	4.71E+04	4.71E+04
598-55-0	9.90E+00	1.00E+01	1.01E+01	9.72E+00	6.06E+01	1.55E-03	1.61E+01	1.61E+01
598-57-2	1.50E+01	1.52E+01	1.54E+01	1.46E+01	1.35E+02	2.04E-02	2.61E+01	2.64E+01
599-79-1	1.86E+03	2.01E+03	2.10E+03	1.62E+03	4.06E+04	3.05E-08	1.21E+03	1.22E+03
60-11-7	4.74E+02	6.65E+02	7.93E+02	1.55E+02	5.46E+04	6.89E-02	7.03E+02	6.88E+02
60-32-2	1.46E+00	1.53E+00	1.58E+00	1.34E+00	3.14E+01	1.36E-11	4.39E+00	8.29E+00
60-56-0	4.57E+00	5.28E+00	5.75E+00	3.39E+00	2.36E+02	2.91E-06	3.07E+01	3.07E+01
600-24-8	7.80E-01	7.80E-01	7.81E-01	7.78E-01	4.98E+01	5.95E-02	2.99E+00	2.99E+00
60102-37-6	6.25E+02	6.44E+02	6.56E+02	5.95E+02	6.34E+03	1.66E-13	4.01E+02	1.02E+03
60142-96-3	4.57E+00	4.87E+00	5.07E+00	4.06E+00	1.16E+02	9.23E-11	4.96E+00	3.25E+01
602-87-9	1.29E+02	1.33E+02	1.36E+02	1.22E+02	1.12E+04	3.97E+00	1.31E+02	1.31E+02
604-75-1	7.11E+02	7.25E+02	7.35E+02	6.87E+02	5.49E+03	1.84E-08	8.44E+02	8.44E+02
60599-38-4	5.48E+01	5.50E+01	5.52E+01	5.44E+01	3.72E+02	6.31E-04	1.52E+02	1.52E+02
607-35-2	1.37E+02	1.42E+02	1.46E+02	1.28E+02	1.52E+03	1.34E-02	1.43E+02	1.42E+02
609-20-1	1.24E+02	1.36E+02	1.44E+02	1.05E+02	4.66E+03	4.64E-03	8.24E+02	8.25E+02
61-76-7	3.59E+01	3.68E+01	3.74E+01	3.45E+01	8.00E+02	1.52E-12	2.70E+02	2.70E+02
61-94-9	1.66E+01	1.69E+01	1.72E+01	1.60E+01	4.40E+02	9.63E-04	1.72E+02	1.72E+02

6109-97-3	6.05E+01	8.50E+01	1.01E+02	1.97E+01	1.73E+04	4.94E-02	4.27E+02	4.27E+02
611-23-4	2.65E-01	2.66E-01	2.67E-01	2.63E-01	2.87E+02	2.13E-02	8.57E+00	8.57E+00
611-32-5	4.03E+00	4.33E+00	4.53E+00	3.53E+00	8.09E+02	4.70E-02	1.25E+01	1.25E+01
612-82-8	9.94E+01	1.16E+02	1.27E+02	7.17E+01	5.53E+03	1.06E-03	7.52E+02	7.52E+02
613-94-5	1.63E+01	1.66E+01	1.69E+01	1.57E+01	1.43E+02	1.14E-04	2.95E+01	3.10E+01
614-00-6	5.19E+01	5.35E+01	5.46E+01	4.91E+01	1.06E+03	8.91E-02	2.66E+02	2.66E+02
614-95-9	3.54E+00	3.69E+00	3.79E+00	3.29E+00	2.03E+02	9.52E-03	2.55E+01	2.55E+01
615-28-1	5.77E+01	5.86E+01	5.92E+01	5.62E+01	1.70E+03	1.70E-03	6.83E+02	6.83E+02
615-53-2	2.47E+00	2.49E+00	2.51E+00	2.43E+00	1.09E+02	2.04E-02	1.90E+01	1.90E+01
616-91-1	9.08E+00	9.88E+00	1.04E+01	7.76E+00	2.91E+02	6.67E-10	4.40E+01	4.40E+01
617-84-5	1.81E+00	1.87E+00	1.91E+00	1.71E+00	3.93E+01	5.04E-04	1.13E+01	1.13E+01
619-17-0	1.82E+02	1.85E+02	1.88E+02	1.75E+02	1.43E+03	3.09E-07	2.74E+02	2.74E+02
62-23-7	2.97E+02	3.02E+02	3.05E+02	2.89E+02	1.87E+03	1.17E-05	3.59E+02	3.59E+02
62-44-2	1.03E+02	1.08E+02	1.12E+02	9.44E+01	1.82E+03	7.12E-05	2.61E+02	2.61E+02
62-54-4	6.77E+00	6.79E+00	6.81E+00	6.72E+00	2.60E+01	1.21E-05	8.73E+00	8.73E+00
621-64-7	2.56E+00	2.60E+00	2.63E+00	2.49E+00	3.00E+02	4.73E-02	6.87E+01	6.87E+01
622-51-5	4.04E+00	4.82E+00	5.34E+00	2.74E+00	2.22E+02	4.32E-05	1.05E+01	1.05E+01
622-97-9	6.43E-02	7.17E-02	7.66E-02	5.21E-02	5.71E+02	4.56E-03	2.88E+00	2.88E+00
624-18-0	7.04E+01	7.11E+01	7.15E+01	6.93E+01	1.97E+03	7.34E-04	8.11E+02	8.11E+02
624-84-0	1.44E+01	1.45E+01	1.45E+01	1.42E+01	8.22E+01	3.11E-04	2.62E+01	2.66E+01
625-89-8	1.95E+00	1.95E+00	1.95E+00	1.95E+00	6.54E+02	8.32E-01	8.35E+01	8.35E+01
627-05-4	7.34E-01	7.35E-01	7.36E-01	7.32E-01	4.87E+01	2.86E-02	1.82E+00	1.82E+00
628-02-4	7.33E+00	7.44E+00	7.51E+00	7.14E+00	6.24E+01	3.31E-04	1.68E+01	1.68E+01
628-36-4	2.42E+01	2.45E+01	2.46E+01	2.38E+01	1.81E+02	3.87E-09	5.84E+01	5.84E+01
628-94-4	9.29E+00	9.39E+00	9.45E+00	9.13E+00	7.82E+01	1.29E-06	2.63E+01	2.63E+01
6294-89-9	1.00E+00	1.01E+00	1.01E+00	9.94E-01	3.23E+01	1.31E-02	6.94E+00	6.98E+00
63412-06-6	1.42E+01	1.63E+01	1.77E+01	1.07E+01	6.66E+02	2.06E-06	6.58E+01	6.95E+01
6358-85-6	2.67E+05	2.84E+05	2.96E+05	2.39E+05	4.96E+06	2.32E-09	2.10E+05	2.10E+05
636-21-5	1.41E+00	1.61E+00	1.74E+00	1.09E+00	4.56E+02	3.74E-03	7.55E+01	7.55E+01
636-23-7	2.25E+01	2.28E+01	2.30E+01	2.21E+01	6.64E+02	2.74E-04	2.68E+02	2.68E+02
6369-59-1	1.32E+02	1.32E+02	1.32E+02	1.31E+02	4.40E+02	3.68E-15	1.84E+02	1.84E+02
637-07-0	2.97E+01	3.00E+01	3.02E+01	2.91E+01	4.04E+03	2.03E+00	6.61E+01	6.61E+01
6373-74-6	1.54E+03	1.55E+03	1.56E+03	1.51E+03	7.65E+03	1.41E-16	2.08E+03	2.08E+03
638-03-9	1.52E+00	1.84E+00	2.06E+00	9.78E-01	6.41E+02	2.83E-03	7.49E+01	7.49E+01
6381-77-7	1.10E+01	1.10E+01	1.11E+01	1.08E+01	5.40E+01	9.20E-21	1.47E+01	1.47E+01
63885-23-4	1.16E+01	1.34E+01	1.47E+01	8.48E+00	4.46E+02	1.53E-08	1.52E+00	1.52E+00
63886-77-1	2.82E+01	3.10E+01	3.29E+01	2.35E+01	2.54E+03	7.66E-02	9.73E+02	9.73E+02
64-77-7	2.04E+02	2.24E+02	2.38E+02	1.70E+02	5.37E+03	4.22E-05	2.41E+02	2.41E+02
64049-29-2	7.09E+02	9.31E+02	1.08E+03	3.39E+02	6.46E+04	1.57E-01	1.42E+03	1.40E+03

