Green Chemistry

Check for updates

CarAT: carbon atom tracing across industrial chemical value chains *via* chemistry language models

Emma Pajak, [a] David Walz, [b] Olga Walz, [b] Laura Marie Helleckes, [a] Klaus Hellgardt [c] and Antonio del Rio Chanona *[a]

The chemical industry is increasingly prioritizing sustainability, with a focus on reducing its carbon footprint to achieve net zero. By 2026, the Together for Sustainability consortium will require reporting the biogenic carbon content (BCC) in chemical products, posing a challenge as the BCC depends on feedstocks, value chain configuration, and process-specific variables. While carbon-14 isotope analysis can measure the BCC, it is impractical for continuous industrial monitoring. This work presents CarAT (Carbon Atom Tracker), an automated methodology for calculating the BCC across industrial value chains, enabling iterative and accurate sustainability reporting. The approach leverages existing Enterprise Resource Planning data in three stages: preparing value chain data, performing atom mapping in chemical reactions using chemistry language models, and applying a linear program to calculate the BCC given known inlet compositions. The methodology is validated on a 27-node industrial toluene diisocyanate value chain. Three scenarios are analyzed: a base case with all fossil feedstocks, a case incorporating a renewable feedstock, and a butanediol value chain with a recycle stream. The results are visualized using Sankey diagrams, showing the flow of carbon attributes across the value chain. The key contribution is a scalable, automated framework for BCC calculation that can update as industrial conditions change. CarAT enables chemical manufacturers to comply with upcoming sustainability mandates while supporting carbon neutrality goals by facilitating the systematic substitution of fossil carbon with biogenic alternatives. By providing transparent, auditable tracking of carbon sources throughout production networks, this framework empowers the broader chemical industry to make data-driven decisions for achieving net-zero targets and accelerating the transition to sustainable manufacturing.

Green foundation

1. This research advances green chemistry by providing an automated framework to trace the BCC in complex industrial value chains, fostering accountability in feedstock sourcing and transitions toward carbon neutrality.

2. Using a synergy of machine-learning-based atom mapping and linear optimization, the framework allows for accurate BCC calculations that can be updated continuously to reflect value chain changes in sourcing, processing, and distribution.

3. In future work, the framework could be leveraged for scenario analysis and value chain optimization to address a market demand for high-BCC products, thus facilitating a transition toward non-fossil feedstocks.

1. Introduction

The global chemical industry faces mounting pressure to achieve net-zero emission targets as part of broader decarboni-zation efforts across all industrial sectors. Chemical manufacturing, which accounts for approximately 5% of global greenhouse gas emissions, plays a critical role in this transition due to its dual positions as both a significant emitter and an enabler of sustainable solutions across multiple value chains. Approaches to advancing sustainability in chemical engineering are varied, with recent developments spanning life cycle-based optimization frameworks,[1] waste valorization through brewing innovations,[2] and intensified separation processes[3] among others.[4,5] This drive towards sustainability has cata-lyzed the development of comprehensive frameworks for

*[a]The Sargent Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK.
E-mail: a.del-rio-chanona@imperial.ac.uk
[b]BASF SE, Ludwigshafen, Germany
[c]Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK*

measuring, reporting, and ultimately reducing the carbon footprint of chemical products throughout their life cycle. Leading this transformation is the Together for Sustainability (TfS) consortium, comprising major chemical manufacturers including BASF, Bayer, Evonik Industries, Henkel, Lanxess, and Solvay.[6] TfS aims to set global standards for sustainability reporting within the chemical industry, building upon principles established by the United Nations Global Compact, the Responsible Care Global Charter, and standards set by the International Labor Organization and the International Organization for Standardization. The consortium has introduced specific requirements for chemical companies to report their products' BCC by 2026, creating an urgent need for standardized assessment methodologies.

1.1 Value chains in the chemical industry

Value chains, first defined by Porter as a "*set of activities that are performed to design, produce, market, deliver, and support its product*",[7] are fundamental to the chemical industry. In this context, a value chain refers specifically to the underlying network of raw materials, production facilities, and products – a system of value-adding chemical reactions and transformations that derives high-value products from feedstock.[8] While supply chains focus on the physical movement and logistics of materials, value chains encompass the entire value-creation process, including the chemical transformations that occur at each stage.

An intrinsic property of chemical value chains is their interconnectedness, arising from synthesis pathways that draw on reactants from various sources, thereby linking the pathways of different products.[9] These chains can be vast and complex, with some synthesis pathways relying on a series of intricate chemical transformations to produce desired compounds. Recycle streams add further complexity by creating additional loops and interactions within the production process. Major chemical manufacturers such as BASF, Dow Chemical, Shell Chemicals, and Mitsubishi Chemical operate extensive global value chains comprising numerous interconnected pathways, encompassing hundreds of thousands of nodes and yielding thousands of diverse commercial products.[10]

1.2 Challenges in molecular-Level sustainability analysis

A fundamental challenge in achieving comprehensive sustainability assessment lies in the molecular-level opacity of existing business systems. Value chains are typically represented through Enterprise Resource Planning (ERP) systems, which integrate various business processes and functions into a unified system. While ERP data encompasses procurement, manufacturing, logistics, sales, and finance operations, it does not provide immediate or transparent access to the molecular-level details of a chemical value chain. Although these systems track the flow of materials and products, they do not inherently reveal the specific molecular species involved at each node.

This lack of molecular transparency poses significant challenges for sustainability assessment and reporting. Without visibility into the chemical transformations occurring at each production stage, it becomes difficult to:

• Track the origin and fate of carbon atoms through complex reaction networks.
• Assess the potential for substituting fossil-derived inputs with biogenic alternatives.
• Calculate accurate sustainability metrics that require a molecular-level understanding.
• Identify optimization opportunities for reducing the environmental impact.

1.3 Sustainability reporting requirements

At the core of sustainability reporting for chemical manufacturers is the Product Carbon Footprint (PCF), a metric that quantifies the total greenhouse gas emissions associated with a product throughout its life cycle.[11] PCF sums up the total greenhouse gas emissions generated by a product over different stages of its life cycle, measured in CO_2 equivalents (CO_2e), a standardized unit that expresses the global warming potential of various greenhouse gases relative to carbon dioxide.

The TfS Guidelines provide a specialized framework for calculating PCFs for chemical products, ensuring adherence to internationally recognized standards for greenhouse gas accounting and environmental assessment. These guidelines align with Principle 7 of Green Chemistry, which advocates for the use of renewable feedstocks rather than depleting ones, by providing mechanisms to track and incentivize the transition from fossil to biogenic carbon sources throughout chemical value chains.

A critical upcoming requirement from the TfS Guidelines is the reporting of a product's Biogenic Carbon Content (BCC), starting in 2026.[6,11] Biogenic carbon is defined by the World Business Council for Sustainable Development as "*carbon derived from living organisms or biological processes, but not fossilized materials or fossil sources*".[12] Typical sources include trees, plants, and soil, which absorb CO_2 as a natural part of their life cycle.[13]

The introduction of BCC reporting serves a crucial purpose: it supports the estimation of end-of-life emissions, which fall outside the cradle-to-gate scope of PCF calculations (see Fig. 1). The cradle-to-gate boundary encompasses all emissions
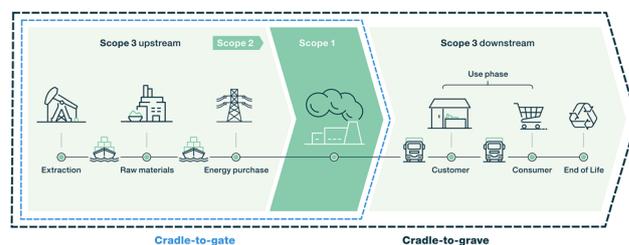

Fig. 1 System boundary definition for Product Carbon Footprint (PCF) calculations showing the cradle-to-gate scope. The PCF includes Scope 1 (direct emissions from owned or controlled sources), Scope 2 (indirect emissions from purchased energy), and upstream Scope 3 emissions (from purchased goods and services). Downstream Scope 3 emissions from product use and end-of-life disposal are excluded from PCF calculations but can be estimated using BCC data. Figure adapted from TfS.[6]

from raw material extraction (Scope 3 upstream) through production processes (Scope 1 and 2) to the factory gate, but excludes downstream emissions from product use and disposal (Scope 3 downstream). For example, a product whose carbon is entirely biogenic would contribute to no fossil CO_2 emissions through combustion or degradation, regardless of the end-of-life treatment.

The ability to quickly calculate and recalculate BCC becomes increasingly critical as manufacturing landscapes evolve. As resilient supply chains expand the availability of biogenic raw materials,[14] manufacturers need rapid BCC assessments for scenario analysis and credible end-of-life emission estimates. Similarly, as net-zero chemical pathways scale up "*carbon capture, low-carbon hydrogen, carbon storage, biomass utilization*"[15] and other technologies, manufacturers will require a BCC framework that can quickly update with changes in routes, feedstocks, and recycling to preserve transparent, auditable product-level attribution.

