Identification of Bioprivileged Molecules: Expansion of a Computational Approach to Broader Molecular Space

Journal:	*Molecular Systems Design & Engineering*
Manuscript ID	ME-ART-02-2021-000013.R1
Article Type:	Paper
Date Submitted by the Author:	11-Apr-2021
Complete List of Authors:	Lopez, Lauren; Northwestern University, Department of Materials Science and Engineering Shanks, Brent; Iowa State University, College of Engineering Broadbelt, Linda; Northwestern University, Department of Chemical Engineering

Identification of Bioprivileged Molecules: Expansion of a Computational Approach to Broader Molecular Space

Lauren M. Lopez[†], Brent H. Shanks[‡], and Linda J. Broadbelt[‖]

Abstract

As interest in biobased chemicals grows, and their application space expands, computational tools to navigate molecule space as a complement to experimental approaches are imperative. This work expands upon previous work that identified candidate bioprivileged molecules from the $C_6H_xO_y$ (C6) subspace. It refines the framework that was developed previously to better refine the molecules according to their biological origin and applies it to three new subspaces of chemical structure: $C_4H_xO_y$ (C4), $C_5H_xO_y$ (C5), and $C_7H_xO_y$ (C7). For C5 and C7, roughly the top 100 bioprivileged candidates were identified, and the enhanced framework was applied to recast slightly the previous list of the top 100 C6 molecules. In addition, all top candidates were analyzed for their key functional moieties using a random forest model, and this algorithm was applied to compare the functional group space occupied by bioprivileged molecules of various databases of molecules with a focus on evaluating how closely the molecules were aligned with those known to biology. Furthermore, with the present work's focus on automation and data science principles, the framework can be easily expanded to include other chemical formulae to screen for bioprivileged candidates. This in turn facilitates the retrosynthesis process inherent in the framework to identify those bioprivileged intermediates in other subspaces that lead to target molecules.

Design, System, and Application

The goal of this research was to refine a strategy for computational discovery of bioprivileged molecules from a generic starting pool of molecules. This work refined and expanded the original application on C6 molecules to C4, C5, and C7 molecules, with the possibility of expanding further to molecules containing other heteroatoms. This strategy will guide future endeavors towards a more bio-based materials future by identifying potentially biologically accessible molecules from databases of existing molecules. With identification of

† Department of Materials Science and Engineering, Northwestern University, 2220 Campus Drive, Evanston, Illinois 60208, United States

‡ Department of Chemical and Biological Engineering, Iowa State University, 1140L BRL, Ames, Iowa 50011, United States

‖ Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, United States

bioprivileged molecules as platform chemicals, focus can thus shift to the chemocatalytic pathways of these molecules. Rapid discovery of bioprivileged molecules through computational means can thus accelerate the identification of novel or existing products of such molecules with targeted functionalities through automated network generation and retrosynthesis.

Introduction

The need to "go green" is ever expanding; data clearly shows the climate is warming, and fossil fuels are a finite resource.[1,2] While some critics argue that this is an unnecessarily pessimistic outlook, it realistically acknowledges the fact that accessing and using these resources have had a detrimental effect on the earth and its inhabitants.[3–6] For example, a ten-fold increase in fracking wells in Oklahoma was followed by a five-fold increase in earthquakes in the state.[7,8] Likewise, the 2010 BP Deepwater Horizon oil spill killed wildlife by the thousands, with estimates as high as 800,000 birds and 173,600 turtles.[9,10]

Biomass is a renewable and potentially environmentally friendly alternative to fossil fuels that also can provide access to novel molecules. While biomass can be processed in many ways, such as pyrolysis or gasification, biomass processed by industrial microorganisms will be the focal point of the present investigation. Improvements in metabolic engineering of microorganisms have been achieved, and now a wide array of chemicals is biologically accessible.[11] Metabolic pathways in so-called "powerhouse" organisms such as *Escherichia coli* (*E. coli*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) can be modified to sustainably and efficiently produce hundreds of compounds.[12–14] This alters the discussion to the intriguing question of "what should be made" instead of "what can be made."

There is already an abundance of studies on the array of biologically available molecules that can be produced in high titer and yield and which are the most valuable. For example, ethanol can be sourced from carbohydrates with high efficiency and can be used as a drop-in replacement or "platform" for the production of polyethylene, a 130-billion-dollar industry.[15,16] A 2004 study by the DOE identified a "Top 12" for chemicals derived from biomass, specifically screening sugar-derived compounds.[17] With an expansive field to scan, there is value in narrowing the possibilities to those molecules that deliver the most value to a biobased economy.

Bioprivileged molecules are a concept introduced in 2016 by Shanks et al. that are described as "biology-derived chemical intermediates that can be efficiently converted to a diversity of chemical products including both novel molecules and drop-in replacements."[18] An example of such a molecule is triacetic acid lactone (TAL). As a molecule originating from biology, TAL can be produced from several microorganisms.[18–23] While not apparently useful on its own, TAL can be converted into a variety of products from antibiotics to pesticides to preservatives with only a small number of chemical catalysis steps.[24–28]

TAL's usefulness and the diversity of its chemical progeny were discovered by a multi-year cycle of experimental effort and a bit of serendipity. There are other molecules which have been experimentally discovered which fall into the category of bioprivileged molecules: 5-hydroxymethylfurfural (5-HMF), muconic acid, levulinic acid, and furfural.[17,18,29–32] Ultimately, these are the type of compounds that would be useful to be targeted by metabolic engineering endeavors to address the "what should be made" question. In order to identify more of these bioprivileged molecules efficiently, computational tools can play a vital role.

After recognizing a dearth of computational tools to guide the identification and development of these molecules, a preliminary framework was introduced for C6 molecules, developed by Zhou et al.[33] The framework used a selection process based on structural cues and logistical information (i.e. patents, vendors, literature) and training molecules to define criteria and boundaries that characterized bioprivileged molecules. The core advance was the utilization of reaction network generation, using automated network generation (NetGen) developed by Broadbelt et al. to quantify the potential for each candidate to be diversified through chemical catalysis.[34] In addition, using on-the-fly estimation of thermodynamic properties by group contribution, the framework screened products based on the feasibility of the reactions that formed them. The resulting framework generated a list of 100 C6 bioprivileged candidates. These molecules represented a wide array of functional groups and structures, falling into nine general structural motifs. The most commonly represented functional groups were carboxylic acids and ring ethers. However, this original framework left two major questions. Are these molecules truly accessible from biology? How can this framework be expanded to include more structures?

This work focused on answering these two questions from a data science perspective by exploring both the data source and the elements comprising the computational framework. A key

advance was re-examining the toxicity screen that the original work used as a final step and replacing it with a new toxicity evaluation that is related to the ability of a molecule to be produced biologically. The second major refinement was the utilization of a random forest model to categorize molecules automatically according to functional groups, allowing the characteristics of the candidate molecules to be compared to the composition of various biological databases as a proxy for biological accessibility. The expanded framework was applied to evaluate and redefine C6 bioprivileged molecules and extended to identify C4, C5 and C7 bioprivileged molecules. For C5 candidates, three benchmark molecules were identified to guide determination of the boundary conditions. For the C4 and C7 candidate space, generalized rules for boundary conditions were set to circumvent the need for benchmark molecules. The details of the new framework and the sets of bioprivileged molecules for C4-C7 molecule spaces are provided herein.

Methods

Exploring and Managing the PubChem Database

The original formulation of the framework for identifying bioprivileged molecules by metabolites, developed by Zhou et al., relied on curating the PubChem database.[33] Thus, the first step in expanding the computational framework proposed by Zhou et al. was to better understand the capacity and the limitations of PubChem. PubChem is an open source database of chemical compounds that has a wide array of information about the compounds, including literature references, physical properties, and patent information.[35,36] The information it contains can be accessed via an html-based application programming interface (API), allowing for streamlined integration of data retrieval and processing. Atomic formulas can be retrieved using a defined syntax; for example, "H-12O1-C6" will return any molecule that has at most 12 hydrogens, at least one oxygen, and exactly six carbons. As seen in

Table **1**, PubChem is continually growing; as of January 2021, it contains 109,291,314 unique compounds. While there are databases such as KEGG, MetaCyc, and others that are more specific to metabolites, PubChem was originally chosen by Zhou et al. as the data source due to its robustness, comprehensiveness, ease of access, and potential for novelty relative to other screening studies of biobased platform chemicals.[33]

Table 1 Growth of the PubChem database over time.[36]

Data Collection	**09/13/2019**	**03/31/2020**	**09/30/2020**	**01/29/2021**
Compounds	96,261,821	102,680,973	111,326,765	109,291,314
Substances	235,485,611	253,334,503	286,669,439	269,783,826
Literature	30,028,905	30,695,548	31,521,997	32,013,955
Patents	3,175,602	3,175,602	24,824,605	24,824,605

Clearly, not all of these compounds are relevant in the scope of "bioprivileged molecules." Additionally, the PubChem API has limits. The more information requested about a structure, the fewer compounds that can be returned. For example, SMILES, a string of letters and characters that can canonically represent a molecule's structure, can be requested on the order of hundreds of thousands, while compound ID numbers (CID) can be requested on the order of tens of millions.[37] In the work by Zhou et al., the PubChem database was queried for the atomic formula $C_6H_yO_x$ to identify bioprivileged candidates.[33] This subspace amounted to approximately 30,000 molecules, or 0.03% of the entire PubChem database, and was a manageable subset that could be fully accessed and refined. In order to understand the potential subspace that should be queried in more detail and whether more narrow categories of bioprivileged molecules needed to be defined as candidates, general statistics of potential $C_xH_yO_zN_w$ candidates were gathered.

