DOI: 10.1111/jiec.13044

METHODS, TOOLS, AND SOFTWARE

JOURNAL OF INDUSTRIAL ECOLOGY WILEY

Improving the reliability of chemical manufacturing life cycle inventory constructed using secondary data

David E. Meyer[1] | Sarah Cashman[2] | Anthony Gaglione[2]

[1] Center for Environmental Solutions and Emergency Response, U.S. Environmental Protection Agency, Cincinnati, Ohio

[2] Eastern Research Group, Inc., Lexington, Massachusetts

Correspondence
David E. Meyer, Center for Environmental Solutions and Emergency Response, U.S. Environmental Protection Agency, 26 W. Martin Luther King Dr., Cincinnati, OH 45268.
Email: Meyer.David@epa.gov

Funding information
Parts of this work were funded through USEPA Contract # EP-D-011-006 by the Chemical Safety and Sustainability National Research Program.

Editor Managing Review: Michael Zwicky Hauschild

Abstract

This study proposes methods to improve data mining workflows for modeling chemical manufacturing life cycle inventory. Secondary data sources can provide valuable information about environmental releases during chemical manufacturing. However, the often facility-level nature of the data challenges their utility for modeling specific processes and can impact the quality of the resulting inventory. First, a thorough data source analysis is performed to establish data quality scoring and create filtering rules to resolve data selection issues when source and species overlaps arise. A method is then introduced to develop context-based filter rules that leverage process metadata within data sources to improve how facility air releases are attributed to specific processes and increase the technological correlation and completeness of the inventory. Finally, a sanitization method is demonstrated to improve data quality by minimizing the exclusion of confidential business information (CBI). The viability of the methods is explored using case studies of cumene and sodium hydroxide production in the United States. The attribution of air releases using process context enables more sophisticated filtering to remove unnecessary flows from the inventory. The ability to sanitize and incorporate CBI is promising because it increases the sample size, and therefore representativeness, when constructing geographically averaged inventories. Future work will focus on expanding the application of context-based data filtering to other types and sources of environmental data.

KEYWORDS
chemical releases, data filtering, data mining, data sanitization, industrial ecology, life cycle inventory (LCI)

1 | INTRODUCTION

The Frank R. Lautenberg Chemical Safety for the 21st Century Act (2016) directs the U.S. Environmental Protection Agency (EPA) to evaluate chemical risk to human and ecological health by considering the full life cycle of chemicals and products as part of its decisions. Life cycle assessment (LCA) provides a complementary approach to traditional risk assessment given its use of multi-attribute impact assessment and comparative analysis based on the function of goods and services in society (Csiszar et al., 2016). Implementing LCA to support chemical decision-making is often limited by the resources required to build the life cycle inventory (LCI), an accounting of all material and energy flows attributed to the chemical manufacture, use, and end-of-life treatment. There are two types of LCA flows, elementary flows and intermediate flows. Elementary flows are the key LCA flows because they are the basis for impact characterization while intermediate flows are used to connect stages within the full life cycle.

Ideally, chemical LCI data should be collected from manufacturers. However, primary data are difficult to obtain for numerous reasons: (a) impractical resource requirements for data collection; (b) manufacturers treat the data as confidential business information (CBI); and (c) the full life cycle value chain can be complex and involve numerous processes and manufacturers. In the absence of primary data, there is a need to estimate the data and LCI modeling has been the focus of numerous studies.

Some of the earliest reported work on LCI modeling is that of Bretz and Frankhauser (1996) who coupled data mining with simple process design principles to estimate inventory for thousands of industrial chemicals. Jiménez-González, Kim, and Overcash (2000a, 2000b) applied process design and rule-of-thumb assumptions to model pharmaceutical processes. This work is interesting because the authors developed methods to model ancillary energy processes (Jiménez-González, 2000b) separate from the chemical of interest (Jiménez-González et al., 2000a). Numerous other examples of using process design in LCI modeling have followed (Alvarado et al., 2019; Geisler, Hofstetter, & Hungerbühler, 2004; Parvatker et al., 2019; Simon et al., 2019; Yao, 2018), with this approach eventually expanding to include full process simulation (Liao, Kelley, & Yao, 2020; Smith et al., 2017). Although effective, process design and simulation often require detailed process knowledge and chemical engineering expertise that may not be practical or readily available (Meyer et al., 2019; Parvatker, 2018).

One approach to circumvent the need for rigorous process modeling is the emerging use of statistical inferencing and machine learning. Wernet, Hellweg, Fischer, Papadokonstantakis, and Hungerbühler (2008) first applied structure-based regression and neural network analysis to predict life cycle impacts and lumped inventory. Song, Keller, and Suh (2017) applied a similar approach to a much larger dataset to improve the predictability of life cycle impacts using molecular structure models. While useful for LCA, the prediction of impacts without an underlying inventory can make it difficult to interpret results because of the black-box nature of the neural network model. The alternative is the use of statistical inferencing based on classification to estimate LCI. Pereira, Hauner, Hungerbühler, and Papadokonstantakis (2018) used classification to estimate process steam requirements while Meyer et al. (2019) applied classification to estimate chemical releases. A challenge for classification is the need for a large set of training data to develop the model.

The other less rigorous approach is data mining. Cashman et al. (2016) presented a method to model the LCI of average chemical manufacturing by developing a workflow to mine publicly available production volume (PV) and chemical release data. While their focus was on U.S. manufacturing and EPA data, consistent and rigorous workflows can be applied to any geographic domain where suitable data are available. The key benefits of a data mining workflow are the potential to automate the process and update the LCI as data sources are updated. A similar data mining workflow has been proposed by Young et al. (2019) for modeling petroleum refinery operations. An LCI is built at the process level while chemical release data are often collected at a facility level and can involve multiple on-site activities. For data mining, the goal is to separate production of the primary chemical from other chemical production and ancillary processes like combustion that can be connected to the primary chemical process as intermediate flows. Cashman et al. manually filtered elementary flows for the primary chemical using detailed knowledge of the required process chemistry (e.g., reactants, solvents, catalysts, and by-products).

The addition of chemistry-based filtering was intended to improve the quality of the LCI in terms of reliability and representativeness (Edelen, 2016). However, the original workflow was susceptible to limitations in practice. Chemistry-based filtering is time intensive to develop and requires intimate knowledge of process chemistries that varies from facility to facility. If the chemical filter includes substances for all industrial synthesis routes, it is possible that the data for a given facility and synthesis route may include an accepted substance that is associated with another on-site activity. In this case, the erroneous substance will not be properly filtered out. When multiple chemicals are produced on-site and involve the same substance, physical allocation is used to distribute the substance across all on-site chemical manufacturing based on each chemical's fraction of total site production. This solution is less desirable because it assumes the magnitude of the flow scales linearly with PV, irrespective of the process. If any of the PV data are withheld as CBI, the corresponding facility data cannot be incorporated into the model because the data mining workflow uses PV to standardize releases and the resulting LCI will not account for total chemical production within a geographic region. These limitations affect the quality of the data, challenges that mirror broader issues within the growing field of data science.

Data science focuses on the extraction of useful knowledge from so-called Big Data. Conceptually, efforts in data science seek to address the five V's: volume, velocity, variety, veracity, and value (Debattista, Lange, Scerri, & Auer, 2015). Of the five, veracity and value capture the LCI modeling limitations described above because they emphasize the need to curate the highest quality data possible to achieve maximum utility (Garcia-Gil, Luengo, Garcia, & Herrera, 2017). All large datasets have the potential for noise, including outliers, anomalies, and duplicates. Higher quality means minimizing such noise and maximizing the value. Preprocessing actions like filtering, sanitization, transformation, and normalization are applied for this purpose (Garcia, Ramírez-Gallego, Luengo, Manuel Benítez, & Herrera, 2016). Of these, data filtering and sanitization are promising options for addressing the data mining workflow issues. Data filtering involves the use of descriptors to identify, correct, and/or remove potentially erroneous data, all of which may affect the reliability, technological correlation, and completeness of the data. Data sanitization removes and protects sensitive information, such as CBI PVs. If successfully implemented, sanitized data maintain the value of the original data and make them useful during data mining. The preservation of the original population size can reduce quality issues related to population sampling.

