In silico exploration of metabolite-derived soft materials using a chemical reaction network: what is possible?

Shruti Iyer[†] and Nicholas E. Jackson[*,†,‡]

[†]*Department of Chemistry, University of Illinois, Urbana-Champaign, Urbana, IL 61801*

[‡]*Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

E-mail: jacksonn@illinois.edu

Abstract

Future soft materials and polymer chemistries will require innovative non-petroleum sourcing pathways to thrive in a sustainable economy. While leveraging microbial metabolites derived from biological feedstocks possesses high potential in many avenues of chemical development, the applicability of this paradigm to the specifics of soft materials chemistry is unclear. Here, we construct a chemical reaction network based on databases of common microbial metabolites and the USPTO reaction set to examine what is possible in the chemical space of metabolite-derived chemistries of relevance to soft materials. We observe that the accessible chemical space of our chemical reaction network possesses strong microbe-specific chemical diversity, and that this space saturates rapidly within three synthetic steps applied to the original microbial metabolites. Importantly, we show that the chemical space accessible from metabolite precursors possesses significant overlap with existing petrochemical building blocks, known and proposed synthetically feasible polymer monomers, and the chemical space of common organic semiconductors, and redox active materials. The biases induced by the metabolite and reaction databases that parameterize our reaction network are analyzed as a function of chemical functional groups, and pathways towards broader sets of chemistries and reactions are outlined. This work introduces a computational framework for exploring a novel paradigm of soft materials discovery with the potential to accelerate the identification of soft materials relevant to metabolic engineering targets and non-petroleum sourcing pathways for existing soft materials.

Introduction

Over the past century, petroleum has been used to produce synthetic building blocks (e.g. ethylene, propylene, benzene, toluene, and xylene) for a wide array of organic chemicals and materials prevalent in fertilizers, fine chemicals, polymers, plastics, pharmaceuticals, detergents, food additives, electronics, sports equipment, clothing, dyes and agrochemicals.[1,2] This dependence on petrochemical derivatives has exacerbated the depletion of finite resources, driven global climate change, and undermined sustainable development.[2,3] Alternative feedstocks and sourcing pathways are critically necessary to move towards a more sustainable chemical and materials economy.

The engineering of microorganisms, designed and optimized to produce chemicals from renewable resources, presents a sustainable alternative to the petrochemical-based synthesis of commodity chemicals.[4,5] However, naturally occurring microorganisms are typically not optimized for the efficient uptake of renewable carbon sources or the high-yield production of target chemicals.[4,6] Through metabolic engineering, these organisms can be reprogrammed to utilize cost-effective renewable substrates and synthesize desired chemicals, including those outside their native metabolic pathways. Metabolic engineering involves first selecting a target product with high demand and promising applications, alongside a host strain that can be engineered to achieve high titer, rate, and yield of the desired product. This is followed by a design-build-test-learn cycle, where native or non-native biochemical pathways are constructed to convert substrates into the target product.[5,7] Key genetic strategies employed to enhance metabolite production for industrial-scale synthesis include increasing precursor supply, optimizing bottleneck enzymes, modifying gene regulation through point mutations that affect transcriptional or allosteric control, and minimizing byproduct formation.[8,9]

While metabolic engineering offers significant potential for sustainable chemical production, the use of hybrid approaches that integrate traditional chemical synthesis steps with metabolic engineering have the potential to augment the discovery of alternative chemical and materials sourcing pathways. Here, biological reactions refer to those which occur within microorganisms via metabolic enzymes and transporters or through enzyme-catalyzed processes in vitro, while chemical reactions involve non-biological methods employing catalysts, solvents, acids, or heat.[10] Biosynthetic methods, enabled by highly selective and efficient enzymes, excel at producing complex natural products while minimizing purification and protection steps. However, they are limited by a narrow reaction repertoire and challenges in engineering pathways for non-natural compounds. In contrast, chemical synthesis offers versatility in modifying diverse molecular structures, but can be less selective, often requiring many synthetic steps and generating significant waste. Combining the complementary strengths of these approaches has the potential to enable the efficient production of complex molecules, leveraging the precision of biosynthesis and the flexibility of chemical synthesis to address a broader range of industrial challenges.[11,12]

Despite significant advancements in this area, most research has focused on finding new pathways to a limited set of well-established, industrially significant chemicals. Typical targets for metabolic engineering are selected based on market demand, societal needs, and the feasibility of biological production. High-demand chemicals or those with promising future applications, including several bulk chemicals, fine chemicals, biofuels, polymers and natural products, are often prioritized.[5] Selection criteria emphasize cost-effectiveness and competitive viability, particularly for bulk chemicals, where achieving high titer, yield and productivity is critical.[6,13,14] Some of the key targets in metabolic engineering include the top biobased molecules identified by the U.S. Department of Energy (DOE), such as succinic acid and 2,5-furan dicarboxylic acid, which are derived from biomass and serve not only as sustainable alternatives to petrochemicals but also as starting points for novel biobased chemicals with enhanced functional properties.[15–18] To achieve desired targets, an integrated bio- and chemo-synthetic approach often begins with metabolic engineering to produce smaller building block chemicals, which are then further transformed through chemical synthesis into the final target molecule. A notable contribution by Lee et al.[10] involved developing a comprehensive metabolic map integrating biological and chemical reactions to facilitate the production of industrial chemicals from renewable resources, underscoring the potential of combining these methodologies.

Despite this exciting potential, the applicability of such sourcing and discovery paradigms to the broader space of soft materials chemistry remains unclear. Impactful developments have begun to emerge in the arena of commodity polymers, where bio-based sourcing pathways have shown promise in replacing petroleum derived precursors. A notable example of this potential for soft materials chemistry is in the production of polyethylene terephthalate (PET), one of the most extensively used polymers in the world, which is traditionally derived from petrochemical building blocks. Monoethylene glycol, one of its monomers, can now be produced by combining biological processes (fermentation of bioethanol) and chemical conversion (bioethanol to monoethylene glycol).[15] Similarly, the other monomer, terephthalic acid (TPA), can be produced by chemically converting biomass to p-xylene, followed by its bioconversion to TPA using engineered microorganisms, achieving efficiencies comparable to traditional chemical processes.[19] To build on the momentum for such transformations in the soft materials field, computational discovery frameworks are needed to identify other existing soft materials chemistries that would be amenable to sourcing modifications via bio-based and metabolic pathways. Moreover, beyond the immediate interest in replacing sourcing pathways of known soft materials, there is the vast untapped potential to expand the functionality of soft materials by leveraging a wider range of biologically derived metabolites as precursors for new, high performing materials, for which exploratory computational frameworks are also critically lacking.

In this study, we integrate bio- and chemo-synthesis by leveraging microbial metabolites as precursor molecules and subjecting them to chemically viable transformations derived from the USPTO patent chemical reaction database.[20] The focus is specifically on replacing petroleum-derived feedstocks for soft materials applications with sustainable, biologically sourced substrates, forming the basis of a chemical reaction network constructed from microbial metabolites. While other components of chemical reactions, such as reaction conditions, solvents and reagents are critical considerations, we concentrate on the substrate sourcing as a foundational step in this work. Two key distinctions define our approach. First, we consider only metabolites natively produced by the organism, capitalizing on their evolved functional compatibility within the metabolic framework and avoiding the additional complexity of engineering new biochemical pathways or enzymes. Second, our core metabolic substrates include both simple molecular building blocks and more complex, underexplored starting materials, hypothesizing that the latter could lead to functionally diverse and useful products in fewer reaction steps. Our investigation examines how the complexity and diversity of the metabolite-derived chemical network evolves with successive reaction steps, enabling an assessment of the breadth of chemistries accessible from these metabolite-derived reactant molecules. The resulting chemical space is then analyzed for its potential in materials applications. By moving beyond conventional targets such as bulk chemicals, fine chemicals, and biofuels, we explore the applicability of metabolite-derived molecules in areas such as polymeric materials, organic semiconductors, photovoltaics and redox-active materials, expanding the scope of sustainable, high-value products derived from biologically sourced precursors.

Methods

In this section, we describe the methods underlying the preparation of the metabolite database and generation of the extended metabolite reaction network (EMRN).

Metabolite dataset preparation

The core of the EMRN is formed from existing databases of microbial metabolic products with known biosynthetic pathways. We selected three commonly studied organisms: *Escherichia coli*, *Saccharomyces cerevisiae* and *Pseudomonas aeruginosa*. *E. coli* is a gram-negative, rod-shaped, facultative anaerobe found in the mammalian gut. The E. coli Metabolome Database (ECMDB)[21,22] contains detailed information on approximately 3,755 small molecules. *S. cerevisiae*, or baker's yeast, is a single-celled eukaryote used in biofuel production, winemaking, baking, and brewing. The Yeast Metabolome Database (YMDB)[23,24] includes data on 870 metabolites. *P. aeruginosa* is an encapsulated, gram-negative, aerobic–facultatively anaerobic rod shaped bacterium, well-known for its ubiquity and adaptability. The P. aeruginosa Metabolome Database (PAMDB)[25] contains information on about 4,370 metabolites. The metabolite molecule SMILES strings were extracted from the organism metabolite databases, combined, and filtered based on the following criteria: molecule SMILES strings should (i) only contain atoms frequently appearing in organic compounds (e.g. H, B, C, N, O, F, Al, Si, P, S, Cl, and Br), (ii) not contain isotopes (e.g. 13C), (iii) not contain explicit hydrogens (e.g. [H]OCCO), (iv) not contain non-zero formal charges (e.g. COC(=O)c1ccccc1[Br+]c1ccccc1), (v) not contain a period punctuation mark to include more than one molecule (e.g. Br.Oc1cc(on1)C1CCNCC1), and (vi) not contain atom-map indices (e.g. [H:4][n:3]1[cH:2][cH:1][c:6]([CH2:7][Cl:8])[n:5]1).[26] After filtering and canonicalizing the SMILES strings, we identified 4,909 unique metabolite molecules across the three organisms, forming the core of our metabolite network.

