Reaction Chemistry
& Engineering

Accelerating Manufacturing for Biomass Conversion via Integrated Process and Bench Digitalization: A Perspective

Journal:	*Reaction Chemistry & Engineering*
Manuscript ID	RE-PER-12-2021-000560.R1
Article Type:	Perspective
Date Submitted by the Author:	22-Jan-2022
Complete List of Authors:	Batchu, Sai Praneet; Univ. of Delaware Hernandez Blazquez, Borja; Univ. of Delaware Malhotra, Abhinav; Univ. of Delaware Fang, Hui; Univ. of Delaware Ierapetritou, Marianthi; University of Delaware, Vlachos, Dion; Univ. of Delaware,

Accelerating Manufacturing for Biomass Conversion via Integrated Process and Bench Digitalization: A Perspective

Sai Praneet Batchu†[a], Borja Hernandez†[a], Abhinav Malhotra†[a], Hui Fang[b], Marianthi Ierapetritou*[a,c], Dionisios G. Vlachos*[a,c]

a) Department of Chemical and Biomolecular Engineering, University of Delaware, 150 Academy St., Newark, DE 19711
b) Department of Electrical and Computer Engineering, University of Delaware, 201B Evans Hall, Newark, DE 19716
c) Catalysis Center for Energy Innovation, RAPID Manufacturing Institute, and Delaware Energy Institute, 221 Academy St., Newark, DE 19716

†Equal contribution, *Corresponding authors: mgi@udel.edu, vlachos@udel.edu

Abstract

We present a perspective for accelerating biomass manufacturing via digitalization. We summarize the challenges for manufacturing and identify areas where digitalization can help. A profound potential in using lignocellulosic biomass and renewable feedstocks, in general, is to produce new molecules and products with unmatched properties that have no analog in traditional refineries. Discovering such performance-advantaged molecules and the paths and processes to make them rapidly and systematically can transform manufacturing practices. We discuss retrosynthetic approaches, text mining, natural language processing, and modern machine learning methods to enable digitalization. Laboratory and multiscale computations automation via active learning are crucial to complement existing literature and expedite discovery and valuable data collection without a human in the loop. Such data can help process simulation and optimization select the most promising processes and molecules according to economic, environmental, and societal metrics. We propose the close integration between bench and process scale models and data to exploit the low dimensionality of the data and transform the manufacturing for renewable feedstocks.

Keywords: Renewables, data science, natural language processing, machine learning, optimization

I. Introduction

Due to global warming,[1] there is mounting societal, economic, and political pressure for sustainable resources and manufacturing.[2, 3] Conversion of lignocellulosic biomass, energy crops, and waste streams, such as agricultural waste and food waste, to materials and chemicals to replace fossil-fuel-based products can address resource sustainability. Specifically, CO_2 capture from the atmosphere via photosynthesis is relatively slow but is scalable. It can lead to a carbon-neutral scenario where CO_2 produced, say from burning fuels, is captured photosynthetically in growing biomass and is eventually converted back into fuels to repeat the entire cycle. Producing solid materials, e.g., renewable terephthalic acid for polyethylene terephthalate (PET) plastic, and chemicals from biomass can lead to carbon capture for extended periods and thus to a negative carbon balance. Building upon this rationale, based on recent roadmap reports put forth by International Renewable Energy Agency (IRENA) in 2018[4] and 2019[5], by 2050, biomass can contribute to ~20% of energy requirements for transportation and total industrial sector (cement, iron and steel, chemical and petrochemical, paper and pulp, aluminium). With the forthcoming electrification of transportation of cars and light trucks, the renewable energy focus should shift to producing jet and marine fuels.[6]

The global catalysis community has taken charge of the molecule and product discovery and catalyst mating with numerous exciting discoveries to platform chemicals and end products (see Figure 1).[7, 8] Platform molecules are building blocks for making many products. Notable examples include biomass-based vinylfuran[9] and furan dicarboxylic acid,[10] instead of fossil-fuel-derived styrene and terephthalic acid[11], respectively. Furanics have been the most exploited platform molecules due to their reactivity and versatility in producing entire product slates, including jet fuels, detergents, lubricants, aromatics, dienes, diols, and diacids, to mention a few.

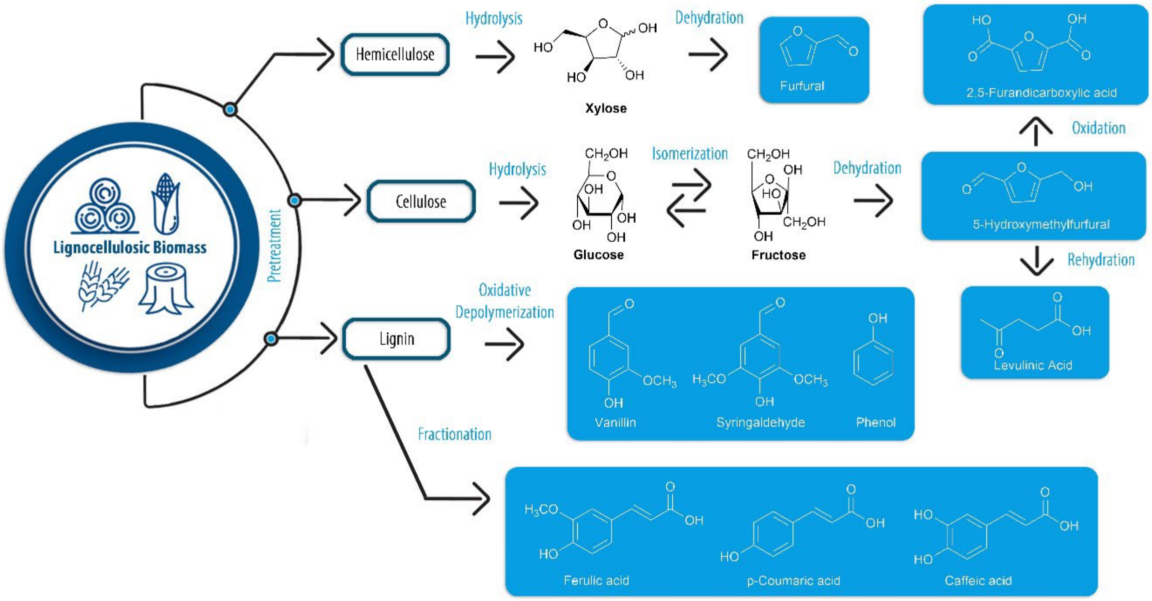


Figure 1. Example routes for lignocellulosic biomass conversion into select platform molecules (green background).[7, 8]

Despite profound advances in product discovery and catalyst development, biomass manufacturing has received less emphasis, and commercialization, beyond the corn and sugarcane ethanol and biodiesel, has been slow. The intense effort in converting a scientist's knowledge and intuition to experimental testing to the culmination of profitable and sustainable manufacturing from biomass feedstocks *motivates* this perspective. A key goal of manufacturing is to be economically viable and environmentally sustainable. Toward this goal, one needs to bring process-scale insights into the bench and use literature data and automation to create data at the laboratory scale to guide process development and systems analysis. We propose that these concepts can soon be met, aided by high-performance computing and artificial intelligence. Indeed, we argue that catalytic reaction engineering, data science, process optimization, and multiscale modeling are all harbingers of the digitalized future of biomass manufacturing. We discuss approaches and tools for bench- and process-scale digitalization. The learnings can form the foundations for a new integrated digitalization framework. We start by summarizing the top ten grand challenges in biomass manufacturing (Figure 2) and expand those amenable to digitalization. We close with the outstanding research topics.

II. Top Ten Grand Biomass Manufacturing Challenges

The top ten biomass manufacturing challenges, based on our experience, are depicted in Figure 2. We discuss select ones next and expand on the products slate that we should manufacture and the reaction paths in the next section as they are crucial but neither well-developed nor reviewed topics.

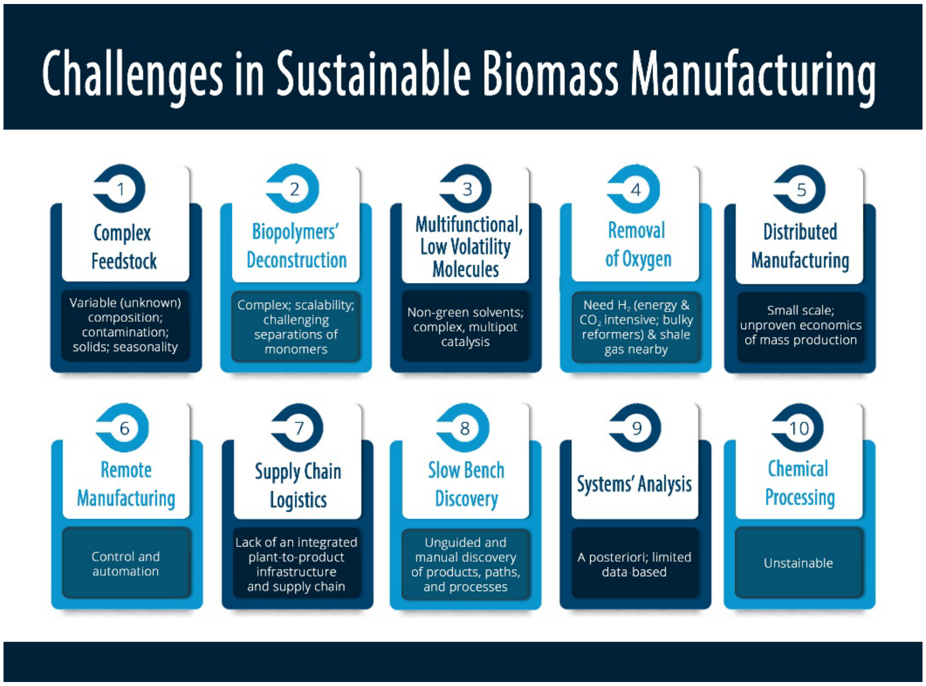


Figure 2. Summary of challenges for sustainable manufacturing from renewable feedstocks. Challenges 1-4 address feedstock diversity, catalysis, separations, and H_2 needs. Challenges 5 and 6 refer to the short transportation distance of wet feedstocks, the unproven economics, and the associated risks. Challenge 7 underscores the lack of a supply chain that starts with the feedstock logistics (production, harvesting, collection, soil maintenance, etc.) and ends up with markets; it creates barriers for individual companies in any part of the chain to enter this space due to lack of upstream and downstream industry. Challenges 8-10 are inspired by the very slow and unsupervised discovery and the lack of a design paradigm where product quality and sustainability are considered from the start rather than as an afterthought.

Feedstock diversity, complex transformations, and unknown chemistry paths. The transition from the current optimized fossil-fuel refineries to biorefineries is ridden with challenges at multiple scales, including the inherent complexity of handling solids and the seasonal and compositional diversity of the feedstock, the inherently complex chemistry of biomass, the use of (non-aqueous) and potentially non-green solvents to solubilize solids and improve processability, the significant developmental time required to identify optimal synthesis paths, solvents, and catalysts, the complexity of integrating processes and information at all scales, and the supply-chain logistics that start with harvesting, collection, and shortage, to pretreatment and multi-step processing to final product formulation. Aside from the lignocellulose, recycling and upcycling of food waste and other waste streams, including tires, biogas, methane leaks, and plastics, are at embryonic stages of development and are expected to face similar challenges.

Distributed, remote, small-scale, semi-supervised manufacturing. Due to the poor economics of transporting wet feedstocks over long (>50 miles) distances, manufacturing will happen at least partially close to the production site in the case of lignocellulose, agricultural waste, and food-producing farms. The limited self-life time of food waste is another driver for small-scale manufacturing. Local manufacturing can benefit rural communities, assisting economic equity. Small-scale, distributed manufacturing is against the established economy of scales paradigm. A century's chemical manufacturing has established that the cost drops with increasing size with an empirical power-law with an exponent of 2/3: the larger the plant, the cheaper the production. The most recent petrochemical plants in Qatar and refineries in the US have been some of the largest ever. Yet, the higher cost and longer time-to-build than the established cost-size 2/3 scaling have put this empirical law to question.

By extrapolating the 2/3 scaling to small plants, one expects the product cost per unit mass in distributed, small-scale manufacturing to be prohibitively higher. This observation makes manufacturing from renewable feedstocks a non-starter. This analysis misses the significantly lower cost stemming from mass production, i.e., the learnings from building big do not apply to building small massively. Several examples illustrate this point: manufacturing a vast number of automobiles, car engines (vs. trucks or planes), smartphones, and other electronics has profoundly dropped the per-unit cost. Small scale also de-risks investments by being inherently smaller. While massive production has seen tremendous success in consumer goods, it has not been tested yet in chemical manufacturing. Plants are much more expensive and complex than an iPhone. Furthermore, there is not always a market pool drive for investment. Continuous, well-integrated, and optimized processes are essential to cut down on utilities and energy use.

