# Application of Automated Network Generation for Retrosynthetic Planning of Potential Corrosion Inhibitors

| | |
|---:|---|
| Journal: | *Molecular Systems Design & Engineering* |
| Manuscript ID | ME-ART-10-2023-000162.R1 |
| Article Type: | Paper |
| Date Submitted by the Author: | 03-Dec-2023 |
| Complete List of Authors: | Lopez, Lauren; Northwestern University, Department of Materials Science and Engineering<br>Zhang, Quan; Northwestern University, Department of Chemical and Biological Engineering<br>Shanks, Brent; Iowa State University, College of Engineering<br>Broadbelt, Linda; Northwestern University, Department of Chemical Engineering |
| | |

# Application of Automated Network Generation for Retrosynthetic Planning of Potential Corrosion Inhibitors

*Lauren M. Lopez[1], Quan Zhang[3], Brent H. Shanks[2], and Linda J. Broadbelt[3,*]*

[1]Department of Materials Science and Engineering, 2220 Campus Drive, Northwestern University, Evanston, Illinois 60208, USA

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

[3]Department of Chemical and Biological Engineering, 2145 Sheridan Road, Northwestern University, Evanston, Illinois 60208, USA

*Corresponding Author: broadbelt@northwestern.edu

## 1.1 Design, System, Application

This work outlines a retrosynthetic framework for generating chemocatalytic libraries given a known list of products and a known list of reactants using automated reaction network generation. The extent of accessible molecular space was assessed using the Tanimoto Similarity Score. This work specifically targeted the synthesis of a list of computationally generated potential organic corrosion inhibitors from a list of candidate bioprivileged molecules. Three systems were explored, with variations in co-reactants or "helper molecules" and amount of reaction families encoded into the network. This work provides a library of reactions for experimental pursuit. With a flexible framework, it allows for retrosynthetic exploration of innumerable reactant or product lists with targeted inclusion of helper molecules or reaction families.

# Application of Automated Network Generation for Retrosynthetic Planning of Potential Corrosion Inhibitors

*Lauren M. Lopez[1], Quan Zhang[3], Brent H. Shanks[2], and Linda J. Broadbelt[3,*]*

[1]Department of Materials Science and Engineering, 2220 Campus Drive, Northwestern University, Evanston, Illinois 60208, USA

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

[3]Department of Chemical and Biological Engineering, 2145 Sheridan Road, Northwestern University, Evanston, Illinois 60208, USA

*Corresponding Author: broadbelt@northwestern.edu

## 1.1 Design, System, Application

This work outlines a retrosynthetic framework for generating chemocatalytic libraries given a known list of products and a known list of reactants using automated reaction network generation. The extent of accessible molecular space was assessed using the Tanimoto Similarity Score. This work specifically targeted the synthesis of a list of computationally generated potential organic corrosion inhibitors from a list of candidate bioprivileged molecules. Three systems were explored, with variations in co-reactants or “helper molecules” and amount of reaction families encoded into the network. This work provides a library of reactions for experimental pursuit. With a flexible framework, it allows for retrosynthetic exploration of innumerable reactant or product lists with targeted inclusion of helper molecules or reaction families.

## 1.2 Abstract

Retrosynthesis is the process of designing chemical pathways from a set of reactants to a set of desired products. However, when both the pools of potential reactants and products grow to a substantial size, this becomes infeasible without the aid of computational tools. This work uses Pickaxe, an automated network generation tool, to perform computational retrosynthesis on a pool of 297 bioprivileged candidate molecules as reactants and 44,003 potential corrosion inhibitors that were generated by a variational autoencoder. Unlike typical approaches in computational synthesis planning, the use of automated network generation allows flexibility in pathways and starting material beyond those that are documented. This work starts by replicating known pathways to corrosion inhibitors from a single bioprivileged candidate molecule and applying the constituent reaction families to the entirety of the reactant pool and concludes by generating networks with a more extensive reaction family list and two sets of co-reactants, or “helper molecules.” Network size, both from the perspective of total reactions enacted and total products formed, was analyzed.

## 1.3 Introduction

Fossil fuels are the building blocks for a range of products, from plastics to fertilizers. However, both environmental and sustainability concerns about their sourcing and uses abound.[1–5] Substitutes are currently being investigated, with a focus on biologically produced alternatives, with some solutions marrying chemical and biological production principles.[6–14] One such solution for a set of diversifiable building blocks that are sourced biologically is bioprivileged molecules, which are defined as “biology-derived chemical intermediates that can be efficiently converted to a diversity of chemical products including both novel molecules and drop-in replacements."[15–17] Among these products, corrosion inhibitors present an ideal target due to the attractiveness of organic alternatives to inorganic corrosion inhibitors that often contain toxic, heavy metals.[18–29] Given the diversity of possible atomic compositions of what could potentially be an organic corrosion inhibitor and the experimental demands of accepted protocols for demonstrating inhibition effectiveness (IE%) through standard ASTM tests, computational approaches are an ideal technique to explore potential molecular space.[30–33] There are two key challenges in developing a portfolio of replacement candidates that can be biologically sourced using a computational framework. The first is developing a list of novel targets that can be reached from biological starting points in a small number of steps, preferably with high yield and high selectivity. The second is to computationally predict inhibition effectiveness (IE%) of a putative corrosion inhibitor. This work focuses on predicting pathways based on chemocatalysis to a set of targets (corrosion inhibitors) from a set of sources (candidate bioprivileged molecules), i.e., how to computationally tackle retrosynthesis. The set of targets is defined based on a pool of candidates derived from known corrosion inhibitors, which based on similarity, obviate the need to couple molecule and pathway design with a prediction of IE%.