64091-91-4	2.93E+02	3.00E+02	3.04E+02	2.82E+02	3.42E+03	6.45E-04	8.57E+02	8.97E+02
6459-94-5	1.03E+03	1.03E+03	1.03E+03	1.03E+03	3.12E+03	4.28E-33	1.48E+03	1.48E+03
6471-49-4	2.45E+03	2.79E+03	3.02E+03	1.88E+03	9.05E+04	1.08E-08	6.08E+01	6.08E+01
6485-34-3	9.01E+01	9.05E+01	9.08E+01	8.95E+01	6.52E+02	2.20E-06	2.72E+02	2.72E+02
66-22-8	4.93E+01	4.98E+01	5.02E+01	4.83E+01	3.58E+02	1.64E-04	1.06E+02	1.06E+02
66-27-3	3.34E+01	3.36E+01	3.37E+01	3.31E+01	1.42E+02	6.70E-03	4.59E+01	4.59E+01
6673-35-4	4.81E+01	5.67E+01	6.25E+01	3.38E+01	2.33E+03	9.55E-10	3.82E+00	3.85E+00
67-20-9	9.48E+02	9.58E+02	9.65E+02	9.31E+02	8.69E+03	7.13E-07	3.02E+03	3.02E+03
67-21-0	1.82E+01	1.87E+01	1.90E+01	1.74E+01	3.59E+02	1.05E-07	1.16E+02	1.15E+02
67-52-7	5.82E+01	5.95E+01	6.03E+01	5.61E+01	4.59E+02	2.28E-04	7.76E+01	7.76E+01
6731-36-8	1.36E-02	2.17E-02	2.71E-02	1.48E-04	2.34E+04	4.46E-05	5.86E-01	5.86E-01
67730-10-3	7.50E+01	9.34E+01	1.06E+02	4.43E+01	5.10E+03	3.18E-05	2.99E+02	4.80E+02
67730-11-4	1.46E+02	1.80E+02	2.03E+02	8.88E+01	9.52E+03	6.52E-05	5.23E+02	7.00E+02
68-23-5	1.58E+03	2.01E+03	2.29E+03	8.72E+02	1.19E+05	1.90E-04	3.52E+03	3.52E+03
68-89-3	1.53E+02	1.53E+02	1.53E+02	1.52E+02	5.15E+02	1.48E-17	2.16E+02	2.16E+02
68844-77-9	7.14E+04	7.97E+04	8.52E+04	5.77E+04	2.23E+06	1.68E-03	1.78E+04	1.78E+04
69-65-8	8.30E-02	9.17E-02	9.76E-02	6.85E-02	2.92E+00	2.44E-14	3.46E-01	3.46E-01
695-53-4	2.64E+01	2.67E+01	2.68E+01	2.60E+01	1.51E+02	3.56E-06	4.24E+01	4.24E+01
6959-48-4	2.51E+01	2.54E+01	2.56E+01	2.46E+01	2.60E+02	9.87E-02	6.14E+01	6.14E+01
70-25-7	1.73E+01	1.92E+01	2.05E+01	1.42E+01	4.27E+02	1.51E-08	4.04E+00	4.05E+00
7003-89-6	1.18E+01	1.19E+01	1.19E+01	1.16E+01	7.11E+01	4.18E-08	2.40E+01	2.40E+01
71125-38-7	1.39E+03	1.50E+03	1.57E+03	1.21E+03	3.06E+04	2.37E-10	8.95E+02	8.95E+02
712-68-5	3.01E+02	3.04E+02	3.06E+02	2.97E+02	2.09E+03	2.34E-07	6.54E+02	6.54E+02
7177-48-2	3.15E+02	3.20E+02	3.23E+02	3.08E+02	2.02E+03	7.37E-13	4.02E+02	4.12E+02
72-33-3	4.40E+03	5.84E+03	6.79E+03	2.01E+03	3.79E+05	1.01E-02	1.59E+03	1.59E+03
720-69-4	3.06E+02	3.13E+02	3.18E+02	2.94E+02	3.02E+03	7.18E-06	6.14E+02	6.14E+02
72254-58-1	2.35E+02	2.73E+02	2.98E+02	1.72E+02	1.27E+04	4.30E-04	1.80E+03	1.80E+03
7227-91-0	7.42E+00	7.46E+00	7.49E+00	7.35E+00	5.23E+02	3.07E-01	8.65E+00	8.53E+00
7235-40-7	1.24E-05	1.41E-05	1.53E-05	9.48E-06	4.60E-04	8.01E-21	1.96E-07	1.96E-07
73-22-3	1.32E+01	1.39E+01	1.44E+01	1.21E+01	4.97E+02	6.78E-10	1.49E+02	1.49E+02
7336-20-1	2.38E+02	2.38E+02	2.38E+02	2.37E+02	8.01E+02	9.96E-21	3.36E+02	3.36E+02
7347-49-1	6.37E+02	7.44E+02	8.15E+02	4.59E+02	3.23E+04	5.74E-03	2.75E+03	2.75E+03
73590-58-6	1.08E+03	1.12E+03	1.14E+03	1.02E+03	1.26E+04	2.81E-07	8.76E+02	1.77E+03
74-31-7	5.59E+02	7.65E+02	9.02E+02	2.15E+02	5.17E+04	1.99E-04	6.44E+02	6.44E+02
74-96-4	2.11E-01	2.11E-01	2.12E-01	2.11E-01	1.49E+02	1.92E-02	3.92E+00	3.92E+00
7411-49-6	1.75E+02	1.75E+02	1.75E+02	1.75E+02	4.60E+03	7.40E-08	2.01E+03	2.01E+03
7422-80-2	2.68E+00	3.38E+00	3.85E+00	1.50E+00	2.05E+02	2.15E-02	1.78E-01	2.37E-01
75-00-3	4.44E-02	4.45E-02	4.46E-02	4.43E-02	4.61E+01	4.56E-03	1.01E+00	1.01E+00
75-01-4	3.67E-03	3.76E-03	3.82E-03	3.53E-03	7.65E+01	3.63E-04	9.08E-01	9.08E-01