Over the past years, the continuous production of pharmaceuticals has been a successful example of transitioning manufacturing. Improved product quality, reduced cost and tighter economic margins, shorter manufacturing times, build where needed, personalized care, flexibility in the product, and small product volumes have been clear drivers that transformed batch processing to intensified, continuous processing. The large volumes of renewable feedstocks and associated products depart though from the pharma industry and the learnings acquired. The larger scale and lower product value create an inherent risk in building new manufacturing facilities. Remote and distributed (farm by farm) manufacturing require an integrated control and automation for remote operation with limited human intervention, reminiscent of semi-supervised machine learning.

Sustainable manufacturing via electrification. Conventional manufacturing using renewable feedstocks would help in sustainability but it still uses energy to run chemical reactions and separations, producing CO_2 and waste streams, making the net carbon balance positive. The net amount of the produced CO_2 per unit mass of a product is a crucial metric. In other words, renewable feedstocks alone can help but may be insufficient to solve the climate problem. A zero-emissions industry needs a paradigm shift that simultaneously considers sustainable renewable resources and sustainable manufacturing using electrification. For example, the production of bioethanol from corn requires significant amounts of energy to separate ethanol from the vast amounts of water—an energy-intensive separation that reduces the photosynthesis carbon capture benefit.[12] The nearly complete removal of oxygen (at least in most cases) from biomass requires profound amounts of H_2 whose current production leads to significant CO_2 emissions and is unsustainable.

III. Build the Transitioning Chemical Industry Better

Meeting the global warming challenge requires an intense effort to transition the chemical industry to zero-emissions. Sustainable resources and recycling combined with electrification are key elements toward this goal. A key question is what to make out of renewable sources. First, the volume of energy and fuels is about two orders of magnitude larger than chemicals. For a substantial environmental impact, one needs to produce fuels and go beyond sugarcane- and corn-based ethanol and biodiesel and produce second-generation biofuels (from lignocellulosic biomass or woody crops, agricultural waste, and dedicated energy crops grown on marginal land). However, the economics of producing fuels is not as attractive. The failure of commercializing large-scale lignocellulosic ethanol underscores precisely the economic challenge. Chemicals provide a much more lucrative target.

Future biofuel production. Given batteries have a (much) lower energy density than fuels, battery-based energy storage for long journeys of airplanes and cargo ships is currently not viable. The European Union announced on July 14, 2021, the ban of new petrol and diesel cars by 2035.[13] Its new ReFuelEU Aviation Initiative mandates blending 5% by 2030, 32% by 2040, and 63% by 2050 sustainable aviation fuel (SAF) with fossil fuels. Renewable and low carbon fuels should be 6-9% of the international maritime transport fuel mix in 2030 and 86-88% by 2050.[6] Whether this will happen using green ammonia, pressurized green hydrogen, or biofuels is still to be determined. We expect SAFS and maritime fuels to be the focus of future biorefineries.

Pyrolysis has been to most mature technology for the production of fuels. It uses thermal energy at high temperatures to unselectively break down bonds leading to a product stream, called bio-oil, containing hundreds of compounds with multiple functionalities. While agnostic to the feedstock and thus versatile (pros), the separation of the product stream into individual compounds is energy-intensive and challenging (cons). Aside from lignocellulose, pyrolysis has been applied to lignin, a major low-value byproduct in the paper industry. However, the economic viability is an issue. Technologies, such as aldol condensation, starting from the building blocks of (hemi-) cellulose, have also been exploited. The furan platform to biofuels of specific molecular weight and structure is compelling.[14] The optimal path to make biofuels will be dictated by several factors including economics, the policy mandates, the scale of biofuels, the chemicals co-produced, etc.

Technoeconomic analysis (TEA) of biomass processes over the past 15 years has underscored that the lower value/higher cost of fuels makes the co-production of chemicals necessary. Biomass consists of biopolymers with building blocks of five- or six-carbon sugars (cellulose and hemicellulose) and aromatic rings connected via small linkers (lignin) that could be ideal platforms to certain chemicals. We discuss these next.

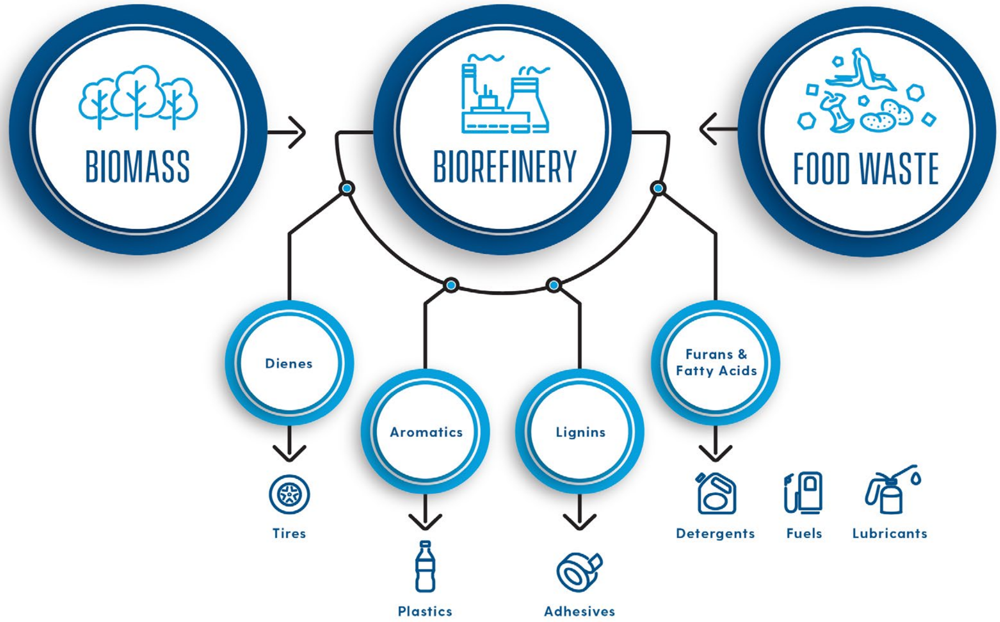


Figure 3. Schematic of feedstocks and select product classes produced via catalytic transformations in the Catalysis Center for Energy Innovation.

Direct replacement chemicals. The vision in the 2000s was to replace refinery chemicals with identical molecules. Given the biomass makeup, the production of chemicals of five to about ten carbons is a clear target. It entails three overall steps: (1) The deconstruction of the biopolymers via C-O and C-C bond scission to produce the monomers and linkers (in the case of lignin); (2) their defunctionalization to remove (some of) the oxygen; and (3) the ring-opening and deoxygenation to make target chemicals or the C-C coupling chemistry to make larger molecules.

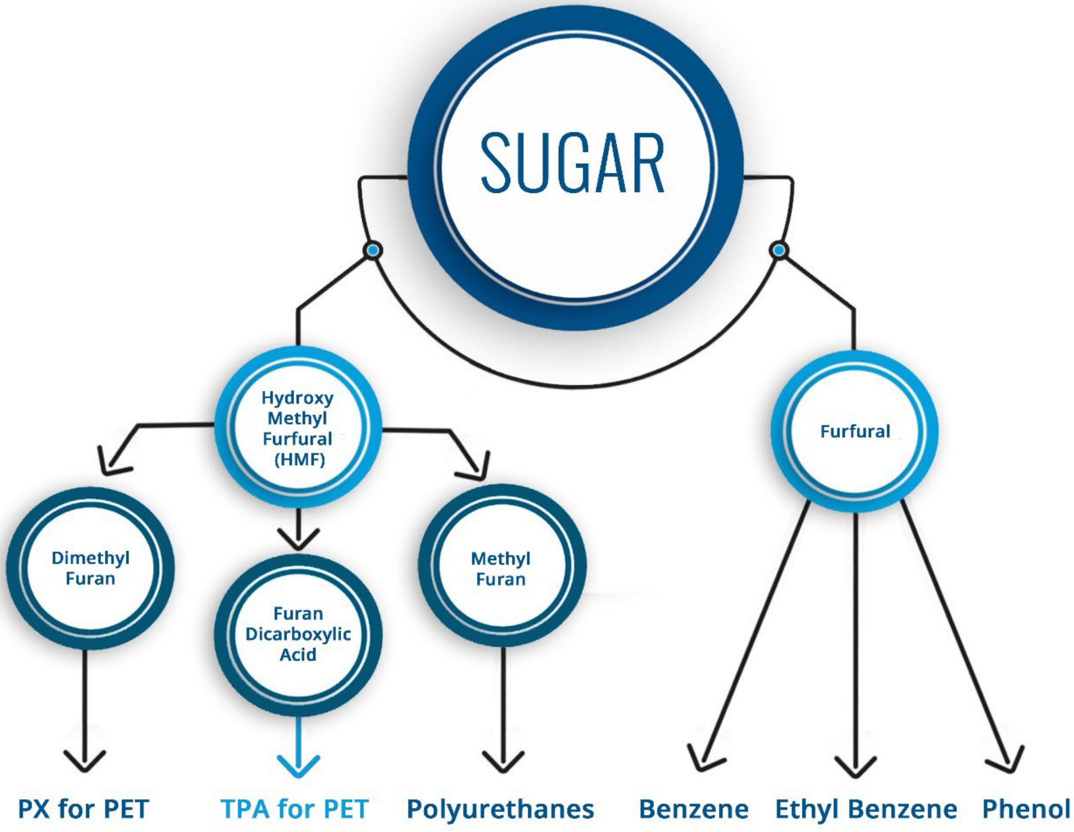
Figure 3 shows potential routes to select replacement targets: dienes, aromatics, diols, diacids (e.g., pentanediol, hexanediol, adipic acid, succinic acid, butadiene, pentadienes).

Figure 4a details paths for producing renewable aromatics from furanics of (hemi-)cellulosic biomass. Specifically, renewable terephthalic acid (TPA) is a bench-scale success for replacing crude-oil-based PET (polyethylene terephthalate polymer) for plastic bottles and other applications. An initial discovery in the Catalysis Center for Energy Innovation (CCEI), using Bronsted acid catalysts led to a modest yield.[15, 16] Advancing fundamental knowledge and catalysts led to a nearly quantitative (97%) yield over a new class of P-containing zeolites,[17] tremendously improving the economic viability. A collateral benefit was the discovery of P-containing and other weak Bronsted acid zeosil catalysts with remarkable selectivity in dehydration chemistry beyond p-xylene production for making dienes[18, 19] and base oil lubricant precursors.[20]

This remarkable optimization took over five years of intensive research from a talented group of investigators with complementary skills in materials synthesis, catalysis, reactors, kinetics, mechanisms, multiscale modeling, and systems analysis. This period is shorter than the usual academic time due to teamwork but still long due to training the future workforce and trial-and-error. Yet, bench-scale concepts need advancement to higher technology readiness levels (TRL) for commercialization more rapidly.

While the platform molecules and replacement chemicals are known, the optimal pathways are not. For example, Figure 4b depicts two paths from HMF to TPA. Path B involves fewer reaction steps and reactors and avoids oxidation and reduction (and importantly, the use of H_2) invoked in Path A; it makes sense from processing and sustainability viewpoints. However, the electron-withdrawing properties of -COOH make the Diels-Alder chemistry in Path B very hard to achieve. Path A leverages the slightly donating $-CH_3$ groups to achieve high rates and quantitative product (para-xylene) yield. This pathfinding was not evident before conducting extensive research. The Diels-Alder reactor employs heptane as a solvent to reduce water (a byproduct of the reaction) around the Bronsted acid sites that hydrolyze the dimethyl furan to (a byproduct) diketone. The choice of heptane, though, was intuitive. Despite Diels-Alder being organic textbook chemistry, significant effort was invested in optimizing the catalyst, solvent, and processing conditions. Production of other replacement chemicals has faced similar development challenges. Digitalization can revolutionalize such optimization.

