Identification of strategic molecules for future circular supply chains using large reaction networks

Jana Marie Weber,[1] Pietro Lió[2*] and Alexei A. Lapkin[1,3*]

[1]*Department of Chemical Engineering and Biotechnology, University of Cambridge, West Cambridge Site, Philippa Fawcett Drive, Cambridge CB3 0AS, UK*

[2]*Department of Computer Science and Technology, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK*

[3]*Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

Abstract

Networks of chemical reactions represent relationships between molecules within chemical supply chains and promise to enhance planning of multi-step synthesis routes from bio-renewable feedstocks. This study aims to identify *strategic molecules* in chemical reaction networks that may potentially play a significant role within the future circular economy. We mine Reaxys®[†] database in order to assemble a network of chemical reactions. We describe molecules within the network by a portfolio of graph theoretical features, and identify strategic molecules with an isolation forest search algorithm. In this work we have identified a list of potential strategic molecules and indicated possibilities for reaction planning using these. This is exemplified by a potential supply chain of functional molecules from bio-waste streams that could be used as feedstocks without being converted to syngas. This work extends the methodology of analysis of reaction networks to the generic problem of development of new reaction pathways based on novel feedstocks.

Keywords: network of chemical reactions; network theory; strategic molecules; circular economy; process development

[*] Corresponding authors: Pietro Lio email: pl219@cam.ac.uk; Alexei Lapkin email: aal35@cam.ac.uk

[†] Copyright © 2019 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.

1. Introduction

The development of greener and more sustainable processes has been a major focus of research and industrial activities in process development due to a number of important reasons: realisation of the contribution of industry on the increase of atmospheric carbon dioxide (CO_2), the depletion of reserves of fossil feedstocks, and increasing problems with waste management.[1,2] A key concept for the future sustainable society is the development of circular industrial structures[2–9] which requires the integration of waste process streams as feedstocks.[10–12] This results in feedstock components that are both new to the supply chain, and are highly functionalised. Provided we can solve the problem of separation of waste product streams, the remaining important question is where best to add these molecules into the current and the emerging circular industrial systems.

We recognize two common approaches to answer this question. First, is to focus on specific direct chemical transformation steps on the feedstock. Previous works have investigated, for example, valorisation of crude-glycerol,[13] conversion of lignocellulosic materials,[14] or hydration reactions on crude sulphate turpentine (CST) components,[15] among many others. The investigations of transformations of specific feedstocks into specific potential platform molecules can be exemplified by the extension of the work on 5-hydroxymethylfurfural (HMF), a molecule derivable from dehydration of fructose, to proposals of new functional platform molecules derived from HMF, such as its cyclopentanone derivatives.[16] The second approach is to follow a more product orientated approach. For instance, Couto gave an overview on the production of industrially relevant metabolites from biological waste,[17] Ravindran et al. outlined the utilisation of vegetable pomace to secondary use,[18] and Van Dyk et al. described compositions of food wastes, valuable compounds and potential products.[19] Both approaches are reasonable with regard to the specific feedstock, but neither consider whole supply chains, nor define precise criteria for these integration points. Without the consideration of the whole supply chain we are unlikely to answer which transformations and reaction pathways have a chance of being practically implemented.

More recently, analysis of reaction networks gained importance in this field. The analysis evaluates best multi-step synthesis alternatives for novel feedstocks. For instance, Zhang et al. investigated reaction networks for the evaluation of biomass-derived polymers.[20] Further, a network-based route evaluation for bio-fuel was performed by Ulonska et al. and later extended with the considerations of supply chain and multi-product decisions.[21,22] These studies pave the way for the consideration of bio-based feedstocks in extended reaction chains. However, the built networks are relatively small and focused on specific applications. They represent the specific problem sets well, but seem less suitable for considerations of whole supply chains or combined branches of industry. Thus, a model built for a larger chemical space appears more applicable.

An abstraction of all known organic synthesis, the Network of Organic Chemistry (NOC), was previously introduced by multiple works in Grzybowskis group[23–26] and described as the "Chemical Internet". Its components are nodes, which represent molecules, and edges, that show chemical reactions from one molecule to another. It consists of a core, a periphery, and unconnected islands. From a statistical perspective, the NOC is similar to other real-world networks.[27] When combined with heuristics, the algorithmic assembly of process routes and evaluation of specific ranking criteria was shown to be useful in multi-objective decision making on the potential process alternatives.[28] Our hypothesis is that a NOC can be used to identify suitable locations, i.e., suitable molecules, in the overall supply chains to introduce the molecular structures available in bio-waste feedstocks.

One way to integrate a bio-feedstock into chemical supply chains is through production of bulk chemicals or of molecules easily transformable into bulk chemicals.[29–31] However, a wrong choice may lead to system-level problems, such as incompatibility of scales (availability of feed vs demand for bulk chemicals), increased demand for 'waste' triggering increased production of other materials, and so on. Another option is to synthesise molecules that can be used in especially short multi-step reactions,[24,32] that is to directly access complex molecular functionality from bio-feedstocks and avoiding building it up in the conventional petrochemical supply chain. Both possibilities may ensure a high contribution of the bio-based molecules in the overall chemical supply chain. Either way, we can say that a bio-feedstock should enter a reaction network in an *optimal* location - be transformed into *strategic* molecules.

There are some works that regard the importance of such molecules in reaction systems. Welsch et al. and Schneider et al. describe the advantage of designing drug compounds from certain molecular frameworks, which they describe as privileged scaffolds.[33,34] The molecular frameworks are formed of geometries, which are suitable to be coupled with side chains. The final product may then bind to target proteins. The suitability of a scaffold is determined in two ways: it is either based on the promiscuity of the most observed scaffolds, which is expressed as Shannon entropy, or it is based on the maximal information content. These scaffolds are stressed for the development of biomedical applications.[33,34] In a very different approach, Aden et al., inspired by the petrochemical industry, focus their work on bio-based building blocks for future supply chains. They consider market-specific selection criteria as well as structural properties of the molecules.[35] More recently Serrano-Ruiz et al. combined flow-chemistry with bio-derived platform molecules. Also in their work market-specific criteria for the selection of bio-based platform molecules were named; e.g. based on availability of commercial technology for their production and for their potential to contribute to both fuels and chemical production.[36] Szymkuć et al. identify *hub* molecules in reaction networks by a popularity function.[24] This means that molecules were identified as hubs due to the number of links in the network and then these molecules were favoured during synthesis planning.[24] All aforementioned approaches have the same aim: they desire general descriptions of *useful* molecules for broad sets of applications. With the goal to extend the application sets to the whole chemical supply chain, we find the method of Szymkuć et al. well suited, because it operates on whole network structures.

In this study we adopt a graph theoretical approach for the identification of hub molecules. The number of links determines if the molecules classify as a hub. In such link-based approaches, important molecules can be detected by methods such as stochastic block models[37,38] or functional network embeddings, such as SCAN[39] or SDNE.[40,41] We argue that the number of links is not the only characteristic that defines a strategic location in a network. Considering, for example, also the links' importance, or the overall shortest pathways through a molecule, provides a more accurate graph theoretical description of a useful molecule for a broad set of applications. In our work, we consider the links' importance and include the number of possible syntheses paths through a molecule as additional characteristics by which we characterise a strategic molecule.

In other fields of network theory, lists of multiple features are used to describe the importance of nodes. For instance, in brain networks[42–44] and web page searches in the Internet,[45,46] different aspects, such as centrality measures are used to classify the nodes. Against this background, integration of such centrality measures may be beneficial for the identification of strategic molecules in chemical networks.

In this work, we use a portfolio of graph theoretical measures to identify strategic molecules in the NOC. If a molecule is a bulk chemical itself, or directly connected to one, its degree, pagerank and HITS y-hubs value will be different to the values of most other nodes. The betweenness centrality is used to measure the efficiency of reaction paths through a molecule. An isolation forest algorithm is applied to identify potential strategic molecules based on a vector of graph metrics. The strategic molecules are evaluated by comparison with common industrial intermediates. Further, we propose a method of reaction path screening over the strategic molecules. Based on these tools, we developed a case study of integration of crude sulphate turpentine (CST) into a supply chain of sample high-value end products, e.g. pharmaceuticals.

The remaining article is organized as follows. In Section 2, we introduce graph theoretical terminology, the assembly of the network, and the measures used. Further, we outline the isolation forest outlier detection algorithm and the pathway screening method. In Section 3, we present results throughout our workflow and finally, discuss the strategic molecules within chemical context. We use a case study to give a qualitative evaluation on the use of strategic molecules in process development.

2. Methods

2.1. Graph theory of chemical reaction networks

In a chemical supply chain, simple molecules are transformed into more functionalised molecules along specific reaction paths. These reaction paths can be illustrated in a network of nodes and edges. The all-to-all wiring scheme, connecting all reactants to all products, and the one-to-one mapping scheme, where only the heaviest reactant is mapped to the heaviest product, have shown to give comparable properties.[23] Different representations of the all-to-all wiring scheme have been introduced in the literature,[25,26,47,48] from which we outline some in Figure 1. With the objective to identify strategic molecules in the network we find the simplified directed network without parallel edges most suited. When practical intersections are desired, information on reaction partners are not required, which leads us to the first simplification from (b) to (c) in Figure 1. Further, duplicate parts of reactions as shown in (a) where molecule A can lead to product B in reaction (1) and in reaction (3) do not enclose additional insight. Thus, we perform the second simplification from (c) to (d) in Figure 1.

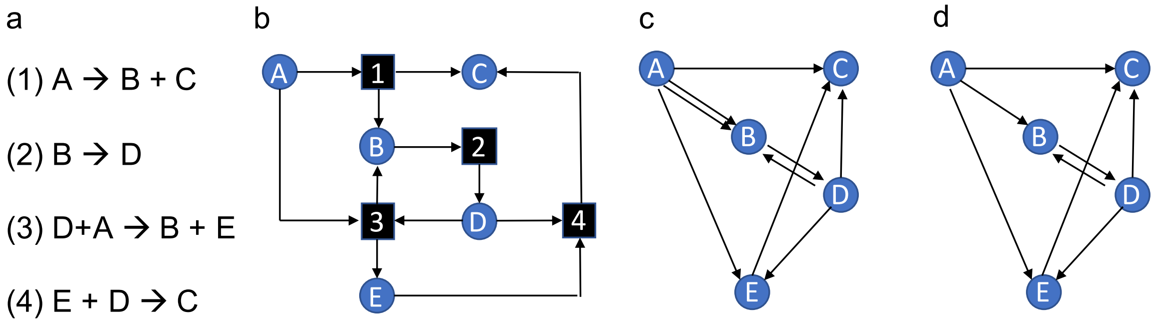


Figure 1. An illustration of reaction representation in different network types. (a) An example representation of a set of chemical reactions. (b) A bipartite network stores information about reaction partners and products. (c) A directed network shows relationships from reactants to products. (d) A directed network without parallel edges further simplifies the reactant-product relationship.

Networks are used to capture real-world problems that often have interactions between large numbers of objects or subjects. A chemical supply chain is a good example of such interacting systems. A simple illustration of a graph is seldom sufficient to understand the relationships within a complex system. In network science a set of metrics is commonly used to describe networks on global and local scales. There are different types of networks and many metrics to characterise them. In this work we describe networks based on their degree distribution and we consider centrality measures to label nodes. For a more profound understanding of network types and theories of network science, we refer the reader to further literature.[49,50]

The main characteristic of a node is its linkage in the network. The number of edges leaving one node is called 'out degree', while the number of incoming edges is called 'in degree'. In scale-free networks the degree distributions are best modelled by power law correlations,[51] meaning that only few nodes have a high in and/or out degree while the other nodes have a comparable low degree. The probability of finding a node with degree k in the network follows Eq 1:[51]

$$P(k) \sim k^{-\gamma} \tag{1}$$

where γ depends on the network under consideration.

Directly connected nodes are neighbours of each other. Nodes reached by an out degree of node u are out neighbours of u, vice versa, nodes reached by an in degree of node u are in neighbours of u. A connection of two nodes, a pair of nodes, in a network is called a path. If every edge on the path is only traversed on once, it is named trail, if a path is the shortest possible connection between two nodes it is called shortest path or geodesic.[50–52]

2.2. Methodologic workflow

While the question of strategic molecules mainly concerns researches within the fields of chemical reaction engineering, the underlying question of finding optimal locations in a network is of considerably broader interest. The presented workflow, see Figure 2, for the identification of strategic molecules is applicable to diverse network problems. The first step of data mining may be performed manually or automatically from any possible data resource. This work is based on an automated download routine from the Reaxys[53] application programming interface (API), which is accessible via an Elsevier license. However, any database with a sufficient number of chemical reactions may be used as a source for a chemical reaction network. With regard to the more general question of strategic locations in networks, data sources outside of the field of chemical reaction engineering may be used. A preferable way of data representation is a network structure as a large number of local and global interactions are covered. A description of all nodes in the network by graph theoretical features is performed via feature engineering. It is important to note that humanly designed features hold both negative and positive implications. On the one hand they introduce human biases to the system, on the other hand they live from prior knowledge, which may be favourable over black-box models. The chosen features characterise a specific question posed from the view of chemical reaction engineering and aim to capture both local and global structures. For arbitrary other network problems, the features may be designed differently.

In the fourth step, we take advantage of the topology of scale-free networks. With regard to the features of interest optimal locations in the network are located in the tail of the power law distribution and can hence be identified as outliers. The method is applicable to common networks as long as the optimal locations rank extremely different in the chosen features. The evaluation step combines field-specific knowledge and graph screening. The following sections explain each part of the methodology in more detail and the Electronic Supplementary Information (ESI) provides information about the general applicable part of the pipeline, which can be found on GitHub.

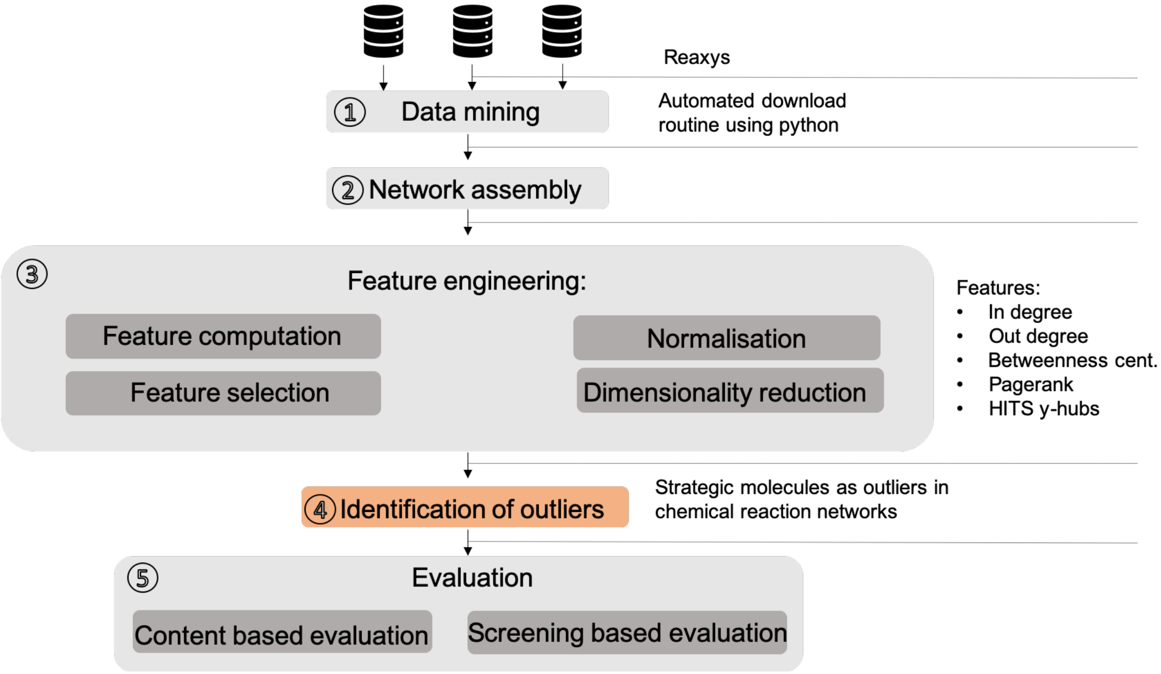


Figure 2. A pipeline for the identification of optimal locations in scale-free networks. The methods in the pipeline are generically applicable. We display our specifications for the identification of strategic molecules on the right-hand side.

2.3. Data mining and Network assembly

The assembly of networks of organic chemistry requires a set of reactions in the chemical region of interest. The set of chemical reactions may be assembled based on chemical knowledge and a literature review, or by mining chemical databases. While this can be performed with reasonable effort on small data sets, it requires an automated routine when regarding large scale interactions. The data for the examined network was obtained using the Reaxys API and an automated download script previously developed in the group.[47] Reaxys records reactions based on Reaxys reaction IDs and the Reaxys IDs of molecules. Every molecule is described by its unique Reaxys ID. In theory, both structural and stereo isomers are assigned different Reaxys IDs. However, in some instances a generic structure of a molecule is described without isomeric specifications. This study does not concern the validation of recorded molecules, their structural arrangements or the further reaction specifications, e.g. yield or selectivity. This work is based on the overall possibility that certain reactions occur and, hence, provides ideas for early stage process development. Further information on the download routine, computer architecture, and versions of python and the main packages is available in the ESI.

2.4. Feature engineering

Ensuring a high contribution of the bio-based molecules in the overall chemical supply chains, requires finding of optimal locations, hence strategic molecules, in the network structure. A molecule may classify as a strategic molecule if it shows to be preferably linked in the network. The linkage might lead to shorter synthesis paths, to commonly used bulk molecules, or scaffolds suitable to connect different regions of chemistry. We wish to describe the characteristics of strategic molecules by their linkage in a graph theoretical framework.

In this work we use graph theoretical metrics to characterise each molecule, i.e. each node, in the graph. The characterisation is based on:

i. the nodes centrality position in the reaction network,
ii. the nodes participation in reactions, and
iii. the nodes direct linkage to molecules that participate in many reactions.

The characteristics are useful to describe local hub behaviour, e.g. the nodes participation in reactions (characteristic ii), and global influences, e.g. the centrality position (characteristic i) and importance-based linkages (characteristic iii).

The central position in the network is described by the betweenness centrality of each node. Betweenness centrality, $C_B(v)$, finds shortest paths, as defined in Section 2.1, between all pairs in the network and counts how many of the shortest paths lead through a specific node. The final result is normalised by the number of nodes in the network. Betweenness centrality, $C_B(v)$, is defined as:[54,55]

$$C_B(v) = \frac{\sum_{s \neq v \neq t \in V} \sigma_{st}(v) / \sigma_{st}}{N} \tag{2}$$

where $\sigma_{st}(v)$ is the number of paths from *s* to *t* via node *v*, while σ_{st} is the number of all shortest paths from node *s* to node *t*. *V* is the set of all nodes in the graph and *N* is the number of all nodes.