75-02-5	5.95E-04	6.17E-04	6.32E-04	5.57E-04	3.14E+01	6.15E-05	2.08E-01	2.08E-01
75-34-3	1.04E-01	1.04E-01	1.04E-01	1.04E-01	4.60E+01	2.06E-02	1.98E+00	1.98E+00
75-38-7	7.52E-04	7.97E-04	8.27E-04	6.77E-04	8.13E+01	6.97E-05	1.92E-01	1.92E-01
75-45-6	3.44E-01	3.44E-01	3.44E-01	3.44E-01	4.48E+01	3.35E-02	9.32E-01	9.32E-01
75-69-4	2.58E-02	2.58E-02	2.59E-02	2.56E-02	1.28E+02	6.75E-03	1.00E+00	1.00E+00
75-71-8	5.17E-03	5.22E-03	5.26E-03	5.08E-03	1.32E+02	9.75E-04	4.74E-01	4.74E-01
75-88-7	2.60E-01	2.60E-01	2.60E-01	2.60E-01	5.01E+01	5.03E-02	1.11E+00	1.11E+00
75104-43-7	1.49E+02	1.65E+02	1.75E+02	1.23E+02	6.69E+03	2.72E-04	1.37E+03	1.37E+03
75330-75-5	3.26E+03	3.68E+03	3.95E+03	2.57E+03	1.10E+05	2.33E-06	6.17E+02	6.17E+02
756-79-6	9.60E+00	9.70E+00	9.77E+00	9.43E+00	6.75E+01	2.54E-03	2.09E+01	2.09E+01
7572-29-4	4.41E-02	4.41E-02	4.42E-02	4.40E-02	5.52E+01	8.69E-03	3.28E+00	3.28E+00
758-17-8	2.60E+00	2.69E+00	2.76E+00	2.44E+00	5.13E+01	1.55E-03	1.15E+01	1.15E+01
759-73-9	1.25E+01	1.27E+01	1.28E+01	1.22E+01	1.01E+02	5.15E-03	2.87E+01	2.87E+01
76-13-1	2.16E-02	2.19E-02	2.21E-02	2.11E-02	6.11E+02	5.24E-03	1.23E+00	1.23E+00
76-25-5	2.32E+03	2.36E+03	2.38E+03	2.26E+03	1.55E+04	6.76E-09	2.98E+03	2.98E+03
76-57-3	1.11E+02	1.27E+02	1.37E+02	8.51E+01	4.11E+03	6.53E-11	6.62E+00	6.64E+00
760-60-1	1.01E+01	1.04E+01	1.05E+01	9.76E+00	1.15E+02	3.20E-03	2.64E+01	2.64E+01
76180-96-6	3.89E+02	4.00E+02	4.08E+02	3.70E+02	6.31E+03	8.64E-06	1.67E+03	1.67E+03
764-41-0	1.60E-01	1.60E-01	1.60E-01	1.60E-01	4.19E+02	2.75E-02	1.73E+01	1.73E+01
765-34-4	3.60E+00	3.68E+00	3.74E+00	3.45E+00	1.84E+02	1.41E-02	4.49E+01	4.49E+01
7681-93-8	2.08E+02	2.09E+02	2.09E+02	2.07E+02	7.23E+02	3.79E-28	2.99E+02	2.88E+02
77-06-5	1.43E+02	1.57E+02	1.67E+02	1.18E+02	3.87E+03	1.20E-13	7.62E+01	7.62E+01
77-09-8	5.56E+02	5.97E+02	6.25E+02	4.86E+02	1.18E+04	9.44E-10	3.77E+02	3.77E+02
77-46-3	8.72E+02	8.81E+02	8.87E+02	8.57E+02	4.80E+03	9.56E-11	1.15E+03	1.15E+03
77-65-6	3.40E+02	3.50E+02	3.57E+02	3.24E+02	3.38E+03	7.11E-04	4.99E+02	4.99E+02
77-79-2	7.31E+00	7.63E+00	7.84E+00	6.78E+00	2.08E+02	9.89E-04	6.39E+01	6.39E+01
77-83-8	6.54E+01	6.66E+01	6.74E+01	6.35E+01	4.43E+03	1.44E+00	2.06E+02	2.06E+02
785-30-8	4.61E+01	5.30E+01	5.76E+01	3.47E+01	2.31E+03	2.14E-08	2.35E+02	3.01E+02
79-24-3	2.59E+00	2.60E+00	2.60E+00	2.59E+00	7.07E+01	1.26E-01	9.78E+00	9.78E+00
79-40-3	5.53E+01	5.64E+01	5.71E+01	5.35E+01	8.75E+02	6.54E-05	2.95E+02	2.95E+02
79-44-7	6.86E+00	7.08E+00	7.22E+00	6.51E+00	1.69E+02	6.97E-03	4.61E+01	4.61E+01
80-07-9	2.39E+03	2.62E+03	2.77E+03	2.01E+03	5.28E+04	4.07E-01	9.02E+02	9.02E+02
80-08-0	8.76E+01	9.62E+01	1.02E+02	7.32E+01	2.92E+03	5.37E-06	3.59E+02	3.59E+02
8015-30-3	5.57E+02	6.65E+02	7.37E+02	3.77E+02	3.20E+04	2.23E-04	2.61E+03	2.61E+03
81-15-2	7.49E+03	8.17E+03	8.62E+03	6.35E+03	1.77E+05	1.99E+01	3.36E+03	3.36E+03
81-16-3	4.19E+01	4.37E+01	4.50E+01	3.88E+01	8.86E+02	4.05E-09	1.52E+02	2.62E+02
81-21-0	5.06E+00	5.15E+00	5.21E+00	4.91E+00	2.01E+02	1.93E-02	4.57E+01	4.57E+01
81-49-2	5.53E+03	6.27E+03	6.77E+03	4.28E+03	1.67E+05	1.36E+00	2.36E+02	2.36E+02
811-97-2	4.20E-01	4.21E-01	4.21E-01	4.20E-01	6.09E+01	8.41E-02	1.30E+00	1.30E+00