(a)



(b)

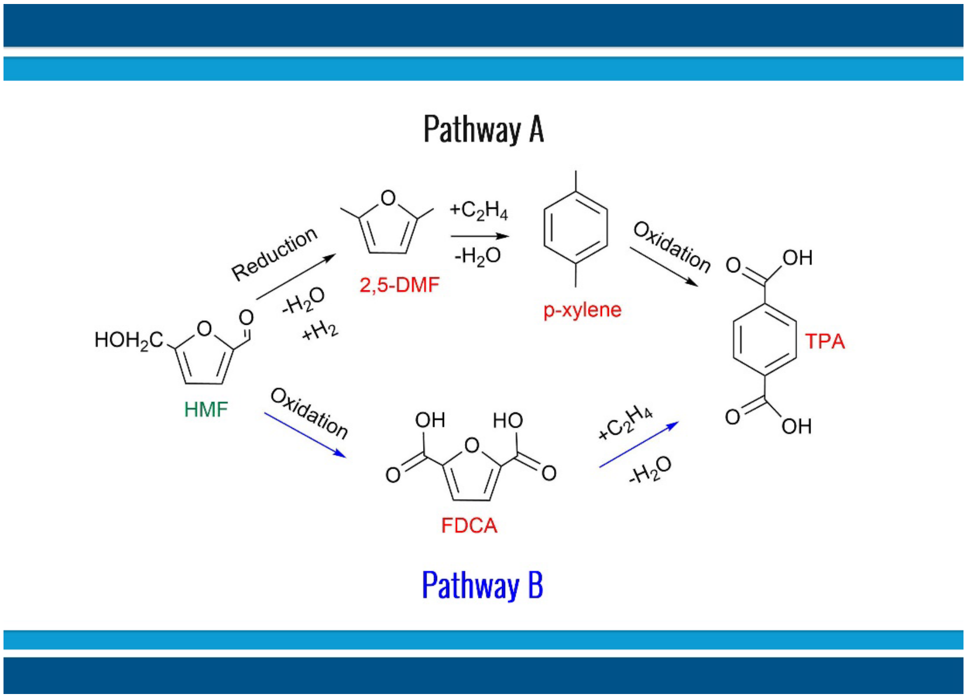


Figure 4. (a) Potential roadmap to replace aromatics from furanic platform molecules in the Catalysis Center for Energy Innovation. (b) Two possible paths to renewable TPA: Path (A) entails the hydrodeoxygenation of HMF to dimethyl furan, followed by the Diels-Alder addition of C_2H_4 *to para-xylene and the mid-century oxidation process to TPA. Path (B) entails the direct oxidation to FDCA followed by the Diels-Alder addition of* C_2H_4 *to TPA.*

Performance-advantaged chemicals and products. The transition of the chemical industry into zero emissions requires a holistic view of the resources, manufacturing practices, electrification, automation, and the targeted products. Notably, the functional groups in platform chemicals can give products with unmatched properties, potential biodegradability, and high value. Here we discuss such examples.

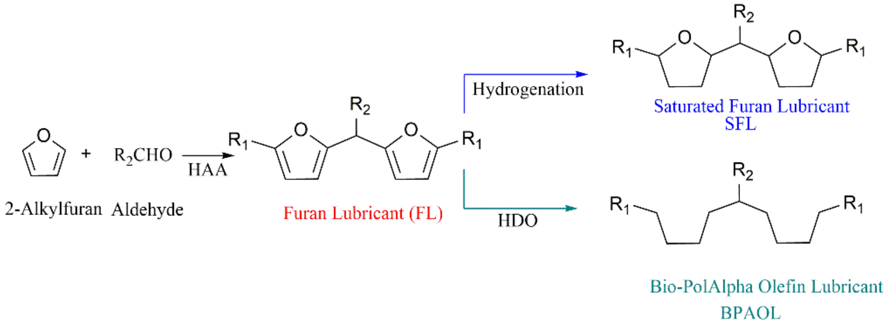
Furan rings have aromatic properties but contain oxygen, and are thus more hydrophilic, and possess higher solubility in polar solvents. As a result, alkyl furan detergents often have better properties than commercial alkylbenzene sulfonates.[21] Similarly, FDCA (2,5-furandicarboxylic acid), the diacid of HMF, can be produced by selective oxidation of HMF to replace TPA and produce renewable PEF (poly(ethylene 2,5-furandicarboxylate) with ethylene glycol, a renewable analog of PET.[22] PEF possesses high performance (diffusion barrier, thermal, and mechanical properties). Recognizing the potential, Avantium started manufacturing PEF for films, fibers, and packaging of soft drinks, water, and alcoholic beverages in 2010.

Turning next to lignin, the monomers 4-propylsyringol and 4-propylguaiacol, obtained from poplar wood lignin, can be functionalized with acrylate or methacrylate groups and then use reversible addition−fragmentation chain-transfer (RAFT) to yield polymers with high glass transition temperatures and good thermal stability. These molecules can serve as low-dispersity triblock polymers for high-performance pressure-sensitive adhesives that outperform commercial products.[23]

Lubricants are ubiquitous in modern life applications. They hold an over $126 billion market

Lubricants are ubiquitous in modern life applications. They hold an over $126 billion market globally[24] with a projection to $183 billion by 2025.[25] Base oils make most of the lubricant (75-99 wt%), with the rest being additives to impart suitable properties in a product. They contain C_{20-50} hydrocarbons. Synthetic base oils of lower viscosity, such as poly-α-olefins (PAOs), are typically used in automobiles; those of high viscosity, like alkylbenzenes, in cooling and refrigeration.[26, 27] Selectively tuning synthetic base oils' molecular size and architecture is challenging. The cationic oligomerization of linear C_{8-12} α-olefins uses homogeneous corrosive acid catalysts (BF_3, HF, or $AlCl_3$) and forms multiple products, leading to expensive separations and a significant environmental footprint. The oxygen of platform molecules enables selective C-C coupling, e.g., via hydroxyalkylation/alkylation (HAA), producing molecules of a precise number of carbons and molecular structure. Their properties can outperform commercial products.[20, 28-30] Figure 5a shows a pathway to renewable base oils using alkylfurans and fatty acids. Figure 5b shows the molecular structures of ~C_{30} from three paths. Remarkably, C_{30}-BPAOL1 and C_{31}-BPAOL1 differ in one methyl only. Yet, their pour points differ dramatically (~-20 vs. -50 °C). This example illustrates the sensitivity of properties on the structure and the importance of having structure-property relations. It also underscores the exquisite selectivity of the underlying chemistry. Importantly, this lubricant technology employs solid catalysts and produces minimal byproducts, reducing separations and energy use. Overall, in addition to using renewable resources, manufacturing is more sustainable. As another example, lignin monomers can be converted to alkylbenzene-based lubricants for refrigerants. A pathway from our recent work is depicted in Figure 6.[31] It illustrates how one can convert lignin (a waste stream from the paper industry) into high-value products.

(a)



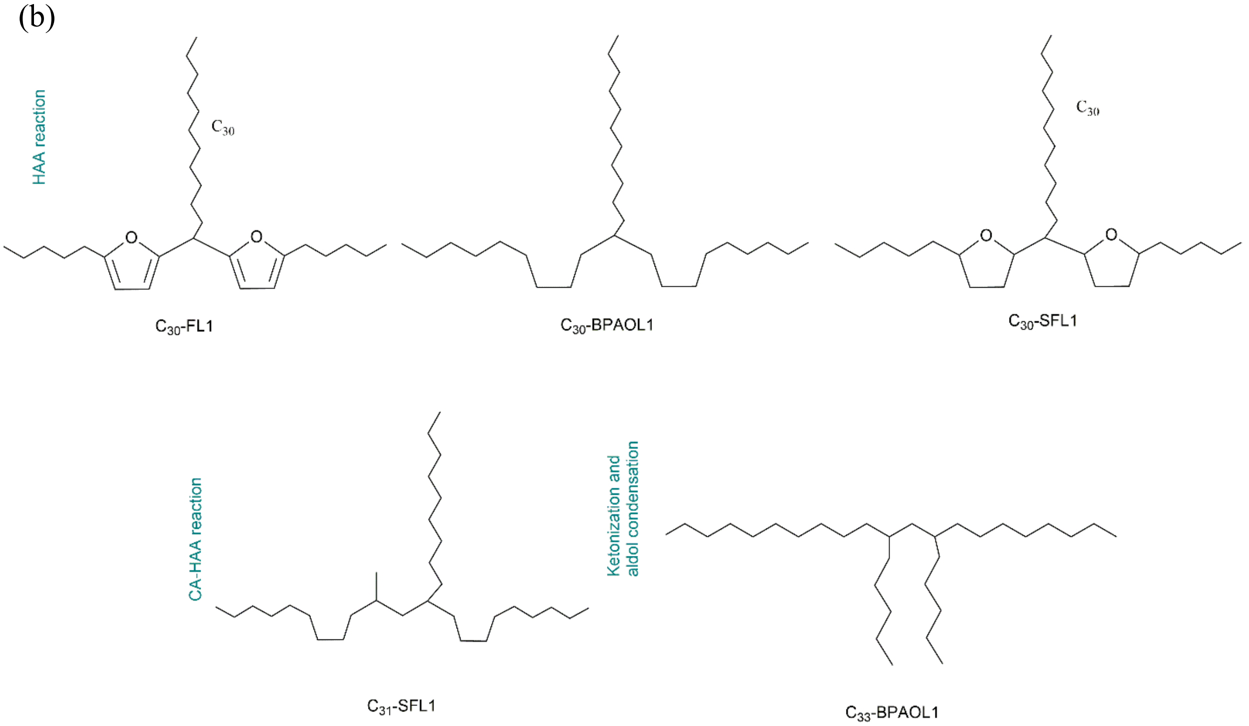


Figure 5. (a) Production of three classes of lubricants from alkyl furans and long-chain aldehydes. (b) Examples of base oil lubricant structures from different chemistry paths and starting chemicals.[20, 28-30]

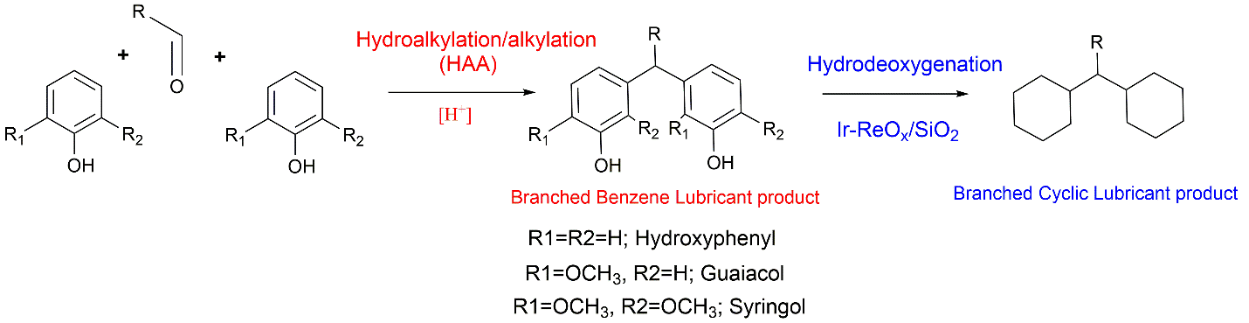


Figure 6. Strategy to produce branched benzene lubricant (BBL) base oils and branched cyclic lubricant (BCL) base oils via Brønsted acid-catalyzed hydroxyalkylation/alkylation (HAA) of lignin-derived monomers (hydroxyphenyls, guaiacols, and syringols) with an aldehyde and the and hydrodeoxygenation (HDO) of the BBL product, respectively.[31]

Food waste is a poorly managed feedstock. Yet, its environmental, societal, and economic impact is enormous. About one-third of the food produced is wasted, resulting in ~34% of the anthropogenically-induced methane emission, loss of 26 EJ energy (equivalent to ¼ of the US energy consumption), and 3.3 billion CO_2 equivalents comparable only to big countries (China and the US). Yet, food waste can be a renewable resource to manufacture bioproducts circularly. Figure 7 shows a roadmap of potential products, applications, and value. Many products are unconventional with unestablished markets. The quantities of extractables are often low and ideal for small volume consumer products. Among them, antioxidants[32] provide tremendous commercialization opportunities unmatched by lignocellulose. A key challenge and opportunity here are to develop economically viable and sustainable extraction processes to recover these high-value molecules before (more aggressive) chemical manufacturing. Following extraction, additional separations to get the main constituents may be necessary. For fruits and vegetables, which constitute a large fraction of food waste and contain a significant portion of the starch, typical deconstruction methods employed in lignocellulosic biomass can be used. These can entail acidified molten salt hydrolysis of starch to produce sugars, acid-catalyzed dehydration of the sugars to furans and their upgrade to value-added products, and pyrolysis of the leftover fraction to produce carbon-based materials as adsorbents, electrodes, etc. Such an integrated, multiple process approach was recently demonstrated for potato peels.[32]

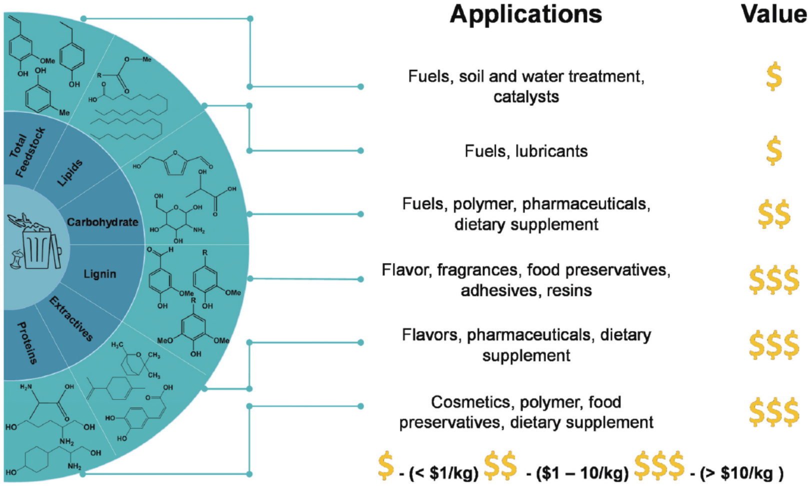


Figure 7. Schematic roadmap of potential products from food waste.[33] The figure is reproduced from Ref. 34 with permission from the Royal Society of Chemistry

Barriers and opportunities for manufacturing renewable products. The established petrochemical products and their properties are mainly known, and the R&D focuses on catalyst and process optimization and product formulation. Biomass manufacturing creates a new set of challenges and opportunities stemming from the diversity of feedstocks and platform molecules and the multiple paths to convert platform molecules (Figure 8) to performance-advantage products, such as the adhesives, detergents, and lubricants mentioned above. Creating new molecules is tedious, human-intensive, and risky.