The participation of molecules in reactions (our second characteristic of a node) is evaluated by the degrees of a molecule. For the directed reaction network, we regard both in and out degree. The different types of degrees have been briefly introduced in Section 2.1.

The connection to highly linked nodes in a network (characteristic iii) is measured by the pagerank, and the HITS y-hubs value.[56] The pagerank and HITS y-hubs of a node depend on the respective values of linked nodes. This is solved by multiple iteration steps in the algorithms. Both measures start by assigning a uniform weight distribution to the nodes. The algorithms compute the pagerank and HITS y-hubs of each node based on the according values of the other nodes from the previous iteration step. If the value of each node changes less than the convergence limit of $1 \cdot 10^{-6}$ in a step the final distribution is found. The weights in pagerank are defined due to the incoming links of the in neighbour of a node. The weights for HITS y-hubs are defined due to the outgoing links of an in neighbour. The pagerank of a node was first introduced in Ref 46 and is implemented in Ref 54:

$$PR(v) = \frac{1-d}{N} + d \sum_{u \epsilon \Gamma^{-(v)}} \frac{PR(u)}{d^{+}(u)} \tag{3}$$

where *d* is a damping factor, *N* is the number of all nodes in the network, *u* is a in neighbour of *v*, $\Gamma^{-(v)}$ is the set of in neighbours of *v*, and $d^{+}(u)$ is the out degree of *u*. Explanations on the different types of degrees and neighbours are given in Section 2.1. The damping factor in the graph-tool implementation[54] is set to 0.85 and the default convergence limit of $1 \cdot 10^{-6}$ is used. The damping factor is used in Ref 46 to account for random changes of websites. With regard to the network, this allows the algorithm to work even if the network has dead-ends and spider traps.[57] The HITS y-hubs, *y(v),* and x-authority, *x(v),* of a node were defined in Ref 45 and an implementation to compute the vectors **x** and **y** of all nodes was done in Ref 54:

$$\boldsymbol{x} = \alpha A \boldsymbol{y} \tag{4}$$

$$\boldsymbol{y} = \beta A^T \boldsymbol{x} \tag{5}$$

where A is the adjacency matrix of the network, and α and β are scaling factors. In Ref 54 β is set to 1 and α the reciprocal value of the largest eigenvalue of the systems cocitation matrix AA^T. The HITS values become especially important when considering nodes with same amount of in degrees. The origin of the in degrees is considered; if it is a hub page, the link is weighted stronger.

Hence, a feature vector, ***z(v)***, is composed of the betweenness centrality, the in degree, the out degree, the pagerank, and the HITS y-hubs of a node. To prevent biased influences of the different metrics, we scale each feature between zero and one, using a maximum/minimum normalisation.[53] For data compression and visualisation purposes, we reduce the dimensions of the input vectors. We performed a principal component analysis (PCA) based on a singular value decomposition as implemented in Ref 58 and obtain a reduced feature representation ***z'(v)***.

2.5. Identification of strategic molecules

Based on ***z'(v)***, an outlier detection algorithm for the identification of the strategic molecules is performed. The feature values in the network are best described by a power law distribution. This means, that most nodes have very similar feature values, but the ones that are different differ a lot. They are exceptional in their feature value. With regard to the considered features, we find that sample strategic molecules rank very high. However, we do not suggest using the absolute values for finding feature specific thresholds for the classification; these will depend on the network size and centre, and introduce more human bias. Instead, we classify them based on the feature distribution.

The strategic molecules differ with their feature values and can thus be regarded as outliers. To detect deviations in all feature values, we use an isolation forest algorithm. In theory, clustering techniques, e.g. clustering techniques for large data sets, may be suited for the same task. However, we encounter many drawbacks using clustering algorithms on a large data set which follows power law distributions. Only naming few: the heterogeneity in the data density leads to especially uneven cluster sizes, distance metrics do not support odd shapes, and there are difficulties finding appropriate mesh sizes for heterogenic data densities for using grid-based clustering approaches.[59,60] We argue that an isolation forest algorithm is well suited because no distance metric is used and the algorithm can take advantage of the heterogenic data density, as anomalous samples are detected easier. The algorithm works on an ensemble of isolation trees, where each branch randomly selects a feature and a feature value between its minimum and its maximum.[61] The decision tree is divided until one data point is isolated from the rest. The average number of branches, in other words the average length of the trees in the forest, is used to characterise the sample. Normal samples require longer isolation steps, while anomalous ones have on average a shorter tree length.[61] Yet, the isolation forest algorithm is a non-deterministic method. Each run of the algorithm on the same data set might result in a slightly different classification of the molecules very close to the border region of in- and outliers as the split is created by randomly branching the tree. Figure 3 illustrates the difference in the required tree length for the isolation of a normal, x_i, and an anomalous, x_x, data point. The algorithm is used as implemented in the python module Scikit-learn.[58] For details on parameterisation see ESI.

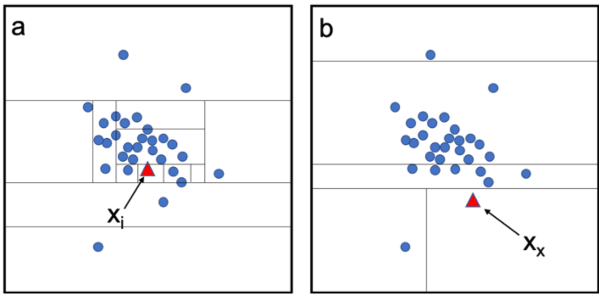


Figure 3. Visualisation of the isolation forest outlier detection algorithm. For an inlier, x_i, which lays in the normal data region, many branches are needed to separate the data point from the rest (a). For an outlier species, x_x, fewer branches isolate it from the rest of the data (b).

2.6. Screening based evaluation: Reaction pathway search over strategic molecules

The second focus of this work is to suggest the integration of the strategic molecules in pathway screening when new chemical processes are developed or, more specifically, an alternative feedstock is sought to produce a known product. A chemical multi-step synthesis from a feedstock to an end product can be resembled by a path in a network from the feedstock node to the end product node. Assembling all sets of the possible paths is commonly done by a screening method, for example depth-first (DF) screening.[62] The total number of paths scales with *O(N!)* where *N* is the number of nodes in the screening set and the algorithm has a time complexity of *O(N+E).*[54]

In this work we propose to include strategic molecules into screening algorithms. We combine a DF search as implemented in Ref 54 from feedstock components to strategic molecules with a DF search from strategic molecules to end products. We call this approach reduced-path depth-first (RPDF) search. We constrain the maximal search depth for both the DF and the RPDF search by cut-off values. A cut-off of one means that only one reaction step from the starting molecules is permitted, a cut-off of two allows two reaction steps and so forth.

3. Results and Discussion

To the best of our knowledge, there is no definitive set of strategic molecules comprising current important industrially-produced molecules and the promising future molecules, or alternatives to the existing one. The topic of product substitution is regularly discussed in the chemical and environmental literature within the contexts of green chemistry and sustainability, and a number of bulk intermediates were proposed for the emerging bio-based chemical supply chain.[35,29] This, however, is a much narrower set than the proposed set of strategic molecules that would cover bulk as well as functionalised intermediates. In the absence of such a list we cannot evaluate our approach by comparing the outputs. One inherent challenge for the identification of strategic molecules is that there is no general 'true' answer. Thus, there is no quantitative method to evaluate our results. Accordingly, we evaluate our results qualitatively. For this we show that common industrial chemicals, important precursors, and potential new building block molecules are at least the subsets of our findings. We consult Ullmanns encyclopaedia of industrial organic chemistry[63] for the evaluation of common industrial intermediates. Additionally, we highlight similarities of our results and the highest ranked privileged scaffolds for biomedical applications.[34] To indicate the algorithms' ability to identify not only petrochemical based strategic locations we consider potential future bio-based building blocks.[35] We further demonstrate the use of strategic molecules in reaction pathway searches by using DF searches on the graph structure to indicate possible multi-step reactions.

3.1. 4-HAP network

The network of interest in this case study is a network centred around the sample strategic molecule, 4-hydroxyacetophenone (4-HAP), referred later as '4-HAP network'. 4-HAP is a common precursor for many pharmaceutical products such as paracetamol, metoprolol, and salbutamol, and can be derived from terpenes, including mixed waste streams such as CST.[64] The 4-HAP network is a scale-free, non-weighted, and directed graph. It consists of 568,270 nodes and 955,905 edges. The degree distributions in the network follow a power law behaviour with the exponents γ_{out}=2.1, and γ_{in} =2.3 (see Equation 1).

3.2. Identification of strategic molecules

We describe each node in the 4-HAP network by ***z(v)*** as described in Section 2.4. After dimensionality reduction via PCA, it was found that two principal components (PCs) express 98% of the data variance. PC0 takes up 79% of the data variance and PC1 19%. Due to the very high coverage of the data variance, we now describe each node with ***z'(v)*** composed of only two components, PC0 and PC1, see Section 2.4. Using the isolation forest algorithm as described in Section 2.5, we demonstrate separation of the outliers. A variety of contamination rates was tested, and a good fit was found by visual inspection. Figure 4 shows the region of the data distribution where both normal and outlier nodes are present (split region) over PC0 and PC1 of ***z'(v)***. Normal species are depicted by red triangles, while blue circles represent the outliers. The aim was to find the best separation between the sparse and the dense regions of the data distribution.

Figure 4 shows some of the tested rates. We find the rates between $0.9\cdot10^{-4}$ and $1\cdot10^{-3}$ to best describe a cut between the sparse and the dense region. With a rate of $1\cdot10^{-3}$, 569 molecules are detected as outliers while a rate of $0.9\cdot10^{-4}$ leads to a classification of 512 molecules as outliers. In the following we consider the contamination rate $1\cdot10^{-3}$ for further discussion. We argue that this list should not be seen as a fixed framework on strategic molecules, but as an ensemble of potentially useful molecules for present and future process developments. Hence, we consider the largest suggested useful subset for further discussion. The list of all 569 identified molecules is given in the ESI.

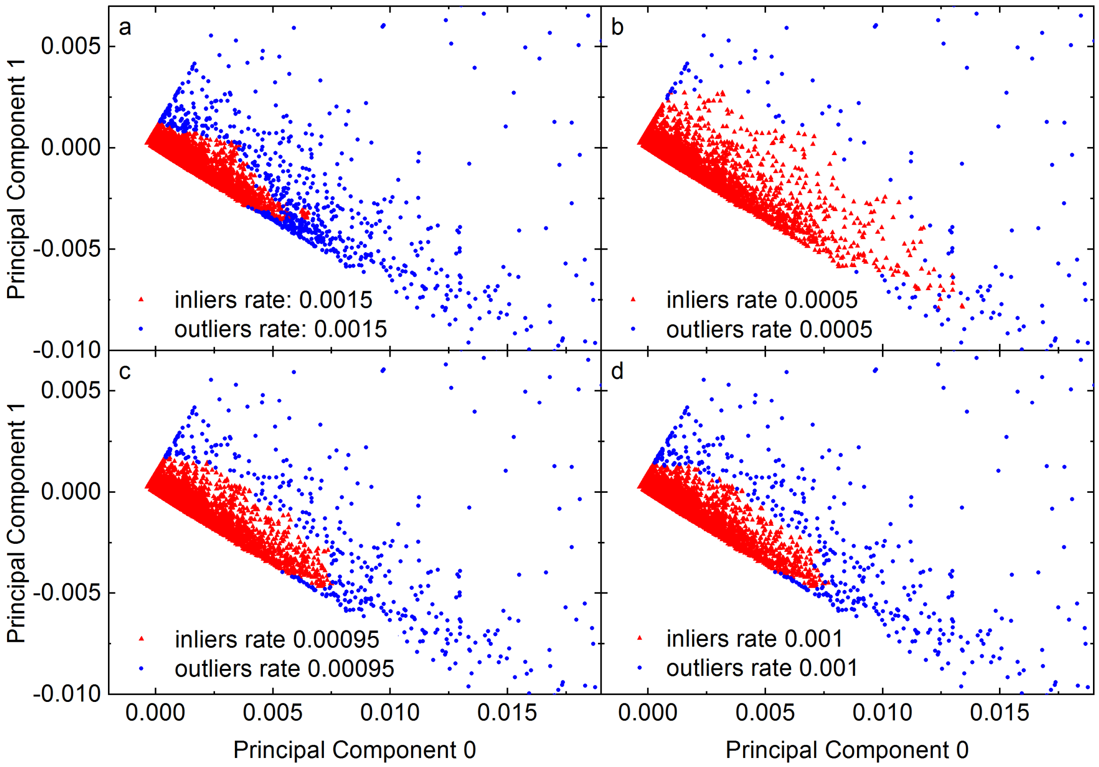


Figure 4. The split region of the normal and the anomalous data of the 4-HAP network identified by the isolation forest algorithm. Red triangles are data assigned to the normal region and blue circles resemble data not grouped with the majority. The isolation forest was performed with a contamination rate of (a) $1.5 \cdot 10^{-3}$, (b) $5 \cdot 10^{-4}$, (c) $9.5 \cdot 10^{-4}$, (d) $1 \cdot 10^{-3}$.

3.3. Evaluation of strategic molecules

We first consider if the list of strategic molecules includes current common industrial intermediates. While searching for strategic molecules we expect to find platform molecules, as well as commonly used co-reactants, because most of the current industrial supply chains are built on these.[65] Here, we discuss some chosen strategic molecules and demonstrate their industrial relevance in the past and the present.

Among the identified strategic molecules, we find ethylene derivatives, e.g. acetaldehyde, ethanol and styrene. Acetaldehyde is a precursor to many further products, e.g. butadiene. Today butadiene is in high demand, but it also had a historical importance in production of synthetic rubber, for which it was produced during World War II. Styrene is widely used in polymer production (homopolymers, copolymers or rubber-modified polymers), while ethanol was a raw material for production of acetic acid; it is now used as a solvent in cleaning, cosmetics and coatings industries, as industrial solvent and reactant, and is also converted to white vinegar.[63] Furthermore, ethanol is used as an octane-enhancer or petrol replacement in the automotive fuels market.[66] All these chemicals have shown major industrial significance in the past and present and are, therefore, correctly identified by the algorithm as strategic molecules. Further, the algorithm detects acetone and phenol as derivatives from propylene. Both are main products from cumene and lead to high value supply chains including the production of nylon 6, epoxy or phenolic resins, polycarbonates, methylmethacrylate, and different solvents.[63] The broad range of uses supports the argument that acetone and phenol should be strategic molecules and, hence, are correctly identified by the algorithm. Acetic and succinic acids are two examples of the derivatives from the C_4 stream that are found by the algorithm. The industrial use of acetic acid includes solvent, precursor to acetate monomer, acetic anhydride, and esters. Succinic acid is used in the food and beverage industry and is a precursor for polymers and solvents.[63] Incidentally, succinic acid is one of key bulk molecules targeted in bio-refining. In summary, the graph-based search finds a selection of important industrial chemicals and shows that different sections of chemical space are covered. Most useful, we notice that molecules at junctions, connecting different subdivisions of industry, are detected.

In addition, the algorithm detects molecules in more specialised branches of chemistry. Benzoyl peroxide as a component for polymerisation reactions is found. It generates free radicals and can be used in many different polymerisation reactions.[67] Moreover, 4-HAP as a precursor for pharmaceutical products is identified. Within few steps, main pharmaceutical compounds can be derived.[64] Naming one last interesting output: 1,1,2,2-tetraphenylethylene is found, which is a potential building block for the assembly of supramolecular frameworks.[68] These frameworks are studied with regard to many different applications, for example, they can be used as metal-organic frameworks with applications in gas adsorption or medical science. Additionally, the algorithm identifies a range of small molecules as CO_2, O_3, CO, and H_2O. These are co-reactants for many chemical reactions, thus highly linked, which explains their appearance as strategic molecules. Moreover, the algorithm detected sucrose and glucose derivatives thus also identifies more functionalised, and three-dimensional structures.

By evaluating the output with chemical knowledge and intuition, we can indicate that common intermediates from industry, small co-reactants for many reactions, and complex and novel structures used in research are respectively identified. We detect a broad range of different chemical scaffolds and find intermediates of historic as well as current relevance. In Figure 5 we show a selection of strategic molecules identified by the isolation forest algorithm.

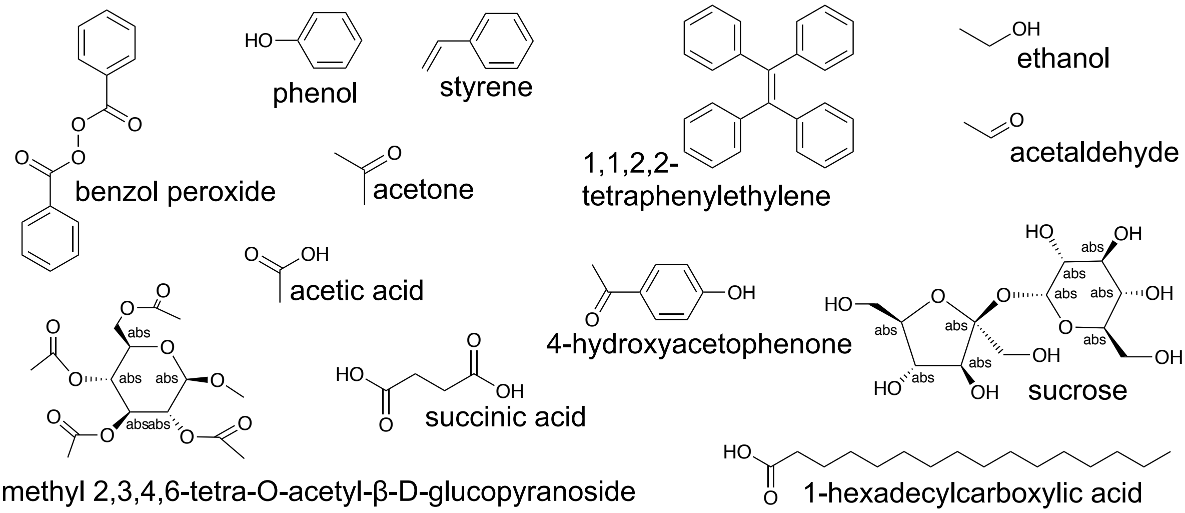


Figure 5. Potential strategic molecules identified by the algorithm with a contamination rate of $1 \cdot 10^{-3}$.

Further, we compare our results with two relevant works in the field. A similar concept of strategic molecules for the design of biomedical applications are so-called privileged scaffolds.[33,34] These molecular frameworks are formed of geometries, which are suitable to be coupled with different side chains. We suggest that highly ranked privileged scaffolds, due to their Shannon entropy, should be a subset of strategic molecules, as these might represent molecules at important junctions in the network. With regard to the highest ranked five scaffolds (see Figure 6), the isolation forest algorithm identifies four of these and a related structure to the fifth one. Quinoline, diphenylmethane, diphenylether, and (benzyloxy)benzene were identified. N-benzylaniline had not been identified, while the algorithm detected a related molecule: N-phenyl benzoyl amide.

