A DATA-CENTRIC FRAMEWORK TO PERFORM END-OF-LIFE CHEMICAL FLOW ANALYSIS

JOSE DANIEL HERNÁNDEZ BETANCUR

A dissertation submitted for the degree of
DOCTOR IN CHEMICAL SCIENCE AND TECHNOLOGY

at the

UNIVERSIDAD DE SALAMANCA

Academic advisor: Prof. Mariano Martín Martín

Co-advisor: Dr. Gerardo José Ruiz Mercado
U.S. Environmental Protection Agency

Doctoral Program in Chemical Science and Technology

Department of Chemical and Textile Engineering

Universidad de Salamanca

June 2022

VNiVERSiDAD
D SALAMANCA
CAMPUS OF INTERNATIONAL EXCELLENCE

El **Dr. D. Mariano Martín Martín**, Profesor Titular de Universidad del Departamento de Ingeniería Química y Textil de la Universidad de Salamanca, y el **Dr. D. Gerardo José Ruiz Mercado**, Investigador Senior de la U.S. Environmental Protection Agency

Informan:

Que la memoria titulada: "Data-centric framework for performing end-of-life chemical flow analysis", que para optar al Grado de Doctor en Ciencia y Tecnología Química con Mención Internacional presenta **D. Jose Daniel Hernández Betancur**, ha sido realizada bajo nuestra dirección dentro del Programa de Doctorado Ciencia y Tecnología Químicas (RD 99/2011) de la Universidad de Salamanca, y que considerando que constituye un trabajo de tesis.

Autorizan:

Su presentación ante la Escuela de Doctorado de la Universidad de Salamanca

Y para que conste a los efectos oportunos, firmamos la presente.

Firmado digitalmente
por MARTIN
MARTIN MARIANO -
DNI 07977093A
Fecha: 2022.07.05
15:58:48 +02'00'

Fdo: Mariano Martín Martín

GERARDO
RUIZ-
MERCADO

Digitally signed by
GERARDO RUIZ-
MERCADO
Date: 2022.07.05
08:53:18 -04'00'

Fdo: Gerardo José Ruiz Mercado

To my family, Katerine, María José, and Sofia.

Natural ability without education has more often raised a man to glory and virtue than education without natural ability.

— Marcus Aurelius

True happiness is to enjoy the present, without anxious dependence upon the future, not to amuse ourselves with either hopes or fears but to rest satisfied with what we have, which is sufficient, for he that is so wants nothing. The greatest blessings of mankind are within us and within our reach. A wise man is content with his lot, whatever it may be, without wishing for what he has not.

— Lucius Annaeus Seneca

ACKNOWLEDGMENTS

This interesting journey that the thesis represents is something that cannot be summed up simply by the title itself. The wonderful people I had the pleasure of meeting along the way are what I will remember for the rest of my life. The experience and knowledge gained from them will be cherished.

I would like to thank my supervisor at the EPA, Dr. Gerardo Ruiz-Mercado, especially for his patience in dealing with the more complicated parts of my personality and for trusting in my abilities. Furthermore, Julius Enriquez and his wife Mariane, as well as their nephew Jacob, for their good times and friendship. I would like to thank my colleagues at the University of Salamanca: Antonio, José Enrique, José Luceño, Carlos, Elena, Guillermo, Sofía, Manuel, and Edgar, and Dr. Mariano for the good times both inside and outside the university. Although it was a short time in Salamanca, I will remember the tradition of the midday coffee as one of the best moments to share with them and forget whatever problem I may have had. I want to give a special thanks to Edgar for the good times during our time in the United States and because he was the bridge to getting to know the wonderful land of Salamanca.

Many people helped with the development of this thesis. I want to give a special thanks to Dr. Gerardo and Dr. Mariano for their teachings and patience during the development of this thesis. Without them, every word, image, table, and model would not have had the best form for the work to reach its greatest impact. I would like to thank Dr. Wesley Ingwersen and Dr. Raymond Smith for sharing their knowledge and experience, which facilitated the development of my project as a researcher at the EPA. I would also like to thank Dr. John Abraham for being the devil's advocate who allowed many regulatory issues to be considered in this thesis. I would also like to thank the EPA for their funding during my 2 years as an Oak Ridge Institute for Science and Education researcher and Banco Santander for funding my 6 month stay in Spain.

I would like to dedicate this thesis to my mother and my grandparents. My cousin Bernardino, his wife Gina, and his daughter Mikaela for their constant support. Juan Esteban, my brother, and Juan Camilo, my cousin, for setting an analytical and investigative example for me.To my friend Juan Serna for his good conversations about different topics. I want to dedicate this to my great friend Sebastián, his wife, Lauris, and little Rafa. It is always good to have great partners along the way.

Last but not least, I would like to dedicate this thesis to my family, Katerine, María José, and Sofía. I know that this thesis will never make up for the time we were away, but I know it will be worth it, as I said, not only for the title but for the transformative journey that these 4 years meant for all of us.

Medellín (Colombia), June 14, 2022

ABSTRACT

Chemicals are involved in commercial, consumer, and industrial activities across the world marketplace. The number of chemicals circulating in the market keeps growing, which can make it difficult for government agencies to manage chemical risk by imposing regulatory restrictions and for businesses to select chemical candidates with safer profiles before launching products into the market or using them in manufacturing processes.

The identification of chemical exposure scenarios and the quantities of chemicals that may be released into the environment are important tasks for risk evaluation. Nonetheless, comprehensive data must be collected for these tasks, making them time-consuming and challenging. In addition, these tasks are even more difficult at the end-of-life (EoL) stage due to the epistemic uncertainty about the exact pathways taken by chemicals through the EoL management chain.

This thesis aims at moving forward to develop a holistic framework to rapidly perform chemical flow analysis (CFA) for release estimations and allocation and exposure scenario identification at the EoL stage. A data-centric approach is proposed where data engineering plays a crucial role in collecting, transforming, harmonizing, and storing data from publicly-accessible, regulatory, and siloed database systems. First, this thesis explored the use of U.S. regulatory data to track chemicals contained in EoL flows transferred by U.S. industrial facilities to off-site locations for further EoL management. Second, this work moves into industrial facilities to identify the potential pollution abatement technologies implemented by industries, thereby improving the CFA and flow allocation. Third, the results of the above two steps are connected and extended to describe the behavior of the EoL management chain and recycling loop, allowing us to identify the inter-industry flow transfers and potential post-recycling exposure scenarios. Fourth, an effort is made to extend the framework beyond U.S. information and incorporate inventory data from other countries and years. Finally, the data obtained in the fourth step is used to explore the development of data-driven models able to identify potential EoL exposure scenarios and be incorporated into the framework for understanding the EoL management chain.

These studies are intended to contribute to the development and implementation of a methodology for the rapid screening of potential EoL chemical exposure scenarios and release estimates, as well as to address the

challenges in performing chemical risk evaluation for regulatory decision-making and selecting safer profile chemicals based on life cycle thinking.

RESUMEN

Los productos químicos están presentes en actividades comerciales, industriales y de consumo en todo el mercado mundial. El número de sustancias químicas que circulan por el mercado no deja de crecer, lo que puede dificultar a los organismos gubernamentales la gestión del riesgo químico mediante la imposición de restricciones reglamentarias y a las empresas la selección de químicos con perfiles más seguros antes de lanzar los productos al mercado o utilizarlos en los procesos de fabricación.

La identificación de los escenarios de exposición a las sustancias químicas y de las cantidades que pueden liberarse en el medio ambiente son tareas importantes para la evaluación del riesgo. Sin embargo, para estas tareas hay que recopilar datos exhaustivos, lo que las convierte en una tarea larga y difícil. Además, estas tareas son aún más difíciles en la fase de fin de vida (EoL, por sus siglas en inglés) debido a la incertidumbre epistémica sobre las vías exactas que siguen las sustancias químicas a través de la cadena de gestión EoL.

Esta tesis pretende avanzar en el desarrollo de un marco holístico que permita realizar rápidamente el análisis del flujo químico (CFA, por sus siglas en inglés) para la estimación y asignación de emisiones y la identificación de escenarios de exposición en la fase de fin de vida. Se propone un enfoque centrado en los datos, en el que la ingeniería de datos desempeña un papel crucial en la recopilación, transformación, armonización y almacenamiento de datos procedentes de sistemas de bases de datos de acceso público, reglamentarios y aislados. En primer lugar, esta tesis explora el uso de los datos reglamentarios de Estados Unidos para rastrear las sustancias químicas contenidas en los flujos de EoL transferidos por las instalaciones industriales estadounidenses a lugares externos para su posterior gestión. En segundo lugar, este trabajo se adentra en las instalaciones industriales para identificar las posibles tecnologías de reducción de la contaminación aplicadas por las industrias, mejorando así el CFA y la asignación de flujos. En tercer lugar, los resultados de los dos pasos anteriores se conectan y amplían para describir el comportamiento de la cadena de gestión de EoL y el bucle de reciclaje, permitiendonos la identificación de transferencias de flujo entre industrias y los posibles escenarios de exposición posteriores al reciclaje. En cuarto lugar, se hace un esfuerzo por ampliar el marco más allá de la información estadounidense e incorporar datos de inventario de otros países y años. Por último, los datos obtenidos en el cuarto paso se utilizan para explorar el desarrollo de modelos basados en datos capaces

de identificar posibles escenarios de exposición a la EoL y de incorporarlos al marco de comprensión de la cadena de gestión de EoL.

Estos estudios están destinados a contribuir al desarrollo e implementación de una metodología para la detección rápida de posibles escenarios de exposición química de EoL y estimaciones de emisiones, así como para abordar los desafíos en la realización de la evaluación de riesgos químicos para la toma de decisiones regulatorias y la selección de productos químicos de perfil más seguro basado en el pensamiento del ciclo de vida.

PUBLICATIONS

This thesis is based on the following publications:

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2021). A data engineering framework for on-site end-of-life industrial operations. *Journal of Cleaner Production*, *327*, 129514. https://doi.org/https://doi.org/10.1016/j.jclepro.2021.129514

Hernandez-Betancur, J. D., Ruiz-Mercado, G. J., Abraham, J. P., Martin, M., Ingwersen, W. W., & Smith, R. L. (2021). Data engineering for tracking chemicals and releases at industrial end-of-life activities. *Journal of Hazardous Materials*, *405*, 124270. https://doi.org/https://doi.org/10.1016/j.jhazmat.2020.124270

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2022). A data engineering approach for sustainable chemical end-of-life management. *Resources, Conservation and Recycling*, *178*, 106040. https://doi.org/https://doi.org/10.1016/j.resconrec.2021.106040

Hernandez-Betancur, J. D., Ruiz-Mercado, G. J., & Martin, M. (2022a). A high-throughput screening of chemical end-of-life transfer scenarios using structure-based classification models (to be submited).

Hernandez-Betancur, J. D., Ruiz-Mercado, G. J., & Martin, M. (2022b). Tracking off-site end-of-life stage of chemicals: A scalable data-centric and chemical-centric approach (to be submited).

CONTENTS

I Understanding the chemical end-of-life management chain and recycling loop

II Towards a cross-country framework

III A step to high-throughput screening of end-of-life exposure scenarios

IV Appendix

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 CHEMICAL RISK EVALUATION ISSUES OVERVIEW

In the 21st century, the chemical industry is one of the largest manufacturing sectors in the world (Morris, 2003). This industry sector is the largest manufacturer in the United States and the second largest one in Europe after food and kindred products (Murmann, 2002). The overall outlook for the chemical industry has been one of growth in recent years. The global sales of chemicals increased from USD 1,220 billion in 2009 to USD 3,820 billion in 2019 (European Chemical Industry Council, 2022). Although the difficulties caused by the COVID-19 outbreak in 2020 caused a reduction in the production of the chemical sector, a growth in demand and chemical production are expected to increase in the following years (Atradius, 2021).

Chemicals are present in products supplied by industry sectors like wholesale, electricity, petroleum, and natural gas and products bought by sectors like health care, plastics, textiles, paper, and rubber (Morris, 2003). Hence, chemicals are everywhere and constitute an important aspect of the daily lives of people around the world (The Organization for Economic Co-operation and Development, 2021). Altough they have many benefits for humankind's development, anthropogenic chemical pollution has the potential to pose one of the largest environmental threats to humanity (Naidu et al., 2021) and some chemicals may present unreasonably high risk of injury if they are not properly managed (Barr et al., 2006).

Chemical risk evaluation has been a widely used tool for the selection and informed use of safer chemicals, materials, and technologies (Whittaker, 2015), driven in recent years by different aspects such as the development of green chemistry and green engineering, life cycle thinking, and public opinion (National Research Council, 2014). Likewise, chemical risk assessment has been a tool for regulatory support to promote the use of safer chemicals and to determine whether they may present a significant risk to human health and the environment throughout their life cycle in the market (Graham, 2008).

In the United States, the Toxic Substances Control Act (TSCA) has given the Environmental Protection Agency (EPA) the legal means to evaluate the risk that a chemical substance may have during its manufacture, processing (including recycling), use in industrial, commercial, and consumer activities, and its end-of-life (EoL) management in the United States (U.S. Environmental Protection Agency, 2017b). Similarly, in Europe, the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) Regulation has empowered the European Chemicals Agency (ECHA) to determine whether chemical risk is under control during manufacture, formulation, or repackaging, final application, and service life (article) or, on the contrary, whether it is necessary to determine risk management measures throughout the territory of the Member States (European Chemicals Agency, 2019a).

Although performing a chemical risk evaluation is necessary, it is still a time-consuming and challenging task, as information needs to be collected about the potential activities or emission sources of the chemicals (e.g., as a reactant), exposure pathways (e.g., water), exposure routes (e.g., inhalation), susceptible receptors (e.g., workers), hazards (e.g., lung cancer), and physicochemical properties (e.g., ethanol/water partition coefficient) needed to perform the engineering analyses and calculations that support such an assessment (see Figure 1.1). The above is more daunting from the regulatory agencies' point of view, as about 80,000 chemicals are regulated by TSCA, of which more than 40,000 substances are active in the U.S. market (U.S. Environmental Protection Agency, 2017a), while more than 106,213 substances are regulated by REACH across the Member States of the European Union and European Economic Area countries (European Chemicals Agency, 2019b). In addition, the EoL activities, i.e. disposal, treatment, energy recovery, and recycling, are a major concern in the framework of chemical risk assessment due to the difficulty in performing a complete traceability and analysis of the material streams of a chemical present in industrial, commercial, and consumer waste streams. The above leads to not anticipate additional exposure pathways that may cause considerable risk to human health and the environment (Ragas, 2011).

Chemical risk, in general, can be separated into its three constituent components, which are exposure, hazard, and vulnerability (Jacob, 2015). Vulnerability is the propensity or predisposition to be adversely affected. A hazard is the potential occurrence of an event or trend or impact that may cause loss of life, injury, or other health impacts, as well as damage and loss to ecosystems and environmental resources. Exposure is the presence of people, livelihoods, species or ecosystems, environmental functions, services, and resources, infrastructure, or economic, social, or cultural assets in places and settings that could be adversely affected (Bojariu et al., 2015). Thus, chemical risk evaluation can be developed by the outcomes of different studies like hazard characterization, hazard identification, risk characterization, and exposure assessment (Wittwehr et al., 2020). An exposure assessment is to assess the exposure or amount of intake of

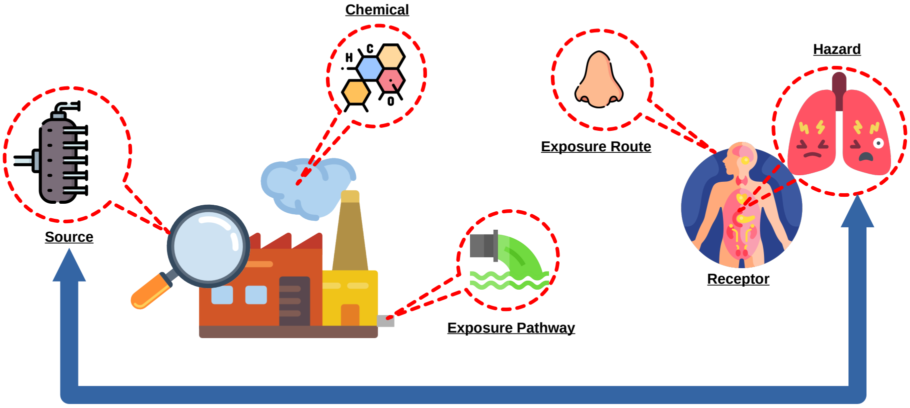


Figure 1.1: An overview to the elements considered during the chemical risk evaluation (developed using Flaticon).

chemicals for humans or organisms in the environment (Whaley et al., 2016). Accounting and tracking of chemical flows are the first steps that must be performed in exposure evaluation (Meyer et al., 2019).

The well-known environmental life cycle assessment (LCA) is somewhat related to chemical risk assessment (Flemström et al., 2005). LCA is a methodology that can be used to evaluate the environmental load of a product, process, or activity throughout its life cycle (P. Roy et al., 2009). While risk assessment focuses on managing the hazard in an exposure, LCA seeks a holistic estimation of the impacts of substances across multiple media (Linkov et al., 2017). Hence, the above two approaches, that dominate environmental policy and decision-making, can be integrated to work together. The potential interoperatibility between the above assessments is because they rely on the creation of life cycle inventories (LCI) tracking the material flows (and energy) throughout product or service life cycle to evaluate its impacts (Guinée et al., 2011). LCI not only provides information about products and services, but it also provides a detailed accounting of environmental releases and flows (Meyer et al., 2020).

A methodology to collect LCI is to perform a material flow analysis (Muller et al., 2011). Indeed, material flow analysis can help to map EoL flow movements and EoL activities (Allesch & Brunner, 2015) and has been used by practitioners and researches to get chemical flow inventories and perform chemical flow analysis (CFA) for a wide range of chemicals (Bornhöft et al., 2013; Gottschalk et al., 2010; van Gils et al., 2020). Some researchers have explored the use of data-driven modelling for high-throughput CFA, chemical exposure assessment, and LCI (Cha et al., 2021; Franzosa et al., 2021; Huang et al., 2021; Meyer et al., 2019; Ring et al., 2019). Thus, data-driven modeling shows promising potential as a tool to accelerate the collection of information to identify exposure scenarios and the traceability of chemical streams. Nonetheless, the development of models and their performance depend on the specific dataset domain (Priatama et al., 2022).

Considering the above, this work aims to develop a methodology and tools based on data, mainly publicly-available, to track the flow of chemicals contained in industrial EoL streams, identify potential EoL generic scenarios, understand the EoL management chain, and perform a rapid CFA. In order to provide a continuous flow of data for the construction of data-driven models, the work seeks to build a data pipeline that allows training and retraining the models periodically for the prediction of potential EoL exposure scenarios for chemical substances, whether they do not belong to the primary data sources or are new to the market. Additionally, the work seeks to propose data models that are suitable for mimicking the EoL supply chain behavior or providing information on the elements that constitute it.

1.2 DATA ENGINEERING ROLE

As mentioned above, both chemical risk evaluation and exposure assessment require comprehensive LCI data in order to identify potential EoL generic exposure scenarios and potential environmental releases. This data is usually stored in siloed data sources in different formats and is comprised of different data types. Due to the above, in recent years, LCI practitioners have been concerned about the development of automatic systems for mining, extracting, and transforming scatter data for using them in different environmental applications (Birney et al., 2022; Li et al., 2022; Young et al., 2022). In addition, data-driven models depends on a vast amount of data for good prediction performance and to capture as many values for input variables as possible (Roh et al., 2021). Thus, a field known as data engineering plays an important role in big data, analytics, and machine learning, in short, in data-driven modeling (Luściński, 2015). Data engineering is the design and building of pipelines that transform and transport data into a format that can be used by data scientists, machine learning engineers, and other users. Hence, by using data engineering, the data is in a highly usable state (Gray & Shenoy, 2000).

The growth and availability of data and the development of more complex state-of-the-art machine learning algorithms have led to the development of machine learning operations (MLOps) that enable collaboration and increase the pace of delivery and quality of model development through monitoring, validation, and governance of machine learning mod-

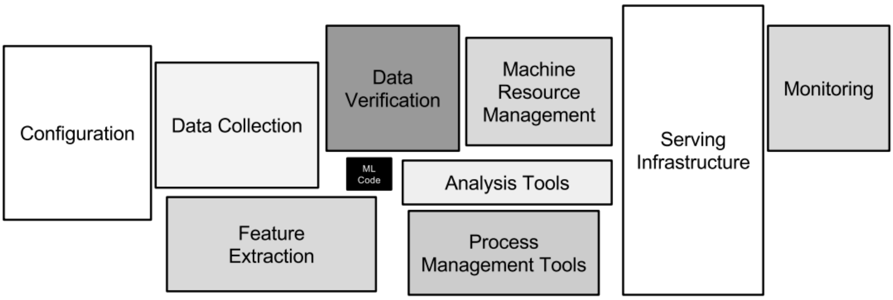


Figure 1.2: Overview of machine learning operations (MLOps) (taken from Sculley et al. (2015))

els (Soh & Singh, 2020). Figure 1.2 depicts an overview of the different elements that constitute MLOps. As shown in the figure, data collection has a more important role in real-world machine learning model development (Sculley et al., 2015). In fact, data collection is the main task of data engineering and greater attention, time, and energy should be paid to this task to guarantee the development of data-driven models able to predict generic EoL exposure scenarios, understand the chemical EoL management chain, and estimate potential environmental releases from EoL activities. In this way, this thesis proposes a data-centric paradigm, i.e., the data is iteratively improved while the model is held fixed (Ng, 2021).

1.3 DATA-DRIVEN MODELLING

Two main modelling paradigms exist. The first one is a mechanistic approach that incorporates the available knowledge of the system into the model. In contrast, data-driven techniques search for relationships between input and output variables, using the available data and without worrying about the underlying process (Ji et al., 2012; T. Zhou et al., 2021). In fact, the best of the approaches have been combined to create hybrid models, i.e., models incorporating both mechanistic and data-driven techniques (Kurz et al., 2022). Data-driven models are built up using statistical or machine learning techniques (Solomatine & Ostfeld, 2008). Figure 1.3 depicts a Venn diagram for the main goals of statistical modelling and machine learning methods and their intersection. A statistical model is the use of statistics to build a representation of the data and then conduct analysis to infer any relationship between variables or discover insights. In contrast, machine learning is the use of mathematical and/or statistical models to obtain a general understanding of the data in order to make predictions (Goldstein et al., 2016; Mendis, 2019). Although both techniques are developed under different ideas and objectives, they share analytical goals like accuracy and precision, potential implementation for future applications, and understanding of data (see Figure 1.3).

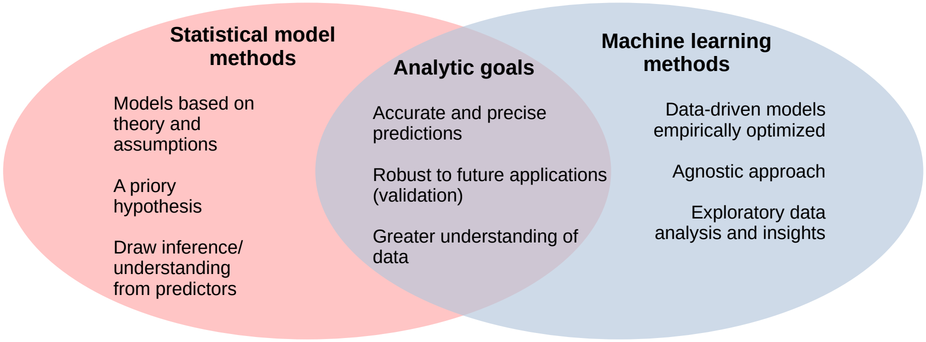


Figure 1.3: One perspective on the intersection of statistical modelling (red) and machine-learning (blue) goals (modified from Goldstein et al. (2016)).

1.3.1 *Probabilistic graphical model*

The EoL supply and management chain for a chemical can be considered a complex system. Likewise, the decision-making process for selecting a pollution abatement unit technology or EoL activity suitable for handling a hazardous chemical or waste can be considered as a complex system. Complex systems are ones with a large effective number of strongly-interdependent aspects. This excludes both low-dimensional systems, and high-dimensional ones where the aspects are either independent, or so strongly coupled that only a few aspects effectively determine all the rest (Shalizi, 2009). The interrelated aspects can be related to the reasoning taks (Koller & Friedman, 2009). For instance, in the chemical EoL management chain, there are multiple possible EoL activities to handle a chemical, environmental regulations that can favour the implementation of an EoL activity, economic aspects such as the chemical price, and many more matters to consider. These domains can be characterized in terms of a set of random variables, where the value of each variable defines an important property of the world (Koller & Friedman, 2009).

To reason probabilistically about the values of one or more of the variables, it is necessary to construct a joint distribution over the space of possible assignments to some set of random variables, thus obtaining tools to answer a wide range of interesting queries. However, constructing a joint distribution for a complex system is quite daunting. Nevertheless, a family of statistical models called probabilistic graphical models (also known as graphical models or structured probabilistic models) provides

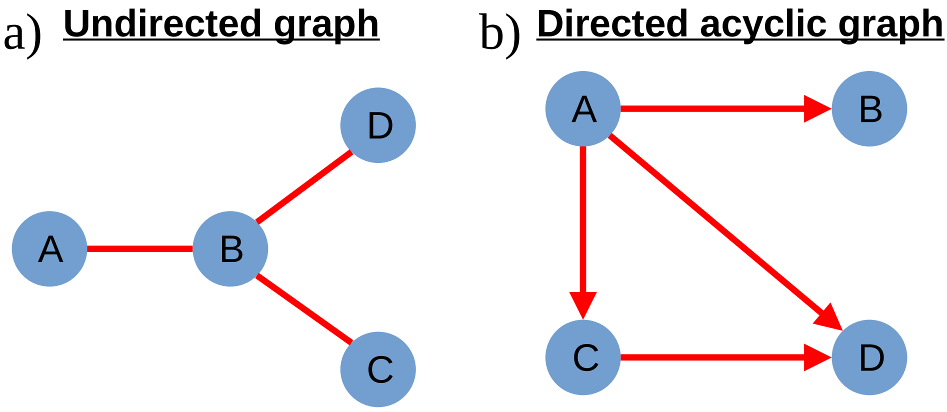


Figure 1.4: Representation of statistical graphical models. (a) an undirected graph representation. (b) a directed acyclic graph representation.

a methodology for exploiting the structure of complex distributions to describe them in a compact form so that they can be constructed and used effectively. Probabilistic graphical models use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space (Koller & Friedman, 2009). In graph theory, a graph is a structure of nodes or vertices connected by edges representing direct interactions (Slutsky, 2014).Graphical models can be largely divided into two categories named directed acyclic graphs and undirected graphs (Goodfellow et al., 2016).

1.3.1.1 *Directed acyclic graph*

Directed acyclic graph also known as belief network or Bayesian network (Pearl, 1988). This models are called directed because their edges are directed, that means, they point from one vertex to another. As shown in Figure 1.4b, an arrow represents the direction. From a statistical point of view, the arrow direction indicates which variable's proability distribution is defined in terms of the others (Goodfellow et al., 2016). For instance, the directed acyclic graph in Figure 1.4b indicates that the statistical distribution of the node D depends directly on C and A. Formally, a directed graphical model defined on variables $\mathbf{x}$ is defined by a directed acyclic graph Θ whose vertices are the random variables in the model, and a set of local conditional probability distributions $p(x_i|Pa_\Theta(x_i))$, where $Pa_\Theta(x_i)$ gives the parents of x_i in Θ. Hence, the probability distribution over $\mathbf{x}$ is given in Equation 1.1 (Goodfellow et al., 2016).

$$p(\mathbf{x}) = \prod_{i=1} p(x_i|Pa_\Theta(x_i)) \tag{1.1}$$

1.3.1.2 *Undirected graph*

Undirect graphical models also known as Markov random fields (MRFs) or Markov networks, is the other class of statistical graphical models whose edges are undirected (Kindermann & Snell, 1980). Directed models are applicable to situations where the causality between the random variables is clearly understood and the causality flows in only one direction. However, the above described situation is rare. When the interactions seem to have no intrinsic direction, or to operate in both directions, it may be more appropriate to use an undirected model (Koller & Friedman, 2009). Figure 1.4a depicts a simple representation for an undirected graph. Unlike directed models, the edge in an undirected model has no arrow and is not associated with a conditional probability distribution (Goodfellow et al., 2016). Formally, an undirected graphical model is a structured probabilistic model defined on an undirect graph Θ. For each clique Λ (a subset of vertices), a factor $\phi(\Lambda)$ (also called a clique potential) measures the affinity of the variables in that clique Λ for being in each of their possible joint states. Like the probabilities, the factors are constrained to be nonnegative. Together they defined an unnormalized probability distribution, as shown in Equation 1.2.

$$\tilde{p}(\mathbf{x}) = \prod_{\Lambda \in \Theta} \phi(\Lambda) \tag{1.2}$$

1.3.2 *Machine learning models*

As shown in Figure 1.3, machine learning models are a good data-driven modelling alternative to be empirically optimized based on available data and their agnosticism allows to apply them in a wide variety of knowledge fields like exposure assessment (Isaacs et al., 2016; Ring et al., 2019). In general, machine learning models look for resolving the problem of the probability of the output variable Y given the input variables X, i.e., the posterior probability or $P(Y|X)$ (Hastie et al., 2009). Based on how the models approach to the posterior probability or $P(Y|X)$, the machine learning models can be classified into two categories known as discriminative and generative models (Goyal, 2021).

$$posterior = \frac{prior \times likelihood}{evidence} \rightarrow P(Y|X) = \frac{P(Y) \times P(X|Y)}{P(X)} \tag{1.3}$$

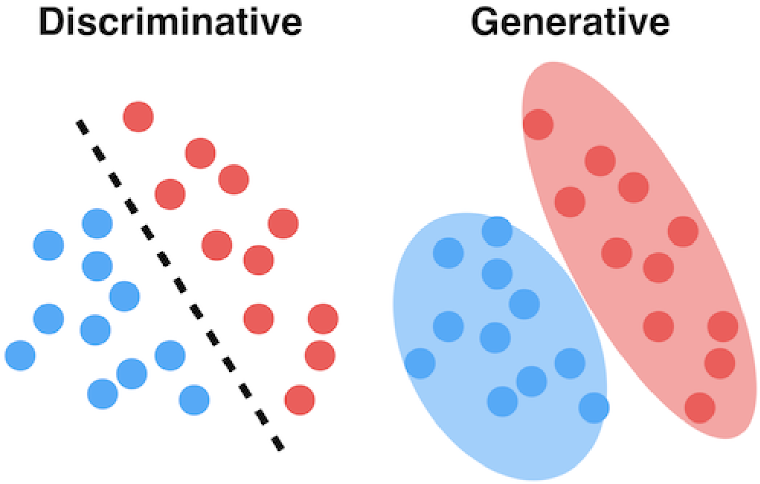


Figure 1.5: Difference between discriminative and generative machine learning models (taken from Goyal (2021)).

1.3.2.1 *Discriminative machine learning models*

Discriminative models (also known as conditional models) make predictions based on new data by using conditional probability. These kinds of models can be used either for classification or regression problems. As shown in Figure 1.5, discriminative models separate classes by learning the boundaries between the classes or labels. Instead of making any assumptions about the data points, discriminative models assume a functional form for the posterior probability or $P(Y|X)$ (see Equation 1.3). Using the available data, they estimate the parameters of $P(Y|X)$ and find the maximum likelihood estimator (Hastie et al., 2009). Some examples of discriminative models available in the literature are linear regression (Kumari & Yadav, 2018), logistic regression (Peng et al., 2002), support vector machines (Evgeniou & Pontil, 2001), traditional neural networks (Grossi & Buscema, 2008), k-nearest neighbor (Cunningham & Delany, 2007), conditional random fields (Sutton & McCallum, 2010), decision trees (Rokach & Maimon, 2005), and random forests (Breiman, 2001; Goyal, 2021). This kind of algorithms are good for tasks with labeled data to try to predict classes or values, i.e., classification and regression what is well-known as supervised learning (Dobbelaere et al., 2021).

RANDOM FOREST: A discriminative machine learning model used in this work is random forest. Random forest has been widely used in applications of chemoinformatics because of its high-performance (Polishchuk et al., 2009; Svetnik et al., 2003). Some advantages of random forest is that it can perform both regression and classification tasks, produces good predictions that can be understood easily, can handle large datasets efficiently, and provides a higher level of accuracy in predicting outcomes over the decision tree algorithm (J. Ali et al., 2012). Figure 1.6 shows a simplified view of a random forest classifier. As shown in the figure, a random forest is composed of a set of tree-based estimators or decision trees. Thus, the unit blocks for a random forest are decision trees, which identify key differentiating factors between classes in a dataset. Formulating yes-no questions, the decision trees form "yes" and "no" paths that will either lead to the next question or to a final output or prediction (Rokach & Maimon, 2005).

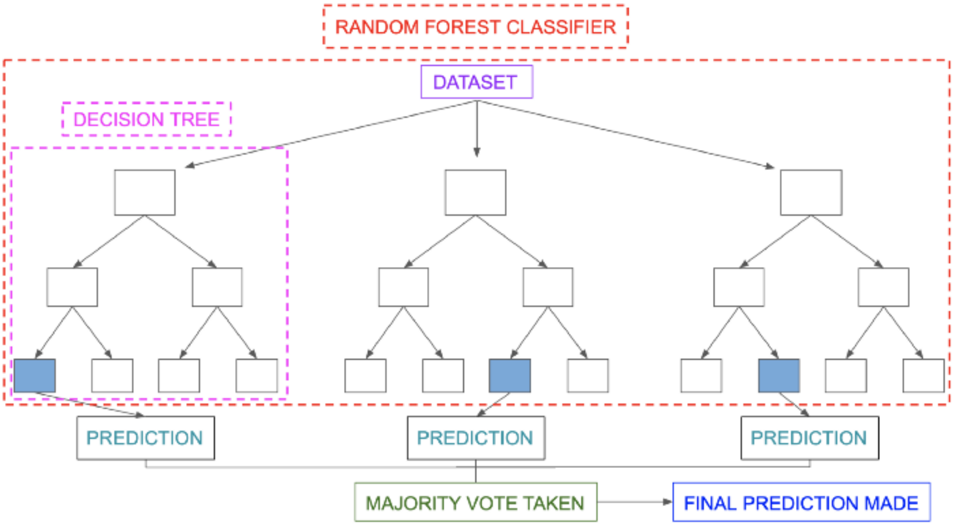


Figure 1.6: A simplified view of a random forest classifier (taken from Kashyap (2019)).

The key terms that are important to understand the development of decision trees are the following (Kashyap, 2019):

1. Classes: the list of different groups to which the data can belong is known as the classes.
2. Nodes: the different questions that make up the decision tree are known as nodes. These can be seen as points in the decision tree where we form a split (or branches). Each of these is a "yes" or "no" question, and once answered, we get one step closer to identifying the class to which the data belongs.
3. Leaves: due to the presence of nodes, there are many alternative paths that are created in any decision tree with even a few layers

> of nodes. The end point of each of these paths is known as a "leaf". These leaves represent the final value or class that is predicted for the given input data.

For discrimination and finding the decision boundary, decision trees use the concepts of information entropy and information gain (Kashyap, 2019). Entropy (E) can be understood as a measure of the homogeneity of a sample from a group of data. Typically, entropy is calculated on a scale of 0 to 1. An entropy of 0 indicates that the data has minimal disorder (very homogeneous/pure), while an entropy of 1 (or a high entropy value) indicates that there is a lot of disorder in the data (Kashyap, 2019). Equation 1.4 presents how to calculate the entropy for a dataset with n classes. In Equation 1.4, p_i is the probability of a data point in the dataset belonging to the class i. In contrast, information gain (IG) ensures that the entropy is reduced by splitting the data with the questions in the decision tree. In the decision tree, the data is split based on the features, thereby creating a subset of the data. Thus, the splitting operation in a decision tree has an entropy associated with it that can be compared with the entropy before the splitting. Tha comparison is what is called the information gain for that split and is calculated as presented in Equation 1.5. In equation 1.5, X represents the target variable, A the attribute on the basis of which this split has been formed, $E(X)$ the entropy of the data at the node before the split, and $E(X, A)$ the weighted sum of the entropies of the two branches formed after the split based on the attribute A. In summary, in a decision tree the objetive is to find a model that minimizes the entropy and maximizes the information gain (Kashyap, 2019).

$$E = \sum_{i=1}^{n} -p_i \times log_2(p_i) \tag{1.4}$$

$$IG(X, A) = E(X) - E(X, A) \tag{1.5}$$

As mentioned before, decision trees are the unit blocks that form a random forest. This collection of decision trees is known as an ensemble, i.e., a random forest is an ensemble method that uses multiple tree-based learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone (Dižo et al., 2022). In a random forest is decision tree is unrelated each other and is traing using a random subset of the training data. This technique for training each tree in different random sample of the data is known as bagging or boostrap aggregation (Altman & Krzywinski, 2017). As shown in Figure 1.6, each individual tree makes a prediction, and the final prediction is then decided by measuring which prediction was made by the greatest number of trees in the forest. In addition to bagging, random forest uses another technique called randomizing attributes or features so that each decision tree does not have access to all of the dataset features. Instead, a random subset of features is given to each decision tree in the random forest (Scornet et al., 2014).

1.3.2.2 *Generative machine learning models*

A generative model focuses on the distribution of a dataset to return a probability for a given example. These kinds of models can be used for tasks like data augmentation (or creation of sintetic data samples), clustering, and dimensionality reduction. As shown in Figure 1.5, generative models focus on understanding the statistical distribution of individual classes in a dataset and model the underlying patterns of data points. To find the posterior probability or $P(Y|X)$, generative models use the available data to estimate prior probability or $P(Y)$ and likelihood probability or $P(Y|X)$, which are the same that the joint probability distribution considering that $P(X, Y) = P(Y) \times P(Y|X)$. Then, based on new data (evidence or $P(X)$), they calculate $P(Y|X)$ (see Equation 1.3). Some examples of generative models available in the literature are Naïve Bayes (Webb et al., 2011), autoregressive models (Dalal et al., 2019), principal component analysis (Mishra et al., 2017), latent dirichlet allocation (Blei et al., 2003), and generative adversarial networks (Goodfellow et al., 2014; Goyal, 2021). From a machine learning and application standpoint, even statistical graphical models such as Bayesian networks and Markov random fields can be classified as generative models (Goodfellow et al., 2016). Unlike discriminative models, this kind of algorithms can be used for clustering, visualization, anomaly or outlier detection, and dimensionality reduction, i.e., for unsupervised learning where the records are unlabeled and the idea is to discover patterns (Dobbelaere et al., 2021).In this thesis, generative models are used for dimensionality reduction, outliers detection, and dataset balancing, as presented in the following paragraphs.

DIMENSIONALITY REDUCTION: As mentioned before, generative models are widely used to find patterns in data and they are good for dimensionality reduction tasks. The idea of dimensionality reduction is to reduce the risk of overfitting problems due to high-dimensionality (Johnstone & Titterington, 2009; Zong et al., 2020). There are several techniques available in the literature to reduce the dimensionality and they can be classified according to their taxonomy and functionality (van der Maaten et al., 2009). A group of dimensionality reduction techniques is based on components or factors analysis. Principal component analysis (PCA) is component-based technique widely used by machine learning practitioners and statisticians in order to reduce the dimensions of large datasets. PCA is a linear method for reducing the dimension of a dataset by using the covariance matrix for features and then getting the eigenvalues and eigenvectors (Mishra et al., 2017). PCA arrives at the final components, which explain most of the dataset variability. Therefore, these new components can be thought of as linear combinations or composites of the original features (Mahmood, 2021). As PCA works well on continuous data, if a dataset is composed of nominal categorical data, a component-based method called multiple correspondence analysis (MCA) is used to detect and represent the dataset's underlying structures and reduce the dimensionality (Khangar & Kamalja, 2017). MCA uses an indicator matrix or a complete disjunctive table (see Equation 1.6). An indicator matrix is a matrix where the rows represent samples and the columns are dummy variables representing categories of the variables (Heiser, 2009). Unfortunately, in the real world, data is a blend of both numerical and categorical data (i.e., mixed dataset). In the above case, a generative technique used for dimensionality reduction is factor analysis of mixed data (FAMD). FAMD does the analysis with a combination of PCA and MCA techniques. Practitioners usually use PCA using encoding techniques over categorical variables. Unlike PCA, the FAMD technique does not require feature encoding to be performed (Visbal-Cadavid et al., 2020).

$$\begin{pmatrix} a & b & c \\ a & d & e \\ a & b & e \\ b & c & e \end{pmatrix} = \left(\begin{array}{ccccc} a & b & c & d & e \\ \hline 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{array}\right) \tag{1.6}$$

OUTLIERS DETECTION: Outliers can cause learning problems for the discriminative machine learning models so that they may misclassify and poorly predict (Verma & Hansch, 2005). As mentioned above, generative machine learning models are also good at detecting outliers in a dataset. Outlier detection techniques may have multiple kinds of categorizations (Smiti, 2020), for instance, distance-based and tree-based methods. Distance-based outlier detection techniques consult the neighbourhood of a data point, which is defined by a given radius (distance threshold). A data point is then considered an outlier if its neighborhood does not have enough other points (Chepenko, 2018). Distance-based outlier detection techniques can perform well for scattered real-world data. However, distance-based outlier detectors may be computationally expensive in large datasets (K. Zhang et al., 2009). Due to the above, tree-based methods are atractive alternatives because of the required time to obtain results. Isolation forest, an unsupervised anomaly detection technique, is a tree-based method widely used by machine learning practitioners for detecting and removing outliers in a large dataset (Liu et al., 2008). Like random forests, isolation forests are built based on decision trees, but unlike random forests, there are no predefined labels. Isolation forest assums that outliers are the data points that are "few and different". In an isolated forest, randomly subsampled data (i.e., boostrap aggregation) is processed into a tree structure based on randomly selected features (i.e., randomizing attributes or features). Samples that go further into the tree are less likely to be anomalies, as they require more cuts to isolate. Similarly, samples that end in shorter branches indicate anomalies, as it was easier for the tree to separate them from other observations (S. Unrau, 2021), see Figure 1.8a. Thus, isolation forest learns patters from the available data to determine the dataset distribution and isolate those data points that are not normal, as shown in Figure 1.8b.

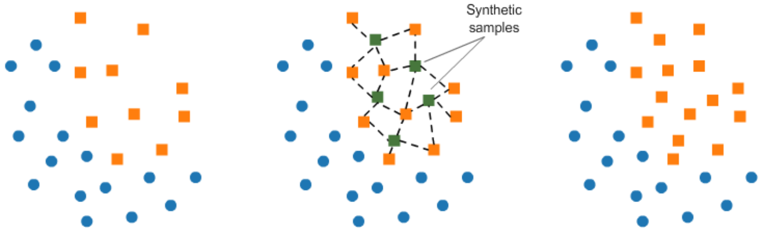


Figure 1.7: Synthetic Sample generation using synthetic minority over-sampling technique (SMOTE) (taken from Das (2019)).

DATASET BALANCING: Supervised classification tasks addressed by discriminative machine learning models can be affected by an imbalanced dataset (Jeatrakul et al., 2010). An imbalanced dataset is one where the number of samples for target classes is not equal, thereby leading to potential type I error (i.e., false positive) and type II error (i.e., false negative) (H. Ali et al., 2019). For dealing with imbalanced datasets, generative machine learning models can support data processing before being fed into discriminative models for classification. Imbalanced dataset methodologies available in the literature can be divided into two categories: undersampling and oversampling. Oversampling, also known as data augmentation techniques, increases the number of samples in the minority class(es), whereas undersampling decreases the number of samples in the majority class(es) (Islam et al., 2022). A method widely used as an undersampling technique is synthetic minority over-sampling technique (SMOTE)

(Blagus & Lusa, 2013). In case of multi-class or binary classification problems, SMOTE creates synthetic data points for the minority class so that it matches up to the majority class (Chawla et al., 2002). SMOTE works by using a k-nearest neighbor algorithm to create synthetic data. SMOTE algorithm can be summarized in the following steps (C. Y. Unrau, 2020):

1. It starts by choosing random data samples from the minority class.
2. It sets the k-nearest neighbors of the data samples.
3. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor.
4. It repeats the process until data is balanced (see Figure 1.7).

Another technique used in multi-class and binary classification problems is NearMiss (J. Zhang & Mani, 2003). Unlike SMOTE, NearMiss is an undersampling technique for dealing with imbalanced datasets. There are three main steps that described the underlying idea of NearMiss (Madhukar, 2020):

1. It calculates the distance between all the data points in the majority class with the data points in the minority class.
2. It selects data points of the majority class that have the shortest distance with the minority class. The n samples need to be stored for elimination.
3. If there are m samples of the minority class then the algorithm will return $m \times n$ samples of the minority class.
4. It repeats the process until data is balanced.

Multi-label classification problems are the ones where a target variable can have more than one possible label or class associated. This is the case when classes are not mutually exclusive. For example, in a photograph, there can appear a person and a car at the same time, so in an image classification problem, both classes are valid for an image (M.-L. Zhang & Zhou, 2014). Another example of multi-label classification problems is that for a chemical over a period of one year, many EoL activities can occur. For this kind of problems, methods like Near-Miss and SMOTE are not applicable. Charte et al. (2015) proposes multilabel synthetic minority over-sampling technique (MLSMOTE), an extension or variant of SMOTE, for producing synthetic data points for imblance multi-lable datasets. The following are the main steps for performing MLSMOTE (Sukhwani, 2020):

1. It selects data to over-sample (i.e., minority class).
2. It chooses an data samples of the data.
3. It finds its k-nearest neighbours of that data points.
4. It chooses a random data point which is in k-nearest neighbours of the selected data point and make a synthetic data point anywhere on the line joining both these points.
5. It repeats the process until data is balanced.

1.3.3 *Quantitative structure-activity relationship*

Quantitative structure-activity/property relationship (QSAR/QSPR) modeling has long been used in medicinal chemistry and computational toxicology. It provides an in silico tool for the development of predictive models towards various activity points and properties of a range of chemicals, using experimentally determined response data and computationally or experimentally derived molecular structure information. Once developed and validated, these models can be used to determine response endpoints of novel and untested chemicals, as well as to obtain a mechanistic interpretation of structure-activity-property relationships. These techniques have been successful in many runway optimization and risk assessment problems. (K. Roy et al., 2015).

1.3.3.1 *QSAR modelling workflow*

As data-driven and predictive models, QSAR development requires a workflow in order to orchestrate the needed steps for running experiments. Figure 1.9 depicts an overview of the QSAR modelling workflow. As shown in the figure, dataset compilation, which involves the retrieval of data from different data sources, is the first QSAR modelling step, which, as mentioned before, is addressed by data engineering pipelines. After collecting data, QSAR needs data processing in order to curate, select features, balance the data, and split the data. Data curation and preprocessing are needed for dropping duplicate records and dealing with null-value either by removing input features/variables or imputing them based on, for example, a central tendency measurement like the mean or median (Lee & Stvilia, 2017).

As presented in Figure 1.9, after data curation and preprocessing, chemical descriptors have to be calculated. Chemical structure information is noramally encoded by using the Simplified Molecular Input Line Entry

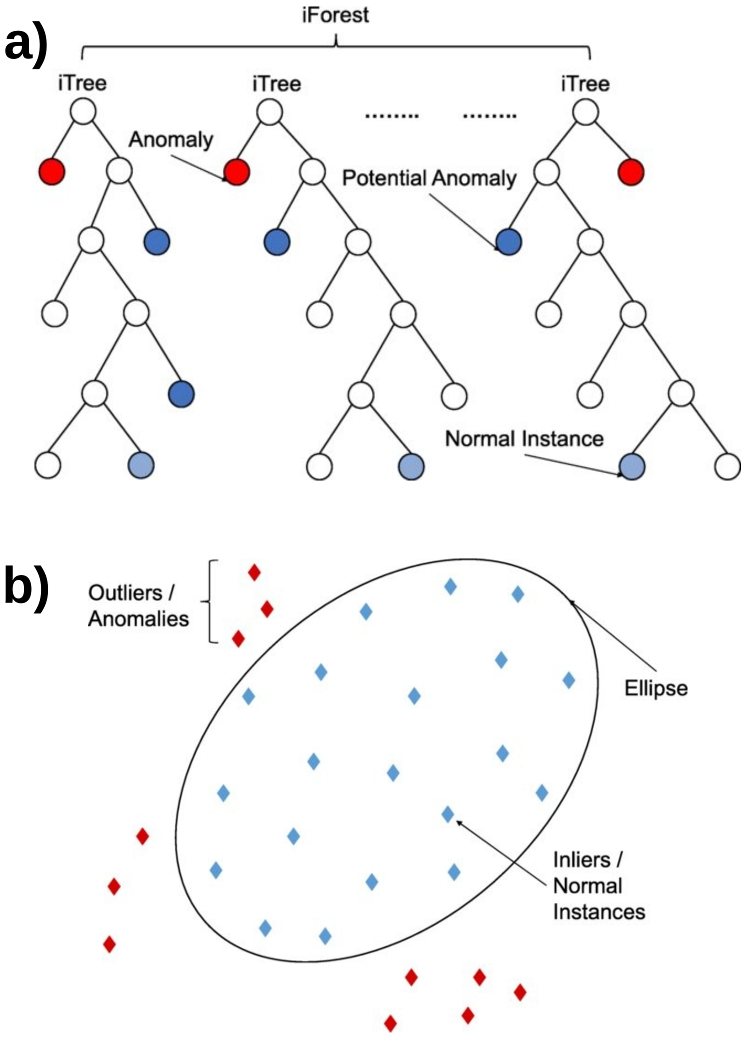


Figure 1.8: Isolation forest underlying idea. (a) decision-making for isolation forest. (b) pattern learn by isolation forest (taken from Regaya et al. (2021)).

System (SMILES), which is a chemical notation widely utilized in chemical information processing (Weininger, 1988). Once chemical descriptors are obtained, feature selection is applied in order to reduce the infrastructure need to deploy QSAR models, reduce redundant features, and mitigate a potential risk of QSAR model over-fitting (Miao & Niu, 2016). If the predictive problem is for classification, the data should be balanced to avoid classification mistakes. Thus, by balancing the data, an equal number of each label is obtained. (Kotsiantis et al., 2005). The final step in dataset processing is splitting the data to divide the data between the train and

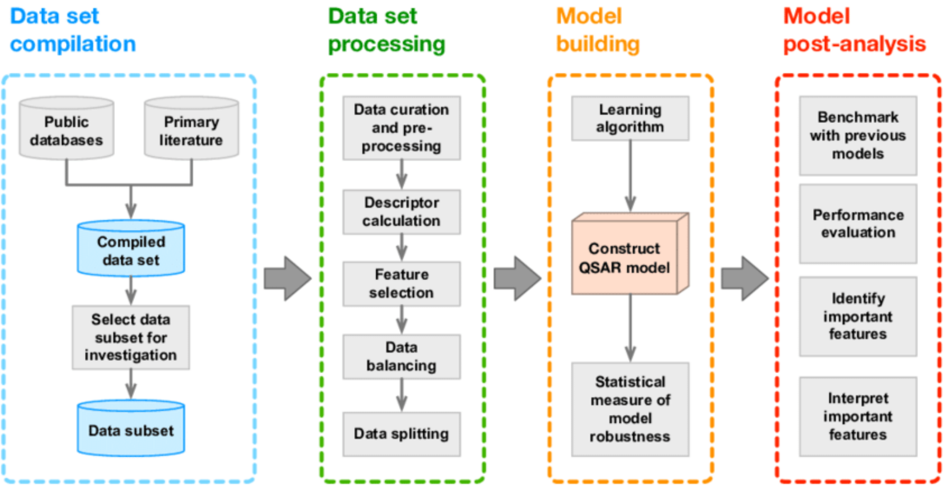


Figure 1.9: Schematic representation of the QSAR modelling workflow (taken from Nantasenamat (2020)).

test datasets. The latter dataset is for external validation and the statistical measure of model robustness (Nantasenamat, 2020).

At the model building step, state-of-the-art machine learning algorithms are selected for running experiments and construct QSAR models (Sarker, 2021). The models used in this part correspond to the discriminative machine learning algorithms and the selection of any model depends on the modelling problem complexity. On the first hand, for regression problems, practitioners commonly develop QSARs using models like multiple linear regressions (Kaya Uyanık & Güler, 2013), partial least squares (Pirouz, 2006), artificial neural networks (Grossi & Buscema, 2008), and Gaussian processes (Rasmussen et al., 2004). On the second hand, for classification problems, they select models like logistic regressions (Peng et al., 2002), linear discriminant analysis (Tharwat et al., 2017), decision trees (Rokach & Maimon, 2005), random forests (Breiman, 2001), k-nearest neighbors (Cunningham & Delany, 2007), probabilistic neural networks (Specht, 1990), and support vector machines (Evgeniou & Pontil, 2001).

After the model is constructed and its parameters are tuned to perform well, the built model has to be evaluated to test whether it will predict properly on new data. Different metrics can be found in the literature based on whether the model is for classification or regression (J. Zhou et al., 2021). The outcomes from a QSAR or machine learning modelling experiment can be used basically in four ways. If a previous model exists, the new one can be compared to determine if it is better than the older one or be discarded if it is not. The QSAR model performance can be evaluated on unseen data by the model, which is called external validation or evaluation on the test dataset (K. Roy et al., 2015). In addition, if the model allows to analyze the feature importance, the feature importance should be analyzyed and understood from the point of view of the problem application to obtain future model improvements (Nantasenamat, 2020).

1.3.3.2 *QSAR and regulatory applications*

The following aspects are addressed by regulatory bodies in order to streamline the chemical risk evaluation via QSAR modelling (K. Roy et al., 2015):

1. Assessment of exposure.
2. Identification and doseresponse characterization of hazard, including classification and labeling of the chemicals.
3. Assessment of hazard and exposure.
4. Identification of persistent, bioaccumulative, and toxic (PBT) as well as very persistent and very bioaccumulative (vPvB) chemicals.