Figure 1 summarizes the characteristics of the PubChem database relevant to the molecule space of interest in this work. Figure 1a summarizes the number of molecules as a function of carbon number for $C_xH_yO_z$ molecules and clearly shows the significant growth resulting in a greater than linear increase in the number of molecules as the carbon number increases; for carbon numbers from four to 12, the number of molecules increases from 4699 to 179,404. Although nitrogen was not included in the analysis of Zhou et al., nitrogen is a common element in biologically-derived chemicals. In Figure 1b, the molecule sets are expanded to include nitrogen atom prospects. As illustrated in Figure 1b, the addition of nitrogen has a profound effect on the number of potential molecules, with a maximum achieved with $C_{12}H_xO_yN_2$ of 534,702 molecules. This quantitative analysis of the PubChem database suggests that full consideration of all the

information available for each molecule and efforts to identify bioprivileged molecules would benefit from analysis of subsets of molecules individually.

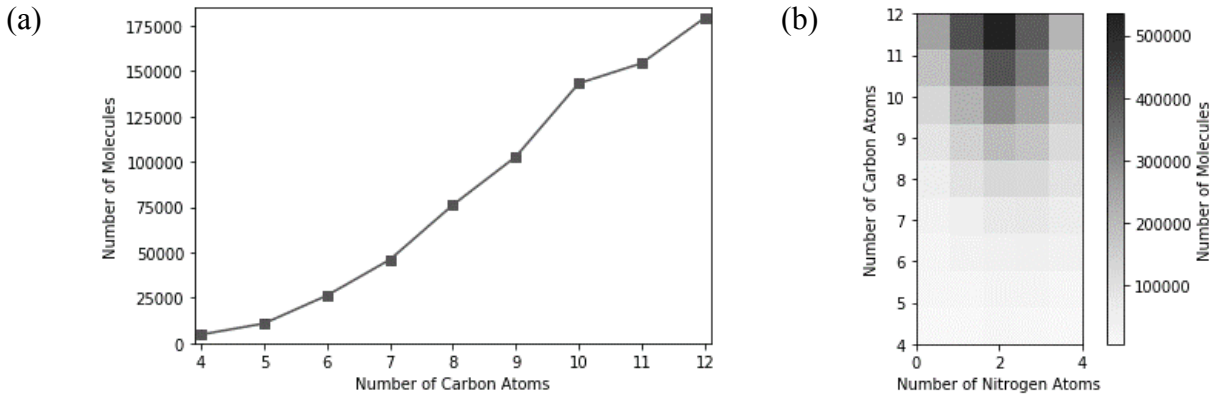
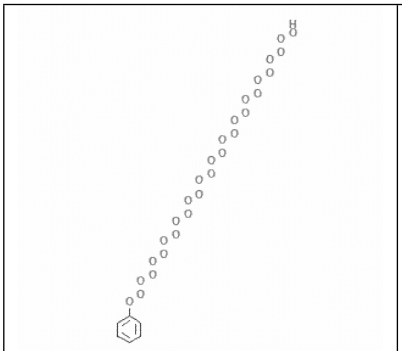
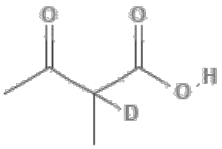
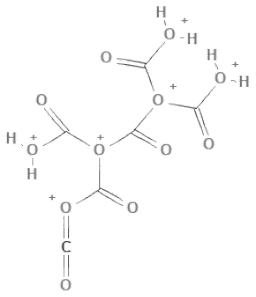
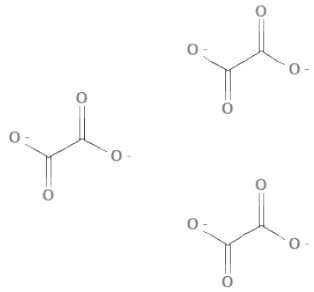


Figure 1 Graphical exploration of PubChem subspace of $C_xH_yO_zN_w$ compounds. (a) Occurrences of hydrogen- and oxygen-containing organic molecules by carbon count. (b) Distributions of nitrogen-containing molecules by carbon count.

Beyond the initial categorization carried out according to atomic formula, additional scrutiny of the PubChem database was required. Table 2 gives an example of some of the entries in the database that are not candidates as bioprivileged molecules or even molecules. Hydrates, ions, isotopes, and unstable compounds are all found within the database. These molecules that are not viable emphasize the importance of several of the steps in the computational framework that are geared to ensure nonsensical structures do not persist through the funneling process. For example, if hydro-$(peroxy)_{14}$-benzene passed through the reactive moiety steps, it would be filtered out when trying to find sufficient literature, vendor, or patent information, i.e., notoriety, pertaining to it since that information does not exist. It is important to note that PubChem records every instance of a structure, so molecules will have separate entries for both ionic and neutral forms.

Table 2 Examples of compounds and their corresponding CIDs found in the PubChem database that were eliminated from consideration.

		
CID: 89175406	CID: 59985446	CID: 23528345
		
CID: 20434779	CID: 57583515	

It is important to note that removal of ions, hydrates, isotopes, and nonsensical molecules in the pre-processing steps was straightforward. However, it is possible that a set of enantiomers, which will be differentiated only by notoriety, pass through the entire funnel. For simplicity, those molecules were reduced to a representative species in the sets before proceeding to automated network generation, based on their shared canonical SMILES since they would have the same products derived from operators that do not consider stereochemistry.

Development of Computational Framework Based on Network Generation

In Zhou et al., the original framework for identifying bioprivileged molecules included 23 steps. The procedure, summarized in Table 3, can be divided into "pre-processing" (steps 1-12) and "post-processing" (steps 13-23) stages, which are delimited by reaction network generation in

which chemical operators expand potential bioprivileged molecules to their progeny.[33] Pre-processing evaluates the potential bioprivileged candidates based on structure, potential network as summarized according to reactive moieties, and notoriety as measured by literature references, patents and vendors. The objective of this stage is to reduce the candidate pool to minimize the computational cost of generating the explicit reaction networks. Since each molecule can have $O(10^2)$ progeny, generating a network for all candidate molecules is computationally expensive while not required to do a first-pass screen of candidates' reactive moieties.

Table 3 Original computational framework as it applied to specifically C6 molecules. [33]

Step	Description
1	download $C_6H_xO_y$ in PubChem
2	remove hydrocarbons C6H
3	remove species with 5/6-membered ring but without ring −O−
4	remove species with total reactive moiety no. < 3
5	remove species with total reactive moiety no. > 6
6	remove species with one same reactive moiety > 3
7	remove species with reactivity index of rank 0 < 8
8	remove species with reactivity index of rank 0 > 14
9	remove species with reactivity index of rank 1 < 65
10	remove species with reactivity index of rank 1 > 150
11	remove radical/ionic species
12	remove species with sum of no. of literature, patent, and vendor ≤ 5
13	remove species with no. of total products < 40
14	remove species with no. of total products > 90
15	remove species with no. of NetGen rxn rank 0 ≤ 6
16	remove species with no. of NetGen rxn rank 1 ≥ 130
17	remove species with no. of rxn ≤ 50 when ΔG < 15 kcal/mol
18	remove species with no. of rxn ≥ 110 when ΔG < 15 kcal/mol
19	remove species with no. of known products < 10
20	remove species with no. of literature of products ≤ 7
21	remove species with ratio of no. of known to novel products ≥ 0.75
22	remove species with ratio of no. of patents to literature < 20
23	remove species with predicted toxicity index > 0.67
24	list of 100 C6 bioprivileged molecule candidates

The Python module RDKit was used to identify reactive moieties, which are functional groups that participate in the chemocatalyic reactions encoded into NetGen. One of the criteria for a bioprivileged molecule is to be both easily diversified and selectively converted. For this

purpose, molecules with excessive or deficient reactive moieties (steps 4-6) were removed. However, this alone cannot predict the reaction networks. Therefore, Zhou and coworkers used a reactivity index to predict the number of progeny of the starting molecule in the first and second generations without explicitly doing network generation. This allows for the removal of structures that would theoretically have networks with too few products or too many products. Specifically, R0 (steps 7-8) refers to the predicted first-generation progeny, and R1 (steps 9-10) refers to the predicted second-generation progeny, the equations for which are given by Equations 1 and 2 in Zhou et al.[33]

Once molecules were filtered based on the criteria specified in steps 1-12, a network was generated for each of the remaining molecules. At this point, the number of products as a function of generation is known, as well as the explicit identity of every product. In the work of Zhou et al., two generations based on the application of 17 different reaction operators were created and evaluated. The operators are summarized in Table 1 of Zhou et al. and reproduced for easy reference here in the Supplementary Information in Table S1. The progeny can then be evaluated to more sufficiently analyze the diversity of the network by examining the thermodynamics of the reaction to form a given product and notoriety of each product. Finally, the candidates were filtered through a toxicity measure. Originally, developmental toxicity was used as the final cutoff using the EPA's TEST software.[38,39] Each of these steps had cutoffs curated based on three "benchmark" molecules: TAL, 5-HMF, and muconic acid. Application of the framework resulted in filtering 100 bioprivileged candidate molecules from a pool of approximately 30,000 molecules. These 100 bioprivileged candidate molecules fell into nine motifs.

In the present work, the 23 steps can be categorized into the overarching categories outlined in Figure 2. Step 12 of the original framework (removal of radicals and ionic species) was combined with step 2 (removal of hydrocarbons). Additionally, steps that had upper and lower limit ranges were combined into one step. The ratio of known to unknown molecules was changed to the percent unknown to provide a more intuitive metric of the progeny subspace. Furthermore, as briefly noted in the previous section, the number of mentions in the literature, patents, and vendors, or "LPV score", was further canonicalized to "notoriety".