Retrosynthesis, or the process of deconstructing a target, has been carried out computationally for many years, but the advent of machine learning has brought forth a new generation of algorithms.[34–38] Computational tools that have embraced machine learning to attempt to tackle this issue include Reaxys, AiZynthFinder, and ASKCOS. [39,40] Reaxys, for example, utilizes artificial intelligence to elucidate pathways to targets based on a library of known reactions. Other studies, such as that by Weber et al., have used Reaxys in their work of analyzing reaction networks.[41] AiZynthFinder uses Monte Carlo methods to deconstruct targets to purchasable precursors based on known reaction templates.[39] ASKCOS, also a freely available tool, has a web interface that additionally predicts the likelihood of a product being formed.[40] While these tools all provide powerful contributions to retrosynthetic planning, the use of AI opens the door to biases present in the training data sets and the libraries of existing reactions. Furthermore, these tools only allow small degrees of customization; particularly relevant to this work is that they are limited by their predefined source list of commercially available chemicals.

Since both the targets and sources in this work are not necessarily commercially available, we sought to not be limited by known reactions and thus used automated network generation. This approach invokes specific reaction families and rules for their application to propagate reactions forward, without those specific reactions having been previously experimentally performed.[42] Automated network generation approaches have proliferated in recent years, and two of these, NetGen and Pickaxe, were developed by us and are based on encoding of specific reaction families, using either matrix representations of connectivity or SMARTS reaction operators, respectively.[42–44] Previous work in which we identified bioprivileged molecules relied on NetGen to explore chemical reactions that ranked the diversification potential of candidate bioprivileged molecules, but it was not used to identify pathways between starting molecules and specific targets, which is a more challenging computational problem because of the combinatorial explosion of potentially long pathways.[16,17,45] Thus, in this work, we transitioned to Pickaxe to produce and evaluate automated reaction networks due to its computational efficiency, use of standard cheminformatics representations which allowed available search algorithms to be deployed, and more compact storage of molecules.[43] The source molecules were the candidate bioprivileged molecules identified by Lopez et al., and the target molecules were corrosion inhibitors distilled from work by Dollar et al., which used variational autoencoders (VAEs) to generate potential corrosion inhibitors from known corrosion inhibitor seed molecules.[16,46] Among these seed molecules were molecules synthesized and tested in Huo et al., which will be computationally reproduced first. The key contribution of this work is the exploration of the accessibility of specific targets from a candidate set of bioprivileged molecules, thereby providing a portfolio of potential corrosion inhibitors, known and novel, that can be derived from biological sources.

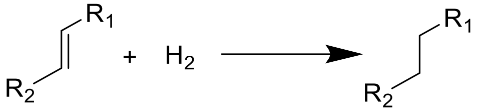
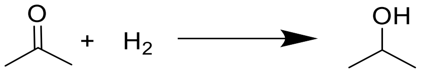
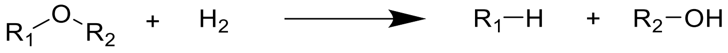
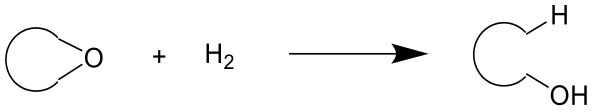
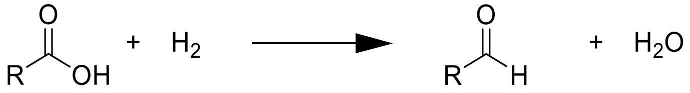
## 1.4 Methods