Calculating the BCC for products within complex value chains presents technical challenges. Currently, this calculation is only feasible when the chemical structure clearly differentiates biogenic carbon atoms from non-biogenic ones. For instance, in an ethoxylated fatty acid – typically synthesized from a fatty acid and ethylene oxide – the fatty acid component is generally derived from vegetable oil (biogenic), while ethylene oxide may be fossil-derived. The biogenic content is then determined by the fraction of carbon originating from the fatty acid.

The complexity increases for products requiring multiple intermediates, which may have their feedstocks replaced with biogenic or recycled materials. Accurate BCC calculation demands comprehensive knowledge of all upstream reactions, including value chain configuration, raw material composition, recycle streams, and process-specific nuances affecting product composition. As BCC depends on these dynamic variables, any upstream changes necessitate recalculation – a significant burden in global value chains where sustainable feedstocks, process setups, and efficiencies frequently change.

1.4 The need for automated BCC calculation

Recent advances in automated PCF calculation methodologies have demonstrated the potential for transforming sustainability assessment in industrial settings. Notable examples include AllocNow,[16] CarbonMinds,[17] and Siemens' SiGreen,[18] which have established automated frameworks for comprehensive carbon accounting across complex value chains. These tools demonstrate key advantages of automation: providing consistent methodologies, enabling certification and traceability, and facilitating scenario analysis for decision-making.

Chemical manufacturers face a critical challenge: they must obtain certification for their sustainability metrics, yet recalculating BCC each time upstream changes occur (such as feedstock modifications or process efficiency improvements) is impractical. An alternative approach involves developing a general computational methodology that decouples the dynamic variables of the value chain from the calculation

process. By certifying the methodology itself, manufacturers automatically receive certification for all subsequent calculations performed using the approved procedure. BASF has successfully employed this strategy for PCF calculations through its SCOTT methodology,[11] creating a strong precedent for developing similar approaches for BCC.

Existing carbon flow accounting approaches have largely operated at aggregated or process levels. For example, Ohno *et al.* (2018)[19] used a waste input–output material flow analysis to quantify materially retained carbon across Japan's economy, offering valuable macro-level insight but relying on historical data that cannot dynamically resolve chemical transformations or carbon origins. Likewise, process-level studies such as Kätelhön *et al.* (2019)[20] have evaluated the climate-change mitigation potential of carbon capture and utilization (CCU) technologies, providing system-wide perspectives on emission reduction. While these static, mass-balance-based frameworks remain essential for understanding aggregate carbon stocks and process optimization, they are not designed to dynamically attribute carbon flows or distinguish between fossil and biogenic origins within chemical reaction networks. This capability is increasingly crucial as manufacturers must continuously reassess carbon attribution in response to changing feedstocks, process configurations, and value-chain designs. The present work addresses this gap by introducing a molecular-level methodology that enables atom-level carbon tracing and produces outputs compatible with broader assessment tools such as LCA and techno-economic analysis (TEA), providing a consistent, data-driven foundation for future sustainability evaluations.

1.5 Aim and scope

This work aims to present a comprehensive framework for calculating the BCC of products within chemical value chains, enabling compliance with upcoming TfS sustainability reporting requirements. The framework extends recent advances in automated PCF calculations to BCC assessment, inheriting key advantages: offering a common and consistent approach, ensuring certifiability and traceability throughout the value chain, and enabling robust scenario analysis and optimization.

The framework will decouple the computational method from value chain data to achieve methodological certification – similar to BASF's SCOTT approach for PCF – rather than relying on product-specific certification. To calculate a product's BCC or any elemental attribute share, the methodology traces atoms through the value chain to their point of origin. The proposed framework derives molecular-level insights by leveraging existing business-focused ERP data, avoiding the need to establish new datasets from scratch.

To achieve these overarching aims, the following objectives are defined:

• Identify and curate an industrial value chain case study that encapsulates challenges faced at scale to demonstrate and validate the methodology.

• Assess and implement an AI-assisted approach to propose atom mappings of value chain reactions.

• Formulate a method for dynamically computing elemental attribute shares of materials based on changing inputs (*e.g.*, feedstock composition, value chain configuration, *etc.*).

Beyond immediate sustainability reporting requirements, this methodology supports broader carbon neutrality goals by facilitating the substitution of fossil-derived inputs with biogenic or recycled alternatives, as discussed by Beer *et al.* (2025).[21] Such transparency provides decision-makers with an auditable basis for reducing Scope 1 and Scope 3 emissions while working toward net-zero targets.

A key aspect of this work is the application of existing machine learning models for atom mapping, which significantly reduces the manual burden of tracking chemical reactions across entire product portfolios. By applying state-of-the-art atom mapping algorithms to industrial value chains, we created a modular, parameter-agnostic workflow that enables rapid recalculations whenever feedstock compositions or value chain parameters change.

Although BCC is an established sustainability metric, CarAT provides the practical means to realize it across large-scale industrial chemical value chains. By bridging molecular-level carbon tracing with process-level assessment tools such as LCA and TEA, the framework operationalizes the BCC calculation in a way that aligns with established sustainability methodologies.

This integration of machine learning and systems-level optimization ensures compliance with TfS requirements while making proactive decarbonization strategies more practical, paving the way for more flexible, transparent, and ultimately greener chemical value chains.

2. Methodology

We present the CarAT (Carbon Atom Tracing) framework as the solution for determining BCC across industrial value chains. Given a value chain of known inlet attributes, *i.e.*, the fossil and biogenic share of the raw materials, CarAT determines the biogenic carbon share of the products. The three-stage approach is shown in Fig. 2, and aligns with the project objectives:

• **Industrial case studies (Section 2.1)**: creating a graph representation of the value chain, and pre-processing value chain data.

• **Atom mapping of chemical reactions (Section 2.2)**: atom mapping chemical reactions using a Chemistry Language model, enabling atom tracing across a production node.

• **Value chain model construction and optimization (Section 2.3)**: formulating and solving a linear program to determine the BCC of each substance in the value chain.

A realistic example, which is representative of the base-to-speciality chemical industry, is used to demonstrate and verify the framework. For this, an atom mapping machine learning model, RXNMapper, published by IBM,[22] is leveraged.

Finally, this framework builds on confidential industrial concepts currently under review in a BASF patent appli-
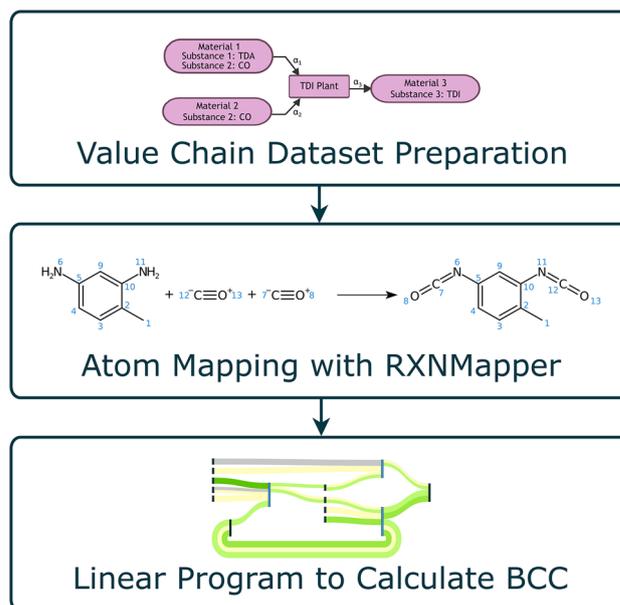

Fig. 2 Project methodology and framework for calculating the BCC across a value chain.

cation.[23] Specifically, the concepts and terms, bill of materials, bill of substances, and bill of atoms (to be introduced in this section) are included within the scope of the application.

2.1 Industrial case studies

An industrial toluene diisocyanate (TDI) value chain of 27 nodes was selected to demonstrate and validate the methodology. Although significantly smaller than a full value chain representing vertically integrated chemical companies such as BASF, Dow, or SABIC, it encapsulates key aspects present in the corresponding large-scale value chains. Furthermore, this work focuses on the transition to carbon neutrality by enabling the replacement of fossil-derived carbon with biogenic or recycled sources. The TDI synthesis pathway represents a realistic candidate for such decarbonization, for example, by substituting the methane used in syngas production with biogas or biomethane.[24] However, TDI synthesis lacks a recycle stream – one of the key challenges in value chain calculations. To address this, a smaller butanediol value chain featuring a recycle loop is also examined, demonstrating the framework's applicability under such complexity.

It is important to note that the industrial ERP dataset used in this work does not explicitly specify the chemical reactions occurring at each production step. Instead, it lists the materials entering and leaving each production node, representing the observed input–output relationships in real manufacturing systems. While this information defines the structural connectivity of the value chain, it does not capture how atoms are redistributed between reactants and products. Therefore, molecular-level tracing is required to determine how carbon atoms are transferred across transformations,

motivating the use of atom mapping as described in Section 2.2.