International regulatory bodies and agencies like the Office of Toxic Substances of the EPA, the European Centre for the Validation of Alternative Methods (ECVAM) of the European Union, and the Agency for Toxic Substances and Disease Registry (ATSDR) and the Council for International Organizations of Medical Science, have fostered the use of QSAR models to perform systemic evaluation of toxicological hazard of existing as well as new chemical and identified QSAR as an alternative method for toxicity testing of animals (K. Roy et al., 2015). In fact, the organization of economic cooperation and development (OECD) has promoted a work to develop tools based on QSAR for chemical hazards identification as the case of the QSAR Toolbox which is a free software application that supports reproducible and transparent chemical hazard assessment (The Organization for Economic Co-operation and Development, 2006).

1.3.3.3 *OECD principals for the validation of QSAR models for regulatory purposes*

Based on the agreement reached by OECD member countries in November 2004 at the 37th Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides, and Biotechnology. The following principles should be associated with a QSAR model for regulatory purposes (The Organization for Economic Co-operation and Development, 2004):

1. A defined endpoint.
2. An unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of–fit, robustness and predictivity
5. A mechanistic interpretation, if possible

1.4 MULTI-CRITERIA DECISION-MAKING

Leveraging the available data, data-driven modelling can be combined with multi-criteria decision-making (MCDM). In fact, researchers have developed holistic frameworks that take advatange of both data-driven modelling and MCDM (Kartal et al., 2016). MCDM can be used to select either EoL activities for chemicals transferred to off-site locations or pollution abatement unit technologies to be implemented during an EoL management activity. MCDM problems can be separated regarding whether there are a finite or infinite number of alternatives (Giove et al., 2008). Due to the finite nature of both the potential EoL activity and pollution abatement technology alternatives, the MCDM problem corresponds to a multi-attribute decision analysis.In this work, the MCDM problem is resolved via a fuzzy analytical hierarchy process (FAHP). FAHP descomposes the MCDM problem into simpler sub-problems, makes pairwise comparisons of alternatives, and reflects the uncertainty and ambiguousness of decision makers (Khorramrouz et al., 2019). FAHP depends on fuzzy mathematics and logic to deal with the epistemic uncertainty (Srichetta & Thurachon, 2012), being triangular fuzzy numbers widely used to perform the different fuzzy operations (Sabaghi et al., 2016) and Shannon's information entropy used to find a consensus between the different parties involved in the MCDM problem (Sitorus & Brito-Parada, 2020).

1.5 THESIS OUTLINE

Figure 1.10 presents a summary of the main topics studied in this thesis. This thesis tries to propose a methodology and a strategy to quickly identify possible generic EoL exposure scenarios and to perform CFA, aiming at the rapid development of chemical risk assessment. This work is structured into three parts. Part I is devoted to understand the chemical EoL management chain and recycling loop based on U.S. environmental regulatory databases; Part II looks for ways to get cross-country data for tracking chemical flows; and Part III proposes the development of data-driven models for high-throughput screening of EoL exposure scenarios.

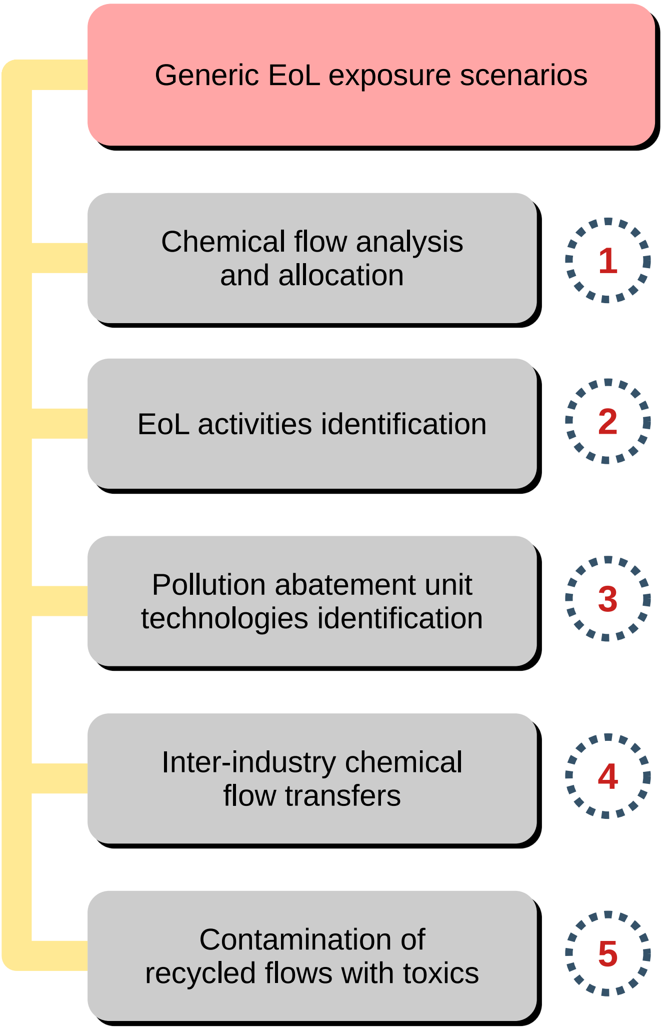


Figure 1.10: Main topics covered in this thesis.

1.5.1 *Part I - Understanding the chemical end-of-life management chain and recycling loop*

CHAPTER 3 - DATA ENGINEERING FOR TRACKING CHEMICALS AND RELEASES AT INDUSTRIAL END-OF-LIFE ACTIVITIES. This chapter explores the use of data engineering for tracking chemical flows transferred to off-site facilities located across the U.S. territories for further EoL management. Moreover, in this chapter, a methodology is proposed to leverage the data for estimating the amount of a transferred chemical that may be released to the environment from EoL activities.

CHAPTER 4 - A DATA ENGINEERING FRAMEWORK FOR ON-SITE END-OF-LIFE INDUSTRIAL OPERATIONS. This chapter focuses on developing data engineering to obtain insights about the pollution abatement unit technologies that a facility may use to handle industrial waste streams containing hazardous chemicals. In addition, an input-output model is proposed to perform CFA and allocate a chemical downstream of a pollution abatement unit technology. A Bayesian network and multi-criteria decision-making are combined to suggest potential pollution abatement unit technology sequences.

CHAPTER 5 - A DATA ENGINEERING APPROACH FOR SUSTAINABLE CHEMICAL END-OF-LIFE MANAGEMENT. This chapter brings together the frameworks proposed in both chapter 3 and chapter 4 to understand the chemical EoL management chain and identify potential post-recycling scenarios that may occur once a chemical is recycled. A Markov random field is suggested to represent the relationship between the different nodes constituting the chemical EoL management chain.

1.5.2 *Part II - Towards a cross-country framework*

CHAPTER ?? - TRACKING OFF-SITE END-OF-LIFE STAGE OF CHEMICALS: A SCALABLE DATA-CENTRIC AND CHEMICAL-CENTRIC APPROACH. This chapter moves forward to harmonize and structure cross-country data to track chemical flow transfers to off-site locations. The framework considers the future building of robust machine learning systems to train and retrain models and deploy them for making predictions on new data. In addition, it analyzes the potential implications and limitations of using the obtained data for building data-driven models.

1.5.3 *Part III - A step to high-throughput screening of end-of-life exposure scenarios*

CHAPTER ?? - A HIGH-THROUGHPUT SCREENING OF CHEMICAL END-OF-LIFE TRANSFER SCENARIOS USING STRUCTURE-BASED CLASSIFICATION MODELS. This chapter gives a step forward in developing data-driven models that allow the rapid identification of potential generic EoL scenarios that could occur for a chemical. Based on QSARs, different modelling strategies are evaluated in order to find a data preparation and modelling pipeline to build robust models for predictions on new data, considering the implications and infrastructure needed to deploy models to be used by stakeholders.

BIBLIOGRAPHY

Ali, H., Salleh, M., Saedudin, R., Hussain, K., & Mushtaq, M. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues(IJCSI)*, *9*.

Allesch, A., & Brunner, P. H. (2015). Material Flow Analysis as a Decision Support Tool for Waste Management: A Literature Review. *Journal of Industrial Ecology*, *19*(5), 753–764. https://doi.org/10.1111/jiec.12354

Altman, N., & Krzywinski, M. (2017). Ensemble methods: Bagging and random forests. *Nature Methods*, *14*(10), 933–934. https://doi.org/10.1038/nmeth.4438

Atradius. (2021). *Global chemicals industry: opportunities and challenges* (tech. rep.). https://group.atradius.com/documents/atradius-industry-trends-global-chemicals-industry.pdf

Barr, D. B., Thomas, K., Curwin, B., Landsittel, D., Raymer, J., Lu, C., Donnelly, K., & Acquavella, J. (2006). Biomonitoring of Exposure in Farmworker Studies. *Environmental Health Perspectives*, *114*(6), 936–942. https://doi.org/10.1289/ehp.8527

Birney, C., Young, B., Li, M., Conner, M., Specht, J., & Ingwersen, W. W. (2022). Flowsa: A python package attributing resource use, waste, emissions, and other flows to industries. *Applied Sciences*, *12*(11). https://doi.org/10.3390/app12115742

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-106

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bojariu, R., Papathoma-Köhle, M., Wendlová, V., & Cica, R.-D. (2015). Changing risks in changing climate.

Bornhöft, N. A., Nowack, B., & Hilty, L. (2013). Material Flow Modelling for Environmental Exposure Assessment - A Critical Review of Four Approaches Using the Comparative Implementation of an Idealized

Example. In S. Verlag (Ed.), *27th international conference on informatics for environmental protection* (pp. 379–388). https://doi.org/https://doi.org/10.5167/uzh-84883

Breiman, L. (2001). *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cha, D., Park, S., Kim, M. S., Kim, T., Hong, S. W., Cho, K. H., & Lee, C. (2021). Prediction of Oxidant Exposures and Micropollutant Abatement during Ozonation Using a Machine Learning Method. *Environmental Science & Technology*, *55*(1), 709–718. https://doi.org/10.1021/acs.est.0c05836

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397. https://doi.org/10.1016/j.knosys.2015.07.019

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chepenko, D. (2018). A Density-based algorithm for outlier detection. https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983#:~:text=Distance%2Dbased%20outlier%20detection%20method,reasonable%20neighbourhood%20of%20the%20object.

Cunningham, P., & Delany, S. (2007). K-nearest neighbour classifiers. *Mult Classif Syst*, *54*. https://doi.org/10.1145/3459665

Dalal, M., Li, A. C., & Taori, R. (2019). Autoregressive models: What are they good for? https://doi.org/10.48550/ARXIV.1910.07737

Das, A. (2019). Oversampling to remove class imbalance using SMOTE. https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

Dižo, J., Blatnický, M., Melnik, R., & https://orcid.org/0000-0003-4677-2535), O. K. (2022). A mathematical model of operation of a semi-trailer tractor powertrain. *Komunikácie*.

Dobbelaere, M. R., Plehiers, P. P., Van de Vijver, R., Stevens, C. V., & Van Geem, K. M. (2021). Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. *Engineering*, *7*(9), 1201–1211. https://doi.org/10.1016/j.eng.2021.03.019

European Chemical Industry Council. (2022). World chemical sales. Retrieved June 13, 2022, from https://cefic.org/a-pillar-of-the-european-economy/facts-and-figures-of-the-european-chemical-industry/profile/%5C#h-china-dominates-chemical-sales-world-ranking

European Chemicals Agency. (2019a). Substance Evaluation. Retrieved June 13, 2022, from https://echa.europa.eu/regulations/reach/evaluation/substance-evaluation

European Chemicals Agency. (2019b). Substances restricted under REACH. Retrieved June 13, 2022, from https://echa.europa.eu/regulations/reach/evaluation/substance-evaluation

Evgeniou, T., & Pontil, M. (2001). Support vector machines: Theory and applications. *2049*, 249–257. https://doi.org/10.1007/3-540-44673-7_12

Flemström, K., Carlson, R., & Erixon, M. (2005). Relationships between life cycle assessment and risk assessment potentials and obstacles.

Franzosa, J. A., Bonzo, J. A., Jack, J., Baker, N. C., Kothiya, P., Witek, R. P., Hurban, P., Siferd, S., Hester, S., Shah, I., Ferguson, S. S., Houck, K. A., & Wambaugh, J. F. (2021). High-throughput toxicogenomic screening of chemicals in the environment using metabolically competent hepatic cell cultures. *NPJ Syst. Biol. Appl.*, *7*(1), 7.

Giove, S., Brancia, A., & Satterstrom, F. (2008). Decision support systems and environment: Role of mcda. https://doi.org/10.1007/978-0-387-09722-0_3

Goldstein, B., Navar, A., & Carter, R. (2016). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, *38*, ehw302. https://doi.org/10.1093/eurheartj/ehw302

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. https://doi.org/10.48550/ARXIV.1406.2661

Gottschalk, F., Scholz, R. W., & Nowack, B. (2010). Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles. *Environmental Modelling & Software*, *25*(3), 320–332. https://doi.org/10.1016/j.envsoft.2009.08.011

Goyal, C. (2021). Deep Understanding of Discriminative and Generative Models in Machine Learning. https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/

Graham, J. D. (2008). *Risk Assessment for Regulatory Decision Making* (tech. rep.). ndiana University School of Public and Environmental Affairs. https://www.oecd.org/gov/regulatory-policy/44800773.pdf

Gray, J., & Shenoy, P. (2000). Rules of thumb in data engineering. *Proceedings of 16th International Conference on Data Engineering (Cat. No.00CB37073)*, 3–10. https://doi.org/10.1109/ICDE.2000.839382

Grossi, E., & Buscema, M. (2008). Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, *19*, 1046–54. https://doi.org/10.1097/MEG.0b013e3282f198a0

Guinée, J. B., Heijungs, R., Huppes, G., Zamagni, A., Masoni, P., Buonamici, R., Ekvall, T., & Rydberg, T. (2011). Life Cycle Assessment: Past, Present, and Future. *Environmental Science and Technology*, *45*(1), 90–96. https://doi.org/10.1021/es101316v

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, Ed.; 2nd ed.).

Heiser, W. (2009). Correspondence analysis in practice by greenacre, m. *Biometrics*, *65*, 997–998. https://doi.org/10.1111/j.1541-0420.2009.01315_11.x

Huang, L., Liu, S., Yang, Z., Xing, J., Zhang, J., Bian, J., Li, S., Sahu, S. K., Wang, S., & Liu, T.-Y. (2021). Exploring deep learning for air pollutant emission estimation. *Geoscientific Model Development*, *14*(7), 4641–4654. https://doi.org/10.5194/gmd-14-4641-2021

Isaacs, K. K., Goldsmith, M.-R., Egeghy, P., Phillips, K., Brooks, R., Hong, T., & Wambaugh, J. F. (2016). Characterization and prediction of chemical functions and weight fractions in consumer products. *Toxicology Reports*, *3*, 723–732. https://doi.org/10.1016/j.toxrep.2016.08.011

Islam, A., Belhaouari, S. B., Rehman, A. U., & Bensmail, H. (2022). KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, *115*, 108288. https://doi.org/10.1016/j.asoc.2021.108288

Jacob, K. H. (2015). Sea level rise, storm risk, denial, and the future of coastal cities. *Bull. At. Sci.*, *71*(5), 40–50.

Jeatrakul, P., Wong, K., Fung, C., & Takama, Y. (2010). Misclassification analysis for the class imbalance problem. *2010 World Automation Congress, WAC 2010*, 1–6.

Ji, J., Choi, C., Yu, M., & Yi, J. (2012). Comparison of a data-driven model and a physical model for flood forecasting. *WIT Trans. Ecol. Environ.*, *159*, 133–142. https://doi.org/10.2495/FRIAR120111

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4237–4253. https://doi.org/10.1098/rsta.2009.0159

Kartal, H., Oztekin, A., Gunasekaran, A., & Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering*, *101*. https://doi.org/10.1016/j.cie.2016.06.004

Kashyap, K. (2019). Machine Learning- Decision Trees and Random Forest Classifiers. Retrieved June 30, 2022, from https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544

Kaya Uyanık, G., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Khangar, N., & Kamalja, K. (2017). Multiple correspondence analysis and its applications. *Electronic Journal of Applied Statistical Analysis*, *10*, 432–462. https://doi.org/10.1285/i20705948v10n2p432

Khorramrouz, F., Kajabadi, N. P., Galankashi, M. R., & Rafiei, F. M. (2019). Application of fuzzy analytic hierarchy process (FAHP) in failure investigation of knowledge-based business plans. *SN Applied Sciences*, *1*(11). https://doi.org/10.1007/s42452-019-1394-3

Kindermann, R., & Snell, J. L. (1980). *Markov random fields and their applications* (Vol. 1). American Mathematical Society. https://doi.org/10.1090/conm/001

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2005). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*, 25–36.

Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, *4*, 33. https://doi.org/10.4103/jpcs.jpcs_8_18

Kurz, S., De Gersem, H., Galetzka, A., Klaedtke, A., Liebsch, M., Loukrezis, D., Russenschuck, S., & Schmidt, M. (2022). Hybrid modeling: towards the next level of scientific computing in engineering. *Journal of Mathematics in Industry*, *12*(1), 8. https://doi.org/10.1186/s13362-022-00123-0

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff (H. Suleman, Ed.). *PLOS ONE*, *12*(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Li, M., Ingwersen, W. W., Young, B., Vendries, J., & Birney, C. (2022). Useeior: An open-source r package for building and using us environmentally-extended input-output models. *Applied Sciences*, *12*(9). https://doi.org/10.3390/app12094469

Linkov, I., Trump, B. D., Wender, B. A., Seager, T. P., Kennedy, A. J., & Keisler, J. M. (2017). Integrate life-cycle assessment and risk analysis results, not methods. *Nature Nanotechnology*, *12*(8), 740–743. https://doi.org/10.1038/nnano.2017.152

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Luściński, S. (2015). Perspectives on data engineering: Emerging role in organisation development. https://doi.org/10.13140/RG.2.1.3275.8886

Madhukar, B. (2020). Using Near-Miss Algorithm For Imbalanced Datasets. https://analyticsindiamag.com/using-near-miss-algorithm-for-imbalanced-datasets/#:~:text=Near%2Dmiss%20is%20an%20algorithm,samples%20from%20the%20larger%20class

Mahmood, M. S. (2021). Factor Analysis of Mixed Data. https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c

Mendis, A. (2019). Statistical Modelling vs Machine Learning. https://www.kdnuggets.com/2019/08/statistical-modelling-vs-machine-learning.html

Meyer, D. E., Bailin, S. C., Vallero, D., Egeghy, P. P., Liu, S. V., & Cohen Hubal, E. A. (2020). Enhancing life cycle chemical exposure assessment through ontology modeling. *Science of The Total Environment*, *712*, 136263. https://doi.org/10.1016/j.scitotenv.2019.136263

Meyer, D. E., Mittal, V. K., Ingwersen, W. W., Ruiz-Mercado, G. J., Barrett, W. M., Gonzalez, M. A., Abraham, J. P., & Smith, R. L. (2019). Purpose-driven reconciliation of approaches to estimate chemical releases. *ACS Sustain. Chem. Eng.*, *7*(1), 1260–1270.

Miao, J., & Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., & Laishram, M. (2017). Principal component analysis. *International Journal of Livestock Research*, 1. https://doi.org/10.5455/ijlr.20170415115235

Morris, P. J. T. (2003). History of the International Chemical Industry: From the 'Early Days' to 2000. *The British Journal for the History of Science*, *36*(2), 253–254. https://doi.org/10.1017/S0007087403365045

Muller, L., Schneider, M., Shenoy, M., & Erkman, S. (2011). Expanding material flow analysis (mfa) data into life cycle inventories (lci): Cases from the silk and cotton textile industry, pig iron industry and paper industry in india.

Murmann, J. P. (2002). Chemical Industries after 1850.

Naidu, R., Biswas, B., Willett, I. R., Cribb, J., Kumar Singh, B., Paul Nathanail, C., Coulon, F., Semple, K. T., Jones, K. C., Barclay, A., & Aitken, R. J. (2021). Chemical pollution: A growing peril and potential catastrophic risk to humanity. *Environment International*, *156*, 106616. https://doi.org/10.1016/j.envint.2021.106616

Nantasenamat, C. (2020). Best Practices for Constructing Reproducible QSAR Models. https://doi.org/10.1007/978-1-0716-0150-1_3

National Research Council. (2014). *A Framework to Guide Selection of Chemical Alternatives*. National Academies Press. https://doi.org/10.17226/18872

Ng, A. (2021). MLOps: From Model-centric to Data-centric AI. Retrieved January 19, 2022, from https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Elsevier. https://doi.org/10.1016/C2009-0-27609-4

Peng, J., Lee, K., & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, *96*, 3–14. https://doi.org/10.1080/00220670209598786

Pirouz, D. (2006). An overview of partial least squares. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1631359

Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., & Kuz'min, V. E. (2009). Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *Journal of Chemical Information and Modeling*, *49*(11), 2481–2488. https://doi.org/10.1021/ci900203n

Priatama, A. R., https://journal.ipb.ac.id/index.php/jmht/article/view/33983/23160, & Setiawan, Y. (2022). Regression models for estimating aboveground biomass and stand volume using landsat-based indices in post-mining area. *J. Manaj. Hutan Trop. (J. Trop. For. Manag.)*, *28*(1), 1–14.

Ragas, A. M. (2011). Trends and challenges in risk assessment of environmental contaminants. *Journal of Integrative Environmental Sciences*, *8*(3), 195–218. https://doi.org/10.1080/1943815X.2011.597769

Rasmussen, C., Bousquet, O., Luxburg, U., & Rätsch, G. (2004). Gaussian processes in machine learning. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures, 63-71 (2004)*, *3176*. https://doi.org/10.1007/978-3-540-28650-9_4

Regaya, Y., Fadli, F., & Amira, A. (2021). Point-denoise: Unsupervised outlier detection for 3d point clouds enhancement. *Multimedia Tools* *and Applications*, *80*, 1–17. https://doi.org/10.1007/s11042-021-10924-x

Ring, C. L., Arnot, J. A., Bennett, D. H., Egeghy, P. P., Fantke, P., Huang, L., Isaacs, K. K., Jolliet, O., Phillips, K. A., Price, P. S., Shin, H.-M., Westgate, J. N., Setzer, R. W., & Wambaugh, J. F. (2019). Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways. *Environmental Science & Technology*, *53*(2), 719–732. https://doi.org/10.1021/acs.est.8b04056

Roh, Y., Heo, G., & Whang, S. E. (2021). A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

Rokach, L., & Maimon, O. (2005). Decision trees. https://doi.org/10.1007/0-387-25465-X_9

Roy, K., Kar, S., & Das, R. (2015). Primer on Qsar/Qspr Modeling.

Roy, P., Nei, D., Orikasa, T., Xu, Q., Okadome, H., Nakamura, N., & Shiina, T. (2009). A review of life cycle assessment (LCA) on some food products. *Journal of Food Engineering*, *90*(1), 1–10. https://doi.org/10.1016/j.jfoodeng.2008.06.016

Sabaghi, M., Mascle, C., Baptiste, P., & Rostamzadeh, R. (2016). Sustainability assessment using fuzzy-inference technique (saft): A methodology toward green products. *Expert Systems with Applications*, *56*, 69–79. https://doi.org/https://doi.org/10.1016/j.eswa.2016.02.038

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Scornet, E., Biau, G., & Vert, J.-P. (2014). Consistency of random forests. https://doi.org/10.48550/ARXIV.1405.2881

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Shalizi, C. R. (2009). The Statistical Analysis of Complex Systems Models. https://www.stat.cmu.edu/$%5Csim$cshalizi/stacs/#:$%5Csim$:text=Complex%20systems%20are%20ones%20with,effectively%20determine%20all%20the%20rest.

Sitorus, F., & Brito-Parada, P. R. (2020). A multiple criteria decision making method to weight the sustainability criteria of renewable energy technologies under uncertainty. *Renewable and Sustainable Energy Reviews*, *127*, 109891. https://doi.org/10.1016/j.rser.2020.109891

Slutsky, D. (2014). The Effective Use of Graphs. *Journal of Wrist Surgery*, *03*(02), 067–068. https://doi.org/10.1055/s-0034-1375704

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100306

Soh, J., & Singh, P. (2020). Machine learning operations. In *Data science solutions on azure: Tools and techniques using databricks and mlops* (pp. 259–279). Apress. https://doi.org/10.1007/978-1-4842-6405-8_8

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics*, *10*(1), 3–22. https://doi.org/10.2166/hydro.2008.015

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, *3*(1), 109–118. https://doi.org/https://doi.org/10.1016/0893-6080(90)90049-Q

Srichetta, P., & Thurachon, W. (2012). Applying fuzzy analytic hierarchy process to evaluate and select product of notebook computers. *International Journal of Modeling and Optimization*, *2*, 168–173. https://doi.org/10.7763/IJMO.2012.V2.105

Sukhwani, N. (2020). Handling Data Imbalance in Multi-label Classification (MLSMOTE). https://medium.com/thecyphy/handling-data-imbalance-in-multi-label-classification-mlsmote-531155416b87

Sutton, C., & McCallum, A. (2010). An introduction to conditional random fields. https://doi.org/10.48550/ARXIV.1011.4088

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *Ai Communications*, *30*, 169–190. https://doi.org/10.3233/AIC-170729

The Organization for Economic Co-operation and Development. (2004). OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

The Organization for Economic Co-operation and Development. (2006). QSAR Toolbox. https://qsartoolbox.org/

The Organization for Economic Co-operation and Development. (2021). The costs and benefits of regulating chemicals. Retrieved March 28, 2022, from https://www.oecd.org/chemicalsafety/risk-management/costs-benefits-chemicals-regulation.htm

Unrau, C. Y. (2020). 5 SMOTE Techniques for Oversampling your Imbalance Data. https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5#:~:text=SMOTE%20works%20by%20utilizing%20a,randomly%20selected%20k%2Dnearest%20neighbour

Unrau, S. (2021). Anomaly detection using Isolation Forest – A Complete Guide. https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/

U.S. Environmental Protection Agency. (2017a). About the TSCA Chemical Substance Inventory. Retrieved June 13, 2022, from https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory

U.S. Environmental Protection Agency. (2017b). Risk Evaluations for Existing Chemicals under TSCA. Retrieved June 13, 2022, from https://cutt.ly/iFDGYtX

van der Maaten, L., Postma, E., & van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

van Gils, J., Posthuma, L., Cousins, I. T., Brack, W., Altenburger, R., Baveco, H., Focks, A., Greskowiak, J., Kühne, R., Kutsarova, S., Lindim, C., Markus, A., van de Meent, D., Munthe, J., Schueder, R., Schüürmann, G., Slobodnik, J., de Zwart, D., & van Wezel, A. (2020). Computational material flow analysis for thousands of chemicals of emerging concern in European waters. *Journal of Hazardous Materials*, *397*, 122655. https://doi.org/10.1016/j.jhazmat.2020.122655

Verma, R. P., & Hansch, C. (2005). An approach toward the problem of outliers in QSAR. *Bioorganic & Medicinal Chemistry*, *13*(15), 4597–4621. https://doi.org/10.1016/j.bmc.2005.05.002

Visbal-Cadavid, D., Mendoza-Mendoza, A., & Hoz-Dominguez, E. D. L. (2020). Use of Factorial Analysis of Mixed Data (FAMD) and Hierarchical Cluster Analysis on Principal Component (HCPC) for Multivariate Analysis of Academic Performance of Industrial Engineering Programs. *Journal of Southwest Jiaotong University*, *55*(5). https://doi.org/10.35741/issn.0258-2724.55.5.34

Webb, G. I., Keogh, E., Miikkulainen, R., Miikkulainen, R., & Sebag, M. (2011). Naïve bayes. In *Encyclopedia of machine learning* (pp. 713–714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005

Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D., Bilotta, G., Coggon, D., Collins, C., Dempsey, C., Duarte-Davidson, R., FitzGerald, R., Galay-Burgos, M., Gee, D., Hoffmann, S., Lam, J., Lasserson, T., Levy, L., Lipworth, S., Ross, S. M., ... Taylor, D. (2016). Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environment International*, *92-93*, 556–564. https://doi.org/10.1016/j.envint.2015.11.002

Whittaker, M. H. (2015). Risk Assessment and Alternatives Assessment: Comparing Two Methodologies. *Risk Analysis*, *35*(12), 2129–2136. https://doi.org/10.1111/risa.12549

Wittwehr, C., Blomstedt, P., Gosling, J. P., Peltola, T., Raffael, B., Richarz, A.-N., Sienkiewicz, M., Whaley, P., Worth, A., & Whelan, M. (2020). Artificial intelligence for chemical risk assessment. *Comput. Toxicol.*, *13*(100114), 100114.

Young, B., Ingwersen, W. W., Bergmann, M., Hernandez-Betancur, J. D., Ghosh, T., Bell, E., & Cashman, S. (2022). A system for standardizing and combining u.s. environmental protection agency emissions and waste inventory data. *Applied Sciences*, *12*(7). https://doi.org/10.3390/app12073447

Zhang, J., & Mani, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proceeding of International Conference on Machine Learning*. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1603053

Zhang, K., Hutter, M., & Jin, H. (2009). A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. https://doi.org/10.1007/978-3-642-01307-2_84

Zhang, M.-L., & Zhou, Z.-H. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593

Zhou, T., Gani, R., & Sundmacher, K. (2021). Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design. *Engineering*, 7(9), 1231–1238. https://doi.org/https://doi.org/10.1016/j.eng.2020.12.022

Zong, P., Jiang, J., & Qin, J. (2020). Study of High-Dimensional Data Analysis based on Clustering Algorithm. *2020 15th International Conference on Computer Science & Education (ICCSE)*, 638–641. https://doi.org/10.1109/ICCSE49874.2020.9201656

2

OBJECTIVE

2.1 SCOPE AND OBJECTIVES OF THE THESIS

2.1.1 *Main objective*

This thesis aims to develop a data-driven methodology from publicly-accessible data for rapid estimation of emissions and exposure to chemicals during generic end-of-life scenarios in order to incorporate and understand exposure pathways during such scenarios to streamline the chemical risk assessment process.

2.1.2 *Specific objectives*

OBJECTIVE I: Develop data engineering pipelines to connect publicly-available and siloed database systems to build a centralized database and datasets that include the necessary elements for the generation of conceptual models of chemical exposure during generic end-of-life scenarios.

OBJECTIVE II: Perform exploratory data analysis for early identification of constraints and implications for building data-driven models from data obtained through data engineering.

OBJECTIVE III: Propose data-driven modeling alternatives in order to identify potential generic end-of-life chemical exposure scenarios, perform a chemical flow analysis and understand the elements that make up the end-of-life management chain.

Part I

UNDERSTANDING THE CHEMICAL END-OF-LIFE MANAGEMENT CHAIN AND RECYCLING LOOP

3

DATA ENGINEERING FOR TRACKING CHEMICALS AND RELEASES AT INDUSTRIAL END-OF-LIFE ACTIVITIES

ABSTRACT

Performing risk evaluation is necessary to determine whether a chemical substance presents an unreasonable risk of injury to human health or the environment across its life cycle stages. Data gathering, reconciliation, and management for supporting risk evaluation are time-consuming and challenging, especially for end-of-life (EoL) activities due to the need for proper reporting and traceability. A data engineering framework using publicly-available databases to track chemicals in waste streams generated by industrial activities and transferred to other facilities across different U.S. locations for waste management is implemented. The analysis tracks chemicals in waste streams generated at industrial processes and handling at off-site facilities and then estimates releases from EoL activities. The final product of this effort is a framework that identifies a set of chemical, activity, and industry sector categories as well as hazardous waste flows, emission factors, and uncertainty indicators to describe EoL activities. This framework helps to identify EoL exposure scenarios that would otherwise not be evaluated. As a case study, methylene chloride, one of the first ten chemicals to undergo risk evaluation under the amended U.S. Toxic Substances Control Act, was evaluated with results highlighting potential additional exposure scenarios.

Keywords: Chemical releases; Chemical risk; Data management; Industrial activities; Waste management

RESUMEN

La evaluación de riesgos es necesaria para determinar si una sustancia química presenta un riesgo excesivo de daño para la salud humana o el medio ambiente en todas las fases de su ciclo de vida. La recopilación, la conciliación y la gestión de datos para apoyar la evaluación de riesgos requieren mucho tiempo y suponen un reto, especialmente en el caso de las actividades de fin de vida (EoL, por sus siglas en inglés), debido a la necesidad de una información y una trazabilidad adecuadas. Se implementa un marco de ingeniería de datos que utiliza bases de datos disponibles públicamente para rastrear las sustancias químicas en los flujos de residuos generados por las actividades industriales y transferidos a otras instalaciones en diferentes lugares de Estados Unidos para la gestión de residuos. El análisis rastrea las sustancias químicas en los flujos de residuos generados en los procesos industriales y la manipulación en instalaciones externas y luego estima las liberaciones de las actividades de EoL. El producto final de este esfuerzo es un marco que identifica un conjunto de categorías de productos químicos, actividades y sectores industriales, así como flujos de residuos peligrosos, factores de emisión e indicadores de incertidumbre para describir las actividades de EoL. Este marco ayuda a identificar los escenarios de exposición de EoL que, de otro modo, no se evaluarían. Como estudio de caso, se evaluó el cloruro de metileno, una de las diez primeras sustancias químicas que se sometieron a una evaluación de riesgos en virtud de la Ley de Control de Sustancias Tóxicas modificada de Estados Unidos, con resultados que ponen de manifiesto posibles escenarios de exposición adicionales.

Palabras clave: Emisiones químicas; Riesgo químico; Gestión de datos; Actividades industriales; Gestión de residuos

3.1 INTRODUCTION

Chemical risk evaluation has been broadly utilized for supporting alternative assessment via a strategy of informed substitution to select chemicals and materials with safer profiles (Whittaker, 2015). The use of risk analysis has increased due to drivers such as green chemistry and engineering, circular economy, and life cycle analysis for avoiding the selection and use of undesirable chemical substitutions (Bodar et al., 2018; National Research Council, 2014). This evaluation has not only been performed in technical fields but also in regulatory decision-making to protect society and the environment (Bernas, 2013), thereby driving the promotion of safer chemical adoption and identifying so-called priority substances (e.g., persistent, bioaccumulative, and toxic chemicals) (National Research Council, 2014). The Toxic Substances Control Act (TSCA), amended in June 2016 by the Frank R. Lautenberg Chemical Safety for the 21st Century Act (U.S. Environmental Protection Agency, 2017b), requires the development of a strategy to determine whether a chemical substance in the U.S. market may pose an unreasonable risk of harming the environment or human health (U.S. Environmental Protection Agency, 2016g). However, there is a continued growth of the TSCA chemical inventory, currently listing 86,406 existing chemicals that can legally be used in U.S. commerce, of which 41,484 are believed to still be in use today, and 90 chemicals are part of the TSCA Workplan near-term review and risk assessment (U.S. Environmental Protection Agency, 2017a). These facts make the chemical risk evaluation a time-consuming and challenging task for anyone to perform. Thus, based on hazards, the potential for exposure, and available information, a prioritization step before risk evaluation is addressed to identify existing chemicals that are of high-priority for subsequent risk analysis (U.S. Environmental Protection Agency, 2016g).

EPA published the list of the first ten chemicals for risk evaluation on December 19, 2016, and, in June 2018, the problem formulation documents of these chemicals were made available. In the problem formulation, EPA develops conceptual models for chemical substances based on evaluation of reasonably available information for physical and chemical properties, environmental fate, exposures, hazards, uses, and consideration of exposure/emission restrictions implemented under other statutes by EPA or other regulatory authorities (U.S. Environmental Protection Agency, 2018a). The models include elements such as exposure pathways (e.g., air), exposure routes (e.g., inhalation), potentially exposed and susceptible subpopulations (e.g., workers), and hazards (e.g., carcinogenic effect), through manufacturing, processing, use (industrial, consumer, and commercial), and disposal of the concerning chemical. Also, EPA identifies potential exposure pathways at the end-of-life (EoL) stage and determines which EoL activities, i.e., recycling, energy recovery, treatment, and disposal (U.S. Environmental Protection Agency, 2014a), will be considered and which will not. Some exposure routes are identified as potentially of concern due to the lack of regulatory requirements or current chemical management methods that may be insufficient to protect human health or the environment. Otherwise, some pathways do not require further analysis, as these do not represent a risk due to the physical properties of the chemical under evaluation. For example, in the risk evaluation of N-methylpyrrolidone, the land-applied biosolids pathway is not considered because N-methylpyrrolidone has a high-water solubility, limited potential for adsorption to organic matter, and biodegrades rapidly (U.S. Environmental Protection Agency, 2018a). Therefore, N-methylpyrrolidone partitions to the aqueous phase, and its residues would not be expected to persist in the environment. Furthermore, other pathways are not included because of the assumption that other EPA regulations provide adequate protection. For instance, air emissions from municipal and industrial waste incineration and energy recovery units are not further assessed as the U.S. Clean Air Act provides standards to regulate all sources of air emissions. Other pathways, like off-site EoL activities, are not under consideration because of the epistemic uncertainty or limited data availability. For instance, during the off-site recycling of waste containing N-methylpyrrolidone, other additional releases of this chemical may occur to the environment, which was acknowledged but not addressed in the N-methylpyrrolidone problem formulation document due to a lack of reasonably available information (U.S. Environmental Protection Agency, 2018a).

Therefore, determining pathways that a chemical flow may follow at the EoL stage is essential since a key step in any chemical risk evaluation is understanding if, where, and how exposure to the chemical may occur. Several studies describe the implementation of different methods and software tools to estimate releases or connect chemical flows with exposure pathways through various life cycle stages (e.g., manufacturing) and activities (e.g., container filling). For instance, among the methods focusing on EoL scenarios, there are those developed by Li et al. (2015) and Clift et al. (2000) based on life cycle analyses to assess human health risk due to solid waste management activities and the EPA model-based approaches for characterizing potential human and ecological health risks from land-based solid waste management units (U.S. Environmental Protection Agency, 2003a).Pizzol et al. (2019) used probabilistic-based strategies to evaluate the risks posed by organic pigments. McNally et al. (2014) and Banerjee et al. (2014) developed Bayesian models for occupational exposure assessment during the manufacturing stage. Conley (2011) applied spatial interaction models for estimating exposure to carcinogenic chemicals associated with lung cancer mortality. Although the latter four frameworks are not focused on EoL, these can be extrapolated to EoL scenario analyses. Likewise, Rosenbaum et al. (2007) proposed a versatile framework based on matrix algebra that can be extended to EoL for connecting emissions to impacts. Nevertheless, all those methods are case-by-case oriented applications, manual-configuration basis, and time-consuming routines for gathering individual information for running the model and estimate chemical releases and exposures.

Cashman et al. (2016) developed a data mining framework for the rapid life cycle inventory, and Meyer et al. (2020) proposed an ontology modeling for assessment of life cycle chemical exposure. Although the former frameworks move beyond the case-by-case and time constraints, these are only applicable to substances that are in the original data source and require some advanced knowledge about the regulatory programs providing the data. However, their underlying idea is extendable to the EoL stage. Delmaar et al. (2005) developed and applied some model-based and data-driven tools to evaluate human exposures to chemical compounds in non-food consumer products, even in the absence of data about exposure to the compounds in the consumer products. Meyer et al. (2019) proposed a machine learning approach to estimate chemical air releases during its manufacturing. These latter approaches advance the prediction of an output state variable (e.g., air emissions) in cases where there is an absence of data. However, the performance of these data-driven models is strongly dependent on the specific dataset domain.

Academic research and government programs have moved forward to provide data and insights for supporting material flow traceability at the EoL stage. For example, Nost et al. (2017) used USPEA's waste records for developing HazMatMapper for visualizing transnational flow transfers in North America. However, HazMatMapper focuses on environmental justice purposes and tracks only total amounts of hazardous waste flows instead of individual chemicals and their risk evaluation. The Organization for Economic Co-operation and Development started working on Pollutant Release and Transfer Registers, because of the United Nations Conference on Environment and Development in Rio de Janeiro in 1992. The Pollutant Release and Transfer Registers are publicly accessible databases with information about releases to air, water, and soil and off-site transfers for further management (Organization for Economic Co-operation and Development, 2015). This set of databases provides information about which chemicals are being transferred or released, where, how much, and by whom. However, in the case of waste brokering, they do not directly give information about the facility in charge of final waste management.

Also, these inventories do not allocate the quantities of chemical releases to the facility receiving the chemical of concern. The EPA's Standardized Emission and Waste Inventories is a collection of Python modules that provide processed EPA emission and waste generation inventory data in standard tabular formats using common identifications for facilities and chemicals (U.S. Environmental Protection Agency, 2017c). However, this last approach does not track chemical transfers nor target chemical risk.

Therefore, our novel data engineering framework for tracking chemical flows at EoL activities would enhance chemical risk assessment by identifying EoL exposure scenarios that would otherwise not be evaluated. This framework effectively gathers, cleans, transforms, and integrates qualitative and quantitative information from multiple publicly-available databases. The framework has five key aims to provide relevant data and describe EoL activities. First, exposure pathways at the EoL stage are identified, considering automatically whether an exposure pathway may be of interest for an environmental regulation and material flow analysis downstream of the waste generation source; therefore, optimizing time and resources. Second, transfers of existing chemicals for the facilities in charge of the final waste management are tracked and mapped, also considering those scenarios where there is waste brokering, which is not a direct task using the Pollutant Release and Transfer Registers databases. Third, leveraging facility-level information, a quick estimate is made of how much of a chemical transferred may be potentially released to the environment during an EoL activity. Fourth, an EoL dataset and machine-readable queries for future automatization are developed, which may be used to supply specific dataset domain for the development of data-driven models and thus, extending the framework to chemicals that are not in the built EoL dataset. Fifth, the framework is made easily accessible and interoperable to connect with exposure, hazard, and subsequent risk assessments.

Currently, this data engineering framework considers waste flows generated by industrial and commercial facility activities and transferred for further waste management at off-site facilities across different U.S. locations. The EoL framework dataset contains records of chemical transfers, and their classification (e.g., acrylamides); type of EoL activity (e.g., recycling); the industry sector of the recycling, energy recovery, treatment & disposal facility (e.g., materials recovery); and the environmental compartment where emission occurs (e.g., surface water, air). Also, the framework can be used for supporting the inclusion of EoL activities in life cycle assessment by providing inventories for existing chemicals in the build EoL dataset (Cashman et al., 2016), using sustainability performance indicators for chemical risk assessment at the EoL stage (Hernandez-Betancur & Ruiz-Mercado, 2019), integrating life cycle assessment and risk analysis results at EoL activities (Linkov et al., 2017), and boosting a safer circular economy of chemicals while minimizing environmental and human health risks (Lahl & Zeschmar-Lahl, 2013).

3.2 GENERIC CONCEPTUAL MODEL FOR CHEMICAL RISK EVALUATION

As part of the development of risk evaluation, the step is the preparation of conceptual models. The models are composed of different elements, like those presented in Figure 3.1. A generic conceptual model includes:

1. *Condition of use (CoU)*: A CoU is the circumstance under which a chemical substance is intended, known, or reasonably foreseen to be manufactured, processed, distributed in commerce, used, or disposed of (U.S. Environmental Protection Agency, 2016j).
2. *Stressors*: Chemical substances that may have an adverse effect on human health or the environment (U.S. Environmental Protection Agency, 2003b).
3. *Exposure pathways and routes*: They are included to understand the relationship between receptors and stressors. An exposure pathway is defined as the physical passage (environmental fate and transport) that a stressor takes from its CoUs to a receptor. In contrast, an exposure route is a way that a stressor enters an organism after contact (e.g., inhalation) (U.S. Environmental Protection Agency, 1992).
4. *Receptors*: Agents exposed to the stressors, for instance, workers in a recycling facility..
5. *Hazards or endpoints*: Effects of a stressor on a receptor, e.g., cancer.

As presented in Figure 3.1, the EoL stage includes recycling, energy recovery, treatment, and disposal activities. However, recycling activities are included at the processing stage in TSCA analysis, while energy recovery, treatment, and disposal activities constitute the TSCA disposal stage. This framework groups these four activities into the EoL stage for analyzing these CoUs of chemical stressor releases separately from the rest of the activities and uses in the life cycle. Figure 3.1 depicts that chemical manufacturing, processing, and use stages can be a source of waste containing a concerning chemical. The framework classifies these waste sources into CoU categories in Table 3.1, based on TSCA Chemical Data Reporting (TSCA CDR). TSCA CDR requires manufacturers and importers to provide information on the chemicals they produce domestically or import into the

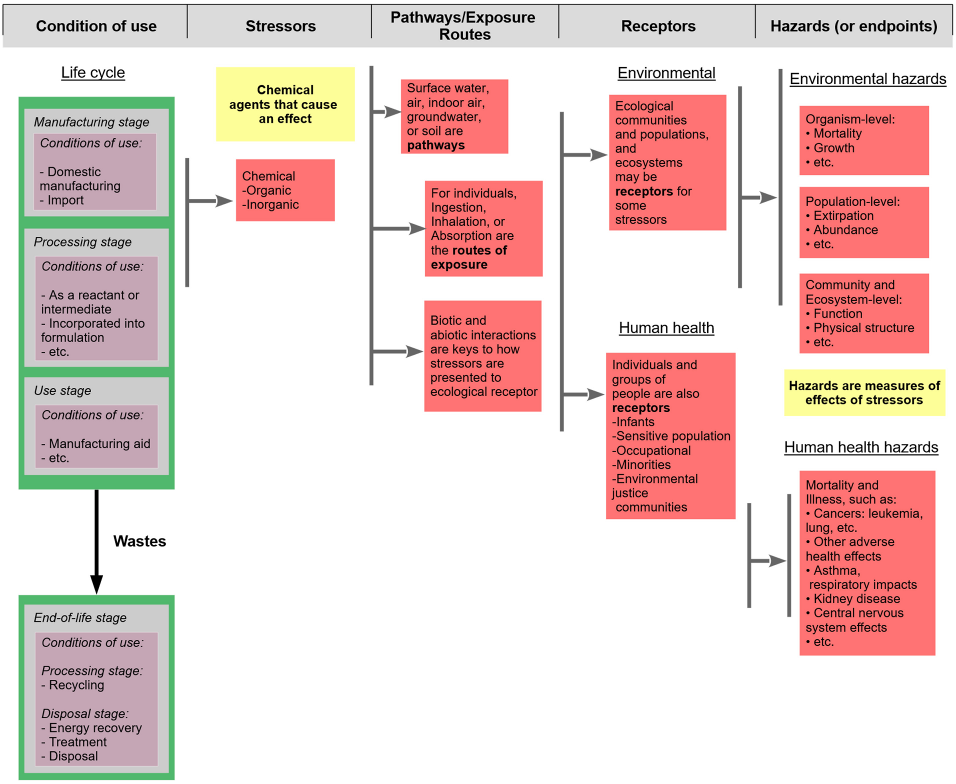


Figure 3.1: A generic conceptual model with examples of possible elements and linkages for chemical risk evaluation (adapted from U.S. Environmental Protection Agency (2003b), and modified based on U.S. Environmental Protection Agency (2003c, 2016g)).

U.S. (U.S. Environmental Protection Agency, 2016d). Nevertheless, this framework is limited to tracking wastes from industrial and commercial sources that are transferred to off-site facilities located across the U.S. But, outcomes from this research can support the inclusion of EoL activities in life cycle assessment studies elsewhere.

The problem formulation document for carbon tetrachloride describes an example of a conceptual model. EPA established this compound could be used as a chemical processing aid in agricultural product manufacturing (U.S. Environmental Protection Agency, 2018b). In this case, the chemical is present at the industrial use stage, and according to Table 3.1, the category of CoU is a non-incorporative activity. This CoU may generate wastewater and liquid waste containing carbon tetrachloride. The resulting waste is then transferred to EoL activities for further handling (i.e., waste handling for disposal, treatment, energy recovery, or recycling), such as industrial wastewater treatment. From those EoL activities, for example,

Table 3.1: Categories of condition of use under which waste may be generated at industrial facilities. The classification is based on Chemical Data Reporting under the Toxic Substances Control Act (U.S. Environmental Protection Agency, 2016d).

Life cycle stage	**Category**	**Comment**
Manufacturing	Domestic manufacturing	A chemical substance is produced in the U.S.
	Import	A chemical substance is imported into the U.S.
Processing	As reactant	A chemical substance is used in chemical reactions for the manufacturing of another chemical substance or product
	Incorporation into a formulation, mixture, or a reaction product	A chemical substance is added to a product (or product mixture) prior to further distribution of the product
	Incorporation into an article	A chemical substance becomes an integral component of an article distributed for industrial, trade, or consumer use
	Repackaging	Preparation of a chemical substance for distribution in commerce in a different form, state, or quantity. This includes transferring the chemical substance from a bulk container into smaller containers
Industrial use[1] /commercial use[2]	Non-incorporative activities	A chemical substance is otherwise used (e.g., as a chemical processing or manufacturing aid, for cleaning or degreasing)

[1] "Industrial use" means use at a site at which one or more chemicals or mixtures are manufactured (including imported) or processed.

[2] "Commercial use" means the use of a chemical or a mixture containing a chemical (including as part of an article) in a commercial enterprise providing saleable goods or services.

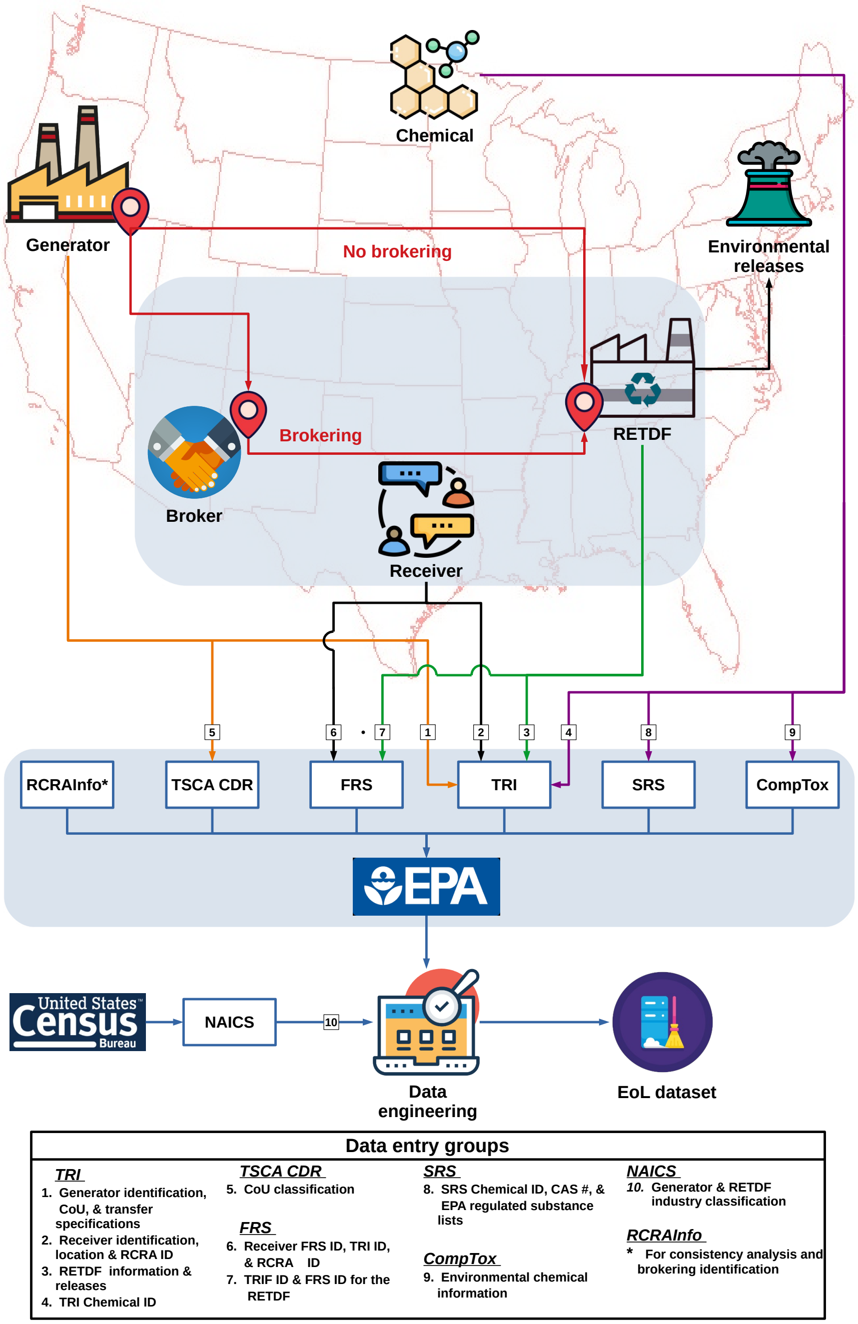
recycling, fugitive emissions to air of the carbon tetrachloride contained in the waste may occur and affect workers by inhaling chemical vapor producing liver problems by acute toxicity (U.S. Environmental Protection Agency, 2018b). Thus, in accordance with 3.1, recycling is the EoL activity, carbon tetrachloride is the stressor, the air is the exposure pathway, inhalation is the exposure route, the workers are the receptors, and liver toxicity is the hazard that carbon tetrachloride (the stressor) can cause in the workers (the receptors) handling the chemical in the above-mentioned waste management activities.

3.3 METHODOLOGY: DATA ENGINEERING FRAMEWORK

Figure 3.2 shows the schematic foundation of the proposed framework. A chemical in a waste stream is tracked from an industrial facility (i.e., waste generator), to a recycling, energy recovery, treatment & disposal facility (RETDF), considering that an intermediate facility or intermediary waste broker facilitates transfer to a RETDF. Also, as presented in Figure 3.2, the entities in the EoL supply and management chain include generators, brokers, and RETDFs. However, the framework considers an "auxiliary" entity named as receiver. The gray block enclosing the receiver, broker, and RETDF denotes that the receiver can be the RETDF or broker. If there is waste brokering, the receiver is the broker; otherwise, it is the RETDF. This tracking is essential to find potential EoL stage locations, exposure scenarios, and the environmental releases (i.e., soil, water, and air) affecting workers or other receptors (e.g., occupational non-users). Furthermore, the framework is used to build a generic EoL dataset to support risk evaluation at the EoL stage, specifically in pathways that might not be further analyzed previously (e.g., off-site recycling) due to lack of an easily accessible and interoperable system that enable exchange, integration, and cooperative use of data. The EoL dataset allows for clustering chemicals according to similar hazards and generators and RETDFs into industry sectors, and EoL activities based on existing classifications to streamline risk evaluation. In summary, the EoL dataset structure includes the following data entries, which are fully listed by database names in Table A.1 of the Appendix A:

1. Chemicals, their chemical category based on hazards, environmental regulations to which they are subject, and their identification number for future queries.