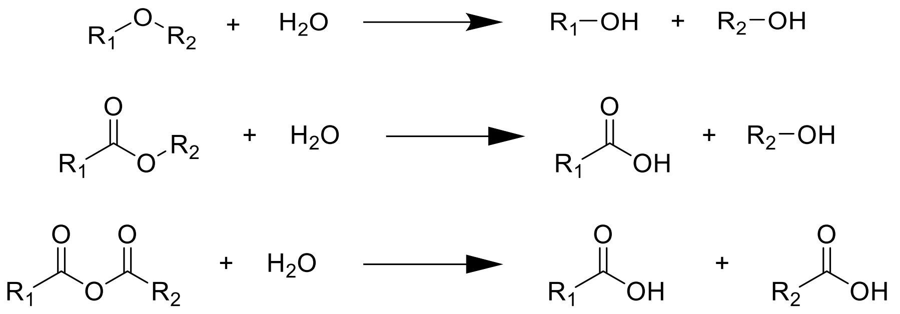
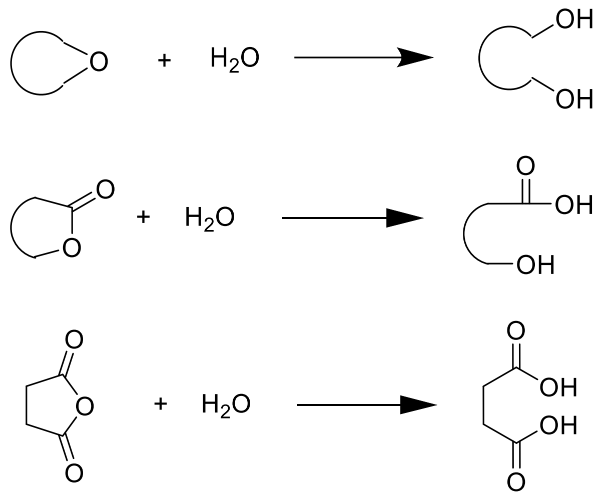
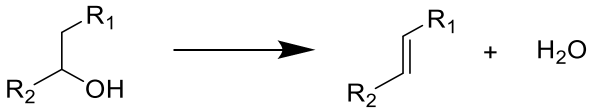
### 1.4.1 Specification of Reaction Families

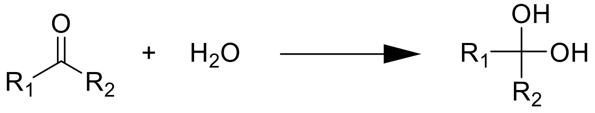
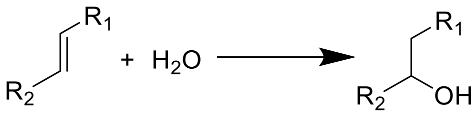
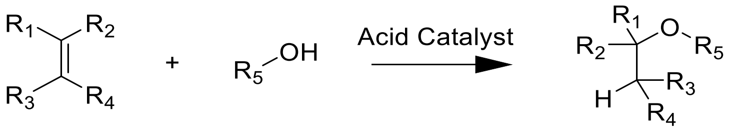
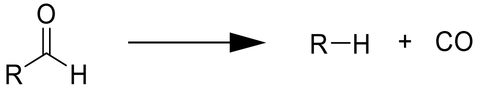
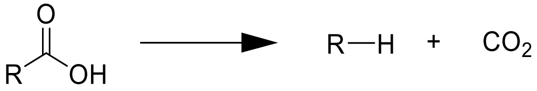
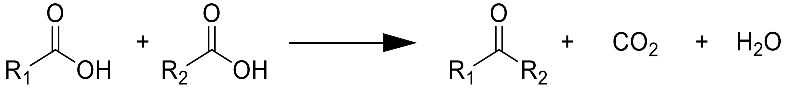
The foundation of automated network generation is specification of the reaction families, or operators, that are allowed. In the first explorations that were performed, which aimed to recreate and expand upon the experimental results in Huo et al., this work used reactions outlined in Zhou et al., Wang et al., and Lopez et al., with additional reaction families added to introduce nitrogen and halogen functionality as summarized in **Table 1**.[16,17,45]

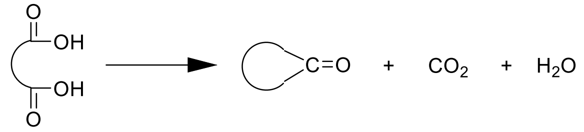
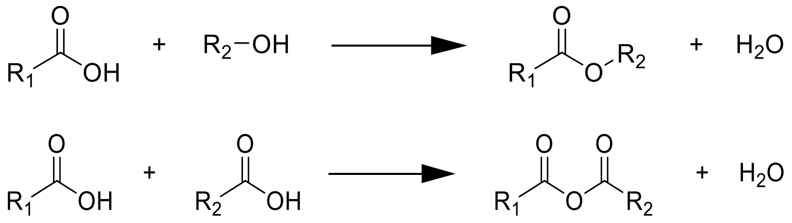
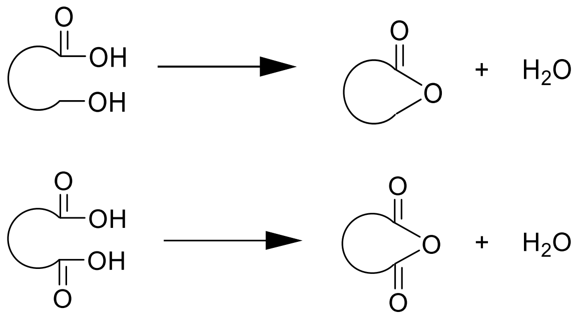
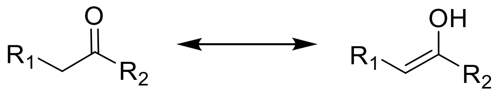
Most of these reaction families require a co-reactant in addition to a specific functional group in a parent molecule, i.e., the potential bioprivileged molecule, or one of its derivatives. For example, hydrogenation needs a molecule containing a carbon-carbon double bond, but it also requires hydrogen to be present. To address this while minimizing the potential combinatorial explosion, the concept of "helper molecules" was introduced, i.e., inclusion of small molecules added to the initial conditions to facilitate reactions. Helper molecules included for the 17 reaction families in the first-generation runs, as well as reactions they might participate in, are also summarized in **Table 1**. Note that these molecules can be changed to be as complex or simple as necessary. In **Table 1**, they are represented by the simplest units possible to ensure the reaction is allowed. Helper molecules are not permitted to react amongst themselves and can only facilitate reactions amongst the original bioprivileged molecule or first-generation progeny. These molecules serve a similar purpose to the ‘strategic molecules’ identified as network hubs in Weber et al.[41]