2.1.1 Value chain graph topology. A value chain can be conceptualized as a bipartite, directed graph, a fundamental mathematical structure in graph theory.[25] A bipartite graph $G = (V, E)$ consists of two distinct sets of vertices, $V = D \cup T$, where D and T are disjoint sets (*i.e.*, $D \cap T = \varnothing$), and every edge in E connects a vertex from D to a vertex in T.

In the context of the value chain, D represents virtual tanks, which serve as mix nodes. These virtual tanks are not actual physical containers where chemical reactions occur; rather, they are conceptual nodes introduced to segregate the convergence of a chemical from different sources before entering the actual chemical reactions. Where $D = \{d_1 d_2 \ldots d_m\}$, each $d_i \in D$ signifies a specific virtual tank (*e.g.* $d_1 d_2$) in Fig. 3.

Conversely, $T = \{t_1 t_2 \ldots t_j\}$ denotes production nodes where chemical reactions and transformations occur. Each $t_j \in T$ represents a production step, such as synthesis, separation, formulation, or relabeling (based on ERP data). A triplet t can have one or more input materials that are consumed, and one or more materials that are produced, where g denotes the main product, and p represents materials (products, byproducts, and reactants). The given value chain structure is such that one production facility can host more than one triplet t_j – this can be a consequence of it being a multi-purpose plant, or there being multiple production versions. Each triplet, t_j, in the value chain is uniquely identifiable by the ERP data code (c, b, g), where c, b, and g represent the company code, business process, and main product, respectively, *e.g.*, $(c_1 b_1 g_1)$ for t_1 in Fig. 3. Furthermore, each product can be further disaggregated into constituent substances s. Each mix node, d_i, is indexed by (c, p) – for instance, $(c_1 p_1)$ is the identifier for d_1 in Fig. 3. Additionally, e denotes the chemical element of interest (*e.g.*, carbon) and a denotes the elemental attributes (*e.g.*, biogenic, fossil, and recycled). In this work, e exclusively refers to carbon, though the CarAT framework is generalizable to other elements.

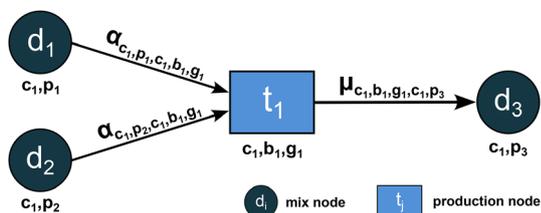Edges E between nodes in D and T represent material flows between virtual tanks and production facilities. For edges $(d_i t_j) \in E$, denoted by $\alpha_{c_i, p_i, c_j, b_j, g_j}$, the attribute α (input ratio) is defined as the kilograms of material from d_i consumed per kilogram of main output at t_j (see Fig. 3). Conversely, for edges $(t_j d_i) \in E$, denoted by $\mu_{c_j, b_j, g_j, c_i, p_i}$, the attribute μ (consumption mix share) indicates the fraction of the mixture in d_i originating from t_j (see Fig. 3).

The value chain is thus modeled as a bipartite directed graph, where edges connect virtual tanks and production nodes. This structure is applied to construct the 27-node TDI value chain. Atom mapping is required only at production nodes, where chemical transformations occur. In contrast, mix nodes (virtual tanks) involve no chemical changes and therefore do not require atom-level tracing.

2.1.2 TDI value chain. The synthesis pathway in this value chain has been comprehensively detailed in the literature and is protected by a patent.[26] The primary synthesis route begins with the nitration of toluene, forming 2,4-dinitrotoluene as the major product:

$$C_6H_5CH_3 + 2HNO_3 \rightarrow C_6H_3(NO_2)_2CH_3 + 2H_2O \qquad (1)$$
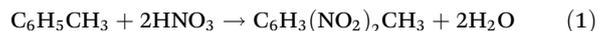
Subsequently, catalytic hydrogenation employing a nickel catalyst reduces the nitro groups of 2,4-dinitrotoluene into amine groups, yielding 2,4-diaminotoluene (TDA):

$$C_6H_3(NO_2)_2CH_3 + 6H_2 \rightarrow C_6H_3(NH_2)_2CH_3 + 2H_2O \qquad (2)$$

In the final stage, TDA undergoes phosgenation to produce TDI alongside hydrochloric acid:

$$C_6H_3(NH_2)_2CH_3 + 2COCl_2 \rightarrow C_6H_3(NCO)_2CH_3 + 4HCl \qquad (3)$$
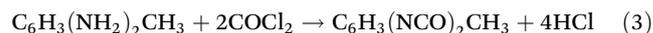
This synthesis requires phosgene, derived through a sub-branch of the value chain beginning with the steam reformation of methane-rich natural gas to form syngas, primarily composed of hydrogen and carbon monoxide. Purified carbon monoxide subsequently reacts with chlorine gas, forming the phosgene necessary for the final synthesis step.

2.1.3 Data structures for value chain representation. When working with value chains, it is necessary to create structures other than the graph topology, such as tables listing the node features, such as products, substances, or even atoms present in a given node. This work frequently refers to set datasets that present the value chain information at different levels, defined as follows:

• Bill of materials: a dataset of the recipes at each production node t_j, indicating the input ratios of each reactant, defined as the kilograms of material from the duplet consumed per kilogram of main output at the connected triplet, along with the corresponding output ratios of products p.

• Bill of substances: this adds a further layer of granularity to the bill of materials; it is a dataset of all substances s for a given production node/set of production nodes.

• Bill of atoms at a substance-level, $\phi_{s's_e}$: this is a dataset that designates the share of atoms with attribute a of a chemical element e in product substance s that originates from a reactant substance s′.

Fig. 3 Exemplary schematic of the bipartite directed-graph representation of a value chain. Mix nodes d_1, d_2, and d_3, indexed by (c, p) – here $(c_1 p_1)$, $(c_1 p_2)$, supply feedstock into the production node t_1, indexed by $(c_1 b_1 g_1)$. Arrows from d_1 and d_2 to t_1 carry input-ratio attributes $\alpha_{c_1 p_1 c_1 b_1 g_1}$ and $\alpha_{c_1 p_2 c_1 b_1 g_1}$, respectively. The arrow from t_1 to output mix node d_3, indexed by $(c_1 p_3)$, represents the output flow with consumption-mix share $\mu_{c_1 b_1 g_1 c_1 p_3}$.

• Bill of atoms at a material-level, $\psi_{p's'pse}$: this is a dataset that designates the share of atoms a of chemical element e in a product substance s in product material p that originates from a reactant substance s′ in reactant material p′; this distinction is particularly important when materials are not pure but are mixtures containing multiple substances.

• Consumption mix table: it is a dataset of all s for a given mix node/set of production nodes.

Note that substances are represented by a Simplified Molecular Input Line Entry System (SMILES), which is a string notation that allows a user to represent a chemical structure in a computer-readable format.[27]

To aid explanation, Tables 3 and 4 show an example bill of materials and bill of substances, respectively, for the TDI production node illustrated in Fig. 4, where toluene diamine (TDA) reacts with CO to form TDI. The corresponding bill of atoms is presented in Section 2.3.1. Note that while toluene appears in the node diagram (Fig. 4), it is omitted from the tables as it was only present in trace amounts below the threshold used in the workflow; substances below this threshold are excluded from further processing to reduce noise and computational burden.

Although not the case here, it is possible to have more than one reactant or product entry with the same substance, i.e., a

Table 1 Notation for indices used in the value chain model

Index	Description
a	Elemental attributes (e.g., biogenic, fossil, etc.)
b	Business process, anonymized coding: PLNT b
c	Company code, anonymized coding: COMP c
e	Chemical element (e.g., carbon)
g	Main product, same structure as p
p	Product, anonymized coding: PROD p
s	Substance, represented by a SMILES

Table 2 Notation for sets used in the graphical representation of the value chain

Notation	Description
D	Set of mix nodes
d	A mix/virtual tank node, or duplet
T	Set of production nodes
t	A production node, or triplet
E	Set of value chain edges
V	Set of all nodes

Table 3 Bill of materials for the TDI production node: t:COMP2| PLNT11|PROD29

Reaction role	Material	Material text	Ratio
Reactant	PROD31	TDA	0.53
Reactant	PROD19	Sodium hydroxide	0.04
Reactant	PROD6	Chlorine	0.46
Reactant	PROD10	Carbon monoxide	0.52
Product	PROD36	HCL	0.63
Product	PROD29	TDI	1.16

Table 4 Bill of substances for the TDI production node: t:COMP2| PLNT11|PROD29

| Reaction role	Material	Material text	SMILES	Ratio
Product	PROD29	TDI	Cc1ccc(N=C=O) cc1N=C=O	1.16
Product	PROD36	HCL	Cl	0.56
Product	PROD36	HCL	O=C=O	0.02
Product	PROD36	HCL	[C−]#[O+]	0.02
Product	PROD36	HCL	N#N	0.03
Reactant	PROD10	Carbon monoxide	[C−]#[O+]	0.52
Reactant	PROD6	Chlorine	ClCl	0.46
Reactant	PROD19	Sodium hydroxide	[Na+]·[OH−]	0.02
Reactant	PROD19	Sodium hydroxide	O	0.02
Reactant	PROD31	TDA	Cc1ccc(N)cc1N	0.53
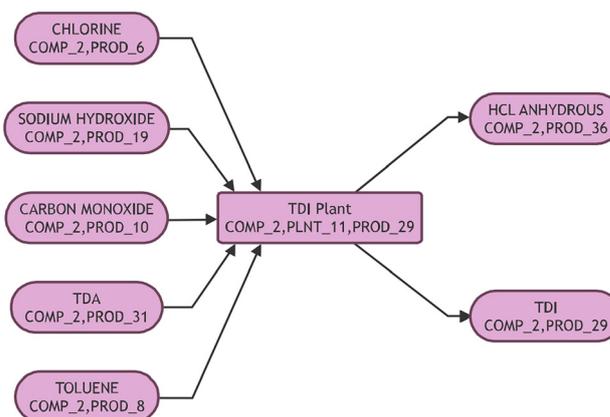
Fig. 4 Production node t:COMP2|PLNT11|PROD29, representing the transformation of TDA into TDI.