Preparing the USPTO Dataset for Reaction Search with the Metabolites as Core Reactants

The USPTO grants database[20] consists of 1.8 million chemical reactions extracted from United States patents published between 1976 and 2016, presented as reaction SMILES strings in the general format of either "reactant>reagent>product" or "reactant-reagent>>product". In databases derived from data mining, differentiating between reactants and reagents on the non-product sides of reactions is often unclear. This distinction is critical for our study, as we need to compare the metabolite molecules with the reactant molecules in the USPTO database to identify reactions involving these metabolites as reactants. To prepare the USPTO database for reaction selection, we performed reaction role mapping for the non-product molecules using a data-driven approach established by Schneider et al,[27] employing fingerprint-based method instead of the traditional, time-consuming atom-to-atom mapping. This method assigns reaction roles - reactant, reagent, or product - by comparing combined reactant/reagent fingerprints with product fingerprints to determine atomic changes. This process resulted in a modified reaction database, with each USPTO reaction having reactants, reagents and products delineated separately.

Extracting Reactions from the USPTO Dataset Given an Initial Metabolite Reactant Set

To generate the EMRN, we implemented a search algorithm that compares the SMILES of metabolite molecules with the reactant SMILES of each reaction in the USPTO database. Upon finding an exact match, the corresponding reaction is stored for that specific metabolite, identifying it as a synthetic pathway utilizing the metabolite as a reactant. This single-step reaction is considered one "reaction round". To generate further reaction rounds, the search algorithm is iteratively applied to the unique product SMILES strings from previous rounds, treating them as reactants for the next round. In the first reaction round, the data is stored with "Metabolite-derived reactant" and "Product" columns, where the "Metabolite-derived reactant" column contains all metabolite molecules with exact matches in the USPTO database, and the "Product" column contains all product molecules from the reactions, mapped to their respective metabolites. For subsequent reaction rounds, the unique molecules in the "Product" column of the previous round are used as the potential "Metabolite-derived reactant" molecules for the next round.

To refine the reaction network, a series of filtering steps were applied. First, we focused on reactions where the metabolite-derived precursor serves as the primary reactant. This was accomplished by retaining only reactions in which the molecular weight of the metabolite-based precursor is greater than that of any other reactant. Additionally, given that the reaction database contains only reaction SMILES and lacks information regarding the relative abundance or yield of the products, we simplified the selection criteria for reaction products. Specifically, reactions that produced a single product were included, as well as those that produced "n" products, where "n−1" of the products contained fewer than three heavy atoms. Molecules with fewer than three heavy atoms are represented as minor byproducts, such as water, ammonia, or other small molecules. Lastly, reactions exhibiting identical molecules on both the reactant and product sides were excluded from the analysis to eliminate trivial cases which do not contribute meaningfully to pathway exploration or design.

In the first reaction round, out of 4,909 metabolite molecules, 335 had exact matches as reactants in the USPTO database, leading to the formation of 2,800 unique products. In the second round, 1,295 of these products had exact matches, generating 8,435 unique products. In the third round, 3,803 of these products served as reactants, forming 15,967 unique products. The fourth round had 7,138 metabolite-derived reactant molecules, resulting in 23,249 products. In the fifth round, 10,192 metabolite-derived reactant molecules led to the formation of 30,918 unique products. The increase in the number of unique products from round four to round five was approximately 32%, compared to the 551% increase observed from the initial metabolites. Provided the increasing saturation of the EMRN as a function of reaction round, along with the anticipated positive correlation of cost with the number synthetic steps, we concluded the expansion of the reaction network after five rounds. All datasets and codes necessary for the reproduction of the results in this paper can be found in our GitHub (https://github.com/TheJacksonLab/ExtendedMetaboliteReactionNetwork, https://doi.org/10.5281/zenodo.14719996).

Results and Discussion

Characterization and Analysis of the Metabolite Database

Understanding the metabolite chemical space is crucial for comprehending the diversity of biochemical pathways across organisms and identifying unique structural characteristics that can inform the reactivity landscape of these molecules. The chemical space defined by the the metabolite pool comprising 4,909 unique compounds, drawn from *Escherichia coli* (1,964 metabolites), *Saccharomyces cerevisiae* (214 metabolites), *Pseudomonas aeruginosa* (1,522 metabolites) and the set of molecules common to all three organisms (1,209 metabolites), was visualized according to their structural characteristics using t-SNE (Fig. 1) with a Morgan fingerprint, the details of which are provided in the Supplementary Information (SI). While metabolites from different organisms share some overlap in chemical space due to canonical metabolic outputs (e.g. amino acids, carbohydrates, lipids, nucleotides and sugar phosphates[28,29]), each organism occupies distinct regions of chemical space, highlighting the intrinsic capability of different organisms to access diverse structural domains. Detailed analysis of the functional group distributions by organism class is provided in the SI - Fig. S1). More generally, this suggests the possibility of broadening the coverage of chemical space via the inclusion of other microorganisms in the future. To further characterize the differences between metabolites from different organisms, we also performed ground and excited-state DFT calculations at the wB97X-D3/def2-SVP, details of which are provided in the SI.

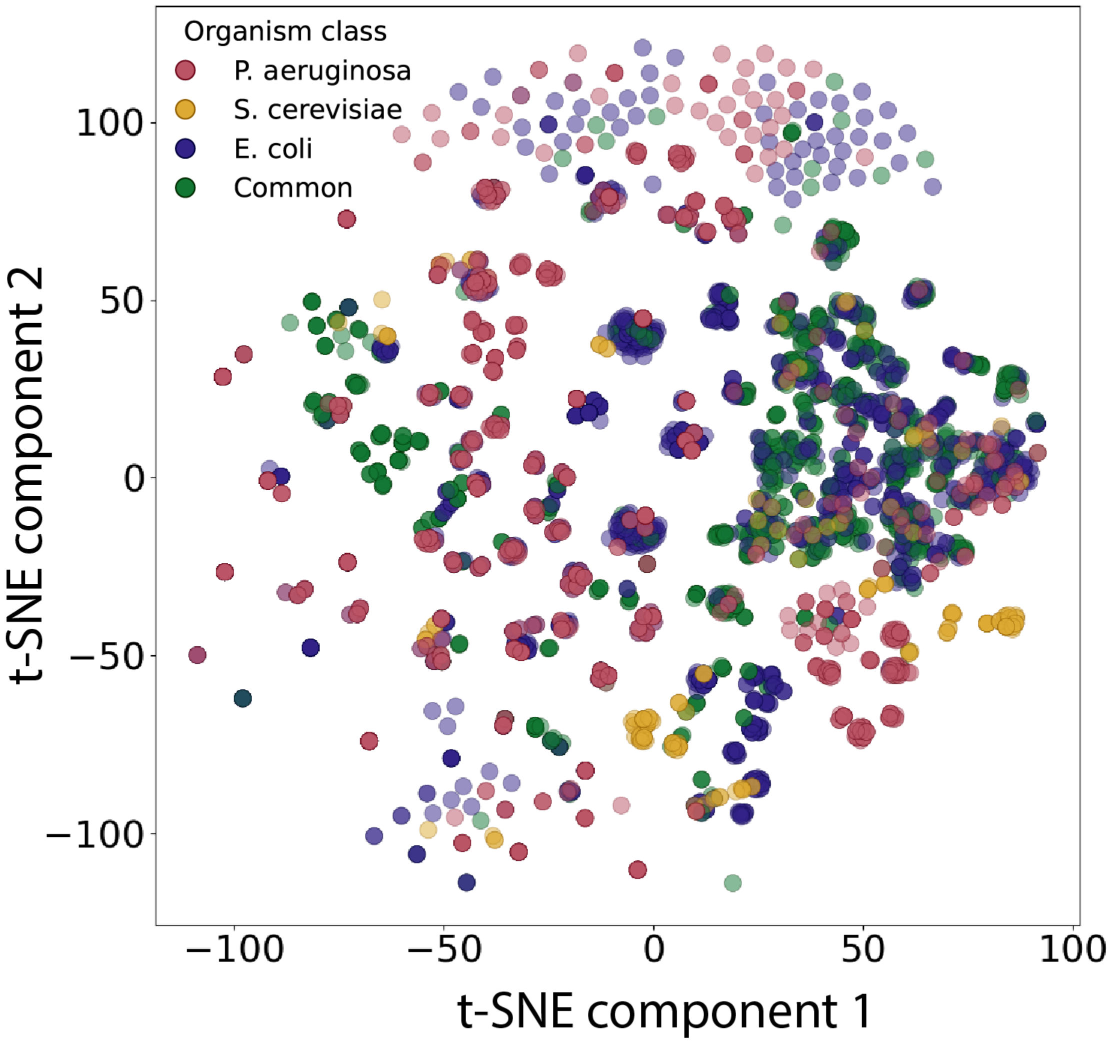

Figure 1: t-SNE plot of all 4,909 metabolite molecules, colored by organism class (*E. coli, S. cerevisiae, P. aeruginosa* and the molecules that were 'common' among the organisms).

Microbial metabolites exhibit substantial structural diversity and complexity, positioning them as promising precursors for the synthesis of complex materials. To quantitatively assess molecular complexity across the metabolite chemical space, two metrics were employed: absolute molecular weight as a proxy for structural complexity, and the Synthetic Complexity Score (SCScore)[30] as an indicator of synthetic feasibility. Molecular weight serves as an intuitive physicochemical descriptor, with larger values generally correlating with more complex molecular architectures.[31] In contrast, the SCScore is a machine learning-derived scoring function for predicting the synthetic accessibility of organic compounds. The molecular weight distribution of these metabolites spans a broad range, from 16 to 8,000 g/mol, reflecting their extensive structural diversity (Fig. 2a). Notably, the most structurally complex metabolite, lipopolysaccharide (LPS) 4-O-antigen (Figure S3) has a molecular weight of 8,069.8 g/mol, while the simplest is methane (CH_4), with a molecular weight of 16.04 g/mol, emphasizing the vast range of chemical structures present in the dataset. Additionally, the SCScore distribution in Fig. 2b reveals that 78% of the metabolites possess a score greater than 3.0, indicating that most metabolites lie above the midpoint of the SCScore scale (ranging from 1 to 5), thereby underscoring the high synthetic complexity inherent to metabolically-derived compounds. Together, these metrics highlight the potential of microbial metabolites to serve as versatile and valuable scaffolds for chemical synthesis, both in terms of structural and synthetic complexity.