The economics and sustainability metrics of performance-advantaged products studied so far look excellent, creating a substantial market pool and economic driver for manufacturing. Yet, two overarching challenges must be overcome, namely the lack of (1) a framework to decide the best products, paths and processes for manufacturing, and (2) structure-property relations of new molecules. Defining a superstructure of all viable paths and processes is a necessity. Finally, laboratory demonstration and validation of the best solutions will be critical. All in all, a systems perspective, which integrates molecule, product, and process design, would be essential. A strong partnership between academia, industry, and markets can be impactful.

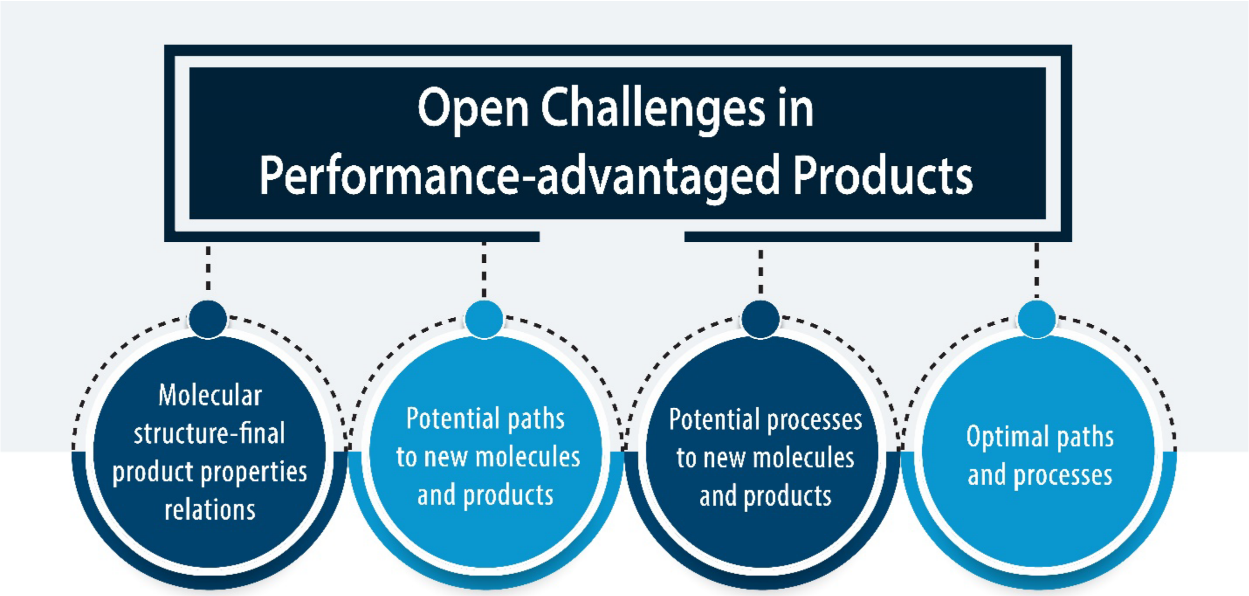


Figure 8. Digitalization enablers for accelerated manufacturing of new molecules and products.

IV. Bench-Scale Digitalization

The idea behind bench-scale digitalization is to extensively utilize the information available in the literature and perform critical experiments in a high-throughput fashion. This enables the generation of high-quality data, leading to optimal processes for scale-up. Figure 9 depicts the workflow for bench-scale digitalization that we touch upon next. Details of retrosynthesis and automated laboratory are given in recent reviews.[34, 35]

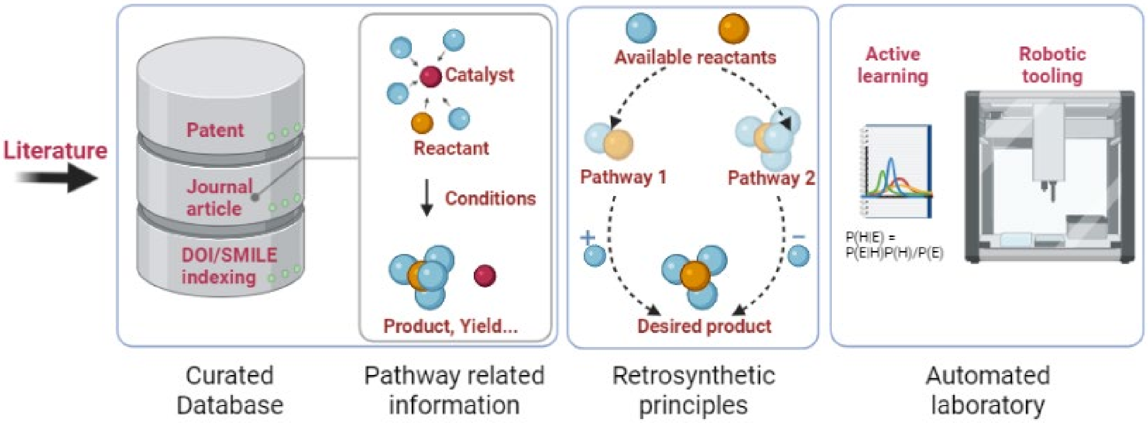


Figure 9. A pathway towards bench-scale digitalization.

A challenge in undertaking a new biomass catalysis project at the lab scale is the massiveness of the scientific literature. This problem is evidenced in a cursory literature search with "biomass" and "catalysis" keywords, which reveals >1,000 relevant papers per year published in the last five years. Under such burgeoning information, it is increasingly difficult to sieve the available knowledge and extract essential information manually.

Retrieval of information and its curation into knowledge databases is emerging in many scientific areas. Cheminformatics, i.e., the extraction of knowledge from molecular structures of reactants, intermediates, and products, has been used since 1975,[36] and is summarized in informative reviews.[37-39] Most pertinent is the curation and description of reaction networks where text mining screened >1.8 million reactions from drug discovery patents.[40] Since then, a multitude of techniques ranging from hard-coded ranking,[41, 42] deep neural networks (NN)[43, 44] using the string format SMILES (the simplified molecular-input line-entry system),[45] TensorFlow,[46] and a graph-convolution NN—harnessing databases, such as Reaxys[47] and SciFinder,[48] or patents[49] have been used.

Yet, these methods typically apply to homogeneous catalysis. Heterogeneous catalysis entails solid materials whose active sites and correlations of active sites with catalyst performance are often unknown. The revolution in data science, the accumulated information in the literature, and the automated workflows on the computer and the laboratory create a tremendous opportunity for discovery. We introduce methods for parsing the literature to a database of biomass conversion.

Literature collection, information extraction, and classification. The first step in the envisioned workflow is literature collection. Domain-specific knowledge bases (PubChem, Chemistry Webbook, etc.) describe molecules with known name variations. Automated data retrieval through querying SciFinder, CrossRef, or Google scholar can enable the collection of relevant literature. The downloaded literature must be uniquely tagged to ensure database consistency, a task made easy with document object identifiers (DOIs).

Upon relevant literature selection, text mining and natural language processing tools can automatically extract and classify information about: (1) reactants, (2) platform molecules, (3) catalysts, (4) operating conditions, (5) solvents, and (6) performance, e.g., yield to products and selectivity and rates. For instance, when operated on a passage[50] like *"5-(Hydroxymethyl) furfural (HMF) and levulinic acid production from glucose in a cascade of reactions using a Lewis acid* (*$CrCl_3$*) *catalyst together with a Bronsted acid (HCl) catalyst in aqueous media is investigated"*, an automated classifier should distinguish product platform chemicals (HMF and levulinic acid) from reactant (glucose), tag the catalysts ($CrCl_3$, HCl), and recognize water as the solvent (aqueous media). Continuing the passage "*Yields of 46% levulinic acid in a single-phase and 59% HMF in a biphasic system have been achieved at moderate temperatures by combining $CrCl_3$ and HCl",* the classifier should ideally identify the performance information (yields) and parse the available qualitative operation conditions (moderate temperatures). The obtained data would most likely be disjoint and relational databases of semi-structured[51] and relational mappings[52], such as NoSQL databases[53] (e.g., MongoDB[54]) would be required for storage.

Such classification is best understood through examples from the mature field of organic synthesis. Cronin and co-workers[55] relied on expert human knowledge heuristics to extract chemical information. In contrast, Laino and co-workers[56] used deep-learning to convert experimental procedures to action sequences without human involvement. Operating between these two extremes, Barzilay and co-workers[57] recently used human intervention to validate the automated classifier of reactants, products, and operating conditions. Considering the nascency of and the complexity inherent to biomass catalysis, such a supervised learning approach could be the first step forward. In such an approach, automated classifier features, including domain-specific word embeddings (e.g., "in the presence of" indicating subsequent catalyst information)[58], and specific patterns for each category (e.g., "heated to" indicating operating conditions) could be utilized. Other specific features would consist of measurement units (e.g., °F or °C for temperature)[59], regular expression matching (e.g., subscripts in chemical formulae),[60] and expert knowledge (e.g., glucose is a typical reactant). These would need, at least initially, human validators to accept or reject the outputs via reinforcement learning. The human-in-the-loop approach also indicates a clear need for an organized and collaborative community effort to create open-source training datasets to avoid duplication.

We have identified several gaps for moving the field forward. First, the content format needs to be easily readable by subsequent models. For instance, older content is primarily available in portable document format (PDF) or images, making extraction challenging. This limits current efforts to articles where a markup language-based text is available. Second, patents tend to be country-specific; having access to databases across countries could be challenging. These aspects curtail the information available for a digitalized framework. Third, the current extraction methodology is restricted to text data. However, images and graphs are vital sources of information. It is typical to report molecular structures as images. Extracting molecular information requires specialized models beyond those described here. Within graphs, there is heterogeneity in visualization (e.g., bar graphs, line graphs, pie charts, etc.), each requiring specialized computer vision models for data extraction. Therefore, tools and frameworks beyond those discussed here are needed to broadly digitalize these important knowledge streams. Beyond new tools, inclusion of underlying graphical data as tables in supplementary information or sharing structured/unstructured databases with the manuscript submission are important moving forward. Digitalization of the data, graphs, and images provided in future publications can alleviate the problem in part; yet, we expect that searching, organizing, and harnessing massive inhomogeneous data-streams would remain a major endeavor, and sustained community cooperation will remain at the heart of enabling a digital future for manufacturing.

Retrosynthetic construction of reaction paths using reaction rules. A database can train models to predict feasible catalytic paths to desired products from biomass platform chemicals, a process known as retrosynthesis in synthetic organic chemistry. Retrosynthesis is an inverse problem: it starts with desired products and works backward to identify suitable reactants. Identifying pathways between known reactants and products is a long-standing problem in organic chemistry, often referred to as scoring or ranking. The visionary work of Corey and Wipke in 1969[61] demonstrated that a computer could do synthesis and retrosynthesis using intuitive reaction rules, also known as reaction templates, and eventually led to the 1990 Nobel Prize. The approach, while pioneering, has seen limited success despite the advances in computer power due to the need for expert organic chemists to define these rules. In the past decade, the automatic generation of reaction rules from a vast database consisting of millions of reactions, e.g., the Route Designer tool, has been demonstrated.[62] The automated extraction of reaction rules entails the following steps: 1) The reaction database is formatted to map atoms in reactants and products; 2) the reaction core, i.e., the set of atoms participating in the reaction, is identified from the atomic map; 3) the reaction cores are extended to include first and second neighboring atoms or functional groups that are vital in the reaction; 4) the extended reaction cores are classified into groups using similarity, including the nature of the bond breaking or forming, charges, coordination environment, hybridization, aromaticity, etc.; and 5) the reaction template representing all reactions in a group constitutes the reaction rule.