SMILES. This can arise when two different materials share a substance, e.g., if a substance has two different sources. For calculations, the bill of substances can be aggregated such that one entry represents the cumulative amount of each substance for reactants and another entry for products. For example, during atom mapping, the stoichiometric coefficients are estimated for each reaction, necessitating the calculation of moles for each chemical species.

2.2 Atom mapping chemical reactions

Atom mapping, or atom-to-atom mapping (AAM), is a computational chemistry technique that tracks the movement of atoms from the reactants to the products in a chemical reaction.[28] In AAM, each atom on the reactant side of the reaction is assigned a unique identifier, which is then transferred to the corresponding atoms on the product side[29] – as visualized in Fig. 2. This process provides a detailed pathway showing how each atom in the reactants transforms into atoms in the products, crucial for understanding reaction mechanisms and optimizing chemical processes.[30] Atom mapping finds wide

applications across various fields, serving as a fundamental tool for advancing areas such as drug design and other chemical studies.[31] It supports the automated identification of reaction centers and the extraction of reaction templates from databases, essential for predicting reaction outcomes[32] and training machine learning models used in single-step retrosynthesis.[33]

Determining the BCC of a molecule requires tracing each carbon atom back to its various source materials, distinguishing between fossil-based and renewable sources. This tracing involves retracing the pathway of each carbon atom from reactants through various chemical transformations to the final product in the value chain. The value chain itself is derived from industrial ERP data, which define material inputs and outputs but do not specify the underlying reaction mechanisms. Therefore, comprehensive atom-to-atom mapping for each chemical transformation within the value chain is essential. This methodology enables the precise tracking of carbon atoms from inlet materials to final products, ensuring an accurate assessment of BCC.

Several commercially available AAM tools could be used to automate the atom mapping of value chain reactions. However, based on comprehensive benchmarking against popular AAM tools including ChemAxon Automapper, Indigo, RDTool, NameRXN, and RXNMapper, the RXNMapper tool distinguishes itself with an efficient unsupervised-learning transformer model approach.[22] It achieved the highest accuracy of the AAM tools and was also the second fastest algorithm – an important factor if such a model were to be deployed at the scale of an entire industrial value chain.[29]

2.2.1 RXNMapper from IBM. RXNMapper is a recent, open-source machine learning (ML) model designed for automatic atom mapping of chemical reactions.[22] Trained using a self-supervised natural language processing (NLP) approach known as masked language modeling,[34] RXNMapper learns to predict obscured atoms in reaction SMILES strings, effectively capturing the grammar and complex patterns of chemical reactions.

RXNMapper was selected due to its demonstrated capability to handle intricate reaction details, including stereochemistry and unbalanced reactions, essential for accurately mapping diverse chemical transformations relevant to this study. Benchmark studies report that RXNMapper achieves high accuracy, correctly mapping 99.4% of a test set comprising 49 000 unbalanced patent reactions sourced from USPTO. Furthermore,

it exhibits superior performance compared to other atom mapping tools such as Indigo[35] and Mappet,[36] providing the fastest inference times at 7.7 ms per reaction using a GPU.[22]

A detailed overview of RXNMapper, including its architecture based on transformer neural networks, training methodology, and performance evaluation, is provided in the SI.

2.2.2 Atom mapping workflow: TDI production node. Fig. 5 illustrates the atom mapping workflow, which transforms ERP data into atom-mapped molecular structures through a three-step process. The TDI production node serves as an example to demonstrate this workflow.

Step 1: ERP data preprocessing

The workflow begins by preprocessing ERP data to identify chemical species and convert them into molecular structures. This process generates a bill of substances for the node (Table 4), which lists all relevant species as canonical SMILES strings. These molecular structures form the foundation for subsequent analysis.

Step 2: Reaction SMILES construction

In the second stage, a reaction SMILES is constructed – a linear string notation that encodes chemical transformations from input substances to output products. The standard reaction SMILES syntax follows the format: [reactants] > [reagents] > [products].

This work adopts a simplified approach: all input substances are included in the reactant section, while the reagent section remains empty. This simplification streamlines the parsing process without affecting the results, as RXNMapper only annotates atoms in reactants and products. This generic structure enables efficient computational analysis, facilitating tasks such as reaction prediction, optimization, and data mining.[37]

Since RXNMapper requires a reaction SMILES with only one product substance, multiple reaction SMILES strings must be constructed for nodes with multiple products. For a production node with j reactant substances and k products, k reaction SMILES strings are generated–each containing the same set of reactants but differing in the product species, as shown in eqn (4). The stoichiometric coefficients are estimated using the mole quantity of each substance (n_s), calculated using eqn (5):

$$\frac{n_1}{n_p}\text{reactant}_1 + \frac{n_2}{n_p}\text{reactant}_2 + \cdots + \frac{n_j}{n_p}\text{reactant}_j \rightarrow \text{product} \quad (4)$$
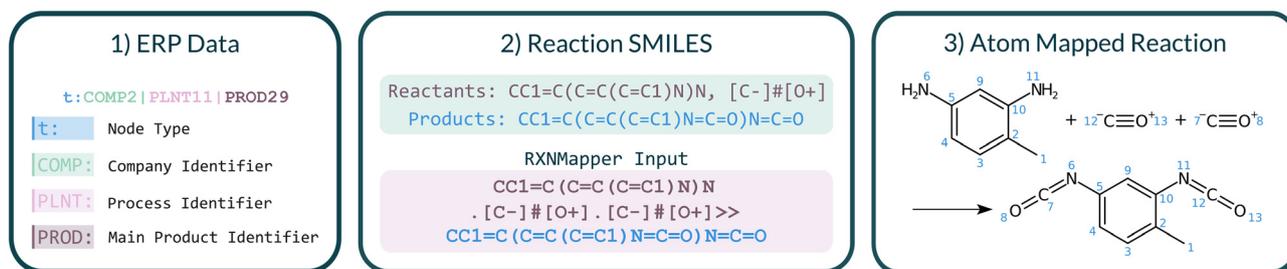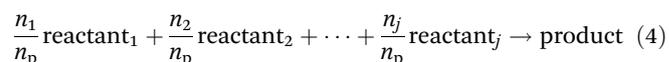

Fig. 5 Atom mapping workflow: convert ERP data to molecular structures, construct reaction SMILES, apply RXNMapper to generate atom-mapped SMILES, and then convert it to a bill of atoms format.

$$n_{\mathrm{s}} = \frac{1}{M_{\mathrm{s}}} \sum_{\mathrm{p}} \alpha_{\mathrm{p}} \cdot \lambda_{\mathrm{ps}} \tag{5}$$

where:

• λ_{ps} is the mass ratio of substance s in product p.

• α_{p} is the input ratio (kg of material from duplet consumed per kg of main output at connected triplet).

• M_{s} is the molar mass of substance s.

Step 3: Atom mapping and generation of a bill of atoms

The constructed reaction SMILES is passed to the RXNMapper model, which returns an atom-mapped reaction SMILES. Fig. 2 displays both the unmapped reaction SMILES and visualization of the mapped reaction output for the TDI production node. For enhanced interpretability, the mapped reactions can be visualized using CDK Depict[38] or RDKit,[39] as shown in the third stage of Fig. 5. To calculate the BCC, the atom mapping must be translated into a "bill of atoms" format.[23] The atom mapping directly yields the substance-level atom bill, denoted $\phi_{\mathrm{s'se}}$, which represents the share of atoms of chemical element e in output substance s that originated from input substance s'. Table 5 presents the complete bill of atoms for the TDI production node. Note that for this framework, the bill of atoms is only required for carbon-containing materials.

2.3 Model construction and optimization

With the atom mapping procedure complete, the necessary data are available to calculate the BCC for a given production node. However, to apply this approach to a value chain involving multiple interconnected mix and production nodes as per the third project objective, a system of equations must be formulated. Solving this system will provide the BCC for every material and substance throughout the value chain.

This section will first present a detailed example of calculating the BCC for a single production node. Subsequently, the value chain system will be defined, and a suitable method for solving the system will be selected and discussed. In addition to the graph notation introduced in Table 2, Table 1 defines indices required for the methodology.