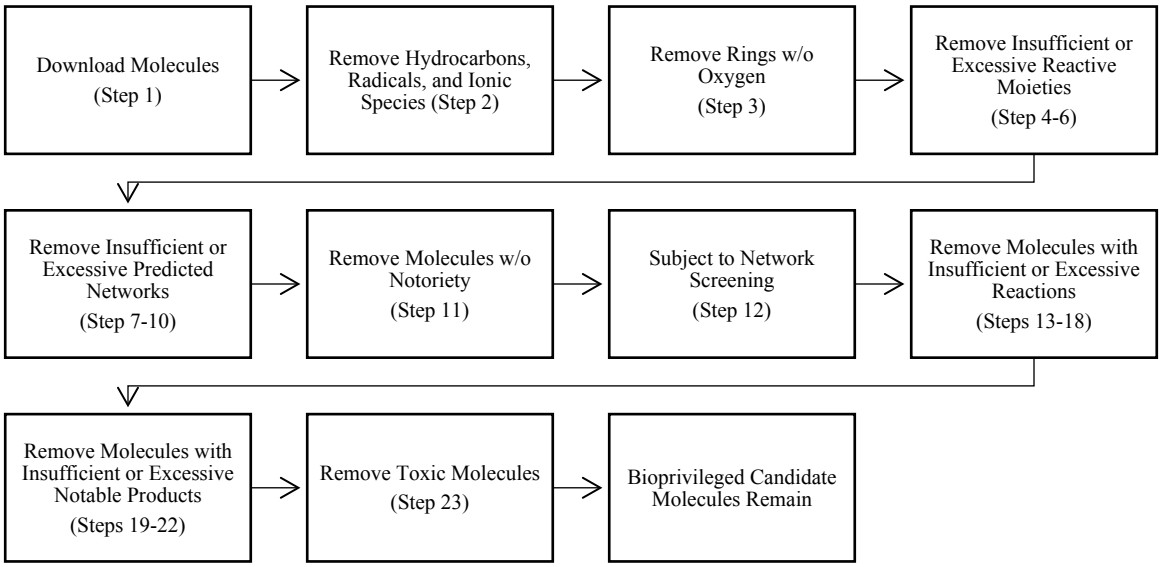


Figure 2 Framework developed by Zhou et al. was generalized into categories of steps for display and analysis of filtering results across different subspaces.[33]

The next major change that was made to the framework was the introduction of a different toxicity measure in order to provide a proxy for the biological feasibility of the molecules. The "bio" part of the bioprivileged molecules in the previous selection framework was implied by structural cues. For example, the criterion that a ring must contain an oxygen atom is a simple proxy for a molecule anticipated to be derived from metabolic processing of sugars. Without fully integrating a metabolic pathway search into the process, which requires the development and detailed application of a companion computational platform that is currently under development, other avenues of better capturing the required biological origins of bioprivileged molecules were explored. Specifically, cytotoxicity is a common metric used in drug discovery and development pipelines to evaluate the potential for a molecule to poison its host. Cytotoxicity can be correlated to hydrophobicity.[40–42] Thus, hydrophobicity was calculated using quantitative structure−activity relationship (QSAR) methods pertaining to the octanol-water partition coefficient using a tool developed by Cheng et al.[43] As the last step in the framework, this can easily be modified.

The final major advance in the framework for bioprivileged molecule discovery was the development of a random forest model for characterization of C6 molecules into one of nine

desirable molecule classes. The original 100 C6 molecules in the work of Zhou et al. were manually grouped into nine structural motifs which are summarized in Figure 3.[33] The goal of the more rigorous classification developed here was twofold. First, classification of molecules according to functional groups is ambiguous when two or more groups are present, so a classification method that assigns each candidate to desired features of known bioprivileged molecules is valuable as a pre- or post-processing step. Second, by developing a method that can quickly and efficiently classify molecules in a database and comparing their statistics against populations of known bioprivileged molecules, we can assess the likelihood these molecules are able to be accessed from biology without directly interfacing with metabolic network generation and ensure our screening protocol is guiding exploration of vast molecule space into a more focused zone that defines bioprivileged molecules.

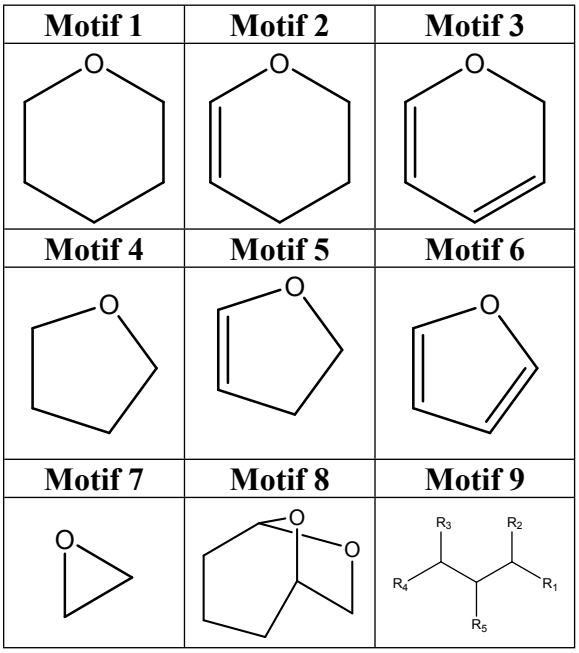


Figure 3 Nine structural motifs identified by manual inspection of the 100 bioprivileged molecules identified by Zhou et al.[33] Motifs 1-3 are oxygen-containing six-membered rings with 0, 1, or 2 double bonds in the ring. Motifs 4-6 are oxygen-containing five-membered rings with 0, 1, or 2 double bonds in the ring. Motifs 7-9 are epoxides, heterocyclic compounds, and "other"/ "linear" structures.

To first classify the entirety of the PubChem C6 database into these nine motifs, a random forest model, from the Python based scikit-learn module, was trained on a sample of 225

molecules: 25 for each of the nine motifs.[44] These 25 were selected by using SMARTS-matching to identify unambiguous examples of each motif. This training set is in the Supplementary Information. SMARTS matching was not used to classify the entire database because while it can provide exact matches, the concept of "closeness" to a certain motif was more desired for the large-scale analysis of molecules with more than one functional group. For example, a four-membered ring would be close to either Motif 1 or Motif 4, but a SMARTS match would put it in Motif 9 or unmatched, depending on the algorithm.

All molecules were converted into Morgan fingerprint arrays to perform this analysis. The accuracy of the model was tested on the 100 candidates output by the original framework of Zhou et al. The random forest gave an accuracy between 93% and 99% when applied to the original 100 candidates. The maximum depth of each tree (*max_depth*) was set to 10, and the number of trees (*n_estimators*) was set to 28. The other parameters were kept at the default values. After training the model, the $C_6H_xO_y$ molecules in three databases (PubChem, MetaCyc, and KEGG) were sorted into the nine motifs.[35,45,46] MetaCyc and KEGG are both biological databases. This was repeated 500 times, and the average occurrence of each motif was selected to represent the data set. Finally, 90% confidence intervals were generated from the set of 500 random forest outputs.

Finally, the revised and expanded computational framework was applied to the C4, C5, and C7 subspaces, and reapplied to the C6 subspace. The $C_5H_xO_y$ subspace had the boundary conditions of its framework developed using the same process as the $C_6H_xO_y$ subspace: select three benchmark molecules, see which values they have at each of the twenty-three steps, and set those values as boundary conditions for the remainder of the subspace. The three benchmark molecules selected for $C_5H_xO_y$ structures were levulinic acid, furfural, and itaconic acid.[17,29,30,47–50] Similar to their C6 predecessors, TAL, 5-HMF, and muconic acid, the three benchmark molecules for C5 were selected based on accessibility from biology and ability to diversify.

Expansion to the C7 and C4 pools used an approach aimed at generalizing the framework without using benchmark molecules to guide the boundary conditions. Extrapolation based on the boundary conditions for the C5 and C6 framework, paired with sensitivity analysis to reduce the starting pools by 95% in the pre-processing stage, provided a generalized rule for creating boundary conditions for other molecules. Not all of the steps needed to change due to carbon counts (e.g., step 12 for notoriety, step 23 for toxicity). Furthermore, these formulas are not

immutable rules and can be modified based on the desired outcome (i.e., how many molecules should pass through each stage). In the following formulas, C_n is the carbon count.

Minimum Total Moiety (step 4):

$$\left\lfloor \frac{C_n}{2} \right\rfloor \quad (1)$$

Minimum R0 Value (step 7):

$$\left\lfloor \frac{C_n}{2} \right\rfloor + 5 \quad (2)$$

Maximum R0 Value (step 8):

$$\left\lfloor \frac{C_n}{2} \right\rfloor + 10 \quad (3)$$

Minimum R1 Value (step 9):

$$C_n \times 10 + 5 \quad (4)$$

Maximum R1 Value (step 10):

$$C_n \times 20 + 5 \quad (5)$$

Results

Evaluation of the Use of Hydrophobicity over Developmental Toxicity

First, the 138 C6 bioprivileged candidates identified in the work of Zhou et al. that made it to step 23 in the pipeline were re-evaluated using hydrophobicity. XLOG3, developed by Cheng et. al., was used to calculate the partition coefficient for each of these candidates.[43] The goal of the framework was to reduce the list to as close to 100 as possible; therefore, a toxicity cutoff was chosen such that either 100 molecules remained, or a benchmark molecule had the highest toxicity, with 0.02 buffer space. In the case of C6, a benchmark molecule came before a list of 100, so the 138 molecules were reduced to a list of 101 candidates, shown in Figure 4, with a partition coefficient cutoff of 0.55; TAL, the most hydrophobic benchmark, has a partition coefficient of 0.53.