The automated rule extraction was applied to an electronic lab notebook containing about 400,000 reactions.[63] The reactions are given string representations known as SMIRKS and those with similar SMIRKS are grouped. The reactions in a group are sorted based on yield, and the reaction with a higher yield is chosen as representative. These groupings are thus converted to a retrosynthesis tool. Apart from the concepts of reaction cores and SMIRKS to represent reactive transformations, the concept of reaction vectors has also been used.[62] Reaction vectors involve the encoding of changes in a reaction (changes in hybridization, bonds, atoms, etc.).[64] An alternative approach to automated rule extraction uses graph theory. The reactants and reactions are represented as nodes, and the edges connecting these nodes represent reactants, reagents, catalysts, solvents, or products. The graph theory suppresses combinatorial explosion and is superior to the rule-based approach in discovering new reactions and capturing the roles of reagents and catalysts.[44] Other alternatives to predict reactions include hard-coded ranking[41, 42] and a deep neural network;[43] the latest advancements borrow concepts from computational linguistics and cast an overall reaction as a translation problem, mapping reactants to products as a neural machine text sequence-to-sequence (seq2seq), using SMILES (the simplified molecular-input line-entry system)[45] and TensorFlow.[46]

The advantages of these data-centric approaches are the (a) use of experimentally based (instead of intuitive) reaction rules, (b) estimation of confidence intervals for the feasibility of a reaction family and individual reactions, based on the frequency of observations, (c) ability to add new experimental data and first-principles calculations to improve prediction, and (d) identification of missing pathways. Despite its application to synthetic organic and medicinal chemistries, retrosynthesis has not been employed for biomass research. This is partly due to the lack of quantitative process models and data built from such retrosynthetic models.

Quantitative knowledge beyond retrosynthetic reaction rules. Extracting reaction rules from literature about the potential transformations is a significant step toward discovery but not sufficient. Unlike pharmaceuticals, whose volume is small and their price is high, the amounts and value of fuels and chemicals are massive and low, respectively. It is thus critical to perform early on material and energy balances along with TEA and LCA (Life Cycle Assessment) to decide the viable paths. Such an analysis requires quantitative reaction data and models. Kinetics information entails reaction rates, apparent activation energies, reaction orders, and rate expressions for all reactants and products. Unfortunately, only product yields are typically reported. Early literature data did not even close carbon balance. Vital information needed includes 1) chemical process information; 2) operating conditions; 3) reaction rates, kinetic data, and reaction network data for the observed compounds; 4) catalyst and solvent information; 5) information on unobserved or unreported species; and 6) molecular-level understanding, especially of elementary reactions crucial for catalyst and solvent selection. Due to significant differences in experimental conditions (operating conditions, solvents, and catalysts) among studies, it is often impossible to compare and rank-order catalytic paths from literature data alone. New tools are desperately needed to close this gap. The automated laboratory discussed next is one means toward this task.

The automated laboratory. The proposed paths should be tested and optimized in the laboratory by conducting experiments at varying continuous variables, such as temperature, pressure, catalyst loading, feed concentration, residence time, and discrete variables, such as the catalyst, solvent, and reactor type. This is essential information for process synthesis and optimization.

At the lab scale, converting an intuition into a viable reaction path to make a product is non-trivial and is a hit-and-miss approach. The solvents, catalysts, starting reactants and synthesis routes, and reaction conditions of many reactions remain unoptimized. One at a time parameter optimization to determine the optimum reaction conditions is a common practice despite the multidimensional optimization nature of these problems. The traditional design of experiments is rarely used and is still time-consuming and costly for high-dimensional problems. Optimizing reaction conditions, as often done, applies to a specific catalyst, solvent, and starting materials. If any of them were to be modified, the entire optimization must be repeated. This emphasizes labor-intensive research and does not prepare the future workforce for creative thinking. With the ever-changing vast literature, early-on determination of leading catalysts, solvents, and starting materials is crucial to saving time and resources. However, literature data has not been leveraged efficiently and effectively toward this goal.

HTE (high throughput experiments) enable rapid screening in miniature well plate batch or continuous flow reactors, leading to diverse data for process model building and optimization. HTE has revolutionized the pharmaceutical industry by accelerating the screening libraries of compounds for therapeutic targets.[65-67] The well plate reactors facilitate the simultaneous screening of several hundreds of solvents and reagents for a specific reaction. The flow reactors enable continuous production of chemicals over many conditions, safe handling of hazardous substances, and uniform mixing and heating; they also can prevent solvent evaporation.[68] HTE has advanced its dispensing capability, e.g., the Acoustic Droplet Injection (ADE), to deliver accuracy and precision in adding reactants at various frequencies while being contactless to avoid impurities.[34, 68, 69] Programmed robots for preparing the reactant mixtures/feed streams and performing reactions at many conditions and in-line analysis via mass and vibrational spectroscopy, chromatography, NMR, etc., and finally extracting and curating the corresponding real-time data.[35, 70-72] Software interfaces can codify the instruction to easily performable actions. For instance, ChemOS[73] interfaces humans and automation for supervised experimentation while incorporating expert knowledge. Alternatively, one can integrate hardware and software modules for automated synthesis.[70] An end-to-end automated HTE framework also requires software to determine the sequence of experiments and carry out the codified experiments, i.e., data-driven active learning algorithms, which provide optimal performance with reduced time and materials. For instance, the open-source software NEXTorch[74] and Gryffin[75], among others[76-79], bring Bayesian optimization to chemical sciences and enable categorical variables while being more effective than the traditional design of experiments.

While optimization of continuous variables is relatively straightforward, that of discrete variables is not. One can use literature and/or HTE data from an automated laboratory to correlate performance with descriptors, such as solvent polarity and H-bonding, acidity for Lewis and Bronsted acid catalysts, electronegativity, d-band center and width for metal catalysis, etc. These descriptors are computable using multiscale simulations. Optimization using the correlations can point to solvents and catalysts to improve performance. While conceptually simple, such a framework is currently lacking.

V. Process Systems Engineering (PSE) for Product and Pathway Discovery and Optimization

A wide gap between the laboratory and process levels exists; closing this gap is essential for commercializing new technologies and products. This section reviews key steps and outstanding challenges toward commercialization.[80] A modeling hierarchy relies on the design maturity presented by the Association for the Advancement of Cost Engineering[81] according to the confidence intervals (indicated by the numbers in parentheses) of the estimated budget: conceptualization (100%), identification of units modeled with parametric models (50%), design of processes with semi-detailed unit costs (30%), detailed unit design (20%), and detailed unit cost and take-off (15%). In most cases, TEA with semi-detailed cost estimations[82, 83] coupled with an LCA for sustainability, including GHG emissions, water, land, and energy consumption, are performed.[84] Early-on TEA and LCA use semi-detailed models[82, 83, 85] and can eliminate processes of low profitability or unfavorable environmental indicators. This approach can be accelerated via bench-scale digitalization discussed above and a computational framework for the design and optimization of processes and products, a method widely established by the Process Systems Engineering (PSE) community.[86] The tools and frameworks depend highly on process maturity and extent of analysis. Below we present the state-of-the-art tools for various stages of process/product development. Process synthesis entails the conceptualization and identification of unit operations and is followed by semi-detailed process design and optimization and detailed design.

Process synthesis. We start with building and comparing different process alternatives using simple models (i.e., linear or parametric models) from literature and experimental measurements. Two overarching approaches can determine the best alternative: hierarchical decomposition and superstructure optimization.[87] The former[88] makes decisions based on experience and rules of thumb; the latter explores alternatives as a mathematical programming problem and is preferred due to covering more alternatives systematically. The scales considered depend on the scope and process maturity. Some process synthesis studies have only considered alternative units.[89-92] Detailed synthesis, on the other hand, includes operating conditions and catalysts but not kinetics.[93, 94] Regardless of the detail, the initial superstructure should be dynamically adapted as new technologies and information emerge.[95] For example, early works focused only on producing biofuels[89, 96, 97] and recent ones value-added products[95, 98] and new catalytic paths.[94] The synthesis involves the following steps:

- Literature data and expert knowledge to create possible alternatives. For a new process, experimental data is necessary.
- Parametric models for process design using available simulators, like Aspen ®.[82, 83] The overall fluxes of raw materials, utilities, emissions, operating (OPEX) and capital (CAPEX) costs are

determined. The OPEX is estimated from the fluxes of raw materials and their prices, and the CAPEX using literature[99-101] and digital tools, e.g., the Aspen Economic Analyser® [102]. The environmental footprint includes direct and indirect impacts. The former is obtained from the mass and energy balances and only requires simple estimates, e.g., the Global Warming Potential of different gas emissions.[103] The latter is associated with the supply chain and waste management, and requires additional literature data[103, 104] or extensive databases like Ecoinvent.[105] Although indicators are not unique because of the different weights to each impact,[106, 107] standardized indicators, such as ReCiPe[108, 109] or Traci,[110, 111] are used in the LCA of biomass processes. An extension entails social impacts using the Social-Life Cycle Assessment (S-LCA), [112-114] but these indicators are subjective. They include the Human Development Index of the United Nations[115, 116] or custom indicators depending on the region. Developed countries have included impacts, such as job generation or health[117, 118]; developing countries include the percent of working children.[119]

- Once the fluxes and costs have been determined, one generates a superstructure[82, 83] (Figure 10). Each alternative is a block with connections to upstream and downstream technologies. The most promising solution is found via optimization. The design is still initial, and the models are linear. Nonlinear models are linearized,[120] leading to a Mixed Integer Linear Programming (MILP) model. Its solution determines the most profitable,[89, 93, 121] environmentally[90, 122] and socially conscious path.[118, 123] Multiple objectives[90, 91, 122, 124] lead to multi-objective optimization and a Pareto trade-off between objectives.

The optimization can also include uncertainties arising from the heterogeneity in biomass composition[125], demand variability,[126] and market prices,[127] as commonly done in superstructure optimization.[128, 129] The uncertainty space is sampled using Latin Hypercube sampling (LHS) via probability distributions.[127, 130] Given the large uncertainty space, the deterministic MILP problem needs to be reformulated.[131] In the synthesis, uncertainty is handled using robust optimization and two-stage stochastic programming. The former includes an additional term in the objective function.[92] The latter fixes the synthesis decisions and adjusts operating variables to account for uncertainties.[127, 132, 133] Fuzzy logic programing has also been used for market uncertainties.[126, 134] The combination of statistical sampling and deterministic optimization has been demonstrated but is computationally demanding.[130, 135]

In summary, process synthesis has become much more widespread in biomass conversion. Despite exciting developments, there are still many challenges for biomass-based process synthesis. Literature information described above can be used to generate a superstructure, including intensified processes, and rank products and processes. However, the qualitative nature of the data leads to numerous alternatives, making identification of the most promising one hard. Clustering and other reduction and aggregation approaches can potentially identify the most promising directions while considering uncertainty and variability systematically. Integrating uncertainty in the resulting optimization problem is non-trivial,[136] as it requires high computational power and efficient algorithms. Furthermore, formulating the multi-objective optimization with economic, environmental, and social targets under uncertainty demands new strategies. In this analysis, the time horizon is an important consideration as it affects the short vs. long term solutions.