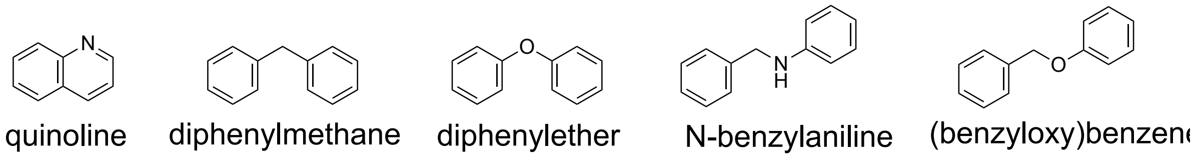


Figure 6. Top-ranked molecular frameworks for medical chemistry based on Shannon entropy.[34]

The third concept describes potential bio-based platform molecules and was introduced, among others, by the National Renewable Energy Laboratory (NREL).[35] The novel building blocks are described as molecules with diverse functional groups that may enable transformation into new families of useful chemistry. Out of more than 300 candidates, 30 molecules were selected based on the model of petrochemical building blocks. We find 27 of these in Reaxys database and 25 in the 4-HAP network. The isolation forest algorithm identified 11 of the aforementioned 25 as strategic molecules. The NREL building blocks were selected based on strategic criteria, e.g. if they are suitable to compete against existing products or if they possess characteristics that can replace current functionality or even give rise to new applications. The 4-HAP network consists of a majority of petrochemical-based reactions and has already shown to identify common industrial intermediates. Against this background we find it highly encouraging that the algorithm detected 11 out of 25 strategic molecules in the given network.

Table 1 gives an overview on the discussed molecules.

Table 1. The NREL Building Blocks which are in Reaxys database and also in the 4-HAP network (4HAP) and identified as strategic molecules (SM).

Molecule	4HAP	SM	Molecule	4HAP	SM	Molecule	4HAP	SM
carbon monoxide	x	x	fumaric acid	x		proline	x	
hydrogen	x	x	malic acid	x		xylitol	x	
glycerol	x	x	succinic acid	x	x	xylonic acid		
3-hydroxypropionic acid	x		threonine	x		aconitic acid	x	
lactic acid	x	x	arabinitol			citric acid	x	x
malonic acid	x	x	furfural	x	x	glucaric acid	x	

propionic acid	x	x	glutamic acid	x		lysine	x	
serine	x		itaconic acid	x		levoglucosan	x	
aspartic acid	x		levulinic acid	x	x	sorbitol	x	x

3.4. Reaction pathway searches over strategic molecules

Here we illustrate the application of the concept of strategic molecules for selection of reaction pathways for potential valorisation of waste CST. CST is a by-product form the Kraft pulping process in the paper processing industry. It consists of an oil mixture of unsaturated and volatile $C_{10}H_{16}$ terpene isomers: α- and β-pinene, Δ-carene, and a mixture of other terpenes,[69] e.g. limonene. Previous work has focused on transformations on CST to fragrance chemicals and into precursors for the pharmaceutical industry.[70,15]

This study uses the strategic molecules previously identified on the basis of 4-HAP network and one other network for the pathway searches. The second network is larger and is centred around one of the feedstock components, limonene. In the following it will be called the "limonene network". 93.25 % of the molecules from the 4-HAP network are included in the limonene network. The main reason for the extension of the network is to achieve a higher coverage of feedstocks and end products regions. The limonene network consists of 12,238,931 nodes and 24,884,365 edges. Its exponents for the degree distributions are γ_{out} =2.0 and γ_{in} =2.5 (see Equation 1). We now perform pathways searches in the described limonene network. We consider three different experiments to show possibilities of integrating the concept of strategic molecules in process development.

Table 2. Designed experiments in the CST case study.

	Description
Experiment 1	DF searches from CST components to a set of strategic molecules and sets of randomly chosen test molecules are conducted. The cut-off values one, two, and three are investigated.
Experiment 2	RPDF searches are conducted over strategic molecules to compare them with each other. Three steps from CST to strategic molecules and

	three steps from the strategic molecules to pharmaceuticals are tested.
Experiment 3	DF and RPDF screening are compared with regard to the number of paths and CPU times needed for each screening.

In experiment 1 we compare strategic molecules with sets of randomly chosen molecules regarding their connectivity from CST feedstock. Figure 7 outlines the pathway search for comparison. We consider three CST components (pinene, limonene and carene) and sets of 569 molecules. We wish to present connectivity in relatively short reaction paths and, hence, stop screening at a cut-off value of three, which is half of the average connectivity in the network.[27] We compare the set of strategic molecules to ten sets of randomly chosen molecules. The aim is to count the number of molecules in the sets that can be reached from the feedstock within each cut-off value, the number of reaction steps. We show the average over ten sets in comparison with the strategic molecules in Figure 8.

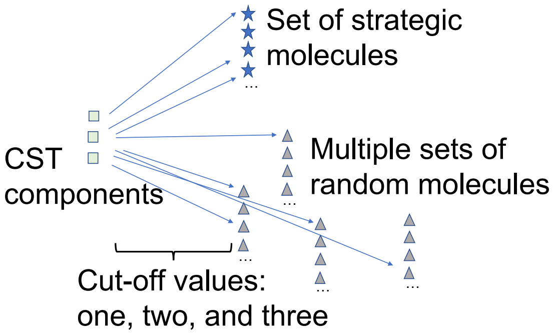

Set of strategic molecules
CST components
Multiple sets of random molecules
Cut-off values: one, two, and three

Figure 7. Scheme for the comparison with sets of random molecules. All sets have the same size. Squares represent the crude sulphate turpentine (CST) feedstock components, stars - the strategic molecules and triangles - the randomly chosen test molecules. One, two, or three reaction steps were allowed.

We find that the CST components are significantly better connected to the strategic molecules than to the sets of random test molecules. Already within the range of one step paths, we connect CST components to 27 out of 569 strategic molecules, some of which with major industrial significance, e.g. benzene, acetophenone, propene, ethene, toluene, and isoprene. Especially for a cut-off value of two, the ratio of connected strategic molecules shows substantial benefit. This becomes less noticeable when allowing more steps. The reason for this is the so-called small-world phenomena.[27] The NOC shows that on average each pair of nodes can be connected in six steps. Thus, differences are most noted in very small screening ranges. Considering a cut-off value of three, we find all strategic molecules connected and 57% of the molecules in random sets being connected. This experiment strongly indicates the importance of strategic molecules as they are better connected to the exemplary feedstock. Further this demonstrates the possibility of rapid screening as an early stage decision making in process development for waste streams. The very small set of one-step reactions can be analysed manually. Figure 8 shows the described results for the sets of randomly chosen test molecules. The standard deviation for cut-off two is 0.6% and the standard deviation for cut-off three is 1.5%. The x-axis shows the cut-off values and the y-axis shows the percentage of molecules in the set that can be reached from CST.

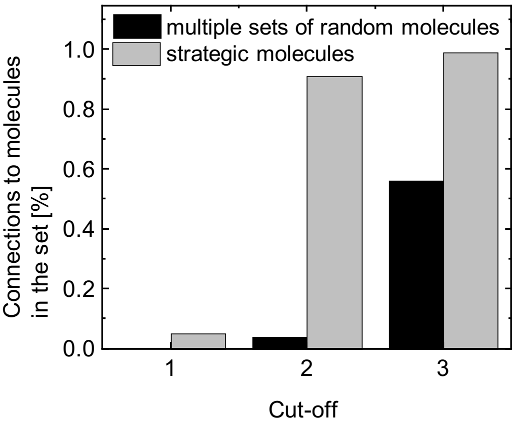


Figure 8. The results of depth-first searches from three components of CST (pinene, carene, and limonene) to sets of molecules are shown depending on the allowed cut-off value. All sets consist of 569 molecules. The set of strategic molecules was assembled by a contamination rate of $1 \cdot 10^{-3}$ on the 4-HAP network. For each cut-off value, ten sets of 569 randomly chosen molecules are assembled.

In experiment 2, we show a possible use of strategic molecules in a scenario that transforms the feedstock to a range of end products. We chose a range of pharmaceuticals as end products based on Ref 71. The connectivity between feedstock components, strategic molecules and end products is investigated by a RPDF search following the scheme outlined in Figure 9 and the description below. Squares represent feedstock components and crosses show end products; a star denotes a strategic molecule. The linkage of strategic molecules is evaluated with regard to:

1. the ratio of feedstock components that can be connected to the strategic molecule,
2. the maximum number of alternative paths to this strategic molecule in comparison to maximum number of alternatives to other strategic molecules,
3. the ratio of end products connected from the strategic molecule, and
4. the maximum number of alternative paths from this strategic molecule to the end products in comparison to maximum number of alternatives from other strategic molecules.

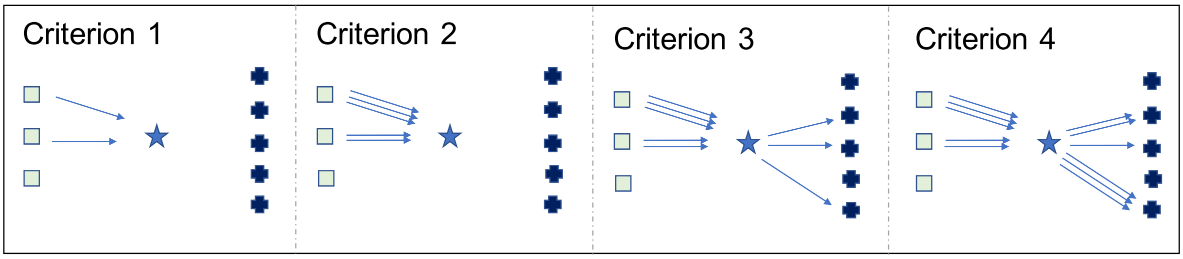


Figure 9. Four criteria for RPDF searches are shown. Squares are feedstock components, the star is a strategic molecule that is tested, and crosses represent a selection of end products. The first criterion measures the ratio of feedstock components connected to the strategic molecule. The second criterion counts the maximum number of paths for the connection from one feedstock to a potential strategic molecule. The third criterion measures the ratio of end products that can be reached from the potential strategic molecule. The forth criterion counts the maximal number of paths to the end products. Criteria one and three only consider the tested strategic molecule, while criteria two and four are set in relation to the maximum paths for other potential strategic molecules.

We test the connection from CST over 56 of the strategic molecules (identified by a contamination rate of $1{\cdot}10^{-4}$) to 115 pharmaceuticals. We allow up to three steps between feedstock components and strategic molecules and strategic molecules to end products. We find, that all feedstock components can connect to the 56 strategic molecules. We also find that many of them have more than one possible path. Both are indicated for selected strategic molecules by the two green bars with stripes in Figure 10. We also show that most of the 56 strategic molecules are also connected to many different pharmaceutical end products with multiple possible synthesis paths. This is indicated by light and dark grey bars in Figure 10 for selected strategic molecules. The experiment 2 demonstrates the suitability of strategic molecules as connections of different regions of chemistry in reaction networks. Further, it indicates a method of generating a smaller subset of most suitable molecules to be investigated for novel process development. The metrics applied on the RPDF search over the strategic molecules compares the strategic molecules with each other. Once again, the large space of organic chemistry is reduced to few options, making manual evaluation possible.

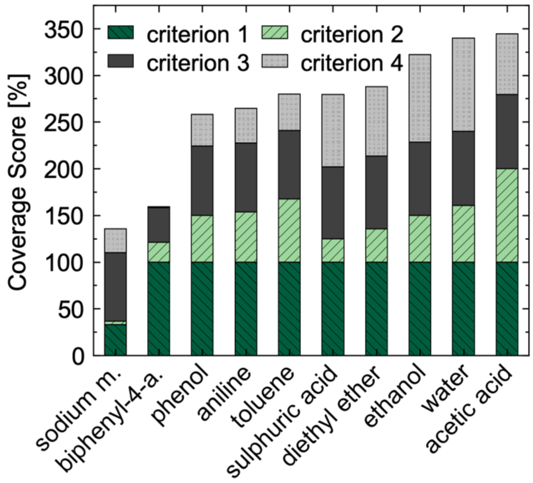


Figure 10. RPDF screening from CST over strategic molecules to a set of 115 pharmaceutical products. Within each bar, up to 100 % can be reached, hence the maximal value of the total score is 400 %. We show the screening over representative strategic molecules, where the abbreviation *sodium m.* stands for sodium methylate and *biphenyl-4-a.* stands for biphenyl-4-acetaldehyde.

The experiment 3 demonstrates the use of the RPDF screening over strategic molecules in comparison to DF searches. We consider a scenario where chemical space is screened for pathways from one specific component to a few end products. For the RPDF search, we screen from pinene, a constituent of CST, to 56 strategic molecules in a cut-off value of two and three. We then search for paths from the strategic molecules to ten selected pharmaceuticals in cut-off of range three. For the DF search, we screen up to a cut-off value of five.

Comparing DF with cut-off five and RPDF with cut-off five, we find that the RPDF algorithm finds a reasonable large subset of the overall possible paths screened by the DF search in noticeable shorter CPU times. The CPU times show a decrease of two orders of magnitude and are comparable with the DF search with cut-off four. Considering RPDF search with cut-off six, we examine that we can now cover more pathway alternatives than with the DF search with cut-off five with the same magnitude of CPU time as before. Figure 11 illustrates these relationships. In order to enhance readability of Figure 11 we use the Reaxys identification number for screened end products. Chemical names of the species can be found in the figures caption.

We find one end product, mycophenolic acid (Reaxys ID: 8644904), where screening via RPDF search does not lead to a route, but the DF search in cut-off five finds one route. This means that from none of the tested strategical molecules mycophenolic acid was accessible in three steps. Further, we find an end product, valepotriate (Reaxys ID: 4339241), where no paths are found by both DF screening in cut-off five and RPDF screening in cut-off six. As we are searching for highly complex and functionalised molecules, e.g. in the case of valepotriate four chiral centres, it is not surprising that neither search finds a path when six steps is the average connectivity for any pair of nodes. Still, the RDPF search presents such negative results in shorter computational times.

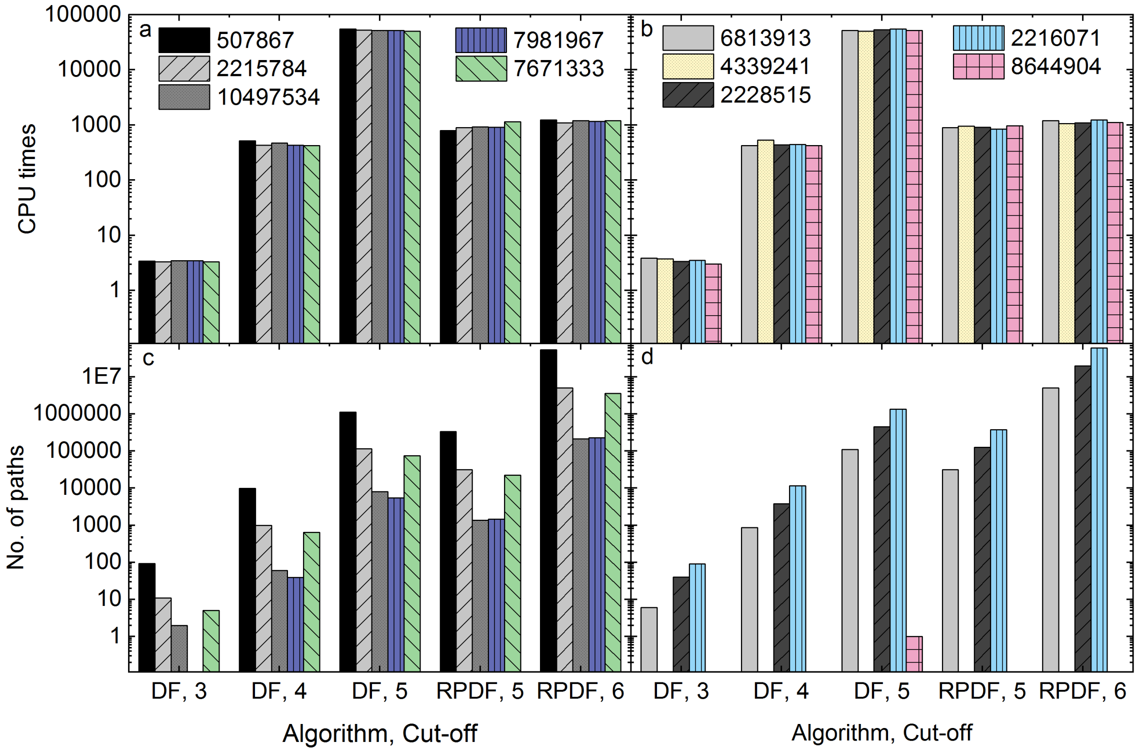


Figure 11. CPU times (a) and (b) and number of paths found (c) and (d) over different cut-off values are shown. The normal depth-first search algorithm runs on cut-off three, four

and five, while the RPDF search covers a cut-off of five (two steps and three steps) and six (two times three steps). The pathways start with pinene and end at 2-amino-1-phenylpropane (Reaxys ID: 507867), 2-diethylamino-N-(2,6-dimethylphenyl)acetamide (Reaxys ID: 2215784), paricalcitol (Reaxys ID: 10497534), ezetimibe (Reaxys ID: 7981967), imatinib (Reaxys ID: 7671333), letrozole (Reaxys ID: 6813913), valepotriate (Reaxys ID: 4339241), L-thyroxine (Reaxys ID: 2228515), ketoprofen (Reaxys ID: 2216071), or mycophenolic acid (Reaxys ID: 8644904).

Based on three experiments we have demonstrated that with regard to short screening ranges the strategic molecules are significantly better connected than sets of random molecules. We further indicate that the RPDF search can be used to compare the strategic molecules with one another. Last but not least, we outline promising reaction options in shorter CPU times than by DF screening.

4. Conclusions

With more than 105 million chemical compounds and 42 million chemical reactions recorded in Reaxys,[53] already the discovered chemical space is reasonably large to allow statistical treatment; the size of yet undiscovered chemical space is open to debate. The chemical space is a relatively new application in the field of network theory and is largely unexplored in this respect. The chemical framework offers opportunities to develop new algorithms, and to bring forward the field of data mining giving essential inputs to neighbouring disciplines, such as synthetic and computational chemistry. In this work, we propose a new method of identification of optimal locations in networks for introduction of new feedstocks, and provide a network-based list of potential strategic molecules, which can be used in developing future circular supply chains. The identified strategic molecules show a variety of chemical structures and an industrial relevance in the past and the present.

In a test example we demonstrate that strategic molecules are better connected to molecules from a waste process stream, CST, than multiple sets of random molecules in up to three-step reactions. The proposed RPDF method is a facile and rapid screening method that redirects pathways over strategic molecules. We illustrate the method in a case study of converting CST components to pharmaceuticals, and show that within a short CPU time a large screening range is achievable; this is compared with the method of finding long reaction paths not including strategic molecules explicitly.