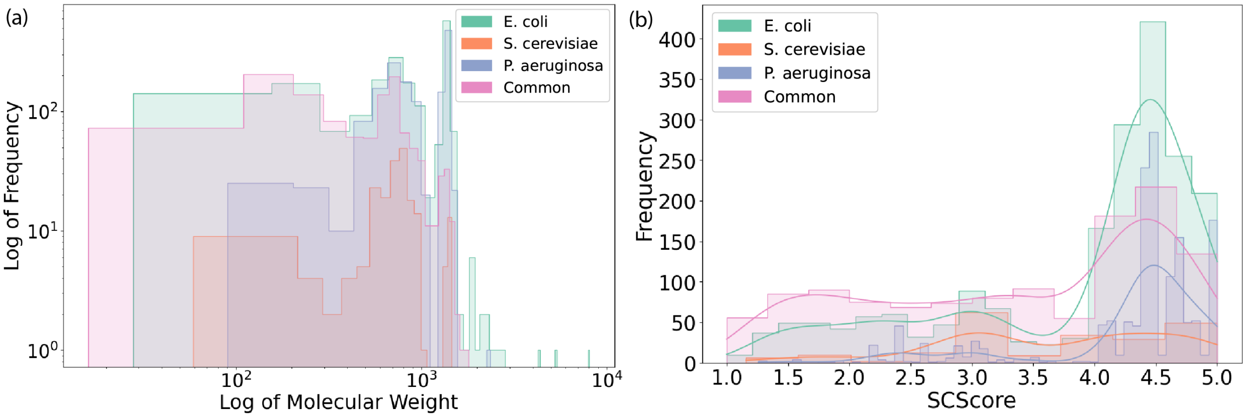

Figure 2: Histograms depicting (a) the log of the distributions of the molecular weights and (b) the distributions of the Synthetic Complexity Score of the metabolites, by organism class.

Analysis of the Multi-step Reaction Network Originating from Metabolite Precursors

To explore the progressive transformation of metabolites through multiple USPTO reaction rounds, we compared the starting metabolites and the resulting product molecules from each round to the reactant SMILES of reactions cataloged in the USPTO database. Fig. 3 examines how the molecular weight and SCScore of the metabolite-derived molecules evolve over five reaction rounds. Despite the broad chemical diversity of the initial metabolite pool, the reactant molecules (i.e. metabolites that have a reaction match in the USPTO database) in the first round of reactions exhibit lower molecular weight and SCScore values compared to the overall metabolite set. The metabolite reactants with the highest and lowest complexity are presented in the SI (Fig S4). This is likely an induced bias from the petrochemical feedstocks used to populate the starting materials of the USPTO reactions, but could also be attributed to the generally smaller and less complex nature of reactant molecules in the USPTO database, with more complex metabolites—such as certain natural products—being underrepresented as reactants in synthetic reaction datasets. Future works may expand to include a more comprehensive reaction dataset, which would enable more complex metabolites to find a direct match. This could include more comprehensive synthetic databases,[32,33] as well as biochemical reaction databases such as KEGG,[34] MetaCyc,[35] etc.[36–38] which offer unique, enzymatically selective transformations under mild conditions that are challenging to achieve through traditional synthetic chemistry.

As the reaction rounds progress, both the median molecular weight and Synthetic Complexity Score (SCScore) of the product molecules increase relative to the metabolite-derived reactant molecules. From the metabolites to the first-round products, the median molecular weight increases by 73.2%, while the SCScore increases by 51.8%. In subsequent rounds, the rate of increase diminishes, with smaller changes observed in the second and third rounds (6.1% and 17.2% for molecular weight and SCScore respectively in the second round, and 5.4% and 3.8% respectively in the third round). By the fifth round of reactions, these increases plateau, indicating a convergence in the structural and synthetic complexity of the products. This suggests that the structural diversity of the chemical space is nearing a saturation point after three reaction rounds, further justifying the truncation of the iteration USPTO reaction rounds in this work after five rounds. Interestingly, these findings align with the seminal analysis of Grzybowski on organic chemistry's network structure, which revealed that the "periphery" of known organic chemical space can be reached within an average of three concerted reactions from the "core" of that space.[39] While methodologically our work is quite distinct from this Grzybowski's, this commonality reinforces the underlying thesis of this work that broad chemical diversity (and functionality) may be accessible using bio-based feedstocks combined with a small number of traditional chemical reaction steps.

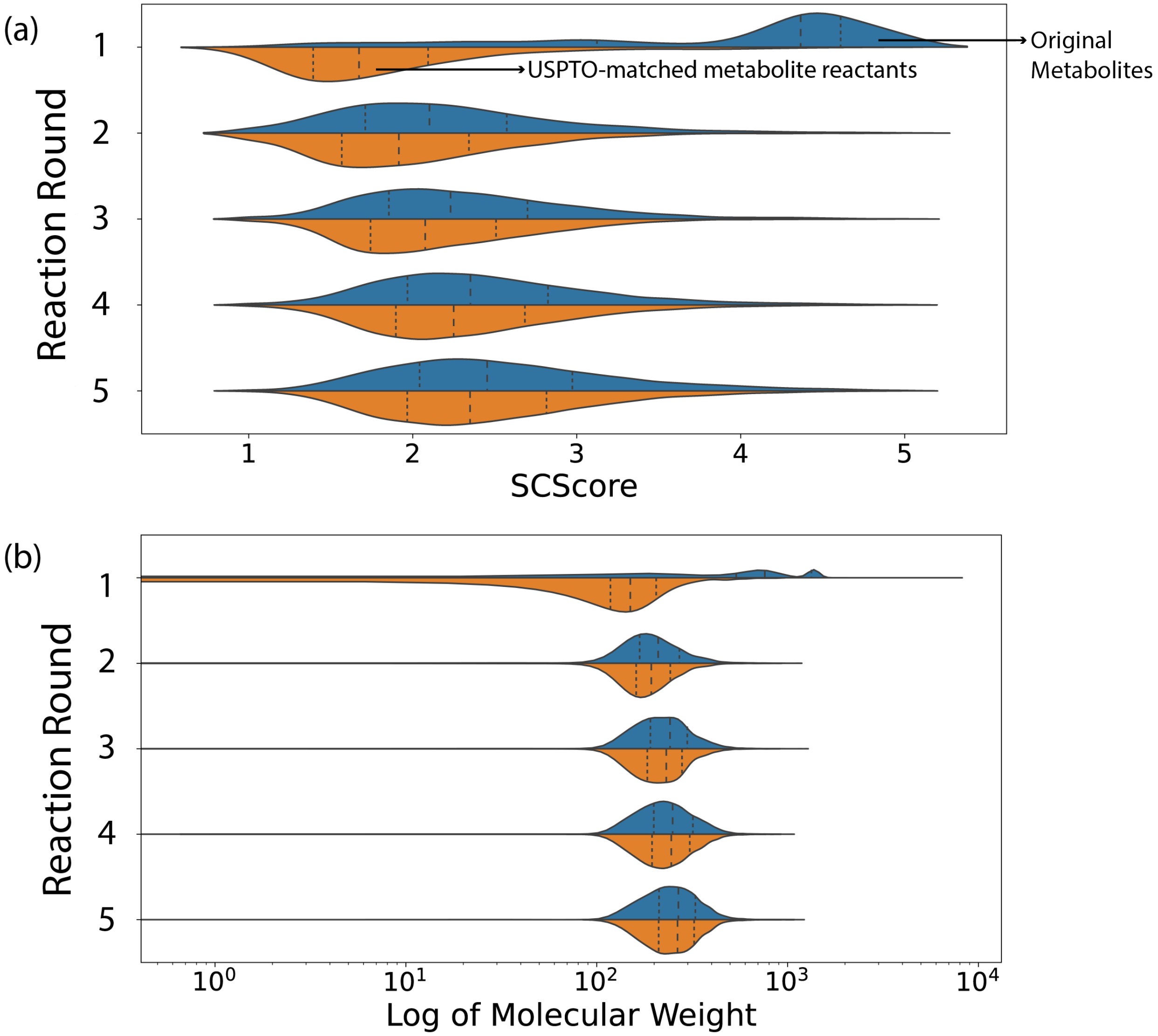

Figure 3: Violin plots depicting the distributions of (a) Molecular weight and (b) SCScore across five reaction rounds. The blue half represents the complete set of potential reactant molecules for a given reaction round (for Round 1, it is the set of all metabolite molecules, for all other rounds it is all the product molecules from the previous round). The orange half represents the molecules that are involved as substrates in a reaction for that given round. The dashed lines represent the quartiles for each group.

The evolution of chemical space coverage with respect to reaction round is visualized by comparing the USPTO-compatible metabolite-derived molecules (i.e. contains a hit in the USPTO database) against the full set of potential metabolite derivatives (all possible reactants which arise from the reaction rounds). Specifically, in the first reaction round, metabolites serve as the initial set of potential reactants for USPTO reactions, but the subset of metabolite molecules that can act as reactants in USPTO reactions occupies a smaller region of the overall metabolite chemical space (Fig. 4a). As the second and third rounds of reactions progress (Fig. 4b, 4c), the chemical space occupied by the USPTO-compatibile metabolite derivatives increasingly overlaps with that of the entire potential set of metabolite derivatives. By the third reaction round, the structural diversity of the metabolite-derived reactant molecules approaches saturation, indicating that the accessible chemical space is nearly fully explored within this iterative process. Additionally, we visualized the expansion of chemical space from reactants to their corresponding products in each reaction round to assess how the structural diversity of the product molecules evolves relative to that of the reactants in a given round (Fig. 4d-f). Similar to the analysis in Fig. 4a-c, the metabolite precursors from the first round occupy a limited region of chemical space, which expands into a more diverse product space. However, this rate of expansion in chemical diversity decreases rapidly, such that after three rounds, both reactants and products begin to occupy similar regions of structural chemical space. This trend further confirms that a sufficiently diverse and complex chemical space can be achieved within just three reaction steps.