2. Generators and RETDF locations in the U.S., their industry sectors based on primary economic activities and unit processes similarity, and their identification numbers for future queries.



Data entry groups

TRI
1. Generator identification, CoU, & transfer specifications
2. Receiver identification, location & RCRA ID
3. RETDF information & releases
4. TRI Chemical ID

TSCA CDR
5. CoU classification

FRS
6. Receiver FRS ID, TRI ID, & RCRA ID
7. TRIF ID & FRS ID for the RETDF

SRS
8. SRS Chemical ID, CAS #, & EPA regulated substance lists

CompTox
9. Environmental chemical information

NAICS
10. Generator & RETDF industry classification

RCRAInfo
* For consistency analysis and brokering identification

Figure 3.2: Schematic explanation about the relationship between the entities in the EoL supply and management chain, data entries, and data sources after refinement, harmonization, and building the EoL dataset. Figure A.1 in the Appendix A presents a detailed connection based on Table A.1 notation.

3. CoUs under which the wastes were generated.
4. Quantities of the chemicals transferred by the generators to RETDFs in a reporting year.
5. Receiver locations in the U.S. and their EPA identification numbers.
6. EoL activity classifications. The EoL activities are classified into three different categorizations and reported as separate columns in the EoL dataset: EPA waste management hierarchy (U.S. Environmental Protection Agency, 2015d), Toxics Release Inventory Program (U.S. Environmental Protection Agency, 2014b), and TSCA reports (U.S. Environmental Protection Agency, 2016g).
7. Maximum amounts of chemicals present at RETDFs, quantities of chemicals released by RETDFs, and quantity of chemicals as waste generated by RETDFs during a reporting year.
8. Reliability indicators of the flow data collected.

Figure 3.3 depicts how to use the EoL dataset to streamline the risk evaluation process. The EoL dataset can relate to hazard assessment using the chemical category ($D_{17} - D_{19}$ in Table A.1) or identification ($D_{11} - D_{15}$) for collecting hazard parameters. The former method could be used even for chemicals out of the built EoL dataset or new in the market. The publicly-available databases with exposure records could be connected utilizing the RETDF identification ($D_{44} - D_{51}$) or industry sector (D_{52}) and the chemical category or identification. Currently, the framework can provide an EoL exposure pathway or EoL activity where it is expected that a chemical is present ($D_{32} - D_{34}$, and D_{57}). The identification of whether an EoL exposure pathway is or is not of interest for an environmental program can be made using the lists of substances of concern ($D_{20} - D_{25}$). For instance, if a chemical is classified as a hazardous air pollutant, i.e., ($D_{20} = Yes$), then the U.S. Clean Air Act provides standards to regulate all sources of air emission, therefore, for that concerning chemical, other environmental regulation would optimize resources by not addressing risk evaluation of stack or fugitive releases from EoL activities such as treatment via incineration.

For existing chemicals in the EoL dataset, the framework leverages the existing information to quantify emission factors to estimate the quantity of the chemicals transferred that might be potentially released to the environment from the RETDF during an EoL activity. Reliability and temporal correlation indicators of each flow in the EoL dataset (D_{31} in Table A.1) can be used to connect with methodologies for uncertainty

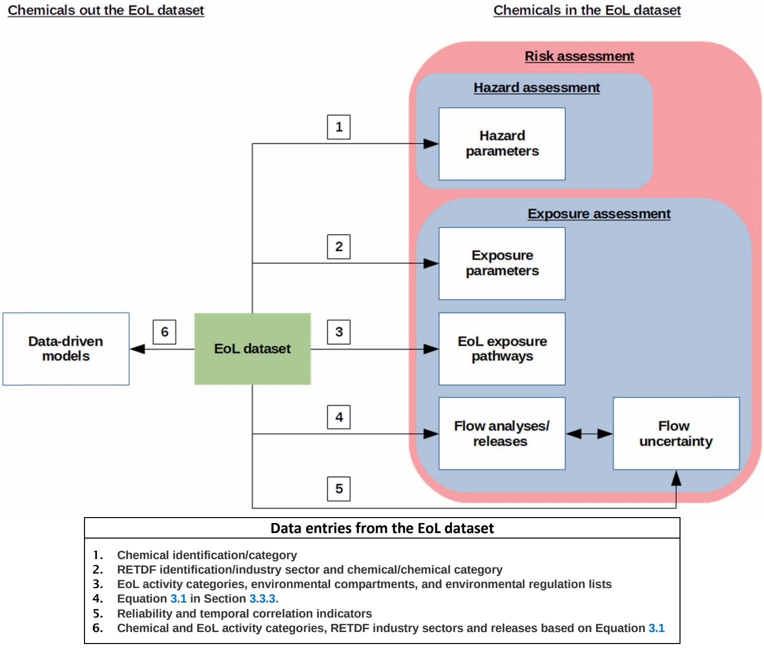


Figure 3.3: Schematic explanation about how to connect the EoL dataset with hazard, exposure, and risk assessment for existing chemicals and with data-driven models for chemicals outside the built EoL dataset.

quantification, which will be explored in future work for integrating into the framework (Edelen & Ingwersen, 2016). As shown in Figure 3.3, the EoL dataset provides domain-specific information for using the relationship of the state variables to build data-driven models. Grouping the EoL activities and chemicals into categories and the RETDFs into industry sectors may facilitate the future development and performance of the models. In this way, the EoL dataset can be used to extend the applicability of the framework to chemicals outside the built dataset, either they are existing but not subject to the EPA programs or new in the U.S. market.

3.3.1 *Data sources: publicly-available databases*

The framework enables exchanging and integrating the information from siloed publicly-available databases such as Toxics Release Inventory (TRI), Facility Registry Service (FRS), Substance Registry Services (SRS), Resource Conservation and Recovery Act Information (RCRAInfo), and the Computational Toxicology (CompTox) Chemicals Dashboard. Figure 3.2 illustrates how the framework connects the databases, which are reporting information from the EoL supply and management chain entities, the chemical of concern properties, and the type of data entries into the EoL dataset. Below, it is explained how each database supports the process of information extraction and refinement, and how these connect with the regulated substance lists:

1. The TRI Program collects information on certain toxic chemicals. Facilities subject to TRI reporting requirements submit forms with information such as pollution prevention, environmental release quantities, including on-site and off-site waste management (i.e., off-site transfers) (Gaona et al., 2020a; Gaona et al., 2020; U.S. Environmental Protection Agency, 2014b). Therefore, annual data entries on TRI chemicals may be retrieved, including specifics such as the quantity of transferred chemical, the purpose of the transfer (e.g., for acid regeneration), and to where (see Figure 3.2).

2. The TSCA CDR database is used to allocate the industrial activities into CoU categories, which is an advantage since the TRI program uses similar chemical activity categories that are selected by facilities when they fill out their TRI reporting forms. Thus, the chemical activity categories of industrial CoUs taken from TSCA CDR and defined in Table 3.1, are used as a data filter in the framework.

3. RCRAInfo provides information about hazardous waste such as waste management and handlers (e.g., generators, transporters, treaters, brokers, and disposers) (U.S. Environmental Protection Agency, 2015c). This information helps track a chemical at the EoL stage, by providing, for instance, the quantity of hazardous waste received or transferred by a facility, where it was transferred to (i.e., physical location), and why (e.g., for wastewater treatment). Although RCRAInfo does not provide any numerical data feature in the framework, a waste that contains a TRI chemical may be classified as hazardous waste under the Resource Conservation and Recovery Act (RCRA), which is the U.S. public law in charge of the proper management of hazardous and non-hazardous wastes. This classification helps for (i) calculating the maximum amount of chemical which may be potentially released from an EoL activity at the RETDF, as further shown in Section 3.3.3; (ii) for tracking the chemical when there are waste brokers, as shown in Figure 3.2; and (iii) for making consistency analyses of flows when a compound is tracked from the waste generator to the RETDF. Furthermore, using RCRAInfo, the RCRA ID for the RETDF may be identified and subsequently used to merge with information from the FRS and TRI.

4. FRS is a central system for information about facilities and sites subjected to EPA regulations, helping to identify them across several EPA programs (U.S. Environmental Protection Agency, 2004). FRS helps one to search for an off-site facility (broker, receiver, or RETDF) that is subject to regulations by EPA programs and is an environmental concern. Thus, facility identifiers in FRS (e.g., FRS IDs) can be used to integrate data from other data sources.

5. SRS has information about substances tracked and regulated by EPA or other organizations. SRS makes it possible to identify a substance across EPA regulatory programs regardless of alternative names (U.S. Environmental Protection Agency, 2016h). Thus, SRS finds TRI chemicals across different EPA regulations by using an internal tracking number. The framework incorporates the following lists of substances of concern from SRS: hazardous air pollutants (U.S. Environmental Protection Agency, 2016c), chemical pollutants that may occur in U.S. biosolids (U.S. Environmental Protection Agency, 2019a), chemical water pollutants for which EPA has developed analytical test methods (U.S. Environmental Protection Agency, 2016i), contaminants that are subject primary drinking water regulation (U.S. Environmental Protection Agency, 2016f) and those which are not currently subject to any primary water regulation, but are known to occur in public water systems (U.S. Environmental Protection Agency, 2016a), and listed hazardous wastes (U.S. Environmental Protection Agency, 2016b). These lists are used in determining whether a TRI chemical is subject to other EPA regulations, leading to having well-established analytical tools for risk assessment in one or more EoL activities and optimizing resources for not analyzing those pathways under TSCA. For example, if a chemical is considered a hazardous waste, then EoL activities, such as underground injection, must satisfy requirements under RCRA Subtitle C, which sets up the framework for the U.S. system of hazardous waste control. Additionally, SRS is used to find whether a chemical is part of TSCA inventory (only non-confidential) (U.S. Environmental Protection Agency, 2015b).

6. CompTox Chemicals Dashboard is a web-based tool that provides valuable information to meet the needs of the environmental sciences and computational chemistry and toxicology communities by integrating the diverse type of relevant domain data (Williams et al., 2017). CompTox is used to retrieve the Simplified Molecular Input Line Entry System (SMILES) for each TRI chemical. SMILES is a chemical notation that encodes information on the molecular structure in a way that can be used by a computer (Toropov et al., 2005).

SMILES is used together with the Analog Identification Methodology (AIM) tool to classify each TRI chemical according to TSCA NCP chemical categories. The EPA's Office of Pollution Prevention and Toxics developed AIM to predict and assess the hazards of new chemical substances by employing an inventory of over 86,000 chemicals analyzing structure to match potential analogs (U.S. Environmental Protection Agency, 2016k). This research searches for clustering chemicals, as is done with the TSCA NCP chemical categories currently used to streamline the EPA review of new chemical substances (Jones, 2010). Furthermore, the Organization for Economic Co-operation and Development integrated the TSCA NCP chemical categories into the Quantitative Structure-Activity Relationship toolbox, which is employed to estimate ecotoxicity and other parameters associated with the assessment of hazardous properties of chemicals (Organization for Economic Co-operation and Development, 2017). Therefore, TSCA NCP categorization may help to rapidly connect the chemical flow tracking with information from hazard assessment.

7. The North American Industry Classification System (NAICS) was developed for use by federal statistical agencies for the collection, analysis, and publication of statistical data related to the US economy. The NAICS uses numerical codes that can range from two to six digits as descriptors of specific industry sectors and subsectors therein. The U.S. Census Bureau uses NAICS to assign a code to each establishment based on primary activity to collect, tabulate, analyze, and publish statistical data related to the U.S. economy (United States Census Bureau, 2017). Most EPA programs require that facilities report their 6-digit NAICS code that represents their industry sectors. The NAICS structure from the U.S. Census Bureau is used to obtain the names of industry sectors for the generator and RETDF. NAICS helps cluster the RETDFs by industry sectors considering their primary economic activity. Additionally, NAICS was developed following a principle of aggregation, which means that producing units using similar production processes should be grouped (United States Census Bureau, 2017). Hence, NAICS may streamline gathering occupational exposure information for RETDFs based on process similarity.

3.3.2 *Data engineering process for tracking chemicals*

This framework for tracking chemical flows at EoL activities and constructing a generic chemical EoL dataset is documented as sequential steps, as illustrated in Figure 3.4. The framework transforms the "raw data" into machine-readable queries for future automatization. As mentioned earlier, the current approach is only applicable to chemical flows generated for industrial and commercial activities (e.g., manufacturing, processing, and use). Furthermore, generator transfers to sewage treatment facilities are not considered since those facilities do not report to the TRI Program. Thus, the following steps, with numbering corresponding to Figure 3.4, are part of the framework:

1. *Is a generator record with CoU & EoL activities of interest?* A TRI reporting year is selected. For this research, the TRI reporting year 2017 is used, which was the most recent data available online at the time of the study. TRI Basic Plus Data Files - File 1a (*Releases and Other Waste Management*) and 1b (*Chemical Activities and Uses*) - are used in this step (U.S. Environmental Protection Agency, 2013). Consistent with the information submitted by the generator (i.e., a facility that reported to TRI), the record is selected if the three following conditions are satisfied: the generator uses the chemical under any CoU listed in Table 3.1; the EoL activity is of interest, i.e., if the generator sends the waste containing the chemical off-site for further waste management; and the receiver (broker or RETDF) is in a U.S. territory. In this step, the TRI data entry groups 1, 2, 4, and 5 shown in Figure 3.2 are gathered.

2. *Check environmental concern lists*: The SRS chemical ID is collected using the TRI chemical ID and then utilized to check whether the chemical is part of any environmental regulated list. In this step, the SRS data entry group 8 in Figure 3.2 is gathered.

3. *Use AIM to categorize chemicals based on TSCA NCP*: The CAS number is used to gather the chemical SMILES from CompTox. The SMILE for each chemical is the input to the AIM tool. The AIM determines which TSCA NCP chemical category (e.g., epoxides) the chemical could belong based on SMILES.

4. *Is the receiver facility listed in FRS?* The TRI data entry group 2 is used to search for the FRS data entry group 6 in Figure 3.2. The TRI data (entry group 2) is used to confirm accurate and complete reporting of RCRA ID information. FRS is searched, and if the RCRA ID submitted with the TRI report matches the FRS record, this ID is used to retrieve the FRS data (entry group 6); otherwise, if the RCRA ID information is not accurate, the receiver name and location are utilized to try to find a match in the FRS records.

5. *Is a record with RCRA receiver, hazardous waste & brokering?* For a record to continue to step 6, it must satisfy these conditions: the receiver has a RCRA ID, the chemical may be a listed hazardous waste (non-null RCRA ID), and the generator reported that the transfer was to a broker (based on the information gathered at step 1). If all three conditions are not met, the record goes to step 8.

6. *Map transfer from generator to RETDF*: RCRAInfo is used to search for the RETDF to which the chemical waste was transferred to, i.e., the entity to which the broker shipped the waste. The search is based on the transfer purpose reported by the generator (i.e., transfer to a broker for energy recovery, for recycling, treatment, or disposal). Thus, this step provides the RCRA ID(s) for the RETDF(s). Although the RCRA ID for the RETDF is not included in Table A.1, this is used to connect with FRS (as described in step 4) and collect the FRS ID and TRIF ID for RETDF. Also, a generator-receiver flow consistency analysis is applied, as further explained in Section 3.3.3.

7. *Confirming identification of a RETDF*: In this step, the success of the search in step 6 is verified. The accomplishment of this step depends on whether the RETDF was found, the result of the flow consistency analysis, and the existence of a TRIF ID for a RETDF, i.e., the RETDF has reported to the TRI program at some point. Thus, records with a successful search pass to step 8, where they are merged with the second subset obtained in step 5.

8. *Identifying the most recent TRI report submitted by the RETDF*: If the RETDF did not report for the reference year 2017, the 2001–2016 TRI reporting years are used to search the most recent TRI report submitted by the RETDF because the RETDF might have reported during previous years. A RETDF stops reporting to the TRI program when it does not meet all three criteria: covered industry sectors (e.g., electric power generation), total employee equivalence is more than 20,000 annual "employee" hours, and TRI chemical activity quantities are above threshold levels in a given year (e.g., 0.1 g for dioxins at manufacturing activities) (U.S. Environmental Protection Agency, 2013). In this way, the most recent TRI report of the RETDF provides the most reasonable exposure scenario conditions for a concerning chemical.

9. *Gathering RETDF data reported to TRI*: The RETDF's reported release quantities and other information (e.g., location, IDs) are collected from the TRI database. Thus, the maximum amount of chemical

present during a reporting year at the RETDF is gathered in this step. This value is reported to the TRI as a range code to indicate the "maximum quantity of the chemical on-site at any one time" in a reporting year. RETDF chemical releases to environmental compartments are also gathered.

10. *Is the search successful?* Even if a RETDF reported to the TRI program, it might have reported other chemicals than the one being tracked. Thus, these records are excluded from the final dataset. In this way, the EoL dataset is obtained by applying the proposed framework.

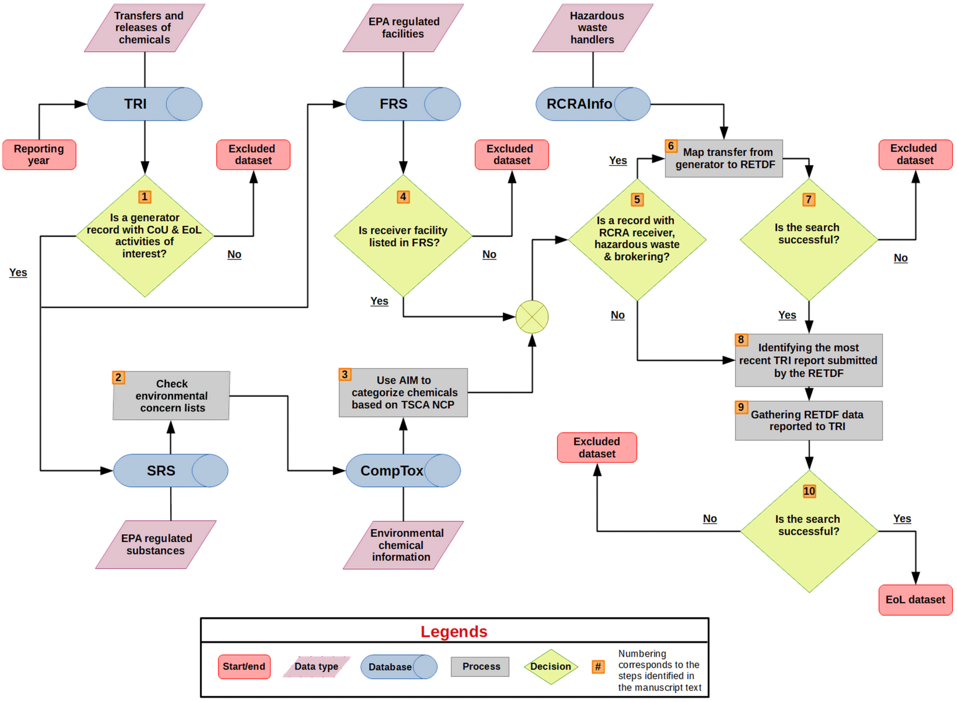


Figure 3.4: Data engineering and refinement process for transforming the TRI database into structures for the EoL dataset and tracking chemical flows at EoL stages.

Detailed versions of Figures 3.2 and 3.4 are shown in Appendix A to provide more technical aspects of the framework and the obtained generic EoL dataset.

3.3.3 *Assumptions and internal estimations to fill knowledge and data gaps*

The data sources considered for this study do not have enough information to cover all the needed aspects of tracking chemical flows at the EoL stage and knowing the potential chemical flow transferred, which may be released to the environment. For example, from the publicly-available data sources, it is not possible to gather information such as the total quantity of chemical present at a facility in a reporting year, chemical quantity that goes into an EoL activity (especially activity other than disposal), number of days that a chemical is present at the facility, and chemical release quantity from an EoL activity. Chemical process reports, life cycle inventories, and computer-aided process simulation can be used to track the concerning chemical, consistent with a "chemical process" systems engineering analysis (Smith et al., 2017). However, this section proposes a method for leveraging available data and screening tools to complete a chemical flow tracking at the EoL stage and estimate how much of a chemical transferred might be released from RETDFs, based on assumptions from available literature to estimate chemical releases based on emission factors, and some data entries like the maximum amount of chemical present at the RETDF during an EoL activity.

The framework follows the track of the chemical until the RETDF. As shown in Figure 3.5, a waste stream containing a chemical of interest may pass through a broker before being received by the RETDF. The framework considers up to two brokers between the generator and the RETDF, Broker 1 and Broker 2 (if applicable), as shown in Figure 3.5.Thus, if there is waste brokering, the Receiver represents Broker 1; otherwise, it represents the RETDF. An i data record is constituted by the quantity of chemical transferred by Generator ($D_{29}^{(i)}$), the maximum amount of chemical present at RETDF during a reporting year ($D_{54}^{(i)}$), the RETDF total generated waste ($D_{55}^{(i)}$), the environmental compartments ($D_{57}^{(i)}$). Furthermore, for record i, $R_1^{(i)}$, $R_2^{(i)}$, and $R_3^{(i)}$ are the total quantities received by Broker 1, Broker 2 (if applicable), and RETDF. In contrast, $T_1^{(i)}$ and $T_2^{(i)}$ are the total quantities shipped by Receiver and Broker 2. Note that $R_1^{(i)}$, $R_2^{(i)}$, $R_3^{(i)}$, $T_1^{(i)}$, and $T_2^{(i)}$ are not in the EoL dataset since these are values only used by the framework to track the chemicals, analyze the consistency of information, and make internal estimations to lead to the finally calculated values. Thus, when a material containing the chemical is tracked from the generator to RETDF (step 6 in Figure 3.4), a flow consistency analysis must be performed. As depicted in Figure 3.5, $D_{29}^{(i)}$ is contained in or equal to $R_1^{(i)}$ ($D_{29}^{(i)} \subseteq R_1^{(i)}$), $T_1^{(i)} \subseteq R_2^{(i)}$, and $T_2^{(i)} \subseteq R_3^{(i)}$ or $D_{29}^{(i)} \subseteq R_3^{(i)}$, then the consistency analysis is satisfied when $D_{29}^{(i)} \leq R_1^{(i)}$, $T_1^{(i)} \leq R_2^{(i)}$, and $T_2^{(i)} \leq R_3^{(i)}$ or $D_{29}^{(i)} \leq R_3^{(i)}$. Otherwise, it does not consider the tracking record.

Moreover, in this framework, it is assumed that a RETDF receives a chemical flow for (i) energy recovery, recycling, treatment (physical/chem-

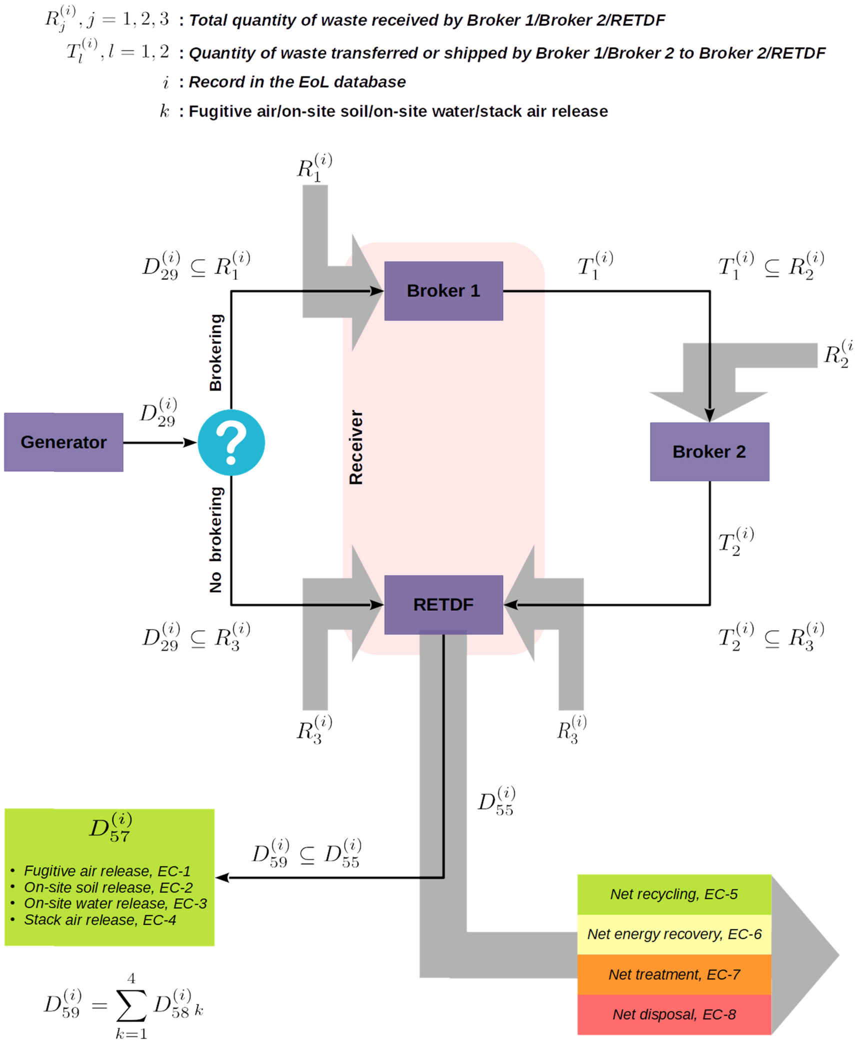


Figure 3.5: EoL supply and management chain entities for tracking a concerning chemical in transfer (D_{29}) from the Generator to EoL environmental compartment (D_{57}) chemical releases (D_{58}).

ical/biological), or disposal, and (ii) the recycled chemical is used as a product or incorporated in a valuable product. Thus, with data entries $D_{29}^{(i)}$, $D_{54}^{(i)}$, $D_{55}^{(i)}$, $D_{57}^{(i)}$, $D_{58}^{(i)}$, and $D_{59}^{(i)}$ as shown in Table A.2, the Generator indirectly releases the chemical to each environmental compartment (e.g., fugitive air release), during an EoL activity, and can be estimated using Equation 3.1. In other words, an indirect release quantity is the mass flow fraction of the chemical transferred off-site by the Generator ($D_{29}^{(i)}$) for EoL management, which may be potentially released during activities occurring at RETDFs. In Equation 3.1, $ER^{(i)}[D_{57\,k}^{(i)}]$ represents the chemical indirect release to the environmental compartments $D^{(i)}_{57\,k}$, and $EF^{(i)}_{k}$ are emission factor equivalents to the loss fraction used by the Chemical Screening Tool for Exposures and Environmental Releases (ChemSTEER), which means, "fraction of an amount of the chemical that is estimated to be released" (U.S. Environmental Protection Agency, 2015a). Also, $D^{(i)}_{29}$ is equivalent to the production volume used by ChemSTEER, which means "the overall 'volume' of the assessed chemical."

$$\begin{gathered} ER^{(i)}[D^{(i)}_{57\,k}] = \left\{ D^{(i)}_{29} \times EF^{(i)}_{k} = D^{(i)}_{29} \times \frac{D^{(i)}_{58\,k}}{\Delta AC^{(i)} + D^{(i)}_{55}}, k = 1,2,3,4 \right\| \\ \Delta AC^{(i)} \in [MAX(D^{(i)}_{59} - D^{(i)}_{55}, -MAX(QM^{(i)})), MAX(QM^{(i)})] \\ \left. \cap QM^{(i)} \in D^{(i)}_{54} \right\} \end{gathered} \tag{3.1}$$

Emission factors, like $EF^{(i)}_{k}$, relate the quantity of the chemical released to the environment divided by a unit weight, volume, or duration of the activity from which release the chemical occurs (Cheremisinoff, 2011). For a record i, the divisor represents that $Input + Generation = \Delta AC^{(i)} + D^{(i)}_{55}$ (i.e., an annual mass balance for a chemical at a RETDF, as demonstrated in Section A.2).$\Delta AC^{(i)}$ is the difference in the amount of chemical present at the RETDF between the end and the beginning of the reporting year, i.e., the annual increment in the amount of the chemical present at the RETDF. $D^{(i)}_{54}$ is a range code used by TRI to describe the maximum quantity of the chemical present on-site at a facility at any one time during a reporting year (U.S. Environmental Protection Agency, 2014b), while $QM^{(i)}$ is a value within the range described by $D^{(i)}_{54}$. Thus, if the range code is 07, the $QM^{(i)}$ is $4,535,924 - 22,679,618 kg$, then $MAX(QM^{(i)}) = 22,679,618 kg$. Also, by knowing $D^{(i)}_{54}$, $D^{(i)}_{55}$, and $D^{(i)}_{59}$, an interval of $\Delta AC^{(i)}$ values for the record i can be determined. Consequently, $\Delta AC^{(i)}$ is a range of values and in the simplest case, sets of uniform statistical distribution numbers can describe it. Table A.2 depicts an example for using Equation 3.1 to address estimations based on data entries for a record from the EoL dataset.

On the other hand, a "high-end scenario" for exposure risk evaluation (U.S. Environmental Protection Agency, 2018a), can be useful when only a few records in the EoL dataset give information about an EoL activity for a chemical of interest. This "high-end scenario" is given by D_{56}, which is the highest potential value of chemical release quantities from the RETDF to the environmental compartments, attributed to the Generator's transfer (D_{29}). Equation 3.2 shows the way to calculate D_{56}. An example is given below for a brokering scenario:

1. A Generator reported to the TRI program a chemical transfer of $500kg/yr$ ($D_{29}^{(i)}$) to a Broker 1 Receiver (or Receiver in Figure 3.5).
2. Broker 1 reported a transfer of $20kg/yr$ ($T_1^{(i)}$ in Figure 3.5) to a RETDF.
3. The RETDF reported to the TRI program total chemical releases to environmental compartments of $60kg/yr$ ($D_{59}^{(i)}$).
4. Even though the RETDF reported total chemical releases to the environmental compartment of $60kg/yr$, the highest value from those releases attributed to the amount taken from the Broker 1 is $20kg/yr$. Therefore, the highest value of RETDF environmental compartment chemical releases attributed to the Generator's transfer is the lowest reported amount of chemical transfers by the EoL supply and management chain entities.
5. Equation 3.2 computes the "high-end scenario" chemical flow ($D_{56}^{(i)}$). Thus, $D_{56}^{(i)}$ is equal to $MIN(500, 20, 60) = 20kg/yr$, i.e., $20kg/yr$ of the Generator's chemical transfer may be potentially released in this pathway.

$$D_{56}^{(i)} = \begin{cases} MIN(D_{29}^{(i)}, T_1^{(i)}, T_2^{(i)}, D_{59}^{(i)}), \textit{ if brokering} \\ MIN(D_{29}^{(i)}, D_{59}^{(i)}), \textit{ if no brokering} \end{cases} \tag{3.2}$$

On the other hand, since is the "high-end scenario," Equation 3.3 estimates the potential releases to the compartment $D_{57\,k}^{(i)}$ under this scenario. Thus, the "high-end scenario" release to the compartment $D_{57\,k}^{(i)}$, is given by multiplying $D_{56}^{(i)}$ by the ratio between the chemical released to the compartment reported by the RETDF ($D_{58\,k}^{(i)}$), and the sum of total releases ($D_{59}^{(i)}$).

$$ER^{(i)}[D_{57\,k}^{(i)}]_{worst-case} = D_{56}^{(i)} \times \frac{D_{58\,k}^{(i)}}{D_{59}^{(i)}} \tag{3.3}$$

3.4 CASE STUDY: METHYLENE CHLORIDE (MC)

In this section, a case study of MC (CAS 75-09-2), also known as dichloromethane, is used to demonstrate the usefulness of the proposed framework for tracking and analyzing chemical flows and estimating potential releases at the EoL stage. MC is a U.S. high production volume chemical (equal to or greater than 453,593kg/yr) (U.S. Environmental

Protection Agency, 2000). It has several uses such as solvent, component of adhesives, colorant, fragrance, flame retardant, food additive, and products designed for children. Moreover, this chemical is associated with hazard endpoints such as acute toxicity, irritation, liver toxicity, and neurotoxicity.

MC belongs to the following lists of substances of concern: (i) hazardous air pollutants, (ii) chemical water pollutants for which EPA has developed analytical test methods, (iii) contaminants subject to U.S. National Primary Drinking Water Regulation, (iv) hazardous wastes, and (v) TRI toxic chemicals. Furthermore, MC is a chemical that belongs to the TSCA inventory of chemical substances and was selected in 2016 as one of the first ten chemicals selected for risk evaluation. In the problem formulation document for MC, EPA found EoL activities such as industrial pretreatment, industrial wastewater treatment, and transfers to sewage treatment facilities. Therefore, this chemical is a relevant case study due to its environmental and health concerns, and the findings can be analyzed with those reported by TSCA (U.S. Environmental Protection Agency, 2018c).

The EoL dataset is used as the information source to support the analysis. The case study employs the following data entries (see Table A.1): the generator industry sector (D_{10}), the generator CoU (D_{28}), the quantity of chemical transferred by the generator (D_{29}), the EoL category under previous TSCA reports (D_{33}) (U.S. Environmental Protection Agency, 2016g), the EoL activity under the waste management hierarchy category (D_{34}) (U.S. Environmental Protection Agency, 2015d), the RETDF industry sector (D_{53}), the maximum amount of chemical present at RETDF (D_{54}), the total chemical generated as waste by the RETDF (D_{55}), and the RETDF chemical releases (D_{58}) to the environmental compartment $D57$. As highlighted in Section 3.3.3, the entry $\Delta AC^{(i)}$ is a range of values describing the annual increment in the amount of the chemical present at the RETDF. This case study assumes that $\Delta AC^{(i)}$ is uniformly distributed, i.e., all the values in the interval have an equal probability of occurrence. The case study development uses the Python script shown in Section A.5 and follows these steps:

1. All the records with information on MC transfers are taken from the EoL dataset.

2. A sample size of 100 is used to obtain 100 random values of $\Delta AC^{(i)}$ for each record i. This sample size is a trade-off between having enough data to achieve statistical significance and computing power limitations.

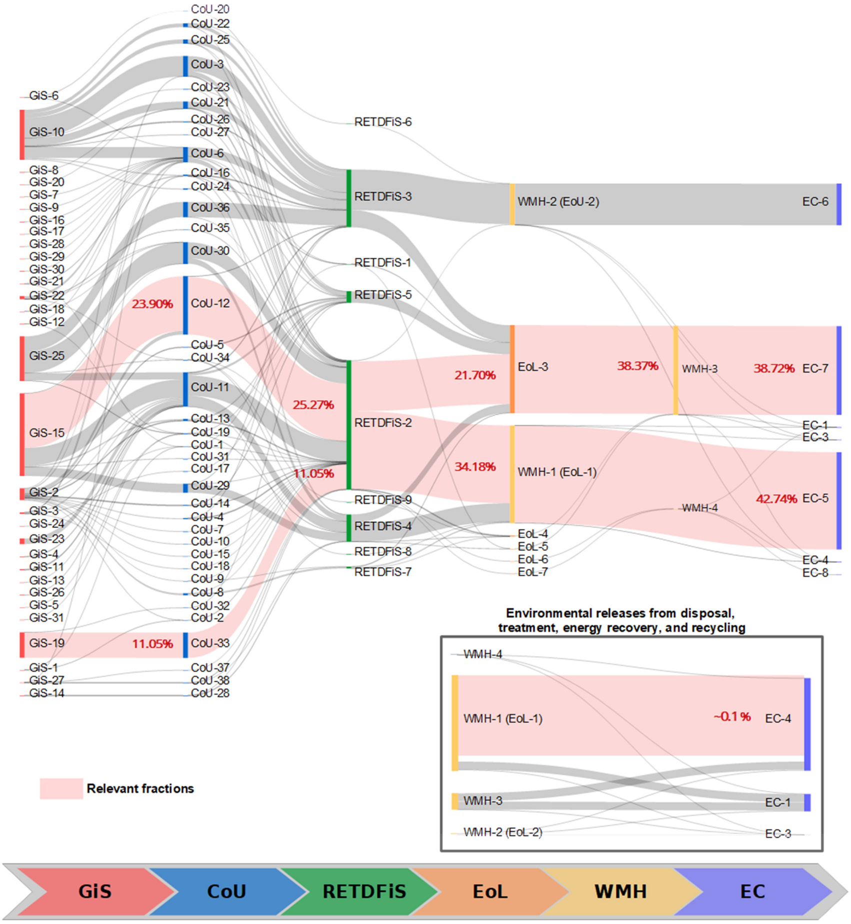
3. For each record i, the 100 values of $\Delta AC^{(i)}$ are used to compute 100 different values of $ER^{(i)}[D^{(i)}_{57\,k}]$ using Equation 3.1. Then, these are

used to calculate the arithmetic average of $ER^{(i)}[D^{(i)}_{57\,k}]$, i.e., $\overline{ER^{(i)}[D^{(i)}_{57\,k}]}$, to represent the chemical flow transferred by each generator which may be potentially released to each environmental compartment $D^{(i)}_{57\,k}$ from EoL activities at each RETDF.

4. The following dataset enables a performance analysis of the chemical flow:
   4.1. Quantity of chemical transferred by the generator or D_{29}.
   4.2. Generator industry sectors (GiS) or D_{10} as shown in Table A.6.
   4.3. CoU of chemical at generator facility or D_{28} as shown in Table A.7.
   4.4. RETDF industry sectors (RETDFiS) or D_{53} as shown in Table A.3.
   4.5. The EoL activities are grouped into three distinct categories, which are in separate columns in the EoL dataset. The EoL activities are listed in Table A.8 for more details. The case study development uses the categorization based on TSCA reports (D_{33}) and the EPA waste management activity hierarchy (D_{34}). To differentiate both categorizations, the first one is simply labeled as EoL and the second one as waste management hierarchy (WMH), as shown in Table A.8. In addition, recycling and energy recovery are in both classification systems; therefore, for energy recovery EoL-2 = WMH-2 and for recycling EoL-1 = WMH-1. Thus, both categorizations for the EoL activities are represented in the chemical flow analysis. Table S8 shows a third category consistent with the TRI Program is included since these are more specific than the others and aids in the discussion of the results.
   4.6. Environmental compartment (EC). EC includes those contained in D_{57}. In addition, net recycling, net energy recovery, net treatment, and net disposal are also included in EC due to mass conservation consistency.
   4.7. The $\overline{ER^{(i)}[D^{(i)}_{57\,k}]}$ is calculated for each record i in step 3

This analysis tracks MC in waste streams generated at industrial manufacture and processing stages to handling at off-site facilities and estimate releases from EoL activities. The tracking of MC is presented in Figure 3.6, which is a 6-level Sankey structure depicting the fraction (relative to the total flow transferred) for allocating MC in a potential pathway throughout the elements conforming to the EoL

Sankey structure. The identification labels for GiSs, CoUs, RETDFiSs, EoL, WMHs, and ECs, can be found in Tables A.3-A.8.



Legend for relevant elements in the 6-level Sankey structure

Label	Label	Label	Label
GiS-10	Hazardous waste treatment and disposal	WMH-4	Disposal
GiS-15	Medicinal and botanical manufacturing	EoL-3	Incineration
GiS-19	Other basic inorganic chemical manufacturing	EC-1	Fugitive air release
GiS-25	Pharmaceutical preparation manufacturing	EC-2	On-site soil release
CoU-12	Chemical processing and manufacturing aid	EC-3	On-site surface water release
CoU-33	As a reactant, as a formulation component, and produce the chemical	EC-4	Stack air release
RETDFiS-2	Hazardous waste treatment and disposal	EC-5	Net recycling
RETDFiS-3	Cement manufacturing	EC-6	Net energy recovery
WMH-1 and EoL-1	Recycling	EC-7	Net treatment
WMH-2 and EoL-2	Energy recovery	EC-8	Net disposal
WMH-3	Treatment		

Figure 3.6: Chemical flow diagram of methylene chloride during end-of-life (EoL) management scenarios in the U.S. using a 6-level Sankey structure. Corresponding categories are the generator industry sector (GiS), the conditions of use (CoU), the RETDF industry sector (RETDFiS), EoL activity (EoL), the waste management hierarchy (WMH), and the environmental compartment (EC).

5. All values of chemical releases to the environmental compartment $D_{57}^{(i)}$ (see Table A.5) can be used to measure central tendency (mean and median) and the descriptive statistics (percentiles and standard deviation) of the 100 annual releases or $ER^{(i)}[D_{57}^{(i)}]$ calculated in step 3 for each record i. The calculation of these statistical values should be performed when a wide variety of data is handled to support risk assessment (U.S. Environmental Protection Agency, 1991).

6. A statistical distribution of MC annual releases from EoL activities (classified based on WMH) can be determined with all values of $ER^{(i)}[D_{57\,k}^{(i)}]$ for all records. For this case study, TRI range codes are selected to group annual release values to provide information about the statistical distribution.

3.5 RESULTS AND DISCUSSION

Tracking the chemical flows aims to determine the material allocation at the EoL stage, under which a chemical exposure is foreseen, thereby finding so-called generic exposure scenarios for use when assessing chemical risk. Also, from a circular economy perspective, tracking the chemical flows aids the building of a material flow analysis to determine a material closed-loop considering the chemical risk associated with the substance under analysis, the industry sectors that may supply the chemical (GiS), and the industry sectors that may acquire it under certain CoUs. Thus, based on the methodology described in Section 3.3 and the case study specific considerations described in Section 3.4, the chemical flow analysis of MC at the off-site EoL stage is performed consistent with the information retrieved from several databases, as shown in Figure 3.2 and transformed as described in Figure 3.4. After identifying the MC EoL activities, the first step is determining the MC flow that may be potentially released. Thus, Section 3.5.1 presents the chemical flow analysis for identifying the possible EoL exposure scenarios for MC. Section 3.5.2 focuses on the variability and distribution of the estimations of releases from the EoL exposure scenarios. Finally, Section 3.5.3 discusses how the framework and chemical flow tracking might extend the borders for risk assessment and release allocation.

3.5.1 *Chemical flow tracking: EoL exposure scenario identification*

Figure 3.6 represents the material flow analysis for MC at the EoL stage when tracking a chemical of interest using the proposed framework.

The red pathways and their percentage values are the highest MC flow fraction relative to the total MC flow transferred. From a descriptive point of view, this figure indicates how the methodology can support risk evaluation by the abstraction of potential exposure pathways at the EoL stage. Also, chemical flows tracking aids to identify systematically the relevant generator industry sector (GiS) of a chemical and how it is used (CoU) by a GiS. Furthermore, the tracking provides information about how the waste containing the chemical of interest is distributed among various locations, managed, and released into the environmental compartments (i.e., RETDFiS, EoL, WMH, and EC).

Table 3.2: Flow allocation between 6-level Sankey structure based on the 12,321,001 kg/yr of MC reported to the TRI Program in 2017. The source and target names are in Figure 3.6. The fractions were obtained by means of flow tracking using the EoL data engineering and saved in the EoL dataset.

Source		Target		Fraction on	Flow
Label	Name	Label	Name	Sankey [%]	[kg/yr]
GiS-15	Medicinal and botanical manufacturing	CoU-12	Chemical processing and manufacturing aid	23.90	2,944,719
GiS-19	Other basic inorganic chemical manufacturing	CoU-33	As a reactant, as a formulation component, and produce the chemical	11.05	1,361,471
CoU-12	Chemical processing and manufacturing aid	RETDFiS-2	Hazardous waste treatment and disposal	25.27	3,113,517
CoU-33	As a reactant, as a formulation component, and produce the chemical	RETDFiS-2	Hazardous waste treatment and disposal	11.05	1,361,471
RETDFiS-2	Hazardous waste treatment and disposal	WMH-1 and EoL-1	Recycling	34.18	4,211,318
RETDFiS-2	Hazardous waste treatment and disposal	EoL-3	Incineration	21.70	2,673,657
EoL-3	Incineration	WMH-3	Treatment	38.37	4,727,568
WMH-1 and EoL-1	Recycling	EC-4	Stack air release	∼0.1	12,321
WMH-1 and EoL-1	Recycling	EC-5	Net recycling	42.74	5,265,996
WMH-3	Treatment	EC-7	Net treatment	38.72	4,770,692

The chi-square hypothesis test with significance level 0.05 and 3-degrees of freedom is used to compare the MC fraction transferred off-site to recycling, energy recovery, treatment, and disposal calculated using both the constructed EoL dataset and the 2017 TRI data for MC available online (see section A.4). According to the test, the hypothesis that the "EoL management allocation for MC calculated based on both sources are equal" is not rejected (see Section A.4). That means, after tracking and finding the MC EoL exposure pathways, its flow distribution trend is sustained without a statistical difference. According to the TRI Program, in 2017, the total quantity of MC transferred for further management was 12,321,001 kg/yr (U.S. Environmental Protection Agency, 2019b). Thus, we use this latter value to contrast what may be found in the case that the MC flows reported to the TRI Program are completely tracked regarding the fraction on the Sankey diagram. Table 3.2 presents the results of this comparison for some of the most relevant fractions on the 6-level Sankey structure calculated by using the EoL dataset.

For MC, it is possible to see that the generator industry sectors that mainly contribute to the off-site transfers of MC are hazardous waste treatment and disposal (GiS-10), medicinal and botanical manufacturing (GiS-15), other basic inorganic chemical manufacturing (GiS-19), and pharmaceutical preparation manufacturing (GiS-25). Also, the main CoUs under which the waste containing MC is generated are the use as chemical processing and manufacturing aid (CoU-12) and as a reactant, as a formulation component, and to produce the chemical (CoU-33). Thus, 23.90% or 2,944,719 kg/yr of the total MC transferred comes from the medicinal and botanical manufacturing sector because of the use as a chemical processing and manufacturing aid, while 11.05% or 1,361,471 kg/yr of MC comes from the basic inorganic chemical manufacturing industry due to the use as a reactant, as a formulation component, and to produce the chemical. Furthermore, according to the EoL dataset, the most significant RETDF industry sector is hazardous waste treatment and disposal (RETDFiS-2). Since MC-containing waste is a hazardous waste, it requires special handling under RCRA Subtitle C requirements. In addition, 11.05% or 1,361,471 kg/yr of the total transferred waste (having MC) is received by hazardous waste treatment and disposal industry sector from activities associated with CoU-33 and shipped by GiS-19, while 25.27% or 3,113,517 kg/yr comes from CoU-15 and shipped by GiS-12 to the same sector or RETDFiS-2.

In general, energy recovery (WMH-2 and EoL-2), treatment (WMH-3) for destruction via incineration (EoL-3), and recycling (WMH-1 and EoL-1) by solvent/organics recovery (see Table A.8) are noticeable EoL activities for MC. As shown in Figure 3.6, energy recovery from waste containing MC is widely employed by the cement manufacturing sector (RETDFiS-3). In contrast, incineration and recycling are utilized by the hazardous waste treatment and disposal sector (RETDFiS-2). Moreover, 34.18% or 4,211,318 kg/yr and 21.70% or 2,673,657 kg/yr of the MC transfers are received by RETDFiS-2 for further management using recycling and incineration (incineration or EoL-3 represents around 99.10% of treatment activities, WMH-3), respectively. Furthermore, data from the EoL dataset would inform exposure pathways excluded from MC risk evaluation due to a previous lack of reasonably available information, such as recycling and solidification/stabilization (included in EoL-5 and WMH-3 as shown in Table A.8). Fortunately, the MC flow tracking and analysis show that these pathways might result in potential fugitive air releases (EC-1), on-site surface water discharges (EC-3), and stack air emissions (EC-4), as shown in Figure 3.6. Therefore, this framework may help to identify EoL exposure scenarios not being evaluated and where there is not a well-established analytical tool for risk evaluation under other EPA regulations (U.S. Environmental Protection Agency, 2018c).

The box at the lower right in Fig. 6 shows that recycling, energy recovery, treatment, and disposal activities affect air due to fugitive releases (EC-1) and stack releases (EC-4). However, the recycling activities have a considerable potential contribution to these compartments, especially to stack releases (0.1% of the total MC flow transferred). If the distribution fraction is conserved, this means that around 12,321 kg/yr of transferred MC might have been potentially released from stacks to the atmosphere during recycling activities. Under RCRA Subpart O, hazardous waste incinerators are required to achieve a minimum destruction and removal efficiency of 99.99% of the principal organic hazardous constituents (U.S. Environmental Protection Agency, 1999c). Similarly, under RCRA Subpart H, energy recovery devices such as boilers and industrial furnaces must satisfy the same destruction and removal efficiency requirement for principal organic hazardous constituents (U.S. Environmental Protection Agency, 1999b). Table 3.2 shows that 38.37% or 4,727,568 kg/yr of the MC transferred was treated via incineration. Therefore, if the equipment for incineration satisfies the destruction requirement, the stack emission from this EoL activity might have been around 473 kg/yr, which is just below the stack air releases from recycling activities and 2,846kg/yr obtained by the tracking considering that incineration represents around 99.10% of treatment activities and 0.02% was the fraction of MC transfer released from the stack during treatment, according to the framework estimations. Thus, this destruction requirement might explain why MC stack releases from recycling are potentially higher than the releases from energy recovery and treatment. However, this aspect makes clear the need to go further and analyze in more depth the RETDFs and understand how the chemical flows perform inside those facilities. That might be done by incorporating additional information from other sources utilizing data engineering to enhance the allocation of releases and incorporate more details about pollution abatement activities.

3.5.2 *Environmental release estimations: variability and distribution*

After identifying the exposure pathways and scenarios at the EoL stage by tracking chemical flows, the developed methodology can help to analyze the potential flow released under the assumptions explained in Section 3.3.3 and considering statistical analysis. Some MC exposure scenarios and pathways meet well-established requirements under regulations other than TSCA. For example, the U.S. Clean Air Act gives specific requirements for the control of hazardous air pollutants during incineration. However, in this contribution, all pathways are examined, since the same exceptions do not apply for all chemicals analyzed by TSCA or even the analysis may be interesting for stakeholders outside EPA.

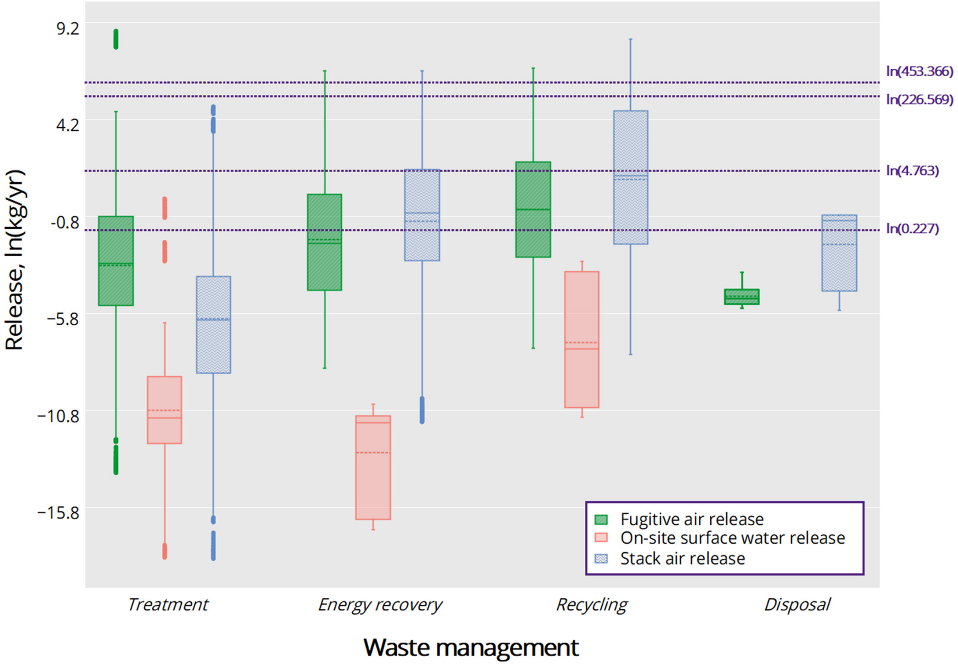


Figure 3.7: Variability of the MC annual releases during EoL management activities. A sample size of 100 was used to obtain 100 different values for the annual increment in the amount of chemical present at each RETDF ($\Delta AC^{(i)}$).

Figure 3.7 describes the mean, median, percentiles (25th and 75th, the median is the 50th percentile), and standard deviations of MC annual releases from EoL activities based on waste management hierarchy classification (see Table A.8). The mean values are the dotted lines inside the boxes, which aim to represent the average of the release values. In contrast, the median values are the horizontal solid lines inside the boxes that aim to show the data sample midpoints of the potential releases. The purple horizontal lines are the logarithmic values for 0.227, 4.762, 226.569, and 453.366 kg/yr, which are the limits of the TRI range codes. These values are references in the box plot for the annual releases calculated in step 3 of the case study. In addition, this figure shows that both the mean and median of stack air releases from recycling are between 0.227 and 4.763 kg/yr. Consequently, more than 50% of the data of stack air releases from recycling activities do not exceed 4.763 kg/yr, and the highest release value is above 453.366 kg/yr. Also, the statistical distributions depicted by histograms in Figure 3.8 show the range of potential annual releases from EoL activities, according to the waste management hierarchy(recycling > energy recovery > treatment > disposal). These two figures illustrate how the framework can support the analysis of the variability of results. However, the current analysis does not incorporate the uncertainty associated with the temporal correlation and reliability of the data, which will be part of future development, as explained in Section 3.3.