The chemistry-based data filter applied by Cashman et al. had limited success removing potential noise because it was developed independent of the data and did not take advantage of metadata, such as process and unit descriptions, captured during facility reporting. As an alternative, the use of metadata to establish filtering rules may provide a better basis for filtering the raw LCI by establishing an activity context that improves the technological correlation and completeness of the LCI. For CBI challenges, it may be possible to use data sanitization to mask facility information

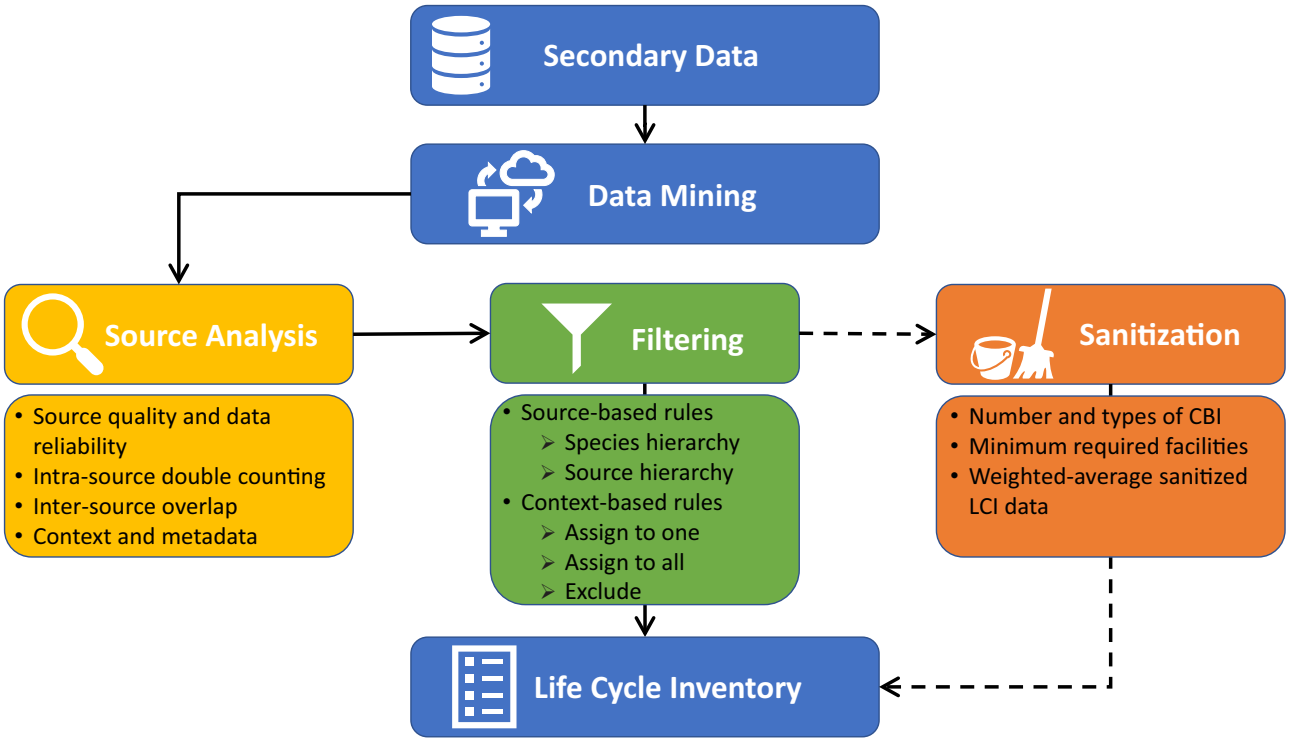


FIGURE 1 A three-step process for improving the quality of LCIs produced by data mining workflows using source analysis, filtering, and sanitization. The optional nature of sanitization is indicated by the dashed arrow

claimed as CBI in a way that allows the data from the facility to still be included in the final LCI and improve its population sampling and technological correlation. Therefore, the objectives of this work are: (a) develop new data processing methods to improve the quality of LCI produced by data mining workflows while still supporting eventual automation; and (b) apply a refined data mining workflow to a set of chemical case studies for chemical production in the United States.

2 | METHODS

2.1 | Methods to enhance the use of existing data in LCI modeling

A three-step approach for developing data mining workflows is shown in Figure 1 and described in more detail in the following subsections. Thorough data source analysis is introduced to provide a better understanding of the context for chemical release data. This knowledge is then applied for rule-based filtering to address attribution of facility releases. An optional final step of sanitization is included but is only necessary when addressing CBI.

2.1.1 | Data source analysis

The first step is to perform a detailed analysis of the targeted data sources. Part of the analysis focuses on understanding reporting requirements, basis of estimates, threshold limits, exemptions, and potential overlaps. These factors are important for properly integrating data from the various sources. For example, if a facility must only report chemical species emitted above a threshold mass for a specific data source, it is possible the resulting LCI built from that data source will be missing species emitted below the threshold. Similar concerns arise when there are chemical species that are exempted from reporting. In some data sources, double counting is possible if the data are not properly processed. The findings from this part of the analysis are used to develop data quality scoring guidelines, a data source hierarchy for inter-source species overlap, and a species hierarchy for intra-source species overlap, all of which are applied during data filtering.

A second part of the analysis focuses on identifying metadata fields that are relevant for establishing context to develop filtering rules in the next step. Here, context refers to the on-site activity (or activities) generating the release. For chemical manufacturing, such information might be

ascertained from something like a description of the equipment generating the release or the process in which the release is occurring. In a more general sense of LCI modeling, context and relevant metadata will vary from LCI to LCI and data source to data source.

2.1.2 | Rule-based data filtering

A first set of rules avoids double counting arising from intra-source species overlap. This situation can occur when facilities simultaneously report releases of chemical groups while reporting releases for individual chemicals within the group. This is most common with volatile organic compounds (VOCs) but must be fully determined when reviewing the species coverage for a data source. For life cycle impact assessment, impact factors are not derived for chemical groups and intra-source overlap rules should focus on a species hierarchy that preferentially selects data for individual species and adjusts releases of corresponding chemical groups accordingly.

A second set of rules resolves issues of inter-source overlap when multiple sources provide release data for the same chemical. These rules should be based on a source hierarchy that considers both data quality and utility. The general data quality of the sources established during their analyses is an obvious first choice to guide the rulemaking because it guarantees the best quality data is used for the LCI. However, the utility of the data depends on the ability to perform context-based filtering, which is subject to the available metadata within a data source. So, if two sources have similar data quality and one has context metadata while the other does not, the source with the metadata should be prioritized.

The final set of rules are the context-based rules. These are perhaps the most important rules because they are used to determine if and how a chemical release should be attributed to the focus chemical. Essentially, there are five general rules: (1) attribute to the focus chemical; (2) attribute to all chemicals produced within the facility; (3) attribute to a group of chemicals within the facility (e.g., organics); (4) attribute to an ancillary process such as heat production; and (5) exclude from the focus chemical. The determination of which rule to set for a chemical release depends on the context that can be ascertained from the source metadata. For example, if a release is from a boiler, it is attributed to an ancillary process (rule 4). If the release is described with a unit or process associated with another on-site chemical, it is excluded from the inventory (rule 5). Distribution to all chemicals (rule 2) is less desirable because it requires the use of allocation as defined in the ISO standards. It should only be necessary when insufficient metadata is available to establish the release context.

2.1.3 | Data sanitization

When dealing with secondary manufacturing data, CBI can apply in numerous ways as companies try to protect trade secrets. This is especially true of PV data, which is necessary for normalizing LCI to production of the focus chemical. Sanitizing this data can be handled in a few ways to allow inclusion of additional facilities in the LCI. A range can be used to mask the PV value, with a distribution specified to support uncertainty analysis. Weighted averaging can be used to mask CBI facilities within an average LCI, provided there are sufficient numbers of known and CBI PVs in the set of modeled facilities.

2.2 | LCI case studies

U.S.-average gate-to-gate chemical manufacturing LCIs were modeled using a data mining workflow developed from the proposed methodology in Section 2.1. Each database incorporated in the workflow supplied unique information about chemical releases and waste flows. The first case study modeled the production of cumene through the reaction of benzene and propylene in an alkylation process (Hwang & Chen, 2010). Propylene, benzene, and carbon dioxide are commonly reported emissions from the manufacture of cumene in commercially available LCI database such as ecoinvent and GaBi (Swiss Centre for Life Cycle Inventories, 2010; Thinkstep, 2016). The second case study modeled the production of sodium hydroxide from a mercury cell, diaphragm cell, or membrane cell (Eggeman, 2011). Caustic soda, salt, and hydrochloric acid are inputs to the chlor-alkali mercury cell process, and water, salt, and hydrochloric acid are inputs to the to the chlor-alkali diaphragm cell process. Chlorine is typically co-produced with sodium hydroxide.

The case studies focused on six EPA databases: PV data from the 2012 Chemical Data Reporting (CDR) database (U.S. EPA, 2020a); air releases from the 2011 National Emissions Inventory (NEI) (U.S. EPA, 2020b) and 2011 Greenhouse Gas Reporting Tool (e-GGRT) (U.S. EPA, 2020c); water releases from the 2011 Discharge Monitoring Report (DMR) (U.S. EPA, 2020d); air and water releases from the 2011 toxics release inventory (TRI) (U.S. EPA, 2020e); and hazardous waste generation from 2011 RCRAinfo Biennial Report (U.S. EPA, 2020f). While newer versions of the data sources are available, the 2012 CDR, representing 2011 chemical production, is the latest version to report chemical PVs by facility to support LCI calculations. A source analysis was first performed on each data source as described in Section 2.1.1. Then, data were collected for each inventory using the data mining workflow depicted in Figure 2. Manufacturing facilities were first identified in the CDR and then cross-walked into the other data sources using the Facility Registry Service (FRS) (U.S. EPA, 2020g). After the raw data was extracted, the results from the source analysis were used to filter the data on a facility-by-facility basis to create facility-level LCIs. This approach was necessary because the metadata varied from facility to