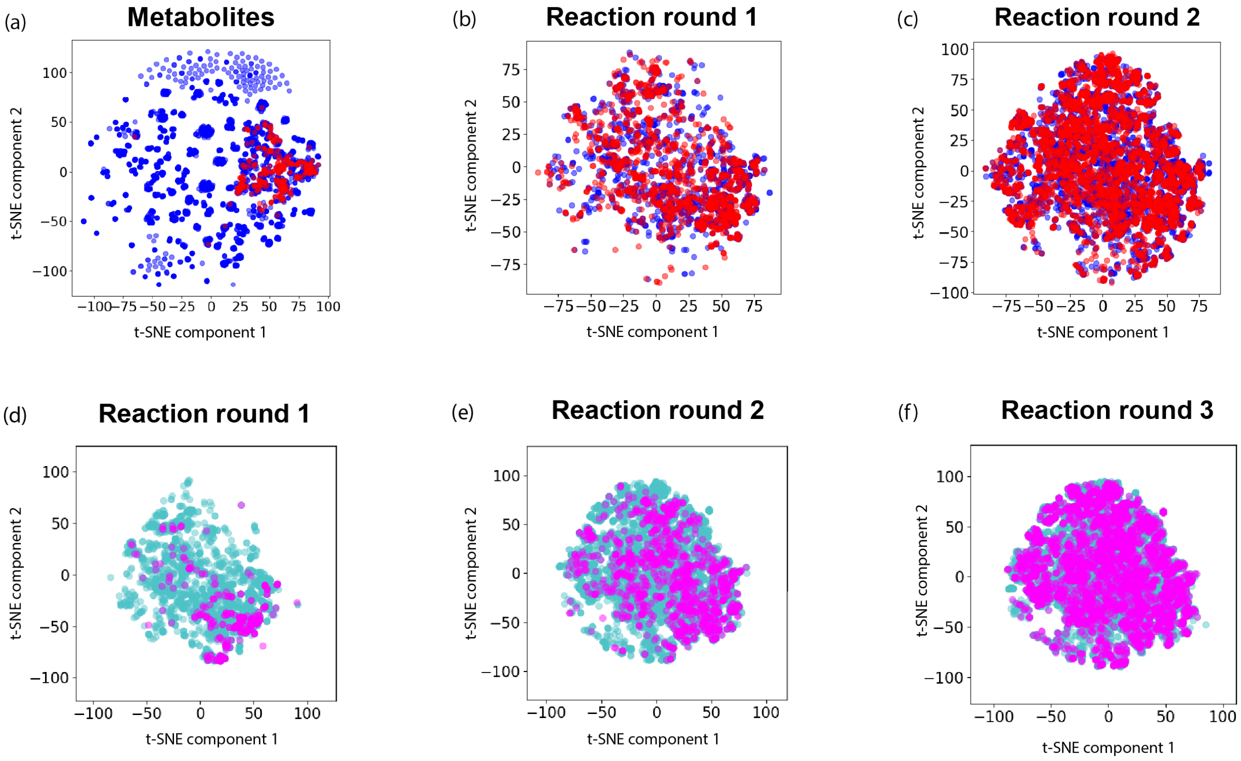

Figure 4: First set of t-SNE plots showing (a) all metabolite molecules, (b) all the product molecules after reaction round 1, (c) all the product molecules after reaction round 2. Molecules involved as substrates in USPTO chemical reactions over the three rounds are highlighted in red. Second set of t-SNE plots illustrating the chemical space of metabolite-derived reactants (magenta) and their corresponding product molecules (cyan) across reaction rounds: (a) round 1, (b) round 2, and (c) round 3.

Biases of the USPTO Dataset and Alternative Reaction Analysis for Functionalizing Metabolites

The utility of the EMRN for soft materials discovery will clearly depend on the robustness of the set of reaction pathways considered. The USPTO patent reaction database, one of the largest open-access repositories of commercially relevant reactions, is one comprehensive option to establish a framework for quantifying the potential of metabolite precursors. However, it is important to acknowledge that it also exhibits significant biases, particularly towards reaction types prevalent in medicinal chemistry. This bias limits its generalizability to other areas of chemistry, including less common reaction types and inorganic reactions.[40–42] Dobbelaere[41] has studied the distribution of reaction types in the USPTO dataset, quantiying the unbalanced distribution of reaction types within the USPTO, with an overrepresentation of heteroatom alkylation and arylation, as well as acylation reactions. The EMRN constructed here may similarly overrepresent certain reaction types and functional groups based on this observation. We expect transferability to applications beyond medicinal chemistry to be improved by incorporating reaction network databases that better represent other possible synthetic pathways.

To further clarify the effect of this bias on our dataset, Fig. 5 presents the distribution of key functional groups in the product molecules across three rounds of reactions. Popular reaction types in medicinal chemistry, such as acylation and alkylation, result in the formation of functional groups like amides, ethers, sulfonamides and amines, which are consequently found in high frequencies in our EMRN.[43,44] Interestingly, alkynes, anhydrides, acyl halides, ethers and sulfonamides, which are only present in low quantities (or entirely absent) in the initial metabolites, become more accessible after two rounds of reactions, rationalizing the observed expansion in chemical space. Additionally, we observe a marked increase in the frequency of aldehydes, ketones, amines, nitriles, ethers, halides and alcohols during the second and third reaction rounds as well. Therefore, despite the limitations of the USPTO reaction database, the resulting reaction network enables access to a broader structural and functional group diversity, presenting significant potential for a wide range of materials applications.

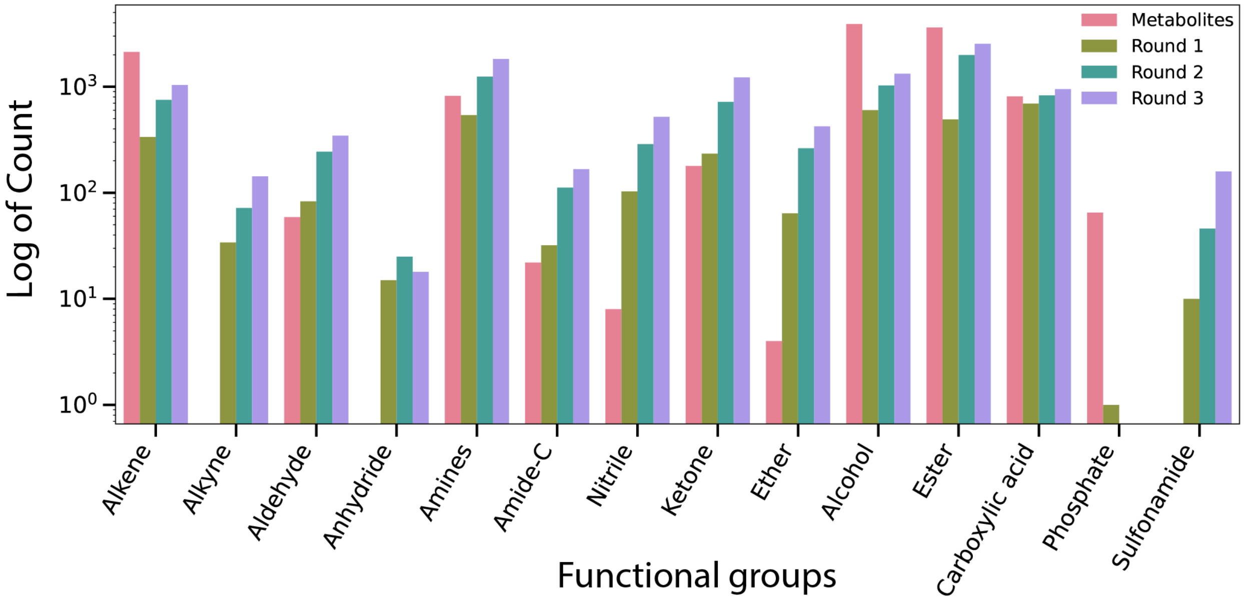

Figure 5: Distributions of common functional groups present in the product molecules across three reaction rounds of the extended metabolite reaction network.

Applications of the EMRN to Petrochemical and Biobased Chemical Building Blocks

We next investigated the extent to which metabolites and metabolite derivatives overlap with the chemical space occupied by both basic petrochemicals and biobased chemicals, which serve as critical platform molecules for commodity materials. The primary petrochemical building blocks include ethylene, propylene, methane, benzene, toluene, xylenes, butene, 1,3-butadiene, isobutylene, and their derivatives. Our analysis reveals that the molecules present in the EMRN constitute 38% of common petrochemical building blocks, including ethylene (with ethanol and ethylene oxide), propylene (with butyraldehyde), methane (with methanol), cumene, benzoic acid and toluene diisocyanate (Fig. 6). These fundamental building blocks serve as precursors for the production of a wide range of essential commodity materials, such as plastics, synthetic fibers and dyes, underscoring the potential of these traditional petrochemical feedstocks being obtained from metabolic sources for the manufacturing of high-value materials.[45,46]

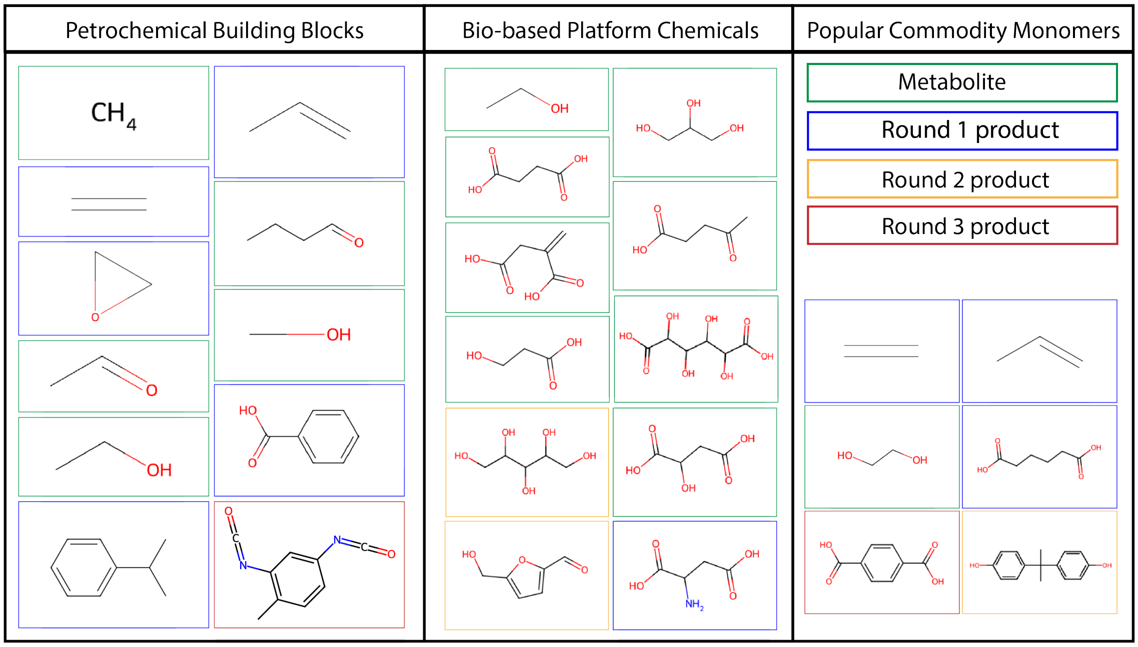
Petrochemical Building Blocks
Bio-based Platform Chemicals
Popular Commodity Monomers
Metabolite
Round 1 product
Round 2 product
Round 3 product
CH4
OH
HO
O
N
NH2

Figure 6: EMRN molecules as potential alternatives to petrochemical building blocks, bio-based platform chemicals and popular commodity materials.