This work contributes to developing methods of decision making for early stage process development that can guide research efforts by considering strategic molecules in synthesis planning. The study indicates benefits gained by inclusion of strategic molecules for multiple planning scenarios. Most notably, this work highlights options for development of sustainable processes. With a fully automated selection of best pathway alternatives in large reaction networks as our long-term goal, this study contributes to the assembly of the pathways. The impact is twofold: we show a method of identifying potential strategic molecules and we suggest a method for the inclusion of these in process development. In future work, the screening method can be extended by evaluation and ranking of pathways, e.g. through the inclusion of reaction yields and mass flows.

Acknowledgements

J. M. Weber would like to thank the Department of Chemical Engineering for funding in the form of a PhD studentship, and Pembroke College Cambridge for additional financial support. We gratefully acknowledge collaboration with RELX Intellectual Properties SA and their technical support, which enabled us to mine Reaxys. Reaxys data were made accessible to our research project via the Elsevier R&D Collaboration Network. This work was funded, in part, by the EPSRC project "Terpene-based manufacturing for sustainable chemical feedstocks" EP/K014889. This project was funded, in part, by the National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program as a part of the Cambridge Centre for Advanced Research and Education in Singapore Ltd (CARES).

Conflict of interest

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at (TO DO: XYZ).

References

1 S. Venkata Mohan, G. N. Nikhil, P. Chiranjeevi, C. Nagendranatha Reddy, M. V. Rohit, A. N. Kumar and O. Sarkar, *Bioresour. Technol.*, 2016, **215**, 2–12.

2 R. Liguori and V. Faraco, *Bioresour. Technol.*, 2016, **215**, 13–20.

3 J. Ying and Z. Li-jun, in *Physics Procedia*, Elsevier Srl, 2012, vol. 25, pp. 1682–1688.

4 A. Murray, K. Skene and K. Haynes, *J. Bus. Ethics*, 2017, **140**, 369–380.

5 S. Sauvé, S. Bernard and P. Sloan, *Environ. Dev.*, 2016, **17**, 48–56.

6 M. Lieder and A. Rashid, *J. Clean. Prod.*, 2016, **115**, 36–51.

7 M. Geissdoerfer, P. Savaget, N. M. P. Bocken and E. J. Hultink, *J. Clean. Prod.*, 2017, **143**, 757–768.

8 L. Reh, *Particuology*, 2013, **11**, 119–133.

9 A. Genovese, A. A. Acquaye, A. Figueroa and S. C. L. Koh, *Omega*, 2017, **66**, 344–357.

10 Y. Geng, Q. Zhu and M. Haight, *Waste Manag.*, 2007, **27**, 141–150.

11 D. Suocheng, K. W. Tong and W. Yuping, *Util. Policy*, 2001, **10**, 7–11.

12 J. A. Mathews and H. Tan, *J. Ind. Ecol.*, 2011, **15**, 435–457.

13 F. Yang, M. A. Hanna and R. Sun, *Biotechnol. Biofuels*, 2012, **5**, 1–10.

14 A. Arevalo-Gallegos, Z. Ahmad, M. Asgher, R. Parra-Saldivar and H. M. N. Iqbal, *Int. J. Biol. Macromol.*, 2017, **99**, 308–318.

15 H. Pakdel, S. Sarron and C. Roy, *J. Agric. Food Chem.*, 2001, **49**, 4337–4341.

16 B. Wozniak, A. Spannenberg, Y. Li, S. Hinze and J. G. de Vries, *ChemSusChem*, 2018, **11**, 356–359.

17 S. Rodriguez Couto, *Biotechnol. J.*, 2008, **3**, 859–870.

18 R. Ravindran and A. K. Jaiswal, *Trends Biotechnol.*, 2016, **34**, 58–69.

19 J. S. Van Dyk, R. Gama, D. Morrison, S. Swart and B. I. Pletschke, *Renew. Sustain. Energy Rev.*, 2013, **26**, 521–531.

20 D. Zhang, E. A. Del Rio-Chanona and N. Shah, *ACS Sustain. Chem. Eng.*, 2017, **5**, 4388–4398.

21 K. Ulonska, A. Voll and W. Marquardt, *Energy and Fuels*, 2016, **30**, 445–456.

22 K. Ulonska, A. König, M. Klatt, A. Mitsos and J. Viell, *Ind. Eng. Chem. Res.*, 2018, **57**, 6980–6991.

23 K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chemie - Int. Ed.*, 2006, **45**, 5348–5354.

24 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chemie - Int. Ed.*, 2016, **55**, 5904–5937.

25 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chemie - Int. Ed.*, 2005, **44**, 7263–7269.

26 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.

27 P. M. Jacob and A. Lapkin, *React. Chem. Eng.*, 2018, **3**, 102–118.

28 P. M. Jacob, P. Yamin, C. Perez-Storey, M. Hopgood and A. A. Lapkin, *Green Chem.*, 2017, **19**, 140–152.

29 J. Van Haveren, E. L. Scott and J. Sanders, *Biofuels, Bioprod. Biorefining Innov. a Sustain. Econ.*, 2008, **2.1**, 41–57.

30 F. Cherubini, *Energy Convers. Manag.*, 2010, **51**, 1412–1421.

31 C. O. Tuck, E. Pérez, I. T. Horváth, R. A. Sheldon and M. Poliakoff, *Science*, 2012, **337**, 695–699.

32 H. C. Kolb, M. G. Finn and K. B. Sharpless, *Angew. Chemie - Int. Ed.*, 2001, **40**, 2004–2021.

33 M. E. Welsch, S. A. Snyder and B. R. Stockwell, *Curr. Opin. Chem. Biol.*, 2010, **14**, 347–361.

34 P. Schneider and G. Schneider, *Angew. Chemie - Int. Ed.*, 2017, **56**, 7971–7974.

35 T. Werpy and G. Petersen, *Top Value Added Chemicals From Biomass: volume I -- results of screening for potential candidates from sugar and synthesis gas*, National Renewable Energy Lab., Golden, CO (US), 2004.

36 J. C. Serrano-Ruiz, R. Luque, J. M. Campelo and A. A. Romero, *Challenges*, 2012, **3**, 114–132.

37 T. P. Peixoto, *arXiv:1705.10225v7*, 2018.

38 T. P. Peixoto, *Phys. Rev. E*, 2017, **95**, 1–21.

39 X. Xu, N. Yuruk, Z. Feng and T. A. J. Schweiger, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, 2007.

40 D. Wang, P. Cui and W. Zhu, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, 2016, pp. 1225–1234.

41 P. Goyal and E. Ferrara, *Knowledge-Based Syst.*, 2018, **151**, 78–94.

42 M. P. van den Heuvel and O. Sporns, *Trends Cogn. Sci.*, 2013, **17**, 683–696.

43 B. C. M. van Wijk, C. J. Stam and A. Daffertshofer, *PLoS One*, 2010, **5**, e13701.

44 E. K. Towlson, P. E. Vertes, S. E. Ahnert, W. R. Schafer and E. T. Bullmore, *J. Neurosci.*, 2013, **33**, 6380–6387.

45 J. Kleinberg, *J. ACM*, 1999, **46**, 604–632.

46 L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Bridging Order to the Web*, Stanford InfoLab, 1999.

47 P. Jacob, Towards algorithmic use of chemical data, PhD Thesis, University of Cambridge, Cambridge, U.K., 2017.

48 P. M. Gleiss, P. F. Stadler, A. Wagner and D. A. Fell, *Adv. Complex Syst.*, 2001, **04**, 207–226.

49 S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, *Phys. Rep.*, 2006, **424**, 175–308.

50 V. Latora, V. Nicosia and G. Russo, *Complex networks: principles, methods and applications*, Cambridge Univeristy Press, 2017.

51 A.-L. Barabási and E. Bonabeau, *Sci. Am.*, 2003, **288**, 60–69.

52 M. Newman and Newman, *SIAM Rev.*, 2003, **45**, 167–256.

53 RELX Intellectual Properties SA, Reaxys - Reaxys is a trademark, copyright owned by RELX Intellectual Properties SA and used under licence., https://www.reaxys.com/, https://www.elsevier.com/solutions/reaxys., (accessed 29 May 2018).

54 T. P. Peixoto, *figshare*, 2014.

55 U. Brandes, *J. Math. Sociol.*, 2001, **25**, 163–177.

56 C. Ding, X. He, P. Husbands, H. Zha and H. D. Simon, *Proc. 2003 SIAM Int. Conf. Data Mining. Soc. Ind. Appl. Math.*, 2003, 249–353.

57 A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, Cambridge, U.K., 2011.

58 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, E. Duchesnay, M. Perrot, M. Brucher, D. Cournapeau, A. Passos and J. Vanderplas, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

59 R. Xu and D. Wunsch, *Theor. Comput. Fluid Dyn.*, 1999, **13**, 129–141.

60 D. Xu and Y. Tian, *Ann. Data Sci.*, 2015, **2**, 165–193.

61 F. T. Liu, K. M. Ting and Z.-H. Zhou, in *2008 Eigth IEEE International Conference on Data Mining*, 2008.

62 R. Tarjan, *SIAM J. Comput.*, 1972, **1**, 146–160.

63 H. A. Wittcoff, B. G. Reuben and J. S. Plotkin, *Industrial organic chemicals*, John Wiley & Sons, New York, USA, 2004.

64 W. Cunnigham, Catalytic Conversion of Terpene Feedstocks into Value-Added Chemicals and Commodity Chemicals, PhD Thesis, University of Bath, Bath, U.K., 2018.

65 B. J. Nikolau, M. A. D. N. Perera, L. Brachova and B. Shanks, *Plant J.*, 2008, **54**, 536–545.

66 H. Ahmed, N. Rask and E. Dean Baldwin, *Biomass*, 1989, **19**, 215–232.

67 T. J. Slaga, A. J. P. Klein-Szanto, L. L. Triplett, L. P. Yotti and J. E. Trosko, *Science*, 1981, **213**, 1023–1025.

68 P. P. Kapadia, Tetraphenylethylene: A versatile supramolecular framework , PhD Thesis, Univeristy of Iowa, Iowa City, U.S.A., 2011.

69 P. Knuuttila, *Fuel*, 2013, **104**, 101–108.

70 D. Helmdach, P. Yaseneva, P. K. Heer, A. M. Schweidtmann and A. A. Lapkin, *ChemSusChem*, 2017, **10**, 3632–3643.

71 N. A. McGrath, M. Brichacek and J. T. Njardarson, *J. Chem. Educ.*, 2010, **87**, 1348–1349.

Identification of strategic molecules for future circular supply chains using large reaction networks

Jana Marie Weber,[1] Pietro Lió[2*] and Alexei A. Lapkin[1,3*]

[1]*Department of Chemical Engineering and Biotechnology, University of Cambridge, West Cambridge Site, Philippa Fawcett Drive, Cambridge CB3 0AS, UK*

[2]*Department of Computer Science and Technology, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK*

[3]*Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

Electronic Supporting Information

Table of Content

Supplementary note 1 **Page 2**
Supplementary note 2 **Page 3**
Supplementary note 3 **Page 4**
Supplementary note 4 **Page 5**
Supplementary material 1 **Page 7**

* Corresponding authors: Pietro Lio email: pl219@cam.ac.uk; Alexei Lapkin email: aal35@cam.ac.uk

Supplementary note 1: Pipeline

The pipeline for the identification of strategic locations in networks is general applicable for other network problems if parameters such as the descriptive features are adapted respectively. Due to license agreements, the download routine to obtain reaction data from Reaxys application programming interface (API), cannot be published (general information on the routine are given in the following Supplementary note). Hence, we start the openly available pipeline from the assumption of an already existing network structure of e.g. chemical reactions. These parts of the pipeline are accessible through GitHub: (https://github.com/Jana-Marie-Weber/strategic_molecules) and start with a network file in graph-tool ".gt" format.

Supplementary note 2: Network assembly

Reaxys database [1] is used to obtain the information needed to build a network of chemical reactions. An automated download script developed by [2] passes a series of search queries to the Reaxys API. Starting from a chosen molecule all reactions that take place with this molecule as reactant are found. Products from these queries are saved and used as starting molecules for the next search step. This procedure is repeated up to a pre-defined search depth. [2] This is called forward search. In addition to that, a one step backward search queries all reactions, which products are the previously saved molecules. Further information, e.g., the yield of reactions, publication year, or reaction conditions are downloaded and saved to the data file as well. Two different mapping schemes have been discussed in the literature. [3] A one-to-one mapping scheme only considers the heaviest reactant and the heaviest product and connects them with an edge, while in the all-to-all scheme all reactants are wired to all products of one reaction. Investigations on the mapping schemes have shown that the choice does not interfere with the network's characteristics. [4] Hence both, the one-to-one and the all-to-all mapping, are possible and equally valid representations of the same problem set. The all-to-all mapping scheme has been chosen for this study. More detailed information on the download routine can be found in. [2] The post processing of the data involves reducing the file by removing duplicate reactions, excluding multi-step reactions and "half" reactions, where either all product fields or all reactant fields are empty.

This work focuses on two networks, one centred around the sample strategic molecule, 4-hydroxyacetophenone (4-HAP). 4-HAP is a common precursor for many pharmaceutical products such as Paracetamol, Metoprolol, and Salbutamol and can be derived from terpenes, including the waste CST. [5] The network around 4-HAP was built using three steps forwards and one step backwards search in Reaxys database. The relatively small data mining thresholds were chosen to highlight local structures and 4-HAP's influence on the reaction network. A second network centred around the molecule limonene was constructed with data from four steps forward and one step backwards searches. We use the smaller 4-HAP data to detect strategic molecules and test the connection of these in the second network around limonene. The network model is implemented in python2.7 with the library graph-tools. [6]

Supplementary note 3: Computer architecture and python version

The code was run using a Linux machine with version 16.04.6 LTS (Xenial Xerus) with 24 cores, width of 64 bits, and 256 GB RAM. It has a Dual Processor Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz. We used the python version 2.7.12 with the lasts updates from November 2018, Scikit-learn version 0.19.2. and Graph-tool version 2.27. Further python requirements may be found on GitHub (https://github.com/Jana-Marie-Weber/strategic_molecules).

Supplementary note 4: Parametrisation in isolation forest

The isolation forest was performed on the default value of 100 base estimators in the ensemble. We train each base estimator on all samples by setting the number of maximum samples to the number of all nodes in the network. We tested all contamination rates shown in Table S*1* and Table S*2* and found respective numbers of outliers. We trained each base estimator on one feature and used the "old" behaviour for the decision function. Please consult scikit-learn or our implementation on GitHub (https://github.com/Jana-Marie-Weber/strategic_molecules) for further information on the algorithm. [7]

Table S1. The number of identified outliers per tested contamination rate is shown for contamination rates between $1{\cdot}10^{-4}$ and $8{\cdot}10^{-4}$.

rate	$1{\cdot}10^{-4}$	$5{\cdot}10^{-4}$	$5.5{\cdot}10^{-4}$	$6{\cdot}10^{-4}$	$6.5{\cdot}10^{-4}$	$7{\cdot}10^{-4}$	$7.5{\cdot}10^{-4}$	$8{\cdot}10^{-4}$
outliers	56	285	313	341	370	398	427	455

Table S2. The number of identified outliers per tested contamination rate is shown for contamination rates between $8.5{\cdot}10^{-4}$and $2{\cdot}10^{-3}$.

rate	$8.5{\cdot}10^{-4}$	$9{\cdot}10^{-4}$	$9.5{\cdot}10^{-4}$	$1{\cdot}10^{-3}$	$1.5{\cdot}10^{-3}$	$2{\cdot}10^{-3}$
outliers	484	512	540	569	853	1137

Bibliography

1 RELX Intellectual Properties SA, Reaxys - Reaxys is a trademark, copyright owned by RELX Intellectual Properties SA and used under licence., https://www.reaxys.com/, https://www.elsevier.com/solutions/reaxys., (accessed 29 May 2018).

2 P. Jacob, Towards algorithmic use of chemical data, PhD Thesis, University of Cambridge, Cambridge, U.K., 2017.

3 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.

4 P. M. Jacob and A. Lapkin, *React. Chem. Eng.*, 2018, **3**, 102–118.

5 W. Cunnigham, Catalytic Conversion of Terpene Feedstocks into Value-Added Chemicals and Commodity Chemicals, PhD Thesis, University of Bath, Bath, U.K., 2018.

6 T. P. Peixoto, *figshare*, 2014.

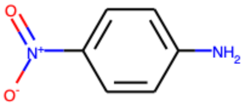
7 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, E. Duchesnay, M. Perrot, M. Brucher, D. Cournapeau, A. Passos and J. Vanderplas, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

Supplementary material 1: Molecular structures of strategic molecules

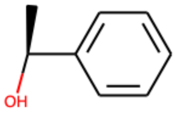
Figure S 1-38 show the molecular structures of 567 potential strategic molecules identified with a contamination rate of $1 \cdot 10^{-3}$. If no structure is shown, it is too large for the grid-wise printing and can be manually retrieved from Reaxys with the given Reaxys ID. Two additional molecules do not have molfiles attached in Reaxys. These molecules are:

1. ethereal hydrogen chloride, and
2. cellulose.

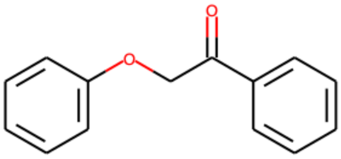
Please find the rest of the strategic molecules below.