816-57-9	5.30E+00	5.41E+00	5.49E+00	5.12E+00	1.16E+02	9.73E-03	2.90E+01	2.90E+01
82-28-0	3.09E+02	3.96E+02	4.53E+02	1.64E+02	2.32E+04	1.87E-02	2.82E+02	2.82E+02
838-88-0	1.91E+02	2.54E+02	2.96E+02	8.51E+01	1.59E+04	2.24E-03	5.70E+02	5.70E+02
842-00-2	1.04E+03	1.06E+03	1.08E+03	9.90E+02	1.06E+04	1.87E-05	1.96E+03	1.96E+03
842-07-9	1.24E+03	1.50E+03	1.67E+03	8.09E+02	6.68E+04	1.16E-01	1.79E+02	1.79E+02
846-50-4	7.94E+02	8.10E+02	8.22E+02	7.66E+02	6.59E+03	9.42E-07	1.07E+03	1.07E+03
853-23-6	1.75E+03	2.38E+03	2.80E+03	7.02E+02	1.59E+05	5.22E-02	2.31E+03	2.31E+03
86-29-3	2.12E+02	2.35E+02	2.49E+02	1.75E+02	6.32E+03	7.94E-02	2.33E+02	2.33E+02
86-86-2	5.17E+01	6.76E+01	7.82E+01	2.51E+01	4.07E+03	2.98E-04	1.44E+01	1.44E+01
86-88-4	3.83E+01	4.79E+01	5.42E+01	2.25E+01	2.63E+03	3.74E-05	2.09E+02	2.09E+02
860-22-0	2.40E+02	2.40E+02	2.40E+02	2.40E+02	7.65E+02	1.83E-20	3.41E+02	3.41E+02
86315-52-8	7.13E+02	7.16E+02	7.19E+02	7.07E+02	3.23E+03	6.46E-09	1.06E+03	1.06E+03
869-01-2	1.48E+01	1.50E+01	1.52E+01	1.43E+01	1.39E+02	2.27E-03	3.28E+01	3.28E+01
87-29-6	1.02E+02	1.40E+02	1.65E+02	3.83E+01	2.17E+04	1.34E-01	1.49E+02	1.49E+02
88-19-7	1.28E+02	1.30E+02	1.31E+02	1.26E+02	7.17E+02	3.82E-03	1.70E+02	1.70E+02
88-96-0	1.29E+02	1.30E+02	1.30E+02	1.28E+02	6.70E+02	6.43E-06	2.26E+02	2.26E+02
88107-10-2	6.55E+02	7.30E+02	7.80E+02	5.30E+02	2.44E+04	1.68E-08	4.98E+02	2.15E+03
89-25-8	3.40E+01	3.86E+01	4.16E+01	2.63E+01	1.20E+03	8.31E-04	2.87E+01	2.87E+01
90-49-3	8.11E+01	8.45E+01	8.68E+01	7.54E+01	1.18E+03	1.99E-06	1.65E+02	1.65E+02
90-94-8	2.33E+02	3.17E+02	3.73E+02	9.36E+01	2.05E+04	3.43E-04	4.66E+02	4.62E+02
91-59-8	3.68E+01	4.94E+01	5.78E+01	1.59E+01	3.87E+03	4.90E-03	1.95E+02	1.92E+02
91-62-3	7.02E+00	8.32E+00	9.18E+00	4.85E+00	8.94E+02	1.99E-02	1.09E+01	1.45E+01
91-76-9	1.49E+02	1.56E+02	1.60E+02	1.38E+02	1.88E+03	2.46E-05	1.74E+02	1.81E+02
91-79-2	7.30E+01	1.01E+02	1.20E+02	2.58E+01	6.97E+03	1.17E-05	2.20E+01	2.26E+01
91-93-0	2.26E+02	3.14E+02	3.73E+02	7.85E+01	4.65E+04	3.49E-01	5.66E+01	5.66E+01
915-67-3	1.50E+02	1.50E+02	1.50E+02	1.50E+02	4.78E+02	4.26E-24	2.13E+02	2.13E+02
92-55-7	1.57E+02	1.62E+02	1.64E+02	1.50E+02	2.21E+03	1.82E-02	5.14E+02	5.14E+02
92-67-1	3.05E+01	3.92E+01	4.49E+01	1.61E+01	5.55E+03	5.45E-02	1.78E+02	1.69E+02
92-84-2	3.50E+02	5.04E+02	6.06E+02	9.31E+01	4.77E+04	4.35E-02	3.27E+02	3.27E+02
924-16-3	2.61E+00	2.68E+00	2.72E+00	2.51E+00	5.51E+02	4.18E-02	1.99E+01	1.99E+01
924-42-5	1.41E+01	1.43E+01	1.45E+01	1.36E+01	1.81E+02	1.05E-06	5.82E+01	5.82E+01
93-46-9	2.67E+03	3.63E+03	4.26E+03	1.09E+03	2.78E+05	9.28E-03	8.66E+01	8.66E+01
930-55-2	1.56E+01	1.57E+01	1.58E+01	1.55E+01	1.27E+02	3.25E-03	4.99E+01	4.99E+01
932-83-2	1.13E+01	1.18E+01	1.21E+01	1.06E+01	3.99E+02	2.47E-02	1.17E+02	1.17E+02
934-00-9	9.08E+00	1.08E+01	1.19E+01	6.23E+00	5.63E+02	1.70E-04	7.29E+01	7.29E+01
938-73-8	4.24E+01	4.52E+01	4.71E+01	3.77E+01	9.94E+02	5.91E-05	1.38E+02	1.38E+02
93957-54-1	1.07E+04	1.12E+04	1.16E+04	9.85E+03	1.59E+05	1.66E-10	1.01E+04	1.01E+04
94-20-2	6.36E+02	6.61E+02	6.77E+02	5.95E+02	8.16E+03	8.94E-05	1.11E+03	1.11E+03
94-26-8	3.90E+01	4.68E+01	5.21E+01	2.58E+01	2.32E+03	6.90E-03	9.81E+00	9.81E+00

94-36-0	1.97E+02	2.07E+02	2.13E+02	1.82E+02	8.35E+03	1.36E+00	1.96E+02	1.96E+02
94-58-6	8.69E-01	8.90E-01	9.04E-01	8.35E-01	1.32E+03	1.18E-01	1.89E+01	1.89E+01
94-59-7	7.23E-01	7.37E-01	7.47E-01	6.99E-01	1.39E+03	9.57E-02	2.48E+01	2.48E+01
95-71-6	3.88E+01	4.30E+01	4.58E+01	3.17E+01	1.44E+03	1.85E-05	1.81E+02	1.81E+02
95-79-4	2.76E+00	2.98E+00	3.12E+00	2.39E+00	1.47E+03	6.88E-02	6.47E+01	6.40E+01
95-83-0	1.12E+01	1.40E+01	1.59E+01	6.53E+00	2.12E+03	1.76E-02	1.87E+02	1.93E+02
959-24-0	2.34E+02	2.39E+02	2.43E+02	2.26E+02	4.08E+03	4.36E-09	1.27E+03	1.27E+03
96-69-5	1.38E+03	1.62E+03	1.78E+03	9.79E+02	6.59E+04	5.91E-02	1.52E+01	1.52E+01
968-81-0	5.78E+02	5.98E+02	6.11E+02	5.45E+02	6.80E+03	2.01E-08	7.86E+02	7.86E+02
97-16-5	1.41E+03	1.56E+03	1.65E+03	1.18E+03	4.71E+04	3.32E+00	4.75E+02	4.75E+02
97-18-7	1.34E+04	1.51E+04	1.62E+04	1.07E+04	4.08E+05	1.06E-02	2.64E+03	2.64E+03
97-56-3	5.99E+02	7.63E+02	8.72E+02	3.26E+02	4.41E+04	5.09E-02	7.11E+02	7.06E+02
97-59-6	1.96E+01	1.99E+01	2.01E+01	1.92E+01	1.98E+02	1.56E-09	6.58E+01	6.58E+01
971-15-3	7.57E+03	8.64E+03	9.35E+03	5.79E+03	2.90E+05	1.48E-05	1.91E+03	1.91E+03
98-85-1	2.84E+00	2.90E+00	2.93E+00	2.74E+00	1.99E+02	1.95E-02	2.36E+01	2.36E+01
98-96-4	3.97E+01	4.01E+01	4.04E+01	3.91E+01	2.14E+02	1.98E-05	5.73E+01	5.73E+01
98319-26-7	1.10E+04	1.12E+04	1.13E+04	1.06E+04	8.36E+04	2.12E-05	1.35E+04	1.35E+04
989-38-8	8.67E+02	9.57E+02	1.02E+03	7.17E+02	2.45E+04	7.41E-15	3.37E+02	3.37E+02
99-50-3	1.97E+01	2.06E+01	2.12E+01	1.82E+01	2.85E+02	1.67E-09	3.42E+01	3.42E+01
99-57-0	1.59E+02	1.62E+02	1.64E+02	1.54E+02	1.38E+03	8.54E-05	2.90E+02	2.90E+02
99-59-2	1.06E+02	1.13E+02	1.18E+02	9.39E+01	2.77E+03	1.24E-01	2.37E+02	2.37E+02