Furthermore, the U.S. Department of Energy (DOE) has identified a set of the top biobased platform chemicals, which highlights biomass-derived chemicals as promising alternatives to reduce reliance on fossil fuels and promote a bioeconomy.[47] These platform chemicals, derived from sugars and other biomass feedstocks, have broad applicability in the production of fuels, materials, polymers, and other chemical products.[15,48] Our analysis shows that the molecules contained within the EMRN comprise 10 out of the 20 biobased platform chemicals, including ethanol, glycerol, aspartic acid, itaconic acid, 3-hydroxypropionic acid, levulinic acid, glucaric acid, hydroxymethylfurfural (HMF), malic acid and 1,4-succinic acid (Fig. 6). The complete list of petrochemical building blocks and biobased platform chemicals considered in this study is given in the SI.

While conceptually straightforward, our analysis demonstrates the significant potential of studying metabolites and their derivatives, not only for the development of novel methods to produce commodity materials, but also as viable alternatives to replace the basic building blocks currently used in the manufacturing of such materials. Expanding the scope of organisms and reactions considered in this analysis presents an obvious future step to enhance the coverage of chemical space, potentially revealing additional metabolites and decomposition reactions allowing for the functional substitution of conventional feedstocks in the production of high-value materials.

Applications of the EMRN to Polymer Synthesis

A key objective in the study of metabolites and their derivatives is to explore their potential for the sustainable synthesis of polymers, which is crucial for reducing the environmental impact of conventional polymer production and advancing towards more biodegradable and recyclable alternatives.[10,49] In addition to their potential to offset dependence on petrochemical precursors, molecules found in the EMRN can substitute several widely used commodity monomers in polymerizations. These include ethylene glycol, adipic acid, propylene, ethylene, terephthalic acid and bisphenol A (Fig. 6), which are integral to a range of applications, from plastics and polyester fibers to adhesives and coatings.[50,51] To further explore the potential of synthetically accessible polymer chemistries derived from metabolites and their derivatives, we examined the overlap between the metabolite network space and the reactant space in the Open Macromolecular Genome (OMG) database.[26] Developed by Kim et al., the OMG database encompasses synthesizable polymer chemistries compatible with established polymerization reactions and commercially available reactants. Our analysis revealed a significant structural overlap between metabolites, the products from the first and second reaction rounds, and the chemical space of OMG reactants (Fig. 7). Specifically, 89.6% (13,221 of 14,753) of the EMRN molecules contained one or more of 17 functional groups (Fig. 8) that are polymerizable via the 17 polymerization reactions covered by the OMG database, including step-growth, chain-growth addition, metathesis, and ring-opening reactions.

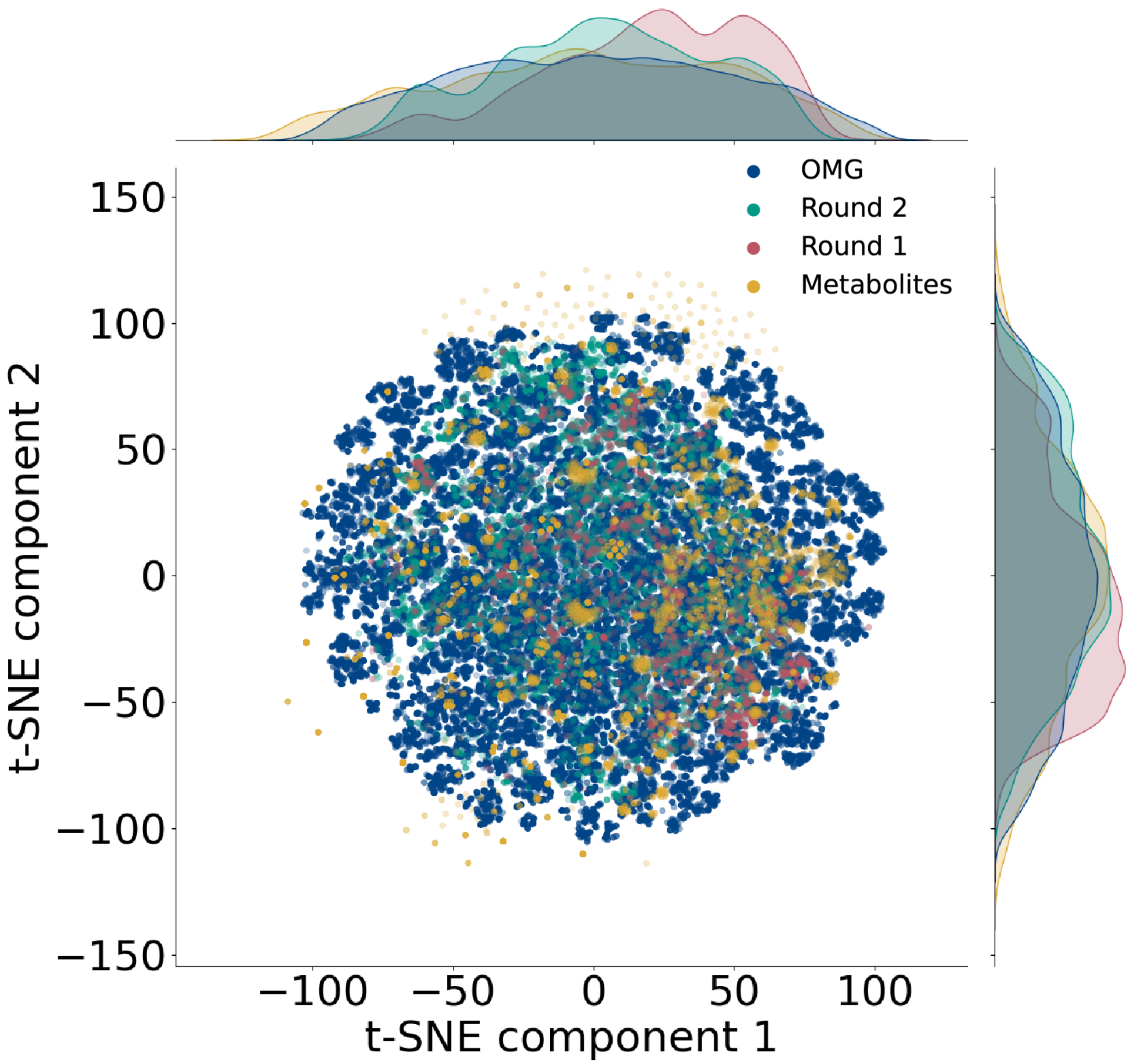

Figure 7: A t-SNE plot depicting the overlap between the original metabolite molecules and the metabolite derivatives with the reactant molecules in the Open Macromolecular Genome database, following the first and second USPTO reaction rounds .

For instance, polyesters—indispensable in textiles, packaging, and biodegradable plastics—are formed through condensation polymerization of diols and dicarboxylic acids.[52–54] The EMRN contains 1,168 diol molecules and 161 dicarboxylic acid molecules, of which 78 diols and 26 dicarboxylic acids are exact matches to OMG reactants, highlighting the potential use case of metabolites for synthesizing a diverse range of polyesters. Similarly, metabolites and their derivatives feature 230 lactone and 131 lactam structures, which are key moieties for the synthesis of high-performance, biodegradable, and recyclable polymers.[55,56] These findings underscore the substantial potential of the metabolite network in enabling sustainable synthesis of commercially relevant polymers.

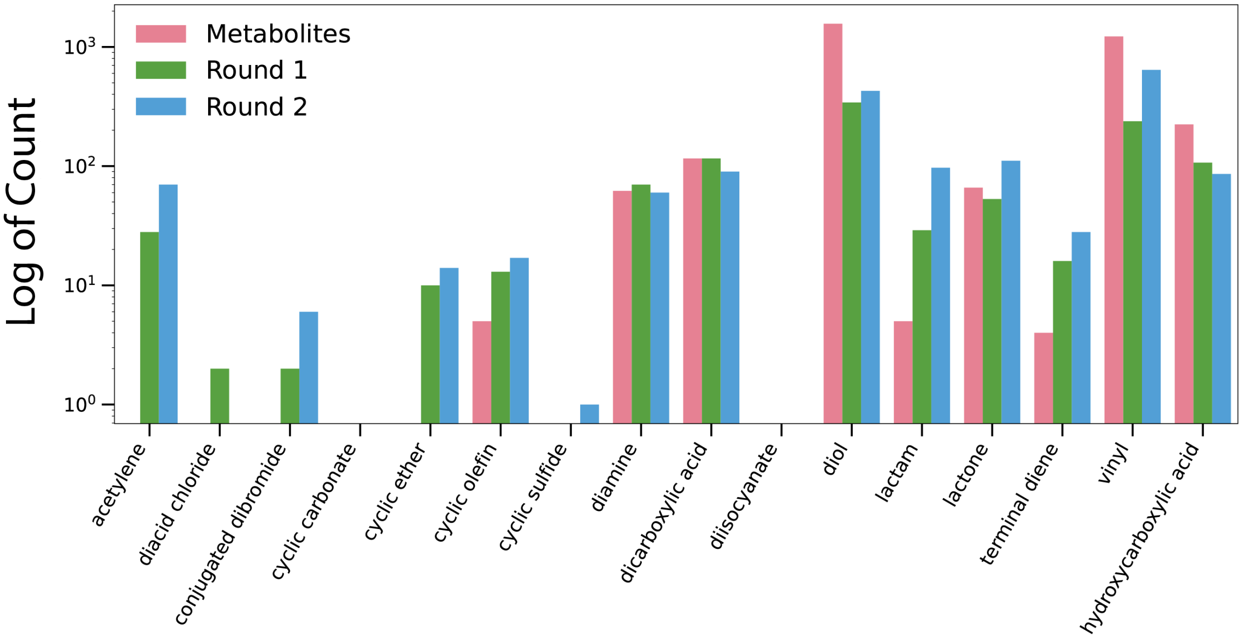

Figure 8: Distribution of the 17 polymerizable functional groups in the OMG database applied across the EMRN molecules.