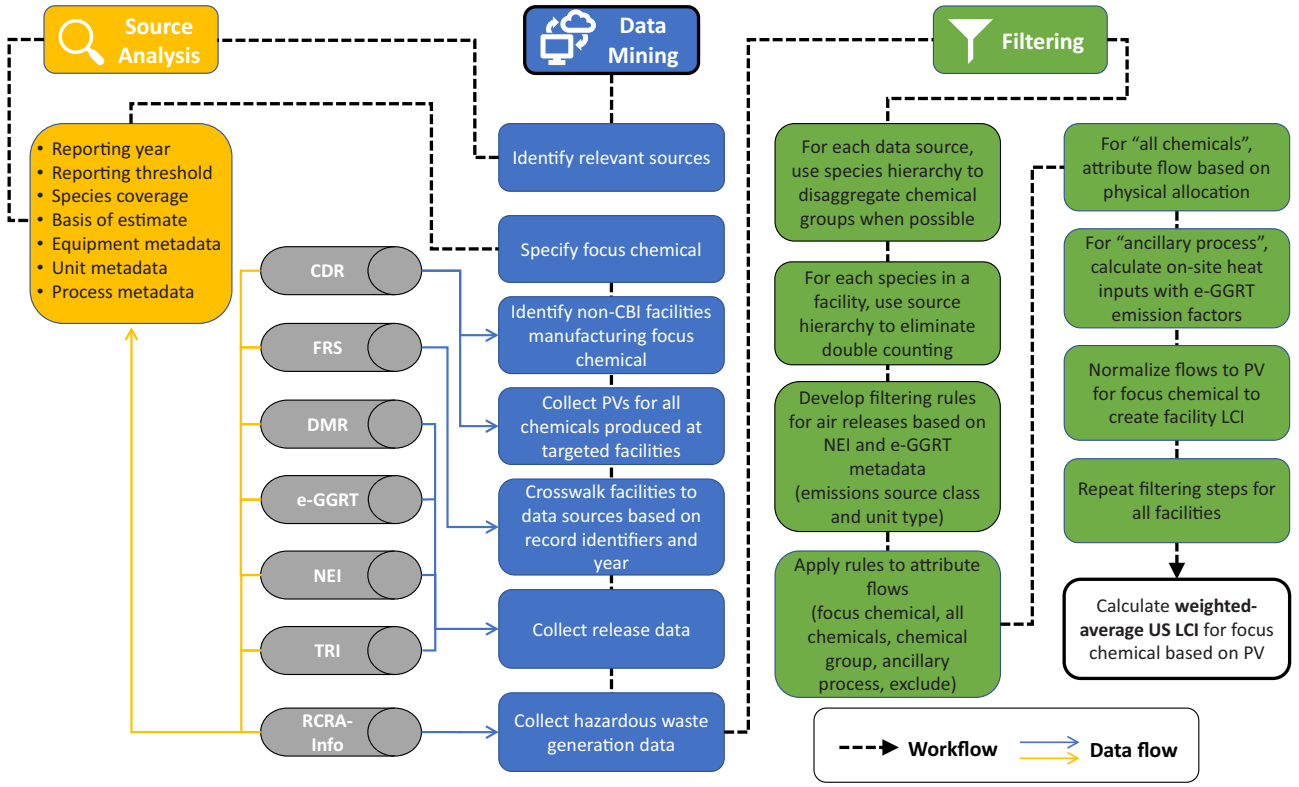


FIGURE 2 Revised data mining workflow to construct gate-to-gate chemical LCI from EPA data sources using data filtering

facility and affected how filtering rules were created for each species. Finally, the facility LCIs were combined into a U.S.-average LCI using PV data and a horizontal weighted-averaging approach (Henriksson, Guinée, Heijungs, de Koning, & Green, 2013).

Chemicals on the Toxic Substances Control Act (TSCA) Inventory produced or imported at 25,000 lb. or greater in a reporting year must be reported to CDR. Because reporters can claim their data as CBI, a key challenge for data mining workflows is the exclusion of CBI PVs and its impact on chemical production coverage. As a last step in this work, an approach based on horizontal weighted averaging was developed to sanitize CBI PV data and increase chemical production coverage, with only facilities below the PV reporting threshold not being represented. The data mining workflow was repeated for each CBI facility identified as a manufacturer of the case study chemicals and the U.S. weighted-average LCIs were recalculated to include the data from these additional facilities. A comparison of the LCIs with and without CBI was performed to demonstrate the effects of sanitization. In addition, the acetic acid case study in Cashman et al. (2016) was revisited as an additional test of the sanitization method. Access to CBI PV data in the CDR was obtained in compliance with guidelines under EPA's TSCA program. Since the sanitization method is pending approval, only mock sanitization results can be presented and discussed here.

3 | RESULTS

3.1 | Establishing data quality through source analysis

A source's flow reliability score, as described in Edelen and Ingwersen (2016), is a data quality indicator (DQI) that indicates the quality of the data generation method and the verification/validation of the data collection methods. For this work, scoring follows the recommendations of Edelen and Ingwersen (2016), with a flow reliability score of 1 (verified measurement) denoting the highest data quality and a score of 5 (undocumented estimate) representing the lowest data quality. When modeling multiple facilities, flow-specific scores can be aggregated across facilities by averaging scores based on the quantity of each exchange (Rousseaux et al., 2001; Edelen, 2018). For the case studies, NEI, e-GGRT, and TRI include a basis of estimate (i.e., how the value was derived) for each reported release that can be mapped to a flow reliability score as demonstrated in Cashman et al. (2016), with the full list of reliability scoring available in Table S1 in Supporting Information. CDR, RCRAInfo, and DMR do not report a flow-specific basis of estimate. Instead, the review of these databases determined that flows are required to be documented based on verified measurements (U.S. EPA, 2017a, 2017b). Therefore, LCI flows originating from these databases are assigned a recommended flow reliability score of 1, which means standardization of flows with CDR data will not result in decreased reliability.

TABLE 1 Species hierarchy rules for case study sources

Data source	Chemical group	Rule	Adjustment
NEI	Particulate matter	Select primary PM10 and PM2.5	$PM10_{Adjusted} = PM10_{PRI} - PM2.5_{PRI}$
NEI	Volatile organic compounds	Select individual species over VOC group totals	$VOC_{Adjusted} = VOC_{Reported} - \sum Species$
NEI	Polycyclic organic matter	Facilities can report by either species or group, but not both	None
DMR	Chemical and biological oxygen demand	Facilities can report both groups	Prioritize COD for chemical sector and filter out BOD

The other DQIs relate to flow representativeness and can be scored from 1 (highest) to 5 (lowest) by comparing characteristics of the data sources with the goal and scope of the inventory:

- Temporal correlation is derived from the database reporting year in relation to the LCI reference year. For example, all case study sources score a 3 for temporal correlation if 2018 is designated as the LCI reference year because the data is more than 6 years and less than 10 years older than the reference year.
- Geographical correlation depends on data source's geographic coverage. National databases will all score a 1 when applied for national average LCIs. For finer geographic resolution, such as regional or municipal production, scoring will depend on the granularity of location data collected by the data source.
- Technological correlation must be determined for each source using unit and process descriptions if available. If such information is not collected, some knowledge can be inferred from industry classification systems, such as the North American Industry Classification System that groups industries with similar processing technologies. In the case studies, NEI and e-GGRT attempted to collect unit and process information while the other sources did not. So only release data coming from these sources could be directly scored. Data from the other sources could only be roughly scored based on industry classification codes. If a workflow captures the majority of facilities producing the chemical in a designated region, a good technology correlation can be inferred for the resulting average LCI.
- Sampling method scoring relates to how much of known chemical production is covered by the data sources. This can depend on both reporting thresholds and the influence of CBI and must be determined on a case-by-case basis. For the case studies, CDR reports total national PV by chemical, which may include imports and CBI production volumes. However, it may not be possible to achieve 100% sampling method correlation because small manufacturers are exempt from reporting to CDR. It would even be difficult to achieve 100% coverage based on the CDR national totals because CBI facilities will be excluded from the workflow without sanitization.

3.2 | Resolving double counting with species hierarchies

The review of substance coverage for the case study data sources identified particulate matter (PM), VOCs, and polycyclic organic matter (POM), including polycyclic aromatic hydrocarbons (PAHs), as the most likely sources of species overlap. For groups like dioxins, xylenes, and cresols, data sources typically allowed either group totals or species to be reported, but not both. This is the approach used by TRI reporting, making the creation of such filtering rules only necessary for NEI and DMR. A summary of these rules is shown in Table 1. For NEI, all PM releases other than PM 2.5-PRI and PM10-PRI are excluded because the primary PM value includes the filterable and condensable PM subgroups. Since TRACI, EPA's North American impact assessment method (Bare, Norris, Pennington, & McKone, 2002), includes characterization factors for both PM10-PRI and PM2.5-PRI, the value of PM10-PRI was adjusted as indicated in Table 1. There is a similar case for VOCs because facilities may report both aggregated (total) VOC releases and speciated VOCs such as ethylbenzene, styrene, and glycol ethers. The VOC filtering rule developed for the case study workflows is to select speciated VOCs whenever possible because these species are typically characterized in TRACI while VOCs in general are not. A full list of speciated VOCs, such as the list derived for NEI from EPA's Industrial, Commercial, and Institutional Fuel Combustion Tool, Version 1.4 (U.S. EPA, 2015b) and provided in Table S2 in Supporting Information, was developed for each source when applicable and used to adjust aggregate VOC data as shown in Table 1. Further examination of reporting requirements for POM/PAH determined facilities could either report individual species or group totals, but not both and no filtering rules or adjustments were required.

The analysis of DMR determined facilities may report several releases related to the same pollutant, such as dissolved and total iron. This is especially common for organic enrichment and nutrient releases. For example, nitrogen may be reported as total nitrogen, TKN, organic nitrogen, and ammonia. The DMR technical documentation (U.S. EPA, 2012b) recommends hierarchies for each species where overlaps arise and these hierarchies were adopted for processing the data in the case study workflows. Adjustments to avoid nutrient discharge overestimation can be directly applied while acquiring the DMR data using the nutrient aggregation function. The one exception to this, as described in Table 1, was the use of

chemical oxygen demand (COD) over biological oxygen demand (BOD) given the focus of the workflow on the chemicals sector. BOD is filtered out to avoid double counting in eutrophication potential impact assessment results.