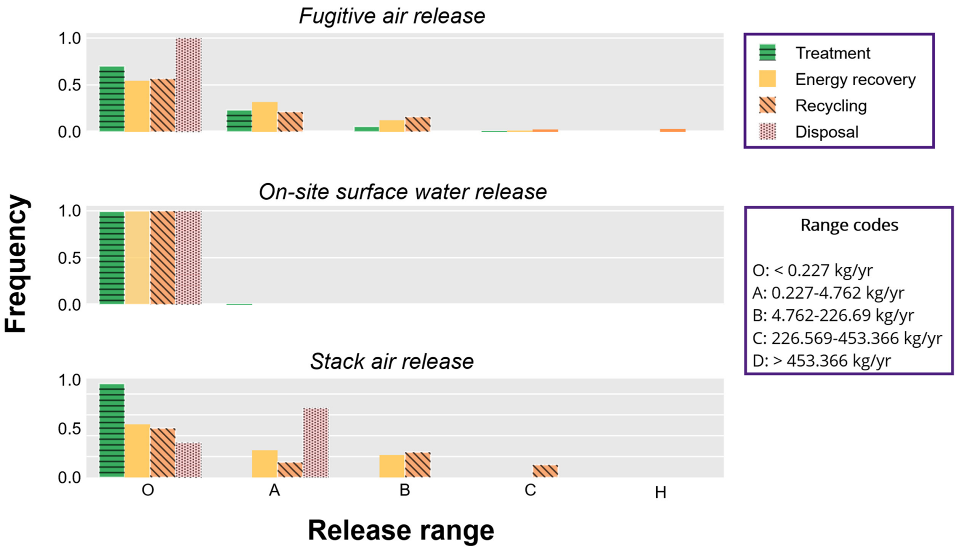


Figure 3.8: The statistical distribution of MC annual release during EoL management. 100 is the sample size to obtain values for the annual increment in the amount of chemical present at each RETDF ($\Delta AC^{(i)}$).

For this case study, a standardized TRI Program consisting of release range codes A, B, and C is employed to show the results of the statistical distribution analysis. The range codes A, B, and C are release values in the following intervals 0.227-4.762 kg/yr, 4.762-226.569 kg/yr, and 226.569-453.366 kg/yr, respectively. Notice that TRI does not use range codes when emissions are less than 0.227 kg/yr and greater than 453.366 kg/yr.

Therefore, O is used here when releases are less than 0.227kg/yr, and H is for releases that are greater than 453.366 kg/yr. For instance, if a released quantity is 100 kg/yr, the range code is B due to the value being within 4.762-226.569 kg/yr. Figure 3.8 depicts that most of the stack air releases from recycling are in the intervals O and A, above 40% and 10%, respectively. Furthermore, it is possible to see from both figures that the total air releases in all the scenarios exceed the average discharges of MC to surface water, which can be because MC is a very volatile organic compound according to EPA classification. Moreover, treatment, energy recovery, and recycling scenarios could potentially result in higher releases than disposal activities, such as class I underground injection wells and hazardous waste landfills. The above may be explained by the fact that MC is a hazardous substance under RCRA; therefore, disposal activities must satisfy the technical requirements described in the RCRA Subtitle C, thus reducing the risk. Also, consistent with both figures, notice that MC releases to soil are not expected from any EoL scenario, and surface water releases may be roughly zero from all scenarios. At room temperature, MC readily partitions from water into the air and does not sorb to soil (Agency for Toxic Substances and Disease Registry, 1990). Therefore, volatilization to air during waste management activities should be expected, especially from recycling activities, which may be in release ranges B and C. In general, the obtained MC chemical flows can be used to assess the exposure of different receptors to MC at RETDFs and their surroundings, for instance, fugitive air releases for occupational receptors, stack air emissions for the general population and terrestrial organisms, and surface water discharge for aquatic organisms.

3.5.3 *Industry sector and facility linkage: indirect effects of chemical flow transfers*

EPA addressed the risk evaluation for MC basing its decisions on the weight of scientific evidence within no more than 3 years. Hence, comprehensive information about the industry sectors across the MC life cycle was collected one by one by considering the reasonably available information (U.S. Environmental Protection Agency, 2020). The framework dentifies those industry sectors at EoL stage (i.e., RETDFiS) and is able to link EoL flows (e.g., recycling) being transferred to industry sectors at upstream life cycle stages (e.g., manufacturing), by considering the chemical uses and activities generating the MC-containing waste and the EoL activities used by the RETDFiS, as presented in Section 3.5.1. The above is essential to run a cradle-to-gate analysis of the potential indirect effect of the chemical flows being managed off-site the waste generation source instead of on-site. From the circular economy point of view, this industry sector linkage may be beneficial to close the recycling loop, considering potential risks and regulatory constraints (U.S. Environmental Protection Agency, 2016e).

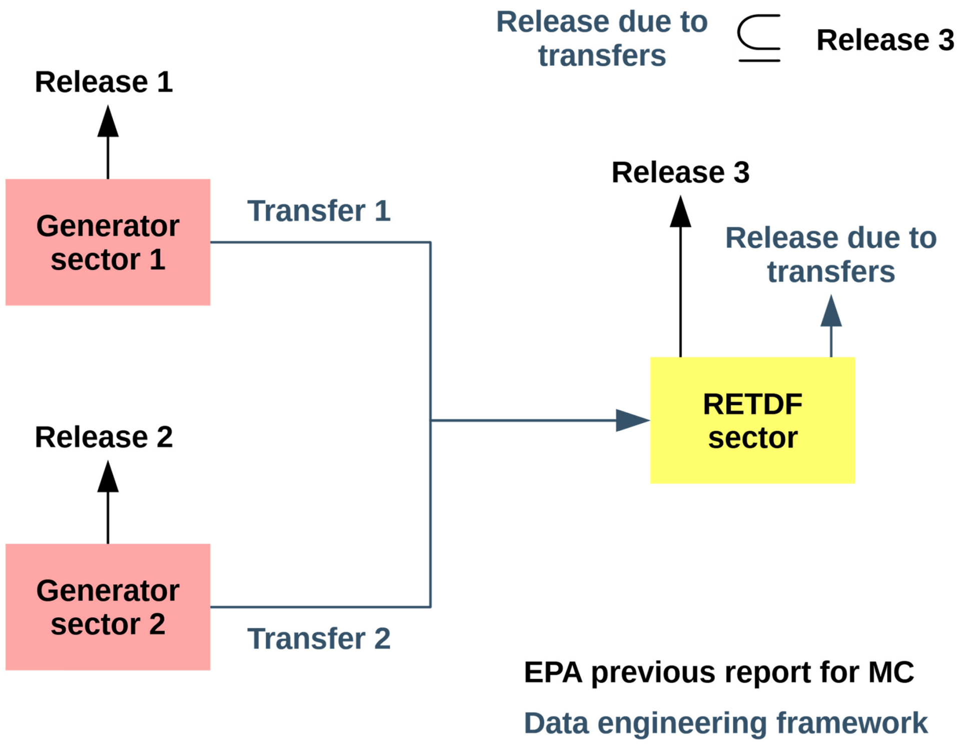


Figure 3.9: Industry sector linkage and the potential indirect effect of the chemical flow transfers.

The Pollutant Release and Transfer Registers databases do not directly provide information on what happens to these chemical flows once they are in the RETDF. Even though TRI is the Pollutant Release and Transfer Register with more data granularity (U.S. Environmental Protection Agency, 2014b), it only gives this information for transfers to sewage treatment facilities (U.S. Environmental Protection Agency, 2013). To overcome this obstacle, the risk evaluation for MC uses the Exposure and Fate Assessment Screening Tool to estimate the chemical concentration in water. However, this tool requires input parameters associated with the releases to water and the estimated removal by the treatment operations to calculate the remaining chemical flow and subsequent chemical concentration in water (U.S. Environmental Protection Agency, 1999a). Instead, the data engineering framework identifies the RETDFiS and reporting facility, which enables to leverage the facilitylevel information to estimate a release range for the amount of chemical transferred that might be released to the environment from the RETDF, as presented in Section 3.5.2. As Figure 3.9 depicts, the chemical flow tracking represents a screening tool to extend the risk evaluation to include and allocate the indirect effects (e.g., exposure scenarios, releases) of chemical transfers. For instance, as presented in Table 3.2, 12,321 kg/yr 12,321 kg/yr of the MC transferred in 2017 might have been released from stack due to recycling activities. By identifying the RETDF and RETDFiS, the framework can enhance the release estimations for going into the RETDFs and searching for additional data of the pollution abatement operations, narrowing the range of emissions estimates regardless the hands-on engineering knowledge and avoiding extensive calculating ()Smith2019.

3.6 CONCLUSIONS

The framework created can aid with tracking chemicals in waste streams generated at industrial activities and handled at off-site facilities, identifying possible exposure scenarios for a concerning chemical. The framework automatically identifies whether a chemical is of concern under a EPA environmental regulation, thereby helping to rule out exposure pathways that may not be of interest under a TSCA analysis. Indeed, the framework builds the EoL dataset, whose machine-readable structure allows for making an annual material balance rapidly for a chemical at RETDF and subsequently calculating emission factors to environmental compartments. Thus, the framework enables estimation of how much of a transferred chemical may be potentially released from the RETDF during an EoL activity. However, a more detailed analysis of the RETDFs will help understand what is occurring inside these facilities. Thus, incorporating additional information about pollution abatement activities at RETDFs from other sources utilizing data engineering might enhance release allocation. Furthermore, the framework considers a "high-end scenario" based on the quantity of transferred chemical that may be potentially released and quantifies central tendency and standard deviation for handling a wide variety of data. Nevertheless, future work will deal with the uncertainty of the measurement based on the reliability and temporal correlation of the data. Currently, this work considers up to two brokers between the generator and the RETDF. However, it is expected to increase the number of brokers considered in future developments.

The case study results for MC show that using the EoL dataset records, the fraction of MC transferred to each element of the EPA’s waste management hierarchy preference is statistically like those obtained using the information from the 2017 TRI data for MC available online. The results also demonstrate that the framework could analyze EoL exposure scenarios like solidification/stabilization and recycling via solvent/organic recovery, which were excluded from MC risk evaluation due to a lack of reasonably available information. Nevertheless, there are some current limitations of this framework; such as when the chemical of interest is not included on the TRI list of toxic chemicals, whether the RETDFs do not appear in the FRS, do not report to the TRI Program, or otherwise do not report the chemical of interest.

Thus, the framework can aid in the rapid identification of hazards for new chemicals and occupational exposure assessment. This is due to the inclusion of the TSCA NCP chemical categories and knowing the management activity at the EoL stage. As a result, it supplies an inventory of emissions from EoL activities for life cycle analysis and risk assessment. Also, the use of data-reconciliation and data analytics approaches to categorize and combine knowledge of a chemical of interest derived from an association with similar chemicals, hazards, or EoL activities will be explored to increase the applicability of this work on future data-driven models to predict chemical releases and exposure. Thus, risk evaluation can be addressed for chemicals that do not report to existing regulatory environmental programs or are new to the U.S. market. Finally, this approach aims to support the rapid TSCA risk evaluations of high priority chemicals.

DISCLAIMER

The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the EPA. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the EPA. The EPA does not endorse any commercial products, services, or enterprises.

ACKNOWLEDGMENTS

This research was supported in by appointment for Jose D. Hernandez-Betancur to the Research Participation Program at the Center for Environmental Solutions and Emergency Response, Office of Research and Development, EPA, administered by the Oak Ridge Institute for Science and Education through an Interagency Agreement between the U.S. Department of Energy and the EPA. The authors express their sincere gratitude and appreciation to Sandra D. Gaona, Mitchell Sumner, and Steve DeVito

of the EPA's Toxics Release Inventory Program, for their critical review of draft versions of our manuscript, their recommendations, and insightful discussions.

BIBLIOGRAPHY

Agency for Toxic Substances and Disease Registry. (1990). *Public Health Statement Public Health Statement - Chloromethane* (tech. rep. November). https://www.atsdr.cdc.gov/ToxProfiles/tp106-c1-b.pdf

Banerjee, S., Ramachandran, G., Vadali, M., & Sahmel, J. (2014). Bayesian Hierarchical Framework for Occupational Hygiene Decision Making. *The Annals of Occupational Hygiene*. https://doi.org/10.1093/annhyg/meu060

Bernas, M. (2013). Assessment and Risk Reduction in Lymphedema. *Seminars in Oncology Nursing*, *29*(1), 12–19. https://doi.org/10.1016/j.soncn.2012.11.003

Bodar, C., Spijker, J., Lijzen, J., Waaijers-van der Loop, S., Luit, R., Heugens, E., Janssen, M., Wassenaar, P., & Traas, T. (2018). Risk management of hazardous substances in a circular economy. *Journal of Environmental Management*, *212*, 108–114. https://doi.org/10.1016/j.jenvman.2018.02.014

Cashman, S. A., Meyer, D. E., Edelen, A. N., Ingwersen, W. W., Abraham, J. P., Barrett, W. M., Gonzalez, M. A., Randall, P. M., Ruiz-Mercado, G., & Smith, R. L. (2016). Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing. *Environmental Science & Technology*, *50*(17), 9013–9025. https://doi.org/10.1021/acs.est.6b02160

Cheremisinoff, N. P. (2011). Pollution Management and Responsible Care. In *Waste* (pp. 487–502). Elsevier. https://doi.org/10.1016/B978-0-12-381475-3.10031-2

Clift, R., Doig, A., & Finnveden, G. (2000). The Application of Life Cycle Assessment to Integrated Solid Waste Management. *Process Safety and Environmental Protection*, *78*(4), 279–287. https://doi.org/10.1205/095758200530790

Conley, J. F. (2011). Estimation of exposure to toxic releases using spatial interaction modeling. *International Journal of Health Geographics*, *10*(1), 20. https://doi.org/10.1186/1476-072X-10-20

Delmaar, J. E., Park, M., & Engelen, v. J. G. M. (2005). *ConsExpo 4.0 - Consumer Exposure and Uptake Models - Program Manual* (tech. rep.).

Edelen, A., & Ingwersen, W. (2016). *Guidance on Data Quality Assessment for Life Cycle Inventory Data* (tech. rep.). https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NRMRL%5C&dirEntryId=321834

Gaona, S. D., Keenan, C., Vallet, C., Reichle, L., & DeVito, S. C. (2020a). The Environmental Impact of Pollution Prevention and Other Sustainable Development Strategies Implemented by the Automotive Manufacturing Industry. In *Green energy to sustainability* (pp. 111–136). Wiley. https://doi.org/10.1002/9781119152057.ch6

Gaona, S. D., Pepping, T., Keenan, C., & DeVito, S. C. (2020). The Environmental Impact of Pollution Prevention, Sustainable Energy Generation, and Other Sustainable Development Strategies Implemented by the Food Manufacturing Sector. In *Green energy to sustainability* (pp. 539–567). Wiley. https://doi.org/10.1002/9781119152057.ch22

Hernandez-Betancur, J. D., & Ruiz-Mercado, G. J. (2019). Sustainability indicators for end-of-life chemical releases and potential exposure. *Current Opinion in Chemical Engineering*, *26*, 157–163. https://doi.org/10.1016/j.coche.2019.09.004

Jones, R. (2010). TSCA New Chemicals Program (NCP) Chemical Categories. https://www.epa.gov/sites/production/files/2014-10/documents/ncp_chemical_categories_august_2010_version_0.pdf [Online; accessed 2020-05-13]

Lahl, U., & Zeschmar-Lahl, B. (2013). Risk based management of chemicals and products in a circular economy at a global scale (risk cycle), extended producer responsibility and EU legislation. *Environmental Sciences Europe*, *25*(1), 3. https://doi.org/10.1186/2190-4715-25-3

Li, H., Nitivattananon, V., & Li, P. (2015). Municipal solid waste management health risk assessment from air emissions for China by applying life cycle analysis. *Waste Management & Research: The Journal for a Sustainable Circular Economy*, *33*(5), 401–409. https://doi.org/10.1177/0734242X15580191

Linkov, I., Trump, B. D., Wender, B. A., Seager, T. P., Kennedy, A. J., & Keisler, J. M. (2017). Integrate life-cycle assessment and risk analysis results, not methods. *Nature Nanotechnology*, *12*(8), 740–743. https://doi.org/10.1038/nnano.2017.152

McNally, K., Warren, N., Fransman, W., Entink, R. K., Schinkel, J., Van Tongeren, M., Cherrie, J. W., Kromhout, H., Schneider, T., & Tielemans, E. (2014). Advanced REACH Tool: A Bayesian Model for Occupational Exposure Assessment. *The Annals of Occupational Hygiene*. https://doi.org/10.1093/annhyg/meu017

Meyer, D. E., Bailin, S. C., Vallero, D., Egeghy, P. P., Liu, S. V., & Cohen Hubal, E. A. (2020). Enhancing life cycle chemical exposure assessment through ontology modeling. *Science of The Total Environment*, *712*, 136263. https://doi.org/10.1016/j.scitotenv.2019.136263

Meyer, D. E., Mittal, V. K., Ingwersen, W. W., Ruiz-Mercado, G. J., Barrett, W. M., Gonzalez, M. A., Abraham, J. P., & Smith, R. L. (2019). Purpose-Driven Reconciliation of Approaches to Estimate Chemical Releases. *ACS Sustainable Chemistry & Engineering*, *7*(1), 1260–1270. https://doi.org/10.1021/acssuschemeng.8b04923

National Research Council. (2014). *A Framework to Guide Selection of Chemical Alternatives*. National Academies Press. https://doi.org/10.17226/18872

Nost, E., Rosenfeld, H., Vincent, K., Moore, S. A., & Roth, R. E. (2017). HazMatMapper: an online and interactive geographic visualization tool for exploring transnational flows of hazardous waste and environmental justice. *Journal of Maps*, *13*(1), 14–23. https://doi.org/10.1080/17445647.2017.1282384

Organization for Economic Co-operation and Development. (2015). Introduction to Pollutant Release and Transfer Registers (PRTRs). Retrieved July 20, 2020, from https://www.oecd.org/env/ehs/pollutant-release-transfer-register/introductionto-pollutant-release-and-transfer-registers.htm

Organization for Economic Co-operation and Development. (2017). Application manual of OECD QSAR Toolbox v.4. Retrieved May 13, 2020, from https://www.oecd.org/chemicalsafety/risk-assessment/TB4_Application_manual_F1.compressed.pdf

Pizzol, L., Hristozov, D., Zabeo, A., Basei, G., Wohlleben, W., Koivisto, A. J., Jensen, K. A., Fransman, W., Stone, V., & Marcomini, A. (2019). SUNDS probabilistic human health risk assessment methodology and its application to organic pigment used in the automotive industry. *NanoImpact*, *13*, 26–36. https://doi.org/10.1016/j.impact.2018.12.001

Rosenbaum, R. K., Margni, M., & Jolliet, O. (2007). A flexible matrix algebra framework for the multimedia multipathway modeling of emission to impacts. *Environment International*, *33*(5), 624–634. https://doi.org/10.1016/j.envint.2007.01.004

Smith, R. L., Ruiz-Mercado, G. J., Meyer, D. E., Gonzalez, M. A., Abraham, J. P., Barrett, W. M., & Randall, P. M. (2017). Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling. *ACS Sustainable Chemistry & Engineering*, *5*(5), 3786–3794. https://doi.org/10.1021/acssuschemeng.6b02724

Toropov, A. A., Toropova, A. P., Mukhamedzhanova, D. V., & Gutman, I. (2005). Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). *Indian Journal of Chemistry - Section A Inorganic, Physical, Theoretical and Analytical Chemistry*, *44*(8), 1545–1552.

United States Census Bureau. (2017). North American Industry Classification System. Retrieved October 7, 2019, from https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf

U.S. Environmental Protection Agency. (1991). *Chemical Engineering Branch Manual for the Preparation of Engineering Assessments* (tech. rep.). https://nepis.epa.gov/Exe/ZyNET.exe/P10000VS.TXT?ZyActionD=ZyDocument%5C&Client=EPA%5C&Index=1991+Thru+1994%5C&Docs=%5C&Query=%5C&Time=%5C&EndTime=%5C&SearchMethod=1%5C&TocRestrict=n%5C&Toc=%5C&TocEntry=%5C&QField=%5C&QFieldYear=%5C&QFieldMonth=%5C&QFieldDay=%5C&IntQFieldOp=0%5C&ExtQFieldOp=0%5C&XmlQuery=

U.S. Environmental Protection Agency. (1992). Guidelines for Exposure Assessment. https://www.epa.gov/sites/production/files/2014-11/documents/frmwrk_cum_risk_assmnt.pdf

U.S. Environmental Protection Agency. (1999a). *Exposure and Fate Assessment Screening Tool (E-FAST) Beta Version Documentation Manual* (tech. rep. No. 68). https://p2infohouse.org/ref/18/17836.pdf

U.S. Environmental Protection Agency. (1999b). *RCRA , Superfund & EPCRA Hotline Training Module Introduction to : Boilers and Industrial Furnaces - EPA Document # EPA530-R-99-042* (tech. rep. October 1999). https://www.epa.gov/sites/production/files/2015-07/documents/bif.pdf

U.S. Environmental Protection Agency. (1999c). *RCRA , Superfund & EPCRA Hotline Training Module Introduction to : Hazardous Waste Incinerators - EPA Document # EPA530-R-99-052* (tech. rep. October 1999). https://www.epa.gov/sites/production/files/2017-02/documents/incin.pdf

U.S. Environmental Protection Agency. (2000). EPA: High Production Volume List. Retrieved April 14, 2020, from https://comptox.epa.gov/dashboard/chemical_lists/EPAHPV

U.S. Environmental Protection Agency. (2003a). 3MRA Methodology. Retrieved November 25, 2019, from https://www.epa.gov/ceam/3mra

U.S. Environmental Protection Agency. (2003b). *Framework for Cumulative Risk Assessment - EPA Document # EPA/630/P-02/001F* (tech. rep. May). https://www.epa.gov/sites/production/files/2014-11/documents/frmwrk_cum_risk_assmnt.pdf

U.S. Environmental Protection Agency. (2003c). *Generic Ecological Assessment Endpoints (GEAEs) for Ecological Risk Assessment - EPA Document # EPA/630/P-02/004F* (tech. rep. October). https://doi.org/EPA/630/P-02/004F

U.S. Environmental Protection Agency. (2004). Facility Registry Service (FRS). Retrieved September 4, 2019, from https://www.epa.gov/frs

U.S. Environmental Protection Agency. (2013). TRI Basic Plus Data Files Guides. Retrieved October 22, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-plus-data-files-guides

U.S. Environmental Protection Agency. (2014a). Guidance on Life-Cycle Thinking and Its Role in Environmental Decision Making. Retrieved May 12, 2020, from https://www.michaeldbaker.com/wp-content/uploads/2014/03/Guidance-on-Life-Cycle-Thinking-031014.pdf

U.S. Environmental Protection Agency. (2014b). TRI Around the World. Retrieved January 17, 2022, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-around-world

U.S. Environmental Protection Agency. (2015a). ChemSTEER User Guide: Chemical Screening Tool for Exposures and Environmental Releases. Retrieved May 11, 2020, from https://www.epa.gov/sites/production/files/2015-05/documents/user_guide.pdf

U.S. Environmental Protection Agency. (2015b). How to Access the TSCA Inventory. Retrieved July 29, 2020, from https://www.epa.gov/tsca-inventory/how-access-tsca-inventory%5C#download

U.S. Environmental Protection Agency. (2015c). RCRAInfo Search User Guide. Retrieved September 4, 2019, from https://www.epa.gov/enviro/rcrainfo-search-user-guide

U.S. Environmental Protection Agency. (2015d). Waste Management Hierarchy and Homeland Security Incidents. Retrieved April 21, 2020, from https://www.epa.gov/homeland-security-waste/waste-management-hierarchy-and-homeland-security-incidents

U.S. Environmental Protection Agency. (2016a). Contaminant Candidate List 4-CCL 4. Retrieved July 29, 2020, from https://www.epa.gov/ccl/contaminant-candidate-list-4-ccl-4-0

U.S. Environmental Protection Agency. (2016b). Defining Hazardous Waste: Listed, Characteristic and Mixed Radiological Wastes. Retrieved July 29, 2020, from https://www.epa.gov/hw/defining-hazardous-waste-listed-characteristic-and-mixed-radiological-wastes

U.S. Environmental Protection Agency. (2016c). Initial List of Hazardous Air Pollutants with Modifications. Retrieved July 29, 2020, from https://www.epa.gov/haps/initial-list-hazardous-air-pollutants-modifications

U.S. Environmental Protection Agency. (2016d). Instructions for the 2016 TSCA Chemical Data Reporting. Retrieved May 11, 2020, from https://www.epa.gov/chemical-data-reporting/instructions-reporting-2016-tsca-chemical-data-reporting

U.S. Environmental Protection Agency. (2016e). Legitimate Hazardous Waste Recycling Versus Sham Recycling. Retrieved September 16, 2020, from https://www.epa.gov/hw/legitimate-hazardous-waste-recycling-versus-sham-recycling%5C#:$%5Csim$:text=EPA's%20definition%20of%20legitimate%20recycling,Regulations%20(CFR)%20section%20260.43.%5C&text=Legitimate%20recycling%20must%20involve%20a,intermediate%20of%20the%20recycling%20process.

U.S. Environmental Protection Agency. (2016f). National Primary Drinking Water Regulations. Retrieved July 29, 2020, from https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations

U.S. Environmental Protection Agency. (2016g). Risk Evaluations for Existing Chemicals under TSCA. Retrieved January 14, 2022, from https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/risk-evaluations-existing-chemicals-under-tsca

U.S. Environmental Protection Agency. (2016h). Substance Registry Services (SRS). Retrieved October 22, 2020, from https://ofmpub.epa.gov/sor_internet/registry/substreg/LandingPage.do

U.S. Environmental Protection Agency. (2016i). Toxic and Priority Pollutants Under the Clean Water Act. Retrieved July 29, 2020, from https://19january2017snapshot.epa.gov/eg/toxic-and-priority-pollutants-under-clean-water-act_.html%5C#priority

U.S. Environmental Protection Agency. (2016j). Toxic Substances Control Act. Retrieved July 28, 2020, from https://uscode.house.gov/view.xhtml?path=/prelim@title15/chapter53%5C&edition=prelim

U.S. Environmental Protection Agency. (2016k). Using Predictive Methods to Assess Hazard under TSCA. Retrieved October 5, 2019, from https://www.epa.gov/tsca-screening-tools/using-predictive-methods-assess-hazard-under-tsca

U.S. Environmental Protection Agency. (2017a). About the TSCA Chemical Substance Inventory. Retrieved September 3, 2019, from https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory

U.S. Environmental Protection Agency. (2017b). Procedures for Chemical Risk Evaluation Under the Amended Toxic Substances Control Act - EPA Document #EPA-HQ-OPPT-2016-0654-0108. https://doi.org/10.1016/0196-335x(80)90058-8

U.S. Environmental Protection Agency. (2017c). Standardized Emission and Waste Inventories (StEWI). Retrieved July 27, 2020, from https://github.com/USEPA/standardizedinventories

U.S. Environmental Protection Agency. (2018a). Draft Risk Evaluation for N-Methylpyrrolidone (2-Pyrrolidinone, 1-Methyl-) (NMP) EPA Document # EPA-740-R1–8009. Retrieved May 16, 2020, from https://www.epa.gov/sites/production/files/2019-11/documents/1_draft_risk_evaluation_for_n-methylpyrrolidone_110419_public.pdf

U.S. Environmental Protection Agency. (2018b). *Problem Formulation of the Risk Evaluation for Carbon Tetrachloride (Methane, Tetrachloro-) - EPA Document # EPA-740-R1-7020* (tech. rep. May).

U.S. Environmental Protection Agency. (2018c). *Problem Formulation of the Risk Evaluation for Methylene Chloride (Dichloromethane, DCM) - EPA Document # EPA 740-R1-7016* (tech. rep. May). https://www.epa.gov/sites/production/files/2018-06/documents/mecl_problem_formulation_05-31-18.pdf

U.S. Environmental Protection Agency. (2019a). *Biosolids Biennial Review Reporting Period 2016–2017 EPA Document # EPA-822R18003* (tech. rep.). https://www.epa.gov/sites/production/files/2019-06/documents/2016-2017-biosolids-biennial-review.pdf

U.S. Environmental Protection Agency. (2019b). TRI Explorer (2018 Updated Dataset (released April 2020)) [Internet database]. Retrieved May 14, 2020, from https://enviro.epa.gov/triexplorer/tri_release.chemical

U.S. Environmental Protection Agency. (2020). *Risk Evaluation for Methylene Chloride (Dichloromethane, DCM) - EPA Document # EPA-740-R1-8010* (tech. rep. June). https://www.epa.gov/sites/production/files/2020-06/documents/1_mecl_risk_evaluation_final.pdf

Whittaker, M. H. (2015). Risk Assessment and Alternatives Assessment: Comparing Two Methodologies. *Risk Analysis*, *35*(12), 2129–2136. https://doi.org/10.1111/risa.12549

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., & Richard, A. M. (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics*, *9*(1), 61. https://doi.org/10.1186/s13321-017-0247-6

4

A DATA ENGINEERING FRAMEWORK FOR ON-SITE END-OF-LIFE INDUSTRIAL OPERATIONS

ABSTRACT

Sustainable initiatives for converting end-of-life (EoL) material flows into feedstocks would make a crucial contribution towards protecting our environment and mitigating the negative impacts of anthropogenic activities. Chemical flow analysis enables decision-makers to identify potential environmental releases and exposure pathways at the EoL stage and, therefore, improves the estimation of chemical exposure. Certain industrial facilities apply on-site pollution abatement operations, thereby constituting nodes of the chemical EoL management chain that can be evaluated and improved to enable greater circularity of materials. This work enhances and extends a recently published EoL data engineering framework by using publicly-available databases, data-driven models, and analytic hierarchy approaches to track chemicals, estimate releases, and potential exposure pathways at on-site industrial pollution management operations. The extended framework develops pollution abatement unit (PAU) technologies and estimates their efficiencies, chemical releases, exposure media, operating expenses, and capital expenditures. Relevant case studies based on the food and pharmaceutical industry sectors illustrate the application of the framework for chemical flow allocation and analysis of a chemical of concern and the benefits of integrating and extending the framework with data-driven and multi-criteria decision-making models. The results show how the enhanced framework designs and evaluates PAU technology systems for managing EoL chemical flows and provides release inventories and pathways for conducting chemical risk evaluation and exposure assessment of potential on-site EoL scenarios.

Keywords: Data engineering; Pollution abatement unit; Chemical releases; Chemical flow analysis

RESUMEN

Las iniciativas sostenibles para convertir los flujos de materiales al final de su vida útil (EoL, por sus siglas en inglés) en materias primas supondrían una contribución crucial para proteger nuestro medio ambiente y mitigar los impactos negativos de las actividades antropogénicas. El análisis de los flujos químicos permite a los responsables de la toma de decisiones identificar las posibles emisiones al medio ambiente y las vías de exposición en la fase de EoL y, por tanto, mejora la estimación de la exposición química. Algunas instalaciones industriales aplican operaciones de reducción de la contaminación in situ, por lo que constituyen nodos de la cadena de gestión de EoL química que pueden evaluarse y mejorarse para permitir una mayor circularidad de los materiales. Este trabajo mejora y amplía un marco de ingeniería de datos de EoL publicado recientemente mediante el uso de bases de datos disponibles públicamente, modelos basados en datos y enfoques de jerarquía analítica para realizar un seguimiento de las sustancias químicas, estimar las emisiones y las posibles vías de exposición en las operaciones de gestión de la contaminación industrial in situ. El marco ampliado desarrolla tecnologías de unidades de reducción de la contaminación (PAU, por sus siglas en inglés) y estima su eficiencia, las emisiones químicas, los medios de exposición, los gastos de funcionamiento y los gastos de capital. Los estudios de casos pertinentes basados en los sectores de la industria alimentaria y farmacéutica ilustran la aplicación del marco para la asignación del flujo de sustancias químicas y el análisis de una sustancia química de preocupación, así como las ventajas de integrar y ampliar el marco con modelos de toma de decisiones basados en datos y criterios múltiples. Los resultados muestran cómo el marco mejorado diseña y evalúa los sistemas de tecnología PAU para la gestión de los flujos de sustancias químicas de EoL y proporciona inventarios de liberación y vías para llevar a cabo la evaluación del riesgo químico y la evaluación de la exposición de posibles escenarios de EoL in situ.

Palabras clave: Ingeniería de datos; Unidad de reducción de la contaminación; Emisiones químicas; Análisis del flujo químico

4.1 INTRODUCTION

Every year, humankind and the environment are exposed to chemical substances. Numerous chemicals may present a risk of injury to health or the environment during production, processing, distribution in commerce, use, or end-of-life (EoL) (European Chemicals Agency, 2019; U.S. Environmental Protection Agency, 2017c). Therefore, regulations exist to track and manage chemicals through their life cycle, evaluating potential impacts, and imposing restrictions if needed (Bodar et al., 2018; National Research Council, 2014). Chemical risk assessment supports the selection of safer-profile chemicals and regulatory decision-making to protect human health and the environment (Bernas, 2013; Whittaker, 2015). However, conducting risk evaluation is a time-consuming and challenging task, especially at the EoL stage, due to the extensive data requirements, data scarcity, problem comprehensiveness, proper reporting, traceability, and epistemic uncertainty to describe the risk (Ragas, 2011). The data gap is especially noteworthy when seeking to prioritize thousands of chemicals based on risk (e.g., the U.S. Toxics Substances Control Act inventory lists about 41,000 commercially-active and non-confidential chemicals (U.S. Environmental Protection Agency, 2017a)).

Using publicly-available information, Hernandez-Betancur, Ruiz-Mercado, et al. (2020) developed a novel EoL data engineering framework for tracking chemicals, estimating releases, and identifying potential exposure scenarios in EoL flows generated at industrial processes and handled at off-site facilities located across the U.S. However, becase of the use of facility-level information for calculating the emission factors, the framework may not allocate releases considering the underlying characteristics of on-site EoL activities and pollution abatement units (PAUs). For instance, the efficiency of a PAU for recovering a chemical of concern from an EoL stream can be essential to allocate chemical flows and exposure downstream of the PAU.

In the literature, it is possible to find methodologies that can be incorporated into the EoL data engineering framework to enhance chemical flow allocation. For example, engineering process design and modeling have represented a toolkit to provide data for life cycle inventory and integrate environmental, health, and safety considerations (Righi et al., 2018; Sugiyama et al., 2008). These approaches can describe a wide variety of PAUs such as boilers, aerobic/anaerobic digesters, wet/dry scrubbers, pH neutralizers, distillers, absorbers, adsorbers, strippers, and biofilters (Baquerizo et al., 2007; Bojarski et al., 2008; Eberle et al., 2017; Jiménez-González et al., 2000; Zhang & Guo, 2013). Likewise, Process System Engineering offers an approach for dealing with complex systems such as solvent recovery and air pollution treatment processes (Cavanagh et al., 2014; Chea et al., 2019; Guerras & Martín, 2019b). Although the above methodologies provide detailed information on the physical, chemical, and biological phenomena underlying each PAU, they require case-by-case development, hands-on application of engineering knowledge, and extensive calculations that can result in a disadvantage compared to the rapid screening provided by the EoL data engineering framework (Smith et al., 2019).

In contrast, data-driven modeling can help to streamline the chemical flow allocation and to overcome the case-by-case limitation (Cashman et al., 2016; Meyer et al., 2019). For example, data mining has proved to reduce the cost associated with screening life cycle assessment (Sundaravaradan et al., 2011), while machine learning addresses data deficiencies (Song, 2019; Zhu et al., 2020). The data-driven method relies on the company data, which can offer some benefits in terms of more realistic insights (Li et al., 2018). However, it can present limitations because of flow allocation and coverage of reported substances, which lead to data gaps (Smith et al., 2017). The performance of these models depends on the specific dataset domain. Diverse sources can supply the datasets for data-driven modeling. Although some databases like EXIOBASE are free to use, others are licensed and proprietary (GreenDelta, 2013), e.g., ecoinvent, which is widely used worldwide for life cycle inventory and impact assessment (Vélez-Henao et al., 2020). Nevertheless, their methodologies focus on relating products and services to cradle-to-grave environmental impacts instead of chemical flow tracking (Frischknecht et al., 2007).

This contribution describes the addition of new data sources, data-driven and multi-criteria decision-making (MCDM) models to enhance and extend the EoL data engineering framework to track chemicals, estimate releases, and potential exposure pathways at facility on-site PAUs. This work has four key features to provide relevant data and PAU techno-economic information into the framework. First, collecting information regarding PAU technologies such as the removal efficiency related to substances of interest and the predominant phase of EoL flow managed by the PAU (e.g., liquid waste). Second, leveraging the data itself and literature for filling data gaps and allocating chemical flows. Third, transforming the information into a machine-readable structure for PAU design, evaluation, and selection using data-driven and MCDM models. Fourth, incorporating the created features into the framework developed by Hernandez-Betancur, Ruiz-Mercado, et al. (2020) to include inside-facility allocation, new data sources, and tools for decision-making. The enhanced framework can offer insights for selecting and recommending PAU technologies to handle chemicals of concern in EoL streams at the early engineering design step to achieve a sustainable process and a circular life cycle (Hassim, 2016).

Furthermore, the enhanced framework may be used as a first layer to provide alternatives for building optimization approaches for pollution abatement systems and networks to deal with chemical pollutants (Chea et al., 2020; Guerras & Martín, 2019a; Yenkie et al., 2019).

4.2 METHODOLOGY

For a chemical of concern c in an EoL input flow to a PAU (F_{input}), the framework allocates the chemical into the waste/environmental release flow ($F_{waste/release}^{(c)}$), the remaining chemical flow leaving the PAU, i.e., effluent ($F_{effluent}^{(c)}$), and determines the chemical flow being destroyed ($F_{destroyed}^{(c)}$), including converted and degraded), removed ($F_{removed}^{(c)}$), or recycled ($F_{recycled}^{(c)}$) as well as the chemical fugitive air release ($F_{fugitive}^{(c)}$). This allocation considers the PAU effect on the chemical like thermal (Lee et al., 1986; National Research Council, 2000; Saxena & Jotshi, 1996), chemical (Fanning, 2000; Huang et al., 1993; Wang et al., 2005), biological (Darvin & Serageldin, 2003; Kaur, 2017; Tay & Zhang, 1999), and physical (Jaeger Products Inc, 2010; Kreith et al., 2001; U.S. Environmental Protection Agency, 2002) treatments, and energy (Energy and Environmtental Analysis Inc., 2005; Mantus, 1992; Ottoboni et al., 1998) and material recovery processes (Bascone et al., 2016; Chmielewski et al., 1997; Hansen et al., 1991; Mular et al., 2002; Shin et al., 2009; Smallwood, 2002; U.S. Environmental Protection Agency, 1978). Additionally, the framework incorporates the predominant phase for F_{input} (α_{input}) because is a key PAU technology selection criterion.

As presented in Figure 4.1, the framework employs an input-output model representation for each PAU. This model must satisfy the annual material balance for a chemical of concern c in Equation 4.1.

$$\begin{gathered}
F_{input}^{(c)} = \eta^{(c)} \times F_{input}^{(c)} + F_{fugitive}^{(c)} + F_{waste/release}^{(c)} + F_{effluent}^{(c)} \\
F_{input}^{(c)} = w^{(c)} \times F_{input} \\
F_{fugitive}^{(c)} = (1 - \eta^{(c)}) \times \beta^{(c)} \times F_{input}^{(c)} \\
\eta^{(c)} \times F_{input}^{(c)} = \begin{cases} F_{recycled}^{(c)}, \textit{ if recycling} \\ F_{removed}^{(c)}, \textit{ if physical treatment} \\ F_{destroyed}^{(c)}, \textit{ if otherwise} \end{cases}
\end{gathered} \tag{4.1}$$

$w^{(c)}$ represents the input concentration of the chemical in F_input, $F_{input}^{(c)}$ the chemical input flow, $\beta^{(c)}$ the emission factor for $F_{fugitive}^{(c)}$, and $\eta^{(c)}$ the PAU abatement efficiency relative to the chemical, i.e., the degree to which the PAU destroys, removes, or recovers it. The framework tracks the chemical regardless of other incidentals that might be generated during the EoL activity, as the case for combustion, chemical, and biological operations that may generate undesired substances.$\beta^{(c)}$ is calculated as described in Hernandez-Betancur, Ruiz-Mercado, et al. (2020), which uses facility-level information for estimating the environmental chemical releases.

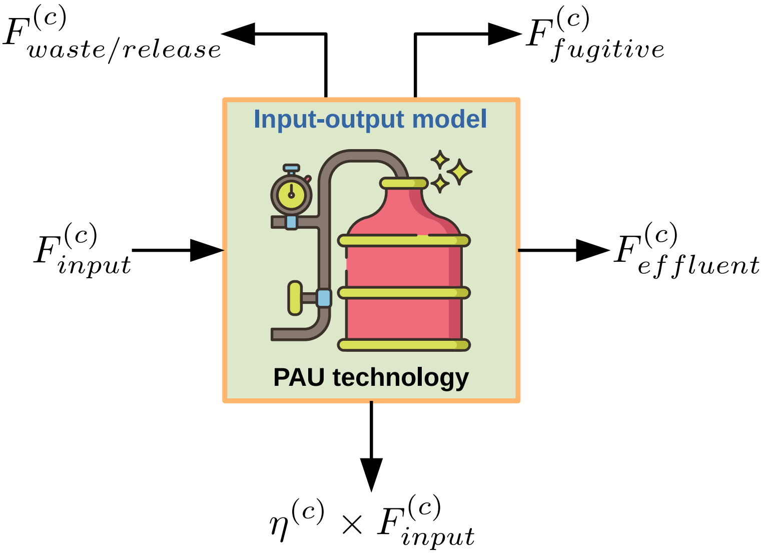


Figure 4.1: Generic PAU diagram for tracking chemical flows based on the PAU effect on a chemical of concern. The framework allocates output flows into $F^{(c)}_{waste/release}$, $F^{(c)}_{effluent}$, $F^{(c)}_{fugitive}$, and $F^{(c)}_{destroyed}$, $F^{(c)}_{removed}$, or $F^{(c)}_{recycled}$.

The framework focuses on recycling, energy recovery, and treatment activities. Also, the framework enables exchanging and integrating information from siloed publicly-available databases and transforms information into a machine-readable structure for future automatization, thereby creating the PAU dataset. A GitHub Repository named PAU4Chem has the Python scripts that build the PAU dataset (Hernandez-Betancur, Martin, et al., 2020b). The PAU dataset information allows input-output modeling to track and allocate chemicals of concern. Moreover, the PAU dataset has chemical unit prices ($UP^{(c)}$) and EoL activity capital expenditures (CAPEX) and operating expenses (OPEX). Hence, the generic framework can provide more realistic insights into the relationship between PAU technologies, expenses, and chemical allocation.

In order to develop the PAU dataset, the framework integrates the following databases: Toxics Release Inventory (TRI) (U.S. Environmental Protection Agency, 2014), Chemical Data Reporting (CDR) (U.S. Environmental Protection Agency, 2016), Facility Registry Service (FRS) (U.S. Environmental Protection Agency, 2004), Statistics of U.S. Businesses (SUSB) (U.S. Census Bureau, 2007, 2011), Annual Survey of Manufactures (ASM) (U.S. Census Bureau, 2009), and Pollution Abatement Costs and Expenditures (PACE) Survey (U.S. Environmental Protection Agency, 2017b). Additionally, it uses the North American Industry Classification System (NAICS) structure to cluster the TRI reporting facility into industry sectors (ISs) to connect the TRI information to the SUSB, ASM, and PACE Survey (U.S. Census Bureau, 2017b). Three modules constitute the PAU dataset: Technologies, OPEX & CAPEX, and ChemPrices. Sections 4.2.1, 4.2.2, and 4.2.3 depict how the framework effectively gathers, cleans, transforms, and integrates qualitative and quantitative information from multiple data sources to build the three modules. Section 4.2.4 shows how to use the PAU dataset modules for data-driven modeling and incorporating MCDM to predict EoL management and chemical flow analysis (CFA).

4.2.1 *Data engineering for collecting PAU information: Technology module*

Figure 4.2 illustrates the steps for building the PAU dataset – Technology module. The framework harnesses the TRI Program (from 1987 to 2018) as this module backbone due to its comprehension and data availability and granularity that enables gathering information for the PAU technologies (U.S. Environmental Protection Agency, 2014).

In step one, PAU technologies and uses/activities for a given chemical c are collected. For treatments, PAU sequences, codes describing α_{input}, $w^{(c)}$ and $\eta^{(c)}$ value ranges, and $\eta^{(c)}$ estimated values are taken. However, $w^{(c)}$ and $\eta^{(c)}$ estimated values were reported from 1987 to 2004 (U.S. Environmental Protection Agency, 2013). If treatment activities for the chemical c were submitted by a facility, typical $w^{(c)}$ and α_{input} are estimated for recycling and energy recovery activities. Otherwise, as NAICS uses a hierarchical structure organized from 2-digit NAICS codes (less specific) to 6-digit NAICS codes (more specific) (U.S. Census Bureau, 2017a), this structure supports estimating $w^{(c)}$ and α_{input}. Moreover, higher weight is given for typical α_{input}. Records with no estimated $w^{(c)}$ and α_{input} are dropped (step two).

In step three, $\eta^{(c)}$ is estimated for recycling and energy recovery activities. Due to energy recovery and incineration activities are combustion operations, incineration activities from TRI are leveraged to obtain a $\eta^{(c)}$ value. Using the NAICS structure as in step one, $\eta^{(c)}$ is assigned. However, if $\eta^{(c)}$ cannot be estimated, this is obtained by considering the threshold limits under hazardous waste and hazardous air pollutant regulations (U.S. Environmental Protection Agency, 1999, 2006).

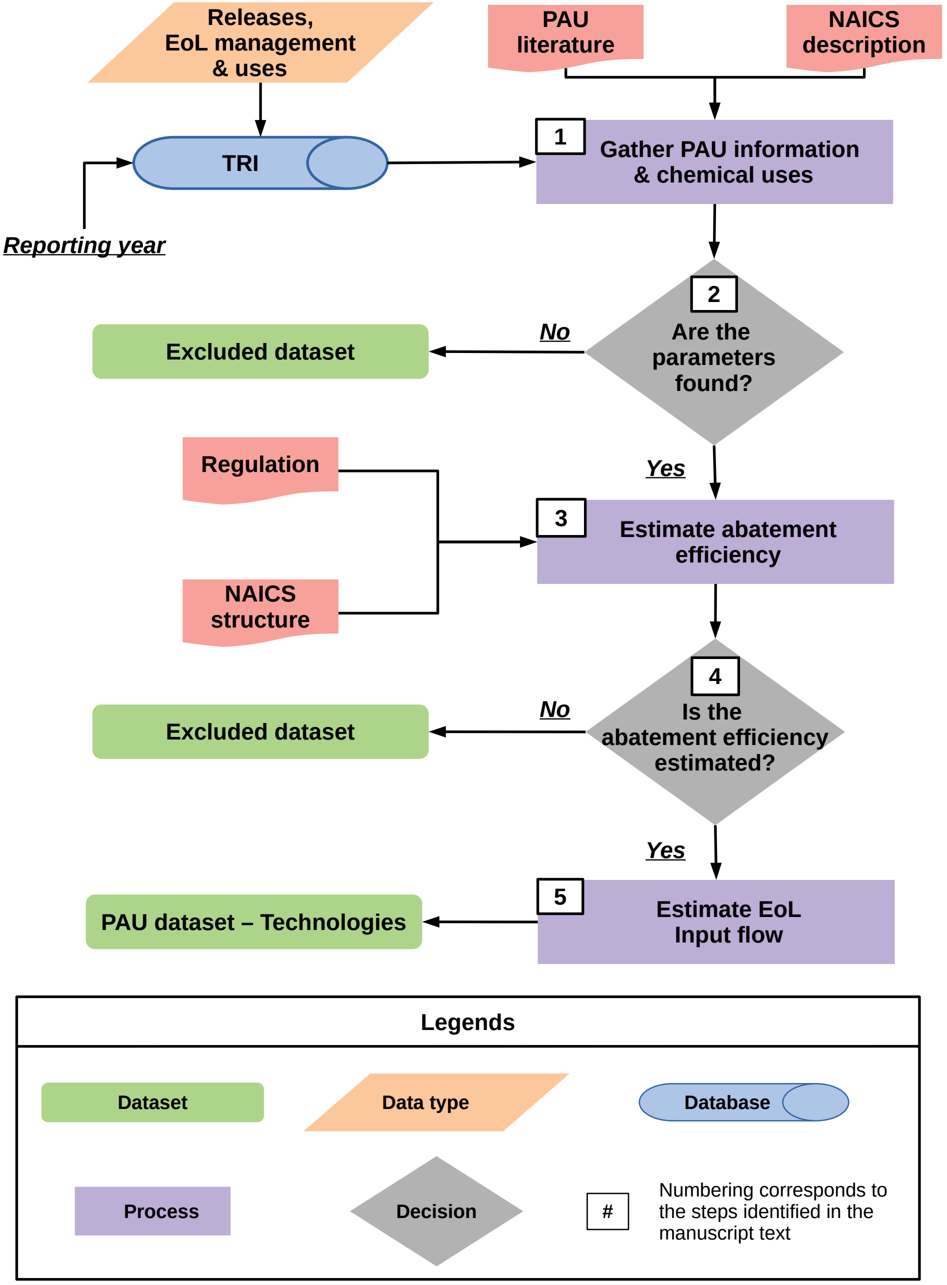


Figure 4.2: Data engineering for transforming information from TRI database into structures for the PAU dataset – Technologies and CFA.

For recycling activities, the EoL chemical flows reported by each facility are used to estimate a $\eta^{(c)}$ potential value. As depicted in Figure 4.3, an EoL chemical flow may have several potential pathways like transfers to sewage treatment plants. The exact pathway a chemical follows in a facility is uncertain. To overcome this uncertainty, for each facility, the non-zero flows for a chemical c are combined obtaining a set of scenarios Θ. Each scenario $\theta_i \in \Theta$ has a possible input flow to recycling $F^{(c)}_{input,\theta_i}$. Using $F^{(c)}_{input,\theta_i}$ and the reported $F^{(c)}_{recycled}$, $\eta^{(c)}_{\theta_i}$ is calculated. Assuming a recycling operation is used to obtain high cost-effectiveness, Equation 4.2 takes the upper value of $\eta^{(c)}_{\theta_i}$ as $\eta^{(c)}$, if this value is not an outlier and its coefficient of variation (CV) is less than 1 (low variance), ensuring a narrow value range for $\eta^{(c)}$. If the above two conditions are not satisfied for a facility, a $\eta^{(c)}$ value is assigned leveraging the NAICS structure again. The assignment considers facilities having the same recycling PAU technology type.

$$\eta^{(c)} = \underset{\theta_i \in \Theta}{MAX}(\eta^{(c)}_{\theta_i}) = \underset{\theta_i \in \Theta}{MAX}\left(\frac{F^{(c)}_{recycled}}{F^{(c)}_{input,\theta_i}} \times 100\right) \tag{4.2}$$

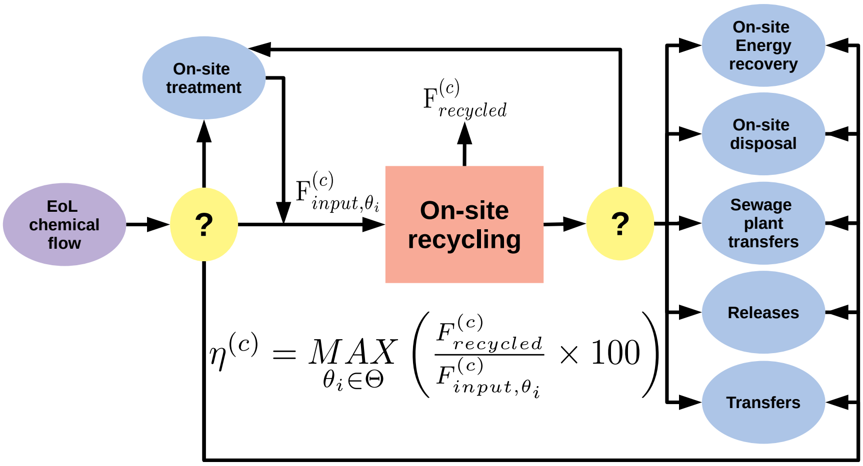


Figure 4.3: Schematic explanation to calculate $\eta^{(c)}$ for recycling using TRI information.

In step four, the framework excludes the PAU dataset records if $\eta^{(c)}$ could not be estimated. Finally, in step five, the value range for F_{input} is calculated using the $w^{(c)}$ range, $\eta^{(c)}$, and $F^{(c)}_{recycled}$. Hence, this module provides F_input ,α_{input}, $w^{(c)}$, $\eta^{(c)}$, PAU technologies, and chemical uses/activities.

4.2.2 *Data engineering for estimating OPEX and CAPEX: OPEX & CAPEX module*

OPEX and CAPEX influence implementing a PAU technology (Collins & Harris, 2002; Gray & Shadbegian, 1995, 1998). Hence, incorporating both parameters would result in realistic insights for inferring/predicting whether an IS affords a specific PAU technology. The framework uses the 2005 PACE – Survey to obtain the CAPEX and OPEX (U.S. Environmental Protection Agency, 2017b). The PACE–survey presents such parameters by media (air, water, and solid waste), EoL activity, and IS (only manufacturers). However, using the publicly-available version is impossible to know each surveyed facility expenses in USD/EoL-flow-kg, e.g., treating air emissions since the information is presented as an aggregated value in USD. Figure 4.4 shows a procedure based on Monte Carlo for mimicking potential CAPEX and OPEX in USD/EoL-flow-kg for each facility to manage contaminants that would otherwise have polluted the environment.