While certain polymer chemistries possess strong overlap with the chemistries in the EMRN, the absence of diisocyanate chemistries in the dataset significantly constrains the synthesis of polyureas and polyurethanes, which are typically produced through the reactions of these diisocyanates with diamines and diols. Additionally, the limited representation of cyclic carbonates, cyclic sulfides, and diacid chlorides in the dataset limits the synthesis of polycarbonates, polythioethers and polyamides. Notably, cyclic sulfides and cyclic carbonates are more prevalent in the metabolites of plants and marine organisms,[57–60] indicating a promising avenue for expanding the EMRN in future studies. The formation of several diacid chlorides, along with cyclic carbonates and cyclic sulfides generated after one or two reaction rounds, further suggests potential for enhanced synthesis capabilities through increased organism diversity.

Application of the EMRN to Organic Electronics and Redox Active Materials

The metabolite molecules feature a high prevalence of conjugated chemistries, which are crucial for applications in organic semiconductors and redox-active materials. We evaluated the potential of metabolite derivatives for these molecule classes by comparing the moieties in conjugated metabolite systems with those commonly found in literature for organic electronics, photovoltaics and redox material applications. Notably, these molecules in the EMRN exhibit a variety of acene, pyrene, porphyrin, purine, phenazine and pteridine motifs (Fig. 9), many of which are common to these application types and are characterized by large, planar aromatic systems with delocalized π-electrons, enabling extended conjugation and promoting π-π stacking interactions.[61–66] The metabolite derivatives also include indoles and benzodiazoles, along with porphyrins and acenes, which are highly suited for organic photovoltaics (Fig. 9). These molecules can absorb light across a broad wavelength range, improving solar cell efficiency. Their tunable band gaps and, in the case of porphyrins, high extinction coefficients, allow for efficient light harvesting and consequently, enhanced photovoltaic performance.[67–71] Furthermore, these metabolites feature various organic redox-active materials such as quinones, naphthoquinones, pyridines, pteridines and phenazines (Fig. 9). Redox-active materials are promising for sustainable electrode materials due to their ability to undergo reversible redox cycles, maintain structural integrity during cycling, and exhibit electron-deficient characteristics.[72–75]

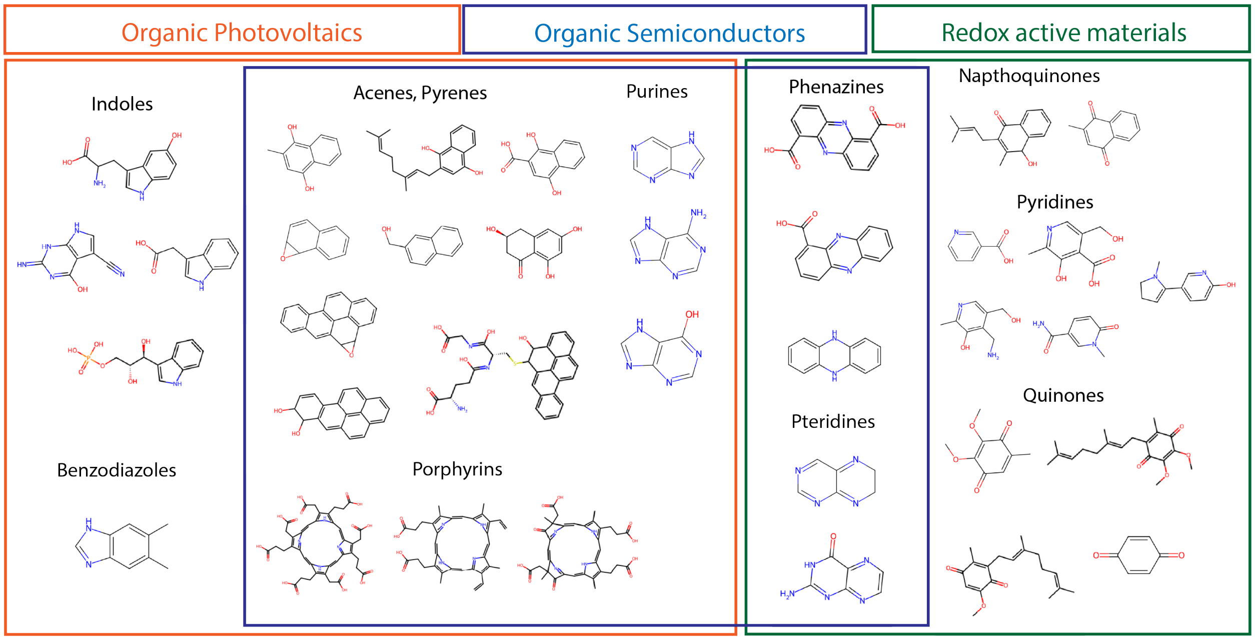

Figure 9: The potential applications of metabolite molecules in organic electronics and redox active materials.

Future Outlook for the Extended Metabolite Reaction Network

This study represents an initial exploration of the chemical space accessible through microbial metabolites, utilizing three metabolite databases from diverse organisms common in metabolic engineering, coupled to a single openly available reaction database. Given that the network is neither comprehensive nor exhaustive, it is important to acknowledge pathways for future development of the EMRN to translate the proposed framework into practical, large-scale usage.

Scaling up metabolic processes for industrial production involves several complex steps and considerations. The process typically begins with laboratory-scale experiments and progresses through pilot-scale testing before reaching full industrial scale. Although leveraging natively produced metabolites, which build on an organism's inherent biosynthetic capabilities, intuitively appears more reliable and efficient than engineering new pathways for novel molecules, it does not necessarily guarantee a simpler scale-up process. Consequently, not all metabolites in the organisms we have considered here have been industrially scaled up. Scaling up metabolite production to a commercial level involves optimizing pathway complexity, selecting suitable host organisms and addressing downstream processing (DSP) challenges, all while ensuring overall economic viability. Achieving this scale requires balancing enzyme expression for efficient metabolic flux, minimizing byproducts and leveraging robust host organisms with established industrial applications. Key challenges include maintaining consistent product quality in large-scale bioreactors, developing efficient DSP methods and ensuring the stability of engineered strains in prolonged cultures, all within the constraints of energy, raw material costs, and waste management.[76–78] Although still a relatively new field, metabolic engineering has made significant strides in scaling up the production of valuable chemicals, with ongoing research and development addressing key challenges to enable commercially viable industrial-scale processes. Understanding which metabolites in our network have successfully been scaled up enables two key objectives: first, it supports the discovery of useful molecules within the subset of industrially scalable metabolites; second, it highlights promising metabolites with potential for scale-up that remain underexplored.

This study aimed to explore alternative bio-based sources for material production as a green replacement for conventional petrochemical feedstocks, however more nuanced treatments of sustainability are crucially needed moving forward. Simply replacing one reactant in a reaction with a metabolite-based precursor does not necessarily create a sustainable process, as sustainability depends on multiple factors beyond feedstock origin. Each reaction also involves additional reagents, solvents, catalysts, and specific reaction conditions such as temperature, pressure, and time - all of which influence sustainability. According to GREENSCOPE (Gauging Reaction Effectiveness for the ENvironmental Sustainability of Chemistries with a multi-Objective Process Evaluator),[79] sustainable chemical processes can be evaluated through four key principles: efficiency, energy, environmental impact and economics. Optimizing each of these principles is essential to identify the most sustainable reaction pathway for a given synthesis. However, evaluating sustainability is complex due to the numerous metrics available within each category, such as atom economy, E-factor, reaction mass efficiency, energy consumption, emissions, waste generation, and process costs, which must be carefully selected based on the specific needs of the reaction process.[80–86] Given that our reaction database comprises only SMILES representations without stoichiometric or process conditions, we relied on molecular weight as a proxy to confirm that the metabolite-derived precursor is the main reactant in each reaction. To further enhance sustainability filtering within the reaction network, we assessed whether the filtered reactions included any bio-based co-reactants to the metabolite-based precursors. The complete details of this exercise are outlined in the SI.

A comprehensive evaluation of reaction sustainability would benefit from including additional metrics from each GREENSCOPE category. However, calculating these metrics often requires detailed reaction information. For example, even basic assessments like atom economy or E-factor require stoichiometric data.[80] Thermodynamic properties, such as reaction free energy, and kinetic factors, such as reaction rates, can be estimated via density functional theory (DFT), while machine learning models offer potential for predicting conditions and yields.[87] Although computationally intensive and sometimes limited in generalizability across diverse datasets, these methods represent a promising direction for developing comprehensive sustainability assessments for reaction pathways. Optimized 3D geometries are provided in our GitHub both for molecular visualization as well as enabling such campaigns in the future.

One specific deficiency of the chemical space accessible within the EMRN relates to its lack of sulfur containing heterocycles. This pattern can be attributed to several factors rooted in the biochemistry and evolutionary adaptations of these microorganisms. Contrastingly, the prevalence of nitrogen heterocycles reflects the abundance and bioavailability of nitrogen in microbial environments, as well as the critical role of nitrogen-containing compounds in essential cellular processes.[88,89] The scarcity of sulfur-containing aromatic rings is consistent with the lower environmental abundance of sulfur and the more specialized metabolic pathways required for its incorporation.[90,91] It is worth noting that some plants and marine organisms produce sulfur heterocycles for specific defensive or signaling purposes, such as the thiophene-containing metabolites in certain Asteraceae plants, which may not be necessary for essential functions in the microorganisms under study.[92] Furthermore, the co-occurrence of sulfur and nitrogen in the few observed heterocycles suggests a synergistic biochemical role, potentially leveraging the unique properties of both elements to fulfill specific functional requirements within these microbial systems. For instance, some fungi produce gliotoxin, a sulfur and nitrogen-containing heterocycles that serves as a virulence factor and contributes to their pathogenicity.[93] Structural comparisons with known motifs for organic electronics, photovoltaics and redox active materials revealed several metabolite-derived structures with significant potential for these applications. Moving forward, the incorporation of microrganisms with biochemistry and more diverse reaction networks will be critical to advancing the EMRN.