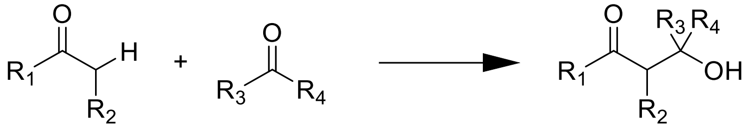
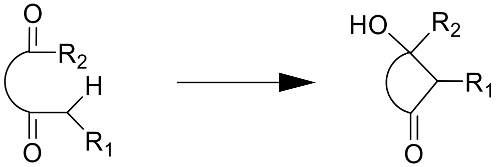
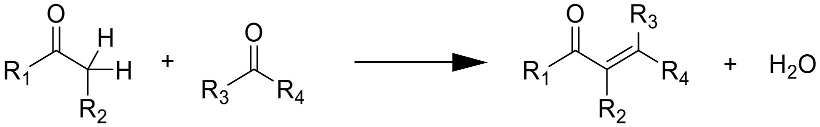
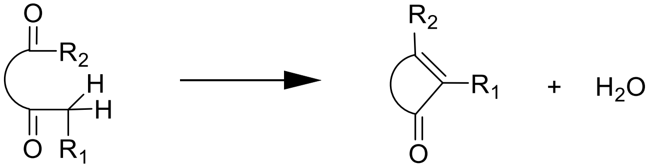
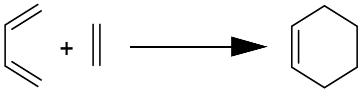
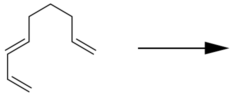
**Table 1** List of 23 initial reaction families and their prototypes used in the first generation of exploration of pathways connecting bioprivileged molecules and potential corrosion inhibitors. Helper molecules of the simplest derivative are shown in the right column.

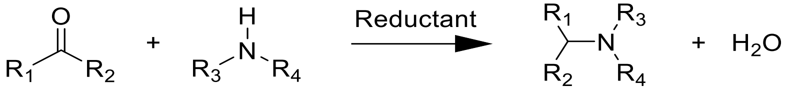
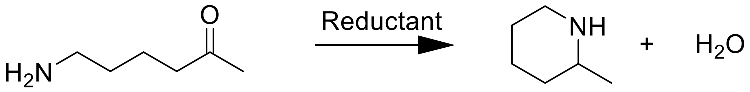
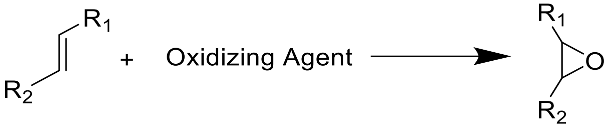
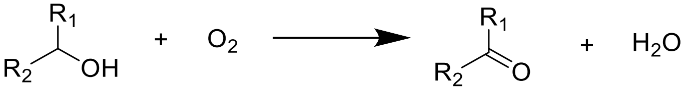
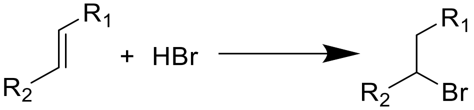
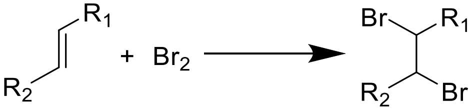
| **Reaction Family** | **SMARTS Representation of Reaction Families and Prototypical Reaction** | **Example Helper Molecule** |
|---|---|---|
| Hydrogenation |  <br>[C+0:1]=[C+0:2].[H][H]>>[*:1][*:2]<br> <br>[C+0!$(*-O):1]=[O+0:2].[H][H]>>[*:1][*:2] | $H_2$ |
| Hydrogenolysis of C-O Bond |  <br>[C,c;+0;!$(*=O):1]!@[O+0:2][C,c;+0;!$(*=O):3].[H][H]>>[*:1].[*:2][*:3]<br> <br>[C,c;+0;!$(*=O):1]@[O+0:2][C,c;+0;!$(*=O):3].[H][H]>>([*:1].[*:2][*:3]) | $H_2$ |
| Hydrodeoxygenation | $R{-}OH + H_2 \longrightarrow R{-}H + H_2O$<br> <br>[C,c;+0:1][O+0H:2].[H][H]>>[*:1].[*:2] | $H_2$ |