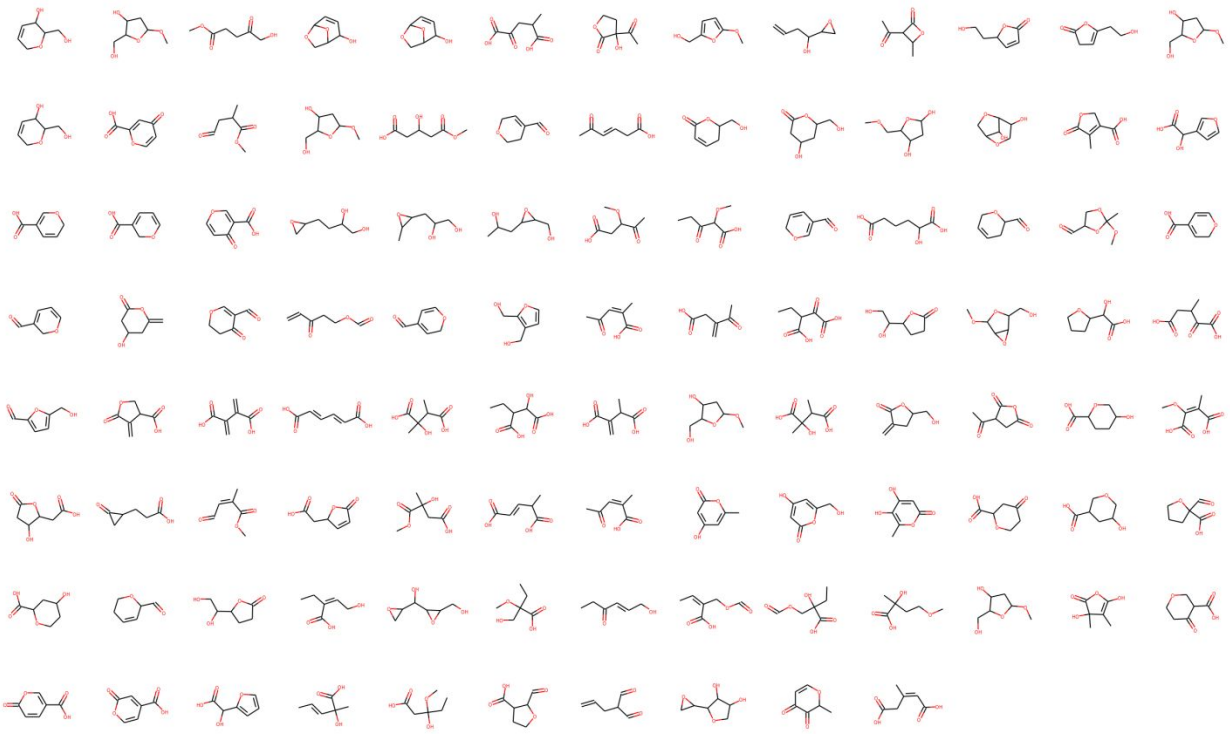


Figure 4 Re-evaluated 101 C6 bioprivileged candidates based on hydrophobicity as opposed to developmental toxicity. The structures of the molecules and their PubChem CID are provided in the Supplementary Information.

It is important to acknowledge the benchmark molecules when studying which motifs are represented. The benchmark molecules for C6 (5-HMF, TAL, muconic acid) fell into Motifs 3, 6, and 9. One would expect a bias towards these three categories, with the exception being Motif 9, which does not inherently have any base moieties and would be disproportionally filtered out due to minimum moiety and reactivity criteria benchmarks. Therefore, while overlap between categories might provide some insight into representation, trends are suggested to be more representative of how close the bioprivileged candidates relate to each category.

From Figure 5, it is clear that the 100 candidates of Zhou and coworkers are modestly better represented by the biological databases than the general database from which they were drawn. For the PubChem database, the candidates' ratios fall into the confidence intervals of Motifs 2 and 5. For MetaCyc and KEGG, they fall into the confidence intervals of Motifs 2, 4, 5, 7, 8, and 9. The first distinction is that more groups fall within the confidence interval of KEGG or MetaCyc.

Of those motifs, three are uniquely represented by those databases – that is, they do not fall within the confidence interval of PubChem for that motif. Furthermore, the χ^2 value for the KEGG, MetaCyc, and PubChem databases compared to the original 100 bioprivileged candidates are 0.40, 0.25, and 0.46, respectively.

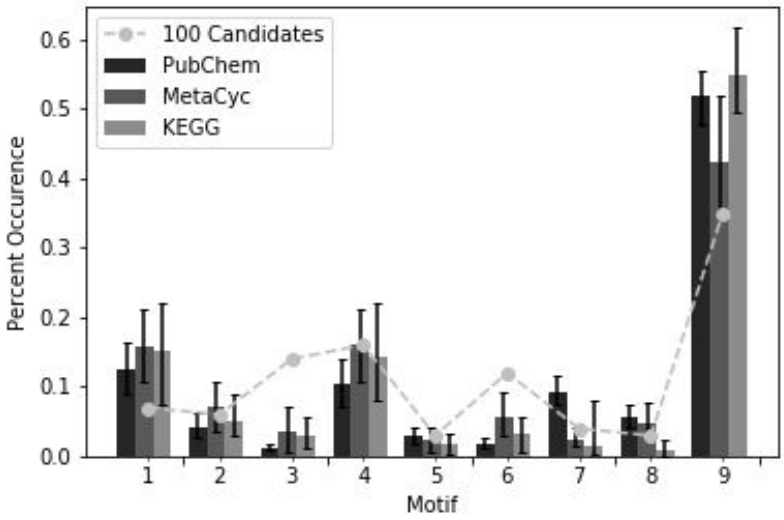


Figure 5 Comparison of motif compositions of 100 bioprivileged candidates proposed by Zhou and coworkers against PubChem, MetaCyc, and KEGG.

It is interesting to examine how the change in the toxicity evaluation altered the final set of bioprivileged molecules and how much overlap existed between the two sets, the union of which was 73 molecules. As seen in Figure 6, 53% of molecules passed both benchmarks. Nine molecules in the 138 passed neither benchmark. The majority of these molecules contained a terminal double bond. At a glance, the most significant change from the original 100 candidate molecules is that Motif 6 decreased. Motif 6 represents furans, which are a perfect example of how a molecule could be biologically sourced yet still toxic.[51]

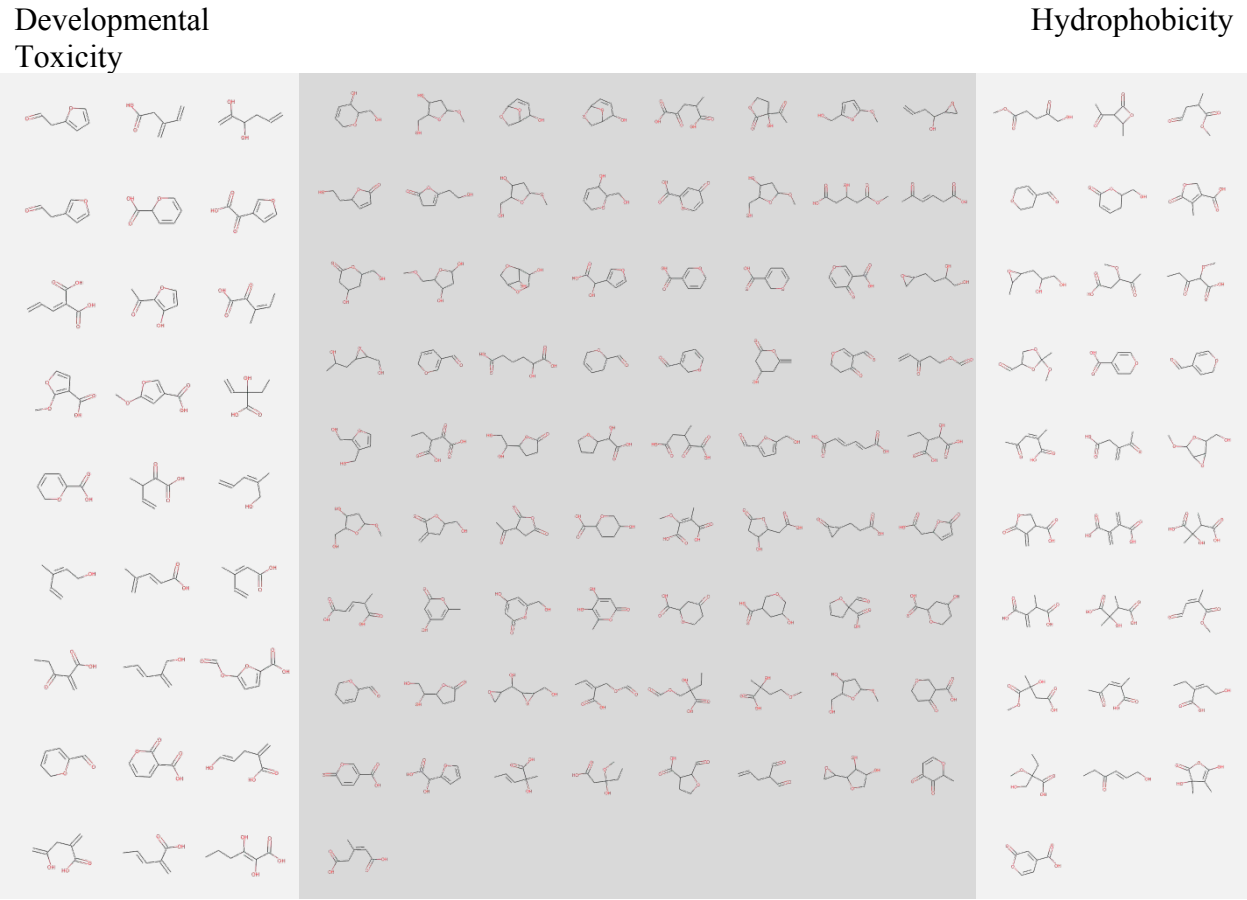


Figure 6 Molecules that passed the developmental toxicity benchmark compared to molecules that passed the hydrophobicity benchmark. The molecules in the darker region in the center met both criteria.