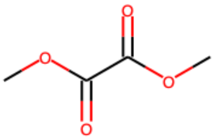
a



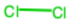
b



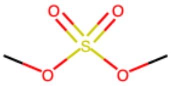
c



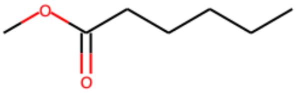
d



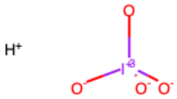
e



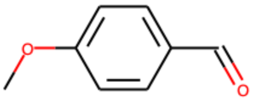
f



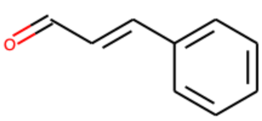
g



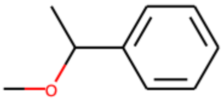
h



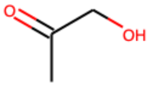
i



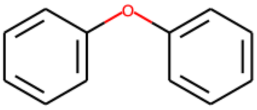
j



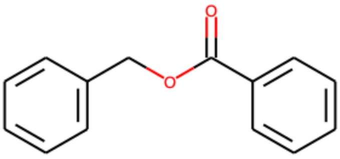
k



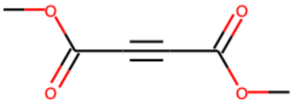
l



m

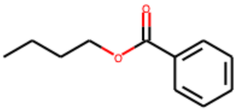


n

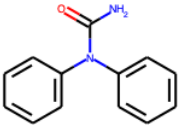


o

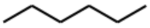
Figure S 1. Molecule set 0



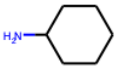
a



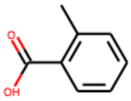
b



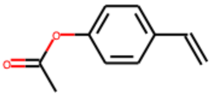
c



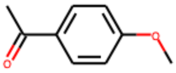
d



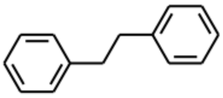
e



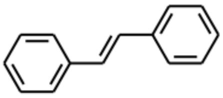
f



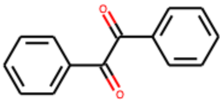
g



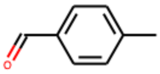
h



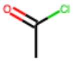
i



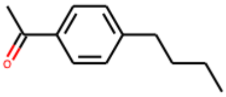
j



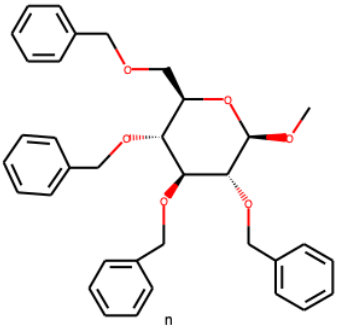
k



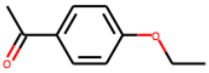
l



m

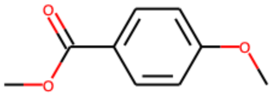


n

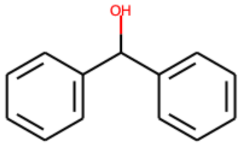


o

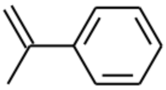
Figure S 2. Molecule set 15



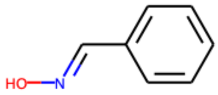
a



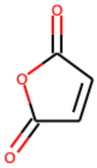
b



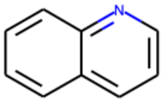
c



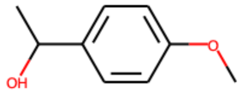
d



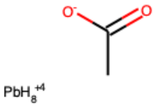
e



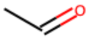
f



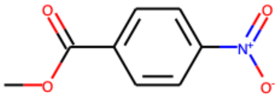
g



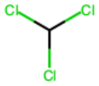
h



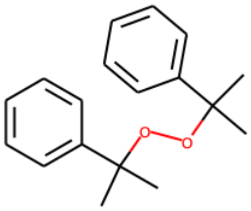
i



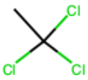
j



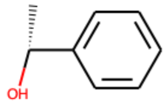
k



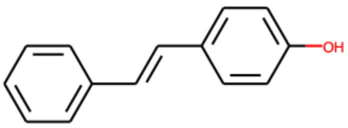
l



m

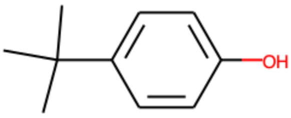


n

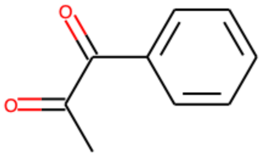


o

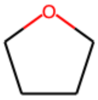
Figure S 3. Molecule set 30



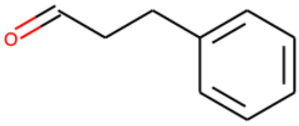
a



b

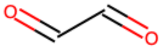


c

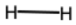


d

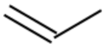
e



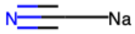
f



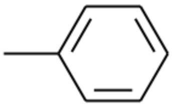
g



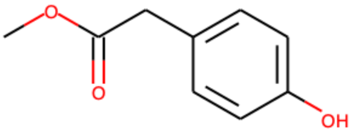
h



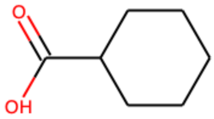
i



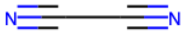
j



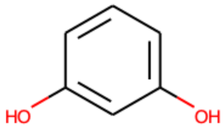
k



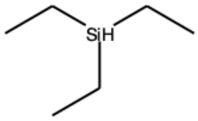
l



m

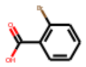


n

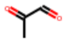


o

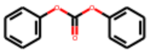
Figure S 4. Molecule set 45



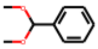
a



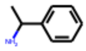
b



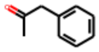
c



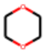
d



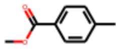
e



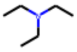
f



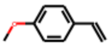
g



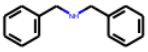
h



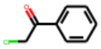
i



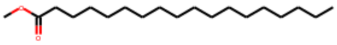
j



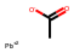
k



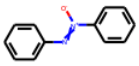
l



m



n



o

Figure S 5. Molecule set 60

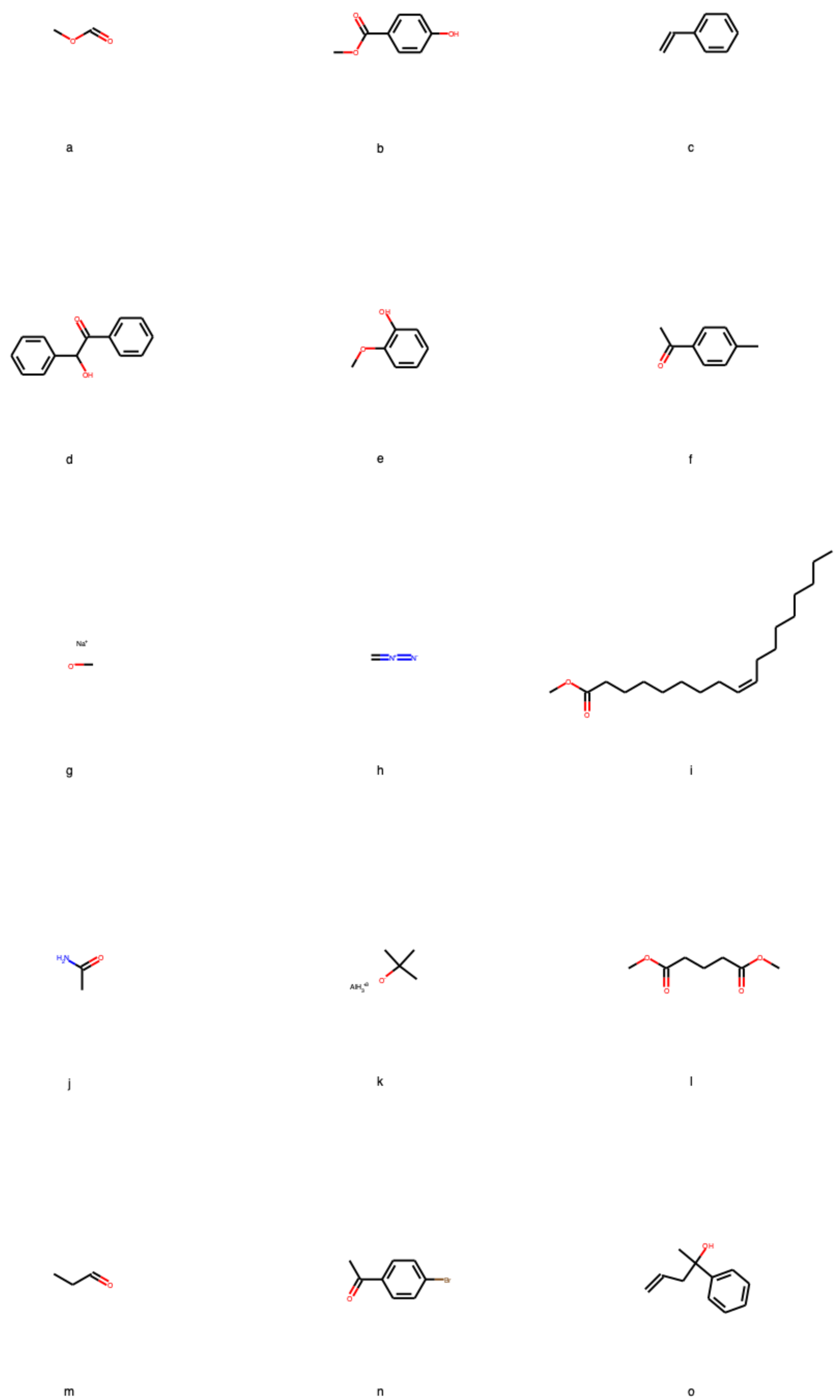
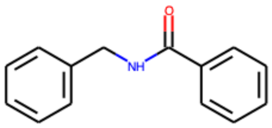
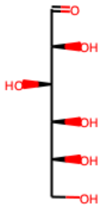


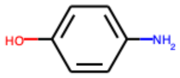
Figure S 6. Molecule set 75



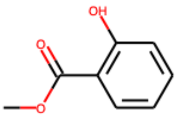
a



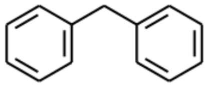
b



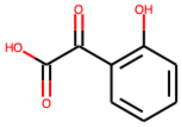
c



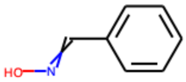
d



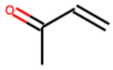
e



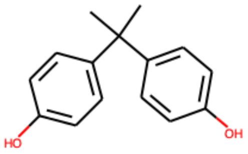
f



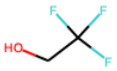
g



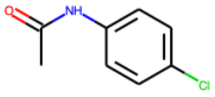
h



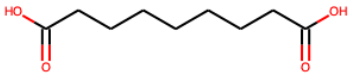
i



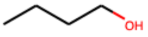
j



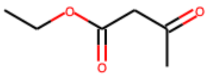
k



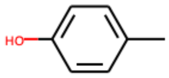
l



m

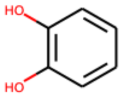


n

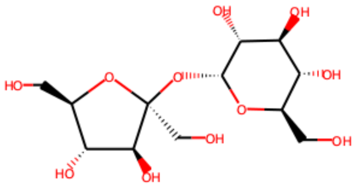


o

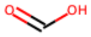
Figure S 7. Molecule set 90



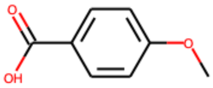
a



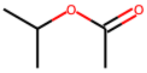
b



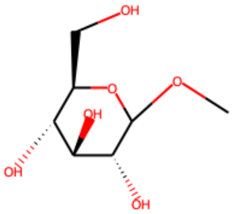
c



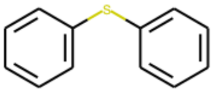
d



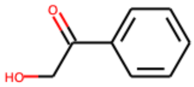
e



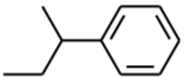
f



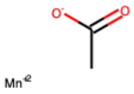
g



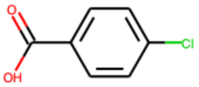
h



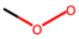
i



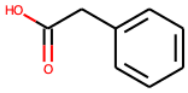
j



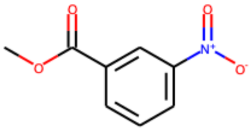
k



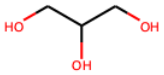
l



m

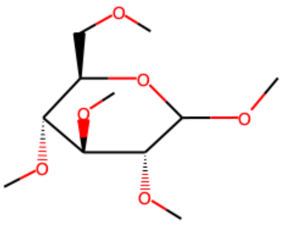


n

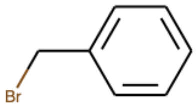


o

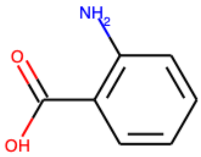
Figure S 8. Molecule set 105



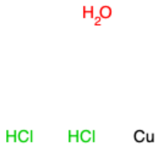
a



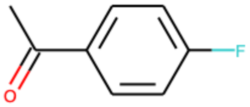
b



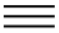
c



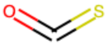
d



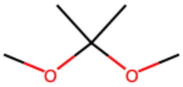
e



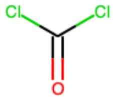
f



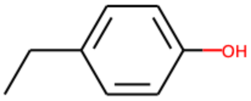
g



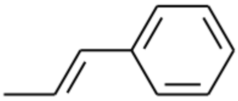
h



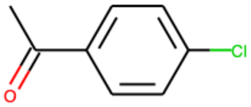
i



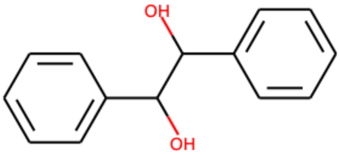
j



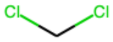
k



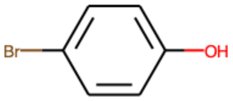
l



m

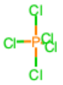


n

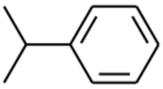


o

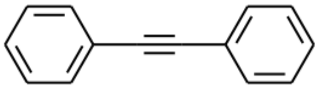
Figure S 9. Molecule set 120



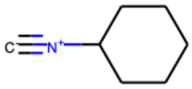
a



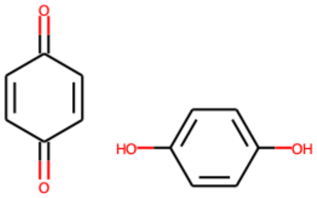
b



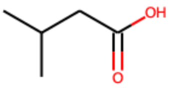
c



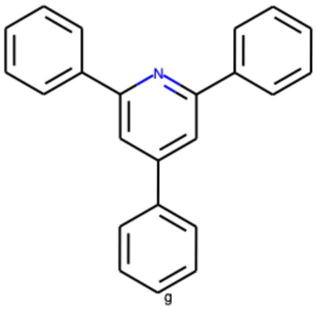
d



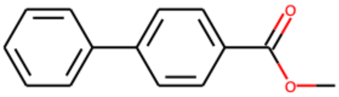
e



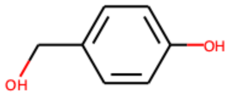
f



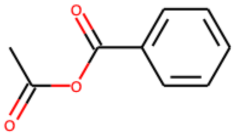
g



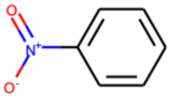
h



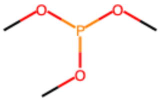
i



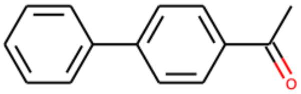
j



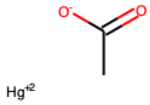
k



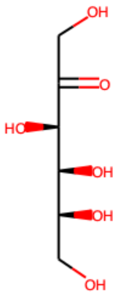
l



m

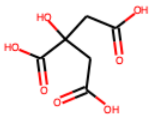


n

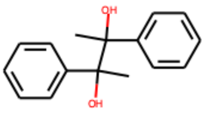


o

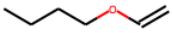
Figure S 10. Molecule set 135



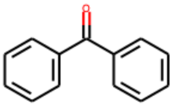
a



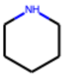
b



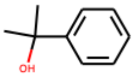
c



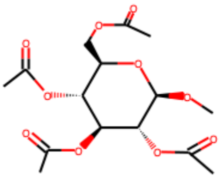
d



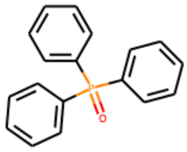
e



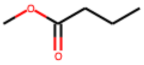
f



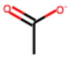
g



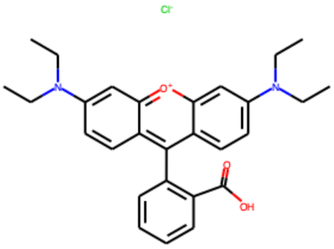
h



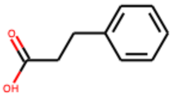
i



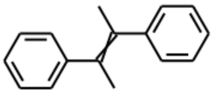
j



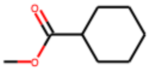
k



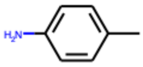
l



m

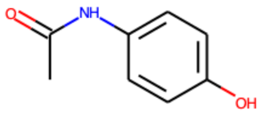


n

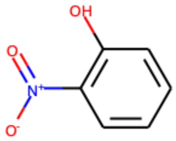


o

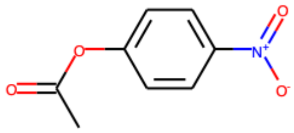
Figure S 11. Molecule set 150



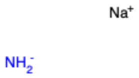
a



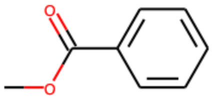
b



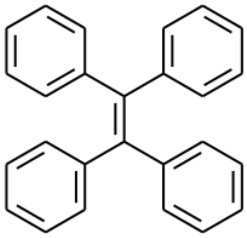
c



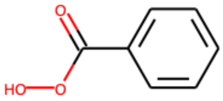
d



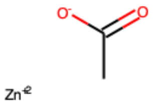
e



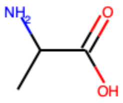
f



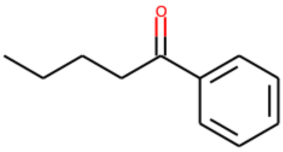
g



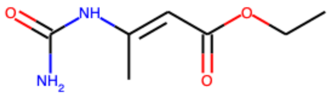
h



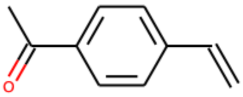
i



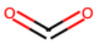
j



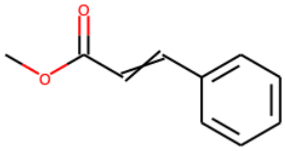
k



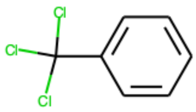
l



m



n



o

Figure S 12. Molecule set 165

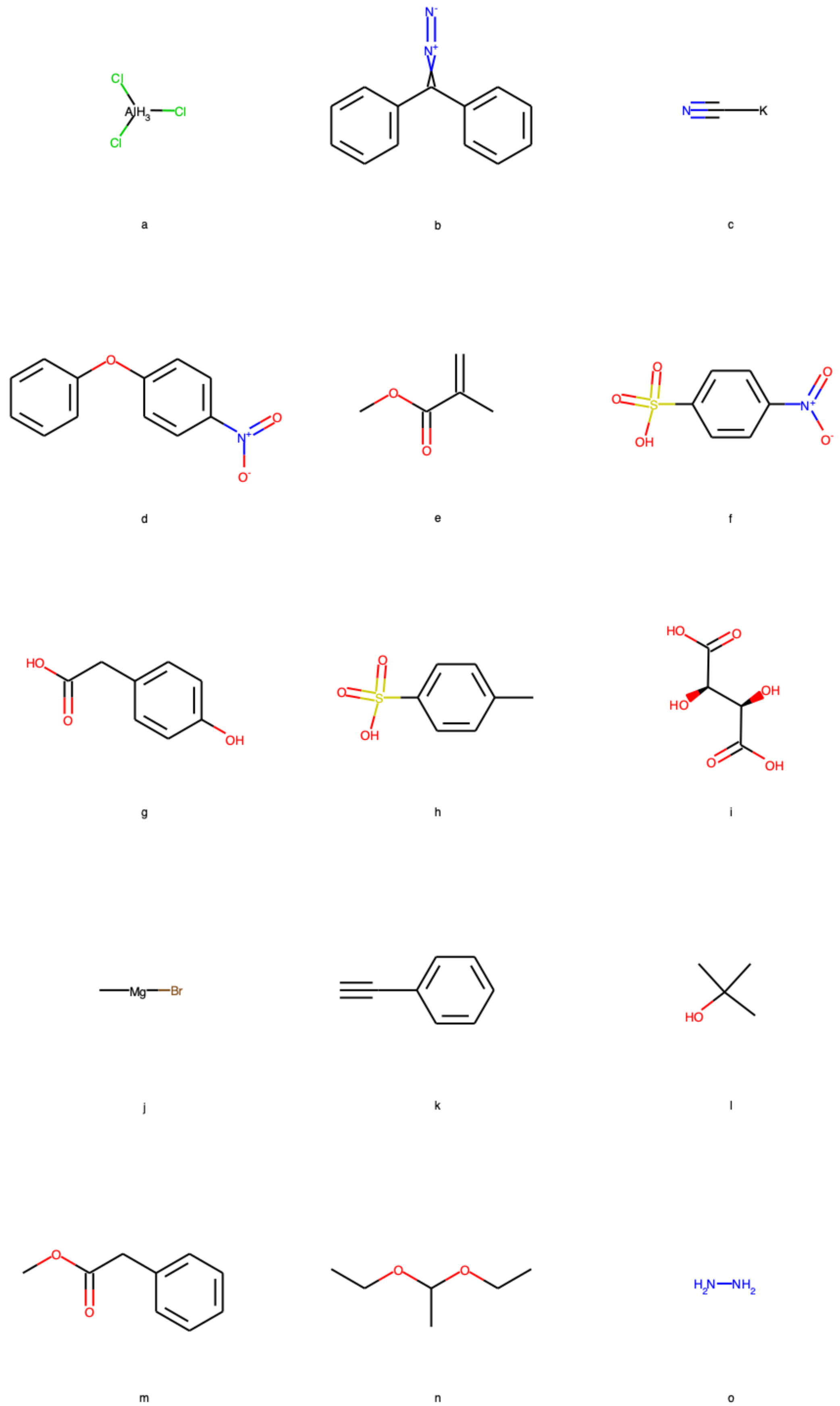
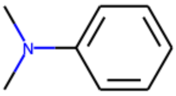
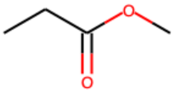