Conclusion

The future of soft material sourcing must shift from petrochemical-based processes to a more sustainable, biobased ecosystem. By constructing an expanded metabolite reaction network (EMRN) through the integration of biological metabolites with commercially viable chemical reactions, we explored a chemical space that bridges biological and synthetic chemistry for potential applications in materials synthesis. In addition to common biochemical metabolites, these organisms produce microbe-specific metabolic outputs that enhance chemical diversity of the extended metabolome. The overall structural diversity of the chemical space saturates within three reaction rounds, demonstrating the potential to achieve maximum diversity from these precursors within a limited number of steps. Despite biases inherent in the USPTO database, the EMRN successfully accessed functional groups beyond those found in the metabolome, broadening the scope of accessible chemistries. The molecules within this network show promise for diverse materials applications, offering alternatives to petrochemical building blocks, biobased platform chemicals and widely used commodity monomers. Furthermore, these molecules exhibit potential in polymer synthesis, organic electronics and redox-active materials, paving the way for innovative solutions in sustainable material development.

References

(1) Smith, P. B.; Payne, G. F. *Renewable and Sustainable Polymers*; American Chemical Society, 2011; p 1–10.

(2) Hou, Q.; Qi, X.; Zhen, M.; Qian, H.; Nie, Y.; Bai, C.; Zhang, S.; Bai, X.; Ju, M. Biorefinery roadmap based on catalytic production and upgrading 5-hydroxymethylfurfural. *Green Chemistry* **2021**, *23*, 119–231.

(3) John, G.; Nagarajan, S.; Vemula, P. K.; Silverman, J. R.; Pillai, C. Natural monomers: A mine for functional and sustainable materials – Occurrence, chemical modification and polymerization. *Progress in Polymer Science* **2019**, *92*, 158–209.

(4) Cho, J. S.; Kim, G. B.; Eun, H.; Moon, C. W.; Lee, S. Y. Designing Microbial Cell Factories for the Production of Chemicals. *JACS Au* **2022**, *2*, 1781–1799.

(5) Volk, M. J.; Tran, V. G.; Tan, S.-I.; Mishra, S.; Fatma, Z.; Boob, A.; Li, H.; Xue, P.; Martin, T. A.; Zhao, H. Metabolic Engineering: Methodologies and Applications. *Chemical Reviews* **2022**, *123*, 5521–5570.

(6) Lee, J. W.; Na, D.; Park, J. M.; Lee, J.; Choi, S.; Lee, S. Y. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical Biology* **2012**, *8*, 536–546.

(7) Nielsen, J.; Keasling, J. Engineering Cellular Metabolism. *Cell* **2016**, *164*, 1185–1197.

(8) Pickens, L. B.; Tang, Y.; Chooi, Y.-H. Metabolic Engineering for the Production of Natural Products. *Annual Review of Chemical and Biomolecular Engineering* **2011**, *2*, 211–236.

(9) Hoff, B.; Plassmeier, J.; Blankschien, M.; Letzel, A.; Kourtz, L.; Schröder, H.; Koch, W.; Zelder, O. Unlocking Nature's Biosynthetic Power—Metabolic Engineering for the Fermentative Production of Chemicals. *Angewandte Chemie International Edition* **2020**, *60*, 2258–2278.

(10) Lee, S. Y.; Kim, H. U.; Chae, T. U.; Cho, J. S.; Kim, J. W.; Shin, J. H.; Kim, D. I.; Ko, Y.-S.; Jang, W. D.; Jang, Y.-S. A comprehensive metabolic map for production of bio-based chemicals. *Nature Catalysis* **2019**, *2*, 18–33.

(11) Rudroff, F.; Mihovilovic, M. D.; Gröger, H.; Snajdrova, R.; Iding, H.; Bornscheuer, U. T. Opportunities and challenges for combining chemo- and biocatalysis. *Nature Catalysis* **2018**, *1*, 12–22.

(12) Wallace, S.; Balskus, E. P. Opportunities for merging chemical and biological synthesis. *Current Opinion in Biotechnology* **2014**, *30*, 1–8.

(13) Ko, Y.-S.; Kim, J. W.; Lee, J. A.; Han, T.; Kim, G. B.; Park, J. E.; Lee, S. Y. Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production. *Chemical Society Reviews* **2020**, *49*, 4615–4636.

(14) Keasling, J. D. Manufacturing Molecules Through Metabolic Engineering. *Science* **2010**, *330*, 1355–1358.

(15) Choi, S.; Song, C. W.; Shin, J. H.; Lee, S. Y. Biorefineries for the production of top building block chemicals and their derivatives. *Metabolic Engineering* **2015**, *28*, 223–239.

(16) Alam, M. I.; Ali, M. A.; Gupta, S.; Ali Haider, M. *Microbial Applications Vol.2*; Springer International Publishing, 2017; p 153–166.

(17) Chia, M.; Schwartz, T. J.; Shanks, B. H.; Dumesic, J. A. Triacetic acid lactone as a potential biorenewable platform chemical. *Green Chemistry* **2012**, *14*, 1850.

(18) Bender, T. A.; Dabrowski, J. A.; Gagné, M. R. Homogeneous catalysis for the production of low-volume, high-value chemicals from biomass. *Nature Reviews Chemistry* **2018**, *2*, 35–46.

(19) Luo, Z. W.; Lee, S. Y. Biotransformation of p-xylene into terephthalic acid by engineered Escherichia coli. *Nature Communications* **2017**, *8*.

(20) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017; `https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1`.

(21) Sajed, T.; Marcu, A.; Ramirez, M.; Pon, A.; Guo, A. C.; Knox, C.; Wilson, M.; Grant, J. R.; Djoumbou, Y.; Wishart, D. S. ECMDB 2.0: A richer resource for understanding the biochemistry ofE. coli. *Nucleic Acids Research* **2015**, *44*, D495–D501.

(22) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. ECMDB: The E. coli Metabolome Database. *Nucleic Acids Research* **2012**, *41*, D625–D630.

(23) Ramirez-Gaona, M.; Marcu, A.; Pon, A.; Guo, A. C.; Sajed, T.; Wishart, N. A.; Karu, N.; Djoumbou Feunang, Y.; Arndt, D.; Wishart, D. S. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Research* **2016**, *45*, D440–D445.

(24) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Man-

dal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, M.; Wishart, D. S. YMDB: the Yeast Metabolome Database. *Nucleic Acids Research* **2011**, *40*, D815–D820.

(25) Huang, W.; Brewer, L. K.; Jones, J. W.; Nguyen, A. T.; Marcu, A.; Wishart, D. S.; Oglesby-Sherrouse, A. G.; Kane, M. A.; Wilks, A. PAMDB: a comprehensive Pseudomonas aeruginosa metabolome database. *Nucleic Acids Research* **2017**, *46*, D575–D580.

(26) Kim, S.; Schroeder, C. M.; Jackson, N. E. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. *ACS Polymers Au* **2023**, *3*, 318–330.

(27) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *Journal of Chemical Information and Modeling* **2016**, *56*, 2336–2346.

(28) Verma, C.; Verma, D. K. *Handbook of Biomolecules: Fundamentals, Properties and Applications*; Elsevier, 2023.

(29) Giera, M.; Yanes, O.; Siuzdak, G. Metabolite discovery: Biochemistry's scientific driver. *Cell Metabolism* **2022**, *34*, 21–34.

(30) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling* **2018**, *58*, 252–261.

(31) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley, 2000.

(32) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. *The Future of the History of Chemical Information*; American Chemical Society, 2014; p 127–148.

(33) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *Journal of the American Chemical Society* **2021**, *143*, 18820–18826.

(34) Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000**, *28*, 27–30.

(35) Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **2013**, *42*, D459–D471.

(36) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research* **2021**, *50*, D693–D700.

(37) Lang, M.; Stelzer, M.; Schomburg, D. BKM-react, an integrated biochemical reaction database. *BMC Biochemistry* **2011**, *12*.

(38) Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Research* **2018**, *47*, D542–D549.

(39) Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The "wired" universe of organic chemistry. *Nature Chemistry* **2009**, *1*, 31–36.

(40) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications* **2021**, *12*.

(41) Dobbelaere, M. R.; Lengyel, I.; Stevens, C. V.; Van Geem, K. M. Rxn-INSIGHT: fast chemical reaction analysis using bond-electron matrices. *Journal of Cheminformatics* **2024**, *16*.

(42) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* **2020**, *11*, 154–168.

(43) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *Journal of Medicinal Chemistry* **2011**, *54*, 3451–3479.

(44) Zhang, Z.-X.; Willis, M. C. Sulfondiimidamides as new functional groups for synthetic and medicinal chemistry. *Chem* **2022**, *8*, 1137–1146.

(45) Matar, S.; Hatch, L. F. *Chemistry of petrochemical processes*; Elsevier, 2001.

(46) Chaudhuri, U. R.; others *Fundamentals of petroleum and petrochemical engineering*; Crc Press Boca Raton, 2011.

(47) from Sugars, C. Top Value Added Chemicals from Biomass. **2004**,

(48) Bozell, J. J.; Petersen, G. R. Technology development for the production of biobased products from biorefinery carbohydrates—the US Department of Energy's "Top 10" revisited. *Green Chemistry* **2010**, *12*, 539.

(49) *Nature Communications* **2018**, *9*.

(50) Mathers, R. T. How well can renewable resources mimic commodity monomers and polymers? *Journal of Polymer Science Part A: Polymer Chemistry* **2011**, *50*, 1–15.

(51) Hayes, G.; Laurel, M.; MacKinnon, D.; Zhao, T.; Houck, H. A.; Becer, C. R. Polymers without Petrochemicals: Sustainable Routes to Conventional Monomers. *Chemical Reviews* **2022**, *123*, 2609–2734.

(52) Pellis, A.; Herrero Acero, E.; Gardossi, L.; Ferrario, V.; Guebitz, G. M. Renewable building blocks for sustainable polyesters: new biotechnological routes for greener plastics. *Polymer International* **2016**, *65*, 861–871.

(53) Zhu, C.; Zhang, Z.; Liu, Q.; Wang, Z.; Jin, J. Synthesis and biodegradation of aliphatic polyesters from dicarboxylic acids and diols. *Journal of Applied Polymer Science* **2003**, *90*, 982–990.

(54) Satti, S.; Shah, A. Polyester-based biodegradable plastics: an approach towards sustainable development. *Letters in Applied Microbiology* **2020**, *70*, 413–430.