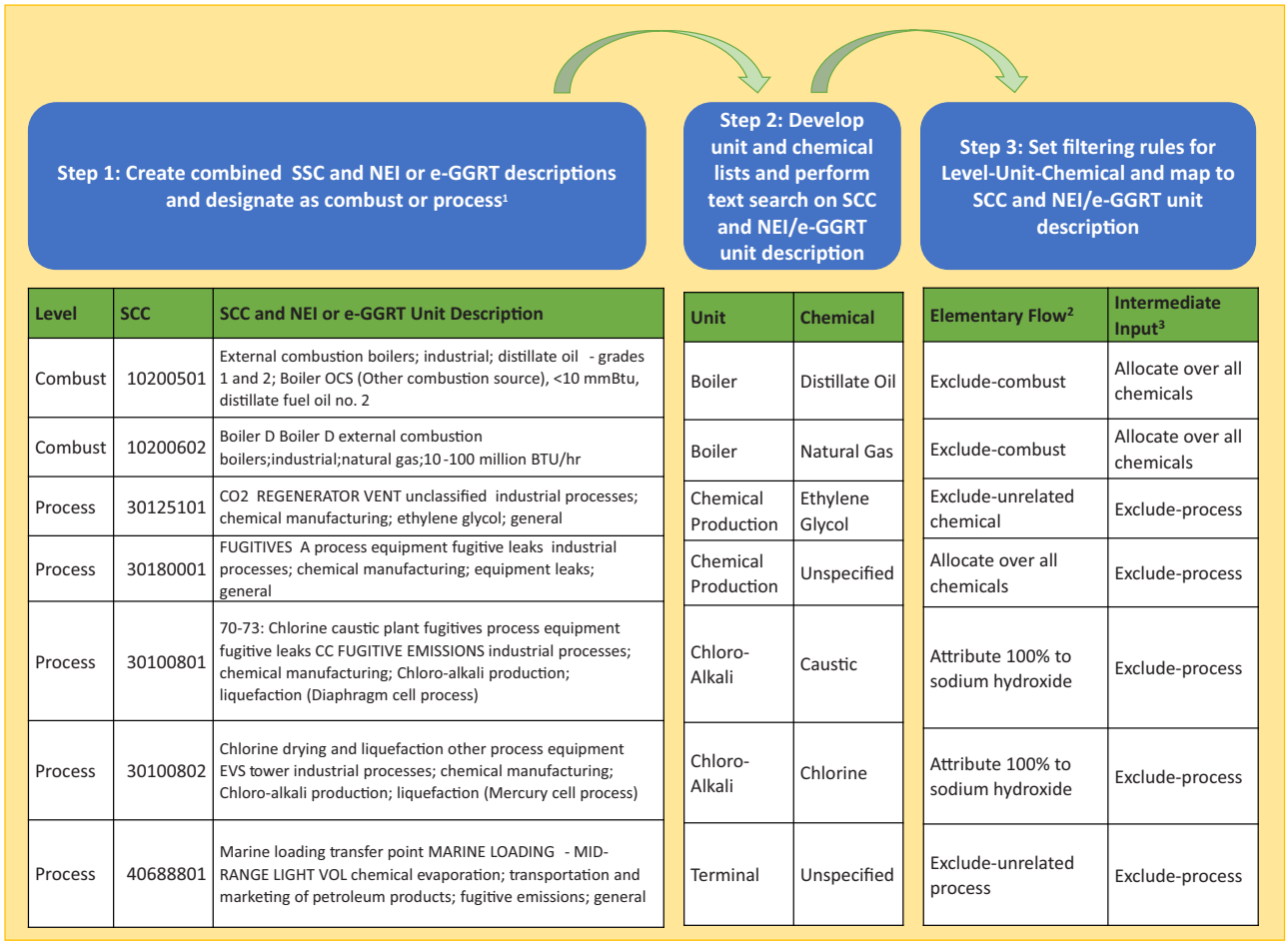
3.3 | Establishing a data source hierarchy for inter-source overlap

In 2014, 43,000 of the 66,000 stationary facilities in the NEI reported to the TRI (Strum et al., 2018). Resolving potential overlaps like this requires consistent rules to govern selection of sources. In general, matching specific release data between sources should be conducted to verify overlapping flow values are either a direct match or within a reasonable level of magnitude. Examples of source hierarchy filtering rules are presented here for handling air and water releases and hazardous waste generation in the case studies. Air releases from NEI are selected over TRI because NEI enables the use of context-based filtering rules while TRI, although offering more reporting accountability, lacks such process specificity. In the absence of context-based filtering, selection would have been based on flow reliability. Overlap of water releases reported in TRI and DMR are possible for ammonia, dioxins, metals, chlorine, polycyclic aromatic compounds, and phosphorus (U.S. EPA, 2015a) and DMR is selected over TRI because there is more information on water quality parameters in DMR. Finally, hazardous waste generation data from RCRAinfo are selected over TRI because RCRAinfo is not constrained to specific chemical constituents like TRI and includes detailed information on the source and management method of the hazardous waste (U.S. EPA, 2017a), which can support hazardous waste process modeling in future refinements of the workflow. Hierarchy rules like these only apply at the facility level when data is available in both sources. Horizontally averaged LCIs can contain data from both sources, as well as both species and chemical groups depending on which sources are included and how facilities are reported to these sources.

3.4 | Context-based data filtering and flow attribution

The development of context-based filtering rules was demonstrated in the case studies for air releases reported in NEI and e-GGRT because these sources provide the necessary metadata. Filtering rules for other data sources, such as the TRI or DMR, are more difficult to create because of the metadata provided in those sources and are the subject of future research. For NEI, the relevant data fields included the Source Classification Code (SCC) and "Emission Unit Description." SCCs are used in NEI to categorize activities that result in air releases based on the underlying source and process. Young et al. (2019) similarly applied SCCs to model refinery processes. Emission unit descriptions provide more specific information on the actual unit where the release occurs and enhance how SCCs can be used. Selected e-GGRT fields included the unit name, unit type, and fuel type because this information is similar to SCCs and emission unit descriptions. Development of context-based filtering rules followed a stepwise approach (Figure 3):

1. Create combined SCC and NEI or e-GGRT unit descriptions based on CDR facility list and assign an attribution of "process," "combustion," or "waste."
   a. Create a unique list of all relevant SCC and emission unit description combinations in NEI.
   b. Concatenate unit name, unit type, and fuel type data and map to SCCs to create a unique list for e-GGRT.
   c. For both lists, assign a designation for the overall unit process "level" by indicating whether each SCC is combustion, waste, or process, with process simply meaning not a waste or combustion SCC. This designation enables combustion and waste-related flows to be excluded from the LCI and attributed to intermediate inputs from ancillary processes.
2. Develop unit and chemical lists for further text searching of SCC-based lists established in Step 1.
   a. Create a "Unit" hierarchy list relevant to chemical production. Examples of unit types identified for the cumene case study include flare, tank, cooling tower, and boiler. An example unit hierarchy is provided in Table S3 in Supporting Information.
   b. Create a "Chemical" hierarchy by searching for chemical names present in the SCC-based lists (Step 1). Examples of chemicals returned from cumene case study searches include cumene, benzene/toluene/aromatics/xylenes, propylene, and phenol. An example chemical hierarchy is also provided in Table S3 in Supporting Information.
   c. Search the text in the SCC-based lists (Step 1) to identify the unit and chemical for each combination and create a combined summary name for the Level, Unit Type, and Chemical Type to assist with setting filtering rules in Step 3.
3. Set filtering rules for unique Level–Unit–Chemical combinations and map to SCC-based lists in Step 1.
   a. Filtering rules attribute releases to five options: the focus chemical, all chemicals, a chemical grouping, ancillary processes, or no chemicals (exclude).
   b. Develop rules separately for elementary flows and intermediate inputs associated with combustion and waste processes because elementary flows associated with intermediate inputs are filtered out of the focus chemical LCI.
   c. Boiler heat is generally attributed to all chemicals because its metadata is not specific to individual chemicals.



Level	SCC	SCC and NEI or e-GGRT Unit Description
Combust	10200501	External combustion boilers; industrial; distillate oil - grades 1 and 2; Boiler OCS (Other combustion source), <10 mmBtu, distillate fuel oil no. 2
Combust	10200602	Boiler D Boiler D external combustion boilers;industrial;natural gas;10-100 million BTU/hr
Process	30125101	CO2 REGENERATOR VENT unclassified industrial processes; chemical manufacturing; ethylene glycol; general
Process	30180001	FUGITIVES A process equipment fugitive leaks industrial processes; chemical manufacturing; equipment leaks; general
Process	30100801	70-73: Chlorine caustic plant fugitives process equipment fugitive leaks CC FUGITIVE EMISSIONS industrial processes; chemical manufacturing; Chloro-alkali production; liquefaction (Diaphragm cell process)
Process	30100802	Chlorine drying and liquefaction other process equipment EVS tower industrial processes; chemical manufacturing; Chloro-alkali production; liquefaction (Mercury cell process)
Process	40688801	Marine loading transfer point MARINE LOADING - MID-RANGE LIGHT VOL chemical evaporation; transportation and marketing of petroleum products; fugitive emissions; general

Unit	Chemical
Boiler	Distillate Oil
Boiler	Natural Gas
Chemical Production	Ethylene Glycol
Chemical Production	Unspecified
Chloro-Alkali	Caustic
Chloro-Alkali	Chlorine
Terminal	Unspecified

Elementary Flow[2]	Intermediate Input[3]
Exclude-combust	Allocate over all chemicals
Exclude-combust	Allocate over all chemicals
Exclude-unrelated chemical	Exclude-process
Allocate over all chemicals	Exclude-process
Attribute 100% to sodium hydroxide	Exclude-process
Attribute 100% to sodium hydroxide	Exclude-process
Exclude-unrelated process	Exclude-process

[1] Combust levels are filtered out of gate-to-gate inventory; intermediate associated with combust processes to link inventories.
[2] Data filtering rules for elementary flows such as air emissions.
[3] Data filtering rules for intermediate inputs such as heat quantities associated with combustion processes.
Acronyms: e-GGRT,Electronic Greenhouse Gas Reporting Tool; NEI, National Emissions Inventory; SCC, Source Classification Code.