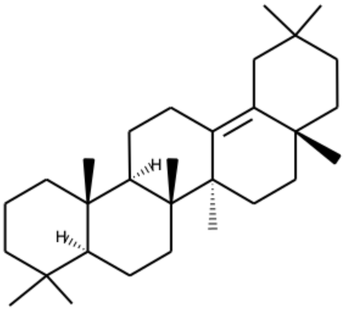
Figure S 13. Molecule set 180



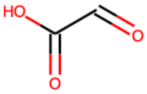
a



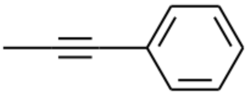
b



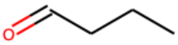
c



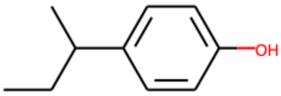
d



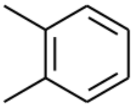
e



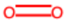
f



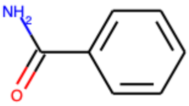
g



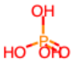
h



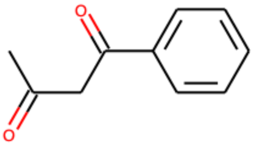
i



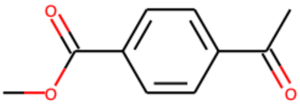
j



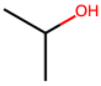
k



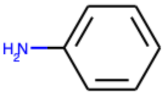
l



m

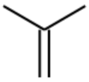


n

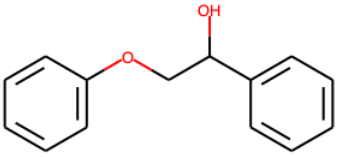


o

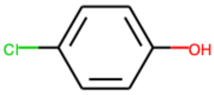
Figure S 14. Molecule set 195



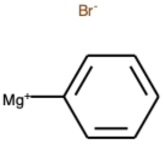
a



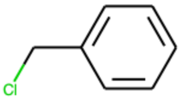
b



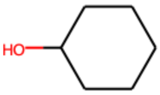
c



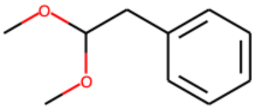
d



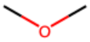
f



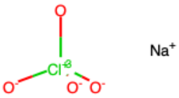
g



h



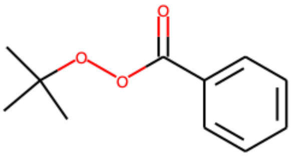
i



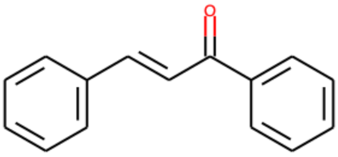
j

HBr

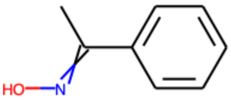
k



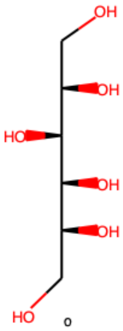
l



m



n



o

Figure S 15. Molecule set 210

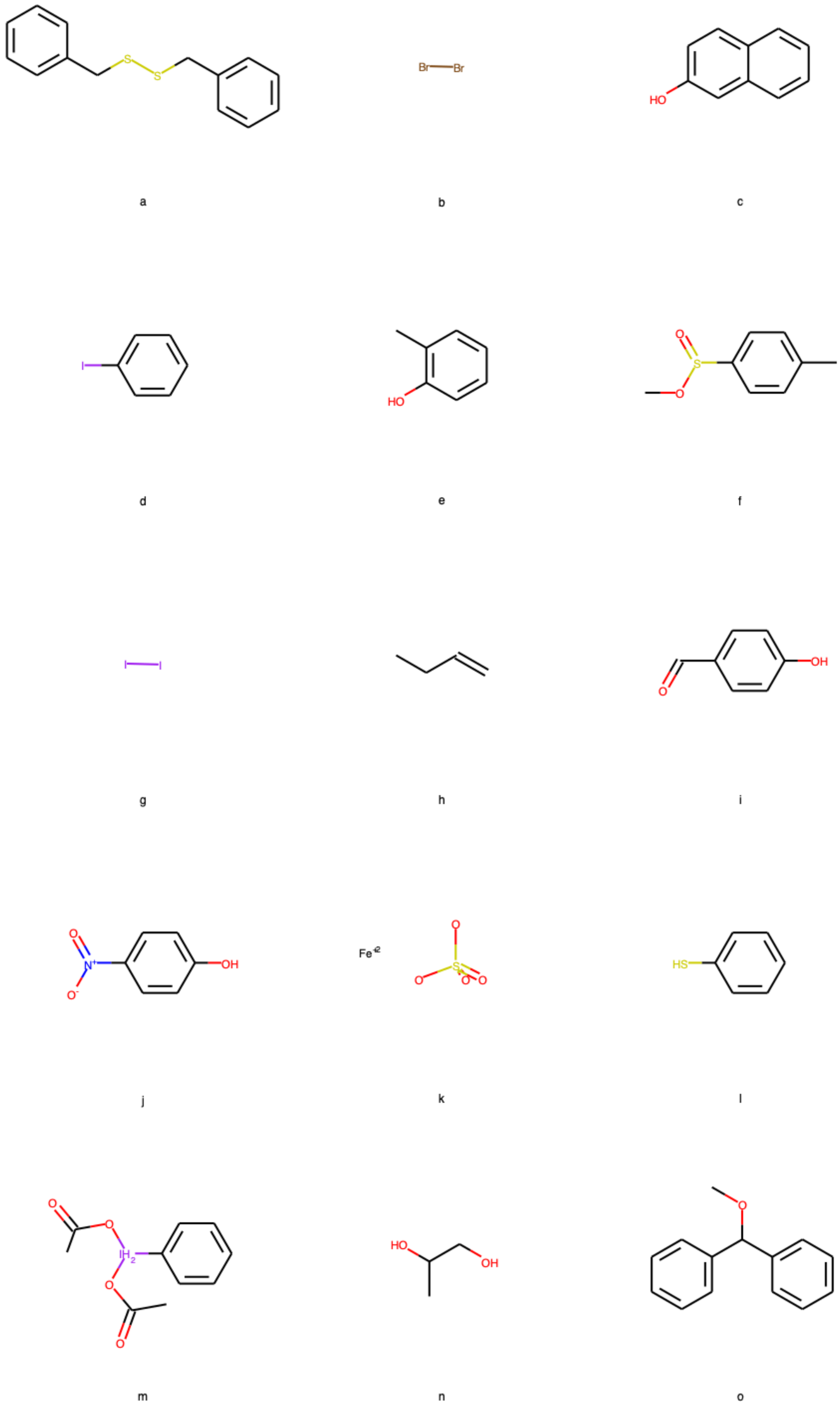
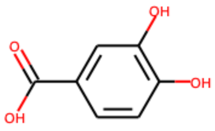
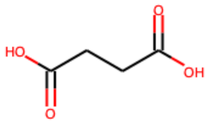


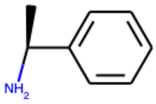
Figure S 16. Molecule set 225



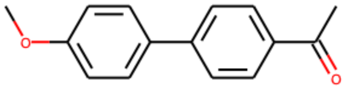
a



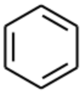
b



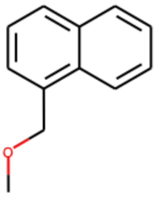
c



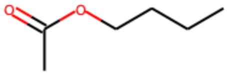
d



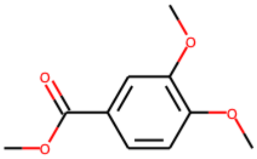
e



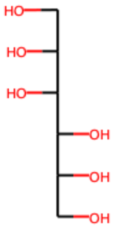
f



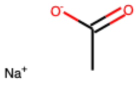
g



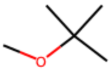
h



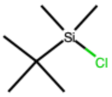
i



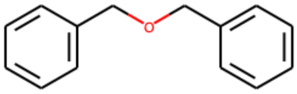
j



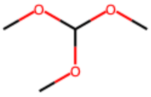
k



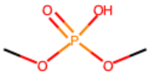
l



m

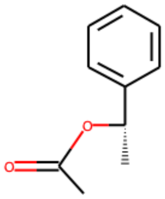


n

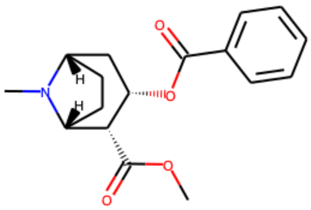


o

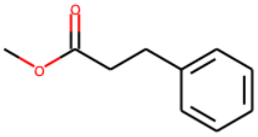
Figure S 17. Molecule set 240



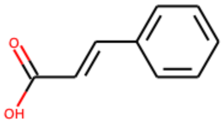
a



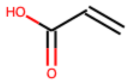
b



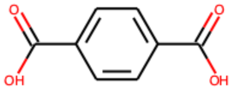
c



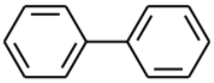
d



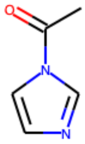
e



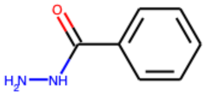
f



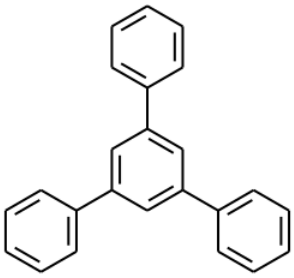
g



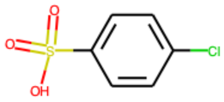
h



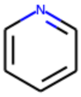
i



j



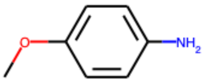
k



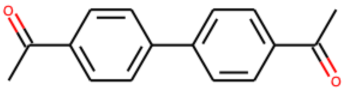
l

HI

m

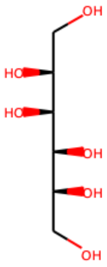


n

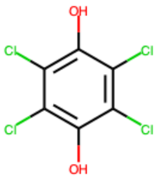


o

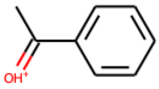
Figure S 18. Molecule set 255



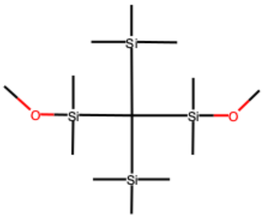
a



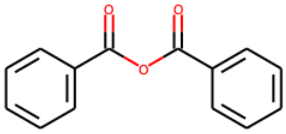
b



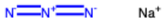
c



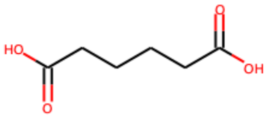
e



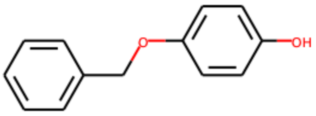
f



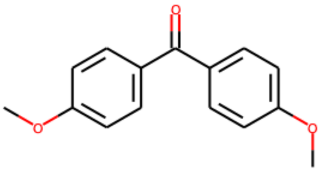
g



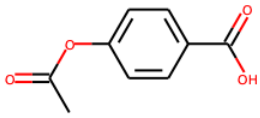
h



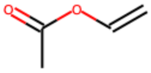
i



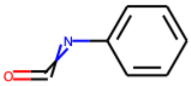
j



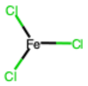
k



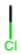
l



m

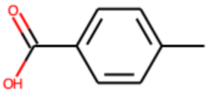


n



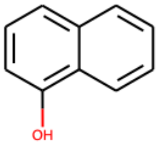
o

Figure S 19. Molecule set 270

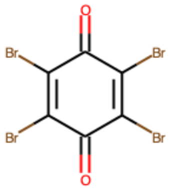


a

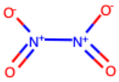
b



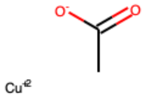
c



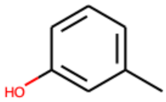
d



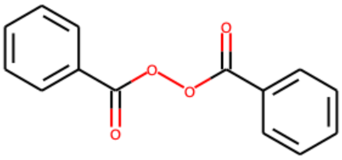
e



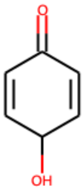
f



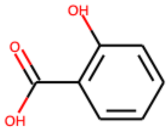
g



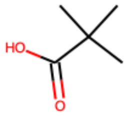
h



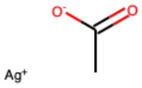
i



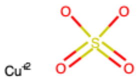
j



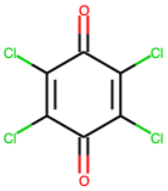
k



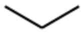
l



m



n



o

Figure S 20. Molecule set 285

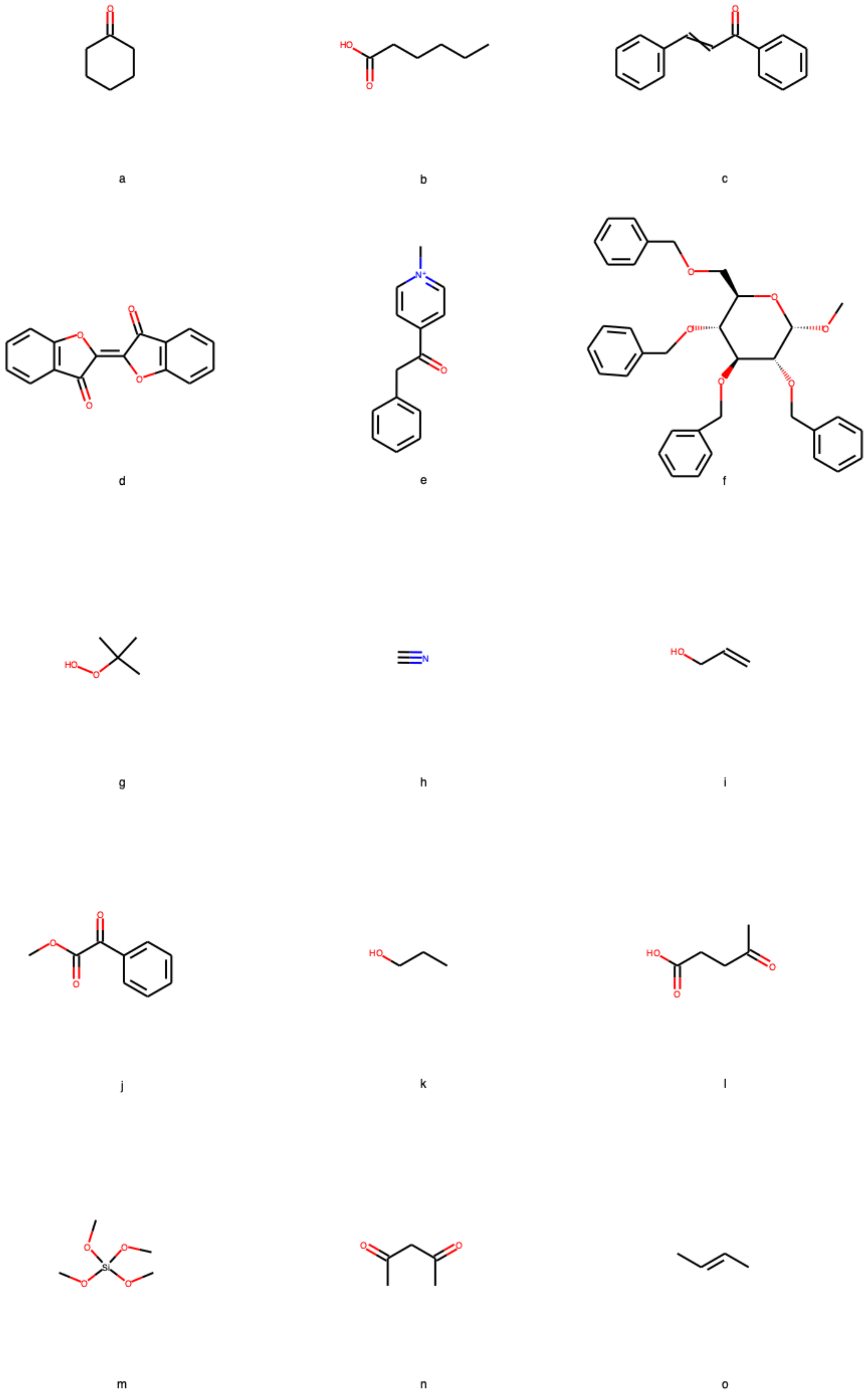
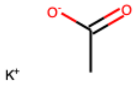
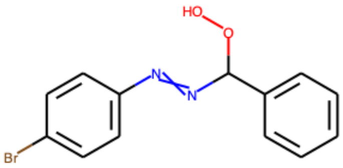