To better evaluate the difference between the hydrophobic and developmental toxicity metrics, the random forest model was revisited with the 101 bioprivileged molecules that were newly selected based on hydrophobicity, with an emphasis on how the representation across the different motifs changed. The new C6 candidates, which required sorting of 28 new molecules, were manually sorted into their groups because the dataset was small enough to not necessitate computational assistance. There was one exception. The saturated four-member ring does not fit into any of the nine motifs. Therefore, it was sorted using the random forest model. In 76% of the random forest instances, the model placed this compound into Motif 4 – the motif with a saturated five-member ring. Although this is a single example, the model is working as intended; intuitively, one would place this molecule in Motif 1 or 4 as it is "closer" to these structures than the linear category of Motif 9.

The classification of the new 101 bioprivileged molecules in comparison to the three databases is shown in Figure 7. There is an even stronger correlation between the metabolic databases and the bioprivileged candidates. For the PubChem database, the candidates did not fall within the confidence interval of any of the motifs. For the MetaCyc and KEGG databases, the candidates' ratios fell within the margin of error of Motifs 2, 4, 6, 7, 8, and 9. Thus the biological databases uniquely represent six motifs of the candidates, indicating a better fit to biologically-sourced molecules. Furthermore, the χ^2 values from comparing the 101 bioprivileged candidates increased from the χ^2 values from the 100 original bioprivileged candidates across the board to 0.36, 0.23, and 0.36 for KEGG, MetaCyc, and PubChem, respectively. Based on this analysis, we recommend that hydrophobicity is a more directed criterion than developmental toxicity as the last filtering stage in the process. The results reported below for C4, C5, and C7 molecules are thus restricted to the application of the revised framework.

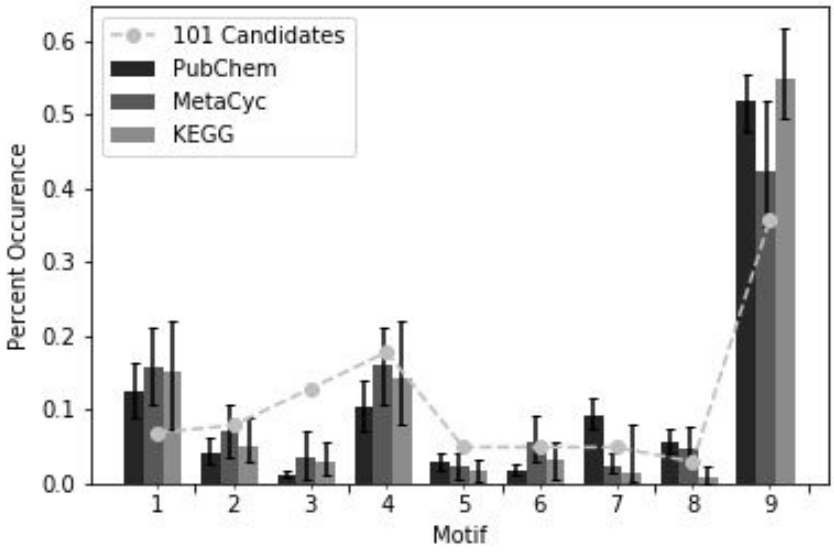


Figure 7 Comparison of motif compositions for 101 re-selected bioprivileged candidates against three databases: PubChem, MetaCyc, and KEGG.

Expanding the Framework to C5 Molecules

The first application in expanding the framework was to C5 molecules. The first step was to establish the parameters that governed the bounds of the framework (Table 3) based on the C5 benchmark molecules, itaconic acid, fufural, and levulinic acid. As shown in Table 4, the C5

benchmark molecules have much tighter ranges of moiety metrics and projected progeny than C6. The range of the highest and lowest number of second-generation progeny expected from the C6 molecules was 78 molecules, compared to C5 benchmarks with a range of 15 molecules.

Table 4 Comparison of parameters characterizing C6 benchmark molecules to C5 benchmark molecules for pre-processing conditions.

Step	Carbon Number	Total Moiety	Total Single Moiety	R0	R1	Notoriety
TAL	6	5	2	8	69	758
5-HMF	6	6	2	13	134	7,984
Muconic Acid	6	6	2	8	147	18,651
Itaconic Acid	5	3	2	11	76	547
Furfural	5	5	2	9	69	35,799
Levulinic Acid	5	3	1	10	84	12,602

In the work of Zhou et al., there were nearly 30,000 compounds adhering to $C_6H_xO_y$ in PubChem. By comparison, there were only fewer than half as many compounds, 13,000, with the $C_5H_xO_y$ formula at the time of this analysis, although this is not surprising given the smaller number of carbon atoms and the number of possible arrangements of fewer atoms. Based on the sample size alone, there is the expectation that to reduce these molecules into a list of approximately 100 candidates, less strict screening would be required. Therefore, although the benchmark molecules for C5 molecules would suggest a smaller range for R0 and R1 values, to not hinder the diversity of potential candidates, these values from the benchmark molecules served as guides for centering the boundary conditions, as seen in Table 5. For example, while the C5 benchmark molecules inhabited a smaller range of second-generation progeny, the benchmarks were set to a minimum of 7 and maximum of 12 for R0, and a minimum of 55 and a maximum of 130 for R1 to encompass a larger space.

Table 5 Comparison of upper and lower limits for ranking criteria based on benchmark molecules.

Carbon Count	Max. Single Moiety	Min. Total Moiety	Max. Total Moiety	Min. R0	Max. R0	Min. R1	Max. R1
C5	3	2	6	7	12	55	130
C6	3	3	6	8	14	65	150

As summarized in Table 4, the notoriety of the C5 benchmark molecules is comparable to or even greater than that of the C6 molecules. Given the success of the notoriety criterion used previously for C6 winnowing, the same boundary was applied to C5 molecules. As long as a molecule had at least five mentions across literature, patents, or vendors, it would pass this step. Although this criterion appears relatively loose, 1,685 C5 compounds were removed at this stage. Overall, the number of molecules remaining at each of the first 11 steps in the pre-processing stage is shown graphically in Figure 8 for both C6 and C5 molecules. At this point, sets of enantiomers are screened to leave only one representative molecule to send to network generation.

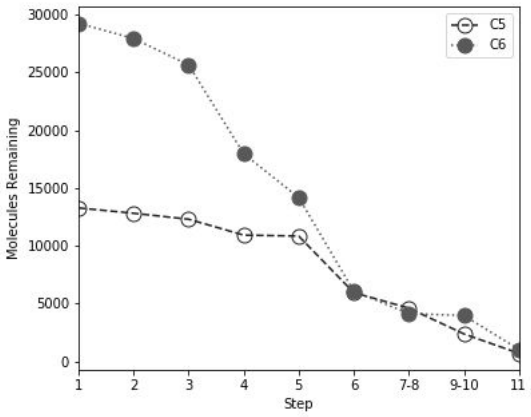


Figure 8 Differences between pre-processing steps for C5 and C6 molecule subspaces. The steps comprising the x-axis are those delineated in Figure .

The post-processing phase of the analysis focuses on the products that are formed from reaction network generation. The first steps in the post-processing stage, those that address the distribution of the products and the thermodynamics of the reactions forming the products, did not change substantially between the C6 and C5 analyses and are not explored in detail. The other post-processing steps are summarized in Table 6. The notoriety of the C5 and C6 derived products, however, showed significant difference. A greater percentage of the C5 progeny is already known,

and the cutoff of 57% used for C6 molecules was too strict. Thus, the C5 cutoff for unknown was altered to a minimum of 25% to account for the progeny occupying a more well-known space, a consequence of a lower carbon atom count.

Table 6 Comparison of post-processing criteria between C5 and C6 benchmarks.

Step	Carbon Number	Known Products	Product Literature	% Unknown	Patents/Lit
TAL	6	10	8	76	134
5-HMF	6	27	94	68	94
Muconic Acid	6	33	307	59	680
Itaconic Acid	5	40	7706	38	26
Furfural	5	24	3306	53	30
Levulinic Acid	5	35	5919	27	27

For the final step of the framework, a partition coefficient was chosen so approximately 100 molecules remained, or until a benchmark had the highest partition coefficient, whichever milestone came first. Furfural, a furan, had the highest partition coefficient at 0.41, which was not unexpected given the toxicity of this class of molecules. In this case, approximately 100 molecules came first. A partition coefficient of 0.53 yielded 102 molecules.

Finally, 102 C5 molecules were output at the end of the filtering process, with the evolution of the post-processing reduction in the number of molecules remaining shown in Figure 9. The process of down selection for the C5 molecules was more gradual than that for the C6 molecules, due to the smaller initial pool overall in the post-processing steps. The C5 pool was most dramatically reduced by step 20, in which the products were removed based on literature mentions. The final set of 102 C5 molecules is shown in Figure 10. As seen, the final 102 C5 bioprivileged candidates give a diverse array of structures and functionalities, similar to what was observed for C6 candidates. Using the C6 motif classifications to bin the C5 molecules, Motifs 1, 2, 4, 5, 6, 7, and 9 are represented in the candidates.

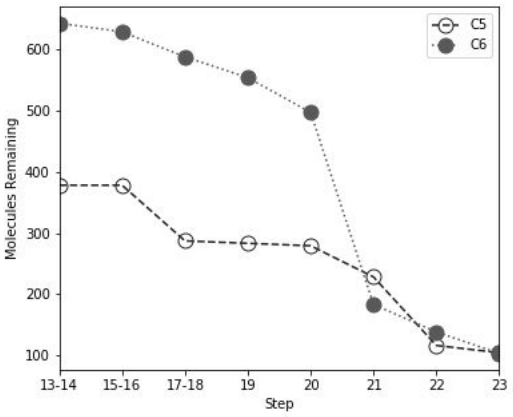


Figure 9 Summary of post-processing stages for down selection of C5 and C6 molecule subspaces. The steps comprising the x-axis are those shown in Figure .