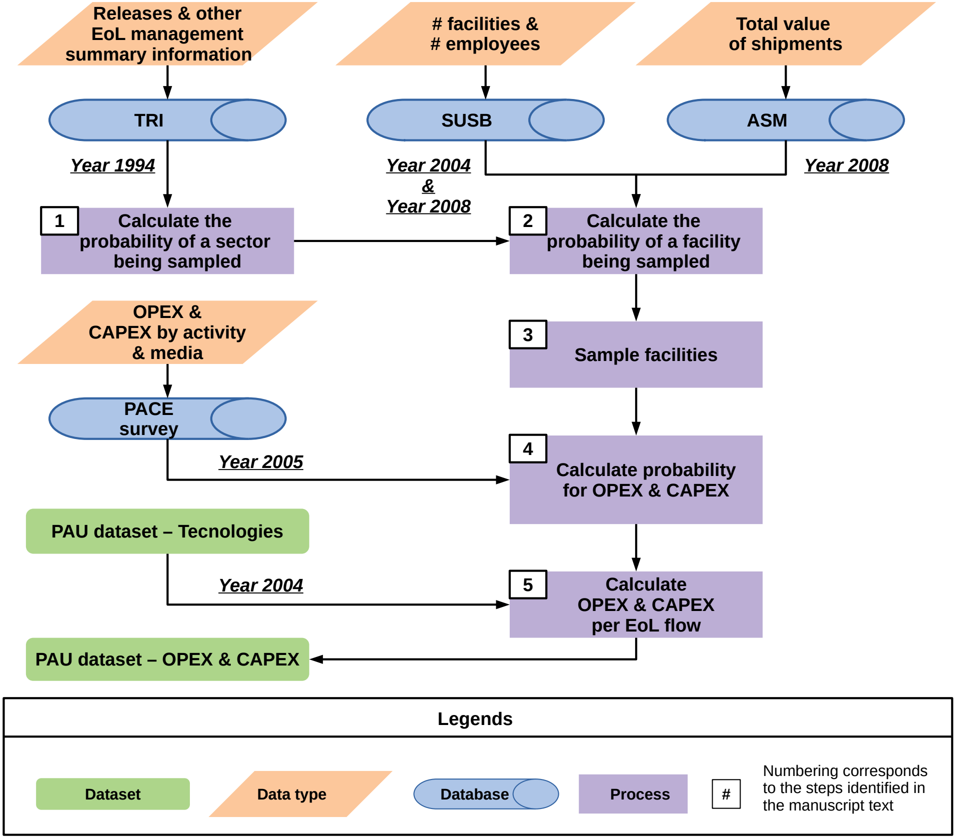


Figure 4.4: Data engineering for mimicking the PACE–survey and obtaining the PAU dataset–OPEX & CAPEX module.

In step one, the probability of sampling an IS is calculated. The PACE – Survey targeted ISs with a high percentage of facilities reporting no OPEX in 1994. Hence, the framework selects 1994 TRI and uses the EoL chemical flows to determine potential facilities in ISs with no OPEX in 1994. In step two, the probability of sampling a facility is estimated, considering that the

2004 SUSB was the PACE–Survey sample frame and facilities with twenty or more employees were eligible.

The CAPEX and OPEX correlate positively with the value of shipments (VoS) (U.S. Environmental Protection Agency, 2017b), which measures the USD of products sold by manufacturers, and it is an IS measure-of-size. Thus, the probability of sampling a facility from an IS is proportional to the VoS. 2008 SUSB and 2008 ASM were selected due to being the closest survey period to 2005. The total VoS, including relative standard error, is taken from the 2008 ASM, meanwhile, the number of facilities from 2008 SUSB. Using these two values, the VoS mean and standard deviation by IS are calculated. After, using the sample frame and the above VoS statistical measures, a lognormal distribution is obtained to assign a VoS for a facility. This distribution has positive values that create a right-skewed curve explaining better the earnings behavior (Heckman & Sattinger, 2015).

In step three, facilities are drawn from the sampling pool, considering the probabilities calculated in steps one and two. In step four, from the PACE – Survey, the CAPEX and OPEX by activity and media are used to calculate the probability that facilities within ISs may have spent on specific activity and media. Thus, using the above probability, the one calculated in step two, and assuming these two events are independent, the number of facilities within ISs having at least a PAU for a specific activity and media are determined.

Finally, in step five, the 2004 PAU dataset – Technologies helps calculate the EoL flow mean and standard deviation by IS, media, and activity to obtain lognormal distribution to assign an EoL flow value for a facility. EoL flow by IS, activity, and media is calculated by the probability obtained in step four and summing up the EoL flow values. This amount normalizes the OPEX and CAPEX for each activity, media, and IS, i.e., OPEX and CAPEX in USD/EoL-flow-kg.

4.2.3 *Data engineering for estimating chemical unit price: ChemPrices module*

ChemPrices module contains information about the relationship between the PAUs and the chemical/chemical category prices ($UP^{(c)}$) in USD/g. $UP^{(c)}$ is obtained from e-commerce sources like SciFinder, Amazon, Alibaba, and Fisher Scientific, considering the currency exchange rate of the U.S. dollar. Due to having $UP^{(c)}$ for multiple suppliers, the framework drops the outliers using the Z score test. For the TRI chemical categories, the framework uses regulatory lists to know potentially-candidate chemicals belonging to them. Such lists can be found in the GitHub repository PAU4Chem (Hernandez-Betancur, Martin, et al., 2020b). The framework uses the FRS, a database containing information about facilities regulated by the U.S. Environmental Protection Agency. FRS provides the alternative name for the reporting facilities in the PAU dataset so the framework can connect them to the CDR database. Hence, the framework identifies those chemicals reported by the facilities to the CDR and allocates these into the TRI chemical categories. If the search is successful, the framework calculates $UP^{(c)}$ for a category using a smaller group of chemicals; otherwise, it uses all the chemicals in the corresponding category. $UP^{(c)}$ attribution is based on the median because it is a central tendency statistic less sensitive to outliers than the mean.

4.2.4 *Estimation of EoL management and CFA*

This section shows a procedure to employ the PAU dataset modules with data-driven and MCDM to suggest PAU technologies and estimate their $\eta^{(c)}$, CAPEX, and OPEX. Moreover, the procedure moves forward to perform CFA after the MCDM, thus, enabling the rapid estimation of chemical releases and output streams from the PAU technologies and providing exposure scenarios for further assessment (Hernandez-Betancur & Ruiz-Mercado, 2019).

Figure 4.5 shows the step-by-step procedure to perform CFA for a chemical of concern in PAU sequences. Bayesian Networks (BNs) infer the potential PAU technologies to manage the chemical (Koller & Friedman, 2009). BN variables are the CAPEX, OPEX, type of EoL management, $\eta^{(c)}$, $w^{(c)}$), F_input, $UP^{(c)}$, α_{input}, whether the chemical is an impurity/by-product ($I_{IB}^{(c)} = Yes/Not$), and *PAU technology*. The BNs structure, variables, and their connections are in the GitHub repository (Hernandez-Betancur, Martin, et al., 2020a). This structure mimics the decision-making process of technical stakeholders selecting and designing a PAU. The three PAU dataset modules supply the data for building the conditional probability tables for revealing relationships between the BN variables, e.g., the conditional probability of obtaining a $\eta^{(c)}$ value given $w^{(c)}$. As shown in Figure 4.5, if F_input has n individual chemicals, n BNs are built. Thus, the inference is based on a case-by-case analysis using the individual chemical information from the PAU dataset. The BNs only use the PAU dataset from 1987 to 2004 since these reporting years have information about $w^{(c)}$. Moreover, the procedure uses the PAU dataset to determine the PAU sequence using only dataset records not reporting PAU sequences to avoid any systematic error. Stakeholders must enter either $\eta^{(c)}$ or PAU technology or optional problem specifications to calculate the probability of either selecting a *PAU technology* or $\eta^{(c)}$ for a chemical.

As Figure 4.5 depicts, if a chemical has more than one potential PAU technology to satisfy the specifications, the Fuzzy Analytic Hierarchy Process (FAHP) with triangular fuzzy numbers supports the MCDM (Hernández-Betancur et al., 2019). The FAHP–selection uses the criteria presented in Figure 4.5, considering the importance of the type of EoL management activity to which the *PAU technology* belongs, i.e., if the *PAU technology* is for recycling (more preferable), energy recovery, or treatment (less preferable) (U.S. Environmental Protection Agency, 2015). The FAHP – selection checks the probability of a type of EoL management, and the PAU technology are selected and designed by the BN. If the F_{input} has more than one chemical, the FAHP – selection considers whether the *PAU technology* can manage several chemicals in F_{input}. Finally, if CAPEX and OPEX are not problem specifications, they are considered as FAHP - selection criteria.

As mentioned before, the framework considers the effect of the *PAU technology* on the chemical of concern. Hence, if more than one *PAU technology* is needed to manage an EoL stream, the FAHP – sequence supports PAU sequences arrangement. FAHP–sequence works using five criteria. Three criteria are chemical flammability, instability, and corrosiveness. These criteria ensure the safety of equipment and process structure. These properties are in a public GitHub repository (Hernandez-Betancur, Martin, et al., 2020c), supporting exposure assessment and circular life cycle endeavors (Hernandez-Betancur et al., 2022). The fourth criterion is the feasibility of finding a similar *PAU technology* sequence in the PAU dataset. The fifth criterion is $F_{input}^{(c)}$, associated with PAU equipment size and cost.

Finally, the CFA is completed considering Equation 4.1 and the procedure for calculating $\beta^{(c)}$ developed by Hernandez-Betancur, Ruiz-Mercado, et al. (2020). As described in Figure 4.5, the CFA complies with boiling and melting points to determine whether a chemical is assessed as a liquid, solid, or gas at standard conditions. The CFA considers if the chemical is a metal to determine whether combustion operations may abate it. Based on PAU technology functioning knowledge, α_{input}, and water solubility, a potential predominant phase for the output streams (α_{input}) is assigned. Meanwhile, the physicochemical properties are in the GitHub repository named above (Hernandez-Betancur, Martin, et al., 2020c). The procedure and data-driven and MCDM depicted in Figure 4.5 are not definitive. In the future, it would be extended and modified to predict potential on-site EoL management for chemicals that are not in the PAU dataset.

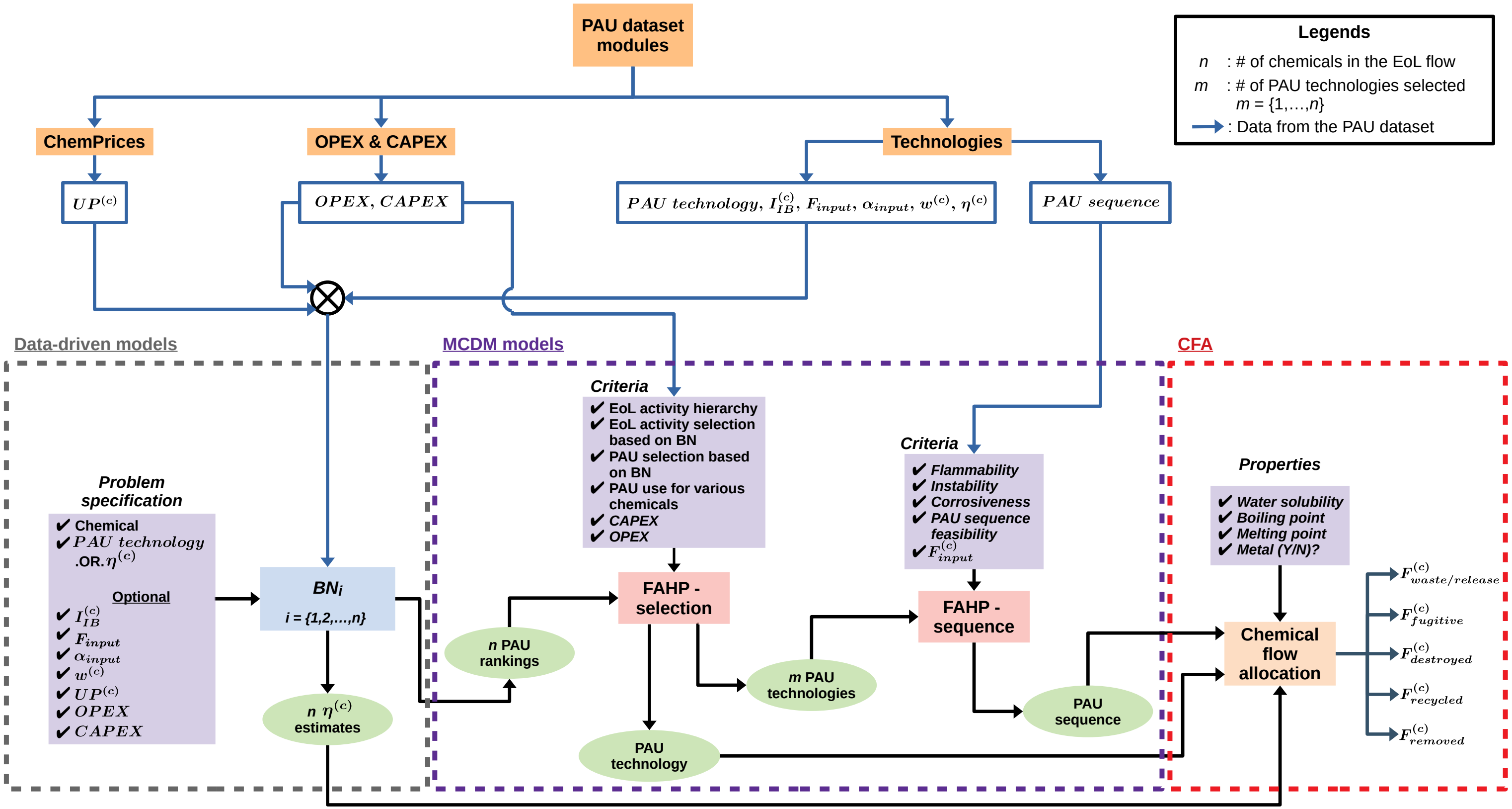


Figure 4.5: A generic step-by-step procedure and schematic explanation to perform CFA in PAU sequences by using the PAU dataset and data-driven and MCDM models.

4.3 CASE STUDIES

Three relevant case studies based on food and pharmaceutical ISs illustrate the application of the enhanced and extended framework with the PAU dataset modules described in Sections 4.1, and the procedure presented in Section 4.2.4. The results show how the enhanced framework designs and evaluates PAU technology systems handling individual and multiple chemicals in EoL flows and provides release inventories and pathways for conducting chemical exposure assessment for potential on-site EoL scenarios. The case studies specify budget (CAPEX and OPEX),. Case study Python scripts can be found in a public GitHub repository (Hernandez-Betancur, Martin, et al., 2020a). Tables describing relevant input/output information for the case studies are in the Appendix B.

The chemicals for the case studies are isopropanol, methanol, ammonia, ethylene glycol, n-hexane, toluene, N,N-dimethylformamide, and dichloromethane. These eight chemicals are part of the Organization for Economic Co-operation and Development List of High Production Volume Chemicals, which contains chemicals produced in amounts equal to or greater than 1,000,000 kg/yr (Organization for Economic Co-operation and Development, 2004). Those chemicals can be found at least in one of the lists of chemicals of concern categories, e.g., as hazardous air pollutants, hazardous wastes, and extremely hazardous substances (U.S. Environmental Protection Agency, 2020). Chemicals like methanol, ammonia, ethylene glycol, and n-hexane are associated with toxic releases from the food manufacturing IS (Gaona et al., 2020). Isopropanol, methanol, toluene, N,N-dimethylformamide, and dichloromethane are related to pharmaceutical manufacturing. These four chemicals are widely released into the air and incinerated (Beck et al., 1978).

Hence, as outcomes from in-depth numerical validation findings, the case study results show features, implications, and limitations of the data engineering framework that should be considered in future developments. Case study 1 demonstrates that, although using historical data can increase data availability, it can dilute most recent data evidencing releases reduction due to using more cleaner technologies. Case study 2 shows potential unintended outcomes on the predictions due to using ranges to describe . Case study 3 shows that the framework can provide data to build models capable of designing PAU technologies and arranging PAU sequences like those used in actual industrial facilities.

4.4 RESULTS AND DISCUSSION

4.4.1 *Case study 1: Celecoxib manufacturing process*

Three EoL flows associated with the isopropanol/water washes, mother liquor or filtrate, and dryer distillates from the Celecoxib manufacturing process are used as a case study (Hounsell et al., 2012; Slater et al., 2008). These streams contain water, methanol, ethanol, and isopropanol. Isopropanol and methanol are the chemicals of concern selected due to their toxicity and presence in the PAU dataset. The GitHub repository has all the problem specifications for the BNs, the criteria values for performing the FAHP decision-making, and the physicochemical properties for completing the CFA (Hernandez-Betancur, Martin, et al., 2020a). The input chemical concentration ($w^{(c)}$) and the predominant input phase (α_{input}) may determine the type of management for each chemical in the EoL flows. Table B.1 presents these specifications. The EoL input flow to a PAU (F_{input}) is $8.05 \times 10^6 kg/yr$. This value is the average F_{input} containing both chemicals and is calculated using the estimated values of F_{input} for the pharmaceutical industry sector (IS) records from the PAU dataset.

A finding from the development of the case studies relates to the best procedure for performing the CFA for the PAU/PAU sequence by following Equation 4.1. A PAU - level approach is employed to perform the CFA and ensures that both the material balance and the required abatement efficiency relative to the chemical ($\eta^{(c)}$) are satisfied (bottom-up in Table B.2). This is an enhancement to the framework developed by Hernandez-Betancur, Ruiz-Mercado, et al. (2020) since using facility-level information to estimate emission factors (top-down in Table B.2) causes an overestimation of and does not meet the requirements.

Comparing between the CFA obtained by both approaches for stream # 1 in Table B.1, e.g., the mean for isopropanol from the batch still distillation is $1.99 \times 10^4 kg/yr$ using the bottom-up approach. In contrast, is $9.86 \times 10^5 kg/yr$ using the top-down approach. The mean for isopropanol is $3.93 \times 10^6 kg/yr$ by the bottom-up approach, which means 99.18% of the isopropanol fed in stream # 1. Instead, for the case of the top-down approach, is $1.97 \times 10^6 kg/yr$, i.e., 49.86% of the isopropanol fed into stream # 1 (see comparison Table B.2 for more details). In this case study, the expected for isopropanol must be 99.50%; therefore, the bottom-up approach can satisfy the established in the material balance. Hence, this exercise demonstrates the effectiveness of using the bottom-up approach to perform all material balances and design PAU systems to estimate chemical releases at on-site EoL management activities, as proposed in this research contribution. Therefore, the inclusion of the on-site EoL activities and PAU technology information enhances the CFA at the EoL stage developed by Hernandez-Betancur, Ruiz-Mercado, et al. (2020).

Considering the problem specifications like , , and for the three streams from the Celecoxib process (see Table B.1), the framework suggests batch still distillation for recycling methanol, while incineration using liquid injection for treating isopropanol, see further details in Table B.3. Hence, the PAU dataset used to develop the BNs would lead to suggest to a stakeholder the use of a technology associated with a destructive process like a liquid injection incinerator. In fact, from a phenomenological perspective, liquid injection incinerator is a suitable technology for wastes with a high-organic content like the three streams from the Celecoxib process (Lee et al., 1986). The BN has a structure whose estimates are based on statistical evidence. Hence, the statistical evidence in the PAU dataset from 1987 to 2004 shows the companies used to destroy the isopropanol instead of recovering it. In fact, researchers have studied isopropanol recovery by pervaporation and distillation as an alternative to incineration (Hounsell et al., 2012; Slater et al., 2008). Methanol has been reported several times, describing successful applications of green chemistry and engineering activities (U.S. Environmental Protection Agency, 2019).

Although both methanol and isopropanol are on the Organization for Economic Co-operation and Development List of High Production Volume Chemicals, the quantities of methanol reported in the U.S. as domestically manufactured, imported, used, and exported far exceed those of isopropanol. For example, the ratio of imported methanol to imported isopropanol is about 51.73 (see Table B.9). This may explain why the statistical distribution of the data used to construct the BN leads to the selection of recycling for methanol but not for isopropanol. However, the quantities may also reflect the importance of the methanol market relative to that of isopropanol. This may be supported by the fact that the global market size of the methanol is around USD 33.69 billion, while for isopropanol it is around USD 2.65 billion (see Table B.9). Nevertheless, unlike BNs, other data-driven models combined with a data preprocessing step for the PAU dataset can get a greener suggestion for managing EoL flow containing isopropanol or even other substances, thereby overcoming the data drift and concept drift that could affect the applicability of the models.

An important implication from a life cycle inventory perspective and chemical releases quantification is that after selecting PAU technologies, the framework suggests potential output streams and their α_{output} where the chemical in the case study may be allocated downstream of the PAUs. This implication is important for understanding the potential indirect risk that may be associated with using a PAU technological to abate a chemical of concern. For example, methanol may generate $F^{(c)}_{waste/release}$ (whose α_{output} is wastewater) from batch still distillation. Both methanol and isopropanol may be in $F^{(c)}_{effluent}$ from the liquid injection operation related to stack air releases. In addition to the potential output streams and α_{output}, the framework also provides a value range for each flow (see Table B.3). For instance, the mean $F^{(c)}_{fugitive}$ for isopropanol from the batch still distillation is $1.99 \times 10^4 kg/yr$ for the stream # 1 and $1.39 \times 10^4 kg/yr$ for stream # 2. The result above means the $F^{(c)}_{fugitive}$ for isopropanol from batch still distillation is higher for stream # 1 than for stream # 2. This result makes sense, considering $w^{(c)}$ for isopropanol is higher for stream # 1, and F_{input} for both cases is $8.05 \times 10^6 kg/yr$. Moreover, the mean value is accompanied by a Coefficient of Variation (CV) for each flow estimate (see Table B.3), enabling the analysis of intervals instead of point values for the results. For example, methanol in $F^{(c)}_{waste/release}$ from batch still distillation can be in the interval of $[0, 9.49 \times 10^0 kg/yr]$ and $[0, 1.31 \times 10^1 kg/yr]$ for stream # 1 and 3. Hence, it may be possible to find methanol quantities in $F^{(c)}_{waste/release}$ from batch still distillation for stream # 1 lower than for stream # 3, even though $w^{(c)}$ for methanol in stream # 1 is higher than for stream # 3, as shown in Table B.1.

4.4.2 *Case study 2: Solvent EoL flows from Food IS*

Methanol, ammonia, ethylene glycol, and n-hexane are four chemicals widely used in the food IS (Gaona et al., 2020). Case studies have been addressed using possible values for the problem specifications (Hernandez-Betancur, Martin, et al., 2020a). Information for these four chemicals reported by this IS was searched in the PAU dataset to identify whether the chemical is an impurity/by-product ($I^{(c)}_{IB} = Yes/Not$), ($w^{(c)}$), and F_{input} the average . Hence, the F_{input} containing the chemicals is fixed to $8.28 \times 10^6 kg/yr$ for methanol, $2.60 \times 10^9 kg/yr$ for ammonia, $8.70 \times 10^9 kg/yr$ for ethylene glycol, and $5.40 \times 10^8 kg/yr$ for n-hexane. Each F_{input} is assumed to be a methanol-water mixture (see Table B.4). The desired output chemical concentration of the case study is based on the lowest ecological benchmark found for each chemical in the Risk Assessment Information System (U.S. Department of Energy, 1998), except for methanol that is based on a study developed for the American Methanol Institute (Malcolm Pirnie Inc, 1999). These values are used to set the required $\eta^{(c)}$ for each circumstance. Hence, setting parameter $I^{(c)}_{IB}$, $w^{(c)}$, $\eta^{(c)}$, and α_{output} to a specific value could lead to suggest a different PAU technology for each F_{input} (see Table B.4 for the values established).

Performing the procedure shown in Figure 4.5 results in the selection of PAU technologies for each solvent F_{input} from the food IS and their CFA for allocating the chemicals downstream (see Tables B.5–B.8 for more details). An aspect to highlight is that according to the results obtained for methanol, see Table B.5, and ammonia, see Table B.7, the selection of the PAU technology depends more on the input concentration of the chemical ($w^{(c)}$) and the required PAU abatement efficiency relative to the chemical ($\eta^{(c)}$) than whether the chemical is an impurity/by-product ($I_{IB}^{(c)} = Yes$). For instance, batch still distillation is selected for managing a solvent F_{input} when $w^{(c)}$ for ammonia is 0.01 %wt/wt regardless of $I_{IB}^{(c)} = Yes$ for ammonia. Likewise, solvent recovery via fractionation is selected when $w^{(c)}$ for ammonia is 0.01 and 0.51 %wt/wt. Hence, for data-driven modeling and decision-making, $w^{(c)}$ has higher relative importance than $I_{IB}^{(c)}$.

Another variable having an important influence on data-driven modeling and decision-making is F_{input}. For example, in case study one, F_{input} containing methanol was $8.05 \times 10^6 kg/yr$, and the result suggested solvent recovery via batch still distillation. Likewise, for streams # 4 and 7, F_{input} is $8.28 \times 10^6 kg/y$, and the suggestion in both cases is batch still distillation for methanol recovery. In contrast, for streams # 11, 12, 15, and 16 containing ammonia, F_{input} is $2.60 \times 10^9 kg/yr$, and the recovery for all the cases would require fractionation. The above may be explained because continuous distillation and fractionation are often preferred to batch operations for large solvent recovery streams (Douglas, 1988; Smallwood, 2002). This aspect may lead to developing future data-driven models that assign a high weight to the parameters associated with $w^{(c)}$ and F_{input}; however, this aspect should be studied thoroughly.

The PAU selection in case of the highest $w^{(c)}$s indicates a high chance of selecting treatment via incineration when $w^{(c)}$ is high. For instance, Table B.5 indicates when $w^{(c)}$ for methanol is 75.00 %wt/wt, liquid injection incineration operation is chosen, which is like what happens with n-hexane at the same $w^{(c)}$, as indicated in Table A.5. In case study one, the same occurred with isopropanol, whose $w^{(c)}$ was between 34.50 and 50.70 %wt/wt, close to high organic content, the desired requirement for combustion operations. Although the above F_{input} contains water, liquid injection incineration can handle high organic-strength aqueous wastes without requiring auxiliary devices (Lee et al., 1986). As explained in case study one, facilities should explore recycling and reuse activities instead of using incineration to reduce the waste volume before disposal.

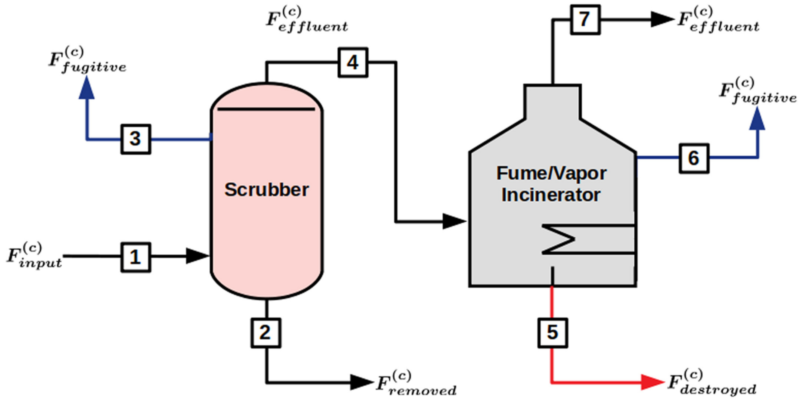
For each solvent, aerobic treatment was selected for managing the solvent. The concentration under the BN and the FAHP suggests that biological treatment is recommended when $w^{(c)}$ is 25.50 %wt/wt for methanol;

1.00×10^{-4} %wt/wt for ammonia; 0.51 and 25.50 %wt/wt for ethylene glycol; and 1.00×10^{-4}, 0.51, and 25.50 %wt/wt for n-hexane. Additionally, F_{input} is wastewater for all the cases. This fact may be explained because several bioenvironmental factors affect bacteria activity and the rate of biological reactions. The biochemical oxygen demand is efficiently treated in the range of 60-500mg/L ($\approx$ 0.01 to 0.05 %wt/wt) (Samer, 2015). However, aerobic treatment suggestion for $w^{(c)}$ at 25.50 %wt/wt might be because 25.50 %wt/wt is in TRI concentration range code 1, which describes $w^{(c)}$ higher than 1.00 %wt/wt. Since $w^{(c)}$ is a crucial design and operating parameter for this treatment technology, employing values range for describing $w^{(c)}$ may be a disadvantage for applying these PAUs in these EoL flows. Another reasonable explanation for the above suggestion is the reporting facilities might have used aerobic wastewater treatment above 500mg/L ($\approx$ 0.05 %wt/wt) biochemical oxygen demand by applying enough dilution rate (Samer, 2015).

As in case study one, after selecting the PAU technologies for each F_{input}, it is possible to allocate chemical flows downstream of each PAU. F_{input} may be larger for F_{input} having ammonia when $I_{IB}^{(c)} = Yes$, e.g., $F_{fugitive}^{(c)}$ is $1.06 \times 10^{0} kg/yr$ and $6.39 \times 10^{-1} kg/y$ for stream # 13 and 17 (see Table B.7). However, considering CV, the intervals for $F_{fugitive}^{(c)}$ is $[0, 2.49 \times 10^{0} kg/yr]$ and $[0, 1.43 \times 10^{0} kg/yr]$ for stream # 13 and 17. From a statistical viewpoint, both values may be the same. This result is coherent because $F_{fugitive}^{(c)}$ is directly proportional to the mass under analysis.

4.4.3 *Case study 3: EoL flow from pharmaceutical preparation manufacturing*

As indicated in Section 4.3, the first two case studies only use the information for individual PAU technologies instead of the PAU sequences found in the PAU dataset. This third case study is employed to demonstrate the framework effectiveness in designing a network of PAUs and estimating the potential releases of the chemical of concern. An existing record from the PAU dataset is randomly selected by considering only records with PAU sequences for comparing this sequence with the one designed and built by BNs and FAHP. The randomly selected record has a gaseous F_{input} having methanol, dichloromethane, toluene, and N,N-dimethylformamide. This record uses a PAU sequence consisting of a condenser, a scrubber, and an absorber. The record belongs to a facility in pharmaceutical IS. The problem specifications for BNs and the value for FAHP criteria can be found in the GitHub repository (Hernandez-Betancur, Martin, et al., 2020a). The problem specifications used for obtaining the results are the same reported by the facility. The PAU sequence obtained following the procedure in Figure 4.5 is compared with the one in the selected record; therefore, testing the framework. Figure 4.6 presents the PAU sequence obtained and the CFA for each chemical.



PAU name	Stream[1]	α_{output}[2]	Chemical	Mean quantity [kg/yr]	CV[3]
Scrubber	2	W	N,N-dimethylformamide	$1.52\text{x}10^{6}$	0.01
			methanol	$1.53\text{x}10^{6}$	0.01
			dichloromethane	$1.53\text{x}10^{6}$	0.01
	3	A	toluene	$8.03\text{x}10^{4}$	0.01
			N,N-dimethylformamide	$7.06\text{x}10^{1}$	1.17
			methanol	$5.00\text{x}10^{3}$	1.18
			dichloromethane	$1.42\text{x}10^{4}$	0.90
	4	A	toluene	$1.53\text{x}10^{6}$	0.01
			N,N-dimethylformamide	$8.01\text{x}10^{4}$	0.01
			methanol	$7.53\text{x}10^{4}$	0.08
			dichloromethane	$6.62\text{x}10^{4}$	0.19
Fume/Vapor Incinerator	5	NA	toluene	$1.45\text{x}10^{6}$	0.01
	6	A	toluene	$9.81\text{x}10^{3}$	0.96
			N,N-dimethylformamide	$4.78\text{x}10^{1}$	1.18
			methanol	$5.23\text{x}10^{3}$	1.48
			dichloromethane	$1.33\text{x}10^{4}$	1.02
	7	A	toluene	$6.65\text{x}10^{4}$	0.14
			N,N-dimethylformamide	$7.99\text{x}10^{4}$	0.01
			methanol	$6.98\text{x}10^{4}$	0.13
			dichloromethane	$5.29\text{x}10^{4}$	0.35

[1] Stream # in the Flow Diagram.

[2] For the TRI Program, A represents air emissions. Additionally, liquid wastes having more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as a nonaqueous material (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[3] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

[4] All chemicals in the randomly selected record have a concentration in the interval 0.01-1 %wt/wt. For the CFA, the concentration was set at the middle of the interval, i.e., 0.051 % wt/wt.

Figure 4.6: PAU selection and sequence and CFA for the record randomly selected from the PAU dataset. F_{input} for this case study is $3.18 \times 10^{9} kg/yr$.

As depicted in Figure 4.6, the PAU network designed consists of a scrubber followed by a fume/vapor incinerator. Although this sequence is different from the selected record one (condenser, scrubber, and absorber), the sequence consists of treatment operations and neither recycling nor energy recovery technologies were selected. Additionally, as presented in the stream # 2 in Figure 4.6, a scrubber was selected for handling methanol, dichloromethane, and N,N-dimethylformamide. It means it is possible that for the record sequence, the reporting facility used a scrubber to remove these three chemicals and the other operations for removing different chemicals from the same F_{input} (e.g., toluene). This inference is coherent because the TRI Program requires reporting the treatment unit or sequence regardless of whether all the units handle the reporting chemical. The BN and FAHP also suggest fume/vapor incineration as the most probable and feasible PAU technology for toluene. Fume/vapor incinerator also called thermal oxidizers are widely used for the treatment of volatile organic compounds and hazardous air pollutants like toluene. In addition, this operation unit performs best at pollutant loading of around $1,500 - 3,000$ ppm (≈ 0.15 to 0.3 %wt/wt) (U.S. Environmental Protection Agency, 2002). Therefore, the PAU statistics for toluene suggests that for the reporting years 1987-2004, facilities frequently reduced EoL flow containing toluene by incineration. The CFA also indicates the stack emission or $F^{(c)}_{effluent}$ consists of $7.09 \times 10^4 kg/yr$ of methanol (CV $= 0.13$), $8.02 \times 10^4 kg/yr$ of N, N-dimethylformamide (CV $= 0.01$), $5.35 \times 10^5 kg/yr$ of dichloromethane (CV $= 0.35$), and $5.97 \times 10^4 kg/yr$ of toluene (CV $= 0.14$). As shown the CVs for the chemical flows from stack, these result values do not present a high variability (CV < 1). Unlike, the above flows, the $F^{(c)}_{fugitive}$s for this case study also have a wide range of possible values. For instance, $F^{(c)}_{fugitive}$ for methanol from fume/vapor incinerator has a mean value of $5.2 \times 10^3 kg/yr$, but its CV is equal to 1.48, which means that the result variability is high (the highest CV for this case study). Thus, the range of potential values for the $F^{(c)}_{fugitive}$ for methanol from fume/vapor incinerator is $[0, 1.30 \times 10^4 kg/yr]$.

4.5 CONCLUSIONS

This work proposed a framework that tracks chemical flows, estimates releases, and identifies potential exposure pathways at on-site EoL industrial activities. The framework integrates multiple publicly-available databases and employs data-driven models and MCDM to develop PAU technology systems and estimates their $\eta^{(c)}$, CAPEX, and OPEX. Nonetheless, further data collection should be required to reduce cost data uncertainty and increase the reliability of the data-driven models or incorporating heuristic equations to estimate costs. The framework leverages the data and the literature to fill information gaps to complete the chemical flow tracking and allocate chemicals of concern inside EoL stage facilities and downstream of the PAU technologies.

Case studies based on food and pharmaceutical ISs were used to illustrate the framework application for CFA and allocation and the benefits of integrating and extending the framework with data-driven models and MCDM. The framework can support PAU technologies selection for managing chemicals of concern in EoL flows, considering the PAU effect on chemicals, and suggesting the EoL management activity sequences. However, some data mining enhancements may help find cross-year data relationships and identify sustainable PAU technologies for chemicals of concern from the PAU dataset. This effort might overcome potential data availability limitations due to reporting requirement changes from the TRI Program post-2004 (no need for reporting $w^{(c)}$). This reporting specification change might affect identifying recent industrial PAU developments and improvements. It is crucial to thoroughly analyze and enhance the framework for designing and recommending biological treatment operations to ensure accurate PAU technology predictions and inferences.

The framework enables the analysis of estimated chemical flow variabilities. In the future, the framework can help support a complete CFA at the EoL life stage for risk assessment, avoiding case-by-case studies, by connecting the on-site EoL framework with the chemical off-site tracking framework and developing and exploring robust data-driven models. The further investigation of robust data-driven models would support CFA and EoL exposure pathway characterization to extend the framework to chemicals outside the PAU dataset. Nonetheless, the models must incorporate techniques to thoroughly assess their uncertainty and reliability.

DISCLAIMER

The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the EPA. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the EPA. The EPA does not endorse any commercial products, services, or enterprises.

ACKNOWLEDGMENTS

This research was supported in by appointment for Jose D. Hernandez-Betancur to the Research Participation Program at the Center for Environmental Solutions and Emergency Response, Office of Research and Development, EPA, administered by the Oak Ridge Institute for Science and Education through an Interagency Agreement between the U.S. Department of Energy and the EPA. The authors express their sincere gratitude and appreciation to Sandra D. Gaona, Mitchell Sumner, and Steve DeVito of the U.S. EPA's Toxics Release Inventory Program, for their critical review of draft versions of our manuscript, their recommendations, and insightful discussions.

BIBLIOGRAPHY

Baquerizo, G., Gamisans, X., Gabriel, D., & Lafuente, J. (2007). A dynamic model for ammonia abatement by gas-phase biofiltration including pH and leachate modelling. *Biosystems Engineering*, *97*(4), 431–440. https://doi.org/10.1016/j.biosystemseng.2007.03.031

Bascone, D., Cipollina, A., Morreale, M., Randazzo, S., Santoro, F., & Micale, G. (2016). Simulation of a regeneration plant for spent pickling solutions via spray roasting. *Desalination and Water Treatment*, *57*(48-49), 23405–23419. https://doi.org/10.1080/19443994.2015.1137146

Beck, D. A., Evans, L. B., & Zobel, K. (1978). *Control of Volatile Organic Emissions From Manufacture of Synthesized Pharmaceutical Products* (tech. rep.). https://www3.epa.gov/airquality/ctg_act/197812_voc_epa450_2-78-029_pharmaceutical_products.pdf

Bernas, M. (2013). Chemical Risk Assessment and Regulatory Decision Making. *Seminars in Oncology Nursing*, *29*(1), 12–19. https://doi.org/10.1016/j.soncn.2012.11.003

Bodar, C., Spijker, J., Lijzen, J., Waaijers-van der Loop, S., Luit, R., Heugens, E., Janssen, M., Wassenaar, P., & Traas, T. (2018). Risk management of hazardous substances in a circular economy. *Journal of Environmental Management*, *212*, 108–114. https://doi.org/10.1016/j.jenvman.2018.02.014

Bojarski, A. D., Guillén-Gosálbez, G., Jiménez, L., Espuña, A., & Puigjaner, L. (2008). Life cycle assessment coupled with process simulation under uncertainty for reduced environmental impact: Application to phosphoric acid production. *Industrial and Engineering Chemistry Research*, *47*(21), 8286–8300. https://doi.org/10.1021/ie8001149

Cashman, S. A., Meyer, D. E., Edelen, A. N., Ingwersen, W. W., Abraham, J. P., Barrett, W. M., Gonzalez, M. A., Randall, P. M., Ruiz-Mercado, G., & Smith, R. L. (2016). Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing. *Environmental Science & Technology*, *50*(17), 9013–9025. https://doi.org/10.1021/acs.est.6b02160

Cavanagh, E. J., Savelski, M. J., & Slater, C. S. (2014). Optimization of environmental impact reduction and economic feasibility of solvent waste recovery using a new software tool. *Chemical Engineering Research and Design*, *92*(10), 1942–1954. https://doi.org/10.1016/j.cherd.2014.02.022

Chea, J. D., Christon, A., Pierce, V., Reilly, J. H., Russ, M., Savelski, M., Slater, C. S., & Yenkie, K. M. (2019). Framework for Solvent Recovery, Reuse, and Recycling In Industries. In *Computer aided chemical engineering* (pp. 199–204). Elsevier B.V. https://doi.org/10.1016/B978-0-12-818597-1.50032-1

Chea, J. D., Lehr, A. L., Stengel, J. P., Savelski, M. J., Slater, C. S., & Yenkie, K. M. (2020). Evaluation of Solvent Recovery Options for Economic Feasibility through a Superstructure-Based Optimization Framework. *Industrial and Engineering Chemistry Research*, *59*(13), 5931–5944. https://doi.org/10.1021/acs.iecr.9b06725

Chmielewski, A. G., Urbański, T. S., & Migdał, W. (1997). Separation technologies for metals recovery from industrial wastes. *Hydrometallurgy*, *45*(3), 333–344. https://doi.org/10.1016/s0304-386x(96)00090-4

Collins, A., & Harris, R. I. D. (2002). Does Plant Ownership Affect the Level of Pollution Abatement Expenditure? *Land Economics*, 171–189. https://doi.org/10.2307/3147267

Darvin, C., & Serageldin, M. (2003). Using Bioreactors To Control Air Pollution. *Nepis.Epa.Gov*, 28. http://nepis.epa.gov/Adobe/PDF/P1005HH8.pdf

Douglas, J. M. (1988). *Conceptual Design Of Chemical Processes*.

Eberle, A., Bhatt, A., Zhang, Y., & Heath, G. (2017). Potential Air Pollutant Emissions and Permitting Classifications for Two Biorefinery Process Designs in the United States. *Environmental Science and Technology*, *51*(11), 5879–5888. https://doi.org/10.1021/acs.est.7b00229

Energy and Environmtental Analysis Inc. (2005). *Characterization of the U.S. industrial/commercial boiler population* (tech. rep. May). https://www1.eere.energy.gov/manufacturing/distributedenergy/pdfs/characterization_industrial_commerical_boiler_population.pdf

European Chemicals Agency. (2019). Substances restricted under REACH. Retrieved August 16, 2020, from https://echa.europa.eu/regulations/reach/evaluation/substance-evaluation

Fanning, J. C. (2000). The chemical reduction of nitrate in aqueous solution. *Coordination Chemistry Reviews*, *199*(1), 159–179. https://doi.org/10.1016/s0010-8545(99)00143-5

Frischknecht, R., Editors, N. J., Althaus, H.-j., Doka, G., Dones, R., Heck, T., Hellweg, S., Hischier, R., Nemecek, T., Rebitzer, G., & Spielmann, M. (2007). *Ecoinvent - overview and methodology* (tech. rep. No. 1). https://www.ecoinvent.org/files/200712_frischknecht_jungbluth_overview_methodology_ecoinvent2.pdf

Gaona, S. D., Pepping, T., Keenan, C., & DeVito, S. C. (2020). The Environmental Impact of Pollution Prevention, Sustainable Energy Generation, and Other Sustainable Development Strategies Implemented by the Food Manufacturing Sector. In *Green energy to sustainability* (pp. 539–567). Wiley. https://doi.org/10.1002/9781119152057.ch22

Gray, W. B., & Shadbegian, R. J. (1995). Pollution abatement costs, regulation, and plant-level productivity. *National Bureau of Economic Research, Inc.*, (January), 65–96. https://doi.org/10.3386/w4994

Gray, W. B., & Shadbegian., R. J. (1998). Environmental Regulation, Investment Timing, and Technology Choice. *The Journal of Industrial Economics*, *46*, 235–256. www.jstor.org/stable/117550.

GreenDelta. (2013). openLCA Nexus. Retrieved September 2, 2020, from https://nexus.openlca.org/

Guerras, L. S., & Martín, M. (2019a). Optimal Flue Gas Treatment for Oxy-Combustion-Based Pulverized Coal Power Plants. *Industrial and Engineering Chemistry Research*, *58*(45), 20710–20721. https://doi.org/10.1021/acs.iecr.9b04453

Guerras, L. S., & Martín, M. (2019b). Optimal gas treatment and coal blending for reduced emissions in power plants: A case study in Northwest Spain. *Energy*, *169*, 739–749. https://doi.org/10.1016/j.energy.2018.12.089

Hansen, G., MacCarthy, L. A., Brooks, C. S., & Brooks, P. L. (1991). *Metal recovery from industrial waste* (1st ed.). Lewis Publishers;CRC Press. shorturl.at/kmxX4

Hassim, M. H. (2016). Comparison of methods for assessing occupational health hazards in chemical process development and design phases. *Current Opinion in Chemical Engineering*, *14*, 137–149. https://doi.org/10.1016/j.coche.2016.10.001

Heckman, J. J., & Sattinger, M. (2015). Introduction to the distribution of earnings and of individual output, by A.D. Roy. *The Economic*

Journal, *125*, 378–402. https://doi.org/https://doi.org/10.1111/ecoj.12226

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2020a). PAU case study. Retrieved August 23, 2020, from https://github.com/jodhernandezbe/PAU_case_study

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2020b). PAU4Chem. https://doi.org/10.5281/zenodo.3990594

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2020c). Properties scraper. Retrieved October 30, 2020, from https://github.com/jodhernandezbe/Properties_Scraper

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2022). A data engineering approach for sustainable chemical end-of-life management. *Resources, Conservation and Recycling*, *178*, 106040. https://doi.org/https://doi.org/10.1016/j.resconrec.2021.106040

Hernandez-Betancur, J. D., & Ruiz-Mercado, G. J. (2019). Sustainability indicators for end-of-life chemical releases and potential exposure. *Current Opinion in Chemical Engineering*, *26*(November), 157–163. https://doi.org/10.1016/j.coche.2019.09.004

Hernandez-Betancur, J. D., Ruiz-Mercado, G. J., Abraham, J. P., Martin, M., Ingwersen, W. W., & Smith, R. L. (2020). Data engineering for tracking chemicals and releases at industrial end-of-life activities. *Journal of Hazardous Materials*, 124270. https://doi.org/10.1016/j.jhazmat.2020.124270

Hernández-Betancur, J. D., Hernández, H. F., & Ocampo-Carmona, L. M. (2019). A holistic framework for assessing hot-dip galvanizing process sustainability. *Journal of Cleaner Production*, *206*, 755–766. https://doi.org/10.1016/j.jclepro.2018.09.177

Hounsell, G., Pilipauskas, D., & Urbanski, F. (2012). Green design alternatives for isopropanol recovery in the celecoxib process, 687–698. https://doi.org/10.1007/s10098-011-0433-6

Huang, C. P., Dong, C., & Tang, Z. (1993). Advanced chemical oxidation: Its present role and potential future in hazardous waste treatment. *Waste Management*, *13*(5-7), 361–377. https://doi.org/10.1016/0956-053X(93)90070-D

Jaeger Products Inc. (2010). *Removal of Organics From Water Using Steam Stripping* (tech. rep. No. 281).

Jiménez-González, C., Kim, S., & Overcash, M. R. (2000). Methodology for developing gate-to-gate Life Cycle Inventory information. *International Journal of Life Cycle Assessment*, *5*(3), 153–159. https://doi.org/10.1007/BF02978615

Kaur, H. (2017). *Sustainable Environmental Biotechnology*. Springer. https://doi.org/10.1007/978-981-10-4041-2_8

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Kreith, F., Jr. Schnelle, K. B., Brown, C. A., & Carelli, C. (2001). *Air Pollution Control Technology Handbook*. CRC Press Inc.

Lee, C. C., Huffman, G. L., & Oberacker, D. A. (1986). An Overview of Hazardous/Toxic Waste Incineration. *Journal of the Air Pollution Control Association*, *36*(8), 922–931. https://doi.org/10.1080/00022470.1986.10466132

Li, S., Feliachi, Y., Agbleze, S., Ruiz-Mercado, G. J., Smith, R. L., Meyer, D. E., Gonzalez, M. A., & Lima, F. V. (2018). A process systems framework for rapid generation of life cycle inventories for pollution control and sustainability evaluation. *Clean Technologies and Environmental Policy*, *20*(7), 1543–1561. https://doi.org/10.1007/s10098-018-1530-6

Malcolm Pirnie Inc. (1999). *Evaluation of the fate and transport of methanol in the enviroment* (tech. rep. January). http://www.methanol.org/wp-content/uploads/2016/06/White-Paper-The-Fate-Transport-of-Methanol-in-the-Environment-1999.pdf

Mantus, E. K. (1992). *All Fired Up: Burning Hazardous Waste in Cement Kilns*. Environmental Toxicology International & the Combustion Research Institute.

Meyer, D. E., Mittal, V. K., Ingwersen, W. W., Ruiz-Mercado, G. J., Barrett, W. M., Gonzalez, M. A., Abraham, J. P., & Smith, R. L. (2019). Purpose-Driven Reconciliation of Approaches to Estimate Chemical Releases. *ACS Sustainable Chemistry and Engineering*, *7*(1), 1260–1270. https://doi.org/10.1021/acssuschemeng.8b04923

Mular, A. L., Halbe, D., & Barratt, D. J. (2002). *Mineral Processing Plant Design, Practice, and Control*. Society for Mining, Metallurgy & Exploration.

National Research Council. (2000). *Environmental transport and exposure pathways of substances emitted from incineration facilities*. https://doi.org/10.17226/5803

National Research Council. (2014). *A Framework to Guide Selection of Chemical Alternatives*. National Academies Press. https://doi.org/10.17226/18872

Organization for Economic Co-operation and Development. (2004). *The 2004 OECD List of High Production Volume Chemicals* (tech. rep.). http://www.oecd.org/chemicalsafety/risk-assessment/33883530.pdf

Ottoboni, A. P., De Souza, I., Menon, G. J., & Da Silva, R. J. (1998). Efficiency of destruction of waste used in the co-incineration in the rotary kilns. *Energy Conversion and Management*, *39*(16-18), 1899–1909. https://doi.org/10.1016/s0196-8904(98)00081-8

Ragas, A. M. (2011). Trends and challenges in risk assessment of environmental contaminants. *Journal of Integrative Environmental Sciences*, *8*(3), 195–218. https://doi.org/10.1080/1943815X.2011.597769

Righi, S., Baioli, F., Dal Pozzo, A., & Tugnoli, A. (2018). Integrating life cycle inventory and process design techniques for the early estimate of energy and material consumption data. *Energies*, *11*(4), 970. https://doi.org/10.3390/en11040970

Samer, M. (2015). *Biological and Chemical Wastewater Treatment Processes, Wastewater Treatment Engineering*. https://doi.org/10.5772/61250

Saxena, S. C., & Jotshi, C. K. (1996). Management and combustion of hazardous wastes. *Progress in Energy and Combustion Science*, *22*(5), 401–425. https://doi.org/10.1016/S0360-1285(96)00007-X

Shin, C. H., Kim, J. Y., Kim, J. Y., Kim, H. S., Lee, H. S., Mohapatra, D., Ahn, J. W., Ahn, J. G., & Bae, W. (2009). Recovery of nitric acid from waste etching solution using solvent extraction. *Journal of Hazardous Materials*, *163*(2-3), 729–734. https://doi.org/10.1016/j.jhazmat.2008.07.019

Slater, C. S., Savelski, M., Hounsell, G., Pilipauskas, D., & Urbanski, F. (2008). Analysis of separation methods for isopropanol recovery in the celecoxib process. *AIChE Annual Meeting, Conference Proceedings*.

Smallwood, I. M. (2002). *Solvent Recovery Handbook*.

Smith, R. L., Ruiz-Mercado, G. J., Meyer, D. E., Gonzalez, M. A., Abraham, J. P., Barrett, W. M., & Randall, P. M. (2017). Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling. *ACS Sustainable Chemistry and Engineering*, *5*(5), 3786–3794. https://doi.org/10.1021/acssuschemeng.6b02724

Smith, R. L., Tan, E. C. D., & Ruiz-Mercado, G. J. (2019). Applying Environmental Release Inventories and Indicators to the Evaluation of Chemical Manufacturing Processes in Early Stage Development. *ACS Sustainable Chemistry & Engineering*. https://doi.org/https://doi.org/10.1021/acssuschemeng.9b01961

Song, R. (2019). *Machine Learning for Addressing Data Deficiencies in Life Cycle Assessment*. https://doi.org/.1037//0033-2909.I26.1.78

Sugiyama, H., Hirao, M., Fischer, U., & Hungerbühler, K. (2008). Activity modeling for integrating environmental, health and safety (EHS) consideration as a new element in industrial chemical process design. *Journal of Chemical Engineering of Japan*, *41*(9), 884–897. https://doi.org/10.1252/jcej.07we263

Sundaravaradan, N., Marwah, M., Shah, A., & Ramakrishnan, N. (2011). Data mining approaches for life cycle assessment. *Proceedings of the 2011 IEEE International Symposium on Sustainable Systems and*

Technology, ISSST 2011, (June). https://doi.org/10.1109/ISSST.2011.5936863

Tay, O.-H., & Zhang, X. (1999). Neural Fuzzy Modeling of Anaerobic Biological Wastewater Treatment Systems. *Journal of Environmental Engineering*, *125*(December), 1149–1159. https://doi.org/https://doi.org/10.1061/(ASCE)0733-9372(1999)125:12(1149)

U.S. Census Bureau. (2007). 2004 SUSB Annual Datasets by Establishment Industry. Retrieved August 20, 2020, from https://www.census.gov/data/datasets/2004/econ/susb/2004-susb.html

U.S. Census Bureau. (2009). 2008 Annual Survey of Manufactures (ASM): Tables. Retrieved August 20, 2020, from https://www.census.gov/data/tables/2008/econ/asm/2008-asm.html

U.S. Census Bureau. (2011). 2008 SUSB Annual Datasets by Establishment Industry. Retrieved August 20, 2020, from https://www.census.gov/data/datasets/2008/econ/susb/2008-susb.html

U.S. Census Bureau. (2017a). NAICS Structure. Retrieved October 23, 2020, from https://www.census.gov/programs-surveys/economic-census/guidance/understanding-naics.html

U.S. Census Bureau. (2017b). *North American Classification System (NAICS)*. https://doi.org/10.1159/000443915

U.S. Department of Energy. (1998). Ecological Benchmark Tool. Retrieved August 24, 2020, from https://rais.ornl.gov/tools/eco_search.php

U.S. Environmental Protection Agency. (1978). *Source Assessment: Reclaiming of Waste Solvent, State of the Art - EPA Document # EPA-600/2-78-004f*. shorturl.at/pAJ46

U.S. Environmental Protection Agency. (1999). *RCRA , Superfund & EPCRA Hotline Training Module Introduction to : Boilers and Industrial Furnaces - EPA Document # EPA530-R-99-042* (tech. rep. October 1999). https://www.epa.gov/sites/production/files/2015-07/documents/bif.pdf

U.S. Environmental Protection Agency. (2002). *EPA Air Pollution Cost Manual - EPA Document # EPA/452/B-02-001* (6th ed.). https://doi.org/10.1126/science.203.4380.500

U.S. Environmental Protection Agency. (2004). Facility Registry Service (FRS). Retrieved October 22, 2020, from https://www.epa.gov/frs

U.S. Environmental Protection Agency. (2006). *Ethylene MACT Compliance Manual* (tech. rep.).