References

1. Rebitzer, G.; Ekvall, T.; Frischknecht, R.; Hunkeler, D.; Norris, G.; Rydberg, T.; Schmidt, W. P.; Suh, S.; Weidema, B. P.; Pennington, D. W., Life cycle assessment Part 1: Framework, goal and scope definition, inventory analysis, and applications. *Environment International* **2004**, *30*, (5), 701-720.
2. Guinee, J. B. In *Life cycle assessment: past, present and future*, International Symposium on Life Cycle Assessment and Construction: Civil Engineering and Buildings, Nantes, FRANCE, Jul 10-12, 2012; Nantes, FRANCE, 2012; pp 9-11.
3. Hellweg, S.; Canals, L. M. I., Emerging approaches, challenges and opportunities in life cycle assessment. *Science* **2014**, *344*, (6188), 1109-1113.
4. Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbuhler, K., Molecular-structure-based models of chemical inventories using neural networks. *Environmental Science & Technology* **2008**, *42*, (17), 6717-6722.
5. Song, R. S.; Keller, A. A.; Suh, S., Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environmental Science & Technology* **2017**, *51*, (18), 10777-10785.
6. Wisthoff, A.; Ferrero, V.; Huynh, T.; DuPont, B. In *Quantifying the Impact of Sustainable Product Design Decisions in the Early Design Phase Through Machine Learning*, ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2016; American Society of Mechanical Engineers: 2016; pp V004T05A043-V004T05A043.
7. Seo, K.-K.; Kim, W.-K. In *Approximate life cycle assessment of product concepts using a hybrid genetic algorithm and neural network approach*, International Conference on Hybrid Information Technology, 2006; Springer: 2006; pp 258-268.
8. Seo, K.-K.; Min, S.-H.; Yoo, H.-W. In *Artificial neural network based life cycle assessment model for product concepts using product classification method*, International Conference on Computational Science and Its Applications, 2005; Springer: 2005; pp 458-466.
9. Park, J. H.; Seo, K.-K., Approximate life cycle assessment of product concepts using multiple regression analysis and artificial neural networks. *Journal of Mechanical Science and Technology* **2003**, *17*, (12), 1969-1976.
10. Park, J.-H.; Seo, K.-K.; Wallace, D. In *Approximate life cycle assessment of classified products using artificial neural network and statistical analysis in conceptual product design*, Environmentally Conscious Design and Inverse Manufacturing, 2001. Proceedings EcoDesign 2001: Second International Symposium on, 2001; IEEE: 2001; pp 321-326.
11. Chen, J. L.; Liao, C.-W. In *A simple life cycle assessment method for green product conceptual design*, Environmentally Conscious Design and Inverse Manufacturing, 2001. Proceedings EcoDesign 2001: Second International Symposium on, 2001; IEEE: 2001; pp 775-780.

12. Azari, R.; Garshasbi, S.; Amini, P.; Rashed-Ali, H.; Mohammadi, Y., Multi-objective optimization of building envelope design for life cycle environmental performance. *Energy and Buildings* **2016**, *126*, 524-534.
13. Nabavi-Pelesaraei, A.; Bayat, R.; Hosseinzadeh-Bandbafha, H.; Afrasyabi, H.; Chau, K.-w., Modeling of energy consumption and environmental life cycle assessment for incineration and landfill systems of municipal solid waste management-A case study in Tehran Metropolis of Iran. *Journal of Cleaner Production* **2017**, *148*, 427-440.
14. Khoshnevisan, B.; Rafiee, S.; Omid, M.; Mousazadeh, H.; Sefeedpari, P., Prognostication of environmental indices in potato production using artificial neural networks. *Journal of cleaner production* **2013**, *52*, 402-409.
15. Ozbilen, A.; Aydin, M.; Dincer, I.; Rosen, M. A., Life cycle assessment of nuclear-based hydrogen production via a copper–chlorine cycle: A neural network approach. *International Journal of Hydrogen Energy* **2013**, *38*, (15), 6314-6322.
16. Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbuhler, K., Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chemistry* **2009**, *11*, (11), 1826-1831.
17. Sundaravaradan, N.; Marwah, M.; Shah, A.; Ramakrishnan, N.; Ieee, Data Mining Approaches for Life Cycle Assessment. *2011 Ieee International Symposium on Sustainable Systems and Technology (Issst)* **2011**.
18. Ramakrishnan, N.; Marwah, M.; Shah, A.; Patnaik, D.; Hossain, M. S.; Sundaravaradan, N.; Patel, C., Data mining solutions for sustainability problems. *Ieee Potentials* **2012**, *31*, (6), 28-34.
19. Marwah, M.; Shah, A.; Bash, C.; Patel, C.; Ramakrishnan, N., Using Data Mining to Help Design Sustainable Products. *Computer* **2011**, *44*, (8), 103-106.
20. Yuan, Y.; Yuan, J.; Du, H.; Li, L., Pareto Ant Colony Algorithm for Building Life Cycle Energy Consumption Optimization. *Life System Modeling and Intelligent Computing, Pt Ii* **2010**, *98*, 59-+.
21. Zhou, Q.; Zhou, H.; Zhu, Y.; Li, T. In *Data-driven Solutions for Building Environmental Impact Assessment*, IEEE 9th International Conference on Semantic Computing, IEEE Computer Society, california, UNITED STATES, 2015 Feb 07-09, 2015; IEEE Computer Society, california, UNITED STATES, 2015; pp 316-319.
22. Frischknecht, R.; Jungbluth, N.; Althaus, H. J.; Doka, G.; Dones, R.; Heck, T.; Hellweg, S.; Hirschler, R.; Nemecek, T.; Rebitzer, G.; Spielmann, M., The ecoinvent database: Overview and methodological framework. *International Journal of Life Cycle Assessment* **2005**, *10*, (1), 3-9.
23. Piao, W.; Kim, C.; Cho, S.; Kim, H.; Kim, M.; Kim, Y., Development of a protocol to optimize electric power consumption and life cycle environmental impacts for operation of wastewater treatment plant. *Environmental Science and Pollution Research* **2016**, *23*, (24), 25451-25466.
24. Chiang, T.-A.; Che, Z.; Wang, T.-T., A design for environment methodology for evaluation and improvement of derivative consumer electronic product development. *Journal of Systems Science and Systems Engineering* **2011**, *20*, (3), 260-274.
25. Chiang, T.-A.; Roy, R., An intelligent benchmark-based design for environment system for derivative electronic product development. *Computers in Industry* **2012**, *63*, (9), 913-929.