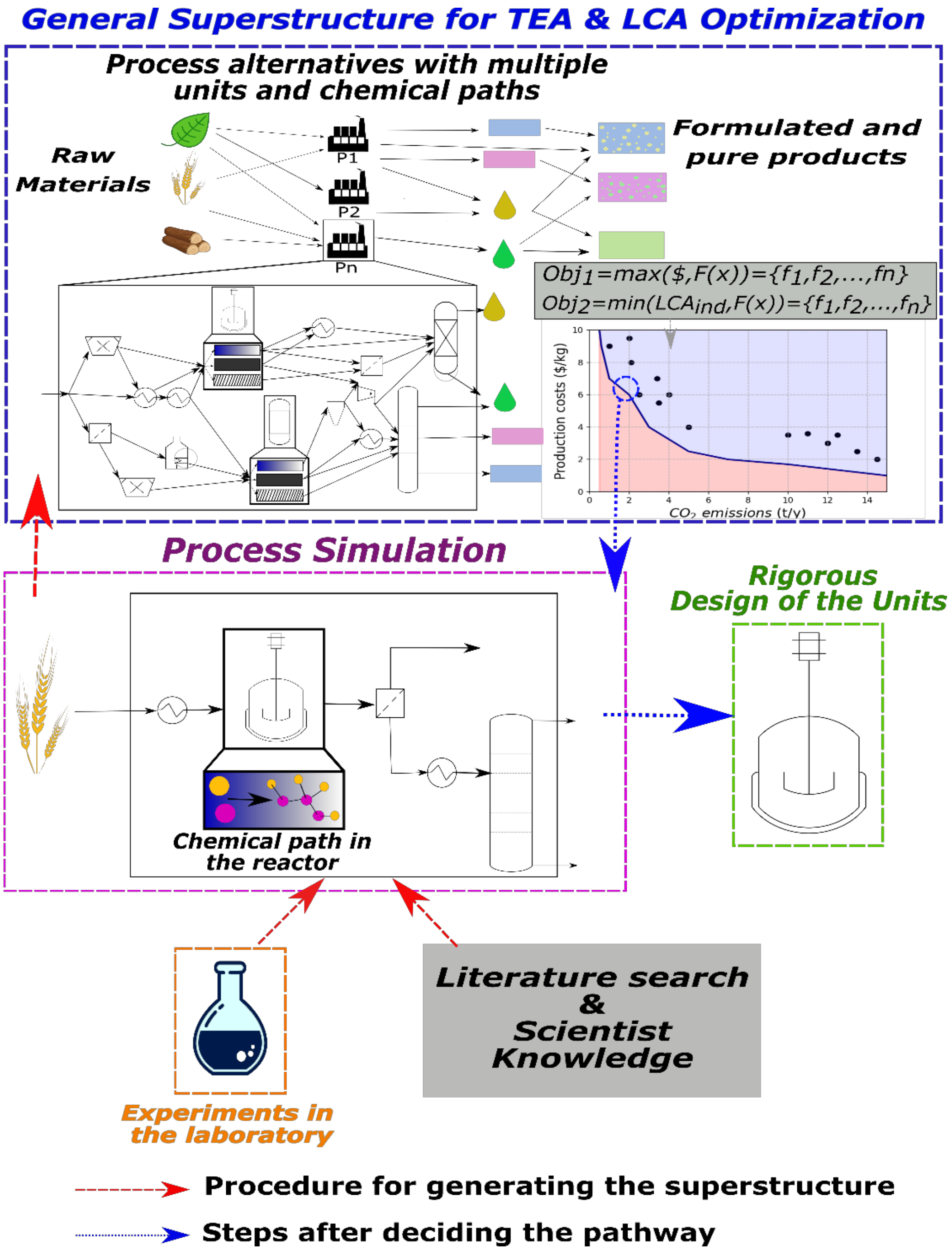


Figure 10. Process synthesis and design steps.

Process and product design. Once the process synthesis has determined the most promising alternative, process design is done (blue arrows in Figure 10). It typically needs more detailed information, such as detailed thermodynamic data, kinetic data, and more elaborated process models.[137-139] The data can be extensive and the phenomena complex, making mechanistic models unsuitable for process simulators. One could employ statistical models with multivariate analysis to determine the most critical variables that are employed in a surrogate model. Multivariable techniques include principal component analysis,[140, 141] partial least squares,[142] Bayesian minimization,[143, 144] and the Akaike's information criteria.[145, 146] Surrogate models include Principal Component Regression, polynomial correlations, Radial Basis Functions, or kriging/gaussian process models.[147, 148] One can size a reactor or a distillation column and reduce uncertainties using detailed mechanistic or surrogate models. This step entails a highly nonlinear programming optimization model to determine the optimal operating conditions, raw materials, and products.[149, 150] Uncertainties due to the heterogeneity of the biomass and markets can also be incorporated,[151] and heat integration can also be carried out. Process integration has been studied extensively in general for specific classes of problems (i.e., Pinch analysis[152-154] and optimal network of heat exchangers and water treatment under economic[153, 155-158] and environmental concerns[159]) but little of this has been applied to biomass conversion. The selection and optimization of catalysts and solvents remain a high toll due to the lack of descriptor-based models or big reactivity and thermodynamic data.

Determining a product with target properties can be achieved using Computer-Aided Molecular Design (CAMD) based on the functional groups or molecular structure.[160-162] CAMD combined with optimization can provide the optimum blend of biofuel. The problem, defined as a Mixed-Integer-Non-Linear (MINLP), is often simplified to a MILP.[163] CAMD has been formulated for various components (i.e., solvents[164], refrigerants[165]). However, product design has not been applied to biorefineries due to the lack of structure-property relations and knowledge about paths identified as gaps (Figure 2).

Rigorous process design. Computational fluid dynamics can optimize unit geometry and process conditions to more accurately estimate CAPEX. This step is common in process intensification. Intensification technologies can be profitable and implementable at an industrial scale.[166] In general, detailed process design has not been conducted due to the lack of detailed process models and commercialization efforts.

Integrated frameworks. Integrating product development with process synthesis and design should be the ultimate goal but this task is at embryonic stages. We provide some examples below.

- Multiple stages of process design. An early effort entailed the Integrated-Computer Aided System tool[167] that combines CAMD[168] and process design. The user specifies the material and reactions to rank the most promising processes and simulate them.[169]
- Product and process design.[170-174] These problems are numerically difficult due to including many nonlinear and integer variables, many components, and complex property estimation. Approaches to overcome the complexity include (1) surrogates[172] and clustering and averaging techniques;[175] (2) a hierarchical approach[175-177] to screen, using the most relevant material properties and processes, and determine the optimal molecule using more rigorous models; and (3) an iterative scheme by formulating a single optimization[170] or two optimizations[171] where

the "master problem" selects the optimal path and a detailed nonlinear optimizer determines the optimal process conditions.

- Supply chain. Even though the supply chain is beyond the scope of this perspective, it affects profitability and is an important barrier (Figure 2). A relevant study is the pooling problem,[178, 179] where the production sites and blending facilities are in different regions, and the optimal locations are determined. Extensions can include graph theory[180] or iterative optimization.[181] We advocate that this is an important frontier for the PSE community.

Barriers and opportunities. We envision that future biomass manufacturing should be supported by a seamless integration framework, as shown in Figure 11. One identifies the target products and properties using market analysis and uses text mining to transform these properties into possible products and technologies. Some ideas in this direction with limited applicability have been illustrated as part of the CAMD strategy. An automated laboratory can translate the available information into potential products and processes. Simulation software has been used extensively for process design at various levels of product maturity outlined above; yet, it is still an R&D exercise, hampering its extensive product and process optimization utilization. Data integration is usually done with specialized software, making integration with process simulators non-trivial. Numerical stability and convergence are still an art in graduate-level education, hindering the adaptation of the technology. Given uncertainty and assumptions, refinements that balance accuracy and time to generate data and integrate with process simulators and optimization software are necessary. One needs to go beyond new algorithms and build a computational environment that integrates existing capabilities easily and refines models and data adaptively. Flexible user interfaces integrating methodologies and software are lacking, leading to underutilized academic developments. Finally, databases and models of consistent and verified thermodynamic and kinetic data are crucial. Digitalization is a key enabler to accelerate this transition.

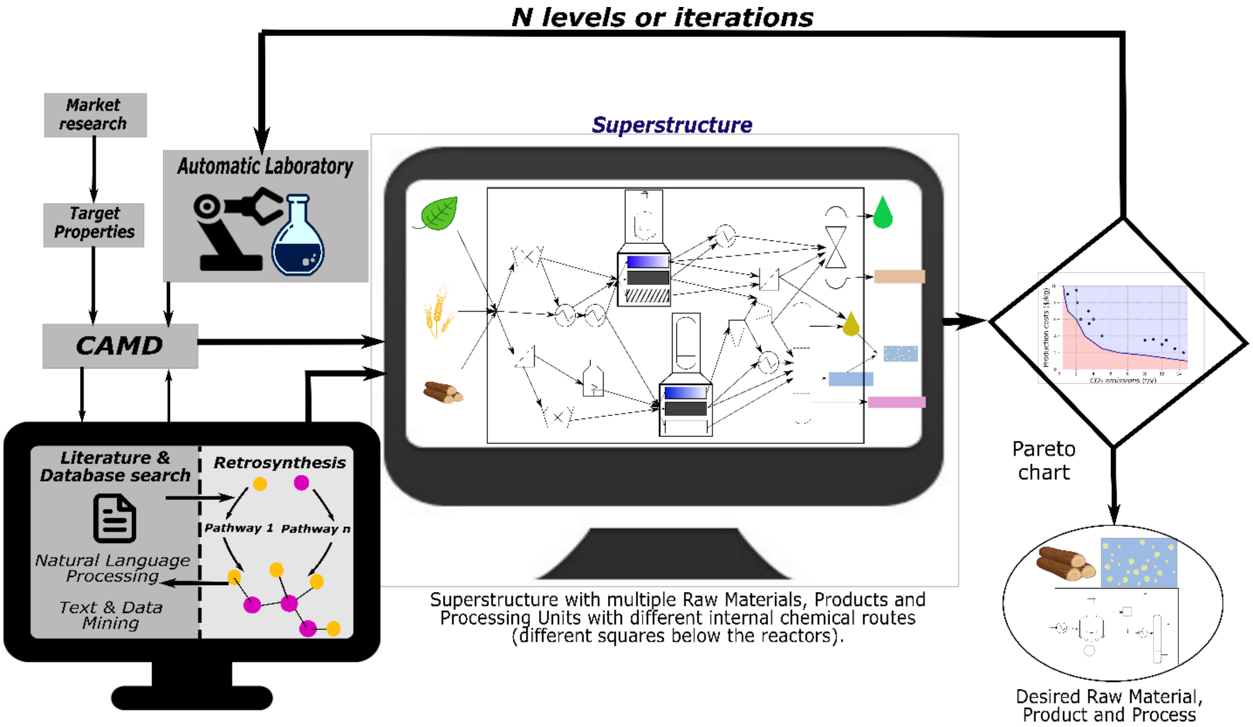


Figure 11. A framework integrating bench-scale digitalization, automatic laboratory experiments and multiscale simulations, and process design for TEA and LCA.

VI. Outlook

While tremendous advances in molecule discovery and catalyst development have been made, progress toward commercialization has been slow, despite the appealing TEA and LCA results. Part of the reason is the lack of an established supply chain, the significant investment risk, the high crude oil price variability, and the lack of supporting energy and climate policies. Data from the laboratory and multiscale simulations are produced slowly and result in publications and patents. Still, little of this information is harnessed to build quantitative models and predict. Digitalization can enable this transformation.

This perspective highlighted digitalization for biomass manufacturing and the challenges in integrating process optimization with economic, environmental, and social objectives for product selection and path, solvent, and catalyst discovery. The integration will require automatic literature search via natural language processing, machine learning tools to organize knowledge, CAMD concepts, process modeling, and optimization. Developing new artificial intelligence tools to extract information from text, images, and graphs using computer vision models will accelerate the identification of alternatives for superstructure optimization. The main challenges in the superstructure formulation will be the smart selection of products and the implementation of physical insights to reduce uncertainties in estimating costs and environmental impacts. An automated laboratory and simulation data via active learning methods can provide key thermochemistry and kinetics data for detailed, validated process models, product structure-property relations, and catalyst and solvent descriptor-performance correlations. The massiveness of unknown molecule space, reaction paths, operating conditions, catalysts, and solvents, along with the uncertainty arising from incomplete and noisy data, feedstock variability, and market fluctuations, make this a hard problem. Yet, most complex problems are inherently low dimensional. A hierarchical approach built on screening, data and model generation, refinement, and coupling across scales could lead to profitable products, processes, and catalytic routes. Such an approach can significantly reduce time to commercialization, minimize uncertainty and risk, and accelerate the path to sustainable manufacturing.

VII. Acknowledgements

The biomass chemistry concepts were supported as part of the Catalysis Center for Energy Innovation, an Energy Frontier Research Center funded by the US Dept. of Energy, Office of Science, Office of Basic Energy Sciences under award number DE-SC0001004. The data science concepts were supported from the Department of Energy's Office of Energy Efficient and Renewable Energy's Advanced Manufacturing Office under Award Number DE-EE0007888-9.5. The Delaware Energy Institute gratefully acknowledges the support and partnership of the State of Delaware toward the RAPID projects. The National Science Foundation (Award number CBET-213447) supported the manufacturing digitalization concepts.

VIII. Conflicts of Interest

There are no conflicts of interest.