FIGURE 3 Example of developing data filtering rules and matching to metadata within environmental release database

d. Upstream chemical processes required for production of the focus chemical at a facility are incorporated within flows attributed to the focus chemical.

Because the filtering rules incorporate all potential metadata for the focus chemical, they can be applied to any facility producing the chemical. After filtering, the remaining release data is converted to a facility LCI by normalizing to the following depending on the flow attribution: focus chemical PV, chemical group PV (e.g., all organic chemicals), or total PV for all chemicals. The average U.S. LCI for the focus chemical can then be constructed by horizontally averaging the facility LCIs using PV weighting.

3.5 | LCIs for case study chemicals

For cumene, data filtering and attribution were applied to 15 intermediate flows and 74 air releases. Attribution was not performed on the 21 air releases only found in TRI or the 203 waste and water flows from DMR, TRI, and RCRAInfo. Table 2 summarizes the effects of filtering and attribution on air releases reported by the seven facilities with public PVs in CDR. Releases of propylene, benzene and cumene notably increased after context-based filtering. This indicates the method is correctly attributing more of the total relevant facility releases to the correct chemical production process. Decreases were seen for other releases unrelated to the production of cumene such as nitrogen oxides and particulate matter. Over 45 substances were excluded, helping to reduce errors from misattributing facility-level releases. Although no toluene releases were directly attributed to cumene, toluene releases increased after filtering because toluene releases at some facilities could only be attributed to organic chemical production. The full raw and filtered cumene LCIs are provided in Tables S4 and S5 in Supporting Information. Flow reliability is not affected by

TABLE 2 Sample cumene inventory results with and without attribution

Air releases	kg per kg cumene produced		% Change with process attribution	Facility count	Flow reliability score	Data source
	Non-attributed	Process attributed				
Toluene	6.3E-07	1.3E-06	110	7.00	2.00	TRI NEI
Cumene	1.0E-05	1.9E-05	94	7.00	2.20	TRI NEI
Benzene	4.3E-06	5.9E-06	37	8.00	2.19	NEI
Propylene	5.9E-06	7.8E-06	33	7.00	2.10	TRI
Volatile organic compounds	6.3E-05	6.2E-05	−2	8.00	2.14	NEI
Ethyl benzene	1.3E-07	1.0E-07	−21	7.00	1.95	NEI
PM10 primary	2.4E-06	1.8E-06	−27	7.00	2.81	NEI
Nitrogen oxides	3.9E-06	1.0E-06	−74	7.00	1.94	NEI

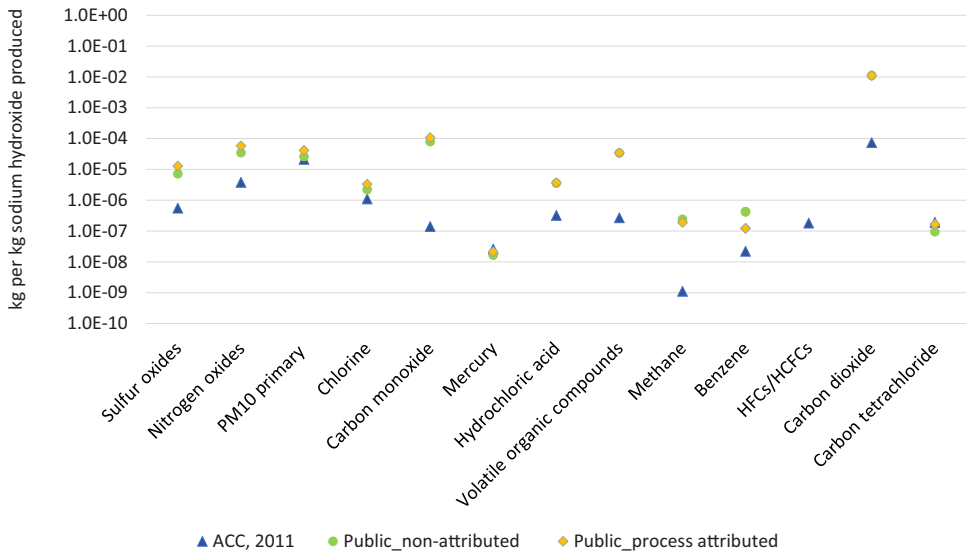


FIGURE 4 Results with and without process attribution for selected air releases during sodium hydroxide production. Included species correspond to those reported by a previous U.S.-average life cycle inventory (ACC, 2011). Underlying data used to create this figure can be found in the Table S8 of Supporting Information

filtering because scores are derived from the underlying basis of estimate. However, the quality of the LCI after filtering improves because technological correlation and completeness are more accurate by better capturing only the relevant flows associated with the focus chemical.

LCIs should communicate if data sources reported true zeros or if zeros represent a lack of reporting because full transparency on this issue improves the overall data quality. Recording the facility count (number of facilities reporting a release of a specific species) as shown in in Table 2 can indicate the likelihood a flow is associated with focus chemical production pathways. True zeroes are included in the flow count, while non-reporting zeros are excluded. A flow count for a species that matches the number of investigated facilities indicates that species typically occurs for production of the focus chemical, with the exception of releases that are reported as zeros. Non-reported values are essentially treated as zero values in the calculation of weighted-average LCI because the results will be artificially scaled up and overestimated if only reporting facilities are included. Minimum values are also reported as zero if non-reported flow values exist.

A previous U.S. LCI for sodium hydroxide production reports chlorine, carbon monoxide, PM, mercury, nonmethane VOCs, hydrogen chloride, and sulfur oxides as typical air releases (ACC, 2011). Figure 4 compares these data with the weighted-average air releases of these species from the workflow for the 24 facilities reporting public PVs. While the data mining releases are larger than those previously reported, several emissions such as mercury, chlorine, PM, and carbon tetrachloride are in reasonable agreement. Additionally, 22 air releases were completely excluded from the LCI using context-based filtering. Full gate-to-gate inventory results with and without process attribution are provided for sodium hydroxide in Tables S6 and S7 in Supporting Information.

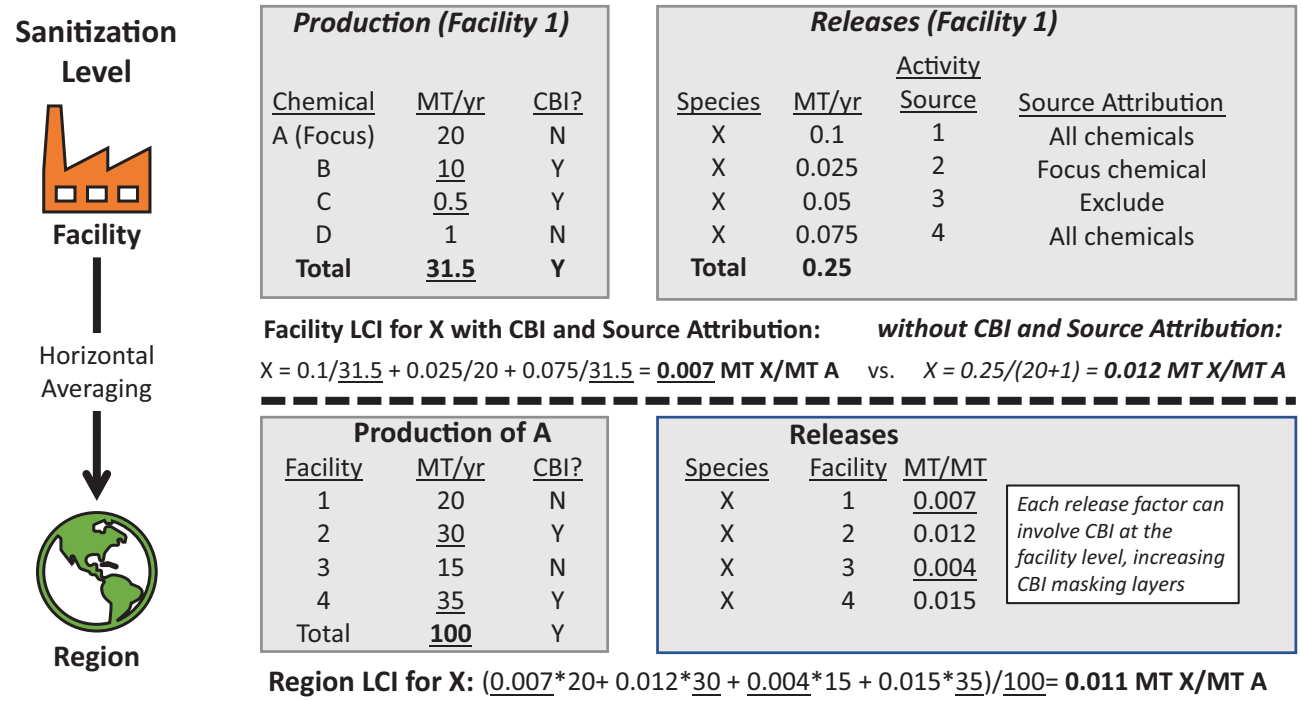


FIGURE 5 Mock CBI sanitization calculations at the facility and regional level. Instances of hypothetical CBI are underlined

3.6 | Optional CBI sanitization

In the case studies, there were multiple ways CBI was claimed by CDR reporters. The focus chemical PV, other chemical PVs, or any combination thereof were encountered as CBI across the numerous modeled facilities. In some cases, facilities even claimed knowledge of their manufacture of the focus chemical as CBI. Therefore, CBI sanitization involves the inclusion of various types of CBI from numerous facilities. The combination of data filtering and attribution with horizontal averaging provides multiple levels in which the various CBI data are collectively sanitized.