(55) Becker, G.; Wurm, F. R. Functional biodegradable polymers via ring-opening polymerization of monomers without protective groups. *Chemical Society Reviews* **2018**, *47*, 7739–7782.

(56) Domb, A. J.; Kost, J.; Wiseman, D. *Handbook of biodegradable polymers*; CRC press, 1998; Vol. 7.

(57) Wu, T.; Salim, A. A.; Capon, R. J. Millmerranones A–F: A Meroterpene Cyclic Carbonate and Related Metabolites from the Australian Fungus Aspergillus sp. CMB-MRF324. *Organic Letters* **2021**, *23*, 8424–8428.

(58) Zhang, H.; Liu, H.-B.; Yue, J.-M. Organic Carbonates from Natural Sources. *Chemical Reviews* **2013**, *114*, 883–898.

(59) Jiang, C.-S.; Müller, W. E. G.; Schröder, H. C.; Guo, Y.-W. Disulfide- and Multisulfide-Containing Metabolites from Marine Organisms. *Chemical Reviews* **2011**, *112*, 2179–2207.

(60) Fukaya, M.; Nakamura, S.; Kyoku, Y.; Nakashima, S.; Yoneda, T.; Matsuda, H. Cyclic sulfur metabolites from Allium schoenoprasum var. foliosum. *Phytochemistry Letters* **2019**, *29*, 125–128.

(61) Kunkel, C.; Margraf, J. T.; Chen, K.; Oberhofer, H.; Reuter, K. Active discovery of organic semiconductors. *Nature Communications* **2021**, *12*.

(62) Wang, C.; Dong, H.; Hu, W.; Liu, Y.; Zhu, D. Semiconducting -Conjugated Systems in Field-Effect Transistors: A Material Odyssey of Organic Electronics. *Chemical Reviews* **2011**, *112*, 2208–2267.

(63) O'Connell, C. E.; Sabury, S.; Jenkins, J. E.; Collier, G. S.; Sumpter, B. G.; Long, B. K.; Kilbey, S. M. Highly fluorescent purine-containing conjugated copolymers with tailored optoelectronic properties. *Polymer Chemistry* **2022**, *13*, 4921–4933.

(64) Richtar, J.; Heinrichova, P.; Apaydin, D. H.; Schmiedova, V.; Yumusak, C.; Kovalenko, A.; Weiter, M.; Sariciftci, N. S.; Krajcovic, J. Novel Riboflavin-Inspired Conjugated Bio-Organic Semiconductors. *Molecules* **2018**, *23*, 2271.

(65) Gazizov, D. A.; Gorbunov, E. B.; Zhilina, E. F.; Slepukhin, P. A.; Rusinov, G. L. Direct C–H/C–H Coupling of the Azoloannulated Pteridines with Electron Rich (Hetero)Aromatic Compounds. *The Journal of Organic Chemistry* **2022**, *87*, 13011–13022.

(66) Chen, J.; Zhang, W.; Wang, L.; Yu, G. Recent Research Progress of Organic Small-Molecule Semiconductors with High Electron Mobilities. *Advanced Materials* **2023**, *35*.

(67) Luke, J.; Yang, E. J.; Labanti, C.; Park, S. Y.; Kim, J.-S. Key molecular perspectives for high stability in organic photovoltaics. *Nature Reviews Materials* **2023**, *8*, 839–852.

(68) Gao, S.; Balan, B.; Yoosaf, K.; Monti, F.; Bandini, E.; Barbieri, A.; Armaroli, N. Highly Efficient Luminescent Solar Concentrators Based on Benzoheterodiazole Dyes with Large Stokes Shifts. *Chemistry – A European Journal* **2020**, *26*, 11013–11023.

(69) Padhy, H.; Huang, J.; Sahu, D.; Patra, D.; Kekuda, D.; Chu, C.; Lin, H. Synthesis and applications of low-bandgap conjugated polymers containing phenothiazine donor and various benzodiazole acceptors for polymer solar cells. *Journal of Polymer Science Part A: Polymer Chemistry* **2010**, *48*, 4823–4834.

(70) Erden, K.; Dengiz, C. 3-Alkynylindoles as Building Blocks for the Synthesis of Electronically Tunable Indole-Based Push–Pull Chromophores. *The Journal of Organic Chemistry* **2022**, *87*, 4385–4399.

(71) Elkanzi, N. A.; Farag, A.; Roushdy, N.; Mansour, A. Design, fabrication and optical characterizations of pyrimidine fused quinolone carboxylate moiety for photodiode applications. *Optik* **2020**, *216*, 164882.

(72) Hong, J.; Lee, M.; Lee, B.; Seo, D.-H.; Park, C. B.; Kang, K. Biologically inspired pteridine redox centres for rechargeable batteries. *Nature Communications* **2014**, *5*.

(73) Tong, L.; Goulet, M.-A.; Tabor, D. P.; Kerr, E. F.; De Porcellinis, D.; Fell, E. M.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. Molecular Engineering of an Alkaline Naphthoquinone Flow Battery. *ACS Energy Letters* **2019**, *4*, 1880–1887.

(74) Zhu, F.; Guo, W.; Fu, Y. Molecular Engineering of Organic Species for Aqueous Redox Flow Batteries. *Chemistry – An Asian Journal* **2022**, *18*.

(75) Er, S.; Suh, C.; Marshak, M. P.; Aspuru-Guzik, A. Computational design of molecules for an all-quinone redox flow battery. *Chemical Science* **2015**, *6*, 885–893.

(76) Tran, V. G.; Mishra, S.; Bhagwat, S. S.; Shafaei, S.; Shen, Y.; Allen, J. L.; Crosly, B. A.; Tan, S.-I.; Fatma, Z.; Rabinowitz, J. D.; Guest, J. S.; Singh, V.; Zhao, H. An end-to-end pipeline for succinic acid production at an industrially relevant scale using Issatchenkia orientalis. *Nature Communications* **2023**, *14*.

(77) Chen, H. et al. High-yield porphyrin production through metabolic engineering and biocatalysis. *Nature Biotechnology* **2024**,

(78) Liew, F. E. et al. Carbon-negative production of acetone and isopropanol by gas fermentation at industrial pilot scale. *Nature Biotechnology* **2022**, *40*, 335–344.

(79) Gonzalez, M. A.; Smith, R. L. A methodology to evaluate process sustainability. *Environmental Progress* **2003**, *22*, 269–276.

(80) Sheldon, R. A. Metrics of Green Chemistry and Sustainability: Past, Present, and Future. *ACS Sustainable Chemistry amp; Engineering* **2017**, *6*, 32–48.

(81) Ruiz-Mercado, G. J.; Smith, R. L.; Gonzalez, M. A. Sustainability Indicators for Chemical Processes: I. Taxonomy. *Industrial amp; Engineering Chemistry Research* **2012**, *51*, 2309–2328.

(82) Neuman, M.; Churchill, S. W. A General Process Model of Sustainability. *Industrial amp; Engineering Chemistry Research* **2011**, *50*, 8901–8904.

(83) Jacob, P.-M.; Yamin, P.; Perez-Storey, C.; Hopgood, M.; Lapkin, A. A. Towards automation of chemical process route selection based on data mining. *Green Chemistry* **2017**, *19*, 140–152.

(84) Zheng, K.; Lou, H. H.; Gangadharan, P.; Kanchi, K. Incorporating Sustainability into the Conceptual Design of Chemical Process-Reaction Routes Selection. *Industrial amp; Engineering Chemistry Research* **2012**, *51*, 9300–9309.

(85) Weber, J. M.; Guo, Z.; Zhang, C.; Schweidtmann, A. M.; Lapkin, A. A. Chemical data intelligence for sustainable chemistry. *Chemical Society Reviews* **2021**, *50*, 12013–12036.

(86) Barrett, W. M.; Takkellapati, S.; Tadele, K.; Martin, T. M.; Gonzalez, M. A. Linking Molecular Structure via Functional Group to Chemical Literature for Establishing a Reaction Lineage for Application to Alternatives Assessment. *ACS Sustainable Chemistry amp; Engineering* **2019**, *7*, 7630–7641.

(87) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. Chemical reaction networks and opportunities for machine learning. *Nature Computational Science* **2023**, *3*, 12–24.

(88) Hemmerling, F.; Hahn, F. Biosynthesis of oxygen and nitrogen-containing heterocycles in polyketides. *Beilstein Journal of Organic Chemistry* **2016**, *12*, 1512–1550.

(89) Walsh, C. T. Nature loves nitrogen heterocycles. *Tetrahedron Letters* **2015**, *56*, 3075–3081.

(90) Francioso, A.; Baseggio Conrado, A.; Mosca, L.; Fontana, M. Chemistry and Biochemistry of Sulfur Natural Compounds: Key Intermediates of Metabolism and Redox Biology. *Oxidative Medicine and Cellular Longevity* **2020**, *2020*, 1–27.

(91) Obaid, R. J.; Naeem, N.; Mughal, E. U.; Al-Rooqi, M. M.; Sadiq, A.; Jassas, R. S.; Moussa, Z.; Ahmed, S. A. Inhibitory potential of nitrogen, oxygen and sulfur containing heterocyclic scaffolds against acetylcholinesterase and butyrylcholinesterase. *RSC Advances* **2022**, *12*, 19764–19855.

(92) Ibrahim, S. R. M.; Omar, A. M.; Bagalagel, A. A.; Diri, R. M.; Noor, A. O.; Almasri, D. M.; Mohamed, S. G. A.; Mohamed, G. A. Thiophenes—Naturally Occurring Plant Metabolites: Biological Activities and In Silico Evaluation of Their Potential as Cathepsin D Inhibitors. *Plants* **2022**, *11*, 539.

(93) Evidente, A. Advances on anticancer fungal metabolites: sources, chemical and biological activities in the last decade (2012–2023). *Natural Products and Bioprospecting* **2024**, *14*.

Acknowledgement

Supporting Information Available

Morgan fingerprints and t-SNE for metabolite analysis; Functional group distribution of metabolites; Methodology for DFT calculations on metabolites; Identifying the metabolites and metabolite precursors with maximum and minimum complexity; Complete list of petrochemical building blocks and biobased platform chemicals considered in the study and Filtering EMRN reactions by bio-based co-reactants methodology.

TOC Graphic

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct.

Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.

The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.

This box and the associated title will always be printed on a separate page at the end of the document.