IX. References

1. IPCC, *Global warming of 1.5ºC*, 2018.
2. EIA, *Manufacturing Energy and Carbon Footprint*, EIA, 2018.
3. Deloitte, *The 2030 decarbonization challenge: The path to the future.*, 2021.
4. I. R. E. Agency, *Global Energy Transformation - A Roadmap to 2050*, 2018.
5. I. R. E. Agency, *Global Energy Transformation: A Roadmap to 2050*, 2019.
6. E. Union, *Regulation of the European Parliament and of the Council on the use of renewable and low-carbon fuels in maritime transport and amending. Directive 2009/16/EC*, 2009.
7. J. A. Poveda-Giraldo, J. C. Solarte-Toro and C. A. C. Alzate, *Renew Sust Energ Rev*, 2021, **138**.
8. A. Hommes, H. J. Heeres and J. Yue, *Chemcatchem*, 2019, **11**, 4671-4708.
9. M. M. Han, X. Liu, X. S. Zhang, Y. Y. Pang, P. Xu, J. W. Guo, Y. D. Liu, S. G. Zhang and S. X. Ji, *Green Chem*, 2017, **19**, 722-728.
10. Z. H. Zhang and K. J. Deng, *Acs Catal*, 2015, **5**, 6529-6544.
11. J. J. Pacheco and M. E. Davis, *P Natl Acad Sci USA*, 2014, **111**, 8363-8367.
12. L. M. Vane, *Biofuel Bioprod Bior*, 2008, **2**, 553-588.
13. L. E. K. Consulting, EU's Green Deal sets out a bold ambition and proposes supportive revisions to three key policy directives, https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en, (accessed 18th January, 2022).

14. S. B. Liu, W. Q. Zheng, J. Y. Fu, K. Alexopoulos, B. Saha and D. G. Vlachos, *ACS Catal.*, 2019, **9**, 7679-7689.
15. C. L. Williams, C.-C. Chang, P. Do, N. Nikbin, S. Caratzoulas, D. G. Vlachos, R. F. Lobo, W. Fan and P. J. Dauenhauer, *ACS Catal.*, 2012, **2**, 935-939.
16. Y.-T. Cheng, J. Jae, J. Shi, W. Fan and G. W. Huber, *Angew. Chemie - Int. Ed.*, 2012, **51**, 1387-1390.
17. H. J. Cho, L. M. Ren, V. Vattipalli, Y. H. Yeh, N. Gould, B. J. Xu, R. J. Gorte, R. Lobo, P. J. Dauenhauer, M. Tsapatsis and W. Fan, *ChemCatChem*, 2017, **9**, 398-402.
18. O. A. Abdelrahman, D. S. Park, K. P. Vinter, C. S. Spanjers, L. Ren, H. J. Cho, D. G. Vlachos, W. Fan, M. Tsapatsis and P. J. Dauenhauer, *ACS Sustain. Chem. Eng.*, 2017, **5**, 3732-3736.
19. Y. C. Ji, A. Lawal, A. Nyholm, R. J. Gorte and O. A. Abdelrahman, *Catal. Sci. Technol.*, 2020, **10**, 5903-5912.
20. S. B. Liu, T. R. Josephson, A. Athaley, Q. P. Chen, A. Norton, M. Ierapetritou, J. I. Siepmann, B. Saha and D. G. Vlachos, *Sci. Adv.*, 2019, **5**, 8.
21. D. S. Park, K. E. Joseph, M. Koehle, C. Krumm, L. M. Ren, J. N. Damen, M. H. Shete, H. S. Lee, X. B. Zuo, B. Lee, W. Fan, D. G. Vlachos, R. F. Lobo, M. Tsapatsis and P. J. Dauenhauer, *ACS Central Science*, 2016, **2**, 820-824.
22. A. F. Sousa, C. Vilela, A. C. Fonseca, M. Matos, C. S. R. Freire, G.-J. M. Gruter, J. F. J. Coelho and A. J. D. Silvestre, *Polymer Chemistry*, 2015, **6**, 5961-5983.
23. S. Wang, L. Shuai, B. Saha, D. G. Vlachos and T. H. Epps, *ACS Central Science*, 2018, **4**, 701-708.
24. G. V. Research, *Lubricants Market Size, Share & Trends Analysis Report by Application (Industrial, Marine, Automotive, Aerospace), By Region (Asia Pacific, North America, Europe, MEA), And Sement Forecasts, 2021-2028*, 2021.
25. P. Newswire, The global lubricants market size is projected to reach USD 182.6 billion by 2025 from USD 157.6 billion in 2020, at a CAGR of 3.0%, https://www.prnewswire.com/news-releases/lubricants-market-size-worth-190-64-billion-globally-by-2028-at-2-7-cagr---exclusive-report-by-the-insight-partners-301459360.html, (accessed 18th January, 2022).
26. W. D. Cooper, R. C. Downing and J. B. Gray, presented in part at the International Compressor Engineering Conference, 1974.
27. K. Takigawa, S. I. Sandler and A. Yokozeki, *Int J Refrig*, 2002, **25**, 1014-1024.
28. S. B. Liu, R. Bhattacharjee, S. Li, A. Danielson, T. Mazal, B. Saha and D. G. Vlachos, *Green Chem.*, 2020, **22**, 7896-7906.
29. A. M. Norton, S. Liu, B. Saha and D. G. Vlachos, *Chemsuschem*, 2019, **12**, 4780-4785.
30. S. B. Liu, B. Saha and D. G. Vlachos, *Green Chem.*, 2019, **21**, 3606-3614.
31. E. O. Ebikade, S. Sadula, S. B. Liu and D. G. Vlachos, *Green Chem*, 2021, **23**, 10090-10100.
32. E. Ebikade, A. Athaley, B. Fisher, K. Yang, C. Q. Wu, M. G. Ierapetritou and D. G. Vlachos, *ACS Sustain. Chem. Eng.*, 2020, **8**, 8124-8136.
33. E. O. Ebikade, S. Sadula, Y. Gupta and D. G. Vlachos, *Green Chem*, 2021, **23**, 2806-2833.
34. Y. B. Shen, J.E. Hardy, M.A. Sarpong, R. Doyle, A.G. Cernak, T., *Nature Reviews Methods Primers*, 2021, **1**.
35. M. Trobe and M. D. Burke, *Angew Chem Int Edit*, 2018, **57**, 4192-4214.
36. B. R. Kowalski, *J Chem Inf Comp Sci*, 1975, **15**, 201-203.

37. A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders and R. Fushimi, *Acs Catal*, 2018, **8**, 7403-7429.
38. S. R. Kalidindi and M. De Graef, *Annu Rev Mater Res*, 2015, **45**, 171-193.
39. R. Lahana, *Drug Discov Today*, 2002, **7**, 898-900.
40. D. M. Lowe, DOI: https://doi.org/10.17863/CAM.16293, University of Cambridge, 2012.
41. H. Satoh and K. Funatsu, *J Chem Inf Comp Sci*, 1995, **35**, 34-44.
42. H. Satoh and K. Funatsu, *J Chem Inf Comp Sci*, 1996, **36**, 173-184.
43. C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *Acs Central Sci*, 2017, **3**, 1237-1245.
44. M. H. S. Segler and M. P. Waller, *Chem-Eur J*, 2017, **23**, 6118-6128.
45. P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas and T. Laino, *Chem Sci*, 2018, **9**, 6091-6098.
46. J. K. Juno Nam, *Journal*, 2016, DOI: arXiv:1612.09529.
47. J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *Acs Central Sci*, 2016, **2**, 725-732.
48. D. D. Ridley, *Information Retrieval: SciFinder and SciFinder Scholar*, John Wiley & Sons, 2002.
49. C. W. Coley, W. G. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem Sci*, 2019, **10**, 370-377.
50. V. Choudhary, S. H. Mushrif, C. Ho, A. Anderko, V. Nikolakis, N. S. Marinkovic, A. I. Frenkel, S. I. Sandler and D. G. Vlachos, *J Am Chem Soc*, 2013, **135**, 3997-4006.
51. M. Quantin, B. Hervy, F. Laroche and A. Bernard, *Procedia CIRP*, 2016, **50**, 436-441.
52. D. C. Jamison, *Current Protocols in Bioinformatics*, 2003, **00**, 9.2.1-9.2.29.
53. J. R. Lourenço, B. Cabral, P. Carreiro, M. Vieira and J. Bernardino, *Journal of Big Data*, 2015, **2**, 18.
54. K. Chodorow and M. Dirolf, *K. Chodorow and M. Dirolf, MongoDB: The Definitive Guide*, O'Rielly Media, Inc., 2010.
55. S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan and L. Cronin, *Science*, 2020, **370**, 101-+.
56. A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller and T. Laino, *Nat Commun*, 2020, **11**.
57. J. Guo, A. S. Ibanez-Lopez, H. Y. Gao, V. Quach, C. W. Coley, K. F. Jensen and R. Barzilay, *J Chem Inf Model*, 2021, **61**, 4124-4124.
58. D. M. Jurafsky, J.H:, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*, 2000.
59. S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit and J. Kim, *Journal of Chemical Information and Modeling*, 2018, **58**, 244-251.
60. R. Krishnamurthy, Y. Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan and H. Y. Zhu, *Sigmod Record*, 2008, **37**, 7-13.
61. E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178-&.
62. H. Patel, M. J. Bodkin, B. N. Chen and V. J. Gillet, *J Chem Inf Model*, 2009, **49**, 1163-1184.
63. C. D. Christ, M. Zentgraf and J. M. Kriegl, *J Chem Inf Model*, 2012, **52**, 1745-1756.
64. Howard Broughton, Peter Hunt and M. MacKey, *US Patent US20030182094A1*, 2003.
65. S. M. Mennen, C. Alhambra, C. L. Allen, M. Barberis, S. Berritt, T. A. Brandt, A. D. Campbell, J. Castanon, A. H. Cherney, M. Christensen, D. B. Damon, J. E. de Diego, S. Garcia-Cerrada, P. Garcia-Losada, R. Haro, J. Janey, D. C. Leitch, L. Li, F. F. Liu, P. C. Lobben, D. W. C. MacMillan, J. Magano, E. McInturff, S. Monfette, R. J. Post, D. Schultz,

B. J. Sitter, J. M. Stevens, J. I. Strambeanu, J. Twilton, K. Wang and M. A. Zajac, *Org Process Res Dev*, 2019, **23**, 1213-1242.

66. S. W. Krska, D. A. DiRocco, S. D. Dreher and M. Shevlin, *Accounts Chem Res*, 2017, **50**, 2976-2985.
67. C. J. Welch, *React Chem Eng*, 2019, **4**, 1895-1911.
68. D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429-434.
69. S. Shaabani, R. X. Xu, M. Ahmadianmoghaddam, L. Gao, M. Stahorsky, J. Olechno, R. Ellson, M. Kossenjans, V. Helan and A. Domling, *Green Chem*, 2019, **21**, 225-232.
70. C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Y. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, 557-+.
71. S. V. Ley, D. E. Fitzpatrick, R. J. Ingham and R. M. Myers, *Angew Chem Int Edit*, 2015, **54**, 3449-3464.
72. M. Peplow, *Nature*, 2014, **512**, 20-22.
73. L. M. Roch, F. Haase, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein and A. Aspuru-Guzik, *Plos One*, 2020, **15**.
74. Y. C. Wang, T. Vlachos, D. , *NEXTorch: A Design and Bayesian Optimization Toolkit for Chemical Sciences and Engineering*, 2021.
75. F. Hase, M. Aldeghi, R. J. Hickman, L. M. Roch and A. Aspuru-Guzik, *Appl Phys Rev*, 2021, **8**.
76. J. Wu and P. I. Frazier, *Adv Neur In*, 2016, **29**.
77. *GPyOpt: A Bayesian OPtimization framework in python*, 2021.
78. N. v. d. H. Knudde, J. Dhaene, T. Couckuyt, I. , *GPflowOpt: A Bayesian Optimization Library using TensorFlow*, 2017.
79. A. P. Paleyes, M. Mahsereci, M. McCollum, C. Lawrence, N. González, J. , presented in part at the Second workshop on machine learning and the physical sciences, NeurIPS 2019, 2019.
80. X. N. Li and A. Kraslawski, *Chem Eng Process*, 2004, **43**, 583-594.
81. A. f. t. A. o. C. Engineering, *Cost estimate classification system-As applied in engineering procurement, and construction for the process industries.*, 2005.
82. D. R. Batsy, C. C. Solvason, N. E. Sammons, V. Chambost, D. L. Bilhartz, M. R. Eden, M. M. El-Halwagi and P. R. Stuart, *Green Chem Chem Eng*, 2013, 3-35.
83. J. Moncada, V. Aristizabal and C. A. Cardona, *Biochem Eng J*, 2016, **116**, 122-134.
84. A. O. Bjorn, M. Molin, C. Hauschild, M.Z., 2017.
85. A. C. Kokossis and A. D. Yang, *Comput Chem Eng*, 2010, **34**, 1397-1405.
86. K. U. Klatt and W. Marquardt, *Comput Chem Eng*, 2009, **33**, 536-550.
87. L. Mencarelli, Q. Chen, A. Pagot and I. E. Grossmann, *Comput Chem Eng*, 2020, **136**.
88. J. M. Douglas, *Aiche J*, 1985, **31**, 353-362.
89. R. C. Baliban, J. A. Elia and C. A. Floudas, *Energ Environ Sci*, 2013, **6**, 267-287.
90. J. Gong and F. Q. You, *Aiche J*, 2014, **60**, 3195-3210.
91. V. Andiappan, A. S. Y. Ko, V. W. S. Lau, L. Y. Ng, R. T. L. Ng, N. G. Chemmangattuvalappil and D. K. S. Ng, *Aiche J*, 2015, **61**, 132-146.
92. J. Gong, D. J. Garcia and F. Q. You, *Acs Sustain Chem Eng*, 2016, **4**, 3160-3173.
93. E. Zondervan, M. Nawaz, A. B. de Haan, J. M. Woodley and R. Gani, *Comput Chem Eng*, 2011, **35**, 1752-1766.