The multi-level sanitization approach is depicted in Figure 5 for a mock focus chemical A releasing species X during production. At the facility level, the emission factor (EF) of species X for Facility 1 is calculated by normalizing the release of X with the appropriate PV for the selected attribution rule. CBI PV data incorporated during normalization may be related to the focus chemical, other chemicals produced on site, or both. The inclusion of CBI and process attribution produces a smaller EF than the original data mining workflow that would have normalized total species X releases to the total publicly available facility PV. The EF calculation is repeated for all other facilities producing chemical A and horizontal averaging is used to obtain a sanitized U.S. EF for the release of species X. Thus, the different forms of CBI are masked through multiple levels of calculations across all facilities, with process attribution increasing the complexity and masking at all levels. The multiple layers of masking typically yield more than three confidential data points, which is the mathematical minimum for aggregating multiple confidential datasets for public release (UNEP-SETAC, 2011).

A summary of CBI in the case studies is shown in Table S9 in Supporting Information. The percentage of facilities claiming CBI for the focus chemical PV ranges from 10% for cumene to 64% for acetic acid. However, the range of facilities claiming at least some form of CBI is much higher at 59% (sodium hydroxide) to 92% (acetic acid). The percentage of total focus chemical production captured with only public CDR data ranged from only 1.2% for acetic acid to 82% for cumene. Although the case studies were selected to test varying levels of CBI for the focus chemical, all case studies included substantial instances of CBI across all chemicals that potentially reduced the technological correlation, completeness, and data collection methods of the resulting LCIs. The inclusion of CBI data improves the data collection methods score for acetic acid and sodium hydroxide from 4 to 1, while cumene is unchanged because 82% of total production is already reported as non-CBI.

A detailed look at cumene provides more insight into the limitations CBI can introduce to a data mining workflow and why it is important to apply sanitization if possible. Of the three focus chemicals, cumene would seem to be the least affected by CBI because 82% of its PV data is available from nine facilities publicly identifying themselves as domestic manufacturers. A tenth facility claims complete CBI (all information) and is omitted from the workflow. While only 1 of the 9 facility PVs is claimed as CBI, 66 PVs for other chemicals produced at 5 of the facilities are claimed as CBI, which is sizeable when considering a total of 348 PVs are associated with the 9 facilities. The fact that 19% of the required PV data is publicly inaccessible limits how data filtering and attribution can be applied. This shows that even for a chemical where total production is highly represented, there can be sufficient quantities of CBI in the underlying calculations for which sanitization is the only means to improve the quality of the resulting LCI.

To further illustrate this point, consider a brief example incorporating mock CBI data into the cumene calculations. The median public PV reported by the 9 cumene facilities was approximately 158,000 metric tons per year in 2011. This value was used to replace all CBI PV data and enable context-based filtering and attribution. The resulting air releases and heat inputs occurring at more than two facilities are shown in Table 3. Of the 47 flows, 19 decreased more than 10%, 21 increased more than 10%, and only 7 changed less than 10%. While one might expect the exclusion of CBI to overestimate EFs due to an artificial reduction in the denominator used to normalize releases to all chemicals, the results in Table 3 do not reflect this. The inclusion of an additional facility previously excluded because of a CBI focus chemical PV actually introduces the possibility of increasing EFs based on the facility's release data. The takeaway from this exercise is the effects of including CBI, although challenging to predict, can be significant in an LCI even when CBI may not seem like a key modeling challenge.

4 | DISCUSSION

The intended benefit of using data mining workflows to estimate chemical LCI is the ability to reduce the required time, resources, and skillset when compared to other methods such as process design/simulation or machine learning (Meyer et al., 2019). The revised workflow discussed here improves on this benefit by incorporating data source analysis, filtering rules, and flow attribution to create higher quality process LCI from facility-level data while being mindful of the desire to automate data processing. Although the case studies were modeled in Microsoft Excel templates to develop the workflow, pieces of the workflow have undergone preliminary automation (U.S. EPA, 2018c). The Standardized Emissions and Waste Inventory (StEWI) algorithm can extract raw (unfiltered) release data from the various EPA data sources as a first step in the workflow, provided the user inputs the list of facilities to be modeled. This same approach to automation can be taken with any publicly accessible data source.

In the downstream steps of the workflow, metadata identified in NEI and e-GGRT during source analysis provides the necessary context to help identify the applicability of each data point to the focus chemical LCI. For NEI, this information resides in the SCC descriptions and the emission unit descriptions, while e-GGRT contains unit names and unit and fuel type information that can be mapped to and combined with SCCs. The value of metadata such as SCCs for LCI modeling has been recently demonstrated in the updates to the Petroleum Refinery Life Cycle Inventory Model (Young et al. (2019)). Deploying this approach on a broader scale for LCI workflows in general can make data mining a more viable option. The caveat for developing context-based filtering rules in an automated fashion will be the ability to parse text description fields and conceptually link the information with the focus chemical. Efforts in text mining for cheminformatics can guide this process. For example, Krallinger, Rabal, Lourenço, Oyarzabal, and Valencia (2017) discuss the challenges of chemical entity recognition as a first step in data extraction and note the need for knowledge resources to deal with the often ambiguous and varying use of chemical names across datasets. For the EPA sources in the case study, this can be more easily addressed because the EPA maintains a substance registry service to describe all the ways chemicals are described in its data. More care will need to be taken when modeling environmental data sources in general, but the task should be manageable if the source analysis step is properly implemented. By combining cheminformatics with manufacturing process terminology, it will be possible to create automated filtering rules to better attribute facility-level data to process LCI. For substances that are not described by suitable metadata related to the focus chemical, improved attribution may require rules like a refined chemistry-based filter, which could be further developed for automated application by leveraging the growing number of chemical reaction databases (Krallinger et al., 2017).

The CBI sanitization method presented here can be easily automated and implemented because it requires no changes to the preprocessing and processing steps of LCI modeling. However, using the method for data mining workflows is not without its challenges because it will be restricted to only those with access to the CBI data. A potential solution for this issue will be to work with data hosts to automate workflows for all relevant focus chemicals in CBI-protected environments and release the final weighted-average LCIs for general use by practitioners. This may become a more viable approach as the use of LCA in environmental management grows and data hosts better understand the need for affordable, high quality LCI. The main drawback to this solution will be the need for the analysis to be repeated every time the data are corrected or updated. Finally, some may question whether the data are actually sanitized if only the final averaged data can be made publicly available. The data mining workflow outlined here meets the criteria for data sanitization because the value of the data is preserved for geographically averaged inventory modeling while the withheld information is protected. Further analysis of sanitization at the facility level is needed to determine if there are conditions when adequately sanitized facility LCI can be publicly released.

Continuing work on the use of data mining workflows for LCI modeling should address the following:

- For the EPA Data Mining Workflow, identifying suitable metadata in sources such as TRI and DMR to extend filtering capabilities to water discharges and air emissions not covered by NEI;
- For data filtering and flow attribution in general, leveraging existing chemical reaction datasets to develop more sophisticated filters that combine metadata with chemical process knowledge;
- For LCI modeling in general, developing workflows for other geographic regions and expanding coverage to other phases of the life cycle such as product manufacturing.

TABLE 3 Example cumene inventory with incorporation of mock CBI

Flow name	Units	Source	Process attribution—Public				Process attribution—CBI example				% Change with CBI
			Average U.S. LCI	Facility count	Maximum	Flow reliability score	Average U.S. LCI	Facility count	Maximum	Flow reliability score	
Products and services:											
Cumene	kg	CDR	1.00	8.00	1.00	NA	1.00	8.00	1.00		
Heat inputs to production:											
Industrial processes; petroleum industry; process heaters; natural gas fired	MJ	e-GGRT	0.14	3.00	4.09	2.00	0.11	3.00	4.09	2.00	−22
External combustion boilers; industrial; natural gas; 10–100 million BTU/hr	MJ	e-GGRT	0.098	2.00	4.83	2.00	0.28	3.00	4.83	2.00	184
External combustion boilers; industrial; process gas; petroleum refinery gas	MJ	e-GGRT	0.20	4.00	2.84	2.00	0.63	5.00	2.62	2.00	212
Industrial processes; petroleum industry; process heaters; process gas fired	MJ	e-GGRT	1.05	3.00	4.26	2.00	0.76	4.00	3.73	2.00	−28
External combustion boilers; industrial; natural gas; >100 million BTU/hr	MJ	e-GGRT	0.54	6.00	1.94	2.00	0.44	6.00	1.94	2.00	−19

(Continues)

TABLE 3 (Continued)