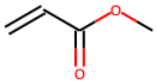
Figure S 21. Molecule set 300



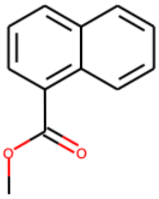
a



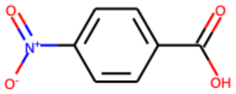
b



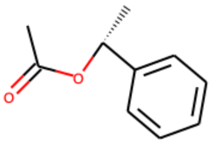
c



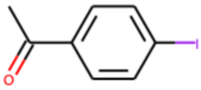
d



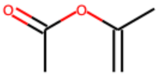
e



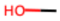
f



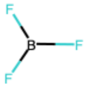
g



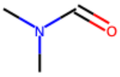
h



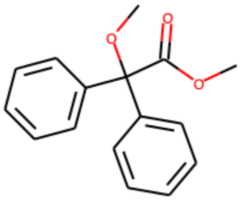
i



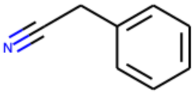
j



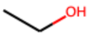
k



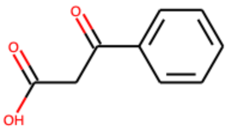
l



m

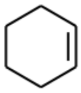


n

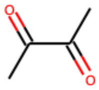


o

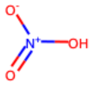
Figure S 22. Molecule set 315



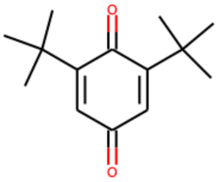
a



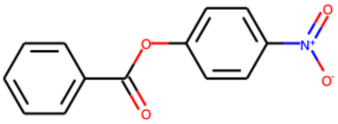
b



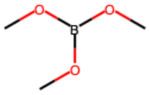
c



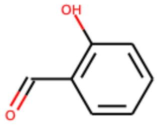
d



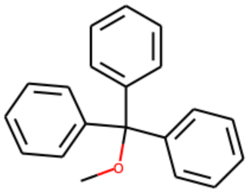
e



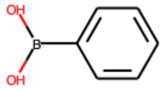
f



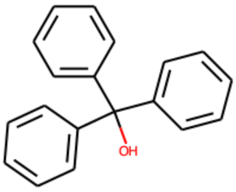
g



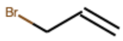
h



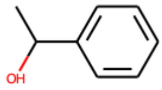
i



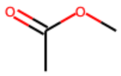
j



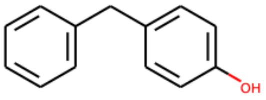
k



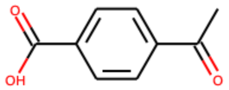
l



m

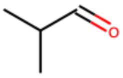


n

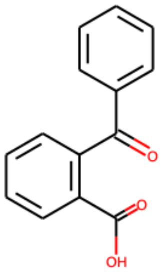


o

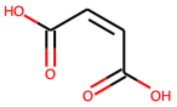
Figure S 23. Molecule set 330



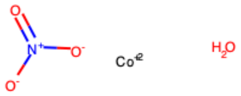
a



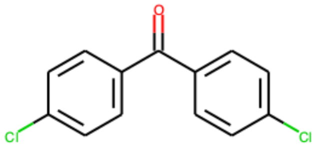
b



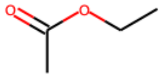
c



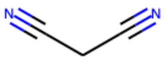
d



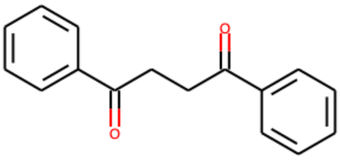
e



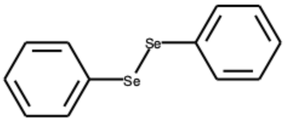
f



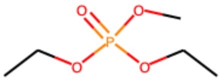
g



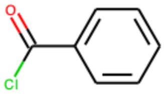
h



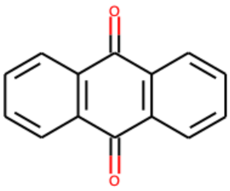
i



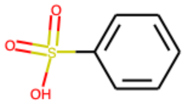
j



k



l

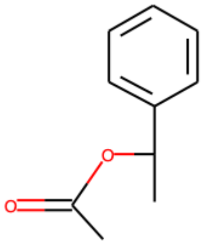


n

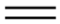
CH_4

o

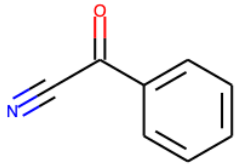
Figure S 24. Molecule set 345



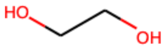
a



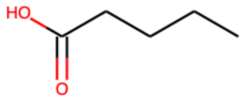
b



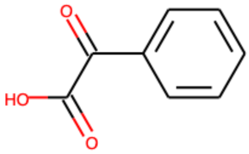
c



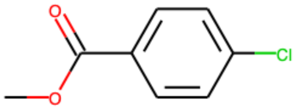
d



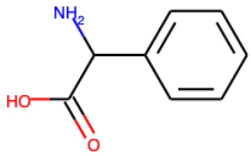
e



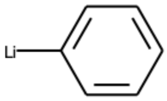
f



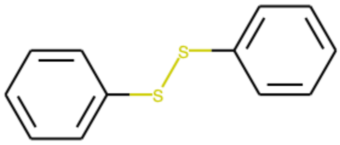
g



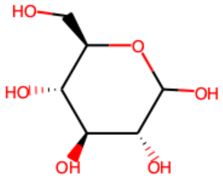
h



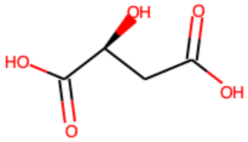
i



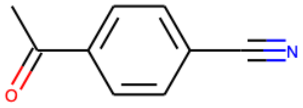
j



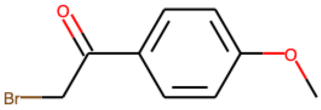
k



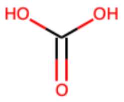
l



m



n



o

Figure S 25. Molecule set 360

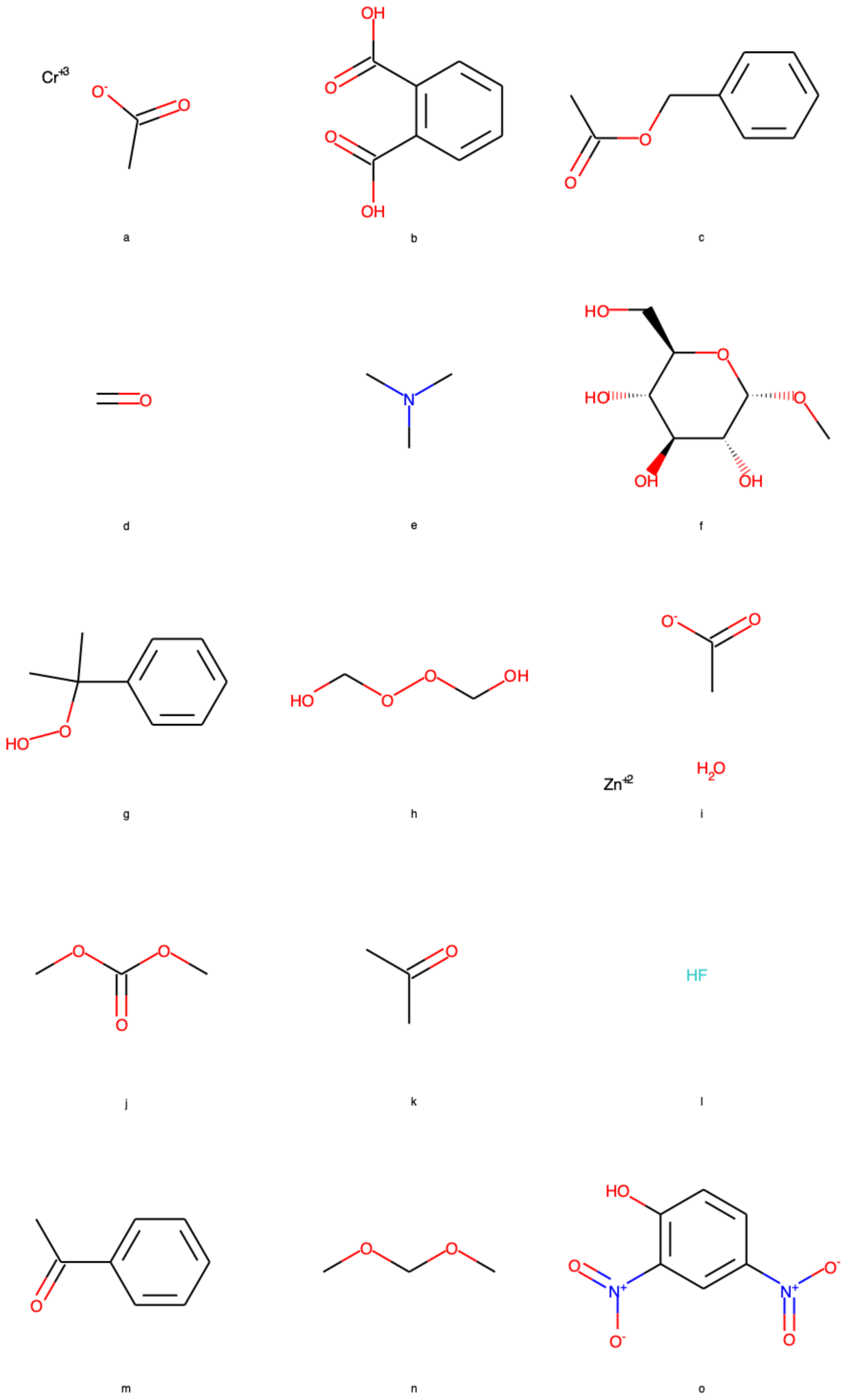
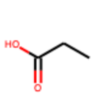
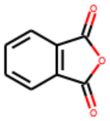


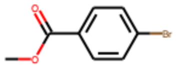
Figure S 26. Molecule set 375



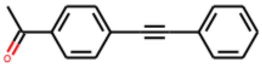
a



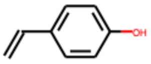
b



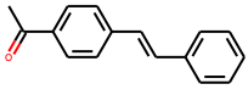
c



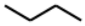
d



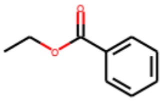
e



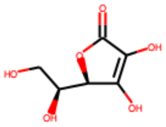
f



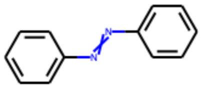
g



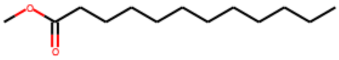
h



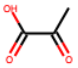
i



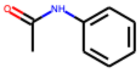
j



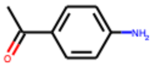
k



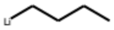
l



m

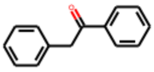


n

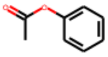


o

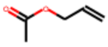
Figure S 27. Molecule set 390



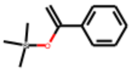
a



b

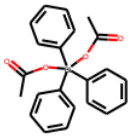


c

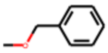


d

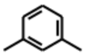
e



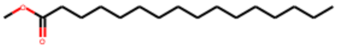
f



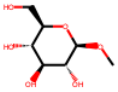
g



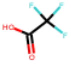
h



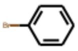
i



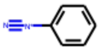
j



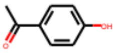
k



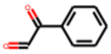
l



m

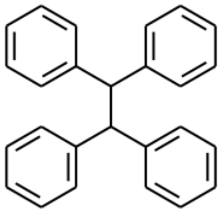


n

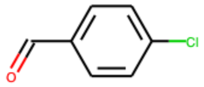


o

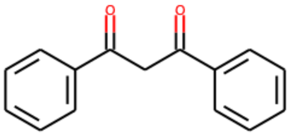
Figure S 28. Molecule set 405



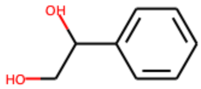
a



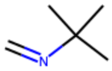
b



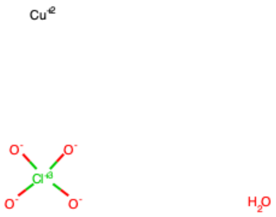
c



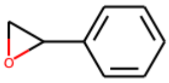
d



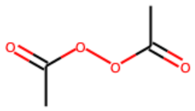
e



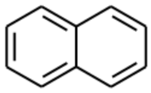
f



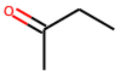
g



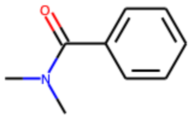
h



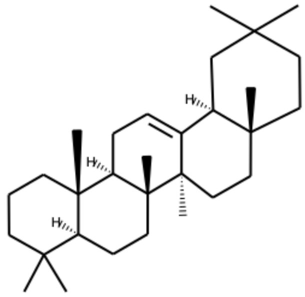
i



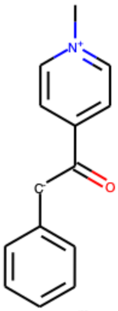
j



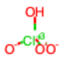
k



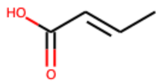
l



m

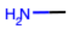


n

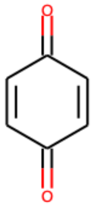


o

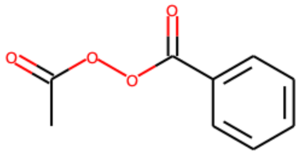
Figure S 29. Molecule set 420



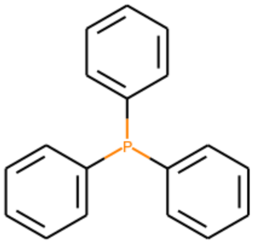
a



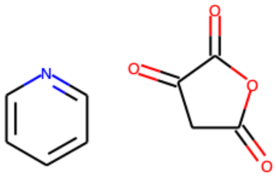
b



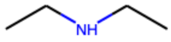
c



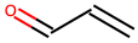
d



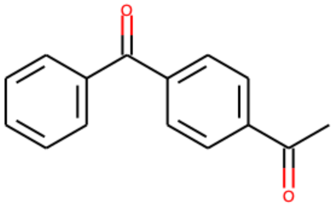
e



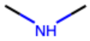
f



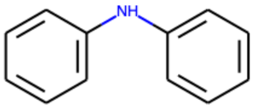
g



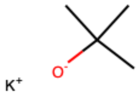
h



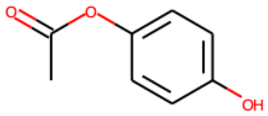
i



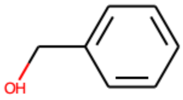
j



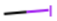
k



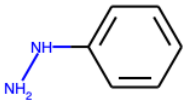
l



m

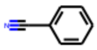


n

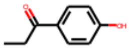


o

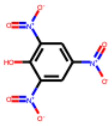
Figure S 30. Molecule set 435



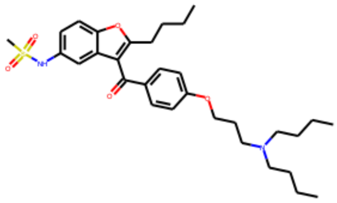
a



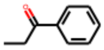
b



c

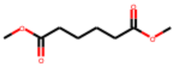


d

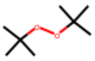


e

f

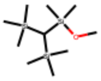


g

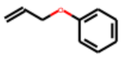


h

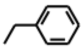
i



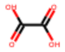
j



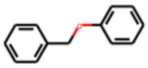
l



m



n



o

Figure S 31. Molecule set 450

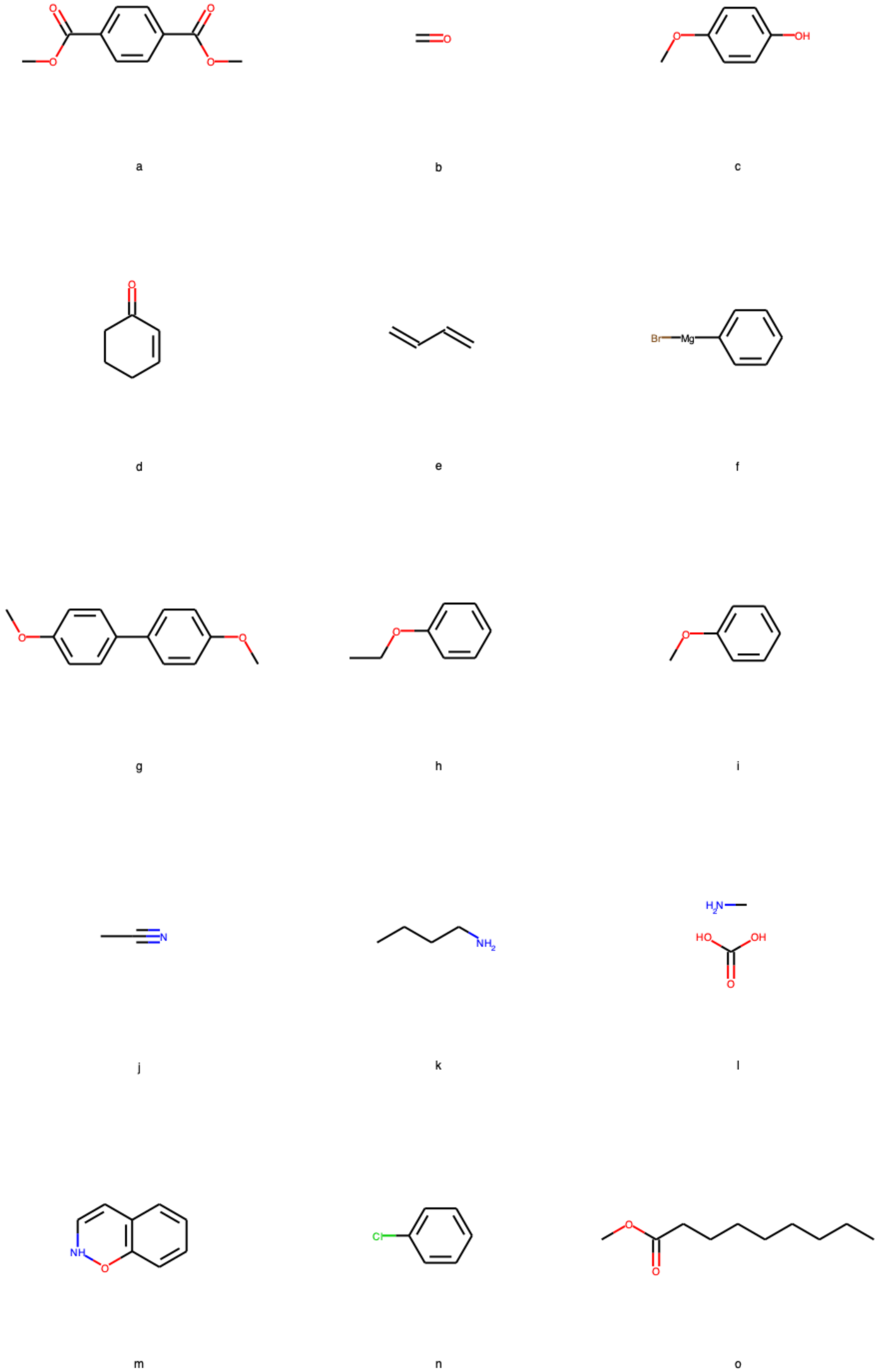
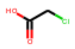
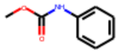