26. Yin, L.; Liao, Y.; Zhou, L.; Wang, Z.; Ma, X. In *Life cycle assessment of coal-fired power plants and sensitivity analysis of CO₂ emissions from power generation side*, IOP Conference Series: Materials Science and Engineering, 2017; IOP Publishing: 2017; p 012055.
27. Suh, S.; Huppes, G., Missing Inventory Estimation Tool using extended Input-Output Analysis. *International Journal of Life Cycle Assessment* **2002**, *7*, (3), 134-140.
28. Frischknecht, R.; Jolliet, O., Global guidance for life cycle impact assessment indicators. *UNEP/SETAC Life Cycle Initiative, Paris* **2016**.
29. Rosenbaum, R. K.; Bachmann, T. M.; Gold, L. S.; Huijbregts, M. A.; Jolliet, O.; Juraske, R.; Koehler, A.; Larsen, H. F.; MacLeod, M.; Margni, M., 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* **2008**, *13*, (7), 532.
30. Hinds, R. d. C.; Weller, J. L., Toxic Substances Control Act. *Environmental Law Practice Guide* **2016**, *4*.
31. Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V., A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP. *Excli Journal* **2009**, *8*, 74-88.
32. Mayo-Bean, K.; Nabholz, J.; Clements, R.; Zeeman, M.; Henry, T.; Rodier, D.; Moran, K.; Meylan, B.; Ranslow, P., Methodology document for the ECOlogical Structure-Activity Relationship Model (ECOSAR) class program: Estimating toxicity of industrial chemicals to aquatic organisms using ECOSAR class program (Ver. 1.1). *US Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention, Office of Pollution Prevention and Toxics, Washington, DC* **2011**.
33. Furuhashi, A.; Toida, T.; Nishikawa, N.; Aoki, Y.; Yoshioka, Y.; Shiraishi, H., Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. *Sar and Qsar in Environmental Research* **2010**, *21*, (5-6), 403-413.
34. Martin, T. *User's guide for TEST (version 4.2)(Toxicity Estimation Software Tool) A program to estimate toxicity from molecular structure*. *US EPA Office of Research and Development, Washington, DC*; EPA/600/R-16/058 Google Scholar: 2016.
35. Plus, S. *ADMET Predictor*.
36. Kostal, J.; Voutchkova-Kostal, A.; Anastas, P. T.; Zimmerman, J. B., Identifying and designing chemicals with minimal acute aquatic toxicity. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, (20), 6289-6294.
37. Voutchkova-Kostal, A. M.; Kostal, J.; Connors, K. A.; Brooks, B. W.; Anastas, P. T.; Zimmerman, J. B., Towards rational molecular design for reduced chronic aquatic toxicity. *Green Chemistry* **2012**, *14*, (4), 1001-1008.
38. Voutchkova, A. M.; Kostal, J.; Steinfeld, J. B.; Emerson, J. W.; Brooks, B. W.; Anastas, P.; Zimmerman, B., Towards rational molecular design: derivation of property guidelines for reduced acute aquatic toxicity. *Green Chemistry* **2011**, *13*, (9), 2373-2379.
39. Melnikov, F.; Kostal, J.; Voutchkova-Kostal, A.; Zimmerman, J. B.; Anastas, P. T., Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals. *Green Chemistry* **2016**, *18*, (16), 4432-4445.

40. Mistry, P.; Neagu, D.; Sanchez-Ruiz, A.; Trundle, P. R.; Vessey, J. D.; Gosling, J. P., Prediction of the effect of formulation on the toxicity of chemicals. *Toxicology Research* **2017**, *6*, (1), 42-53.
41. Gomes, A. I.; Pires, J. C. M.; Figueiredo, S. A.; Boaventura, R. A. R., Multiple linear and principal component regressions for modelling ecotoxicity bioassay response. *Environmental Technology* **2014**, *35*, (8), 945-955.
42. Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29*, (6-7), 476-488.
43. Miller, T. H.; Gallidabino, M. D.; MacRae, J. I.; Hogstrand, C.; Bury, N. R.; Barron, L. P.; Snape, J. R.; Owen, S. F., Machine Learning for Environmental Toxicology: A Call for Integration and Innovation. *Environmental Science & Technology* **2018**, *52*, (22), 12953-12955.
44. Singh, K. P.; Gupta, S.; Kumar, A.; Mohan, D., Multispecies QSAR Modeling for Predicting the Aquatic Toxicity of Diverse Organic Chemicals for Regulatory Toxicology. *Chemical Research in Toxicology* **2014**, *27*, (5), 741-753.
45. Li, F. X.; Fan, D. F.; Wang, H.; Yang, H. B.; Li, W. H.; Tang, Y.; Liu, G. X., In silico prediction of pesticide aquatic toxicity with chemical category approaches. *Toxicology Research* **2017**, *6*, (6), 831-842.
46. Sala, S.; Marinov, D.; Pennington, D., Spatial differentiation of chemical removal rates from air in life cycle impact assessment. *International Journal of Life Cycle Assessment* **2011**, *16*, (8), 748-760.
47. Marvuglia, A.; Kanevski, M.; Benetto, E., Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. *Environment international* **2015**, *83*, 72-85.
48. Marvuglia, A.; Kanevski, M.; Leuenberger, M.; Benetto, E. In *Variables selection for ecotoxicity and human toxicity characterization using Gamma Test*, International Conference on Computational Science and Its Applications, 2014; Springer: 2014; pp 640-652.
49. Marvuglia, A.; Leuenberger, M.; Kanevski, M.; Benetto, E., Random Forest for Toxicity of Chemical Emissions: Features Selection and Uncertainty Quantification. *Journal of Environmental Accounting and Management* **2015**, *3*, (3), 229-241.
50. Birkved, M.; Heijungs, R., Simplified fate modelling in respect to ecotoxicological and human toxicological characterisation of emissions of chemical compounds. *International Journal of Life Cycle Assessment* **2011**, *16*, (8), 739-747.
51. Hou, P.; Cai, J. R.; Qu, S.; Xu, M., Estimating Missing Unit Process Data in Life Cycle Assessment Using a Similarity-Based Approach. *Environmental Science & Technology* **2018**, *52*, (9), 5259-5267.
52. Suh, S.; Huppes, G., Methods for life cycle inventory of a product. *Journal of Cleaner Production* **2005**, *13*, (7), 687-697.
53. Souma, W.; Fujiwara, Y.; Aoyama, H., Complex networks and economics. *Physica a-Statistical Mechanics and Its Applications* **2003**, *324*, (1-2), 396-401.
54. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavezf, M.; Hwang, D. U., Complex Networks: Structure and Dynamics. *Complex Systems and Complexity Science* **2007**, *4*, (1), 49-92.