94. J. Han, S. M. Sen, J. S. Luterbacher, D. M. Alonso, J. A. Dumesic and C. T. Maravelias, *Comput Chem Eng*, 2015, **81**, 57-69.
95. M. O. Bertran, R. Frauzem, A. S. Sanchez-Arcilla, L. Zhang, J. M. Woodley and R. Gani, *Comput Chem Eng*, 2017, **106**, 892-910.
96. V. Pham and M. El-Halwagi, *Aiche J*, 2012, **58**, 1212-1221.
97. M. Martin and I. E. Grossmann, *Ind Eng Chem Res*, 2013, **52**, 3044-3064.
98. V. Aristizabal-Marulanda, C. A. C. Alzate and M. Martin, *Comput Chem Eng*, 2019, **126**, 35-53.
99. M. S. T. Peters, K.D. West, R.E., *Plant Design and Economics for Chemical Engineers*, 2003.
100. R. S. Turton, J. Bhattacharyya, D. Whiting, W., *Analysis, Synthesis, and Design of Chemical Processes*, 2018.
101. Matches, Process Equipment Cost Estimates, https://www.matche.com/equipcost/Default.html (accessed 18th January, 2022).
102. Aspentech, *Aspen Economic Analyser*, 2021.
103. P. Forster, V. Ramaswamy, P. Artaxo, T. Berntsen, R. Betts, D. W. Fahey, J. Haywood, J. Lean, D. C. Lowe, G. Myhre, J. Nganga, R. Prinn, G. Raga, M. Schulz and R. Van Dorland, *Ar4 Climate Change 2007: The Physical Science Basis*, 2007, 129-234.
104. U. S. E. P. Agency, *Inventory of U.S. Greenhouse Gases Emissions and Sinks: 1990-2014*, 2016.
105. Ecoinvent, https://ecoinvent.org/, (accessed 18th January, 2022).
106. OECD, *Handbook on Constructing Composite Indicators: Methodology and User Guide.*, 2008.
107. S. Van Schoubroeck, M. Van Dael, S. Van Passel and R. Malina, *Renew Sust Energ Rev*, 2018, **94**, 115-126.
108. G. Sandin, F. Royne, J. Berlin, G. M. Peters and M. Svanstrom, *J Clean Prod*, 2015, **93**, 213-221.
109. E. Gnansounou, P. Vaskan and E. R. Pachon, *Bioresource Technol*, 2015, **196**, 364-375.
110. J. C. Bare, *Journal of Industrial Ecology*, 2008, DOI: https://doi.org/10.1162/108819802766269539.
111. R. Shi and J. S. Guest, *Acs Sustain Chem Eng*, 2020, **8**, 18903-18914.
112. U. N. E. Programme, *Guidelines for Social Life Cycle Assessment of Products and Organizations.* , 2020.
113. E. Cadena, F. Rocca, J. A. Gutierrez and A. Carvalho, *J Clean Prod*, 2019, **238**.
114. A. Azapagic and S. Perdan, *Process Saf Environ*, 2000, **78**, 243-261.
115. U. N. D. Program, *Human development reports.* , 2020.
116. B. E. Dale and R. G. Ong, *Biofuel Bioprod Bior*, 2014, **8**, 487-503.
117. J. A. Schaidle, C. J. Moline and P. E. Savage, *Environ Prog Sustain*, 2011, **30**, 743-753.
118. M. Martin, *Energy*, 2016, **114**, 833-845.
119. V. S.-T. Aristizabal-Marulanda, J.C. Alzate, C.A. , *Bioresource Technology Reports*, 2020, **9**.
120. C. A. Floudas, *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications*, 1995.
121. A. M. Niziolek, O. Onel, J. A. Elia, R. C. Baliban, X. Xiao and C. A. Floudas, *Ind Eng Chem Res*, 2014, **53**, 17002-17025.

122. C. Vadenbo, G. Guillen-Gosalbez, D. Saner and S. Hellweg, *Resour Conserv Recy*, 2014, **89**, 41-51.
123. H. S. Bamufleh, J. M. Ponce-Ortega and M. M. El-Halwagi, *Clean Technol Envir*, 2013, **15**, 185-197.
124. Z. Zore, L. Cucek and Z. Kravanja, *Comput Chem Eng*, 2017, **102**, 139-155.
125. K. A. Pyrgakis and A. C. Kokossis, *Comput Chem Eng*, 2019, **122**, 326-355.
126. H. Kasivisvanathan, D. K. S. Ng, G. Poplewski and R. R. Tan, *Ind Eng Chem Res*, 2016, **55**, 4035-4044.
127. P. Cheali, A. Quaglia, K. V. Gernaey and G. Sin, *Ind Eng Chem Res*, 2014, **53**, 6021-6032.
128. I. E. Grossmann, R. M. Apap, B. A. Calfa, P. Garcia-Herreros and Q. Zhang, *Comput Chem Eng*, 2016, **91**, 3-14.
129. C. G. Li, I.E. , *Frontiers in Chemical Engineering*, 2021, DOI: https://doi.org/10.3389/fceng.2020.622241
130. M. Rizwan, M. Zaman, J. H. Lee and R. Gani, *Comput Chem Eng*, 2015, **82**, 362-373.
131. N. V. Sahinidis, *Comput Chem Eng*, 2004, **28**, 971-983.
132. A. Bhosekar, A. Athaley and M. Ierapetritou, *Ind Eng Chem Res*, 2021, **60**, 12956-12969.
133. A. Geraili and J. A. Romagnoli, *Aiche J*, 2015, **61**, 3208-3222.
134. C. L. Sy, A. T. Ubando, K. B. Aviso and R. R. Tan, *J Clean Prod*, 2018, **170**, 496-509.
135. J. E. M.-R. Santibanez-Aguilar, R. Gonzalez-Campos, J.B. Ponce-Ortega, J. M., *Chemical Engineering Transactions*, 2015, **45**, 367-372.
136. R. M. Apap and I. E. Grossmann, *Comput Chem Eng*, 2017, **103**, 233-274.
137. J. A. Quintero and C. A. Cardona, *Ind Eng Chem Res*, 2011, **50**, 6205-6212.
138. M. B. Nikoo and N. Mahinpey, *Biomass Bioenerg*, 2008, **32**, 1245-1254.
139. K. Rajendran, H. R. Kankanala, M. Lundin and M. J. Taherzadeh, *Bioresource Technol*, 2014, **168**, 7-13.
140. C. A. Aceves-Lara, E. Latrille, P. Buffiere, N. Bernet and J. P. Steyer, *Chem Eng Process*, 2008, **47**, 1968-1975.
141. H. S. Choi, Y. S. Choi and H. C. Park, *Renew Energ*, 2012, **42**, 131-135.
142. S. J. You, N. Park, E. D. Park and M. J. Park, *J Ind Eng Chem*, 2015, **21**, 350-355.
143. G. Schwarz, *Ann Stat*, 1978, **6**, 461-464.
144. J. Remon, F. Broust, G. Volle, L. Garcia and J. Arauzo, *Int J Hydrogen Energ*, 2015, **40**, 5593-5608.
145. H. Akaike, *IEEE Transactions on Automatic Control*, 1974, **AC-19**, 716-723.
146. L. Li, Y. Y. Wang, J. T. Xu, J. R. V. Flora, S. Hoque and N. D. Berge, *Bioresource Technol*, 2018, **262**, 284-293.
147. A. Bhosekar and M. Ierapetritou, *Comput Chem Eng*, 2018, **108**, 250-267.
148. A. H. Sánchez, B. Martín, M. , in *Process Systems Engineering for Biofuels Development*, 2020, pp. 49-83.
149. M. Martin and I. E. Grossmann, *Aiche J*, 2011, **57**, 3408-3428.
150. B. Hernandez, E. Leon and M. Martin, *Chem Eng Res Des*, 2017, **121**, 163-172.
151. R. Morales-Rodriguez, A. S. Meyer, K. V. Gernaey and G. Sin, *Comput Chem Eng*, 2012, **42**, 115-129.
152. J. J. Klemes and Z. Kravanja, *Curr Opin Chem Eng*, 2013, **2**, 461-474.
153. J. J. Klemes, P. S. Varbanov, T. G. Walmsley and X. X. Jia, *Renew Sust Energ Rev*, 2018, **98**, 439-468.
154. R. Smith, *Chemical Process Design and Integration*, John Wiley & Sons., 2016.

155. C. Fu and T. Gundersen, *Aiche J*, 2015, **61**, 3770-3785.
156. B. Linnhoff and J. R. Flower, *Aiche J*, 1978, **24**, 633-642.
157. T. F. Yee, I. E. Grossmann and Z. Kravanja, *Comput Chem Eng*, 1990, **14**, 1185-1200.
158. E. Ahmetovic and I. E. Grossmann, *Aiche J*, 2011, **57**, 434-457.
159. S. Perry, J. Klemes and I. Bulatov, *Energy*, 2008, **33**, 1489-1497.
160. N. D. Austin, N. V. Sahinidis and D. W. Trahan, *Chem Eng Res Des*, 2016, **116**, 2-26.
161. R. Gani, *Chem Eng Res Des*, 2004, **82**, 1494-1504.
162. L. Zhang, H. T. Mao, Q. L. Liu and R. Gani, *Curr Opin Chem Eng*, 2020, **27**, 22-34.
163. C. D. Maranas, *Ind Eng Chem Res*, 1996, **35**, 3403-3414.
164. S. Jonuzaj and C. S. Adjiman, *Chem Eng Sci*, 2017, **159**, 106-130.
165. N. V. Sahinidis, M. Tawarmalani and M. R. Yu, *Aiche J*, 2003, **49**, 1761-1775.
166. V. Francia, L. Martin, A. E. Bayly and M. J. H. Simmons, *Chem Eng Sci*, 2015, **134**, 399-413.
167. R. Gani, F. Chen, C. Jaksland, P. Harper and M. Hostrup, *Setac Sp P*, 1997, 143-159.
168. R. Gani, B. Nielsen and A. Fredenslund, *Aiche J*, 1991, **37**, 1318-1332.
169. A. K. Tula, M. R. Eden and R. Gani, *Comput Chem Eng*, 2015, **81**, 245-259.
170. H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Siougkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, *Nat Chem*, 2013, **5**, 952-957.
171. M. Folic, C. S. Adjiman and E. N. Pistikopoulos, *Comp Aid Ch*, 2004, **18**, 175-180.
172. J. Burger, V. Papaioannou, S. Gopinath, G. Jackson, A. Galindo and C. S. Adjiman, *Aiche J*, 2015, **61**, 3249-3269.
173. N. Mac Dowell, A. Galindo, G. Jackson and C. S. Adjiman, *Comput-Aided Chem En*, 2010, **28**, 1231-1236.
174. Y. S. Lee, E. J. Graham, A. Galindo, G. Jackson and C. S. Adjiman, *Comput Chem Eng*, 2020, **136**.
175. A. I. Papadopoulos, P. Seferlis and P. Linke, *Chem Eng Sci*, 2017, **159**, 154-176.
176. P. Limleamthong, M. Gonzalez-Miquel, S. Papadokonstantakis, A. I. Papadopoulos, P. Seferlis and G. Guillen-Gosalbez, *Green Chem*, 2016, **18**, 6468-6481.
177. A. I. Papadopoulos, S. Badr, A. Chremos, E. Forte, T. Zarogiannis, P. Seferlis, S. Papadokonstantakis, A. Galindo, G. Jackson and C. S. Adjiman, *Mol Syst Des Eng*, 2016, **1**, 313-334.
178. R. Misener and C. A. Floudas, *Appl Comput Math-Bak*, 2009, **8**, 3-22.
179. N. Julka, I. Karimi and R. Srinivasan, *European Symposium on Computer Aided Process Engineering - 12*, 2002, **10**, 895-900.
180. A. M. Sampat, E. Martin, M. Martin and V. M. Zavala, *Comput Chem Eng*, 2017, **104**, 296-310.
181. M. Taifouris, M. Martin, A. Martinez and N. Esquejo, *Comput Chem Eng*, 2021, **152**.