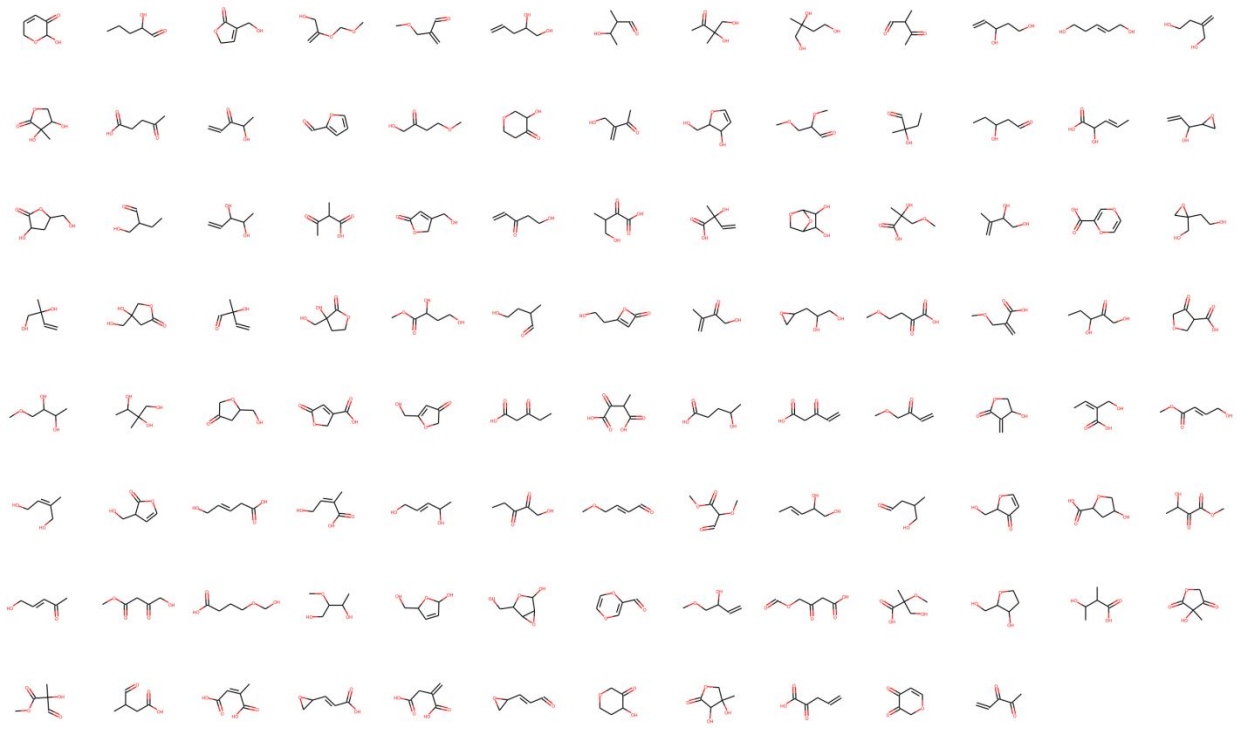


Figure 10 Final 102 C5 bioprivileged molecule candidates. The structures of the molecules and their PubChem CID are provided in the Supplementary Information.

Expansion to C7 and C4 Molecules

For C6 and C5 molecules, the filtering was informed by preselected molecules known to fit the bioprivileged paradigm. A fully automated process would bypass the need for "benchmark"

molecules and select boundaries based on carbon count. The values for these benchmarks are shown in Table 7, which are based off Equations 1-5 listed in the Methods section.

Table 7 Extrapolated boundary conditions for C4 and C7 pre-processing stage based on formulas provided in Equations 1-5.

Carbon Count	**4**	**7**
Max. Moiety	3	3
Min. Total Moiety	2	3
Max. Total Moiety	5	7
Min. R0	7	8
Max. R0	12	14
Min. R1	45	75
Max. R1	110	170

Figure 11 depicts the C7 molecules remaining in the pipeline after each filtering step as compared to the C5 and C6 analyses. Similar to C5 and C6, the reduction of the molecule pool in the pre-processing phase is approximately 95%. Therefore, even though C7 started with approximately 60,000 molecules and underwent a "wider" filter, the percentage reduction of the data pool was proportional to C5 and C6 molecules. This indicates that the use of an equation instead of benchmark molecules to guide the filtering process is a reasonable methodology. There are 150 molecules remaining after the 23rd step in the process. These remaining candidates were then filtered to the final list of 100 C7 candidates using a partition coefficient of 0.57 and are shown in Figure 12.

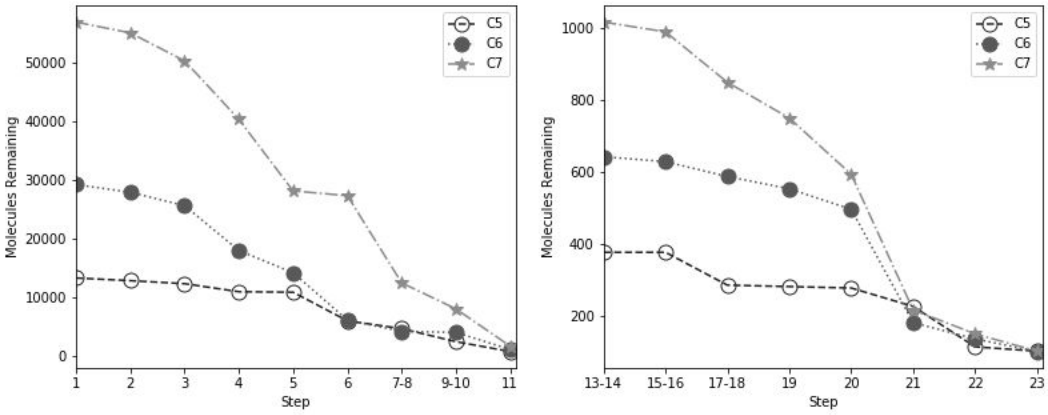


Figure 11 Reduction of approximately 60,000 C7 molecules on PubChem to 100 C7 candidate bioprivileged molecules. The steps comprising the x-axis are those outlined in Figure .

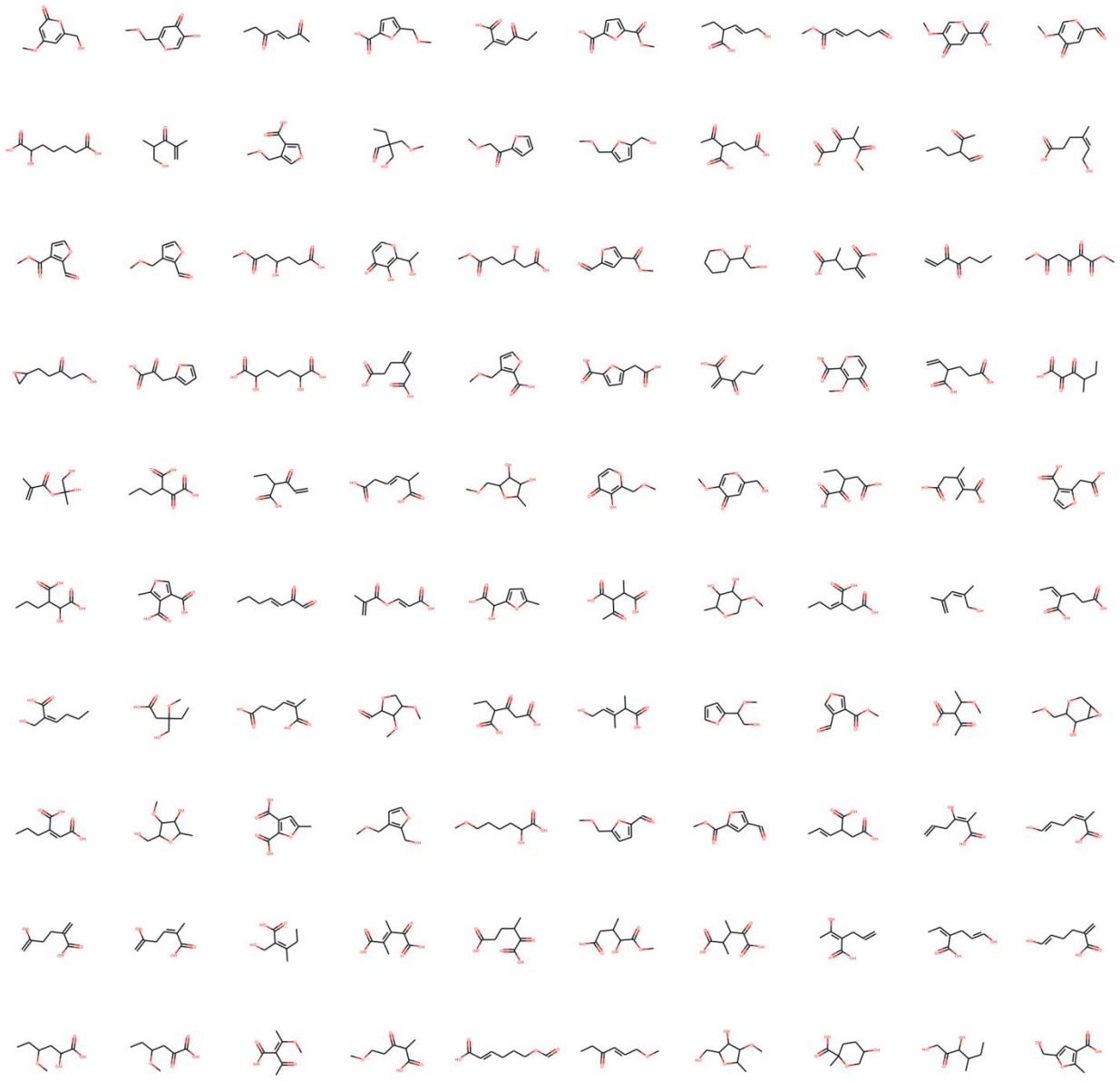


Figure 12 Final 100 C7 bioprivileged molecule candidates. The structures of the molecules and their PubChem CID are provided in the Supplementary Information.