U.S. Environmental Protection Agency. (2013). TRI Basic Plus Data Files Guides. Retrieved October 22, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-plus-data-files-guides

U.S. Environmental Protection Agency. (2014). TRI Around the World. Retrieved September 16, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-around-world

U.S. Environmental Protection Agency. (2015). Waste Management Hierarchy and Homeland Security Incidents. Retrieved April 21, 2020, from https://www.epa.gov/homeland-security-waste/waste-management-hierarchy-and-homeland-security-incidents

U.S. Environmental Protection Agency. (2016). Chemical Data Reporting under the Toxic Substances Control Act. Retrieved August 20, 2020, from https://www.epa.gov/chemical-data-reporting

U.S. Environmental Protection Agency. (2017a). About the TSCA Chemical Substance Inventory. Retrieved August 16, 2020, from https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory

U.S. Environmental Protection Agency. (2017b). Pollution Abatement Costs and Expenditures: 2005 Survey. Retrieved August 20, 2020, from https://www.epa.gov/environmental-economics/pollution-abatement-costs-and-expenditures-2005-survey

U.S. Environmental Protection Agency. (2017c). Risk Evaluations for Existing Chemicals under TSCA. Retrieved August 16, 2020, from https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/risk-evaluations-existing-chemicals-under-tsca

U.S. Environmental Protection Agency. (2019). Green Chemistry Activities. Retrieved August 24, 2020, from https://www.epa.gov/trinationalanalysis/green-chemistry-activities

U.S. Environmental Protection Agency. (2020). *List of Lists: Consolidated List of Chemicals Subject to the Emergency Planning and Community Right- To-Know Act (EPCRA), Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) and Section 112(r) of the Clean Air Act* (tech. rep.). https://www.epa.gov/sites/production/files/2015-03/documents/list_of_lists.pdf

Vélez-Henao, J.-A., García-mazo, C.-m., Freire-gonz, J., & Font Vivanco, D. (2020). Environmental rebound effect of energy efficiency improvements in Colombian households. *Energy Policy*, *145*(August). https://doi.org/10.1016/j.enpol.2020.111697

Wang, L. K., Hung, Y.-T., & Shammas, N. K. (2005). *Physicochemical Treatment Processes* (Vol. 3). Humana Press. https://doi.org/10.1385/159259820x

Whittaker, M. H. (2015). Risk Assessment and Alternatives Assessment: Comparing Two Methodologies. *Risk Analysis*, *35*(12), 2129–2136. https://doi.org/10.1111/risa.12549

Yenkie, K. M., Burnham, S., Dailey, J., Cabezas, H., & Friedler, F. (2019). Generating Efficient Wastewater Treatment Networks: an integrated approach comprising of contaminant properties, technology suitability, plant design, and process optimization. In *Computer aided chemical engineering* (pp. 1603–1608). Elsevier B.V. https://doi.org/10.1016/B978-0-12-818634-3.50268-X

Zhang, M., & Guo, Y. (2013). Process simulations of NH3 abatement system for large-scale CO2 capture using aqueous ammonia solution. *International Journal of Greenhouse Gas Control*, *18*, 114–127. https://doi.org/10.1016/j.ijggc.2013.07.005

Zhu, X., Ho, C. H., & Wang, X. (2020). Application of Life Cycle Assessment and Machine Learning for High-Throughput Screening of Green Chemical Substitutes. *ACS Sustainable Chemistry and Engineering*. https://doi.org/10.1021/acssuschemeng.0c02211

5

A DATA ENGINEERING APPROACH FOR SUSTAINABLE CHEMICAL END-OF-LIFE MANAGEMENT

ABSTRACT

The presence of chemicals causing significant adverse human health and environmental effects during end-of-life (EoL) stages is a challenge for implementing sustainable management efforts and transitioning towards a safer circular life cycle. Conducting chemical risk evaluation and exposure assessment of potential EoL scenarios can help understand the chemical EoL management chain for its safer utilization in a circular life-cycle environment. However, the first step is to track the chemical flows, estimate releases, and potential exposure pathways. Hence, this work proposes an EoL data engineering approach to perform chemical flow analysis and screening to support risk evaluation and exposure assessment for designing a safer circular life cycle of chemicals. This work uses publicly-available data to identify potential post-recycling scenarios (e.g., industrial processing/use operations), estimate inter-industry chemical transfers, and exposure pathways to chemicals of interest. A case study demonstration shows how the data engineering framework identifies, estimates, and tracks chemical flow transfers from EoL stage facilities (e.g., recycling and recovery) to upstream chemical life cycle stage facilities (e.g., manufacturing). Also, the proposed framework considers current regulatory constraints on closing the recycling loop operations and provides a range of values for the flow allocated to post-recycling uses associated with occupational exposure and fugitive air releases from EoL operations.

Keywords: Circular life cycle; End-of-life management chain; Chemical flow analysis; Data pipeline

RESUMEN

La presencia de productos químicos que causan efectos adversos significativos para la salud humana y el medio ambiente durante las etapas de fin de vida (EoL, por sus siglas en inglés) es un reto para la aplicación de los esfuerzos de gestión sostenible y la transición hacia un ciclo de vida circular más seguro. Llevar a cabo una evaluación del riesgo químico y de exposición de los posibles escenarios de EoL puede ayudar a entender la cadena de gestión de EoL de las sustancias químicos, para su utilización más segura en un entorno de ciclo de vida circular. Sin embargo, el primer paso es rastrear los flujos químicos, estimar las liberaciones y las posibles vías de exposición. Por lo tanto, este trabajo propone un enfoque de ingeniería de datos de EoL para llevar a cabo el análisis del flujo químico y la selección, con el fin de apoyar la evaluación de riesgos y exposición para el diseño de un ciclo de vida circular más seguro de los productos químicos. Este trabajo utiliza datos disponibles públicamente para identificar posibles escenarios de post-reciclaje (por ejemplo, operaciones de procesamiento/uso industrial), estimar las transferencias químicas entre industrias y las vías de exposición a las sustancias químicas de interés. Una demostración de un caso práctico muestra cómo el marco de ingeniería de datos identifica, estima y rastrea las transferencias de flujo de sustancias químicas desde las instalaciones de la etapa EoL (por ejemplo, el reciclaje y la recuperación) a las instalaciones donde ocurre un etapa de post-reciclaje para un químico (por ejemplo, la fabricación). Asimismo, el marco propuesto tiene en cuenta las actuales restricciones normativas sobre las operaciones de cierre del bucle de reciclado y proporciona valores para el flujo de químico que es reciclado y procede de operaciones de EoL, el cual podría estar asociados exposición ocupacional y emisiones fugitivas al aire.

Palabras clave: Ciclo de vida circular; Cadena de gestión del fin de la vida útil; Análisis del flujo químico; Canalización de datos

5.1 INTRODUCTION

Sustainable materials management and circular economy have underpinned many initiatives to achieve sustainable economic systems and industrial value chains by improving resource productivity, source reduction, reuse, recycling, and prevention and minimizing waste generation and disposal rate (Corona et al., 2019; United Nations Industrial Development Organization, 2019). Governments worldwide have enacted policies to promote these initiatives (European Environment Agency, 2016; U.S. Environmental Protection Agency, 2012), which have created jobs and generated wages and tax revenues (U.S. Environmental Protection Agency, 2016c). However, the transition towards a circular economy still has many challenges to overcome (Bressanelli et al., 2019; Hopkinson et al., 2018). One of them is the presence of hazardous chemicals in end-of-life (EoL) flows that may enter the environment during recycling or unexpectedly accompany the ingredients of an industrial, commercial, or consumer product (Bodar et al., 2018). Additionally, the uncertainty at the EoL stage is a barrier both to know how much of a chemical flow can be recycled (Henriksson et al., 2010) and to characterize the risk that the circular life cycle may pose to human beings and the environment due to exposure to chemicals of concern (Alaranta & Turunen, 2020; de Römph & van Calster, 2018; Wassenaar et al., 2017). For example, bromine and brominated flame retardants that may be contained in plastic products may affect the quality of waste plastics as secondary materials; therefore, recycling may reintroduce bromine and brominated flame retardants into the new plastic product cycle and lead to increased exposure levels (Pivnenko et al., 2016).

Material flow analysis has been widely used for EoL flow tracking in a circular life cycle (Chertow & Park, 2015; Tanzer & Rechberger, 2019) and chemical release quantification during exposure assessment (Bornhöft et al., 2013; Gottschalk et al., 2010). Previous work developed a data engineering framework based on publicly-available databases to track chemical flows generated by industrial facilities and transferred to other off-site locations for further EoL management. The framework leverages facility-level information to estimate the potential release quantity of a transferred chemical from EoL operations (e.g., recycling) at an off-site facility (Hernandez-Betancur et al., 2020). Another work employs a data engineering approach on a bottom-up chemical flow analysis (CFA) of industrial pollution abatement units (PAUs) (e.g., batch still distillation). This PAU-level approach allocates a chemical downstream of a PAU. The allocation considers the PAU effect on a chemical of interest and the expected flows from a PAU (e.g., wastewater from solvent recovery) (Hernandez-Betancur et al., 2021). Although both frameworks could work synergistically to perform an EoL

CFA rapidly, they do not provide the potential paths and destination of a recycled chemical flow.

The databases which are part of the openLCA Nexus, like ecoinvent (GREENDELTA, 2013), which is the most widely used database by LCA practitioners (Vélez-Henao et al., 2020), have supported analyses towards a safer circular life cycle (Bech et al., 2019; Lucchetti et al., 2019; Meys et al., 2020; Rigamonti et al., 2017). However, these databases focus on assessing environmental impacts associated with the inputs and outputs for obtaining products and services instead of tracking chemical flows for potential release estimation and exposure assessment (Wedema et al., 2013). Efforts from policy-makers, academia, and corporations like RISKCYCLE (Waste Management Alternatives Development and Exchange Forum, 2012), Chemical Leasing (Perthen-Palmisano & Jakl, 2004; United Nations Industrial Development Organization, 2013), and others have looked for designing sustainable chemical supply chains and accelerating the shift to a circular life cycle, considering the traceability and risk of chemicals (De Groene Zaak, 2015; Together for Sustainability, 2013). Nevertheless, the studies developed started targeting specific industry sectors like textiles in order to find hazardous chemicals in their products (Grundmann et al., 2013; Lahl & Zeschmar-Lahl, 2013), or searching for particular chemicals/chemical categories in recycled products, for example, phthalates or other persistent organic pollutants in recycled plastics (Leslie et al., 2016; Pivnenko et al., 2016) or polycyclic aromatic hydrocarbons in recycled rubber (European Chemicals Agency (ECHA), 2017; Pronk et al., 2020). Although these analyses can provide accurate results, they are case-by-case approaches and time-consuming tasks, and they require significant resources for running experiments. These aspects are disadvantages for conducting a rapid chemical risk evaluation and exposure assessment of a chemical of interest.

This work proposes an EoL data engineering framework to connect siloed publicly-available database systems to estimate releases and track chemical flow transfers from EoL stage facilities (e.g., recycling and recovery) to upstream chemical life cycle stage facilities (e.g., manufacturing). Also, the current approach performs a screening to identify potential sustainable EoL management and circular life cycle exposure scenarios for recycled chemicals and could be extended by using data-driven models. The scenarios consider whether the recycled chemical might end up in industrial, commercial, and consumer uses and processing operations. Additionally, the framework proposes a methodology that leverages the data to identify if the recycling activities target the chemical under analysis besides the other chemicals constituting the EoL multicomponent material flow transfers. This methodology can be connected to publicly-available frameworks for near-field exposure assessment (U.S. Environmental Protection Agency, 2015c, 2015d). The advantage of the framework is that it automatizes the above searches reducing the need for extensive user input and hands-on knowledge (Smith et al., 2019). This contribution also describes the estimation of chemical flows for including EoL performance indicators and release estimations (Hernandez-Betancur & Ruiz-Mercado, 2019).

5.2 BACKGROUND

5.2.1 *Tracking chemical flow transfers: facility-level information*

A data engineering framework previously developed tracks chemicals contained in industrial EoL flows generated due to the chemical manufacturing (e.g., as a byproduct), importing, processing operations (e.g., as a reactant), and industrial uses (e.g., used as a cleaner) (Hernandez-Betancur et al., 2020). The framework tracks chemical flow transfers from the EoL flow Generator facility to the recycling, energy recovery, treatment & disposal facility (RETDF) situated across different locations in the U.S. The off-site chemical flow tracking maps the RETDF in case of EoL brokering scenarios. In the case of recycling, the identification of the RETDF can help determine whether the RETDF recycles a chemical for sale, distribution, or on-site use. As a result of the tracking, the framework provides the type of EoL activity (e.g., solvents/organics recovery) and both the Generator and RETDF industry sectors. Identifying the industry sectors determines the link between them as nodes of the chemical EoL management chain. Furthermore, the framework provides chemical and RETDF identification numbers to connect with additional information like the PAU functionality (see Section 5.2.2). Moreover, the framework leverages facility-level information to estimate how much of a chemical flow transferred to a RETDF might be released to the environment during an EoL activity. Equation 5.1 is how the framework estimates emission factors by leveraging facility-level information (Hernandez-Betancur et al., 2020).

$$\begin{gathered} TR_k = \left\{ F_T \times EF_k = F_T \times \frac{F_{R-k}}{\Delta AC + F_{TW}}, k = 1,2,3,4 \right\| \\ \Delta AC \in [MAX(F_{TR} - F_{TW}, -MAX(QM)), MAX(QM)] \\ \left. \cap QM \in P_{MC} \right\} \end{gathered} \tag{5.1}$$

In Equation 5.1, k represents the release compartments considered by the framework, i.e., fugitive air, on-site surface soil, on-site surface water, and stack air releases. Furthermore, TR_k represents the indirect chemical release to the compartment k, F_T is the annual quantity of chemical transferred by a Generator, F_{TW} is the RETDF total generated waste, F_{R-k} is the RETDF chemical flow releases to compartment k, ΔAC is a range of potential values for the annual increment in the amount of the chemical present at the RETDF, F_{TR} is the RETDF total chemical releases to compartments, P_{MC} is a range code that represents the maximum amount of chemical present at RETDF during a reporting year, and QM is a value within the range described by P_{MC}.

5.2.2 *Chemical flow allocation at PAU technologies: PAU-level information*

A second framework uses an equivalent principle to the one presented in Section 5.2.1 (Hernandez-Betancur et al., 2021). However, this framework relies on PAU-level information (process units) instead of aggregated facility-level information. This framework regards the effect of a PAU technology on a chemical of interest; for instance, an incinerator destroys organic chemicals but not metals. Additionally, using equipment process design catalogs, guidelines, and reports about PAUs, the framework determines generic PAU output flows. Thus, the framework allocates a chemical of interest that passes through a PAU technology into fugitive air release flow, effluent flow, destroyed, recycled, or recovered flow, and release flows. The effect and abatement efficiency of a PAU on a chemical determine the destroyed, recycled, or recovered flows. The framework considers the chemical water solubility, melting and boiling point, and whether the chemical is a metal compound for the allocation. Moreover, the framework suggests a phase for the output flows (e.g., liquid waste). Equation 5.2 is the generic chemical flow balance proposed by the PAU framework. In Equation 5.2, F_{input} is the chemical input flow, η is the PAU abatement efficiency relative to the chemical of interest, β is the emission factor for the chemical fugitive air release, $F_{waste/release}$ is the chemical flow in the PAU waste or release (excluding fugitive air), and $F_{effluent}$ is the chemical flow in the effluent. Therefore, this framework determines η given the input chemical concentration (Hernandez-Betancur et al., 2021).

$$F_{input} = \eta \times F_{input}^{(c)} + (1-\eta) \times \beta \times F_{input} + F_{waste/release} + F_{effluent} \tag{5.2}$$

5.3 METHODOLOGY

The developed frameworks can operate alone to build data-driven models, perform generic CFA for PAUs and chemical transfers, and incorporate decision-making tools into chemical EoL management. Therefore, the new proposed development combines both frameworks and extends the CFA to track chemicals, estimate releases, and potential exposure pathways in flow transfers from EoL stage facilities (e.g., recycling) to upstream chemical life cycle stage facilities (e.g., manufacturing). This framework can also determine recycling activities and their chemical flows that might be transformed into new products in a circular world while possibly assessing, reducing, or eliminating hazards.

5.3.1 *Previous methodology enhancements*

5.3.1.1 *ΔAC statistical distribution*

As presented in Equation 5.1, ΔAC is a range of values so that one can take ΔAC as an uncertain variable. In addition, information entropy is a measure of the uncertainty and is maximum when the outcome of a random process is equally likely, which is the case for a uniform distribution (MacKay, 2002). Therefore, to reduce the ΔAC uncertainty, this work assumes the mode value of the statistical distribution that describes ΔAC should be around zero. This means that it is more likely the annual chemical material balance at RETDF is at steady state, which is a reasonable assumption for operating periods above 1 yr (Honrath, 2020). Nevertheless, the framework computes and places the mode value of ΔAC (ΔAC_{mode}) around zero by using information from the Toxics Release Inventory (TRI) about the source reduction activities for a chemical and the annual change of the production or activity that is the primary influence on the quantity of chemical at RETDF (U.S. Environmental Protection Agency, 2013). Hence, for each chemical at each RETDF, the framework gathers information associated with whether the RETDF uses source reduction activities to decrease the mass of chemical present at the facility, the estimated percentage of reduction (S_R), and the ratio of production or activity in the reporting year divided by production or activity in the previous year (PA_R). Equation 5.3 shows how the framework leverages the RETDF information to place ΔAC_{mode}. In Equation 5.3, L_{mode} is a value which depends on both PA_R and S_R. Equation 5.3 uses a logistic function to allocate the ΔAC_{mode} value inside the interval described in Equation 5.1. This work uses a triangular distribution to accurately represent the whole range of potential ΔAC values. The triangular distribution enables to work with functions either symmetric (i.e., $\Delta AC_{mode} = 0$ = 0 and $|MAX(F_{TR} - F_{TW}, -MAX(QM))| = MAX(QM)$ or non-symmetric (e.g., $\Delta AC_{mode} = 0$ but $|MAX(F_{TR} - F_{TW}, -MAX(QM))| \neq MAX(QM)$.

$$\Delta AC_{mode} = \begin{cases} 0, \, if \, L_{mode} = 0 \\ MAX(QM) \times L_{mode}, \, if \, L_{mode} > 0 \\ MAX(F_{RT} - F_{TW}, -MAX(QM)) \times |L_{mode}|, \, if \, L_{mode} < 0 \end{cases}$$
$$L_{mode} = logistic(PA_R) - logistic(S_R + 1) \tag{5.3}$$

5.3.1.2 *Uncertainty for brokering scenario pathways*

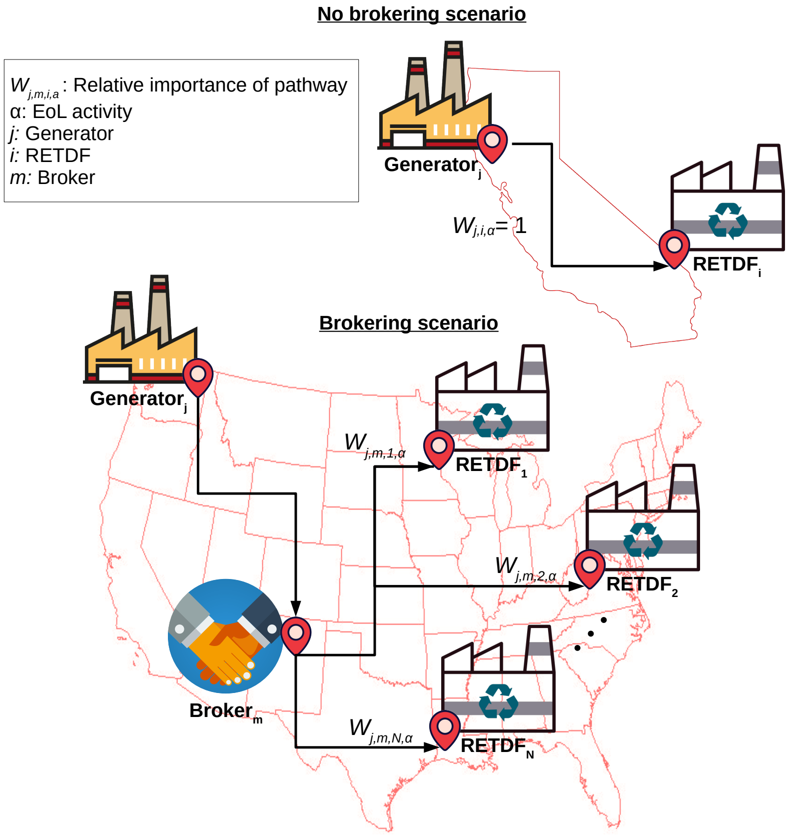


Figure 5.1: Analysis for the degree of certainty or relative importance for both no brokering and brokering scenarios, considering the geographical location of the Generators, Brokers, and RETDFs.

Although the framework developed by Hernandez-Betancur et al. (2020) maps the potential RETDFs for brokering scenarios, it does not indicate information about the uncertainty that a chemical flow transfer by a Generator passes through one of the identified RETDFs. This work develops a methodology to assign a degree of uncertainty or relative importance to each potential pathway that a chemical flow transfer can follow in EoL brokering cases. Figure 5.1 depicts the analysis for no brokering and brokering scenarios. For no brokering, $Generator_j$ transfers a chemical directly to $RETDF_i$ $(i = 1, \ldots, N)$ for EoL activity α. Therefore, the relative importance of the pathway ($W_{j,i,\alpha}$) is equal to one, i.e., the pathway is clearly defined. For brokering, $W_{j,m,i,\alpha}$ is less than one and greater than or equal to zero and depends on the $Broker_m$.

TRI is the most comprehensive worldwide publicly-available Pollutant Release and Transfer Register regulatory system. Its year of data availability and granularity of release data elements enable gathering helpful information for tracking chemical transfers, considering chemical quantities transferred, EoL activity, reporting year, facilities, and the location of the facilities involved in these transfers (U.S. Environmental Protection Agency, 2014). Thus, this work uses this information to overcome the epistemic uncertainty associated with brokering scenarios using a Fuzzy Analytic Hierarchy Process (FAHP) and considering four criteria to determine potential EoL pathways (Hernández-Betancur et al., 2019). Each pathway found for brokering scenarios is an alternative assessed by FAHP under the four criteria to obtain a value for $W_{j,m,i,\alpha}$ by applying fuzzy operations (Sabaghi et al., 2016).

Hernandez-Betancur et al. (2020) mapped a potential $RETDF_i$ for $Generator_j$ using the available reporting years for the publicly-available information like TRI and selecting the $RETDF_i$ reporting information closest to the reference year. Therefore, the first criterion for the FAHP is a temporal correlation indicator assigned to each $RETDF_i$, based on the data uncertainty analysis (quality pedigree matrix) (Edelen & Ingwersen, 2016). The second criterion considers how many times the relationship $Broker_m$ - $RETDF_i$ appears in the available information. The third criterion considers the times the EoL activity α associated with $RETDF_i$ appears in the records for the chemical of interest. The EoL activity α is why $Broker_m$ transfers the chemical to the $RETDF_i$. The fourth and last criterion is the cost of transporting a EoL chemical from $Generator_j$ to $RETDF_i$.

The transportation cost considers F_T as well as the length of the traveling route. The traveling route is based on the geographical location of facilities, as presented in Figure 5.1. The framework collects facility latitude and longitude from TRI and the other sources used by Hernandez-Betancur et al. (2020). In cases not found in those sources, the addresses, zip codes, cities, and counties help make queries to the open-source geocoding API of the Nominatim project, thereby obtaining the facility latitudes and longitudes (contributors, 2011a). The current work uses the shortest route between geographical locations by ground transportation, retrieving it via the OSRM project API (contributors, 2011b). Hence, F_T (i.e., the annual quantity of chemical transferred) and the shortest ground transportation distance are employed together in a methodology developed by the U.S. Environmental Protection Agency for calculating ground transportation cost of hazardous waste by stake trucks (Abkowitz et al., 1985). The framework includes maritime transportation costs, e.g., for transfers from Florida to Puerto Rico. Thus, the haversine formula calculates the maritime transport distance (Shylaja, 2015), considering the locations of container and tonnage ports in the U.S. (U.S. Department of Transportation, 2018). Furthermore, the framework takes from the Organization for Economic Co-operation and Development statistics, the average maritime transportation cost of manufacturing goods that include commodities like organic chemicals (Organization for Economic Co-operation and Development, 2010). The framework uses the degree of uncertainty in the methodology explained in Section 5.3.3.

5.3.2 *Chemical activities and uses for recycled chemicals*

Closing the recycling loop requires information about the possible uses and activities for a chemical of interest once recycled. The framework leverages the information from the Chemical Data Reporting (CDR) database, which contains information about chemicals domestically manufactured in the U.S. or imported into U.S. territory (including byproducts and impurities) (U.S. Environmental Protection Agency, 2015a). This database information connects to the frameworks presented in Sections 5.2.1 and 5.2.2 by using the Facility Registry Service (U.S. Environmental Protection Agency, 2004) and the Substance Registry Services (U.S. Environmental Protection Agency, 2016d). The former service system provides the identification numbers and industry sectors for the CDR reporting facilities. The industry sector classification for the CDR facilities is according to the 6-digit code provided by North American Industry Classification System (NAICS) (U.S. Census Bureau, 2017b), which is used by the framework as shown in Sections 5.2.1 and 5.2.2. Hence, CDR information links to the RETDFs, either using the facility identification number (preferable) or NAICS code. The second service system supports relating CDR substances to TRI chemicals by a unique chemical identification number (Hernandez-Betancur et al., 2020).

Thus, CDR supplies information about potential product categories for commercial and consumer uses (e.g., personal care products). Moreover, CDR provides the industrial processing and operations (e.g., repackaging), industrial function categories for the chemical compounds (e.g., odor agents), and industry sectors for facilities downstream the CDR reporting facilities (U.S. Environmental Protection Agency, 2016a). The framework uses these industry sectors to determine what are called the post-recycling industry sectors, which may buy the materials containing the recycled chemicals sold by the REDTFs. Unfortunately, in some cases, CDR reporting facilities only provide a description of the industry sectors or use an industry sector classification that is different from the NAICS codes (U.S. Environmental Protection Agency, 2009). The framework processes and transforms the description texts and uses web automatization to search for the potential NAICS codes which satisfy those descriptions. Thus, the framework standardizes the industry sector classification across the chemical EoL management chain and recycling loop, i.e., the Generator, RETDF, and post-recycling industry sectors. In some instances, the NAICS codes obtained for the post-recycling industry sectors are not the 6-digit ones, which is the most specific in the NAICS hierarchical structure (U.S. Census Bureau, 2017a). Therefore, this research uses the number of industry sector facilities from the 2017 Statistics of U.S. Businesses (U.S. Census Bureau, 2020). In the methodology described in Section 5.3.3, this information helps to select a 6-digit NAICS code for the potential post-recycling industry sectors retrieved from CDR, based on the probabilities proportional to the number of facilities. For instance, the Bituminous Coal Underground Mining industry sector (6-digit NAICS code 212112) might be chosen if the post-recycling industry sector, according to CDR, is Mining, Quarrying, and Oil and Gas Extraction (2-digit NAICS code 21).

5.3.3 *Relationship between the EoL management chain elements and recycling loop*

Figure 5.2A depicts a generic CFA for a chemical across its life cycle stages, i.e., manufacturing or importing, processing, uses (industrial, commercial, and commercial), and EoL. The framework focuses on the chemical flows represented by the continuous black arrows in Figure 5.2A. The framework tracks the chemical flow transfers according to Section 5.2.1 and allocates them as Section 5.2.2 describes. The chemical EoL management chain and the recycling loop consider seven critical elements represented by the light blue circles in Figure 5.2A. The generator, the RETDF industry sectors, and the EoL activity engage in the chemical flow transfer. Furthermore, in the case of recycling, the EoL activity and the RETDF industry sector enable connecting with the other four elements in the recycling loop, i.e., the post-recycling industry sector, industrial function category, industrial processing or use operation, and commercial and consumer product category, described in Section 5.3.2. The last four elements depict the chemical flow transfer, which goes back into the value chain and closes the recycling loop.

The framework incorporates a Markov Random Field (MRF), a probabilistic graphical model with undirected edges (Koller & Friedman, 2009), to represent the correlation between the chemical EoL management chain elements and the recycling loop. Unlike directed graphical models like Bayesian Networks, MRF allows the framework to depict the relationship between the EoL management chain and recycling loop elements without forcing a specific direction to the influence between them. For example, although the paper manufacturing industry sector recycles chlorine, chlorine recycling is not restricted to this sector and vice versa. In addition, MRF is less data-demanding than an undirected discriminative graphical model like Conditional Random Fields (Li et al., 2021). Conditional Random Fields focus on the posterior distribution of a label sequence for a known observation array, while MRF lets describe the network and its joint statistical distribution by factorization (Koller & Friedman, 2009; Li et al., 2021). Thus, MRF results adequate considering that the exact sequence is not well-known for an example and the relationship between the elements in the chemical EoL management chain elements and the recycling loop come from siloed database systems. Figure 5.2B presents the MRF structure composed of the seven elements described before. The elements are the blue light nodes in the MRF, and the undirected red edges are their relationships. Although the three sectors (i.e., generator, RETDF, and post-recycling) may be the same, their representation as separate nodes in the MRK allows such a result to be possible, but not a restriction. Therefore, MRF can establish the relationship between the chemical EoL management chain elements and the recycling loop. Also, the framework employs Markov Chain Monte Carlo (MCMC) to obtain local statistical distributions for determining potential pathways between the EoL management chain and recycling loop elements, thereby inferring about the global distribution for the MRF (joint probability distribution) (Winkler, 2003).

The framework presented here is not definitive and future developments can incorporate additional data sources and more robust models for data-driven modeling and decision-making to reduce uncertainty, improve estimates, and extend the framework to chemicals that are not in the data source. The models presented here would enable identify limitations and implications for future developments, and they should be seen for an

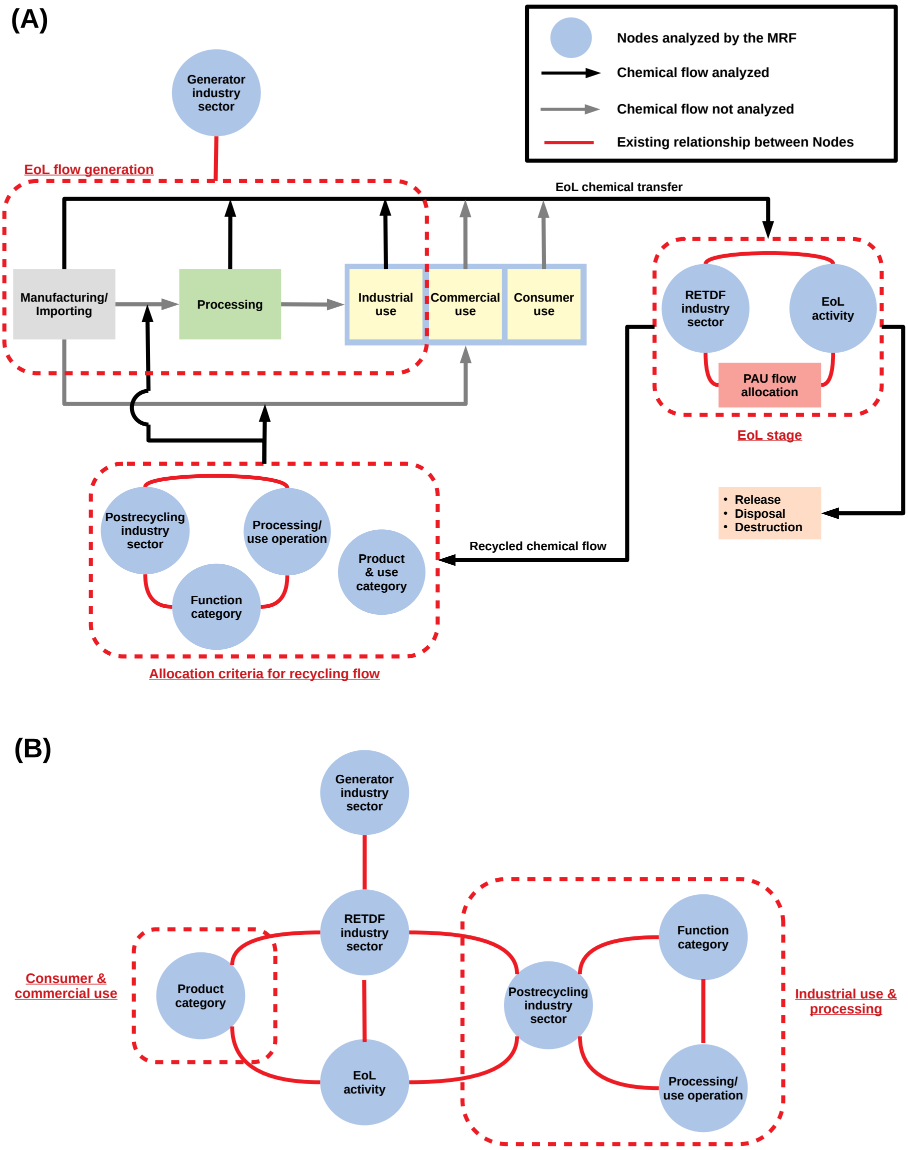


Figure 5.2: A schematic explanation for the proposed methodology. (A) Generic CFA across a chemical life cycle and its connection with the proposed data engineering approach. (B) Markov random field (MRF) establishes the relationship between the chemical EoL management chain elements and the recycling loop.

exploratory data analysis. Additionally, although the framework takes advantage of publicly-available data to determine whether RETDFs, in the case of recycling, can sell the recycled chemical, the framework does not incorporate information to determine the waste management capacity of

RETDFs to either fully manage the waste received or to sell and ship a portion to other RETDFs yet. Figure C.1 in the Appendix C depicts a generic scheme that outlines the uncertainty propagation both for identifying the EoL management chain and recycling loop elements and performing the CFA.

5.4 CASE STUDIES

This section presents a case study based on n-hexane (CAS 110–54–3) to demonstrate the framework usefulness for performing CFA, tracking the chemical across the chemical EoL management chain, and closing the recycling loop. N-hexane is part of the Organization for Economic Co-operation and Development list of high production volume chemicals, which are chemicals produced at levels greater than $1.00 \times 10^6 kg/yr$ in at least one organization member (Organization for Economic Co-operation and Development, 2004). N-hexane major global use is as a component in fuels and other petroleum products; however, its physicochemical properties like high solubility and low cost make n-hexane ideal for many petroleum products applications. For instance, as an edible-oil extractant for various seed crops like soybeans, cottonseed, and corn germ (Canadian Environmental Protection Agency, 2009) and solvent in biodiesel production (Martín & Grossmann, 2012). Moreover, this chemical is associated with toxicity hazards to aquatic life and damage to organs and fertility functions (U.S. Department of Health and Human Services, 1999). Therefore, this chemical results in a relevant case study due to its applications and environmental and health concerns.

This case study uses the information for n-hexane obtained by the framework explained in Sections 5.2 and 5.3. The reference reporting year is 2018, which corresponds to the last TRI report at the study time. MRF enables analyzing the chemical EoL management chain and recycling loop for n-hexane. Due to a trade-off between achieving statistical significance and computing power limitations, the case study takes 100 cycles to infer and draw samples by the MCMC. Python scripts support data engineering, web automation, FAHP, MRF and inference, and CFA. These scripts are in a public GitHub repository (see Section C.2 in the Appendix C). The case study analyzes first the global state of the n-hexane EoL management chain. The case study also gives an idea of the potential performance of the n-hexane recycling loop and determines if n-hexane is a chemical of interest in the recycling activities for which the generator industry sector transfers the chemical to the RETDF sector. The last aspect is based on the chemicals found in the information submitted by the RETDFs to the TRI

Program, considering whether the recycling PAU technologies reported for n-hexane process other chemicals, e.g., a RETDF reports the use of fractionation for n-hexane and 1-butanol.

5.5 RESULTS AND DISCUSSION

Tracking chemical flow transfers enables identifying potential exposure scenarios and estimating releases in the chemical EoL management chain. Moreover, it allows moving forward to analyze potential pathways for chemicals in the recycling loop, considering potential risk due to the unintended recycling of their toxic contents, and identifying if the recycling activities mainly target the chemical under analysis besides the other chemicals constituting the multicomponent EoL flow transfers. Therefore, Section 5.5.1 shows the most probable scenario both for the n-hexane EoL management chain and for the recycling loop. Section 5.5.2 presents aspects of the inter-industry EoL transfers and recycled flows containing n-hexane. Finally, Section 5.5.3 depicts a way to analyze potential risks due to the unintended recycling of toxic contents in the recycling loop.

5.5.1 *Generic performance for the EoL management chain and recycling loop*

A finding from the development of the case study is associated with representing the MRF factor or potential functions by using the data. These potential functions depict the affinity or the strength of the relationship occurrence; for example, the automobile manufacturing industry sector transfers EoL chemical flows to the materials recovery facilities. The number of times a relationship appears in the data can misinterpret the affinity between the EoL chemical flow generator and the receiver. Table C.1 (in Section C.3 in the Appendix C) presents examples for the values of two variables in the MRF for n-hexane, i.e., generator and RETDF industry sectors. Table C.1 shows that according to the times of appearance, n-hexane transfers from facilities in the hazardous waste treatment & disposal sector to facilities in the same sector have a greater affinity (5 times) than from all other basic organic chemical manufacturing to the hazardous waste treatment & disposal (3 times). However, if the flow transfer is considered, the last relationship would have a higher affinity ($9.72 \times 10^4 kg/yr$) than the initial one ($1.72 \times 10^3 kg/yr$).

Additionally, using the number of times of appearance leads to the belief that transfers from the other chemical and allied products merchant wholesaler sector would have the same affinity than from all other basic organic chemical manufacturing to the hazardous waste treatment & disposal (3 times). Nevertheless, for this case, the former relation has $1.14 \times 10^2 kg/yr$ of affinity, i.e., only 0.12% of the last relation affinity. The above finding leads to select the multiplication between the number of appearances and the transfer flow as the value for the MRK potential function. The flow transfer can reflect and enable the inclusion of regulatory aspects associated with legitimate hazardous waste recycling operations like tolling and contractual agreements, speculative accumulation before recycling, and adaptation of the facilities to recycle hazardous wastes (U.S. Environmental Protection Agency, 2015b). Also, it can show whether the chemical is a processing or manufacturing impurity.

Table 5.1: General results for the CFA of n-hexane by using data engineering and the MRK. This table presents n-hexane transfers for waste management in 2018 from TRI Explorer (data released April 2020) (U.S. Environmental Protection Agency, 2020).

Query	Value
Total n-hexane transferred according to data engineering [kg/yr][1]	3.75×10^6
Total n-hexane transferred according to the TRI Explorer [kg/yr][1,2]	4.23×10^6
Total n-hexane transferred for recycling according to data engineering [kg/yr][1]	1.70×10^5
Total n-hexane transferred for recycling according to the TRI Explorer [kg/yr][1]	2.49×10^5
Average recycled flow for industrial activities [kg/yr]	1.49×10^5
Average recycled flow sold [kg/yr]	9.19×10^4

[1] The differences between values reported by TRI Explorer and the ones obtained by data engineering are because the RETDFs could not be tracked (Hernandez-Betancur et al., 2020).

[2] Excluding transfers to sewage treatment plants. These transfers are not part of the data engineering framework developed by Hernandez-Betancur et al. (2020).

Table 5.1 presents the data engineering framework tracking n-hexane transfers for further waste management is equivalent to $3,75 \times 10^6 kg/yr$, 88.86% of the total n-hexane is transferred off-site for further waste management reported in the TRI Explorer (U.S. Environmental Protection Agency, 2020). Moreover, the framework tracks $1.70 \times 10^5 kg/yr$ of n-hexane transferred for recycling, 68.53% of the value reported in the TRI Explorer. Hence, tracking the flows transferred off-site for recycling and identifying the RETDF in charge of this activity is more challenging than other EoL activities. Nevertheless, for this chemical, the ratio between recycling flow transfers and the total one is similar for both information sources, 4.54% and 5.88% for data engineering and TRI Explorer, respectively. These numerical differences between values might be because some RETDFs do not need to report to the TRI Program. Therefore, the framework cannot close the recycling loop. However, future development and further research is expected to overcome this limitation.

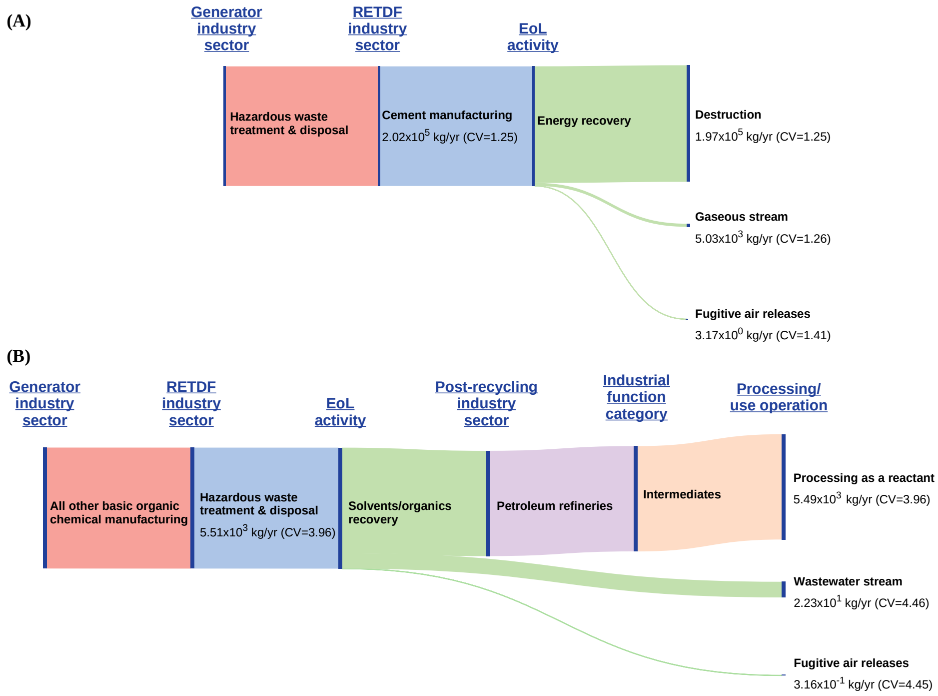


Figure 5.3: CFA for n-hexane in the EoL management chain and recycling loop. (A) CFA for the most probable pathway in the MRK. (B) CFA for the most probable recycling path determined by the MRK. CV represents the coefficient of variation of the quantity presented.

Due to the length limits of the manuscript and to facilitate the visualization and explanation of the results in this section, not all the potential pathways in the n-hexane EoL management chain and recycling loop are presented. Instead, Figure 5.3 depicts the two scenarios, whose most relevant details are discussed in this section. Figure 5.3A shows the CFA for the most probable n-hexane EoL pathway according to the MRF described in Figure 5.2B, while Figure 5.3B represents the most probable n-hexane pathway when it is established in the MRF that the pathway must contain recycling activities.

In Figure 5.3A, the most probable n-hexane EoL pathway is associated with energy recovery from EoL flows containing n-hexane transferred by the hazardous waste treatment and disposal industry sector (generator industry sector) and harnessed by the cement manufacturing industry sector (RETDF industry sector). Because of the generator sector, the waste containing n-hexane can be categorized as hazardous, leading to satisfying special regulatory requirements and making it unsafe for recycling activities (U.S.

Department of Health and Human Services, 1999). Moreover, the regulation considers the chemical heating value to differentiate between legitimate energy recovery and incinerator (U.S. Environmental Protection Agency, 2019). Thus, the combustion heat of liquid n-hexane ($-4,163 \pm 20 kJ/mol$) (National Institute of Standards and Technology, 1987) and its average flow transferred for this scenario ($2.02 \times 10^5 kg/yr$) can make energy recovery a feasible EoL activity.

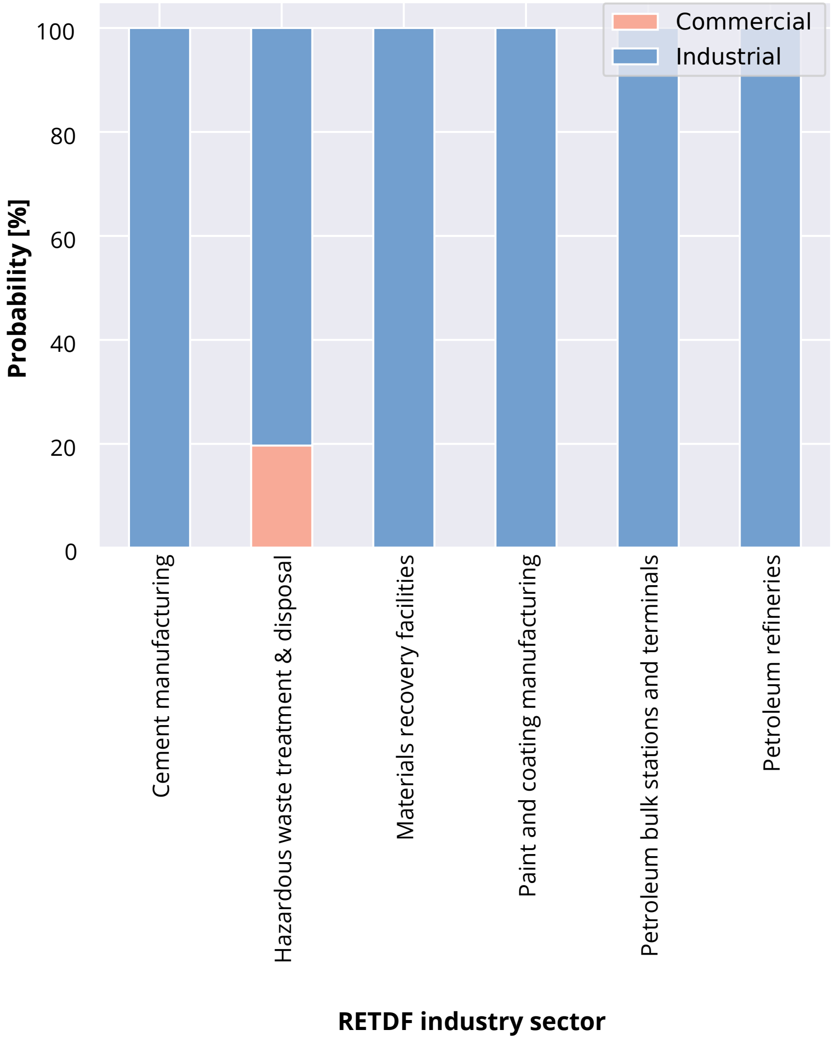


Figure 5.4: The probability the recycled n-hexane flow ends in industrial processing/use operations, commercial uses, or consumer uses given the RETDF industry sector.

Furthermore, industrial kilns from the cement manufacturing industry sector are considered legitimate waste-to-energy management equipment (U.S. Environmental Protection Agency, 2019). The above facts prove that considering the linkage between the industry sectors involved in an EoL chemical transfer may help understand and predict potential EoL scenarios for chemicals. Figure 5.3A presents the results of the n-hexane allocation for this scenario. On average, the energy recovery processes do not destroy around $5.03 \times 10^{3} kg/yr$ of n-hexane, which is 2.49% of the average flow transferred for this scenario ($2.02 \times 10^{5} kg/yr$). This undestroyed gaseous flow may end up as stack air releases if there is no further treatment. This value presents a high variability, as the coefficient of variation (CV) is equal to 1.26. Hence, by adding and subtracting the average value times its CV ($\pm 5.0310^{3}1.26$) to the average value ($5.03 \times 10^{3} \pm 5.03 \times 10^{3} \times 1.26$) the expected gaseous flow values are in the interval $[0, 1.14 \times 10^{4} kg/yr]$; therefore, providing low-end and high-end scenarios for a further exposure assessment (U.S. Environmental Protection Agency, 1991).

As presented in Figure 5.3A, the most probable current scenario for n-hexane is not associated with recycling. However, as mentioned above, 4.54% of the amount of n-hexane transferred was recycled in 2018. Figure 5.3B depicts that the most potential recycling pathway for n-hexane is related to transfers from facilities in all the other basic organic chemical manufacturing sector (generator industry sector) to facilities in the hazardous waste treatment & disposal sector (RETDF industry sector). The generator sector transfers on average $5.51 \times 10^{3} kg/yr$ (CV = 3.96) of n-hexane for solvent/organic recovery. Moreover, for the recycling scenario presented in Figure 5.3B, petroleum refineries are the post-recycling industry sector of choice. Since the RETDF and the post-recycling industry sectors are different for this case, this scenario consists of the RETDF sector selling the recycled n-hexane instead of using it on-site (e.g., as a reactant or solvent). As shown in Table 5.1, the average recycled flow sold by RETDFs is $9.19 \times 10^{4} kg/yr$, i.e., corresponding to 54.06% of the n-hexane recycled. Therefore, there is a probability just above 50% that an RETDF sells the recycled n-hexane.

According to Figure 5.3B, on average, for this scenario, $5.49 \times 10^{3} kg/yr$ would be the net n-hexane recycled amount, i.e., around 99.64% of the n-hexane transfer. The petroleum refineries sector may use such a recycled n-hexane as an intermediate (industrial function category) in a chemical reaction (processing/use operation). Hence, recycled n-hexane may go back to industrial activities, which would be expected due to the industrial source of the EoL flow containing n-hexane to reduce the risk associated with hazardous waste recycling (U.S. Environmental Protection Agency, 2015b). Likewise, Table 5.1 presents that on average $1.49 \times 10^{5} kg/yr$ of the n-hexane transferred for recycling ends up in industrial activities. Figure 5.3B allows visualizing that this scenario generates wastewater and fugitive air releases. Hence, the wastewater generated relates to potential on-site surface water releases or further transfers to sewage treatment plants if there is not further on-site treatment at the RETDFs. As described for the undestroyed gaseous flow, by adding and subtracting the average value times its CV ($\pm 2.23 \times 10^{1} \times 4.46$) to the average value ($2.23 \times 10^{1}$), this wastewater flow may be in the interval described by $[0, 1.22 \times 10^{2} kg/yr]$. Also, fugitive air releases that would affect occupational receptors at the RETDFs, with an average value of 3.16×10^{-1} and CV of 4.45, may lie in the interval $[0, 1.72 \times 10^{0} kg/yr]$.

Like Table 5.1, Figure 5.4 depicts the RETDF industry sectors mainly recycle n-hexane sent back to industrial processing/use operations. Nevertheless, Table 5.1 indicates that on average $2.10 \times 10^{4} kg/yr$ of the recycled n-hexane may flow to commercial/consumer uses (12.35% of the recycled flow). As shown in Figure 5.4, this linkage of recycled n-hexane and fresh n-hexane comes from the hazardous waste treatment & disposal sector. The probability the n-hexane recycled by this sector goes to commercial uses is around 20.00%. As shown in Section 5.5.2, this RETDF industry sector receives a large amount of n-hexane flow transferred for recycling, increasing the chance of such flows being transferred to commercial activities. Additionally, the MRF in Figure 5.2B does not include potential composition and functionality differences between recycled and brand-new product flows. Future research should incorporate this aspect into the MRF variables that describe the chemical EoL management chain and its recycling loop.

5.5.2 *Inter-industry sector transfers for n-hexane*

As presented in Section 5.5.1, there is a high chance the n-hexane flow transfers for recycling go back to industrial processing/use operations. An important aspect to capture the behavior of the EoL management chain and recycling loop is understanding the inter-industry sector transfers. Hence, drawing only the samples from the MCMC that indicate transfers that would end up into industrial processing/use operation, Figure 5.5 is elaborated for inter-industry transfer flows whose mean value is greater or equal to $1.00 \times 10^{3} kg/yr$ for visualization purposes. Figure 5.5A is for inter-industry EoL n-hexane transfer for recycling, while Figure 5.5B is for n-hexane recycled transfers. Figure 5.5 shows that a fragment on the outer part of the circular layout represents each industry sector (a chord diagram node). In the figure, the arcs represent the inter-industry sector connections. The size of each arc is proportional to the inter-industry sector transfer flow importance. Also, the arc color is the same as the color of the parent node.