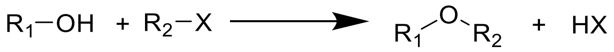
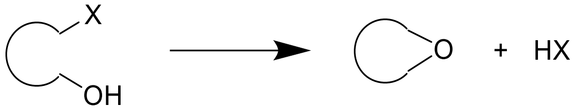
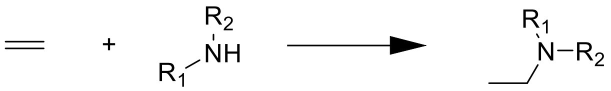
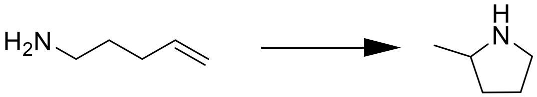
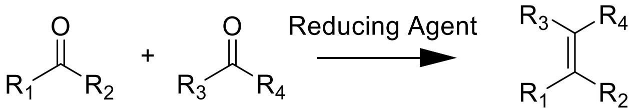
| | | |
|---|---|---|
| Hydrolysis | <br>[C,c;+0:1][O+0:2]-!@[C,c;+0:3].[O+0H2:4]>>[*:1][*:2].[*:3][*:4]<br><br>[C,c;+0:1][O+0:2]-@[C,c;+0:3].[O+0H2:4]>>([*:1][*:2].[*:3][*:4]) | $H_2O$ |
| Dehydration | <br>[C+0!H0:1][C+0:2][O+0H:3]>>[*:1]=[*:2].[*:3] | |

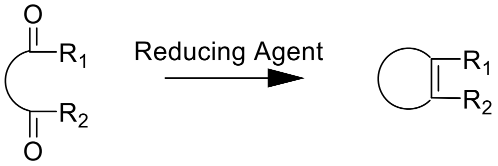
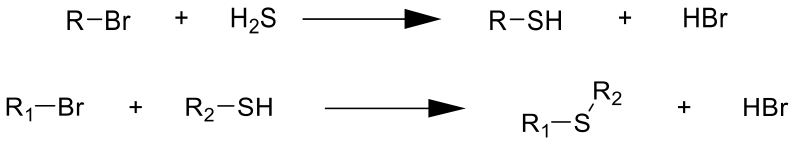
| | | |
|---|---|---|
| Hydration | <br>[C+0;!$(*-O):2]=[O+0:3].[O+0H2:4]>>[*:2]([*:3])[*:4]<br><br><br> <br>[C+0:1]=[C+0:2].[O+0!H0;!$(*-C=O):3]>>[*:1][*:2][*:3] | $H_2O$ |
| Decarbonylation | <br>[*+0:1][C+0H:2]=[O+0:3]>>[*:1].[*-:2]#[*+:3] | |
| Decarboxylation | <br>[*+0:1][C+0:2](=[O+0:3])[O+0H:4]>>[*:1].[*:3]=[*:2]=[*:4] | |
| Ketonization | <br>[*+0:1][C+0:2](=[O+0:3])[O+0H:4].[*+0:5][C+0:6](=[O+0:7])[O+0H:8]>>[*:1][*:2](=[*:3])[*:5].[*:4]=[*:6]=[*:7].[*:8] | $CH_3COOH$ |

| | | |
|---|---|---|
| | <br>([*+0:1][C+0:2](=[O+0:3])[O+0H:4].[*+0:5][C+0:6](=[O+0:7])[O+0H:8])>>[*:1][*:2](=[*:3])[*:5].[*:4]=[*:6]=[*:7].[*:8] | |
| Esterification | <br>[C+0:2](=[O+0:3])[O+0H:4].[C,c;+0:5][O+0H:6]>>[*:2](=[*:3])[*:6][*:5].[*:4]<br><br>([C+0:2](=[O+0:3])[O+0H:4].[C,c;+0:5][O+0H:6])>>[*:2](=[*:3])[*:6][*:5].[*:4] | $CH_3OH$ |
| Keto-enol Tautomerization | <br>[C,N;+0!H0:1][C+0:2]=[O,N;+0:3]>>[*:1]=[*:2][*:3] | |