2.3.1 BCC calculation for TDI production node. From the atom mapping section of the workflow, the bill of atoms at a

substance level, $\phi_{\mathrm{s'se}}$, is calculated, which denotes the share of atoms of chemical element e in outlet substance s that originate from the inlet substance s'. However, the atom bill is required on a material level, $\psi_{\mathrm{p's'pse}}$ for instances where the materials are not pure, *i.e.*, contain more than one substance. The material-level atom bill denotes the share of chemical element e in outlet substance s, in outlet material p that originates from the inlet substance s' within inlet material p'. Eqn (6) shows how the material-level atom bill can be calculated given the substance-level atom bill, where $\lambda_{\mathrm{p's'}}$ represents the mass fraction of inlet substance s' in inlet material p'.

$$\psi_{\mathrm{p'\,s'\,pse}} = \frac{\alpha_{\mathrm{p'}} \cdot \lambda_{\mathrm{p'\,s'}}}{\sum_{\mathrm{p'}} \alpha_{\mathrm{p'}} \cdot \lambda_{\mathrm{p'\,s'}}} \phi_{\mathrm{s'\,se}} \tag{6}$$

With the material-level atom bill in place, two further equations are required to fully define a system to determine the BCC of a value chain system. These equations focus on calculating the elemental attribute share β. Eqn (7) is specific to calculating the share of attribute a (*e.g.*, fossil, biogenic, *etc.*) of chemical element e in substance s in material p within the production node (c, b, g). It does so by summing the attribute contributions from each incoming mix node denoted (c', p'), weighted by the material-level atom bill $\psi_{\mathrm{p's'pse}}$.

$$\beta_{\mathrm{cbgpsea}} = \sum_{\mathrm{p'\,s'}} \psi_{\mathrm{p'\,s'\,pse}} \beta_{\mathrm{cp'\,s'\,ea}} \tag{7}$$

Eqn (8) calculates the attribute share a of element e in substance s in material p within the mix node (c, p). It does so by summing the attribute contributions from each incoming pro-
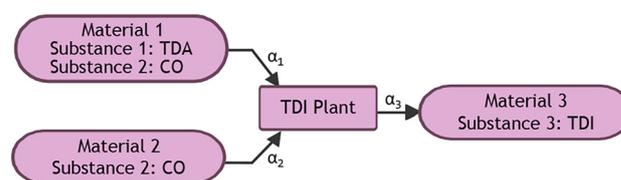

Fig. 6 Example of a TDI production node with impure inlet materials, corresponding to triplet *t* = (c, b, g).

Table 5 Bill of atoms for the TDI production node: t:COMP2|PLNT11|PROD29

Reactant material	Reactant SMILES	Product material	Product SMILES	Element	Atom count	Atom share
PROD10	[C–]#[O+]	PROD36	O=C=O	O	1	0.50
PROD10	[C–]#[O+]	PROD36	O=C=O	C	1	1.00
PROD31	Cc1ccc(N)cc1N	PROD29	Cc1ccc(N=C=O)cc1N=C=O	C	7	0.78
PROD31	Cc1ccc(N)cc1N	PROD29	Cc1ccc(N=C=O)cc1N=C=O	H	6	1.00
PROD31	Cc1ccc(N)cc1N	PROD29	Cc1ccc(N=C=O)cc1N=C=O	N	2	1.00
PROD10	[C–]#[O+]	PROD29	Cc1ccc(N=C=O)cc1N=C=O	O	2	1.00
PROD10	[C–]#[O+]	PROD29	Cc1ccc(N=C=O)cc1N=C=O	C	2	0.22
PROD6	ClCl	PROD36	Cl	Cl	1	1.00
PROD6	ClCl	PROD36	Cl	H	1	1.00
PROD31	Cc1ccc(N)cc1N	PROD36	N#N	N	2	1.00
PROD19	O	PROD36	O=C=O	O	1	0.50
PROD10	[C–]#[O+]	PROD36	[C–]#[O+]	O	1	1.00
PROD10	[C–]#[O+]	PROD36	[C–]#[O+]	O	1	1.00

duction node, denoted (c′, b′, g′), weighted by the consumption mix share (i.e., how much from the mix node comes from each production node), $\mu_{c'b'g'cp}$.

$$\beta_{cpsea} = \sum_{c'\,b'\,g'} \mu_{c'\,b'\,g'\,cp}\beta_{c'\,b'\,g'\,psea}. \qquad (8)$$

2.3.2 Example: one node calculation. This example demonstrates the BCC calculation for a production node $t = (c, b, g)$, as illustrated in Fig. 6, where the main one-product reaction for the formation of TDI is:

$$C_7H_{10}N_2 + 2CO \rightarrow C_9H_6N_2O_2$$

TDA, carbon monoxide (CO), and TDI are represented as substances 1, 2, and 3, respectively. In this case, CO is present in both input materials:

- Material 1: 80% TDA, 20% CO → $\lambda_{11} = 0.8$, $\lambda_{12} = 0.2$
- Material 2: 100% CO → $\lambda_{22} = 1$

Let the corresponding input ratios be:

$$\alpha_1 = \frac{20}{60}, \quad \alpha_2 = \frac{40}{60}$$

The substance-level atom bill, calculated using the atom mapper tool, $\phi_{s'se}$, is:

$$\phi_{13C} = \frac{7}{9}, \quad \phi_{12C} = \frac{2}{9}$$

Since substance 1 (TDA) is only present in material 1, the material-level and substance-level atom bills are equivalent:

$$\psi_{1133C} = \phi_{13C} = \frac{7}{9}$$

However, for CO, which is split across both materials, the material-level atom bills are calculated using eqn (6):

$$\psi_{1233C} = \frac{\alpha_1\lambda_{12}}{\alpha_1\lambda_{12} + \alpha_2\lambda_{22}} \cdot \phi_{12C} = \frac{\frac{20}{60}\cdot 0.2}{\frac{20}{60}\cdot 0.2 + \frac{40}{60}\cdot 1} \cdot \frac{2}{9} = \frac{2}{99}$$

$$\psi_{2233C} = \frac{\alpha_2\lambda_{22}}{\alpha_1\lambda_{12} + \alpha_2\lambda_{22}} \cdot \phi_{12C} = \frac{\frac{40}{60}\cdot 1}{\frac{20}{60}\cdot 0.2 + \frac{40}{60}\cdot 1} \cdot \frac{2}{9} = \frac{20}{99}$$

Assuming material 1 is entirely fossil-derived and material 2 is entirely biogenic, let A be the set of elemental attributes considered (e.g., fossil, biogenic), and in this example, let $x \in A$ denote biogenic carbon:

$$\beta_{c11Cx}, \beta_{c12Cx} = 0, \quad \beta_{c22Cx} = 1$$

The BCC for this TDI node is then computed as:

$$\beta_{c33Cx} = \beta_{c11Cx} \cdot \psi_{1133C} + \beta_{c12Cx} \cdot \psi_{1233C} + \beta_{c22Cx} \cdot \psi_{2233C}$$

$$\beta_{c33Cx} = 0 \cdot \frac{7}{9} + 0 \cdot \frac{2}{99} + 1 \cdot \frac{20}{99} = \frac{20}{99}.$$

2.3.3 Selection of a linear program approach. Section 2.3.2 demonstrated the procedure for calculating the BCC for a single product in a production node. However, in a value chain, the BCC is dependent on the BCC of all preceding nodes. Additionally, due to the presence of recycle streams, it would also be infeasible to sequentially determine the BCC by carrying out the calculation on a one-node-at-a-time basis. Hence, to compute the BCC across the value chain, a linear system of equations is defined, with eqn (7) and (8) at the core of this system. A Linear Program (LP) is chosen to solve the system of equations – a practical choice that enables the problem to be posed as slack minimization. A linear optimization model is characterized by the following criteria: only continuous variables, a single linear objective function, and only linear equality or inequality constraints.

The BCC system can be formulated as a feasibility problem, wherein the objective is not to optimize a particular function, but rather to identify values for the elemental attribute shares of each substance that satisfy a set of constraints. In practice, however, industrial datasets often contain inconsistencies or incomplete information that may render the constraint set infeasible. To accommodate such cases, slack variables are introduced, allowing for controlled violations of the constraints. This enables the model to yield a solution even when exact feasibility is not attainable, while also quantifying the extent of any deviations.

In this formulation, the objective function is defined to minimise the total system slack. This drives the solution towards that of the original feasibility problem under the assumption of fully consistent and accurate data. Moreover, the magnitude and location of slack values provide diagnostic insight by identifying specific constraints where data limitations are most pronounced. As such, the use of slack variables offers both computational robustness and practical interpretability, making the approach particularly valuable in industrial contexts where data uncertainty is common.

The LP formulation was implemented using the Python MIP package,[40] using the CBC (COIN-OR branch and cut) solvers[41] – as it is open-source and suitable for LPs.