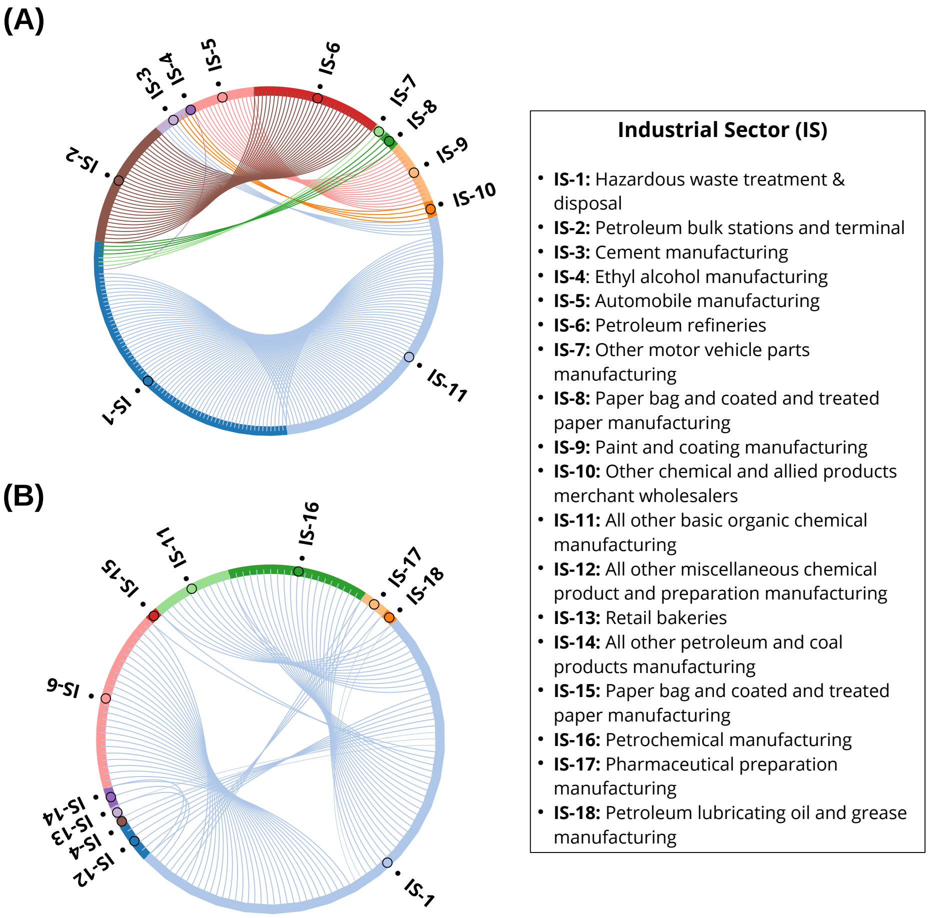


Figure 5.5: Inter-industry n-hexane transfer flows. A fragment on the outer part of the circular layout represents each industry sector (a chord diagram node). The arcs represent the inter-industry sector connections. The size of each arc is proportional to the inter-industry sector transfer flow importance. Also, the arc color is the same as the color of the parent node. (A) transfer of EoL n-hexane flows between industry sectors for recycling. This plot only considers transfer flows for recycling that would end up in industrial processing/use operations, and their mean value is greater or equal to $1.00 \times 10^3 kg/yr$. (B) transfer of n-hexane recycled flow between industry sectors. This plot only considers the recycled flows used in industrial processing/use operations and whose mean value is greater or equal to $1.00 \times 10^3 kg/yr$.

Figure 5.5A depicts that hazardous waste treatment & disposal (IS-1), petroleum refineries (IS-6), and paint and coating manufacturing (IS-9) are the RETDF industry sectors that received a significant part of the EoL n-hexane transfers from recycling. As the dark blue chord in Figure 5.5A indicates, the hazardous waste treatment & disposal sector receives most of the n-hexane transferred for recycling operations, followed by petroleum refineries. The hazardous waste treatment & disposal sector mainly receives transfers from four generator industry sectors (IS-4, IS-7, IS-8, and IS-11). The all other basic organic chemical manufacturing industry sector (IS-11) is the largest transfer generator. This fact supports that in Figure 5.3B, the MRK most probable path under recycling activities involves these two sectors as generator and RETDF.

Figure 5.5B presents the hazardous waste treatment & disposal industry sector is related to the inter-industry sector shipments of recycled n-hexane flows whose arithmetic average is greater or equal to $1.00 \times 10^3 kg/yr$. It is consistent with Figure 5.5A that presents this RETDF industry sector as the most relevant EoL n-hexane flow receptor. This RETDF sector recycles n-hexane to ship it to the post-recycling industry sectors involved in petroleum refining and petrochemical manufacturing. As presented in Figure 5.5A, the former post-recycling sector is involved in recycling operations. The petroleum refineries sector receives a large amount of flow from the petroleum bulk stations and terminals (see Figure 5.5A), but it is not involved in the transfer of recycled n-hexane (see Figure 5.5B). The above may signify there are scenarios under which petroleum refineries receive EoL n-hexane flow and recycle it for on-site use.

Nevertheless, as indicated in Figure 5.5B, other scenarios exist under which petroleum refineries can play as the post-recycling industry sector for the n-hexane recycled by the hazardous waste treatment & disposal sector. This last scenario may indicate that petroleum refineries sign tolling or contractual agreements to receive n-hexane and recycle it from hazardous secondary materials (U.S. Environmental Protection Agency, 2015b). This inter-industry transfer behavior can explain why the petroleum refineries sector is the post-recycling industry sector for the MRF most probable path in Figure 5.3B. Additionally, Figure 5.5B depicts the hazardous waste treatment & disposal sector ships a considerable amount of recycled n-hexane to all other basic organic chemical manufacturing industry sector. Figure 5.5A indicates that all the EoL n-hexane flow received by the former sector comes from the last one. Hence, it could point out that an agreement may exist between facilities in these two sectors. The recycling operation is performed by a facility in the hazardous waste treatment & disposal sector but under any facility control in the other basic organic chemical manufacturing sector (U.S. Environmental Protection Agency, 2015b).

5.5.3 *Risk due to the potential unintended recycling of toxic content*

The inadequate design of a recycling loop can lead to recirculate hazardous chemicals that are not expected to appear in a product or service. Up to this point, n-hexane was analyzed as if it was an intentionally recycled compound. Nevertheless, as presented in the above analyses, the inter-industry amount of n-hexane transferred annually can be below $1.00 \times 10^3 kg/yr$. These small flow transfers may be associated with manufacturing or processing impurities generated by the generator industry sector. For example, only one time, the ground or treated mineral and earth manufacturing sector transfers $7.67 \times 10^{-1} kg/yr$ of EoL n-hexane flow to the hazardous waste treatment & disposal sector (see Table C.1). In addition, due to tracking only n-hexane and not all the EoL material flow, large amounts of n-hexane may be at a low concentration in industrial EoL flows.

5.5.3.1 *A screening approach for identifying unintended recycled chemicals*

This work moves forward to proposing a methodology based on publicly-available data to identify unintended recycling of toxic content. The screening seeks to identify whether the chemical being tracked (n-hexane) is intentionally recycled or is otherwise an impurity of the recycled product. Hence, Figure 5.6 is a "heatmap" whose components represent a different visualization dimension:

- The horizontal axis contains the commercial product categories for n-hexane (e.g., adhesives and sealants).
- The vertical axis contains toxic chemicals that may be potentially recycled with n-hexane (e.g., dichloromethane).
- The rightmost vertical bar presents the mean value of the mass ratio between the other toxic chemicals and n-hexane on a logarithmic scale. This logarithmic scale is used to display the wide variety of ratios in a compact way. Negative values (blue) on the bar mean the amount of n-hexane is greater than the amount of the other chemical indicated on the vertical axis. Positive values (red) mean n-hexane is present in a smaller amount than the other chemical. Finally, a value equal to 0 indicates that both n-hexane and the other chemical are in equal amount.´

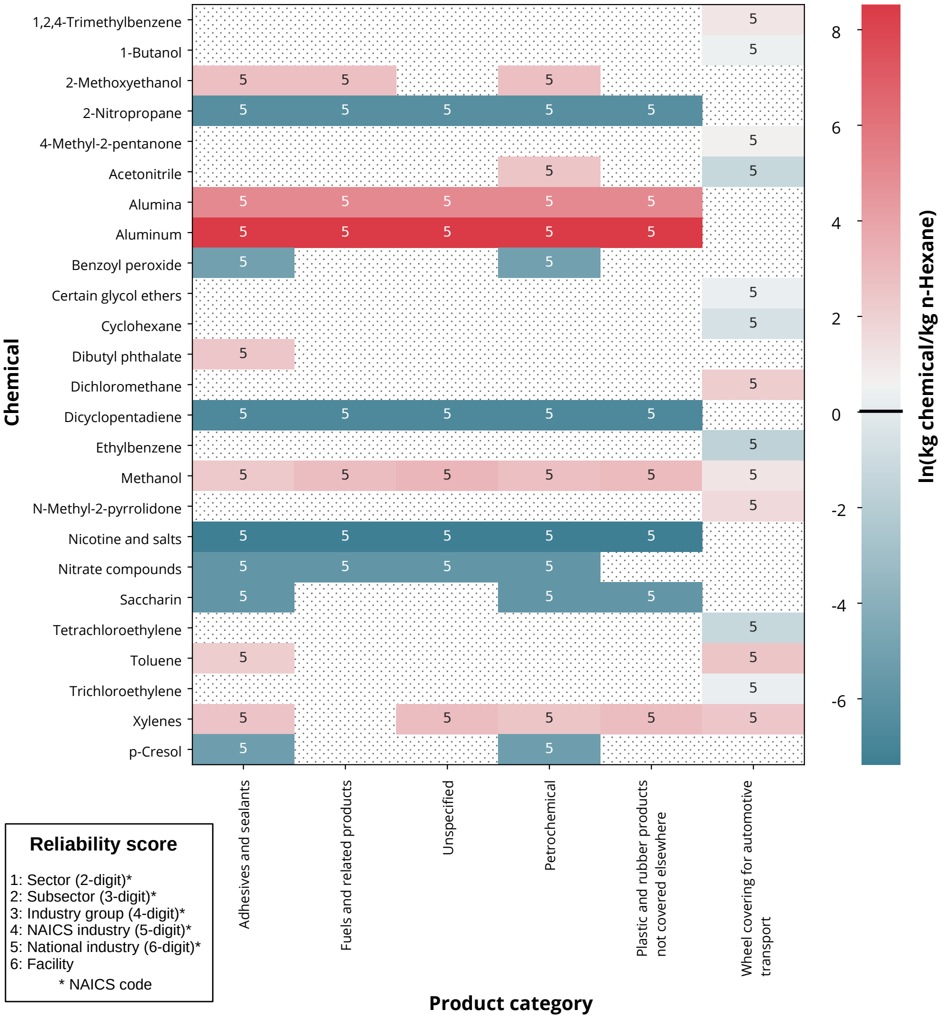


Figure 5.6: Analysis of potential chemicals associated with unintended recycling once n-hexane is recycled and goes back to commercial or consumer products. The analysis is based on the ratio between the flow recycled of another chemical and n-hexane flow.

- ¸Inside each cell of the "heatmap" there is a score that says how reliable is the ratio value according to the data used for estimating it (e.g., using the same facility information). The reliability score is similar to the technological correlation score used in the Pedigree Matrix for life cycle assessment (Edelen & Ingwersen, 2016). Nonetheless, the reliability score uses the taxonomy of the NAICS hierarchical structure to indicate the presence of toxic chemicals that may be potentially recycled together with the chemical of interest (e.g., n-hexane) (U.S. Census Bureau, 2017a). The NAICS taxonomy ranges from 2-digit

> (Sector) to 6-digit codes (National Industry), where the 6-digit NAICS code is the most specific classification. However, as shown in Figure 5.6, the score indicates whether the toxic chemicals are found based on the same reporting facility information. Thus, the reliability score ranges from 1 for less reliable (2-digit NAICS code) to 6 for more reliable (same facility).

The empty spaces (dotted pattern) in Figure 5.6 mean that non relationship between the toxic chemical and the product categories was found using the data. According to Table 5.1, 87.65% of the recycled n-hexane may go back to industrial processing/use operations and the remaining to commercial/consumer activities. Also, Figure 5.4 indicates the n-hexane recycled by facilities in the hazardous waste treatment & disposal sector has a 20.00% chance of going into commercial uses for producing valuable goods in the generic CDR product categories (described in Figure 5.6). All the findings presented in Figure 5.6 have a reliability score of 5, which means they are found based on a 6-digit NAICS code or National Industry. Although this figure does not show the full granularity of the results, i.e., PAU technology and probability that recycled n-hexane flows back to each product category, it depicts the overall result of each product category. For example, Figure 5.6 indicates that p-cresol may accompany n-hexane in adhesive and sealants and petrochemical products. However, the p-cresol/n-hexane mass ratios are below 1 (blue), i.e., the average amount of n-hexane is greater than that of p-cresol. In contrast, Figure 5.6 shows that aluminum may flow along n-hexane in all the product categories (excluding wheel covering for automotive transport); nonetheless, the dark red color warns the aluminum/n-hexane mass ratio would be above 1.

If only the mass ratio is considered as an indicator of unintended recycling, based on Figure 5.6, someone could decide that n-hexane is an impurity contaminating products containing aluminum. Therefore, one could also conclude that n-hexane is not recycled to be incorporated into any category presented in Figure 5.6, e.g., adhesive and sealant products category, especially considering that aluminum would not be in the above category (U.S. Environmental Protection Agency, 2016b). Unlike the adhesive and sealant category, aluminum could be part of fuels and related products (Pawel et al., 2012), with n-hexane being one constituent. Hence, concluding unintended recycling requires further effort to thoroughly incorporate additional information about the product category formulations and chemical uses (Dionisio et al., 2018; Isaacs et al., 2020). Moreover, incorporating the PAU technology information as a MRF variable instead of only performing the CFA and allocation can support and enhance such a critical task. Likewise, the information describing activities and uses (e.g., as a manufacturing impurity) of the chemical under study by the industry sector generating the EoL chemical transfers can provide the framework some additional insights to compare the composition and functionality between recycled and brand-new products.

5.5.3.2 *Using publicly-available data with product formulations*

As mentioned above, incorporating information about products formulation may help to improve the results soundness. Information from Chemicals and Products Database (CPDat) can be used for comparison with results in Figure 5.6 (see Table C.2). Using CPDat data, summary information (e.g., mean) for the ratio value on logarithmic scale is estimated for chemicals in products whose categorization/description is related to the adhesive and sealants product category in Figure 5.6 and n-hexane is included in their formulation (see Table C.2). For example, comparing the chemicals in Table C.2 and the ones in Figure 5.6 for adhesive and sealants category, there are chemicals like 2-methoxyethanol (CAS 109-86-4) that are not reported in Table C.2. Due to the mean ratio value on logarithmic scale, 2-methoxyethanol/n-hexane is much greater than zero and 2-methoxyethanol is not reported in any product formulation with the adhesive and sealants category, someone would conclude that the RETDF did not have intention to recycle n-hexane. Instead, the RETDF intentionally recycles 2-methoxyethanol, and n-hexane may have contaminated the recycled products. Moreover, chemicals that are in the adhesives and sealants product category may not have been tracked by the framework. For instance, acetone (CAS 67-64-1) can appear in the product category formulation (see Table C.2), but it is not in Figure 5.6 (nor for other product categories). The reason is because that chemical is not part of the TRI Program. Additionally, both toluene and n-hexane are part of the formulation of this product category. According to Figure 5.6, the ratio for toluene/n-hexane is much greater than 1 (red); however, the range value for the ratio in Table C.2 is [-2.01, 0.09] for seven samples. Then at this point, it is not feasible to obtain results with statistical significance. Hence, in summary, future developments must overcome the database reporting limitations and confidentiality, include more data sources and/or building most robust data models, and define their domain of applicability.

5.6 CONCLUSIONS

The developed data engineering framework tracks EoL chemical transfers, performs a screening to identify potential EoL exposure scenarios, and allocates chemical flows at PAU technologies. Also, the framework supplies quantitative and qualitative information from on-site PAU technologies for allocating chemical flows inside a RETDF facility, which is essential to understand the chemical EoL management chain performance, releases, and potential exposure pathways. Therefore, this contribution describes and demonstrates developments and enhancements to the data engineering framework by incorporating the EoL flow generator, the RETDF, and post-recycling sectors for identifying and assessing the current chemical EoL management chain and their recycling loops. The n-hexane case study demonstrates the framework capability of understanding current inter-industry relationships and determining feasible flow transfers between EoL and manufacturing industry sectors under existing regulatory constraints and contractual agreements. Additionally, the MRF provides the link and affinity between the chemical EoL management chain and their recycling loops, vital for their design, comprehension, and optimization. The enhanced framework can incorporate potential industrial function, use, industrial processing/use operations, post-recycling industry sector, and commercial and consumer uses. Also, the approach advances the data engineering framework in determining potential unintended recycling of toxic contents in a circular life cycle. Nevertheless, future work should analyze the integration of the chemical weight composition of products, which is relevant to understanding the near-field exposure associated with the circular life cycle and comparing the composition and functionality between recycled and brand-new products. Moreover, future work must overcome the reporting criteria restrictions of the publicly-available database systems for adequately connecting the data sources. In addition, future research should carefully analyze whether post-recycling scenarios are economically viable and if not, try to incorporate features such as OPEX, CAPEX, chemical unit price, and potential product category costs. Finally, some upcoming work should extend the framework applicability by integrating additional data and developing data-driven models for multi-scale and multi-level analyses of the chemical EoL management chain and their recycling loops.

DISCLAIMER

The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the EPA. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the EPA. The EPA does not endorse any commercial products, services, or enterprises.

ACKNOWLEDGMENTS

This research was supported in by appointment for Jose D. Hernandez-Betancur to the Research Participation Program at the Center for Environmental Solutions and Emergency Response, Office of Research and Development, EPA, administered by the Oak Ridge Institute for Science and Education through an Interagency Agreement between the U.S. Department of Energy and the EPA.

BIBLIOGRAPHY

Abkowitz, M., Eiger, A., & Srinivasan, S. (1985). Estimating the Release Rates and Costs of Transporting Hazardous Waste, 22–30.

Alaranta, J., & Turunen, T. (2020). How to Reach a Safe Circular Economy? - Perspectives on Reconciling the Waste, Product and Chemicals Regulation. *Journal Of Environmental Law*, 1–24. https://doi.org/10.1093/jel/eqaa016

Bech, N. M., Birkved, M., Charnley, F., Kjaer, L. L., Pigosso, D. C., Hauschild, M. Z., McAloone, T. C., & Moreno, M. (2019). Evaluating the environmental performance of a product/service-system business model for Merino Wool Next-to-Skin Garments: The case of Armadillo Merino®. *Sustainability (Switzerland)*, *11*(20), 1–21. https://doi.org/10.3390/su11205854

Bodar, C., Spijker, J., Lijzen, J., Waaijers-van der Loop, S., Luit, R., Heugens, E., Janssen, M., Wassenaar, P., & Traas, T. (2018). Risk management of hazardous substances in a circular economy. *Journal of Environmental Management*, *212*, 108–114. https://doi.org/10.1016/j.jenvman.2018.02.014

Bornhöft, N. A., Nowack, B., & Hilty, L. M. (2013). Material Flow Modelling for Environmental Exposure Assessment – A Critical Review of Four Approaches Using the Comparative Implementation of an Idealized Example. *Proceedings of the 27th EnviroInfo 2013 Conference*, (September), 379–388. https://doi.org/10.5167/uzh-84883

Bressanelli, G., Perona, M., & Saccani, N. (2019). Challenges in supply chain redesign for the Circular Economy: a literature review and a multiple case study. *International Journal of Production Research*, *57*(23), 7395–7422. https://doi.org/10.1080/00207543.2018.1542176

Canadian Environmental Protection Agency. (2009). Screening Assessment for the Challenge Hexane. Retrieved October 23, 2020, from https://www.ec.gc.ca/ese-ees/default.asp?lang=En%5C&n=BCBE839D-1

Chertow, M., & Park, J. (2015). Material Flow Analysis and Waste Management. In *Taking stock of industrial ecology* (pp. 87–116). https://doi.org/10.1007/978-3-319-20571-7_5

contributors, O. (2011a). Nominatim. Retrieved October 22, 2020, from https://nominatim.org/

contributors, O. (2011b). OSRM. Retrieved October 22, 2020, from http://project-osrm.org/

Corona, B., Shen, L., Reike, D., Rosales Carreón, J., & Worrell, E. (2019). Towards sustainable development through the circular economy—A review and critical assessment on current circularity metrics. *Resources, Conservation and Recycling*, *151*(May 2019), 104498. https://doi.org/10.1016/j.resconrec.2019.104498

De Groene Zaak. (2015). Governments going circular. Retrieved October 20, 2020, from http://www.govsgocircular.com/

de Römph, T. J., & van Calster, G. (2018). REACH in a circular economy: The obstacles for plastics recyclers and regulators. *Review of European, Comparative and International Environmental Law*, *27*(3), 267–277. https://doi.org/10.1111/reel.12265

Dionisio, K. L., Phillips, K., Price, P. S., Grulke, C. M., Williams, A., Biryol, D., Hong, T., & Isaacs, K. K. (2018). Data Descriptor: The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Scientific Data*, *5*, 1–9. https://doi.org/10.1038/sdata.2018.125

Edelen, A., & Ingwersen, W. (2016). *Guidance on Data Quality Assessment for Life Cycle Inventory Data* (tech. rep. June). https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=321834%5C#0Ahttps://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=528687

European Chemicals Agency (ECHA). (2017). *An evaluation of the possible health risks of recycled rubber granules used as infill in synthetic turf sports fields* (tech. rep. No. 1.01). https://echa.europa.eu/documents/10162/13563/annex-xv_report_rubber_granules_en.pdf/dbcb4ee6-1c65-af35-7a18-f6ac1ac29fe4

European Environment Agency. (2016). *More from less - material resource efficiency in Europe: 2015 overview of policies, instruments, and targets in 32 countries* (tech. rep.). https://doi.org/https://doi.org/10.2800/652446

Gottschalk, F., Scholz, R. W., & Nowack, B. (2010). Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles. *Environmental Modelling and Software*, *25*(3), 320–332. https://doi.org/10.1016/j.envsoft.2009.08.011

GREENDELTA. (2013). openLCA Nexus. Retrieved October 21, 2020, from https://nexus.openlca.org/databases

Grundmann, V., Bilitewski, B., Zehm, A., Darbra, R. M., & Barceló, D. (2013). Risk-based management of chemicals and products in a circular economy at a global scale- Impacts of the FP7 funded project RISKCYCLE. *Environmental Sciences Europe*, *25*(1), 2–7. https://doi.org/10.1186/2190-4715-25-14

Henriksson, G., Åkesson, L., & Ewert, S. (2010). Uncertainty regarding waste handling in everyday life. *Sustainability*, *2*(9), 2799–2813. https://doi.org/10.3390/su2092799

Hernandez-Betancur, J. D., Martin, M., & Ruiz-Mercado, G. J. (2021). A data engineering framework for on-site end-of-life industrial operations. *Journal of Cleaner Production*, *327*, 129514. https://doi.org/https://doi.org/10.1016/j.jclepro.2021.129514

Hernandez-Betancur, J. D., & Ruiz-Mercado, G. J. (2019). Sustainability indicators for end-of-life chemical releases and potential exposure. *Current Opinion in Chemical Engineering*, *26*(November), 157–163. https://doi.org/10.1016/j.coche.2019.09.004

Hernandez-Betancur, J. D., Ruiz-Mercado, G. J., Abraham, J. P., Martin, M., Ingwersen, W. W., & Smith, R. L. (2020). Data engineering for tracking chemicals and releases at industrial end-of-life activities. *Journal of Hazardous Materials*, 124270. https://doi.org/10.1016/j.jhazmat.2020.124270

Hernández-Betancur, J. D., Hernández, H. F., & Ocampo-Carmona, L. M. (2019). A holistic framework for assessing hot-dip galvanizing process sustainability. *Journal of Cleaner Production*, *206*, 755–766. https://doi.org/10.1016/j.jclepro.2018.09.177

Honrath, R. E. (2020). Mass and Energy Balances. Retrieved October 22, 2020, from https://pages.mtu.edu/$%5Csim$reh/courses/ce251/251_notes_dir/node3.html

Hopkinson, P., Zils, M., Hawkins, P., & Roper, S. (2018). Managing a Complex Global Circular Economy Business Model: Opportunities and Challenges. *California Management Review*, *60*(3), 71–94. https://doi.org/10.1177/0008125618764692

Isaacs, K. K., Dionisio, K., Phillips, K., Bevington, C., Egeghy, P., & Price, P. S. (2020). Establishing a system of consumer product use categories to support rapid modeling of human exposure. *Journal of Exposure Science & Environmental Epidemiology*, *30*(1), 171–183. https://doi.org/10.1038/s41370-019-0187-5

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Lahl, U., & Zeschmar-Lahl, B. (2013). Risk based management of chemicals and products in a circular economy at a global scale (risk cycle), extended producer responsibility and EU legislation. *Environmental Sciences Europe*, *25*(1), 1–12. https://doi.org/10.1186/2190-4715-25-3

Leslie, H. A., Leonards, P. E., Brandsma, S. H., de Boer, J., & Jonkers, N. (2016). Propelling plastics into the circular economy - weeding out the toxics first. *Environment International*, *94*, 230–234. https://doi.org/10.1016/j.envint.2016.05.012

Li, Y., Li, C., Li, X., Wang, K., Rahaman, M. M., Sun, C., Chen, H., Wu, X., Zhang, H., & Wang, Q. (2021). *A Comprehensive Review of Markov Random Field and Conditional Random Field Approaches in Pathology Image Analysis*. Springer Netherlands. https://doi.org/10.1007/s11831-021-09591-w

Lucchetti, M. G., Paolotti, L., Rocchi, L., & Boggia, A. (2019). The Role of Environmental Evaluation within Circular Economy: An Application of Life Cycle Assessment (LCA) Method in the Detergents Sector. *Environmental and Climate Technologies*, *23*(2), 238–257. https://doi.org/10.2478/rtuect-2019-0066

MacKay, D. J. (2002). *Information Theory, Inference, and Learning Algorithms* (1st, Vol. 13). Cambridge University Press.

Martín, M., & Grossmann, I. E. (2012). Simultaneous Optimization and Heat Integration for Biodiesel Production from Cooking Oil and Algae. *Industrial & Engineering Chemistry Research*, *51*(23), 7998–8014. https://doi.org/10.1021/ie2024596

Meys, R., Frick, F., Westhues, S., Sternberg, A., Klankermayer, J., & Bardow, A. (2020). Towards a circular economy for plastic packaging wastes – the environmental potential of chemical recycling. *Resources, Conservation and Recycling*, *162*(June), 105010. https://doi.org/10.1016/j.resconrec.2020.105010

National Institute of Standards and Technology. (1987). n-Hexane. Retrieved October 27, 2020, from https://webbook.nist.gov/cgi/cbook.cgi?ID=C110543&Mask=1EFF

Organization for Economic Co-operation and Development. (2004). *The 2004 OECD List of High Production Volume Chemicals* (tech. rep.). http://www.oecd.org/chemicalsafety/risk-assessment/33883530.pdf

Organization for Economic Co-operation and Development. (2010). Maritime Transport Costs. Retrieved October 22, 2020, from https://stats.oecd.org/Index.aspx?DataSetCode=MTC%5C#

Pawel, S. J., Thomson, J. K., Wilson, D. F., Kass, M. D., Meyer III, H. M., & Haynes, J. A. (2012). *Engine Materials Compatibility with Alternate Fuels* (tech. rep. May). Oak Ridge National Laboratory. http://

www1.eere.energy.gov/vehiclesandfuels/pdfs/merit_review_2012/propulsion_materials/pm039_pawel_2012_0.pdf

Perthen-Palmisano, B., & Jakl, T. (2004). Chemical Leasing - Cooperative business models for sustainable chemicals management - Summary of research projects commissioned by the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management (5 pp). *Environmental Science and Pollution Research 2005 12:1*, *12*(1), 49–53. https://doi.org/10.1065/ESPR2004.12.227

Pivnenko, K., Eriksen, M. K., Martín-Fernández, J. A., Eriksson, E., & Astrup, T. F. (2016). Recycling of plastic waste: Presence of phthalates in plastics from households and industry. *Waste Management*, *54*, 44–52. https://doi.org/10.1016/j.wasman.2016.05.014

Pronk, M. E., Woutersen, M., & Herremans, J. M. (2020). Synthetic turf pitches with rubber granulate infill: are there health risks for people playing sports on such pitches? *Journal of Exposure Science and Environmental Epidemiology*, *30*(3), 567–584. https://doi.org/10.1038/s41370-018-0106-1

Rigamonti, L., Falbo, A., Zampori, L., & Sala, S. (2017). Supporting a transition towards sustainable circular economy: sensitivity analysis for the interpretation of LCA for the recovery of electric and electronic waste. *International Journal of Life Cycle Assessment*, *22*(8), 1278–1287. https://doi.org/10.1007/s11367-016-1231-5

Sabaghi, M., Mascle, C., Baptiste, P., & Rostamzadeh, R. (2016). Sustainability assessment using fuzzy-inference technique (SAFT): A methodology toward green products. *Expert Systems with Applications*, *56*, 69–79. https://doi.org/10.1016/j.eswa.2016.02.038

Shylaja, B. S. (2015). From Navigation to Star Hopping : Forgotten Formulae. *Resonance*, (April), 352–359.

Smith, R. L., Tan, E. C. D., & Ruiz-Mercado, G. J. (2019). Applying Environmental Release Inventories and Indicators to the Evaluation of Chemical Manufacturing Processes in Early Stage Development. *ACS Sustainable Chemistry & Engineering*. https://doi.org/https://doi.org/10.1021/acssuschemeng.9b01961

Tanzer, J., & Rechberger, H. (2019). Setting the common ground: A generic framework for material flow analysis of complex systems. *Recycling*, *4*(2). https://doi.org/10.3390/recycling4020023

Together for Sustainability. (2013). Together for sustainability Initiative. Retrieved October 20, 2020, from https://tfs-initiative.com/

United Nations Industrial Development Organization. (2013). Chemical leasing: a performance-based business model. Retrieved October 20, 2020, from https://chemicalleasing.org/publications

United Nations Industrial Development Organization. (2019). Circular economy. Retrieved October 19, 2020, from https://www.unido.org/our-focus-cross-cutting-services/circular-economy

U.S. Census Bureau. (2017a). NAICS Structure. Retrieved October 23, 2020, from https://www.census.gov/programs-surveys/economic-census/guidance/understanding-naics.html

U.S. Census Bureau. (2017b). *North American Classification System (NAICS)*. https://doi.org/10.1159/000443915

U.S. Census Bureau. (2020). 2017 SUSB Annual Data Tables by Establishment Industry. Retrieved October 23, 2020, from https://www.census.gov/data/tables/2017/econ/susb/2017-susb-annual.html

U.S. Department of Health and Human Services. (1999). Toxicological Profile for n-Hexane. Retrieved October 23, 2020, from https://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=393%5C&tid=68%5C#top

U.S. Department of Transportation. (2018). Port Performance Freight Statistics Program. Retrieved October 22, 2020, from https://www.bts.gov/ports

U.S. Environmental Protection Agency. (1991). *Chemical Engineering Branch Manual for the Preparation of Engineering Assessments* (tech. rep.). https://nepis.epa.gov/Exe/ZyNET.exe/P10000VS.TXT?ZyActionD=ZyDocument%5C&Client=EPA%5C&Index=1991+Thru+1994%5C&Docs=%5C&Query=%5C&Time=%5C&EndTime=%5C&SearchMethod=1%5C&TocRestrict=n%5C&Toc=%5C&TocEntry=%5C&QField=%5C&QFieldYear=%5C&QFieldMonth=%5C&QFieldDay=%5C&IntQFieldOp=0%5C&ExtQFieldOp=0%5C&XmlQuery=

U.S. Environmental Protection Agency. (2004). Facility Registry Service (FRS). Retrieved October 22, 2020, from https://www.epa.gov/frs

U.S. Environmental Protection Agency. (2009). *Inventory Update Reporting (IUR) Technical Support Document Replacement of 5-digit NAICS Codes with Industrial Sector (IS) Codes* (tech. rep. October). https://www.epa.gov/sites/production/files/documents/replacingnaicswithis.pdf

U.S. Environmental Protection Agency. (2012). Sustainable Materials Management. Retrieved October 19, 2020, from https://www.epa.gov/smm

U.S. Environmental Protection Agency. (2013). TRI Basic Plus Data Files Guides. Retrieved October 22, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-plus-data-files-guides

U.S. Environmental Protection Agency. (2014). TRI Around the World. Retrieved November 13, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-around-world

U.S. Environmental Protection Agency. (2015a). Chemical Data Reporting under the Toxic Substances Control Act. Retrieved October 22, 2020, from https://www.epa.gov/chemical-data-reporting

U.S. Environmental Protection Agency. (2015b). Identification and listing of hazardous waste. Retrieved November 4, 2020, from https://www.ecfr.gov/cgi-bin/text-idx?SID=cafa8bb6d96bac714147972899e592e4%5C&mc=true%5C&node=pt40.26.261%5C&rgn=div5%5C#se40.26.261_14

U.S. Environmental Protection Agency. (2015c). Rapid Chemical Exposure and Dose Research. Retrieved October 21, 2020, from https://www.epa.gov/chemical-research/rapid-chemical-exposure-and-dose-research%5C#1

U.S. Environmental Protection Agency. (2015d). Using Predictive Methods to Assess Exposure and Fate under TSCA. Retrieved October 21, 2020, from https://www.epa.gov/tsca-screening-tools/using-predictive-methods-assess-exposure-and-fate-under-tsca

U.S. Environmental Protection Agency. (2016a). Instructions for the 2016 TSCA Chemical Data Reporting. Retrieved May 11, 2020, from https://www.epa.gov/chemical-data-reporting/instructions-reporting-2016-tsca-chemical-data-reporting

U.S. Environmental Protection Agency. (2016b). Product and Use Categories (PUCs) - Aluminum. Retrieved November 5, 2020, from https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID3040273%5C#product-use-categories

U.S. Environmental Protection Agency. (2016c). Recycling Economic Information (REI) Report. Retrieved October 19, 2020, from https://www.epa.gov/smm/recycling-economic-information-rei-report

U.S. Environmental Protection Agency. (2016d). Substance Registry Services (SRS). Retrieved October 22, 2020, from https://ofmpub.epa.gov/sor_internet/registry/substreg/LandingPage.do

U.S. Environmental Protection Agency. (2019). *Toxic Chemical Release Inventory reporting forms and instructions - revised 2019 version* (tech. rep.). https://ofmpub.epa.gov/apex/guideme_ext/guideme/file/ry_2019_rfi.pdf

U.S. Environmental Protection Agency. (2020). Waste Transfer Reports. Retrieved October 27, 2020, from https://enviro.epa.gov/triexplorer/tri_transfer.facility

Vélez-Henao, J.-A., García-mazo, C.-m., Freire-gonz, J., & Font Vivanco, D. (2020). Environmental rebound effect of energy efficiency improvements in Colombian households. *Energy Policy*, *145*(August). https://doi.org/10.1016/j.enpol.2020.111697

Wassenaar, P., Janssen, N., de Poorter, L., & Bodar, C. (2017). *Substances of very high concern and the transition to a circular economy : An initial inventory* (tech. rep.). https://rivm.openrepository.com/bitstream/handle/10029/620905/2017-0071.pdf?sequence=1%5C&isAllowed=y

Waste Management Alternatives Development and Exchange Forum. (2012). Risk-based management of chemicals and products in a circular economy at a global scale - RISKCYCLE. Retrieved October 20, 2020, from http://www.wadef.com/projects/riskcycle/

Wedema, B. P., Bauer, C., Hischier, R., Mutel, C., Nemecek, T., Reinhard, J., Vadenbo, C. O., Wenet, G., Weidema, B. P., Bauer, C., Hischier, R., Mutel, C., Nemecek, T., Reinhard, J., Vadenbo, C. O., & Wernet, G. (2013). *Data quality guideline for the ecoinvent database version 3. Ecoinvent Report 1 (v3)* (tech. rep. No. 1). http://www.ecoinvent.org/database/methodology-of-ecoinvent-3/methodology-of-ecoinvent-3.html

Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer-Verlag Berlin Heidelberg.

Part II

TOWARDS A CROSS-COUNTRY FRAMEWORK

Part III

A STEP TO HIGH-THROUGHPUT SCREENING OF END-OF-LIFE EXPOSURE SCENARIOS

6

CONTRIBUTIONS AND FUTURE DIRECTIONS

6.1 CONTRIBUTIONS

The key contribution of this work is the development of a data-driven methodology that leverages publicly-accessible environmental regulatory database systems (e.g., the U.S. Toxics Release Inventory or TRI) in order to streamline the chemical risk evaluation process by rapid identification of potential end-of-life exposure scenarios. The following specific contributions can be made in this context:

1. A data engineering framework that connects different siloed database systems and saves that transformed data into a machine-readable structure for automation, either for exploratory data analysis or building and testing data-driven models to predict potential end-of-life activities and exposure scenarios.

2. A harmonization framework for industry sectors, chemicals, and end-of-life activities that facilitates crosswalking between different international Pollution Release and Transfer Register systems. Moreover, it can easily be connected to additional sources like the OECD statistics in order to gain economic and environmental context for decision-making.

3. The data-driven framework advances by interconnecting database systems to better understand end-of-life inter-industry chemical flow transfers and the inderect chemical environmental releases that may occur when a chemical is transferred to off-site locations for end-of-life management. Moreover, the framework contributes to the state of the art by developing data-driven solutions for screening risk due to the potential unintended recycling of toxic content.

4. The framework proposes a Markov random field, a statistical graphical model, that can help to understand the end-of-life management chain and recycling loop. This statistical model is less data-demanding than an undirected discriminative graphic model, and it is suitable considering that the exact sequence for the relationship between the elements in the chemical end-of-life management chain and the recycling loop comes from siloed database systems.

5. This thesis explores the development of data preparation pipelines and the building of QSAR-inspired machine learning models called Quantitative Structure-Transfer Relationship or QSTR to predict potential end-of-life activities. Those models can be included as part of the Markov random field to provide the model factors for chemicals that are not reported in the primary data sources used by the framework.

6.2 DIRECTIONS FOR FUTURE WORK

Multiple directions for future research can be developed to build a multi-scale and multi-level model for rapid chemical flow tracking and end-of-life exposure scenario identification. Some future lines of work are drawn below:

1. The development of data-driven models to predict pollution abatement technologies for chemicals that are not reported in the primary data sources. Those data-driven models can be integrated with multi-criteria decision-making tools like the Fuzzy Analytical Hierarcy Process or FAHP and criteria like flammability, as this thesis presents.

2. Chemical risk evaluation reports usually assess and consider high-end, central tendency, and low-end parameters that represent potential scenarios that may occur during chemical exposure. Thus, three separate data-driven models to predict those three different scenarios for environmental chemical releases should be developed: worst-case, middle-case, and best-case scenario models.

3. A multi-label classification modelling strategy should be explored to develop data-driven models or QSTRs to predict end-of-life transfer scenarios for chemicals. However, the multi-label approach should be developed, considering techniques like label powerset or transformation chain. Those models should be developed under a data-centric paradigm, keeping in mind that the result of a modelling experiment strongly depends on the dataset and the model.

4. The models developed for predicting the pollution abatement units, estimating the environmental chemical releases, and screening potential end-of-life chemical flow transfer scenarios should be integrated and incorporated into the Markov random field model, in order to perform a rapid chemical flow analysis and understand the end-of-

life management chain for chemicals that are not in the primary data sources presented in the thesis.

Part IV

APPENDIX

A

APPENDIX A: SUPPORTING INFORMATION OF CHAPTER 3

A.1 DATA ENGINEERING FRAMEWORK

Table A.1 and Figure A.1 describe the collected data entry (i.e., EoL dataset features or columns) names, sources, data type, and the entities in the EoL supply and management chain. Please note that data symbol entries (first column, A.1) are used to show their entity sources in Figure A.1.

The data engineering process to transform the data into structures for the framework is depicted as sequential steps in Figure A.2 and Figure A.3. Figure A.2 presents the data engineering process used to retrieve, clean, filter, and transform information from TRI considering waste brokering, while Figure A.3 describes a similar approach for RCRAInfo. The following step-by-step instructions describe the proposed approach to build the EoL dataset and correspond to those processing steps shown in Figure A.2 and Figure A.3, respectively. Notice that Figure 3.4 is a simplified version of the data engineering and refinement process after combining Figure A.2 and Figure A.3 for transforming the TRI database into structures for the EoL dataset and tracking chemical flows at EoL stages.

DATA ENGINEERING AND REFINEMENT PROCESS OF TRI, FIGURE A.2

1. A TRI reporting year (D_1) is selected. For this work, the TRI reporting year 2017 is used, which is the most recent data year available online at the time of the study. The documents that described the content of the data reported to the TRI Program can be found in https://rb.gy/tca04t (last access May 15, 2020). Note that the database that generates these files is updated 4-6 times per year. Thus, analysts might not be able to reproduce the case study with the same data source.

2. Generator TRIF ID, location, and primary NAICS code (D_2-D_9), TRI chemical I.D. (D_{12}), TRI chemical name (D_{15}), TRI classification (D_{26}), and generator condition of use (D_{28}) are retrieved from TRI File 1a and TRI File 1b (*Release and Other Waste Management and Chemical Activities and Uses, respectively*). In contrast, the quantity transferred

Table A.1: Data entries collected from several sources (as shown in Figure A.1) for reconciliation, harmonization, and built the EoL dataset. The 3rd and 6th columns describe the data type for each entry in the EoL dataset: I for integer, F for float type, and C for character data (alphanumeric).

Data symbol	**Data name**	**Data type**	**Data symbol**	**Data name**	**Data type**
D_1	TRI reporting year[1]	I	D_2	Generator TRIF ID	C
D_3	Generator name	C	D_4	Generator street	C
D_5	Generator city	C	D_6	Generator county	C
D_7	Generator state	C	D_8	Generator zip	I
D_9	Generator primary NAICS code	I	D_{10}	Generator primary NAICS name	C
D_{11}	SRS chemical ID	I	D_{12}	TRI chemical I.D.	C
D_{13}	RCRA chemical I.D.	C	D_{14}	CAS number	C
D_{15}	TRI chemical name	C	D_{16}	SMILES	C
D_{17}	TSCA NCP chemical category 1[3]	C	D_{18}	TSCA NCP chemical category 2[3]	C
D_{19}	TSCA NCP chemical category 3[3]	C	D_{20}	Is a HAP under the CAA?[4]	C
D_{21}	Is it identified in biosolids by the CWA?[5]	C	D_{22}	Is a priority pollutant under the CWA?[6]	C
D_{23}	Is it part of CCL under the SDWA?[7]	C	D_{24}	Is it part of NPDWR under the SDWA?[8]	C
D_{25}	Is a chemical in the TSCA non-confidential inventory?[9]	C	D_{26}	TRI classification[10]	C
D_{27}	Metal indicator	C	D_{28}	Generator CoU based on TRI and matching TSCA CDR categories	C
D_{29}	Quantity transferred by the generator	F	D_{30}	Unit of measurement	C
D_{31}	Reliability and temporal correlation of the quantity transferred	F	D_{32}	EoL activity category based on TRI	C
D_{33}	EoL activity category under TSCA reports	C	D_{34}	EoL activity category under the WMH	C
D_{35}	Receiver FRS ID[2]	I	D_{36}	Receiver TRIF ID[2]	C
D_{37}	Receiver RCRA ID	C	D_{38}	Receiver name	C
D_{39}	Receiver street	C	D_{40}	Receiver city	C
D_{41}	Receiver county	C	D_{42}	Receiver state	C
D_{43}	Receiver zip	I	D_{44}	RETDF FRS ID[2]	I
D_{45}	RETDF TRIF ID[2]	C	D_{46}	RETDF name	C
D_{47}	RETDF street	C	D_{48}	RETDF city	C
D_{49}	RETDF county	C	D_{50}	RETDF state	C
D_{51}	RETDF zip	I	D_{52}	RETDF primary NAICS code	I
D_{53}	RETDF primary NAICS name	C	D_{54}	Maximum amount of chemical present at RETDF	I
D_{55}	Total chemical generated as waste by RETDF	F	D_{56}	High-end scenario chemical flow	F
D_{57}	Environmental compartments, EC-1-EC-4	C	D_{58}	RETDF chemical flow releases to the compartment D_{57}[11]	F
D_{59}	RETDF total chemical release[11]	F			

---

[1] D_1 is the year when the generator reported the off-site transfer in the TRI program. It is the starting point for data engineering (see Figure A.1).

[2] For traceability, if $D^{35} = D^{44}$ and $D^{36} = D^{45}$, then the RETDF was the receiver; therefore, there is not a brokerage. On the other hand, when a transfer was to a waste broker $D^{35} \neq D^{44}$ and $D^{36} \neq D^{45}$.

[3] For the starting point of hazard assessment, D_{17}, D_{18}, and D_{19} are the categories in which a chemical may belong under TSCA NCP. However, when a chemical does not belong to any TSCA NCP category, then this is classified according to the TRI program, that means, $D^{17} = D^{26}$.

[4] Link: https://rb.gy/a97rbp (access June 2, 2020).

[5] Link: https://rb.gy/zmox5e, Appendix A, Table A1 (access June 2, 2020).

[6] Link: https://rb.gy/d4zapo (access June 2, 2020)

[7] Link: https://rb.gy/xkoxxo (access June 2, 2020).

[8] Link: https://rb.gy/jc6u59 (access June 2, 2020).

[9] Link: https://rb.gy/ej09ah (access June 2, 2020).

[10] TRI (general EPCRA Section 313 chemical), Persistent Bioaccumulative and Toxic, and Dioxin or Dioxin-like compound.

[11] D_{58} is the total emission to each environmental compartment that was reported by the RETDF to the TRI (see Section 3.3.3) instead of being the indirect release of a chemical that may occur after it is transferred to an off-site facility for EoL management. Moreover, D_{59} is the sum of all releases to each compartment from the RETDF.

off-site (D_{29}), EoL activity category based on TRI (D_{32}), receiver name and location (D_{38}-D_{43}), and in some cases receiver RCRA ID (D_{37}, when it has the correct structure of the I.D. assigned by EPA to the physical location of a hazardous RETDF), are gathered from TRI File 3a (*Off-site Transfers*).

3. If D_{28} belonging to TSCA CDR industrial CoU, the receiver is located in the U.S., and D_{32} is an EoL activity of interest, the records are selected to form a partial dataset called TRI dataset 1.

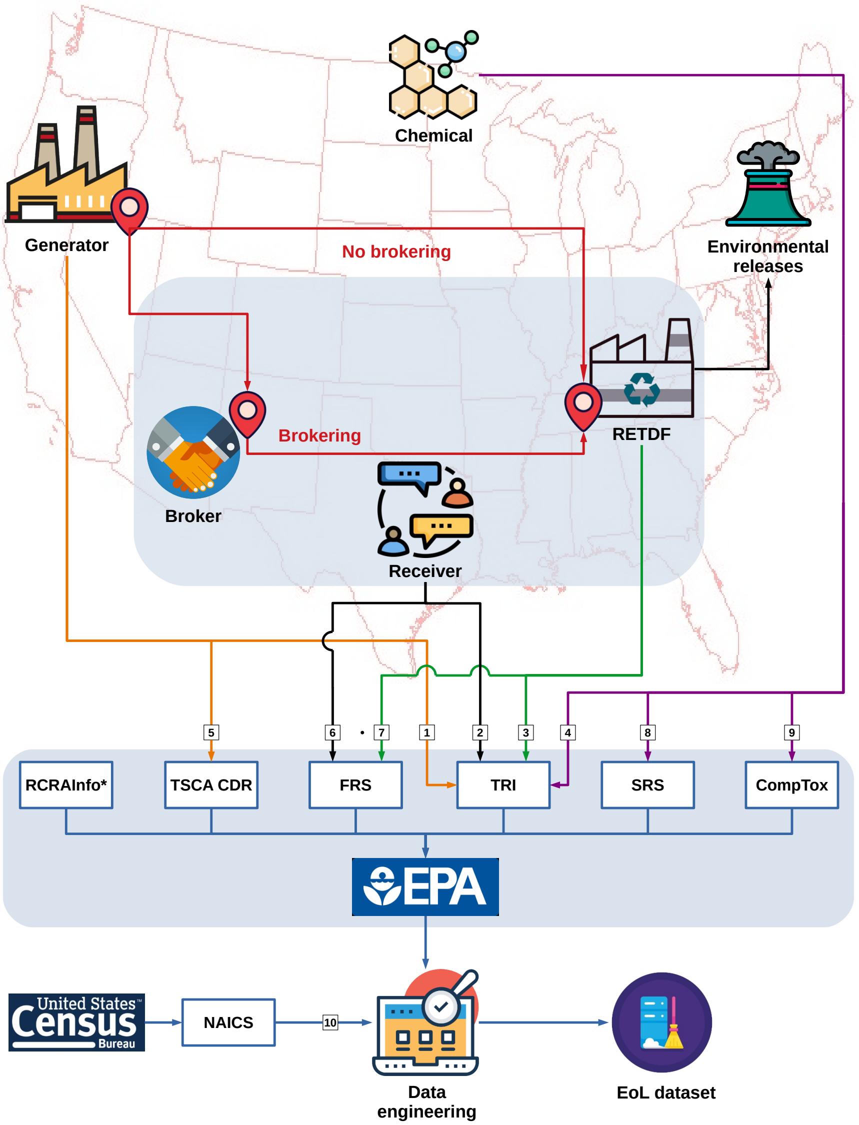
4. As shown by the red diamond 1, D_{37} or D_{38}-D_{43} are used to access to FRS. When D_{37} has the correct reporting structure, this is used to retrieve receiver FRS ID (D_{35}) and subsequently receiver TRIF ID (D_{36}). Otherwise, D_{38}-D_{43} are compared to find whether or not these are similar to any facility location recorded in the FRS, and after gathering D_{35}, D_{36}, and D_{37}.

5. To be able to continue the tracking, records are selected when the receivers have reported at least to TRI or RCRAInfo (D_{36} and D_{37} are not null) to form TRI dataset 2.

6. As shown by the red diamond 2, D_{12} is used to connect SRS and then obtain SRS chemical I.D. (D_{11}) and CAS number (D_{14}). After that, as indicated by the red diamond 3, D_{11} is used to confirm using D_{20}-D_{25} whether or not a TRI chemical belongs to one of the lists of substances of concern named in Section 3.3.1 (e.g., RCRA listed hazardous waste). Thus, TRI dataset three is obtained.

7. According to the red diamond 4, D_{12} is used to retrieve the SMILES (D_{16}) of each TRI chemical from CompTox. After that, as presented by the red diamond 5, SMILES of each chemical is the input so that the AIM tool, predicts whether a chemical may belong to any TSCA NCP categories (D_{17}-D_{19}). Thus, TRI dataset four is built.

8. As shown by the red diamond 6, D_{12}, D_{13}, D_{32}, D_{36}, and D_{37} are used to merge TRI database 4 with RCRA database 3 (resulting as explained below in the Data engineering and refinement process of RCRAInfo). Thus, obtaining RETDF location and primary NAICS code (D_{46}-D_{52}), the maximum amount of chemical at RETDF (D_{54}), total chemical generated as waste at by RETDF (D_{55}), and chemical flow released to a compartment from RETDF and sum of releases to all compartments (D_{57}-D_{59}), which are data entry group 3 in Figure A.1. Moreover, the next sub-steps shall be followed:



Data entry groups

TRI

1. $D_1 - D_9, D_{28} - D_{32}$

2. $D_{37} - D_{43}$

3. $D_{46} - D_{52}, D_{54}, D_{55}, D_{57} - D_{59}$

4. $D_{12}, D_{15}, D_{26}, D_{27}$

TSCA CDR

5. D_{28}

FRS

6. $D_{35} - D_{37}$

7. D_{44}, D_{45}

SRS

8. $D_{11}, D_{13}, D_{14}, D_{20} - D_{25}$

CompTox

9. D_{16}

NAICS

10- D_{10}, D_{53}

RCRAInfo

* For consistency analysis and brokering identification

Figure A.1: Schematic explanation about the relationship between the entities in the EoL supply and management chain and data entries and their source after refinement, harmonization, and building of the EoL dataset.

9.
    9.1. Brokering case: If D_{13} and D_{37} are non-null; these are used to check whether the TRI chemical may be considered a listed hazardous waste, and the receiver has reported to RCRAInfo. Then, D_{13}, D_{37} and D_{32} are used to access to RCRA dataset 3. If several data records satisfy the search for D_{13}, D_{32}, and D_{37}, then the one closest to the year 2017 is selected, as far as it satisfies the consistency analysis explained in Section 3.3.3.

    9.2. No brokering case: If the previous search is not successfully completed and the record has a non-null D_{36} (i.e., the receiver has reported to TRI), it goes through a recursive search from the TRI reporting year 2001 to 2017. The same applies when a receiver has not reported to RCRAInfo (null D_{32}), the TRI chemical is not listed hazardous waste (null D_{13}), or both. For this recursive search, D_{12}, D_{32}, and D_{36} (for this case D_{44}=D_{36}) are used. If various records are found having the same D_{12}, D_{32}, and D_{36} (or D_{44}), then the record closest to the year 2017 is selected.

10. Finally, the maximum possible flow of chemical transferred (D_{56}) is computed as described in Section 3.3.3. As was explained before, D_{56} is the maximum amount of chemical transferred that could end up as emissions at an off-site facility during its EoL. As explained in Section 3.3.3, it is given by the minimum flow found by tracking a chemical using the procedures described in Figure A.2 and Figure A.3.

DATA ENGINEERING AND REFINEMENT PROCESS OF RCRAINFO, FIGURE A.3

1. A reporting year of the RCRA report is selected as a starting point. However, the procedure in Figure A.3 is done for each RCRA biennial report from 2001 to 2017 (odd years). The idea is to try to find an off-site facility (broker(s), receiver, and RETDF) in the information stored to date, such that it is possible to track a chemical flow.

2. Using the amount of hazardous waste received by a facility and the reported management method code (at the red diamond 1 stage), the Nonbrokers dataset 1 and Brokers dataset 1 are built. Therefore, If the record has an amount received other than 0, this goes to Nonbrokers dataset 1, unless it has a management method code equal to H141, which means that a facility receives the waste only for storage and then transfer, therefore, it would go to Brokers dataset 1.