Similar to C5, a diverse group of structures is represented in the C7 candidates: Motifs 1, 3, 4, 6, 7, and 9 are all present. Since no benchmark molecules were chosen, the products were examined for a compound of known biological origin, to evaluate the robustness of the framework. A notable member of the C7 candidates is opuntinol.[52] While opuntinol can be readily accessed from one of our C6 benchmark molecules via transformations, it can also be derived from biological sources directly.

Finally, the framework was expanded to smaller molecules. There were approximately 6,000 C4 molecules in the PubChem database; this is less than half of C5, which was already a reasonably fleshed out chemical space as evaluated according to the metrics imposed in this framework. During automated network generation, the C4 molecules produced very few unknown molecules, which was not unexpected given the types of chemocatalytic transformations encoded, with little growth in carbon number. The results of this analysis are shown in Figure 13, which compares the percentage of unknown progeny for candidates of each of the four carbon counts.

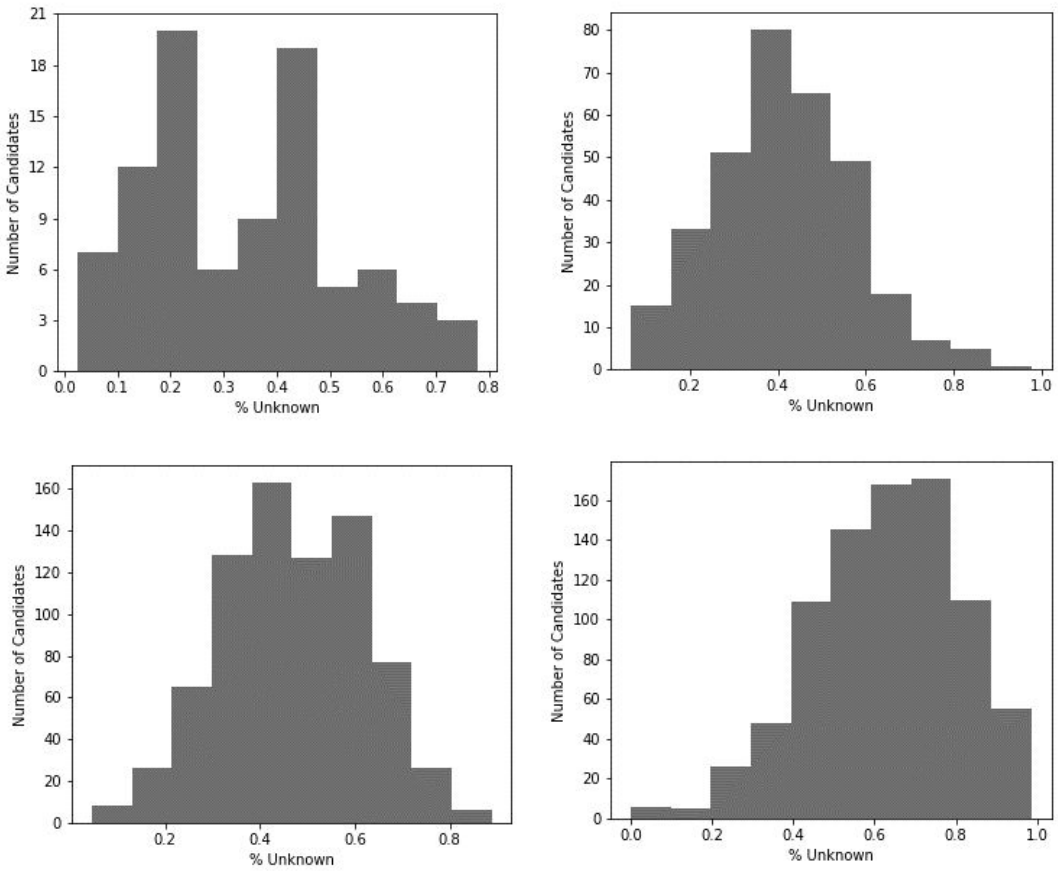


Figure 13 Histograms of percent of unknown products for C4 (top left), C5 (top right), C6 (bottom left), and C7 (bottom right) candidates that survived until the 19th step of the framework and were then used as starting molecules for reaction network generation for two generations. The total number of candidates that were used as starting molecules was 91, 324, 773, and 843, for C4, C5, C6, and C7, respectively. The mean percent of unknown products for these subspaces are 33%, 42%, 47%, and 63%, respectively.

It is important to note that a key feature of bioprivileged molecules is the presence of novel molecules in their chemocatalyic space. Thus, filtration of the C4 molecules implied a lower limit for applying this framework. To be clear, however, this is not because C4 molecules are not valuable as bioprivileged molecules, but because the space generated in the current work is small enough that computation does not lend the same advantages that are apparent for the larger and more diverse, yet undiscovered, molecule spaces characteristic of higher carbon numbers. For this reason, the 'final' molecules presented in Figure 14 should be scrutinized with the selection rules applied in mind, as there were fewer degrees of freedom to modify the boundary conditions. The results, however, are not unexpected. For example, it is well within the framework for five- and six- membered rings to appear, i.e., Motifs 1-6. In order for these structures to appear in the C4 space, heteroatoms have to be included to build the ring. Specifically, the list contains several dioxanes, which, while hydrophilic, are known carcinogens and for that reason are questionable as bioprivileged molecules.[53]

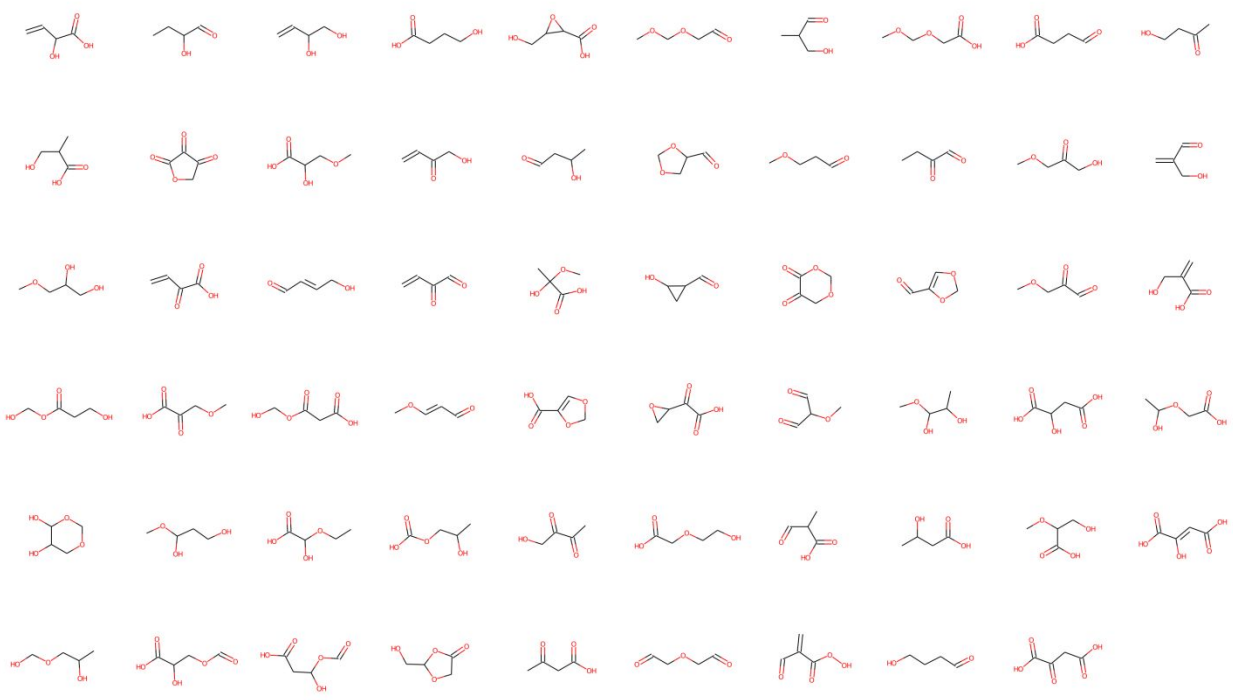


Figure 14 Final 59 C4 bioprivileged molecule candidates. The structures of the molecules and their PubChem CID are provided in the Supplementary Information.

Summary

Bioprivileged molecules are an intriguing class of molecules that offer a diverse group of compounds with tremendous potential. By improving the computational tools to expedite the discovery of such molecules, we can work towards extrapolating and exploring the space their products can occupy, from antivirals to corrosion inhibitors to flame retardants. Modification of the procedure introduced in Zhou et al. led to a more generalized framework for identifying bioprivileged molecules from existing databases. Replacement of a developmental toxicity metric with a hydrophobicity metric led to more biologically feasible structures, which was then verified through the implementation of a random forest model to classify sets of candidate molecules and diverse databases, including two that are biological in nature, KEGG and MetaCyc.

The new framework was applied to revisit the classification of C6 bioprivileged molecules and develop new molecule sets for C4, C5, and C7 bioprivileged candidates. The filtration of C5 and C6 molecules was guided by benchmark molecules. By extrapolating the characteristics of these molecules as a function of carbon number, the C4 and C7 molecule spaces were also explored. However, exploration of the progeny of the C4 candidates, which offered a limited number of novel molecules, showed that there is a lower limit on the number of carbon atoms that warrants exploration computationally. The C7 molecules space, in contrast, was dramatically pruned from approximately 60,000 starting molecules to a set of 100 candidates, which in turn produced a large novel product subspace. Overall, the work delivers a list of 303 C5, C6, and C7 bioprivileged candidates that are recommended for further exploration as efforts to transition to a biobased economy continue to push forward.