Flow name	Units	Source	Process attribution—Public				Process attribution—CBI example				% Change with CBI
			Average U.S. LCI	Facility count	Maximum	Flow reliability score	Average U.S. LCI	Facility count	Maximum	Flow reliability score	
Releases:											
1,2,4-Trimethylbenzene	kg	TRI	5.4E-08	5.00	7.0E-07	2.15	2.8E-07	6.00	1.3E-06	1.88	411
1,3-Butadiene	kg	NEI	1.0E-07	5.00	4.1E-07	1.62	7.6E-08	6.00	3.6E-07	1.69	−26
2,2,4-Trimethylpentane	kg	NEI	6.4E-08	4.00	1.5E-07	2.00	5.0E-08	5.00	1.5E-07	2.00	−23
Acetophenone	kg	NEI	1.2E-06	3.00	0.0026	2.25	6.0E-07	3.00	3.6E-05	2.25	−50
Ammonia	kg	TRI NEI	3.4E-07	6.00	0.0017	2.33	5.4E-08	7.00	2.4E-05	2.96	−84
Benzene	kg	NEI	5.9E-06	8.00	6.0E-04	2.19	1.0E-05	9.00	2.3E-05	2.18	72
Benzo[g,h,i,]perylene	kg	TRI NEI	8.1E-13	4.00	4.4E-12	3.39	1.3E-11	5.00	6.5E-11	3.77	1451
Carbon dioxide	kg	e-GGRT	0.013	5.00	0.052	2.10	0.0096	6.00	0.046	2.03	−26
Carbon monoxide	kg	NEI	6.2E-07	7.00	2.6E-04	1.83	1.8E-06	8.00	4.7E-06	2.03	185
Chlorine	kg	NEI TRI	1.5E-10	3.00	7.4E-09	4.52	5.4E-10	4.00	7.4E-09	3.89	271
Cumene	kg	NEI TRI	1.9E-05	7.00	1.3E-04	2.21	1.9E-05	8.00	1.3E-04	2.19	−1
Cumene hydroperoxide	kg	TRI	3.0E-08	3.00	5.3E-06	1.31	1.8E-08	3.00	5.3E-06	1.31	−40
Cyclohexane	kg	TRI	7.0E-08	6.00	2.3E-07	1.99	6.5E-08	7.00	2.2E-07	1.80	−6
Diethanolamine	kg	TRI NEI	3.8E-08	3.00	9.5E-08	2.00	1.1E-07	4.00	4.5E-07	2.00	189
Dinitrogen monoxide	kg	e-GGRT	4.5E-08	5.00	1.8E-07	1.99	3.6E-08	6.00	1.5E-07	1.94	−21
Dioxin and dioxin compounds	kg	TRI	4.2E-15	3.00	1.1E-12	2.98	2.5E-14	4.00	1.1E-12	2.59	488
Ethyl benzene	kg	NEI	1.0E-07	7.00	1.3E-04	1.95	1.8E-07	8.00	6.6E-07	1.99	74
Ethylene	kg	TRI	1.8E-06	5.00	7.5E-06	2.12	1.4E-06	6.00	6.6E-06	1.86	−26
Ethylene glycol	kg	NEI TRI	1.2E-09	2.00	5.0E-09	3.14	8.7E-10	3.00	4.4E-09	3.81	−28
Formaldehyde	kg	NEI	1.6E-09	2.00	9.1E-09	2.00	1.9E-09	3.00	9.1E-09	2.00	13
Hexane	kg	NEI	1.8E-07	6.00	4.8E-07	2.01	1.9E-07	7.00	4.2E-07	2.03	5
Hydrochloric acid	kg	NEI TRI	1.5E-08	4.00	5.3E-06	3.74	1.1E-08	5.00	5.3E-06	2.98	−23
Hydrogen fluoride	kg	NEI	4.7E-12	3.00	3.2E-11	4.08	9.2E-10	4.00	5.0E-09	3.43	19631

(Continues)

TABLE 3 (Continued)

Flow name	Units	Source	Process attribution—Public: Average U.S. LCI	Facility count	Maximum	Flow reliability score	Process attribution—CBI example: Average U.S. LCI	Facility count	Maximum	Flow reliability score	% Change with CBI
Lead	kg	TRI NEI	2.9E-11	4.00	1.2E-09	4.32	3.2E-11	5.00	1.1E-09	4.53	8
Mercury	kg	NEI TRI	1.6E-10	5.00	6.7E-10	2.30	1.3E-10	6.00	6.5E-10	2.94	−20
Methane	kg	e-GGRT	1.0E-05	5.00	5.0E-05	2.15	1.0E-05	6.00	4.5E-05	2.08	3
Methanol	kg	NEI TRI	1.5E-07	5.00	1.2E-06	2.51	1.1E-07	5.00	1.2E-06	2.51	−28
Methyl tert-butyl ether	kg	NEI	4.7E-09	3.00	1.9E-08	2.00	5.6E-09	3.00	1.7E-08	2.00	20
Molybdenum trioxide	kg	TRI	5.7E-11	2.00	5.2E-10	5.00	4.6E-11	3.00	5.2E-10	5.00	−19
Naphthalene	kg	NEI	1.2E-08	5.00	3.4E-08	2.00	3.3E-08	6.00	1.3E-07	2.02	181
Nickel	kg	NEI	2.5E-10	3.00	1.7E-09	4.08	3.6E-10	4.00	1.7E-09	4.37	43
Nitrogen oxides	kg	NEI	1.0E-06	7.00	3.2E-04	1.91	1.5E-06	8.00	2.2E-05	1.87	48
Phenol	kg	NEI TRI	4.8E-07	5.00	1.0E-05	2.66	4.0E-07	6.00	9.8E-04	2.56	−16
PM10 primary	kg	NEI	1.8E-06	8.00	6.4E-05	2.79	4.0E-06	8.00	1.0E-04	3.03	128
Polycyclic aromatic compounds	kg	TRI NEI	1.7E-09	5.00	7.0E-09	3.21	1.4E-09	6.00	6.9E-09	2.84	−19
Propylene	kg	TRI	7.8E-06	7.00	2.2E-05	2.10	6.4E-06	8.00	2.2E-05	1.90	−19
Styrene	kg	NEI	3.7E-09	3.00	1.5E-08	2.00	7.0E-09	4.00	2.3E-08	2.00	86
Sulfur dioxide	kg	NEI	6.1E-07	7.00	2.5E-06	2.25	3.2E-06	7.00	1.4E-05	2.66	423
Sulfuric acid	kg	TRI	2.0E-07	2.00	1.3E-06	1.33	5.5E-07	3.00	2.1E-06	1.58	179
Toluene	kg	TRI NEI	1.3E-06	7.00	5.2E-06	2.00	1.4E-06	8.00	5.0E-06	2.03	5
Volatile organic compounds	kg	NEI	6.2E-05	8.00	0.055	2.14	6.4E-05	8.00	3.5E-04	2.15	4
Xylenes (mixed isomers)	kg	NEI	2.6E-07	7.00	9.3E-07	2.00	1.8E-06	7.00	8.6E-06	2.02	580

Notes: Results shown for heat inputs and air releases that occur at more than two facilities. Results include the process attribution refinement. Demonstration of the CBI inclusion approach uses a mock value of 158 thousand metric tons per year for any CBI chemical production instance.

CBI, confidential business information; CDR, Chemical Data Reporting; e-GGRT, Electronic Greenhouse Gas Reporting Tool; NEI, National Emissions Inventory; LCI, life cycle inventory; TRI, toxics release inventory.

DISCLAIMER

The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

ACKNOWLEDGMENTS

The authors would like to acknowledge the following individuals: Ben Morelli and Stephen Treimel from ERG for their help during the development of the process allocation approach; Ben Young and Stacie Enoch from ERG for help with harmonizing database species; Scott Sherlock, Greg Macek, and Darryl Ballard from USEPA's Office of Pollution Prevention and Toxics for answering questions about CBI sanitization under TSCA, data collection under TSCA, and the CDR database, respectively; and John Abraham from USEPA's Office of Research and Development for reviewing various aspects of the project.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

David E. Meyer https://orcid.org/0000-0001-5753-6103
Sarah Cashman https://orcid.org/0000-0001-9859-9557

REFERENCES

Alvarado, V. I., Hsu, S.-C., Wu, Z., Lam, C.-M., Leng, L., Zhuang, H., & Lee, P.-H. (2019). A standardized stoichiometric life-cycle inventory for enhanced specificity in environmental assessment of sewage treatment. *Environmental Science & Technology*, *53*(9), 5111–5123. https://doi.org/10.1021/acs.est.9b01409

American Chemistry Council (ACC). (2011). *Revised final appendices: Cradle-to-gate life cycle inventory of nine plastic resins and four polyurethane precursors*, : . Prairie Village, KS: Franklin Associates. https://plastics.americanchemistry.com/LifeCycle-Inventory-of-9-Plastics-Resins-and-4-Polyurethane-Precursors-APPS-Only/

Bare, J., Norris, G. A., Pennington, D. W., & McKone, T. (2002). TRACI: The tool for the reduction and assessment of chemical and other environmental impacts. *Journal of Industrial Ecology*, *6*, 49–78.

Bretz, R., & Frankhauser, P. (1996). Screening LCA for large numbers of products. *The International Journal of Life Cycle Assessment*, *1*, 139–146. https://doi.org/10.1007/BF02978941

Cashman, S. A., Meyer, D. E., Edelen, A., Ingwersen, W. W., Abraham, J. P., Barrett, W. M., … 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*, 9013–9025.

Clean Air Act (1990). 42 United States Code, Chapter 85, Sections 7401–7626, As amended November 15, 1990.

Csiszar, S. A., Meyer, D. E., Dionisio, K. L., Egeghy, P., Isaacs, K. K., Price, P. S., … Bare, J. C. (2016). Conceptual framework to extend life cycle assessment using near-field human exposure modeling and high-throughput tools for chemicals. *Environmental Science & Technology*, *50*, 11922–11934.