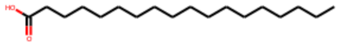
Figure S 32. Molecule set 465



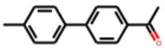
a



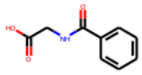
b



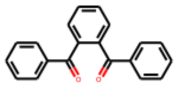
c



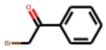
d



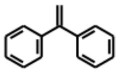
e



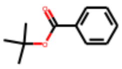
f



g

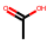


h

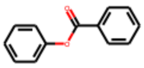


i

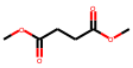
j



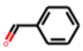
k



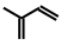
l



m

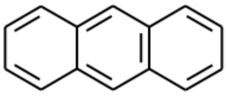


n

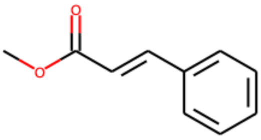


o

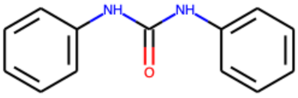
Figure S 33. Molecule set 480



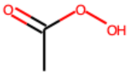
a



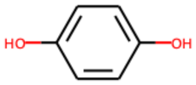
b



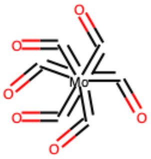
c



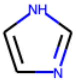
d



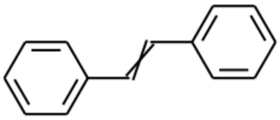
e



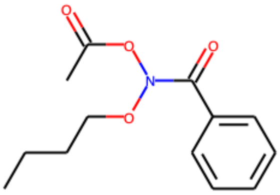
f



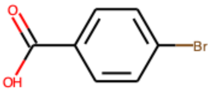
g



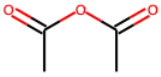
h



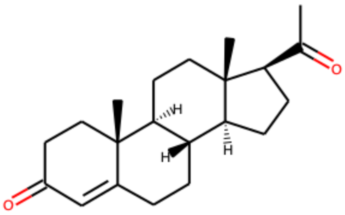
i



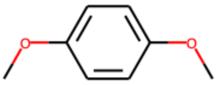
j



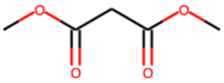
k



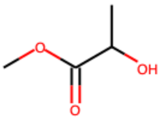
l



m

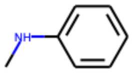


n

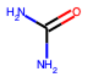


o

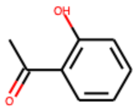
Figure S 34. Molecule set 495



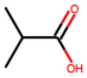
a



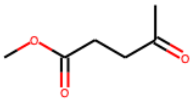
b



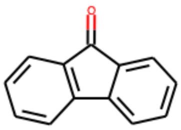
c



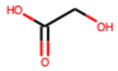
d



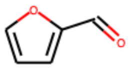
e



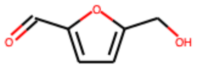
f



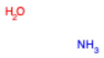
g



h

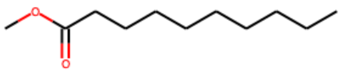


i

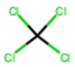


j

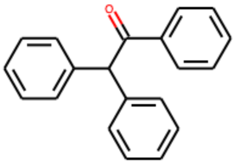
k



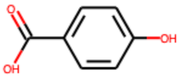
l



m

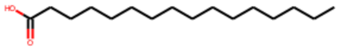


n



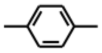
o

Figure S 35. Molecule set 510

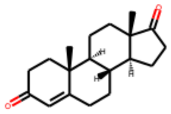


a

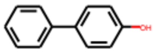
b



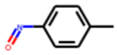
c



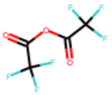
d



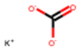
e



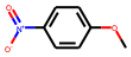
f



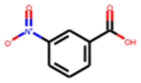
g



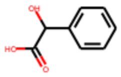
h



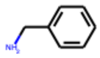
i



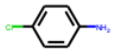
j



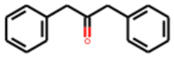
k



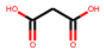
l



m

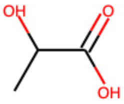


n

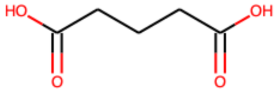


o

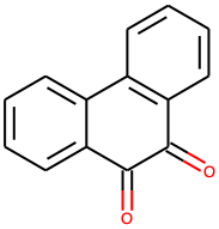
Figure S 36. Molecule set 525



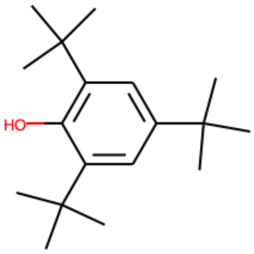
a



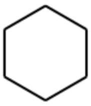
b



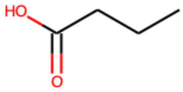
c



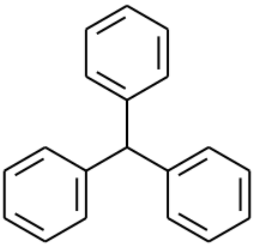
d



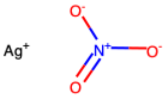
e



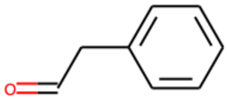
g



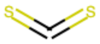
h



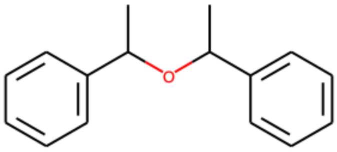
i



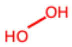
j



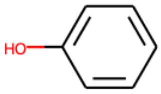
k



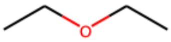
l



m



n



o

Figure S 37. Molecule set 540

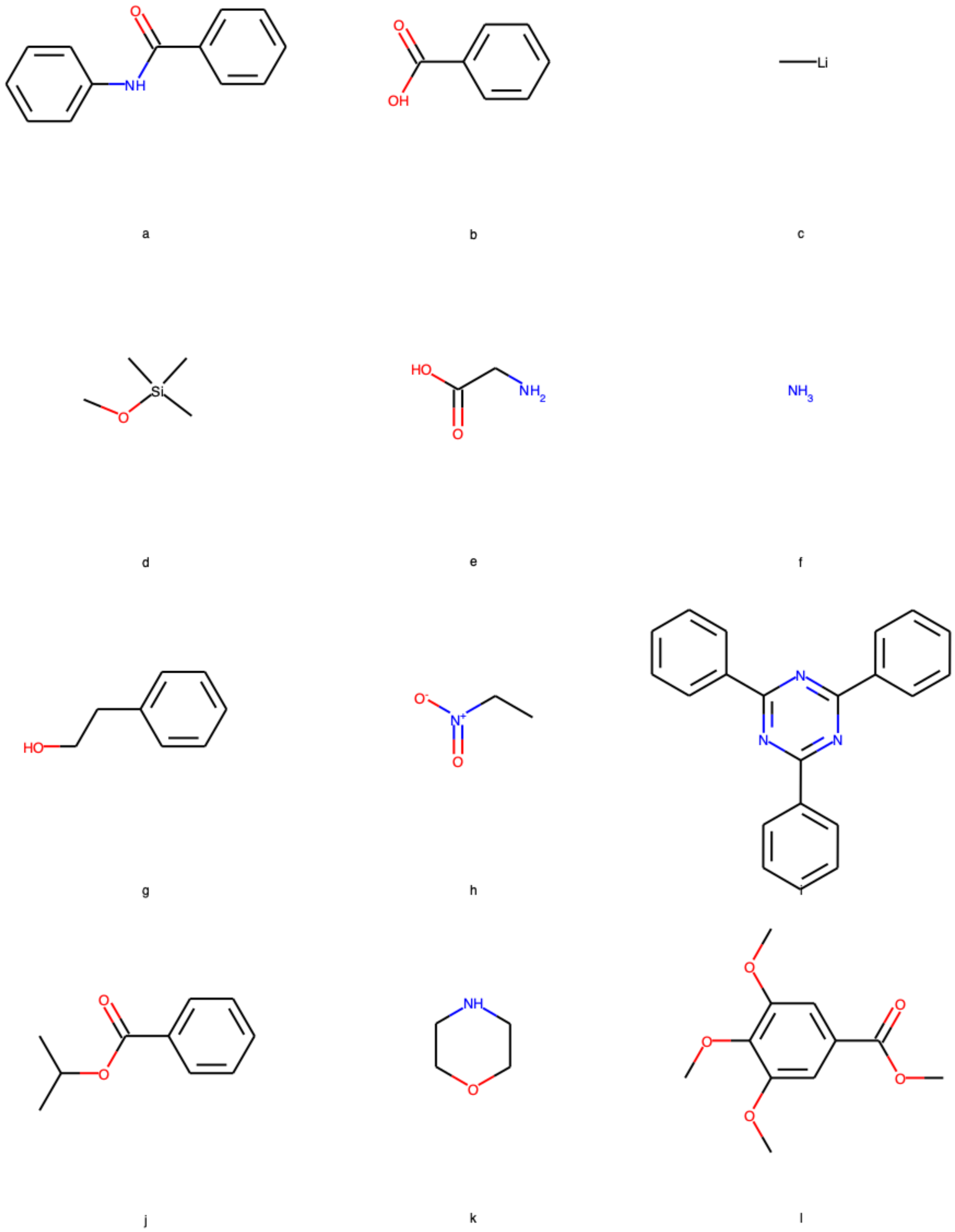


Figure S 38. Molecule set 555

Table S 3. Reaxys IDs of molecule sets 0 to 90

Set	0	RX-ID	15	RX-ID	30	RX-ID	45	RX-ID	60	RX-ID	75	RX-ID	90	RX-ID
	a	508690	a	1867073	a	2208571	a	1817334	a	971266	a	1734623	a	2211315
	b	2039797	b	2101214	b	1424379	b	878450	b	906750	b	509801	b	1724615
	c	1873445	c	1730733	c	969405	c	102391	c	1074863	c	1071236	c	385836
	d	1071744	d	471175	d	1209596	d	1071910	d	2044501	d	391839	d	971516
	e	3902968	e	1072103	e	106909	e	1098214	e	636127	e	508112	e	1904982
	f	635994	f	1862793	f	107477	f	1732463	f	742120	f	606053	f	2691288
	g	1744683	g	742313	g	2043521	g	3587189	g	102551	g	3592982	g	774138
	h	4933662	h	508068	h	3595640	h	1696878	h	1100609	h	102415	h	506021
	i	471382	i	1616740	i	505984	i	3556020	i	605283	i	1727037	i	1107700
	j	1071571	j	608047	j	1912198	j	635760	j	1098935	j	1071207	j	1733203
	k	1906758	k	385772	k	1731042	k	2087538	k	909664	k	4148229	k	509638
	l	605368	l	605303	l	2056090	l	970529	l	507950	l	1771444	l	1101094
	m	1364620	m	2245771	m	1731614	m	1732464	m	1786213	m	506010	m	969148
	n	2049280	n	1303311	n	2039798	n	906905	n	3692537	n	386015	n	385838
	o	607063	o	636783	o	2045489	o	1098278	o	743984	o	1363772	o	1305151

Table S 4. Reaxys IDs of molecule sets 105 to 195

Set	105	RX-ID	120	RX-ID	135	RX-ID	150	RX-ID	165	RX-ID	180	RX-ID	195	RX-ID
	a	471401	a	1425521	a	2036449	a	782061	a	2208089	a	1209238	a	507140
	b	90825	b	385801	b	1236613	b	1878154	b	775403	b	744112	b	1737628
	c	1209246	c	471803	c	606478	c	1560217	c	515874	c	3593645	c	2222141
	d	508910	d	15497285	d	1563093	d	1238185	d	4933359	d	2051911	d	741891

	e	1740761	e	386013	e	3919222	e	102438	e	1072099	e	605459	e	1071474
	f	1423685	f	906677	f	1098522	f	1905012	f	789087	f	2214219	f	506061
	g	1907932	g	1697284	g	216615	g	97217	g	1210120	g	1448766	g	1364714
	h	507951	h	635678	h	1954046	h	745854	h	3563830	h	472690	h	1815558
	i	1903902	i	1098367	i	1858967	i	1740743	i	635807	i	1725147	i	3587218
	j	3692531	j	1363317	j	2357699	j	1901470	j	1907717	j	3535220	j	385876
	k	907196	k	1361672	k	507540	k	4091619	k	5789190	k	605461	k	1921286
	l	2035876	l	386014	l	956570	l	907515	l	1856201	l	906698	l	742413
	m	1099647	m	2050813	m	1101615	m	2043485	m	1900390	m	878795	m	2093095
	n	392449	n	1730800	n	3563831	n	1306359	n	774921	n	1098310	n	635639
	o	635685	o	1680024	o	1239004	o	471281	o	508152	o	878137	o	605631

Table S 5. Reaxys IDs of molecule sets 210 to 300

Set	210	RX-ID	225	RX-ID	240	RX-ID	255	RX-ID	270	RX-ID	285	RX-ID	300	RX-ID
	a	773645	a	1110443	a	1448841	a	2044321	a	1721898	a	507600	a	385735
	b	2212664	b	3587193	b	1754069	b	91034	b	1876374	b	1730716	b	773837
	c	507004	c	742134	c	2204907	c	1909591	c	4660199	c	1817321	c	509985
	d	4720968	d	1446140	d	1874672	d	1905952	d	3599431	d	1912183	d	25122
	e	14282231	e	506917	e	969212	e	635743	e	4976010	e	3903637	e	3547996
	f	471308	f	2208131	f	2554695	f	1909333	f	516726	f	3595638	f	1303312
	g	906744	g	3587194	g	1741921	g	1634058	g	1209320	g	506719	g	1098280
	h	879360	h	1098262	h	2215244	h	108425	h	1209788	h	984320	h	1718793
	i	1730743	i	471352	i	8496933	i	471797	i	1958305	i	2040548	i	605307
	j	3568367	j	1281877	j	3595639	j	1912744	j	1878026	j	774890	j	1100868
	k	3587158	k	4933679	k	1730942	k	2209486	k	1871997	k	969480	k	1098242
	l	1342734	l	506523	l	505999	l	103233	l	1209327	l	3595636	l	506796

	m	742765	m	1879369	m	1911156	m	3587159	m	471391	m	4933678	m	1699658
	n	1562059	n	1340498	n	969215	n	471556	n	3903066	n	393006	n	741937
	o	1721899	o	1911158	o	1702242	o	2050577	o	1696839	o	1730718	o	1718756

Table S 6. Reaxys IDs of molecule sets 315 to 405

Set	315	RX-ID	330	RX-ID	345	RX-ID	360	RX-ID	375	RX-ID	390	RX-ID	405	RX-ID
	a	3595449	a	906737	a	605330	a	3196868	a	4134730	a	506071	a	1072876
	b	1821801	b	605398	b	1107841	b	1730731	b	608199	b	118515	b	636458
	c	605396	c	3587310	c	605762	c	1072101	c	1908121	c	2045132	c	1742050
	d	1868204	d	1911512	d	13149477	d	505945	d	1209228	d	2262644	d	1306914
	e	973593	e	2054709	e	643345	e	969454	e	956566	e	506844	e	3535004
	f	3196867	f	1697939	f	506104	f	606718	f	81568	f	3118345	f	4713419
	g	1857412	g	471388	g	773697	g	1564310	g	1908117	g	969129	g	1099242
	h	1280347	h	1914067	h	782937	h	608018	h	1733451	h	1908172	h	605441
	i	1098229	i	970972	i	2047179	i	506502	i	3732513	i	84272	i	1780973
	j	3535140	j	1460837	j	1704568	j	639794	j	635821	j	742609	j	81567
	k	605365	k	605308	k	471389	k	1281604	k	635680	k	1767780	k	742035
	l	2699534	l	1905149	l	390030	l	1723541	l	3535002	l	506211	l	1236661
	m	385941	m	1736662	m	4921393	m	1932887	m	605842	m	606468	m	1447765
	n	1718733	n	1449572	n	742513	n	743112	n	1697025	n	471493	n	774355
	o	2207336	o	2207355	o	1718732	o	1900717	o	1246142	o	1209227	o	1854721

Table S 7. Reaxys IDs of molecule sets 420 to 510

Set	420	RX-ID	435	RX-ID	450	RX-ID	465	RX-ID	480	RX-ID	495	RX-ID	510	RX-ID
	a	2054084	a	741851	a	506893	a	1107185	a	605438	a	1905429	a	741982
	b	385858	b	773967	b	907511	b	1900508	b	2208085	b	386468	b	635724
	c	514910	c	1911081	c	423400	c	507924	c	608585	c	782650	c	386123
	d	1306723	d	610776	d	8176529	d	1280477	d	2093062	d	1098464	d	635770
	e	1815927	e	5789338	e	606215	e	605258	e	1073987	e	605970	e	1754008
	f	14770794	f	605268	f	506008	f	606081	f	789077	f	9125638	f	1636531
	g	108582	g	741856	g	1707443	g	2047018	g	606474	g	103853	g	1209322
	h	1753975	h	2258952	h	1735581	h	636270	h	1099062	h	1904445	h	105755
	i	1421310	i	605257	i	2037554	i	506892	i	2042392	i	4191822	i	110889
	j	741880	j	508755	j	2236517	j	741857	j	3593646	j	1906923	j	4933243
	k	774955	k	3556712	k	11341079	k	605269	k	506007	k	385737	k	3587155
	l	3121203	l	1940871	l	1905622	l	5859534	l	1566346	l	1915950	l	1759170
	m	3548893	m	878307	m	1901871	m	13195391	m	956776	m	774605	m	1098295
	n	1209425	n	969135	n	385686	n	605632	n	471223	n	774261	n	1913036
	o	1719943	o	606080	o	2045713	o	1754521	o	969158	o	1720586	o	970950

Table S 8. Reaxys IDs of molecule sets 525 to 555

Set	525	RX-ID	540	RX-ID	555	RX-ID
	a	607489	a	1209341	a	1102980
	b	3587190	b	1209725	b	636131
	c	1901563	c	608838	c	3587162
	d	2059239	d	1913256	d	1731490
	e	1907452	e	1900225	e	635782
	f	1854613	f	17008030	f	3587154

	g	746197	g	906770	g	1905732
	h	4267587	h	1909753	h	1209324
	i	1865361	i	4933628	i	239186
	j	908644	j	385791	j	2044384
	k	510011	k	1098293	k	102549
	l	741984	l	1967145	l	2218156
	m	471359	m	3587191	m	
	n	974767	n	969616	n	
	o	1751370	o	1696894	o	