| | | |
|---|---|---|
| Aldol Condensation |  O<br>R$_1$<br>H<br>R$_2$<br>+<br>O<br>R$_3$<br>R$_4$<br>O<br>R$_3$ R$_4$<br>R$_1$<br>OH<br>R$_2$<br>[C+0:2](=[O+0:3])[C+0!H0:4].[O+0:5]=[C+0X3:6]>>[*:2](=[*:3])[*:4][*:6]([*:5])<br> <br>([C+0:2](=[O+0:3])[C+0!H0:4].[O+0:5]=[C+0X3:6])>>[*:2](=[*:3])[*:4][*:6]([*:5])<br> O<br>H<br>R$_1$<br>H<br>R$_2$<br>+<br>O<br>R$_3$<br>R$_4$<br>O<br>R$_3$<br>R$_1$<br>R$_4$<br>R$_2$<br>+<br>H$_2$O<br>[C+0:2](=[O+0:3])[C+0;H3,H2:4].[O+0:5]=[C+0X3:6]>>[*:2](=[*:3])[*:4]=[*:6].[*:5]<br> O<br>R$_2$<br>H<br>H<br>O<br>R$_1$<br>R$_2$<br>R$_1$<br>O<br>+<br>H$_2$O<br>([C+0:2](=[O+0:3])[C+0;H3,H2:4].[O+0:5]=[C+0X3:6])>>[*:2](=[*:3])[*:4]=[*:6].[*:5] | CH₃CHO |
| Diels-Alder Reaction |  <br>[C+0:1]=[C+0:2][C+0:3]=[C+0:4].[C+0:5]=[C+0:6]>>[*:1]1[*:2]=[*:3][*:4][*:6][*:5]1<br><br>([C+0:1]=[C+0:2][C+0:3]=[C+0:4].[C+0:5]=[C+0:6])>>[*:1]1[*:2]=[*:3][*:4][*:6][*:5]1 | C₂H₄ |

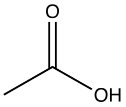
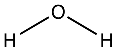
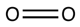
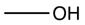
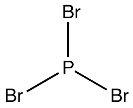
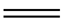
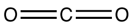
| | | |
|---|---|---|
| Ketone Amination |  O<br>R1<br>R2<br>+<br>H<br>N<br>R3<br>R4<br>Reductant<br>R1<br>R3<br>N<br>R2<br>R4<br>+<br>H2O<br>[CX3!$(*[O,S,N])+0:2](=[O+0:3]).[N+0X3!H0:6].[H][H]>>[*:2][*:6].[*:3]<br> <br>([CX3!$(*[O,S,N])+0:2](=[O+0:3]).[N+0X3!H0:6]).[H][H]>>[*:2][*:6].[*:3] | $NH_3$ |
| Epoxidation |  <br>[C+0:1]=[C+0:2].[O+0:3]=[O+0]>>[*:1]1[*:2][*:3]1 | $O_2$ |
| Selective Oxidation |  <br>[C+0!H0:1][O+0H:2].[O+0:3]=[O+0]>>[*:1]=[*:2].[*:3] | |
| Hydrohalogenation of Alkenes |  <br>[C+0:1]=[C+0:2].[F,Cl,Br,I;+0H:3]>>[*:1][*:2][*:3] | HBr |
| Halogenation of Alkenes |  <br>[C+0:1]=[C+0:2].[F,Cl,Br,I;+0:3][F,Cl,Br,I;+0:4]>>[*:1]([*:3])[*:2][*:4] | $Br_2$ |

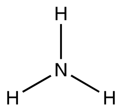
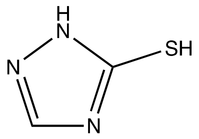
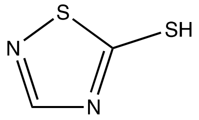
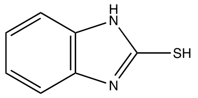
| | | |
|---|---|---|
| Halogenation of Alcohols | $R{-}OH + H{-}X \longrightarrow R{-}X + H_2O$<br>[C+0:1][O+0H:2].[F,Cl,Br,I;+0H:3]>>[*:1][*:3].[*:2]<br>$R{-}OH + PBr_3 \longrightarrow R{-}Br + PBr_2OH$<br>[C,c;+0:1][O+0H:2].[Br;+0:3][P+0:4]>>[*:1][*:3].[*:2][*:4] | PBr₃ |
| Williamson Ether Synthesis |  <br>[C,c;!$(*=O);+0:1][O+0H:2].[CX4!H0,c;+0:3][F,Cl,Br,I;+0:4]>>[*:1][*:2][*:3].[*:4]<br> <br>([C,c;!$(*=O);+0:1][O+0H:2].[CX4!H0,c;+0:3][F,Cl,Br,I;+0:4])>>[*:1][*:2][*:3].[*:4] | |
| Hydroamination |  <br>[C+0:1]=[C+0:2].[N+0X3!H0:3]>>[*:1][*:2][*:3]<br> <br>([C+0:1]=[C+0:2].[N+0X3!H0:3])>>[*:1][*:2][*:3] | NH₃ |
| McMurry Reaction |  <br>[*+0:1][C+0;!$(*-O):2](=[O+0:3]).[*+0:5][C+0;!$(*-O):6](=[O+0:7]).[H][H]>>[*:1][*:2]=[*:6][*:5].[*:3].[*:7] | $CH_3CHO$ |

| | | |
|---|---|---|
| |  <br>([*+0:1][C+0;!$(*-O):2](=[O+0:3]).[*+0:5][C+0;!$(*-O):6](=[O+0:7])).[H][H]>>[*:1][*:2]=[*:6][*:5].[*:3].[*:7] | |
| Thiol-Sulfur Reaction |  <br>[C,c;+0:1][F,Cl,Br,I;+0:2].[S!H0X2+0:3]>>[*:1][*:3].[*:2] | $H_2S$ |

Specifically, the helper molecules used in the first portion of this work are outlined in **Table 2**. The last three molecules in this table correspond to molecules #1, #2, and #9 in Scheme 1 in Huo et al. and can partake in the thiol-sulfur reaction in lieu of $H_2S$ (**Table 1**). An additional condition to note includes the use of $PBr_3$ in lieu of HBr to recreate the halogenation pathway of the alcohol group on triacetic acid lactone (TAL) used in Huo et al. Each C5, C6, and C7 bioprivileged molecule from our previous work underwent its own network expansion, i.e., every network was initiated with one candidate bioprivileged molecule and the list of helper molecules, resulting in 297 total networks in the first instantiation.