55. Wang, X. F., Complex networks: Topology, dynamics and synchronization. *International Journal of Bifurcation and Chaos* **2002**, *12*, (5), 885-916.
56. Liben-Nowell, D.; Kleinberg, J., The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **2007**, *58*, (7), 1019-1031.
57. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.-C., Bipartite network projection and personal recommendation. *2011 International Conference on Applied Social Science (Icass 2011), Vol Iii* **2011**, 489-+.
58. Zhou, T.; Kuscsik, Z.; Liu, J. G.; Medo, M.; Wakeling, J. R.; Zhang, Y. C., Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, (10), 4511-4515.
59. Zeng, W.; Shang, M.-S.; Zhang, Q.-M.; Lue, L.; Zhou, T., CAN DISSIMILAR USERS CONTRIBUTE TO ACCURACY AND DIVERSITY OF PERSONALIZED RECOMMENDATION? *International Journal of Modern Physics C* **2010**, *21*, (10), 1217-1227.
60. Zhang, Q.-M.; Shang, M.-S.; Zeng, W.; Chen, Y.; Lue, L., Empirical comparison of local structural similarity indices for collaborative-filtering-based recommender systems. *International Conference on Complexity and Interdisciplinary Sciences: 3rd China-Europe Summer School on Complexity Sciences* **2010**, *3*, (5), 1887-1896.
61. Ben Schafer, J.; Konstan, J. A.; Riedl, J., E-commerce recommendation applications. *Data Mining and Knowledge Discovery* **2001**, *5*, (1-2), 115-153.
62. Goldberg, D. S.; Roth, F. P., Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, (8), 4372-4376.
63. Clauset, A.; Moore, C.; Newman, M. E. J., Hierarchical structure and the prediction of missing links in networks. *Nature* **2008**, *453*, (7191), 98-101.
64. Lue, L.; Zhou, T., Link prediction in complex networks: A survey. *Physica a-Statistical Mechanics and Its Applications* **2011**, *390*, (6), 1150-1170.
65. Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A., The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, (11), 3747-3752.
66. Newman, M. E. J., Analysis of weighted networks. *Physical Review E* **2004**, *70*, (5).
67. Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B., Theecoinvent database version 3 (part I): overview and methodology. *International Journal of Life Cycle Assessment* **2016**, *21*, (9), 1218-1230.
68. Weidema, B. P.; Bauer, C.; Hischer, R.; Mutel, C.; Nemecek, T.; Reinhard, J.; Vadenbo, C.; Wernet, G. *Overview and methodology: Data quality guideline for the ecoinvent database version 3*; Swiss Centre for Life Cycle Inventories: 2013.
69. Vershynin, R., Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* **2010**.
70. U.S. Life Cycle Inventory Database. In 2012; <https://www.lcacommons.gov/nrel/search>.
71. Cooper, J. S.; Kahn, E., Commentary on issues in data quality analysis in life cycle assessment. *International Journal of Life Cycle Assessment* **2012**, *17*, (4), 499-503.
72. McKone, T. E.; Nazaroff, W. W.; Berck, P.; Auffhammer, M.; Lipman, T.; Torn, M. S.; Masanet, E.; Lobscheid, A.; Santero, N.; Mishra, U.; Barrett, A.; Bomberg, M.;

- Fingerman, K.; Scown, C.; Strogon, B.; Horvath, A., Grand Challenges for Life-Cycle Assessment of Biofuels. *Environmental Science & Technology* **2011**, *45*, (5), 1751-1756.
73. Wang, M. Q. *REET 1.5-transportation fuel-cycle model-Vol. 1: methodology, development, use, and results*; Argonne National Lab., IL (US): 1999.
74. Aggarwal, C. C., *Neural networks and deep learning*. Springer: 2018.
75. Holland, J. H., Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of Michigan Press* **1975**, 439-444.
76. Pham, D.; Karaboga, D., *Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks*. Springer Science & Business Media: 2012.
77. Benardos, P.; Vosniakos, G.-C., Optimizing feedforward artificial neural network architecture. *Engineering Applications of Artificial Intelligence* **2007**, *20*, (3), 365-382.
78. Ritchie, M. D.; White, B. C.; Parker, J. S.; Hahn, L. W.; Moore, J. H., Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC bioinformatics* **2003**, *4*, (1), 28.
79. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M., Application of a genetic algorithm and an artificial neural network for global prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Monatshefte Fur Chemie* **2009**, *140*, (11), 1279-1288.
80. Drgan, V.; Zuperl, S.; Vracko, M.; Como, F.; Novic, M., Robust modelling of acute toxicity towards fathead minnow (*Pimephales promelas*) using counter-propagation artificial neural networks and genetic algorithm. *Sar and Qsar in Environmental Research* **2016**, *27*, (7), 501-519.
81. Payet, J., Assessing toxic impacts on aquatic ecosystems in life cycle assessment (LCA). *Ecole Polytechnique Fédérale de Lausanne, Lausanne* **2004**.
82. Hauschild, M. Z.; Huijbregts, M.; Jolliet, O.; MacLeod, M.; Margni, M.; van de Meent, D. V.; Rosenbaum, R. K.; McKone, T. E., Building a model based on scientific consensus for life cycle impact assessment of chemicals: The search for harmony and parsimony. *Environmental Science & Technology* **2008**, *42*, (19), 7032-7037.
83. Henderson, A. D.; Hauschild, M. Z.; van de Meent, D.; Huijbregts, M. A. J.; Larsen, H. F.; Margni, M.; McKone, T. E.; Payet, J.; Rosenbaum, R. K.; Jolliet, O., USEtox fate and ecotoxicity factors for comparative assessment of toxic emissions in life cycle analysis: sensitivity to key chemical properties. *International Journal of Life Cycle Assessment* **2011**, *16*, (8), 701-709.
84. Stieger, G.; Scheringer, M.; Ng, C. A.; Hungerbuhler, K., Assessing the persistence, bioaccumulation potential and toxicity of brominated flame retardants: Data availability and quality for 36 alternative brominated flame retardants. *Chemosphere* **2014**, *116*, 118-123.
85. Fantke, P. E.; Bijster, M.; Guignard, C.; Hauschild, M.; Huijbregts, M.; Jolliet, O.; Kounina, A.; Magaud, V.; Margni, M.; McKone, T. E.; Posthuma, L.; Rosenbaum, R. K.; van de Meent, D.; van Zelm, R., USEtox® 2.0 user manual (Version 1). <http://usetox.org> **2017**.
86. Kienzler, A.; Barron, M. G.; Belanger, S. E.; Beasley, A.; Embry, M. R., Mode of Action (MOA) Assignment Classifications for Ecotoxicology: An Evaluation of Approaches. *Environmental Science & Technology* **2017**, *51*, (17), 10203-10211.

87. Patlewicz, G.; Jeliaskova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B., An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *Sar and Qsar in Environmental Research* **2008**, *19*, (5-6), 495-524.
88. Verhaar, H. J.; Van Leeuwen, C. J.; Hermens, J. L., Classifying environmental pollutants. *Chemosphere* **1992**, *25*, (4), 471-491.
89. Todeschini, R.; Consonni, V., *Handbook of molecular descriptors*. John Wiley & Sons: 2008; Vol. 11.
90. Schrödinger, L., QikProp, version 3.5. *New York, NY* **2012**.
91. Haykin, S.; Network, N., A comprehensive foundation. *Neural networks* **2004**, *2*, (2004), 41.
92. MacKay, D. J., Bayesian interpolation. *Neural computation* **1992**, *4*, (3), 415-447.
93. Foresee, F. D.; Hagan, M. T. In *Gauss-Newton approximation to Bayesian learning*, Neural networks, 1997., international conference on, 1997; IEEE: 1997; pp 1930-1935.
94. Demuth, H. B.; Beale, M. H.; De Jess, O.; Hagan, M. T., *Neural network design*. Martin Hagan: 2014.
95. Heermann, P. D.; Khazenie, N., Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Transactions on Geoscience and Remote Sensing* **1992**, *30*, (1), 81-88.
96. Gori, M.; Tesi, A., On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1992**, *14*, (1), 76-86.
97. Kingma, D. P.; Ba, J., Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
98. Duchi, J.; Hazan, E.; Singer, Y., Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **2011**, *12*, (Jul), 2121-2159.
99. Zeiler, M. D., ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* **2012**.
100. Dozat, T., Incorporating nesterov momentum into adam. **2016**.
101. Whitley, D.; Starkweather, T.; Bogart, C., Genetic algorithms and neural networks: Optimizing connections and connectivity. *Parallel computing* **1990**, *14*, (3), 347-361.
102. Yao, X.; Liu, Y., A new evolutionary system for evolving artificial neural networks. *IEEE transactions on neural networks* **1997**, *8*, (3), 694-713.
103. García-Pedrajas, N.; Hervás-Martínez, C.; Muñoz-Pérez, J., COVNET: a cooperative coevolutionary model for evolving artificial neural networks. *IEEE Transactions on Neural Networks* **2003**, *14*, (3), 575-596.
104. Lee, C.-Y.; Yao, X., Evolutionary programming using mutations based on the Lévy probability distribution. *IEEE Transactions on Evolutionary Computation* **2004**, *8*, (1), 1-13.
105. Oong, T. H.; Isa, N. A. M., Adaptive evolutionary artificial neural networks for pattern classification. *IEEE Transactions on Neural Networks* **2011**, *22*, (11), 1823-1836.
106. Montana, D. J.; Davis, L. In *Training Feedforward Neural Networks Using Genetic Algorithms*, IJCAI, 1989; 1989; pp 762-767.
107. Zi-wu, R.; Ye, S., Improvement of real-valued genetic algorithm and performance study [j]. *Acta Electronica Sinica* **2007**, *2*, 017.