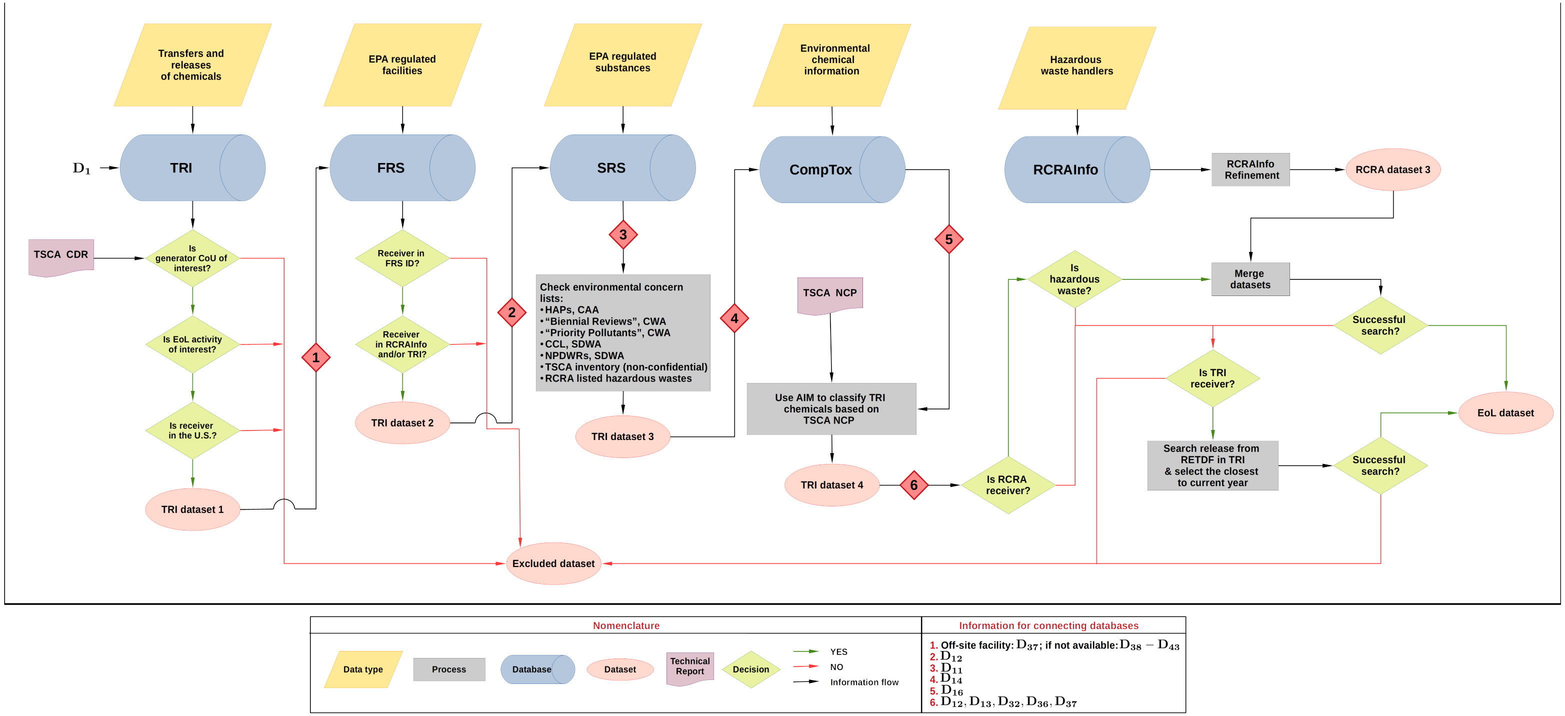


Figure A.2: Data engineering and refinement process for transforming the TRI database into structures for the EoL dataset and tracking chemical flows at EoL stages. RCRAInfo, data engineering process, is presented in Figure A.3.

3. With the information collected at the red diamond 2 stage, the Brokers dataset 1 passes through the grey block composed of 4 stages, obtaining a cleaner data set called Brokers dataset 2 by:
    3.1. Searching the quantity of waste containing the chemical of concern shipped by a broker if source code is G61, that means, the waste was received from off-site for storage and then transfer. Here, information about "to where" and "for what" the waste was shipped is identified as well.
    3.2. Knowing the management activity for which a hazardous waste was shipped (e.g., biological treatment) and where it was transferred, the off-site facility which received the hazardous waste is searched in the information stored in RCRAInfo to date.
    3.3. If the searching is successful, a flow consistency analysis is done to ensure the flow shipped by a broker is less than the flow received by the off-site facility.
    3.4. If several records are found having the same off-site facility and management activity, it is selected the record closest to the reporting year of the RCRA report, which was selected in step 1.
    3.5. Step 3 is done until the RETDF is found.
4. Broker dataset 2 and Nonbrokers dataset 1 are combined.
5. After that, RCRA IDs of facilities, at the red diamond 3, are used to find their FRS IDs. Subsequently, as shown the red diamond 4, these are used to check if a facility has reported to the TRI program at any time. Thus, if a facility satisfies the above, a record is kept in RCRA dataset 1.
6. As shown at the red diamond 5, RCRA IDs of hazardous wastes are used to retrieve the SRS IDs. They are then used to verify if hazardous waste is a TRI chemical, which means if it has a TRIF ID. Thus, RCRA dataset 2 is generated at this step.
7. Finally, the TRIF ID of RETDFs and hazardous wastes are used to connect with TRI reports (the red diamond 7). Then, information about data entry group 9 in Figure A.1 is searched in the existing and available TRI reports to date. If the search is successful, the record is kept; otherwise, it is discarded. Subsequently, RCRA dataset 3 is obtained and can be used by step 8.1 of the Data engineering and refinement process of TRI.

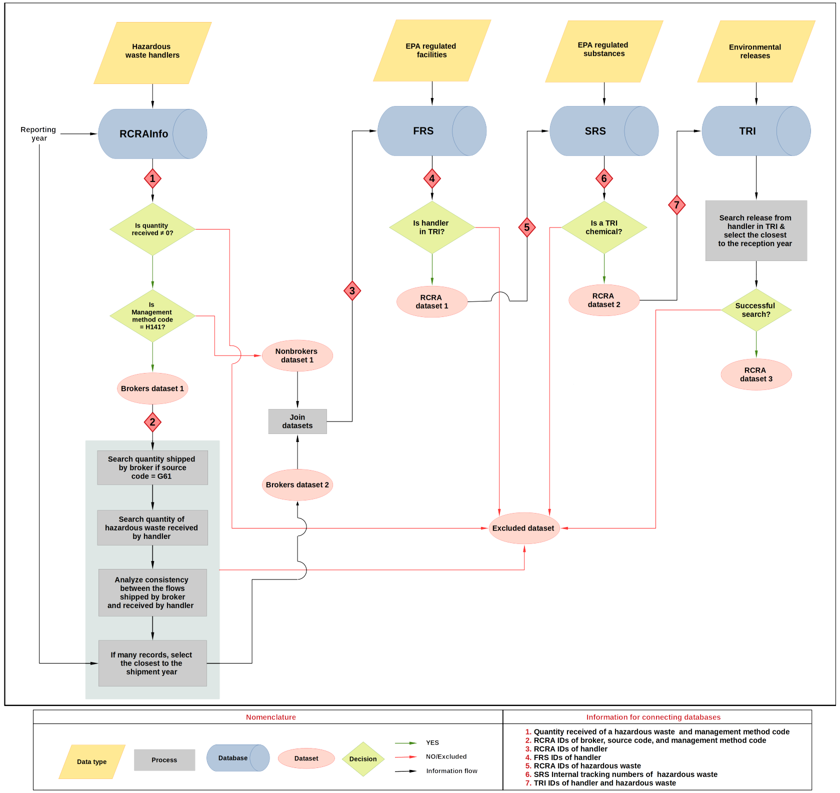


Figure A.3: Data engineering and refinement process for transforming the RCRAInfo database into structures for the EoL dataset and tracking chemicals in industrial wastes. Management method code H141 “describes the type of hazardous waste management system used to treat, recover, or dispose of hazardous waste.” The source code G61 “describes the type of process or activity (i.e., source) where hazardous waste was generated.”

Note that both processes shown in Figure A.2 and Figure A.3 can be done in parallel, at least until the process presented in Figure A.2, requires RCRA dataset 3. It is essential to clarify that the recursive search shown in Figure A.2 for all existing and available TRI reports to date, is independent of the data engineering process presented in Figure A.3 for RCRAInfo, making it possible to address both procedures in parallel.

A.2 DEVELOPMENT OF EQUATION 3.1

The framework assumes that a RETDF receives a chemical flow for (i) energy recovery, recycling, treatment (physical/chemical/thermal/biological), or disposal, and (ii) the recycled chemical is used as a product or incorporated in a valuable product. Additionally, as shown in Figure A.4(A), the RETDF may also generate wastes containing a chemical. The total waste having the chemical can be recycled, treated, transferred, released, disposed of, used for recovering energy. Thus, Equation A.1 presents the annual mass balance for the chemical. $\Delta AC^{(i)}$ is the annual change of the amount of chemical present at the RETDF, which can be calculated as $\Delta AC^{(i)} = AC^{(i)}_{end} - AC^{(i)}_{start}$ using the amount of chemical present at the start of the year ($AC^{(i)}_{start}$) and at the end of the year ($AC^{(i)}_{end}$).

$$\Delta AC^{(i)} = Input + Generation - D^{(i)}_{55} \leftrightarrow Input + Generation = \Delta AC^{(i)} + D^{(i)}_{55} \tag{A.1}$$

As depicted in Figure A.4(B), the amount of chemical present or accumulated at the RETDF can change over the reporting year. Thus, if $D^{(i)}_{54}$ is the range code representing the maximum amount of chemical present at RETDF at any moment over the reporting year and $QM^{(i)}$ is a value that belongs to D_{54}, $AC^{(i)}_{start}$, and $AC^{(i)}_{end}$ would be $AC^{(i)}_{start} \leq QM^{(i)}$ and $AC^{(i)}_{end} \leq QM^{(i)}$ and even they can be 0. Thus, $\Delta AC^{(i)}$ must be estimated to calculate the emission factor for each record i. Also, two boundary-value cases are:

1. Figure A.5(A) shows a first case where $AC^{(i)}_{start} = QM^{(i)}$ and $AC^{(i)}_{end} = 0$ with $QM^{(i)} \in D^{(i)}_{54}$. If the highest value for $QM^{(i)}$ is taken from $D^{(i)}_{54}$, $\Delta AC^{(i)} = -MAX(QM^{(i)})$. This case represents the lowest possible value that $\Delta AC^{(i)}$ can take.
2. Figure A.5(B) shows a second case where $AC^{(i)}_{start} = 0$ and $AC^{(i)}_{end} = QM^{(i)}$. Similarly, taking the highest value for $QM^{(i)}$, $\Delta AC^{(i)} = MAX(QM^{(i)})$. This value is the highest possible value for$\Delta AC^{(i)}$.

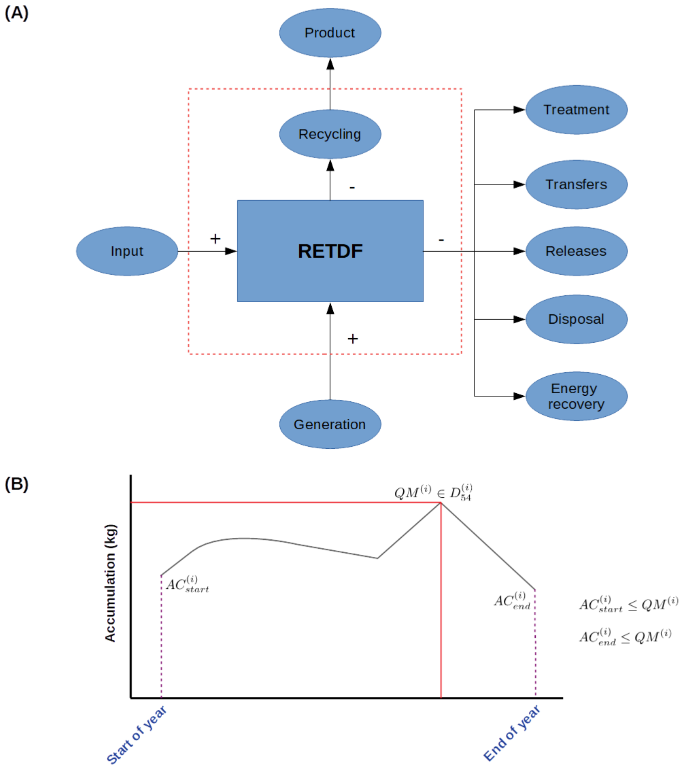


Figure A.4: (A) Summary of the elements included in the annual mass balance for a chemical at a RETDF based on the assumptions presented in Section 3.3.3 (B) The hypothetical curve is representing the amount of a chemical present at a RETDF over the course of a reporting year as a time function.

Due to the above boundary-value cases, $\Delta AC^{(i)}$ has a constraint such $\Delta AC^{(i)} \in [-MAX(QM^{(i)}), MAX(QM^{(i)})]$. However, this is the first of three constraints for $\Delta AC^{(i)}$, as shown in Figure A.5(C). The other two constraints for $\Delta AC^{(i)}$ are obtained as:

1. The term $Input + Generation$ cannot be negative, and the denominator of the emission factor must not be 0 to avoid indetermination. Thus, using the Equation A.1, a second constraint for $\Delta AC^{(i)}$ is obtained, $\Delta AC^{(i)} > -D_{55}^{(i)}$, i.e., $\Delta AC^{(i)} \in (-D_{55}^{(i)}, +\infty)$, which is depicted in Figure A.5(C).

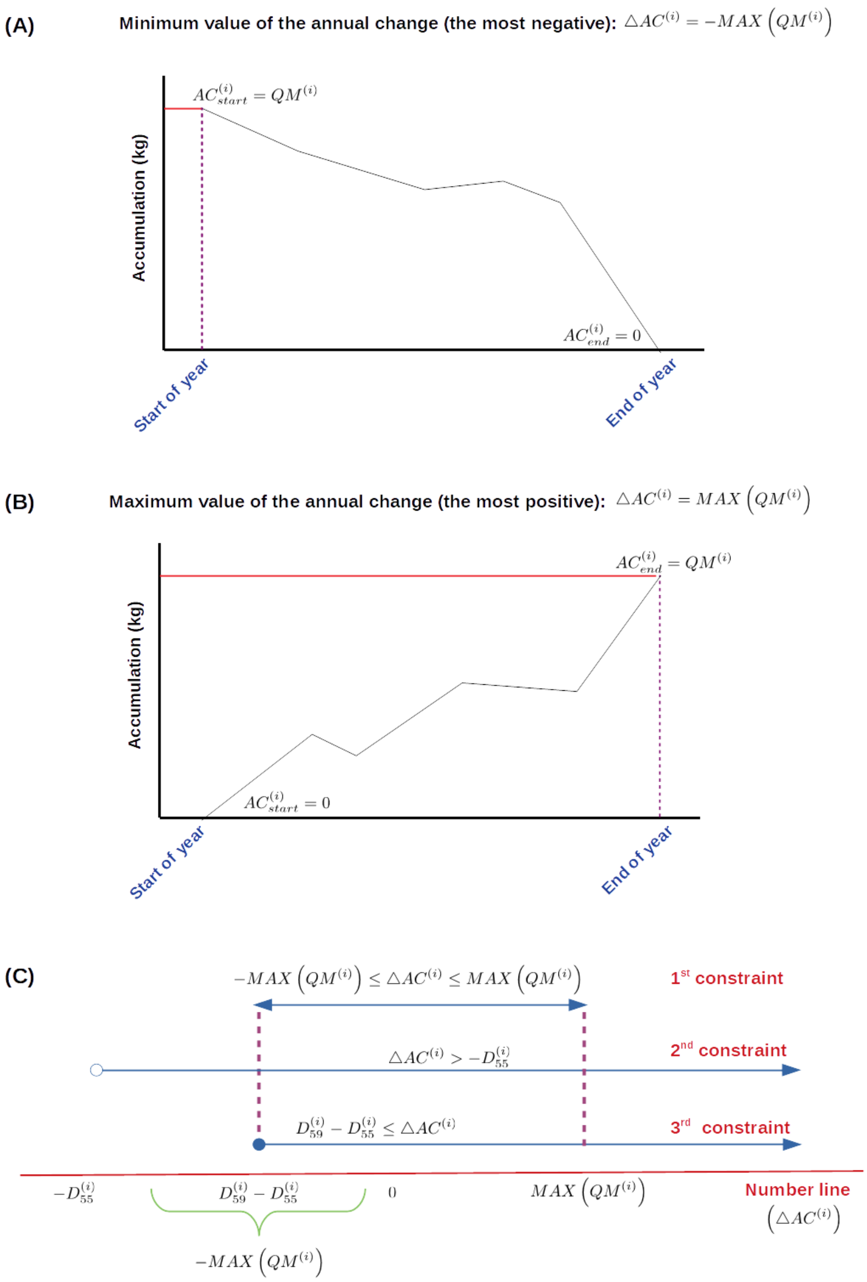


Figure A.5: (A) Hypothetical case with the lowest value for $\Delta AC^{(i)}$ (more negative).(B) Hypothetical case with the highest value for $\Delta AC^{(i)}$ (more positive).(C) Intersection of the intervals obtained by applying the constraints.

2. Additionally, the sum of the emission factors must be a value between 0 to 1. As Figure 3.5 shows, the framework considers four

Table A.2: Estimating indirect release quantities using Equation 3.1. A record from the EoL dataset is used as an example.

Data from the EoL dataset						
$D^{(i)}_{29}$	$D^{(i)}_{54}$	$MAX(QM^{(i)})$	$D^{(i)}_{55}$	$D^{(i)}_{57}$	$D^{(i)}_{58}$	$D^{(i)}_{59}$
784.261kg	05 (45,359.237 to 453,591.916kg)	453,591.916kg	348,373.733kg	fugitive air release	170.319kg	176.234kg
				on-site soil release	0.000kg	
				on-site surface water release	0.023kg	
				stack air release	5.892kg	

Internal calculations to obtain quantities released to each environmental compartment

- $D^{(i)}_{59} - D^{(i)}_{55} = -348,197.499kg$
- $MAX(D^{(i)}_{59} - D^{(i)}_{55}, -MAX(QM^{(i)})) = MAX(-348,197.499kg, -453,591.916kg) = -348,197.499kg$
- Thus, $\Delta AC^{(i)} \in [-348,197.499kg, 453,591.916kg]$
- As an example, uniform distribution and a sample size of 3 is used to obtain three values of $\Delta AC^{(i)}$ and subsequently three values of $EF^{(i)}_k$ for each environmental compartment using Equation 3.1:
  1. Fugitive air release, EC-1: 6.44×10^{-5}, 0.0001, and $4.88 \times 10^{-5} kg/kg$
  2. On-site soil release, EC-2: all are equal to $0kg/kg$
  3. On-site surface water release, EC-3: 8.57×10^{-9}, 1.34×10^{-8}, and $6.50 \times 10^{-9} kg/kg$
  4. Stack air release, EC-4: 2.23×10^{-6}, 3.47×10^{-6}, and $1.69 \times 10^{-6} kg/kg$
- Using the values obtained above it is possible to calculate $ER^{(i)}[D^{(i)}_{57\,k}]$:
  1. Fugitive air release, EC-1: 5.05×10^{-2},7.87×10^{-2}, and $3.83 \times 10^{-2} kg$
  2. On-site soil release, EC-2: all are equal to $0kg$
  3. On-site surface water release, EC-3: 6.72×10^{-6}, 1.05×10^{-5}, and $5.10 \times 10^{-6} kg$
  4. Stack air release, EC-4: 1.75×10^{-3}, 2.72×10^{-3}, and $1.32 \times 10^{-3} kg$
- For each compartment an arithmetic average for $ER^{(i)}[D^{(i)}_{57\,k}]$, i.e., $\overline{ER^{(i)}[D^{(i)}_{57\,k}]}$ can be computed:
  1. Fugitive air release, EC-1: $5.58 \times 10^{-2} kg$
  2. On-site soil release, EC-2: $0.0kg$
  3. On-site surface water release, EC-3: $7.43 \times 10^{-6} kg$
  4. Stack air release, EC-4: $1.93 \times 10^{-3} kg$
- Additionally, these values can be added to obtain the average total amount of the chemical transferred by the generator that may be potentially released from the RETDF: $5.78 \times 10^{-2} kg$

- Additionally, these values can be added to obtain the average total amount of the chemical transferred by the generator that may be potentially released from the RETDF: $5.78 \times 10^{-2} kg$

compartments: indoor air (EC-1), on-site soil (EC-2), on-site surface water (EC-3), and outdoor air (EC-4). Due to this criterion in the sum of the emission factors, a third constraint can be obtained for $\Delta AC^{(i)}$. Thus, developing the Equation A.2 results in the constraint $-D^{(i)}_{55} \leq D^{(i)}_{59} - D^{(i)}_{55} \leq \Delta AC^{(i)}$. However, considering the second constraint, the third constraint for $\Delta AC^{(i)}$ can be expressed as $D^{(i)}_{59} - D^{(i)}_{55} \leq \Delta AC^{(i)}$, i.e. $[D^{(i)}_{59} - D^{(i)}_{55}, +\infty)$, as presented in Figure A.5(C).

Figure A.5(C) shows the intersection of the three intervals which represent the three constraints for $\Delta AC^{(i)}$. Thus, from the discontinued vertical purple line on the right, $MAX(QM^{(i)})$ is the upper value for the intersection. However, as indicated by the green curly bracket, the lower value for the intersection depends on the largest value between $D^{(i)}_{59} - D^{(i)}_{55}$ and $MAX(QM^{(i)})$, i.e., $MAX(D^{(i)}_{59} - D^{(i)}_{55}, -MAX(QM^{(i)}))$. Thus, the domain for $\Delta AC^{(i)}$ is $[MAX(D_{59}^{(i)} - D_{55}^{(i)}, -MAX(QM^{(i)})), MAX(QM^{(i)})]$. Thus, for a record i, $\Delta AC^{(i)}$ has several values in that interval so that $\Delta AC^{(i)}$ may be represented by a random number.

$$\begin{gathered}
0 \leq \sum_{k=1}^{4} EF_{k}^{(i)} \leq 1 \\
0 \leq \sum_{k=1}^{4} \frac{D_{58\,k}^{(i)}}{\Delta AC^{(i)} + D_{55}^{(i)}} \leq 1 \\
0 \leq \frac{1}{\Delta AC^{(i)} + D_{55}^{(i)}} \times \sum_{k=1}^{4} D_{58\,k}^{(i)} \leq 1 \\
0 \leq \leq 1 \\
0 \leq \frac{D_{59}^{(i)}}{\Delta AC^{(i)} + D_{55}^{(i)}} \leq 1 \leftrightarrow -D_{55}^{(i)} \leq D_{59}^{(i)} - D_{55}^{(i)} \leq \Delta AC^{(i)}
\end{gathered} \tag{A.2}$$

A.3 TABLES WITH LABELS FOR THE CASE STUDY: TRACKING OF MC FLOWS AT THE EOL STAGE AT OFF-SITE LOCATIONS (TRANSFERS)

Table A.3: NAICS Classification for the RETDF industry sector (RETDFiS) and used in the chemical flow analysis shown in Figure 3.6.

Name	Label
All other miscellaneous chemical product and preparation manufacturing	RETDFiS-1
Hazardous waste treatment and disposal	RETDFiS -2
Cement manufacturing	RETDFiS -3
Materials recovery facilities	RETDFiS -4
Solid waste combustors and incinerators	RETDFiS -5
Petroleum lubricating oil and grease manufacturing	RETDFiS -6
All other basic organic chemical manufacturing	RETDFiS -7
Photographic film, paper, plate, and chemical manufacturing	RETDFiS -8
Pesticide and other agricultural chemical manufacturing	RETDFiS -9

Table A.4: TRI activities and uses of chemicals (TRIU) for defining the CoU categories described in Table A.7.

Name	Label
Added as a formulation component	TRIU-1
Used as a chemical processing aid	TRIU-2
Repackaging	TRIU-3
Ancillary or other use	TRIU-4
Produce the chemical	TRIU-5
Used as a reactant	TRIU-6
As a process impurity	TRIU-7
Used as a manufacturing aid	TRIU-8
Import the chemical	TRIU-9
Used as an article component	TRIU-10

Table A.5: Environmental compartment (E.C.) labels for the data engineering approach, and shown in Figure 3.6.

Name	Label
[1]Fugitive air release	EC-1
[1]On-site soil release	EC-2
[1]On-site surface water release	EC-3
[1]Stack air release	EC-4
Net recycling	EC-5
Net energy recovery	EC-6
Net treatment	EC-7
Net disposal	EC-8

[1]These are the environmental compartments which are part of D_{57} in the EoL dataset.

Table A.6: NAICS classification for the generator industry sector (GiS) used in the chemical flow analysis shown in Figure 3.6.

Name	Label
Adhesive manufacturing	GiS-1
All other basic organic chemical manufacturing	GiS-2
All other miscellaneous chemical product and preparation manufacturing	GiS-3
Ammunition (except small arms) manufacturing	GiS-4
Artificial and synthetic fibers and filaments manufacturing	GiS-5
Cement manufacturing	GiS-6
Copper rolling, drawing, extruding, and alloying	GiS-7
Ethyl alcohol manufacturing	GiS-8
Guided missile and space vehicle propulsion unit and propulsion unit parts manufacturing	GiS-9
Hazardous waste treatment and disposal	GiS-10
In–vitro diagnostic substance manufacturing	GiS-11
Industrial gas manufacturing	GiS-12
Irradiation apparatus manufacturing	GiS-13
Materials recovery facilities	GiS-14
Medicinal and botanical manufacturing	GiS-15
Metal coating, engraving (except jewelry and silverware), and allied services to manufacturers	GiS-16
Metal crown, closure, and other metal stamping (except automotive)	GiS-17
National security	GiS-18
Other basic inorganic chemical manufacturing	GiS-19
Other chemical and allied products merchant wholesalers	GiS-20
Paint and coating manufacturing	GiS-21
Pesticide and other agricultural chemical manufacturing	GiS-22
Petrochemical manufacturing	GiS-23
Petroleum lubricating oil and grease manufacturing	GiS-24
Pharmaceutical preparation manufacturing	GiS-25
Photographic and photocopying equipment manufacturing	GiS-26
Plastics material and resin manufacturing	GiS-27
Precision turned product manufacturing	GiS-28
Special die and tool, die set, jig, and fixture manufacturing	GiS-29
Sporting and athletic goods manufacturing	GiS-30
Surgical and medical instrument manufacturing	GiS-31

Table A.7: Chemical condition of use (CoU) categories (at generator facility) employed in the chemical flow analysis shown in Figure 3.6. Also, Table A.4 describes the corresponding TRI activities and uses of chemicals (TRIU) combinations.

Label[1]	**Combination**
CoU-1	TRIU-1
CoU-2	TRIU-1 + TRIU-2
CoU-3	TRIU-1 + TRIU-3 + TRIU-4
CoU-4	TRIU-1 + TRIU-3 + TRIU-2
CoU-5	TRIU-1 + TRIU-3 + TRIU-2 + TRIU-4
CoU-6	TRIU-4
CoU-7	TRIU-5
CoU-8	TRIU-5 + TRIU-3
CoU-9	TRIU-5 + TRIU-6 + TRIU-1 + TRIU-7 + TRIU-2 + TRIU-8 + TRIU-4
CoU-10	TRIU-3 + TRIU-2 + TRIU-4
CoU-11	TRIU-2
CoU-12	TRIU-2 + TRIU-8
CoU-13	TRIU-2 + TRIU-8 + TRIU-4
CoU-14	TRIU-6 + TRIU-1 + TRIU-3 + TRIU-2
CoU-15	TRIU-6 + TRIU-2 + TRIU-4
CoU-16	TRIU-1 + TRIU-4
CoU-17	TRIU-5 + TRIU-3 + TRIU-7
CoU-18	TRIU-6 + TRIU-2
CoU-19	TRIU-8
CoU-20	TRIU-3 + TRIU-7 + TRIU-4
CoU-21	TRIU-3
CoU-22	TRIU-7
CoU-23	TRIU-7 + TRIU-4
CoU-24	TRIU-5 + TRIU-1
CoU-25	TRIU-5 + TRIU-9 + TRIU-3 + TRIU-7 + TRIU-4
CoU-26	TRIU-5 + TRIU-3 + TRIU-4
CoU-27	TRIU-3 + TRIU-4
CoU-28	TRIU-9 + TRIU-1 + TRIU-4
CoU-29	TRIU-5 + TRIU-2 + TRIU-4
CoU-30	TRIU-2 + TRIU-4
CoU-31	TRIU-6 + TRIU-1
CoU-32	TRIU-1 + TRIU-3
CoU-33	TRIU-5 + TRIU-6 + TRIU-1
CoU-34	TRIU-6
CoU-35	TRIU-6 + TRIU-7 + TRIU-2 + TRIU-8 + TRIU-4
CoU-36	TRIU-5 + TRIU-9 + TRIU-3 + TRIU-2 + TRIU-8 + TRIU-4
CoU-37	TRIU-1 + TRIU-7 + TRIU-2
CoU-38	TRIU-5 + TRIU-1 + TRIU-2 + TRIU-4

[1] In the TRI program, facilities do not report quantities of chemicals generated at each activity; instead of that, they submit a form with all the aggregated quantities.

Table A.8: Classification of EoL activities used in the chemical flow analysis shown in Figure 3.6. The 1st column is the categorization used by the TRI program for representing off-site transfers. Notice that EoL management by combustion is captured in both energy recovery and treatment via incineration. Also, this table shows the resulting TRI classification categories after applying the EoL data engineering framework for the MC case study. Other chemical substances might have other resulting categories from the full set of TRI classification categories https://rb.gy/d2hwez (access June 2, 2020).

TRI classification[1]	**TSCA classification[2]**	**Label**	**EPA WMH[3]**	**Label**
Solvents/organics recovery Other reuse or recovery	Recycling[4]	EoL-1	Recycling[4]	WMH-1
Energy recovery	Energy recovery[4]	EoL-2	Energy recovery[4]	WMH-2
Incineration/insignificant fuel value[5] Incineration/thermal treatment[5]	Incinerators	EoL-3	Treatment	WMH-3
Wastewater treatment (excluding POTWs) —non–metals	Industrial wastewater treatment	EoL-4		
Solidification/stabilization — treatment —non–metals Other waste treatment	Other treatment	EoL-5		
RCRA Subtitle C landfills	Landfill (municipal, hazardous, or other land disposals)	EoL-6	Disposal	WMH-4
Underground injection class I wells	Underground injection	EoL-7		

[1] These correspond to D_{32}.
[2] These correspond to D_{33}.
[3] These correspond to D_{34}.
[4] These classifications are part of D_{33} and D_{34}.
[5] For a facility "to claim that a reported EPCRA Section 313 chemical sent off-site is used for the purpose of energy recovery and not for treatment for destruction, the EPCRA Section 313 chemical must have a significant heating value and must be combusted in an energy recovery unit such as an industrial boiler, furnace, or kiln". Thus, the "incineration/insignificant fuel value" category represents that the chemical goes into a "legitimate energy recovery unit," but it does not "contribute to the heating value of the waste."

A.4 DATA OBTAINED FROM TRI EXPLORER

Based on https://rb.gy/ot4mph (access May 15, 2020), the TRI off-site transfers for further management of waste containing MC were gathered and summarized in Table A.9 using percentages. That information was obtained by the TRI administrators by means of 257 facilities in all industries, for MC in the U.S., 2017. In addition, from the EoL dataset, 259 pathways for MC were obtained after tracking using the data engineering approach, which were used to calculate the values in the 3rd column of Table A.9. Using chi-square hypothesis test with 3-degree freedom and significance level 0.05, the critical value is 7.81 and the test statistic 1.9011, that means, it is not rejected the hypothesis that the obtained material distribution using the EoL dataset records are similar to those obtained using the information available online from the TRI Explorer tool since 1.9011 < 7.81.

Table A.9: MC percentage distribution for each type of off-site waste management using the pathways found using the data engineering approach and the information from TRI administrators.

Management type	%		χ^2
	From TRI	**From EoL**	
Transfers to energy recovery	19.016	18.215	0.0337
Transfer to recycling	36.720	42.779	0.9998
Transfers to disposal	0.348	0.017	0.3148
Transfers to treatment	43.916	38.989	0.5528
Total Transfers off-site for further waste management	100.000	100.000	1.9011

A.5 PYTHON SCRIPT THAT WAS WRITTEN FOR THE MC CASE STUDY

The Python code was uploaded to the following GitHub repository:

¸ https://github.com/jodhernandezbe/MC_Case_Study

B

APPENDIX B: SUPPORTING INFORMATION OF CHAPTER 4

B.1 CASE STUDY 1: CELECOXIB MANUFACTURING PROCESS

Table B.1: Specifications of EoL streams from the Celecoxib manufacturing process.

Stream	α_{input}[1]	**Chemical**	**[%wt/wt]**
1 (isopropanol/water washes)	L	Isopropanol	49.20
		Methanol	0.70
2 (mother liquor or filtrate)	W	Isopropanol	34.50
		Methanol	8.45
3 (dryer distillates)	L	Isopropanol	50.70
		Methanol	0.47

[1] For the TRI Program, liquid wastes having more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as nonaqueous waste (L).

Table B.2: Comparison between CFA performed by the bottom-up and top-down approaches. The results in the table are for stream # 1 in Table B.1.

PAU name	**Output stream**	**Chemical**	**Mean quantity - bottom-up [kg/yr]**	**Mean quantity - top-down [kg/yr]**
Solvents/Organics Recovery - Batch Still Distillation	$F^{(c)}_{waste/release}$	methanol	2.79×10^0	4.46×10^0
	$F^{(c)}_{recycled}$	methanol	5.07×10^4 (88.48 %)[1]	4.74×10^4 (82.72%)[1]
	$F^{(c)}_{fugitive}$	methanol	3.67×10^2	3.62×10^3
	$F^{(c)}_{effluent}$	isopropanol	1.99×10^4	9.86×10^5
		methanol	5.26×10^3	5.26×10^3
		isopropanol	3.95×10^6	2.99×10^6
Liquid Injection Incinerator	$F^{(c)}_{destroyed}$	isopropanol	3.93×10^6 (99.18 %)[2]	1.97×10^6 (49.85%)[2]
	$F^{(c)}_{fugitive}$	methanol	3.79×10^2	3.39×10^2
		isopropanol	4.55×10^3	1.19×10^6
	$F^{(c)}_{effluent}$	methanol	4.90×10^3	4.90×10^3
		isopropanol	1.52×10^4	9.92×10^3

[1] The expected $\eta^{(c)}$ for methanol is 90.00%.
[2] The expected $\eta^{(c)}$ for isopropanol is 99.50%.

Table B.3: PAU selection and sequence arrangement and CFA for the EoL flows from the Celecoxib manufacturing process in Table B.1.

Stream	PAU name	Position	Output stream	α_{input}[1]	Chemical	Mean quantity [kg/yr]	CV[2]
1	Solvents/Organics Recovery - Batch Still Distillation	1	$F^{(c)}_{waste/release}$	W	methanol	2.79×10^0	2.40
			$F^{(c)}_{recycled}$	L	methanol	5.07×10^4	0.01
			$F^{(c)}_{fugitive}$	A	methanol	3.67×10^2	1.22
					isopropanol	1.99×10^4	0.01
			$F^{(c)}_{effluent}$	L	methanol	5.26×10^3	0.08
					isopropanol	3.95×10^6	0.01
	Liquid Injection Incinerator	2	$F^{(c)}_{destroyed}$	NA	isopropanol	3.93×10^6	0.01
			$F^{(c)}_{fugitive}$	A	methanol	3.79×10^2	1.36
					isopropanol	4.55×10^3	0.79
			$F^{(c)}_{effluent}$	A	methanol	4.90×10^3	0.13
					isopropanol	1.52×10^4	0.24
2	Solvents/Organics Recovery - Batch Still Distillation	1	$F^{(c)}_{waste/release}$	W	methanol	4.16×10^1	2.80
			$F^{(c)}_{recycled}$	L	methanol	6.12×10^5	0.01
			$F^{(c)}_{fugitive}$	A	methanol	3.99×10^3	1.26
					isopropanol	1.39×10^4	0.01
			$F^{(c)}_{effluent}$	W	methanol	6.39×10^4	0.07
					isopropanol	2.77×10^6	0.01
	Liquid Injection Incinerator	2	$F^{(c)}_{destroyed}$	NA	isopropanol	2.75×10^6	0.01
			$F^{(c)}_{fugitive}$	A	methanol	4.46×10^3	1.34
					isopropanol	3.20×10^3	0.82
			$F^{(c)}_{effluent}$	A	methanol	5.99×10^4	0.12
					isopropanol	1.07×10^4	0.25
3	Solvents/Organics Recovery - Batch Still Distillation	1	$F^{(c)}_{waste/release}$	W	methanol	2.84×10^0	3.61
			$F^{(c)}_{recycled}$	L	methanol	3.41×10^4	0.01
			$F^{(c)}_{fugitive}$	A	methanol	2.23×10^2	1.34
					isopropanol	2.04×10^4	0.01
			$F^{(c)}_{effluent}$	L	methanol	3.56×10^3	0.08
					isopropanol	4.07×10^6	0.01
	Liquid Injection Incinerator	2	$F^{(c)}_{destroyed}$	NA	isopropanol	$4.05 \times^1 0^6$	0.01
			$F^{(c)}_{fugitive}$	A	methanol	2.37×10^2	1.36
					isopropanol	5.86×10^3	0.91
			$F^{(c)}_{effluent}$	A	methanol	3.31×10^3	0.13
					isopropanol	1.45×10^4	0.37

[1] For the TRI Program, A represents air emissions. Additionally, liquid wastes holding more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as a nonaqueous liquid (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[2] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

B.2 CASE STUDY 2: SOLVENT EOL FLOWS FROM FOOD IS

Table B.4: Parameters in the EoL flow that differentiate how the procedure in Figure 4.5 deals with each solvent EoL amount from the food IS[1].

Chemical	**Stream**	$I_{IB}^{(c)}$	α_{input}[2]	$\eta^{(c)}$ **[%]**	$w^{(c)}$ **[%]**
methanol	4	Yes	W	89.55	0.51
	5	Yes	W	99.85	25.50
	6	Yes	L	99.98	75.00
	7	No	W	89.55	0.51
	8	No	W	99.85	25.50
	9	No	L	99.98	75.00
ammonia	10	Yes	W	99.00	1.00×10^{-4}
	11	Yes	W	99.99	0.01
	12	Yes	W	100.00	0.51
	13	Yes	W	100.00	25.50
	14	No	W	99.00	1.00×10^{-4}
	15	No	W	99.99	0.01
	16	No	W	100.00	0.51
	17	No	W	100.00	25.50
ethylene glycol	18	No	W	96.22	0.51
	19	No	W	99.94	25.50
n-hexane	20	No	W	42.00	1.00×10^{-7}
	21	No	W	99.88	1.00×10^{-4}
	22	No	W	100.00	0.01
	23	No	W	100.00	0.51
	24	No	W	100.00	25.50
	25	No	L	100.00	75.00

[1] These streams are assumed to be a binary mixture.

[2] For the TRI Program, liquid wastes having more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as a nonaqueous liquid (L).

Table B.5: PAU selection and CFA for the methanol EoL flows from the food IS in Table B.4.

Stream	**PAU name**	**Output stream**	α_{output}[1]	**Mean quantity [kg/yr]**	**CV**[2]
4	Solvents/Organics Recovery - Batch Still Distillation	$F^{(c)}_{waste/release}$	W	4.00×10^{0}	1.64
		$F^{(c)}_{recycled}$	L	3.75×10^{4}	0.01
		$F^{(c)}_{fugitive}$	A	1.73×10^{2}	1.19
		$F^{(c)}_{effluent}$	W	4.19×10^{3}	0.05
5	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	2.11×10^{6}	0.01
		$F^{(c)}_{fugitive}$	A	1.97×10^{2}	1.40
		$F^{(c)}_{effluent}$	W	3.08×10^{3}	0.09
6	Liquid Injection Incinerator	$F^{(c)}_{destroyed}$	NA	6.21×10^{6}	0.01
		$F^{(c)}_{fugitive}$	A	7.12×10^{1}	1.37
		$F^{(c)}_{effluent}$	A	1.03×10^{3}	0.10
7	Solvents/Organics Recovery - Batch Still Distillation	$F^{(c)}_{waste/release}$	W	3.07×10^{0}	2.72
		$F^{(c)}_{recycled}$	L	3.75×10^{4}	0.01
		$F^{(c)}_{fugitive}$	A	2.55×10^{2}	1.22
		$F^{(c)}_{effluent}$	W	4.12×10^{3}	0.07
8	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	2.11×10^{6}	0.01
		$F^{(c)}_{fugitive}$	A	2.00×10^{2}	1.19
		$F^{(c)}_{effluent}$	W	3.07×10^{3}	0.08
9	Liquid Injection Incinerator	$F^{(c)}_{destroyed}$	NA	6.21×10^{6}	0.01
		$F^{(c)}_{fugitive}$	A	7.31×10^{1}	1.24
		$F^{(c)}_{effluent}$	A	1.02×10^{3}	0.09

[1] For the TRI Program, A represents air emissions. Additionally, liquid wastes containing more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as nonaqueous material (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[2] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

Table B.6: PAU selection and CFA for the ethylene glycol EoL flows from the food IS in Table B.4.

Stream	**PAU name**	**Output stream**	α_{output}[1]	**Mean quantity [kg/yr]**	**CV**[2]
18	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	4.23×10^{7}	0.01
		$F^{(c)}_{fugitive}$	A	3.50×10^{4}	1.55
		$F^{(c)}_{effluent}$	W	1.63×10^{6}	0.04
19	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	2.22×10^{9}	0.01
		$F^{(c)}_{fugitive}$	A	2.57×10^{4}	1.38
		$F^{(c)}_{effluent}$	W	1.22×10^{6}	0.03

[1] For the TRI Program, A represents air emissions. Additionally, liquid wastes containing more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as nonaqueous material (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[2] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

Table B.7: PAU selection and CFA for the ammonia EoL flows from the food IS in Table B.4.

Stream	**PAU name**	**Output stream**	α_{output}[1]	**Mean quantity [kg/yr]**	**CV**[2]
10	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	1.29×10^{3}	0.01
		$F^{(c)}_{fugitive}$	A	1.23×10^{0}	1.17
		$F^{(c)}_{effluent}$	W	1.18×10^{1}	0.12
11	Solvents/Organics Recovery - Fractionation	$F^{(c)}_{waste/release}$	W	1.29×10^{-1}	2.56
		$F^{(c)}_{recycled}$	L	1.32×10^{5}	0.01
		$F^{(c)}_{fugitive}$	A	1.32×10^{0}	1.24
		$F^{(c)}_{effluent}$	W	1.14×10^{1}	0.17
12	Solvents/Organics Recovery - Fractionation	$F^{(c)}_{waste/release}$	W	4.10×10^{-2}	1.55
		$F^{(c)}_{recycled}$	L	1.31×10^{7}	0.01
		$F^{(c)}_{fugitive}$	A	1.46×10^{0}	1.14
		$F^{(c)}_{effluent}$	W	1.14×10^{1}	0.15
13	Neutralization	$F^{(c)}_{destroyed}$	NA	6.64×10^{8}	0.01
		$F^{(c)}_{fugitive}$	A	1.06×10^{0}	1.35
		$F^{(c)}_{effluent}$	W	8.64×10^{0}	0.16
14	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	1.29×10^{3}	0.01
		$F^{(c)}_{fugitive}$	A	8.44×10^{-1}	1.21
		$F^{(c)}_{effluent}$	W	1.22×10^{1}	0.08
15	Solvents/Organics Recovery - Fractionation	$F^{(c)}_{waste/release}$	W	3.91×10^{-2}	2.61
		$F^{(c)}_{recycled}$	L	1.31×10^{5}	0.01
		$F^{(c)}_{fugitive}$	A	8.89×10^{-1}	1.32
		$F^{(c)}_{effluent}$	W	1.20×10^{1}	0.10
16	Solvents/Organics Recovery - Fractionation	$F^{(c)}_{waste/release}$	W	3.45×10^{-2}	2.37
		$F^{(c)}_{recycled}$	L	1.31×10^{7}	0.01
		$F^{(c)}_{fugitive}$	A	8.37×10^{-1}	1.24
		$F^{(c)}_{effluent}$	W	1.20×10^{1}	0.09
17	Neutralization	$F^{(c)}_{destroyed}$	NA	6.64×10^{8}	0.01
		$F^{(c)}_{fugitive}$	A	6.39×10^{-1}	1.24
		$F^{(c)}_{effluent}$	W	9.05×10^{0}	0.09

[1] For the TRI Program, A represents air emissions. Additionally, liquid wastes containing more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as nonaqueous material (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[2] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

Table B.8: PAU selection and CFA for the n-hexane EoL flows from the food IS in Table B.4.

Stream	PAU name	Output stream	α_{output}[1]	Mean quantity [kg/yr]	CV[2]
20	Oil Skimming	$F^{(c)}_{removed}$	L	2.27×10^{-1}	0.01
		$F^{(c)}_{fugitive}$	A	2.57×10^{-2}	1.25
		$F^{(c)}_{effluent}$	W	2.87×10^{-1}	0.11
21	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	2.70×10^{2}	0.01
		$F^{(c)}_{fugitive}$	A	2.41×10^{-2}	1.15
		$F^{(c)}_{effluent}$	W	2.89×10^{-1}	0.10
22	Solvents/Organics Recovery - Solvent Extraction	$F^{(c)}_{waste/release}$	W	3.00×10^{-4}	2.33
		$F^{(c)}_{recycled}$	L	2.72×10^{4}	0.01
		$F^{(c)}_{fugitive}$	A	2.53×10^{-2}	1.27
		$F^{(c)}_{effluent}$	W	2.87×10^{-1}	0.11
23	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	2.72×10^{6}	0.01
		$F^{(c)}_{fugitive}$	A	2.13×10^{-2}	1.28
		$F^{(c)}_{effluent}$	W	2.90×10^{-1}	0.09
24	Aerobic Treatment	$F^{(c)}_{destroyed}$	NA	1.38×10^{8}	0.01
		$F^{(c)}_{fugitive}$	A	1.97×10^{-2}	1.46
		$F^{(c)}_{effluent}$	W	2.14×10^{-1}	0.13
25	Liquid Injection Incinerator	$F^{(c)}_{destroyed}$	NA	4.05×10^{8}	0.01
		$F^{(c)}_{fugitive}$	A	6.10×10^{-3}	1.41
		$F^{(c)}_{effluent}$	A	7.22×10^{-2}	0.12

[1] For the TRI Program, A represents air emissions. Additionally, liquid wastes containing more than 50.00 %wt/wt of water are considered wastewater (W); otherwise, they are taken as nonaqueous material (L). NA is not applicable, and it is used for destroyed or converted mass streams.

[2] The bright red color represents the flows have high variance or variability, i.e., CV > 1.

B.3 COMPARATIVE DATA FOR METHANOL AND ISOPROPANOL

Table B.9: Data for the comparison of methanol and isopropanol for the case study one.

Data name	**Methanol**	**Isopropanol**	**Ratio kg methanol/kg isopropanol**
Domestic Manufacturing Production [kg/yr][1]	4.60×10^{8}	3.20×10^{7}	14.34
Imported Volume [kg/yr][1]	3.91×10^{9}	7.56×10^{7}	51.73
Volume Used [kg/yr][1]	8.07×10^{6}	3.64×10^{6}	2.22
Volume Exported [kg/yr][1]	4.1×10^{7}	3.85×10^{6}	10.77
OECD High Production Volume[2]	Yes	Yes	NA
Market size [USD]	33.69 billion in 2020[3]	2.65 billion in 2019[4]	NA

[1] Source: https://bit.ly/2YMyyhq

[2] Source: https://bit.ly/3AOSOfL

[3] Source: https://bwnews.pr/3joXjgZ

[4] Source: https://bit.ly/3AB8935

C

APPENDIX C: SUPPORTING INFORMATION OF CHAPTER 5

C.1 UNCERTAINTY PROPAGATION

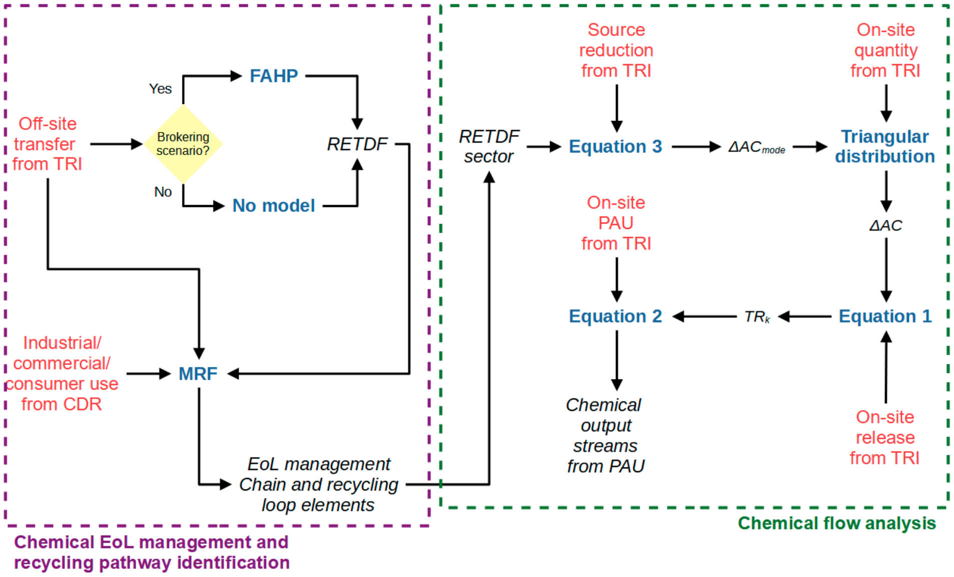


Figure C.1: Generic visualization diagram of uncertainty propagation. The **red captions** represent data entries from the data sources (e.g., TRI), the **blue captions** are for the models for handling uncertainty or equations. The **black captions** are for the outputs/results from models or equations. The **purple dotted-line area** encloses the generic scheme of uncertainty propagation for identifying chemical EoL management and recycling pathways, while the **green dotted-line area** encloses the generic uncertainty propagation for the CFA.

C.2 GITHUB REPOSITORY

The supplementary GitHub repository associated with this article can be found in the public GitHub repository named EoL4Chem at the following link:

https://github.com/jodhernandezbe/EoL4Chem.

C.3 EXAMPLE OF MRF FACTOR OR POTENTIAL FUNCTION

Table C.1: Comparison between the MRF factors using the number of times a relationship appears and the transfer flows for n-hexane. The light blue cells are the highest value for the MRF factor or potential function.

Generator industry sector	RETDF industry sector	Transfer [kg/yr]	Times
All other basic organic chemical manufacturing	Hazardous waste treatment & disposal	9.72×10^4	3
Other motor vehicle parts manufacturing	Hazardous waste treatment & disposal	3.07×10^3	1
Hazardous waste treatment & disposal	Hazardous waste treatment & disposal	1.72×10^3	5
Ethyl alcohol manufacturing	Hazardous waste treatment & disposal	1.12×10^3	1
Commercial printing (except screen and books)	Cement manufacturing	8.01×10^2	1
Adhesive manufacturing	Hazardous waste treatment & disposal	6.32×10^2	2
Other chemical and allied products merchant wholesalers	Hazardous waste treatment & disposal	1.14×10^2	3
Ground or treated mineral and earth manufacturing	Hazardous waste treatment & disposal	7.67×10^{-1}	1

C.4 DATASET FROM CPDAT

In the link CPDat.csv, you can find the processed dataset extracted from the Chemical and Products Database (CPDat). This database contains information mapping more than 49,000 chemicals to a set of terms categorizing their usage or function in 16,000 consumer products (e.g. shampoo, soap, etc.) based on their chemical content (CPDat). The processed dataset contains information of products having n-hexane and the other chemicals that are part of their formulation. After obtaining the processed dataset, the information is used to get the values presented in Table C.2.

Table C.2: Summary information obtained for the product formulations containing n-hexane and product category related to generic CDR product category named adhesives and sealants.

Name	CAS	ln(wf chemical/wf n-Hexane)				
		sample	min	max	mean	median
1,1,1-Trichloroethane	71-55-6	3	-1.61	1.53	0.44	1.40
1,3-Butadiene, homopolymer	9003-17-2	1	-0.69	-0.69	-0.69	-0.69
1,3-Butadiene-styrene copolymer	9003-55-8	3	-0.92	0.00	-0.31	0.00
Acetone	67-64-1	3	-2.01	0.00	-0.67	0.00
Bentonite	1302-78-9	1	-0.69	-0.69	-0.69	-0.69
Benzene, 1,3-diethenyl-, polymer with 1,3-butadiene and ethenylbenzene	26471-45-4	2	0.00	0.00	0.00	0.00
Butane	106-97-8	3	-1.10	-1.10	-1.10	-1.10
Calcium carbonate	471-34-1	3	-0.92	-0.18	-0.67	-0.92
Carbon dioxide	124-38-9	1	-1.50	-1.50	-1.50	-1.50
Cyclohexane	110-82-7	3	-1.90	0.00	-1.26	-1.90
Dichloromethane	75-09-2	2	-1.74	-1.70	-1.72	-1.72
Dimethyl ether	115-10-6	1	0.00	0.00	0.00	0.00
Dimethyl polysiloxane	63148-62-9	4	-2.71	-1.00	-1.99	-2.12
Distillates (petroleum), steam-cracked, polymers with light steam-cracked petroleum naphtha	68410-16-2	3	-0.92	0.69	-0.07	0.00
Heptane	142-82-5	4	0.00	0.00	0.00	0.00
Isobutane	75-28-5	14	-1.79	1.61	-0.65	-0.85
Isopropanol	67-63-0	1	-1.00	-1.00	-1.00	-1.00
Kaolin	1332-58-7	2	1.54	1.54	1.54	1.54
Methylcyclohexane	108-87-2	2	-0.98	-0.98	-0.98	-0.98
Petroleum resins	64742-16-1	1	-1.10	-1.10	-1.10	-1.10
Propane	74-98-6	17	-1.35	0.17	-0.59	-0.57
Quartz-alpha (SiO2)	14808-60-7	1	-2.39	-2.39	-2.39	-2.39
Resin acids and Rosin acids, esters with glycerol	8050-31-5	2	0.00	0.00	0.00	0.00
Solvent naphtha, petroleum, light aliph.	64742-89-8	9	-0.11	1.39	0.62	0.69
Stoddard solvent	8052-41-3	2	-2.01	-0.92	-1.47	-1.47
Stoddard solvent IIC	64742-88-7	1	-2.35	-2.35	-2.35	-2.35
Tetraethyl orthosilicate	78-10-4	3	-2.89	-2.89	-2.89	-2.89
Toluene	108-88-3	7	-2.01	0.09	-0.68	-0.49
Zinc oxide	1314-13-2	1	-2.90	-2.90	-2.90	-2.90