2.3.4 Linear program formulation.

$$\min_{\theta} \sum_{c,b,g,p,s,e}\left(z_{cbgpse} - q_{cbgpse}\right) + \sum_{c,p,s,e}\left(z_{cpse} - q_{cpse}\right) \qquad (9a)$$

s.t.

$$\beta_{cbgpsea} = \sum_{p'\,s'} \psi_{p'\,s'\,pse}\beta_{cp'\,s'\,ea}, \qquad \forall c, b, g, p, s, e, a \quad (9b)$$

$$\beta_{cpsea} = \sum_{c'\,b'\,g'} \mu_{c'\,b'\,g'\,cp}\beta_{c'\,b'\,g'\,psea}, \quad \forall c, p, s, e, a \qquad (9c)$$

$$\sum_a \beta_{cbgpsea} - z_{cbgpse} - q_{cbgpse} = 1, \qquad \forall c, b, g, p, s, e \quad (9d)$$

$$\sum_a \beta_{cpsea} - z_{cpse} - q_{cpse} = 1, \quad \forall c, p, s, e \qquad (9e)$$

$$\beta_{cbgpsea} \in [0, 1], \qquad \forall c, b, g, p, s, e, a \quad (9f)$$

$$\beta_{cpsea} \in [0, 1], \qquad \forall c, p, s, e, a \qquad (9g)$$

$$z_{cbgpse} \in R^+, \quad q_{cbgpse} \in R^-, \qquad \forall c, b, g, p, s, e \quad (9h)$$

$$z_{cpse} \in R^+, \quad q_{cpse} \in R^-, \qquad \forall c, p, s, e \qquad (9i)$$

Table 6 Decision variables, slack variables, and parameters

Notation	Description
β_{cbgpsea}	Fraction of elemental attribute a of chemical element e in substance s, material p, at production node (c, b, g)
β_{cpsea}	Fraction of elemental attribute a of chemical element e in substance s, material p, at mix node (c, p)
z_{cbgpse}	Positive slack variable for chemical element e in substance s, material p, at production node (c, b, g)
q_{cbgpse}	Negative slack variable for chemical element e in substance s, material p, at production node (c, b, g)
z_{cpse}	Positive slack variable for chemical element e in substances, material p, at mix node (c, p)
q_{cpse}	Negative slack variable for chemical element e in substances, material p, at mix node (c, p)
$\mu_{\text{c'b'g'cp}}$	Mix node share, *i.e.*, the fraction of a virtual tank (c, p) sourced from a production node (c', b', g')
$\psi_{\text{p's'pse}}$	Bill of atoms, *i.e.*, the fraction of chemical element e in substance s in product p, sourced from substance s' in product p'

N.B. c' denotes the inlet company code, whereas c denotes the outlet company code, and θ is the set of decision variables β_{cbgpsea}, β_{cpsea}, z_{cbgpse}, q_{cbgpse}, z_{cpse}, and q_{cpse} for all company codes c, business processes b, main products g, materials p, substances s, chemical elements e, and elemental attributes a. The formulation assumes that the bill of atoms $\psi_{\text{p's'pse}}$, derived from the atom mapping stage, is internally consistent and correct.

The objective function (9a) minimizes slack across the entire value chain, encouraging efficient use or elimination of slack variables. Table 6 summarizes the decision variables and parameters used in the LP. Slack variables are denoted z (positive) and q (negative), with subscripts indicating context: z_{cpse} and q_{cpse} for duplets, and z_{cbgpse} and q_{cbgpse} for triplets.

Inlet conditions. Solving this formulation is dependent on the inlet nodes entering the value chain subgraph having known values of β_{cpsea}. To express this, we define a set of inlet mix nodes, denoted by D_0. These are nodes in the set of all mix nodes D that have no incoming edges. We formalize this as:

$$D_0 = \{d_i \in D | \delta^-(d_i) = \varnothing\} \tag{10}$$

Here, d_i refers to a particular mix node in D, and $\delta^-(d_i)$ denotes the set of inlet edges into node d_i; if this set is empty, then d_i is an inlet node. For simplicity in formulation, a structural assumption is imposed: the value chain must start and end with mix nodes. Hence, the values of β_{cpsea} at the inlet mix nodes in D_0 must be specified and set to a known constant h:

$$\beta_{\text{c}_{d_i}\text{p}_{d_i}\text{sea}} = h \quad \forall d_i \in D_0, \, >\forall\, > \text{s, e, a} \tag{11}$$

Constraints. • Constraints (9b) and (9c) define elemental attribute share equations for production and mix nodes, respectively.

• Eqn (9d) and (9e) ensure that for any chemical substance, the sum of carbon attributes (including fossil and biogenic) equals one, incorporating slack variables.

• Eqn (9f) and (9g) set bounds on elemental attribute shares for production and mix nodes, ensuring that β values are non-negative and do not exceed one.

• Eqn (9h) define non-negative bounds for positive and negative slack variables in production nodes. Eqn (9i) establishes analogous bounds for mix nodes.

3. Results

Building on the methodology presented in Section 2, we apply CarAT to an industrial TDI value chain and, to demonstrate recycle-stream handling, a butanediol value chain. These examples were selected for their industrial relevance and contrasting topologies. Three scenarios are analyzed:

• **Base case – SI**: A TDI value chain with entirely fossil-derived carbon input. TDI product BCC = 0%.

• **Case 1 – section 3.1**: A TDI value chain with 100% biogenic natural-gas feed. TDI product BCC = 22%.

• **Case 2 – section 3.2**: A BDO value chain featuring a recycle stream; 75% biogenic acetylene and 50% biogenic BDO inlet. Butanediol product BCC = 38%.

The three-stage CarAT methodology is fully implemented in a Python package and demonstrated here through three worked scenarios. For each case, the linear program results are visualized as a Sankey diagram, which clearly conveys the carbon attribute flows and the bipartite graph structure of the value chain. The edge widths are illustrative and not scaled to mass flow, as proportional weighting was found to reduce interpretability. Importantly, the objective function – *i.e.*, the total system slack – for all three scenarios is numerical zero, indicating successful LP convergence. The complete analysis, including value chain data, solver setup, and visualization scripts, is available in a public GitHub repository,† ensuring full reproducibility and enabling others to apply CarAT to new chemical systems. The base case scenario, being structurally simple and yielding a zero BCC by design, is discussed in the SI (Section 6).

To better visualize the flow of elemental attributes, particularly the flow of biogenic carbon, a color-coding scheme was implemented. In the diagram, dark blue bars represent mix nodes and light blue bars represent production nodes. Pale yellow links represent the flow of non-carbon-containing compounds, which are still included to provide a full picture of the value chain. To differentiate between carbon and non-carbon-containing compounds, a link is shown for each substance (SMILES) transferred between the two nodes. The thickness of the links entering the blue mix nodes is proportional to the consumption mix share of that substance $\mu_{\text{c'b'gcp}}$. Similarly, the thickness of the links entering the red production nodes is

---

† https://github.com/EmPajak21/CarAT.

proportional to the input ratio of that substance at that node (c, b, g).