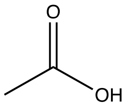
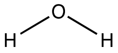
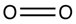
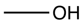
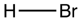
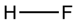
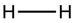
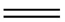
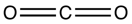
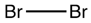
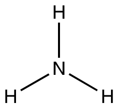
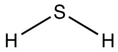
**Table 2** Helper molecules for the first portion of the work. Notably, singular bromine atoms are introduced using $PBr_3$ and sulfur atom introduction comes from the heterocyclic structures outlined in Huo et al.

| Name | SMILES | Structure |
|---|---|---|
| Acetic Acid | CC(O)=O |  |
| Water | O |  |
| Oxygen | O=O |  |
| Methanol | CO |  |
| Phosphorus Tribromide | BrP(Br)Br |  |
| Hydrogen | [H][H] | |
| Carbon Monoxide | [C-]#[O+] | |
| Ethylene | C=C |  |
| Carbon Dioxide | O=C=O |  |
| Bromine | [Br][Br] | |

| | | |
|---|---|---|
| Ammonia | N |  |
| 1H-1,2,4-triazole-3-thiol | SC1=NC=NN1 |  |
| 1,2,4-thiadiazol-5-thio | SC1=NC=NS1 |  |
| 2-mercaptobenzimidazole | SC(N1)=NC2=C1C=CC=C2 |  |

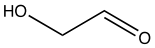
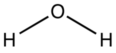
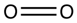
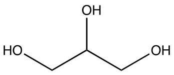
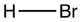
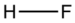
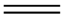
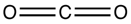
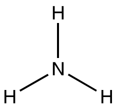
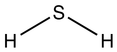
The limited set of reaction operators in **Table 1** was originally conceived in the context of bioprivileged molecules containing carbon, hydrogen and oxygen based on conversion of typical biomass-derived sugars and related compounds. However, corrosion inhibitors may have more diverse functionality, and the small set of operators in **Table 1** may not be diverse enough to connect VAE-derived corrosion inhibitors with bioprivileged molecules. To explore the effect of the diversity of chemistry allowed, the list of possible reaction families was expanded to 242, while enforcing expansion for only two generations. The list of this expanded set of operators is provided in the Supplemental Information. For many of these reaction families, like in the first portion of the work, helper molecules needed to be present to enable bimolecular reactions. To acknowledge the complexity introduced in the additional reaction families, certain helper molecules were simplified from the initial runs. Notably, $PBr_3$ was replaced with HBr and HF, and the introduction of sulfur was simplified from the heterocyclic compounds to $H_2S$. This leaves a list of 13 helper molecules listed in **Table 3**. Once again, every bioprivileged candidate molecule underwent individual network expansion, i.e., the initiating molecules to each network were a single candidate bioprivileged molecule and the list of helper molecules in **Table 3**.

**Table 3** Helper molecules used for the second portion of this work. Sulfur introduction is simplified to $H_2S$ and halogen introduction is changed from the use of $PBr_3$ to the use of HBr and HF as helper molecules.

| **Name** | **SMILES** | **Structure** |
|---|---|---|
| Acetic Acid | CC(O)=O |  |
| Water | O |  |
| Oxygen | O=O |  |
| Methanol | CO |  |
| Hydrobromic Acid | Br |  |
| Hydrofluoric Acid | F |  |
| Hydrogen | [H][H] |  |
| Carbon Monoxide | [C-]#[O+] | |
| Ethylene | C=C |  |
| Carbon Dioxide | O=C=O |  |
| Bromine | [Br][Br] |  |
| Ammonia | N |  |
| Hydrogen Sulfide | S |  |

The purpose of helper molecules is to ensure the propagation of a reaction in the absence of moieties in the primary reactive species, and the logic of choosing the smallest unit of reactive moiety in a helper molecule is to preserve carbon economy. However, a final analysis of network breadth was performed with more biologically relevant helper molecules. Notably, glycerol was used to introduce alcohol functionality that was incorporated in earlier runs using methanol as a helper molecule, and glycolaldehyde was used to incorporate aldehyde functionality instead of carboxylic acid functionality from acetic acid. Both chemicals are accessible through biological sources; glycerol is a byproduct of biodiesel production and glycolaldehyde is a byproduct of the pyrolysis of glucose.[47–49]

**Table 4** Helper molecules used for the third portion of this work. Acetic acid was replaced by glycolaldehyde, and methanol was replaced by glycerol. Both of these helper molecules provide alcohol and aldehyde functionalities from biologically accessible molecules.