Debattista, J., Lange, C., Scerri, S., & Auer, S. (2015). Linked 'Big' data: Towards a manifold increase in big data value and veracity. *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, 92–98. https://doi.org/10.1109/BDC.2015.34

Edelen, A., & Ingwersen, W. (2016). *Guidance on data quality assessment for life cycle inventory data*. United States Environmental Protection Agency, EPA/600/R-16/096.

Edelen, A., & Ingwersen, W. (2018). The creation, management, and use of data quality information for life cycle assessment. *The International Journal of Life Cycle Assessment*, *23*, 759–772.

Eggeman, T. (2011). *Sodium hydroxide*. Kirk-Othmer Encyclopedia of Chemical Technology. https://doi.org/10.1002/0471238961.1915040905070705.a01.pub2

Emergency Planning and Community Right-to-Know Act of 1986 (1986). 42 United States Code, Chapter 116, Sections 11001–11050, October 17, 1986.

Frank R. Lautenberg Chemical Safety for the 21st Century Act (2016). Public Law Number 114–182, 90 Stat. 2003 (December 18, 2016).

Garcia, S., Ramírez-Gallego, S., Luengo, J., Manuel Benítez, J., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics*, *1*, 9.

Garcia-Gil, D., Luengo, J., Garcia, S., & Herrera, F. (2017). Enabling smart data: Noise filtering in big data classification. Cornell University e-Print Service, arXiv:1704.01770v2 [cs.DB].

Geisler, G., Hofstetter, T. B., & Hungerbühler, K. (2004). Production of fine and speciality chemicals: Procedure for the estimation of LCIs. *The International Journal of Life Cycle Assessment*, *9*, 101–113. https://doi.org/10.1007/BF02978569

Henriksson, P. J. G., Guinée, J. B., Heijungs, R., de Koning, A., & Green, D. M. (2013). A protocol for horizontal averaging of unit process data – including estimates for uncertainty. *The International Journal of Life Cycle Assessment*, *19*, 429–436.

Hwang, S. Y., & Chen, S. S. (2010). *Cumene*. Kirk-Othmer Encyclopedia of Chemical Technology. https://doi.org/10.1002/0471238961.0321130519030821.a01.pub3

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

Jiménez-González, C., & Overcash, M. (2000b). Energy sub-modules applied in life-cycle inventory of processes. *Clean Products and Processes*, *2*, 57–66. https://doi.org/10.1007/s100980050051

Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., & Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, *117*, 7673–7761.

Liao, M., Kelley, S., & Yao, Y. (2020). Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. *ACS Sustainable Chemistry & Engineering*, *8*, 1252—1261. https://doi.org/10.1021/acssuschemeng.9b06522

Meyer, D. E., Mittal, V. K., Ingwersen, W. W., Ruiz-Mercado, G. J., Barrett, W. M., Gonzalez, M. A., … 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

Parvatker, A. G., & Eckelman, M. J. (2018). Comparative evaluation of chemical life cycle inventory generation methods and implications for life cycle assessment results. *ACS Sustainable Chemistry & Engineering*, *7*, 350—367. https://doi.org/10.1021/acssuschemeng.8b03656

Parvatker, A. G., Tunceroglu, H., Sherman, J. D., Coish, P., Anastas, P., Zimmerman, J. B., & Eckelman, M. J. (2019). Cradle-to-gate greenhouse gas emissions for twenty anesthetic active pharmaceutical ingredients based on process scale-up and process design calculations. *ACS Sustainable Chemistry & Engineering*, *7*(7), 6580–6591. https://doi.org/10.1021/acssuschemeng.8b05473

Pereira, C., Hauner, I., Hungerbühler, K., & Papadokonstantakis, S. (2018). Gate-to-gate energy consumption in chemical batch plants: Statistical models based on reaction synthesis type. *ACS Sustainable Chemistry & Engineering*, *6*(5), 5784–5796. https://doi.org/10.1021/acssuschemeng.7b03769

Rousseaux, P., Labouze, E., Suh, Y., Blanc, I., Gaveglia, V., & Navarro, A. (2001). An overall assessment of life cycle inventory quality. *The International Journal of Life Cycle Assessment*, *6*(5), 299–306.

Simon, T., Yang, Y., Lee, W. J., Zhao, J., Li, L., & Zhao, F. (2019). Reusable unit process life cycle inventory for manufacturing: Stereolithography. *Production Engineering*, *13*, 675–684. https://doi.org/10.1007/s11740-019-00916-0

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.

Song, R., Keller, A. A., & Suh, S. (2017). Rapid life-cycle impact screening using artificial neural networks. *Environmental Science & Technology*, *51*(18), 10777–10785. https://doi.org/10.1021/acs.est.7b02862

Strum, M., Wu, C. Y., Banas, R., Basnight, D., Griffin, S., Drukenbrod, J., … Wilbanks, C. (2018). *SLT/NEI/TRI R&D team final report and recommendations*. https://www.epa.gov/sites/production/files/2019-02/documents/final-report-phase2-tri-nei-slt.pdf

Swiss Centre for Life Cycle Inventories. (2010). *Ecoinvent database*, version 2.2. Dübendorf, Switzerland: Ecoinvent Centre.

Thinkstep. (2016). *GaBi databases 2016*. Retrieved from http://www.gabi-software.com/support/gabi/gabi-database-2016-lci-documentation/

UNEP-SETAC (2011). Aggregated data development. *Global guidance principles for life cycle assessment databases*. New York: United Nations Environment Programme. https://www.lifecycleinitiative.org/wp-content/uploads/2012/12/2011%20-%20Global%20Guidance%20Principles.pdf

U.S. EPA. (2012a). *Instructions for the 2012 TSCA chemical data reporting*. Retrieved from https://www.epa.gov/chemical-data-reporting/instructions-2012-tsca-chemical-data-reporting

U.S. EPA. (2012b). *Technical users background document for the discharge monitoring report (DMR) pollutant loading tool*, Version 1.0, January 2012.

U.S. EPA. (2015a). *TRI national analysis 2013: Updated January 2015*. Retrieved from https://www.epa.gov/sites/production/files/2017-01/documents/2013-tri-national-analysis-complete_1_0.pdf

U.S. EPA. (2015b). *USEPA Industrial, commercial, and institutional (ICI) fuel combustion tool*, Version 1, December 2015.

U.S. EPA. (2016). *2016 Chemical data reporting results*. Retrieved from https://www.epa.gov/chemical-data-reporting/2016-chemical-data-reporting-results

U.S. EPA. (2017a). *Biennial report overview*. Retrieved from https://rcrapublic.epa.gov/rcrainfoweb/action/modules/br/main/broverview on May 28, 2020.

U.S. EPA. (2017b). *Clean water act analytical methods*. Retrieved from https://echo.epa.gov/trends/loading-tool/water-pollution-search

U.S. EPA. (2018a). *ChemView*. Retrieved from https://chemview.epa.gov/chemview

U.S. EPA. (2018b). *Enforcement and compliance history inline (ECHO) water pollutant loading tool*. Retrieved from https://echo.epa.gov/trends/loading-tool/water-pollution-search

U.S. EPA (2018c). *Standardized emission and waste inventories (StEWI)*. Retrieved from https://github.com/USEPA/standardizedinventories

U.S. EPA (2020a). *2012 Chemical data reporting database*. Retrieved from https://www.epa.gov/chemical-data-reporting/access-cdr-data#2012

U.S. EPA (2020b). *2011 National emissions inventory (NEI) data*. Retrieved from https://www.epa.gov/air-emissions-inventories/2011-national-emissions-inventory-nei-data

U.S. EPA (2020c). *Greenhouse gas reporting program (GHGRP)*. Retrieved from https://www.epa.gov/ghgreporting

U.S. EPA (2020d). *Water pollution search*. Retrieved from https://echo.epa.gov/trends/loading-tool/water-pollution-search/

U.S. EPA (2020e). *TRI basic plus data files: Calendar years 1987–2018*. Retrieved from https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-plus-data-files-calendar-years-1987-2018

U.S. EPA (2020f). *RCRAInfo public extract*. Retrieved from https://rcrapublic.epa.gov/rcra-public-export/?outputType=CSV

U.S. EPA (2020g). *Facility registry service (FRS)*. Retrieved from https://www.epa.gov/frs

Wernet, G., Hellweg, S., Fischer, U., Papadokonstantakis, S., & Hungerbühler, K. (2008). Molecular-structure-based models of chemical inventories using neural networks. *Environmental Science & Technology*, *42*(17), 6717–6722. https://doi.org/10.1021/es7022362

Yao, Y., & Masanet, E. (2018). Life-cycle modeling framework for generating energy and greenhouse gas emissions inventory of emerging technologies in the chemical industry. *Journal of Cleaner Production*, *172*, 768–777. https://doi.org/10.1016/j.jclepro.2017.10.125

Young, B., Hottle, T., Hawkins, T., Jamieson, M., Cooney, G., Motazedi, K., & Bergerson, J. (2019). Expansion of the petroleum refinery life cycle inventory model to support characterization of a full suite of commonly tracked impact potentials. *Environmental Science & Technology*, *53*(4), 2238–2248. https://doi.org/10.1021/acs.est.8b05572

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Meyer DE, Cashman S, Gaglione A. Improving the reliability of chemical manufacturing life cycle inventory constructed using secondary data. *J Ind Ecol*. 2020;1–16. https://doi.org/10.1111/jiec.13044