Importantly, fossil carbon is shown in gray, and green links represent biogenic carbon, with darker green signifying a larger share of biogenic carbon and paler green representing a smaller share. As expected, the carbon in the resulting TDI product stream from this value chain is 100% fossil-based.
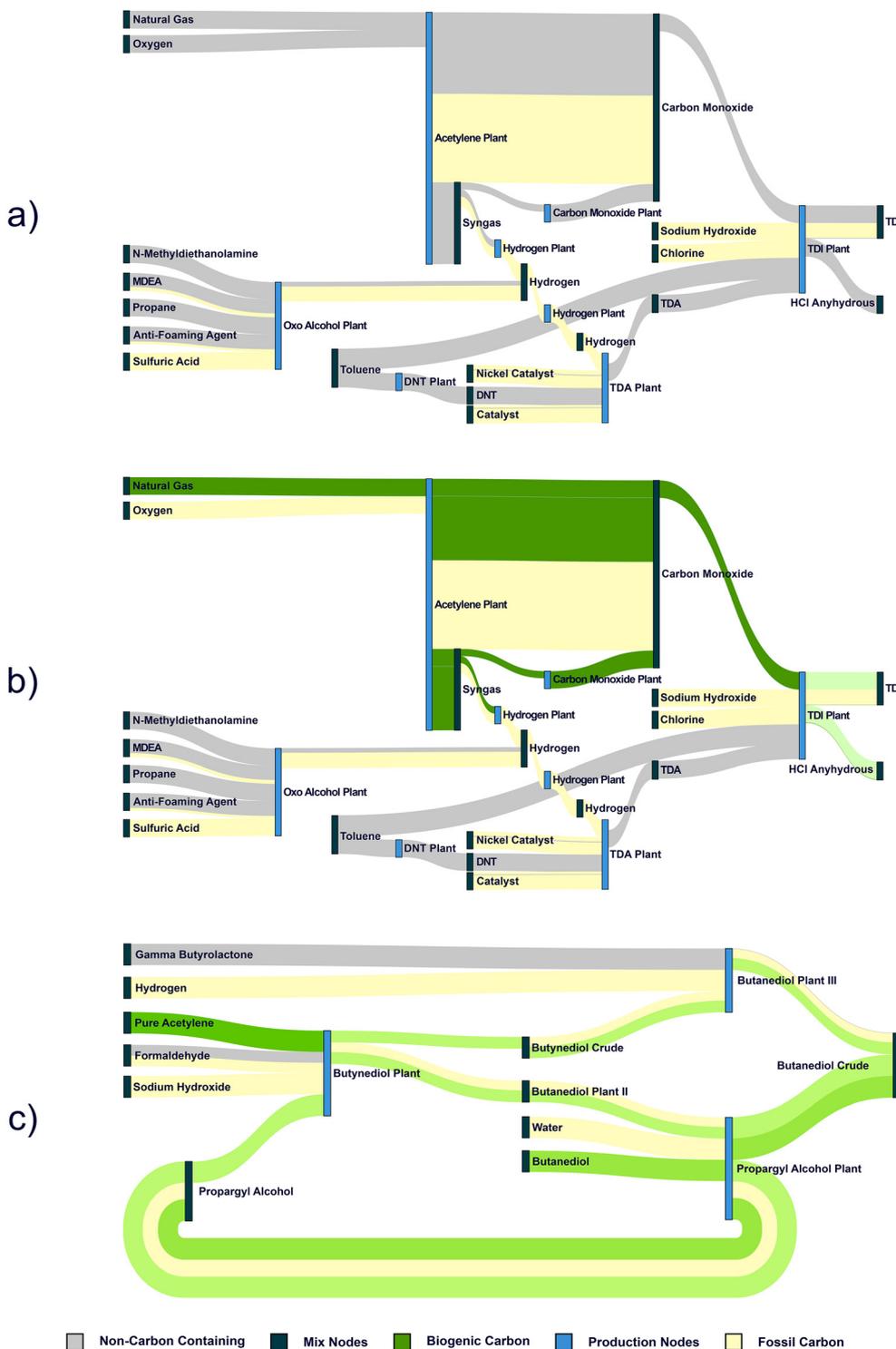
Fig. 7 Sankey diagrams for (a) Base Case, (b) Case 1, and (c) Case 2. Node colors: dark blue = mix and light blue = production. Link colors: gray = fossil carbon, green gradient = biogenic carbon (darker green indicates a higher biogenic carbon share), and yellow = non-carbon streams. Note: edge widths are illustrative and not proportional to mass flow. Overall BCCs: 0%, 22%, and 38%, respectively.

3.1 Case 1: biogenic methane feedstock

Case 1 (Fig. 7b) features a 100% biogenic natural gas feedstock, shown in dark green. The biogenic methane and ethane enter a production node to form carbon monoxide. The resulting carbon monoxide enters the TDI production node, where it reacts with chlorine to form phosgene, which subsequently reacts with TDA to yield TDI, alongside the byproduct anhydrous hydrochloric acid.

Since TDI consists of nine carbon atoms – two sourced from the 100% biogenic carbon monoxide and seven from the fossil-derived TDA – the resulting TDI has a BCC of 22%, indicated by a lighter shade of green. This scenario demonstrates the capability of calculating the BCC of a product when using mixed sources of carbon feedstock in its synthesis.

3.2 Case 2: butanediol value chain with a recycle stream

In Case 2 (Fig. 7c), based on the butanediol value chain, all carbon is fossil-derived except for a 75% biogenic pure acetylene feedstock and a 50% biogenic butanediol stream. The 75% biogenic acetylene reacts with two equivalents of fossil-derived formaldehyde, producing butynediol with a BCC of 38%. Following hydrogenation, the resulting butanediol retains this BCC of 38%. In the recycle stream, two carbon-containing substances are present: 50% biogenic butanediol and propargyl alcohol, a byproduct of butynediol synthesis.

The recycle stream adds complexity and reflects the reality of industrial value chains where such flows are common. The framework correctly apportions biogenic carbon through the recycle path, demonstrating its robustness to cyclic topologies.

4. Discussion and limitations

The results above confirm that CarAT quantifies the BCC of value chains ranging from a simple linear TDI route to a cyclic BDO system. Two practical constraints now merit discussion: (i) atom-mapping accuracy and (ii) data completeness.

4.1 Atom-mapping accuracy

RXNMapper delivers a high overall accuracy,[29,42] yet occasional mis-mappings are inevitable. We envisage a two-tier mitigation strategy:

1. Domain expert supersession of automated mapping. In industrial deployment, chemists can flag and correct mismapped reactions during data onboarding, ensuring that critical pathways are traced correctly. Having some form of confidence in atom mapping – such as that provided by the more recent LocalMapper[30] – could help guide targeted spot checks.

2. Reaction-class-specific fine-tuning. If systematic errors are observed for a given reaction class, additional training on that subset can be leveraged to improve accuracy – see Section 5.1.

RXNMapper's default token limit (512) was sufficient for the value chains analyzed in this work. However, as reactions become more complex downstream in an industrial value chain, the impact of impurities, solvents, and other non-reacting compounds could make the size limitation more significant. To address this, initial efforts could involve using RXNMapper without considering stoichiometry, as the model has demonstrated good accuracy even with unbalanced reactions.[43] Furthermore, careful curation of the reaction SMILES might be beneficial, such as removing non-essential compounds while preserving the core chemical reaction and essential reactants.

4.2 Data completeness

A more fundamental bottleneck is missing or non-standardized molecular identifiers. CarAT requires a canonical SMILES for every carbon-containing species; where none exists, the framework is unable to trace atoms. Conversion from CAS numbers could alleviate part of this gap, but a persistent lack of digital representations for proprietary intermediates remains a barrier to full-chain BCC accounting. Expanding internal chemical inventories or adopting open-identifier policies will therefore be as important as future algorithmic improvements.

5. Conclusion

This research addresses an urgent challenge facing the chemical industry: the need for a scalable, generalizable, and dynamic methodology to calculate the BCC of chemical products. As the TfS consortium prepares to mandate BCC reporting by 2026, companies must be able to track the origin of carbon atoms across complex and evolving value chains. Existing methods are static and product-specific, limiting their use in an industry characterized by changing feedstocks, recycling, and distributed manufacturing.

To meet this challenge, we have developed CarAT, a framework that integrates enterprise-level data with a chemistry language model and linear optimization to automate carbon tracing at scale. By mapping carbon atoms from feedstocks through each stage of production and solving for BCC *via* a linear program, CarAT decouples data from methodology. This structure allows manufacturers to obtain methodological certification, enabling automatic recalculations as operational parameters evolve.

Validation was carried out across three scenarios that increased in complexity: a fossil-only TDI value chain, the same chain with a 100% biogenic natural gas inlet, and a butanediol value chain incorporating a recycle stream and partial biogenic inputs. These scenarios tested the framework's capacity to adapt to key industrial features, including mixed feedstocks and recycling loops, with results confirming both robustness and scalability. Having demonstrated its robustness through these case studies, CarAT is now being implemented within BASF to proactively trace the carbon attributes of its product portfolio. This industrial adoption affirms CarAT's potential for large-scale deployment and highlights the chemical sector's pressing demand for scalable carbon-tracing solutions.

Further to supporting compliance with sustainability reporting requirements, CarAT enables informed decision-making for decarbonization. Tracing the biogenic carbon content and quantifying the impact of raw material composition on the product-level BCC facilitate the substitution of fossil-based carbon with biogenic alternatives. This, in turn, guides internal decisions and enhances value chain transparency, enabling Scope 3 emission reductions through more informed upstream raw material choices.

The transition to a low-carbon chemical industry will not occur through a single, definitive shift between fossil and biogenic feedstocks, but rather through a gradual and heterogeneous evolution in carbon sourcing. In practice, manufacturers will have to evaluate feedstocks and process routes with differing PCFs, considering factors such as availability, regional infrastructure, energy intensity, and overall life-cycle performance. CarAT provides a consistent and auditable basis for calculating BCC and tracing carbon flows across all such scenarios, including hybrid pathways where feedstocks of different origins and carbon intensities are combined within complex production networks. Importantly, these results are intended to complement process-level assessments: the atom-level tracing enabled by CarAT can be coupled with techno-economic or life-cycle analysis frameworks to evaluate trade-offs between the carbon origin, energy use, and yield. In particular, integration with established LCA platforms such as OpenLCA or Brightway would allow CarAT-calculated BCC to be imported and directly incorporated into cradle-to-grave assessments. This coupling would create a seamless link between molecular-level carbon tracing and full life-cycle environmental evaluation. By providing molecular-level transparency across value chains, CarAT supports chemical manufacturers in making informed, data-driven decisions toward lower-carbon production systems.

By clearly aligning with key Green Chemistry principles, particularly around substituting fossil feedstocks with biogenic or recycled sources, CarAT also offers a concrete step toward net-zero goals. Its capacity for rapid recalculation facilitates real-time adjustments in sourcing and operational strategies, making sustainable innovation both more transparent and more feasible across the chemical industry.

