6 Supplementary Information

TDI Value Chain Case Study

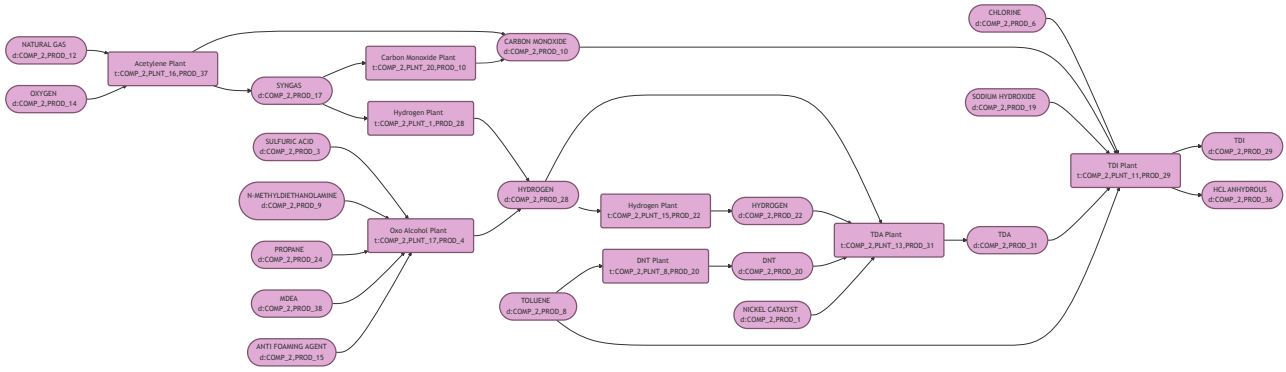

Fig. 8 Bipartite directed graph of the toluene diisocyanate value chain case study, comprising 27 nodes

Overview of RXNMapper

Natural Language Processing (NLP) is a subfield of machine learning that enables computers to understand, interpret, and generate human language. NLP includes tasks such as text analysis, language translation, and question answering[44]. Within chemistry, NLP is facilitated by SMILES notation, a text-based representation that encodes chemical structures into strings, thus enabling computational applications such as property prediction and chemical compound generation[45].

RXNMapper employs a self-supervised NLP technique called masked language modeling, trained on 2.8 million chemical reactions[34]. In this approach, certain atoms within reaction SMILES strings are obscured, prompting the model to predict the missing atoms based on contextual clues from surrounding atoms. This method allows RXNMapper to implicitly learn chemical grammar and complex reaction patterns from the data itself.

Transformer Neural Networks

Transformer neural networks, introduced in the landmark paper "Attention Is All You Need"[46], have emerged as a state-of-the-art technique in NLP. Transformers differ significantly from traditional Recurrent Neural Networks (RNNs) through their use of self-attention, enabling simultaneous processing of input sequences and effectively handling long-range dependencies.

RXNMapper utilizes the ALBERT (A Lite BERT) architecture, a variant of the widely used BERT model known for bidirectional context processing. ALBERT shares weights across layers during training, resulting in a smaller model size that retains consistent functionality across different layers and inputs[43,47]. This capability is particularly valuable for accurately modeling complex chemical reactions.

Performance assessments indicate RXNMapper's high accuracy, achieving correct atom mapping in 99.4% of tested reactions, including diverse reaction types such as Diels-Alder reactions, methylene transfers, and epoxidations[22]. However, the model occasionally demonstrates inaccuracies, particularly regarding atom ordering within rings, azide compounds, and the mapping of oxygen atoms in reductions or Mitsunobu reactions.

Base Case: Entirely Fossil Feedstock

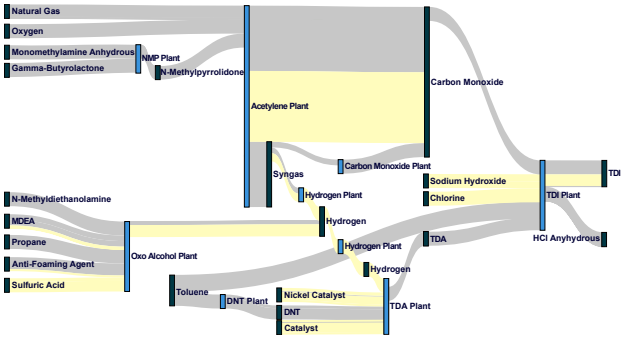

Fig. 9 Sankey diagram visualizing results of base case linear program solution. All carbon is fossil-sourced

In the Base Case, all carbon inputs are fossil-derived, giving a product BCC of 0%. The Sankey diagram confirms that carbon flows track exclusively through gray links; no biogenic carbon enters the chain.

The base case scenario with the TDI value chain serves as the first proof of concept for the framework. While developing the linear program optimization for the final stage of the framework, a step-by-step approach was taken to ensure the validity of the method. The first use of the linear program was at the scale of one node, allowing the results to be verified by hand to ensure that the optimization program and implementation were working as expected.

After confirming this, the next step was to trial a control study, or base case, in which all carbon is fossil-derived. The purpose of this was to check that the linear program was well-posed and that the results were as expected. This also provided an opportunity to consider how best to represent the results, as the immediate output of the linear program is often unintuitive.