108. Sattar, M. A.; Islam, M. M.; Murase, K. In *A new constructive algorithm for designing and training artificial neural networks*, International Conference on Neural Information Processing, 2007; Springer: 2007; pp 317-327.
109. Islam, M. M.; Sattar, M. A.; Amin, M. F.; Yao, X.; Murase, K., A new constructive algorithm for architectural and functional adaptation of artificial neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2009**, *39*, (6), 1590-1605.
110. Subirats, J. L.; Franco, L.; Jerez, J. M., C-Mantec: A novel constructive neural network algorithm incorporating competition between neurons. *Neural Networks* **2012**, *26*, 130-140.
111. Engelbrecht, A. P., A new pruning heuristic based on variance analysis of sensitivity information. *IEEE transactions on Neural Networks* **2001**, *12*, (6), 1386-1399.
112. Lauret, P.; Fock, E.; Mara, T. A., A node pruning algorithm based on a Fourier amplitude sensitivity test method. *IEEE transactions on neural networks* **2006**, *17*, (2), 273-293.
113. Alvarez, A., A neural network with evolutionary neurons. *Neural processing letters* **2002**, *16*, (1), 43-52.
114. Kim, H. B.; Jung, S. H.; Kim, T. G.; Park, K. H., Fast learning method for back-propagation neural network by evolutionary adaptation of learning rates. *Neurocomputing* **1996**, *11*, (1), 101-106.
115. Guo, Z.; Uhrig, R. E. In *Using genetic algorithms to select inputs for neural networks*, Combinations of Genetic Algorithms and Neural Networks, 1992., COGANN-92. International Workshop on, 1992; IEEE: 1992; pp 223-234.
116. Bishop, C. M., *Neural networks for pattern recognition*. Oxford university press: 1995.
117. Maitra, S.; Yan, J., Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models* **2008**, *79*, 79-90.
118. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A., Conditional variable importance for random forests. *BMC bioinformatics* **2008**, *9*, (1), 307.
119. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the Acm* **2017**, *60*, (6), 84-90.
120. Pereira, S.; Pinto, A.; Alves, V.; Silva, C. A., Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *Ieee Transactions on Medical Imaging* **2016**, *35*, (5), 1240-1251.
121. Friedman, J.; Hastie, T.; Tibshirani, R., *The elements of statistical learning*. Springer series in statistics New York: 2001; Vol. 1.
122. Aldrich, C., *Exploratory analysis of metallurgical process data with neural networks and related methods*. Elsevier: 2002; Vol. 12.
123. Muhlbacher, T.; Piringer, H.; Gratzl, S.; Sedlmair, M.; Streit, M., Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. *Ieee Transactions on Visualization and Computer Graphics* **2014**, *20*, (12), 1643-1652.
124. Bloom, A. D.; de Serres, F., *Ecotoxicity and human health: a biological approach to environmental remediation*. CRC Press: 1995.
125. Reis, C.; Paiva, L.; Moutinho, J.; Marques, V. M. In *Genetic Algorithms and Sensitivity Analysis in Production Planning Optimization*, 10th WSEAS International Conference on Applied Informatics and Communications/3rd WSEAS International Conference on

- Biomedical Electronics and Biomedical Informatics, Taipei, TAIWAN, Aug 20-22, 2010; Taipei, TAIWAN, 2010; pp 246-+.
126. Pinel, F.; Danoy, G.; Bouvry, P., Evolutionary Algorithm Parameter Tuning with Sensitivity Analysis. *Security and Intelligent Information Systems* **2012**, 7053, 204-216.
 127. Xu, M.; Yang, J.; Gao, Z. Y., Parameters Sensitive Analyses for Using Genetic Algorithm to Solve Continuous Network Design Problems. *8th International Conference on Traffic and Transportation Studies (Ictts)* **2012**, 43, 435-444.
 128. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **2014**, 15, 1929-1958.
 129. Goh, G. B.; Hodas, N. O.; Vishnu, A., Deep learning for computational chemistry. *Journal of Computational Chemistry* **2017**, 38, (16), 1291-1307.
 130. Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Korner, R.; Vogt, J.; Tetko, I. V., ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chemical Research in Toxicology* **2016**, 29, (5), 768-775.
 131. Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V., Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* **2015**.
 132. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **2003**, 43, (6), 1947-1958.
 133. Baskin, II, Machine Learning Methods in Computational Toxicology. *Computational Toxicology: Methods and Protocols* **2018**, 1800, 119-139.
 134. Breiman, L., Random forests. *Machine learning* **2001**, 45, (1), 5-32.
 135. Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E., Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *Journal of Chemical Information and Modeling* **2009**, 49, (11), 2481-2488.
 136. Li, N.; Qi, J.; Wang, P.; Zhang, X.; Zhang, T. L.; Li, H., Quantitative structure-activity relationship (QSAR) study of carcinogenicity of polycyclic aromatic hydrocarbons (PAHs) in atmospheric particulate matter by random forest (RF). *Analytical Methods* **2019**, 11, (13), 1816-1821.
 137. DiCiccio, T. J.; Efron, B., Bootstrap confidence intervals. *Statistical science* **1996**, 189-212.
 138. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A., Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA* **2005**, 33, 155-173.
 139. Strobl, C.; Boulesteix, A. L.; Zeileis, A.; Hothorn, T., Bias in random forest variable importance measures: Illustrations, sources and a solution. *Bmc Bioinformatics* **2007**, 8.
 140. Altmann, A.; Tolosi, L.; Sander, O.; Lengauer, T., Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010**, 26, (10), 1340-1347.
 141. Breiman, L., *Classification and regression trees*. Routledge: 2017.
 142. Stekhoven, D. J.; Buhlmann, P., MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, 28, (1), 112-118.
 143. Hotelling, H., Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **1933**, 24, (6), 417.

144. EPA, U., Estimation Program Interface (EPI) Suite. In Ver: 2010.
145. Schapire, R. E.; Freund, Y., *Boosting: Foundations and Algorithms*. 2012; p 1-527.