5.1 Future work

Future work will proceed along three key avenues. First, improving the accuracy of atom mapping is essential for broader adoption. As the atom-mapping stage constitutes the main source of uncertainty within the framework, industrial deployment will incorporate domain-expert validation to review and correct mappings in key reactions, ensuring reliability in practice. Should such analyses reveal systematic issues across particular reaction classes, targeted datasets can then be curated to enable fine-tuning of the model and improve accuracy for those transformations. In parallel, extending the context window size will be explored to accommodate a longer or more complex reaction SMILES that exceeds the current model limit.

Second, the linear program underlying CarAT is generalizable beyond carbon. With suitable data, the same methodology could be extended to trace other elemental attributes such as nitrogen, recycled content, or toxic elements. This would enable more comprehensive sustainability assessments across chemical value chains.

Third, future work will explore the inverse optimization problem: how to allocate biogenic raw materials across a value chain to achieve a desired biogenic carbon content in the final product. This has clear relevance for setting and meeting emission targets across value chains.

Taken together, these developments will support the evolution of CarAT into a more robust and flexible tool for sustainability analysis within value chains, with the potential to assist industry in its transition towards net zero.

Author contributions

Emma Pajak: conceptualization, methodology, writing – original draft, software, and visualization. David Walz: conceptualization, data curation, resources, and writing – review & editing. Olga Walz: conceptualization, data curation, resources, and writing – review & editing. Laura Marie Helleckes: methodology, supervision, software, and writing – review & editing. Klaus Hellgardt: conceptualization, supervision, writing – review & editing, and project administration. Antonio del Rio Chanona: conceptualization, supervision, writing – review & editing, and project administration.

Conflicts of interest

There are no conflicts to declare.

Data availability

CarAT, the tool developed in this work, is available as open-source software on GitHub; an archived release is deposited on Zenodo (https://doi.org/10.5281/zenodo.16851777). The repository includes the complete case study used in this paper.

Supplementary information (SI) is available. The SI includes a schematic of the full TDI value chain case study used in this work, along with furhter details on the choice of chemistry language model (RXNMapper) and the base case Sankey diagram (all fossil carbon inlet streams). See DOI: https://doi.org/10.1039/d5gc04348d.

Acknowledgements

References

1 F. Lechtenberg, R. Istrate, V. Tulus, A. Espuña, M. Graells and G. Guillén-Gosálbez, *J. Ind. Ecol.*, 2024, **28**, 1449–1463.

2 L.-P. Merkouri, L. F. Bobadilla, J. L. Martín-Espejo, J. A. Odriozola, A. Penkova, G. Torres-Sempere, M. Short, T. R. Reina and M. S. Duyar, *Appl. Catal., B*, 2025, **361**, 124610.

3 Y. Tian, V. Meduri and E. N. Pistikopoulos, *Comput. Chem. Eng.*, 2022, **160**, 107679.

4 J. García-Serna, L. Pérez-Barrigón and M. J. Cocero, *Chem. Eng. J.*, 2007, **133**, 7–30.

5 M. Baldea, E. E. Endler, E. Hale, C. T. Maravelias, M. Barolo, I. Harjunkoski, M. Mercangoz, S. L. Shah, M. Soroush, B. R. Young and Q. Zhang, *Ind. Eng. Chem. Res.*, 2025, **64**, 16466–16478.

6 TfS_PCF_guidelines_2024_EN_pages-low.pdf, https://www.tfs-initiative.com/app/uploads/2024/03/TfS_PCF_guidelines_2024_EN_pages-low.pdf.

7 M. E. Porter, *Competitive Advantage: Creating and Sustaining Superior Performance*, Free Press, 1985.

8 M. Holweg and P. Helo, *Int. J. Prod. Econ.*, 2014, **147**, 230–238.

9 R. Blackburn, J. Kallrath and S. T. Klosterhalfen, *Int. Trans. Oper. Res.*, 2015, **22**, 385–405.

10 BASF, *Ludwigshafen Site 2022 in Figures*, 2023, https://www.basf.com/global/documents/en/news-and-media/publications/reports/2023/BASF_Ludwigshafen_site_2022_in_figures.pdf.assetinline.pdf.

11 *Product Carbon Footprint: Customer Information*, https://chemicals.basf.com/global/en/Intermediates/sustainability/product-carbon-footprint-customer-information.html.

12 Pathfinder-Framework-Version-2.0.pdf, https://www.wbcsd.org/wp-content/uploads/2023/01/Pathfinder-Framework-V\%ersion-2.0.pdf.

13 Z. M. Harris, S. Milner and G. Taylor, *Greenhouse Gas Balances of Bioenergy Systems*, Academic Press, 2018, pp. 55–76.

14 H. Paulo, M. Vieira, B. S. Gonçalves, T. Pinto-Varela and A. P. Barbosa-Póvoa, *Industrial Engineering and Operations Management*, Cham, 2023, pp. 29–40.

15 P. Gabrielli, L. Rosa, M. Gazzani, R. Meys, A. Bardow, M. Mazzotti and G. Sansavini, *One Earth*, 2023, **6**, 682–704.

16 3con Management Consultants GmbH, *Sustainability Report* 2023, 3con management consultants gmbh sustainability report, 2024.

17 L. Stellner, A. A. Kalousdian, R. Meys and A. Kätelhön, *Carbon Footprints and LCA Data: Embracing New Trends with Enhanced Data for Chemical Value Chains*, 2024, https://www.carbon-minds.com/, White Paper.

18 J. Hohlweck, *SiGREEN – Dynamic Product Carbon Footprints*, Siemens ag technical white paper, 2023.

19 H. Ohno, H. Sato and Y. Fukushima, *Environ. Sci. Technol.*, 2018, **52**, 3899–3907.

20 A. Kätelhön, R. Meys, S. Deutz, S. Suh and A. Bardow, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 11187–11194.

21 K. Beer, M. Böcher, C. Ganzer, A. Blöbaum, L. Engel, T. De Paula Sieverding, K. Sundmacher and E. Matthies, *Forest Policy Econ.*, 2025, **177**, 103521.

22 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.

23 D. Walz, System and Method for Generating a Production Parameter Based on Chemical Compositions, (Patent Application) PCT/EP2025/054843, 2025.

24 A. Kalinichenko, V. Havrysh and V. Perebyynis, *Ecol. Chem. Eng. S*, 2016, **23**, 387–400.

25 A. S. Asratian, T. M. J. Denley and R. Häggkvist, *Bipartite Graphs and their Applications*, Cambridge University Press, 1998.

26 J. Baldyga, E. Molga, S. Szarlik, W. Wójcik, P. Machniewski, L. Rudniak, S. Piechota, A. Slawatycki, W. Chrupala, J. Lachmajer, L. Ruczynski, L. Wójcik and J. Stuczynski, A method of producing toluene diisocyanate (TDI) in the process of the toluene diamine (TDA) phosgenation reaction in the gaseous phase, *European Patent* EP2463272A1, 2012, https://patentimages.storage.googleapis.com/dc/6e/73/0468f41759f5be/EP2463272A1.pdf.

27 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

28 W. L. Chen, D. Z. Chen and K. T. Taylor, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2013, **3**, 560–593.

29 A. Lin, N. Dyubankova, T. I. Madzhidov, R. I. Nugmanov, J. Verhoeven, T. R. Gimadiev, V. A. Afonina, Z. Ibragimova, A. Rakhimbekova, P. Sidorov, A. Gedich, R. Suleymanov, R. Mukhametgaleev, J. Wegner, H. Ceulemans and A. Varnek, *Mol. Inf.*, 2022, **41**, 2100138.

30 S. Chen, S. An, R. Babazade and Y. Jung, *Nat. Commun.*, 2024, **15**, 2250.

31 E. E. Litsa, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett and L. E. Kavraki, *J. Chem. Inf. Model.*, 2019, **59**, 1121–1135.

32 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.

33 M. H. S. Segler and M. P. Waller, *Chem. – Eur. J.*, 2017, **23**, 5966–5971.

34 H. Wang, J. Li, H. Wu, E. Hovy and Y. Sun, *Engineering*, 2023, **25**, 51–65.

35 Indigo Toolkit, https://lifescience.opensource.epam.com/indigo/#documentation.

36 W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.

37 Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.

38 CDK Depict, https://www.simolecule.com/cdkdepict/depict.html.

39 G. Landrum, *RDKit: Open-source cheminformatics*, 2006, https://www.rdkit.org, Accessed: 2025-06-18.

40 Python-MIP, https://www.python-mip.com/.

41 J. Forrest, T. Ralphs, S. Vigerske, H. G. Santos, J. Forrest, L. Hafer, B. Kristjansson, Jpfasano, E. Straver, M. Lubin, J. Willem, R. Jpgoncal1, S. Brito, H.-I. Gassmann, C. M. Saltzman, Tosttost, B. Pitrus and F. Matsushima and to st, coin-or/Cbc: Release releases/2.10.11, 2023, https://zenodo.org/records/10041724.

42 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Unsupervised Attention-Guided Atom-Mapping*, 2020, https://chemrxiv.org/engage/chemrxiv/article-details/60c74b2aee301c3c2cc79dac.

43 P. Schwaller, PhD thesis, Universität Bern, 2021.