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Bioenergy Technologies Office Funding Opportunity Announcement DE-FOA-0001916 BioEnergy Engineering for Products Synthesis (BEEPS)", Award Number DE-EE0008492. We are grateful to Mr. Kevin Shebek for providing a list of molecules from PubChem based on SMARTS pattern matching, and we thank Professor George Kraus for offering suggestions for C5 benchmark molecules.

References

1. Kan, H., Chen, R. & Tong, S. Ambient air pollution, climate change, and population health in China. *Environment International* **42**, 10–19 (2012).

2. Huang, J. *et al.* Recently amplified arctic warming has contributed to a continual global warming trend. *Nature Climate Change* **7**, 875–879 (2017).

3. Greene, D. L., Hopson, J. L. & Li, J. Have we run out of oil yet? Oil peaking analysis from an optimist's perspective. *Energy Policy* **34**, 515–531 (2006).

4. Adelman, M. A. Comment on: R.W. Bentley, "Global oil & gas depletion", Energy Policy 30 (2002) 189–205. *Energy Policy* **31**, 389–390 (2003).

5. Bentley, R. W. Global oil and gas depletion: An overview. *Energy Policy* **30**, 189–205 (2002).

6. Kampa, M. & Castanas, E. Human health effects of air pollution. *Environmental Pollution* vol. 151 362–367 (2008).

7. Ellsworth, W. L. Injection-Induced Earthquakes. *Science* **341**, 1225942–1225942 (2013).

8. Keranen, K. M., Savage, H. M., Abers, G. A. & Cochran, E. S. Potentially induced earthquakes in Oklahoma, USA: Links between wastewater injection and the 2011 Mw 5.7 earthquake sequence. *Geology* **41**, 699–702 (2013).

9. Haney, J., Geiger, H. & Short, J. Bird mortality from the Deepwater Horizon oil spill. II. Carcass sampling and exposure probability in the coastal Gulf of Mexico. *Marine Ecology Progress Series* **513**, 239–252 (2014).

10. Williams, R. *et al.* Underestimating the damage: interpreting cetacean carcass recoveries in the context of the Deepwater Horizon/BP incident. *Conservation Letters* **4**, 228–233 (2011).

11. Lee, J. W., Kim, H. U., Choi, S., Yi, J. & Lee, S. Y. Microbial production of building block chemicals and polymers. *Current Opinion in Biotechnology* vol. 22 758–767 (2011).

12. Zhang, X., Tervo, C. J. & Reed, J. L. Metabolic assessment of E. coli as a Biofactory for commercial products. *Metabolic Engineering* **35**, 64–74 (2016).

13. Krivoruchko, A., Siewers, V. & Nielsen, J. Opportunities for yeast metabolic engineering: Lessons from synthetic biology. *Biotechnology Journal* **6**, 262–276 (2011).

14. Hong, K. K. & Nielsen, J. Metabolic engineering of Saccharomyces cerevisiae: A key cell factory platform for future biorefineries. *Cellular and Molecular Life Sciences* vol. 69 2671–2690 (2012).

15. Bechthold, I., Bretz, K., Kabasci, S., Kopitzky, R. & Springer, A. Succinic Acid: A New Platform Chemical for Biobased Polymers from Renewable Resources. *Chemical Engineering & Technology* **31**, 647–654 (2008).

16. Sheldon, R. A. Green and sustainable manufacture of chemicals from biomass: State of the art. *Green Chemistry* vol. 16 950–963 (2014).

17. Werpy, T. & Petersen, G. *Top Value Added Chemicals from Biomass: Volume I -- Results of Screening for Potential Candidates from Sugars and Synthesis Gas*. *Us Nrel* http://www.osti.gov/servlets/purl/15008859/ (2004) doi:10.2172/15008859.

18. Shanks, B. H. & Keeling, P. L. Bioprivileged molecules: Creating value from biomass. *Green Chemistry* **19**, 3177–3185 (2017).

19. Xie, D. *et al.* Microbial synthesis of triacetic acid lactone. *Biotechnology and Bioengineering* **93**, 727–736 (2006).

20. Cardenas, J. & da Silva, N. A. Metabolic engineering of Saccharomyces cerevisiae for the production of triacetic acid lactone. *Metabolic Engineering* **25**, 194–203 (2014).

21. Tang, S. Y. *et al.* Screening for enhanced triacetic acid lactone production by recombinant Escherichia coli expressing a designed triacetic acid lactone reporter. *Journal of the American Chemical Society* **135**, 10099–10103 (2013).

22. Saunders, L. P., Bowman, M. J., Mertens, J. A., da Silva, N. A. & Hector, R. E. Triacetic acid lactone production in industrial Saccharomyces yeast strains. *Journal of Industrial Microbiology and Biotechnology* **42**, 711–721 (2015).

23. Markham, K. A. *et al.* Rewiring yarrowia lipolytica toward triacetic acid lactone for materials generation. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 2096–2101 (2018).

24. Kraus, G. A., Wanninayake, U. K. & Bottoms, J. Triacetic acid lactone as a common intermediate for the synthesis of 4-hydroxy-2-pyridones and 4-amino-2-pyrones. *Tetrahedron Letters* **57**, 1293–1295 (2016).

25. Chia, M., Schwartz, T. J., Shanks, B. H. & Dumesic, J. A. Triacetic acid lactone as a potential biorenewable platform chemical. *Green Chemistry* **14**, 1850–1853 (2012).

26. de March, P. *et al.* Reactions of triacetic acid lactone with carbonyl compounds. X-Ray structure determination of 3-acetoacetyl-2-chromenone and 3,6,9,12-tetramethyl-1 *H* ,6 *H* ,7 *H* ,12 *H* -6,12-methanodipyrano[4,3- *b* :4,3- *f*]-dioxocin-1,7-dione. *Journal of Heterocyclic Chemistry* **23**, 1511–1513 (1986).

27. Bacardit, R. & Moreno-Mañas, M. Hydrogenations of triacetic acid lactone. A new synthesis of the carpenter bee (Xylocopa hirsutissima) sex pheromone. *Tetrahedron Letters* **21**, 551–554 (1980).

28. Kraus, G. A. & Wanninayake, U. K. An improved aldol protocol for the preparation of 6-styrenylpyrones. *Tetrahedron Letters* **56**, 7112–7114 (2015).

29. Bozell, J. J. *et al.* Production of levulinic acid and use as a platform chemical for derived products. *Resources, Conservation and Recycling* **28**, 227–239 (2000).

30. Wright, W. R. H. & Palkovits, R. Development of heterogeneous catalysts for the conversion of levulinic acid to γ-valerolactone. *ChemSusChem* **5**, 1657–1667 (2012).

31. Saha, B. & Abu-Omar, M. M. Advances in 5-hydroxymethylfurfural production from biomass in biphasic solvents. *Green Chemistry* vol. 16 24–38 (2014).

32. Zhang, Y., Zhang, J. & Su, D. 5-Hydroxymethylfurfural: A key intermediate for efficient biomass conversion. *Journal of Energy Chemistry* **24**, 548–551 (2015).

33. Zhou, X. *et al.* Computational Framework for the Identification of Bioprivileged Molecules. *ACS Sustainable Chemistry and Engineering* **7**, 2414–2428 (2019).

34. Broadbelt, L. J., Stark, S. M. & Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Industrial and Engineering Chemistry Research* **33**, 790–799 (1994).

35. Kim, S. *et al.* PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research* **47**, D1102–D1109 (2019).

36. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Research* **46**, W563–W570 (2018).

37. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

38. EPA, U. S. User's Guide for TEST (version 4.2)(Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure. (2019).

39. Martin, T. Toxicity estimation software tool (TEST). *US Environmental Protection Agency: Washington DC* (2016).

40. Screnci, D. *et al.* Relationships between hydrophobicity, reactivity, accumulation and peripheral nerve toxicity of a series of platinum drugs. *British Journal of Cancer* **82**, 966–972 (2000).

41. D. Cronin, M. The Role of Hydrophobicity in Toxicity Prediction. *Current Computer Aided-Drug Design* **2**, 405–413 (2006).

42. Veith, G. D. Relationships Between Descriptors For Hydrophobicity And Soft Electrophilicity In Predicting Toxicity. *SAR and QSAR in Environmental Research* **1**, 335–344 (1993).

43. Cheng, T. *et al.* Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *Journal of Chemical Information and Modeling* **47**, 2140–2148 (2007).

44. Pedregosa, F. *et al. Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

45. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* vol. 28 27–30 (2000).

46. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research* **46**, D633–D639 (2018).

47. Willke, T. & Vorlop, K. D. Biotechnological production of itaconic acid. *Applied Microbiology and Biotechnology* vol. 56 289–295 (2001).

48. Klement, T. & Büchs, J. Itaconic acid - A biotechnological process in change. *Bioresource Technology* vol. 135 422–431 (2013).

49. Jang, Y. S. *et al.* Bio-based production of C2-C6 platform chemicals. *Biotechnology and Bioengineering* vol. 109 2437–2459 (2012).

50. Chheda, J. N., Román-Leshkov, Y. & Dumesic, J. A. Production of 5-hydroxymethylfurfural and furfural by dehydration of biomass-derived mono- and poly-saccharides. *Green Chemistry* **9**, 342–35 (2007).

51. Ventura, S. P. M. *et al.* Evaluating the toxicity of biomass derived platform chemicals. *Green Chemistry* **18**, 4733–4742 (2016).

52. Loganayagi, C., Kamal, C. & Sethuraman, M. G. Opuntiol: An active principle of Opuntia elatior as an eco-friendly inhibitor of corrosion of mild steel in acid medium. *ACS Sustainable Chemistry and Engineering* **2**, 606–613 (2014).

53. Kano, H. *et al.* Carcinogenicity studies of 1,4-dioxane administered in drinking-water to rats and mice for 2 years. *Food and Chemical Toxicology* **47**, 2776–2784 (2009).