| Name | SMILES | Structure |
|---|---|---|
| Glycolaldehyde | C(C=O)O |  |
| Water | O |  |
| Oxygen | O=O |  |
| Glycerol | OCC(O)CO |  |
| Hydrobromic Acid | Br |  |
| Hydrofluoric Acid | F |  |
| Hydrogen | [H][H] | |
| Carbon Monoxide | [C-]#[O+] | |
| Ethylene | C=C |  |
| Carbon Dioxide | O=C=O |  |
| Bromine | [Br][Br] | |
| Ammonia | N |  |
| Hydrogen Sulfide | S |  |

## 1.4.2 Using Automated Network Generation to Mimic Retrosynthesis

While retrosynthesis, by definition, implies deconstructing a product, the operator set formulated for automated reaction network generation of chemocatalytic reactions considers reactions in a “forward” direction as defined by typical reactions carried out in conventional practice. While some reaction families that are practiced routinely in both directions may exist as forward and reverse pairs via two distinct operators, not all reaction operators in **Table 1** or Table S3 have a reverse counterpart. Therefore, networks were generated from a list of desired reactants, i.e., the bioprivileged molecules, and these networks were searched for the creation of desired products, i.e., the target corrosion inhibitors. Specifically, the reactants were the list of candidate bioprivileged 5-carbon (C5), 6-carbon (C6), and 7-carbon (C7) molecules that were identified previously and which are summarized in the Supplementary Information in Table S1.[16] As the reaction operators are not stereospecific, formation of distinct stereoisomers is not delineated, so the 303 candidates proposed in Lopez et al. were reduced to 297 molecules, which will be the number of networks referenced from this point forward. The products were assembled from a list of 44,003 unique molecules, i.e., not accounting for stereochemistry, generated via VAEs using known corrosion inhibitor seeds[46] and are summarized in a separate Supplemental Information file.

Finally, to provide insight to traversing molecular space and the likelihood of forming the desired products, Tanimoto similarity scores, or the Tanimoto index, between bioprivileged reactants and target products were calculated.[50,51] The Tanimoto index is a metric of comparison or "similarity score" given by the following equation:

$$T(A,B) = \frac{(A\ \cap B)}{(A + B - A \cap B)} \tag{1}$$

where *A* and *B* are the molecular fingerprints of the two species being compared, and the Tanimoto index falls inclusively between 0 and 1: "0" means no similar bits, and "1" means all the bits match. These values are calculated using RDKit, with molecules being converted to 2048-bit RDKit molecular fingerprints.

## 1.5 Results and Discussion

### 1.5.1 Molecular and Atomic Characteristics of Bioprivileged Molecules and Target Corrosion Inhibitors

Due to the extensive size of the pool of target corrosion inhibitors and the vastness of the chemical space that could potentially be explored using network generation, we first carried out an analysis to examine the characteristics of both the source and target molecules independently and how they relate to each other. The first step was examining the sources and targets and identifying if any of the proposed targets, i.e., corrosion inhibitors, were identical to any of the 297 sources, i.e., candidate bioprivileged molecules. This is a simple, yet still useful, outcome of this analysis, since the candidate bioprivileged molecules are all known compounds derived from PubChem, and thus the production of these molecules and their potential of being produced from biomass-derived sources is catalogued. Indeed, three candidate bioprivileged molecules, one C5 and two C6 bioprivileged molecules, were identical to molecules in the target pool, depicted in **Figure 1**. As a related important point, their existence within the target pool does not preclude their participation in reaction network generation, as they could form other targets from chemical expansion.

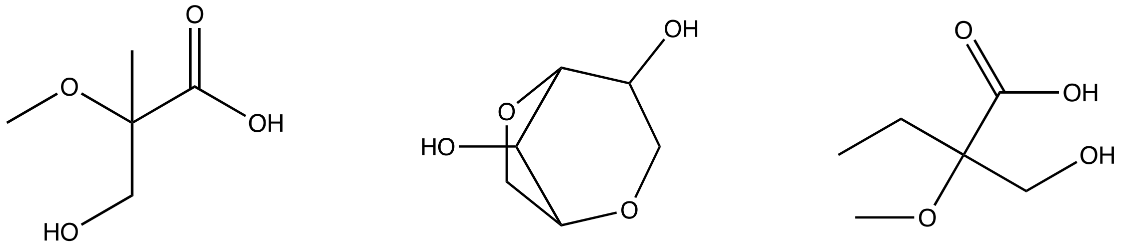


**Figure 1** Three bioprivileged molecules among the 297 C5, C6 and C7 candidates that were identical to three molecules among the 44,003 target molecules predicted to be corrosion inhibitors by the variational autoencoder.

The next analysis of the molecule pools that was carried out was tallying the characteristics of the number and types of atoms in the target molecule pool. Differences in the length and arrangement of the carbon skeleton and diversity of the elemental composition in the target molecules will greatly impact which products from a limited C5-C7 pool of bioprivileged source molecules, where oxygen is the only heteroatom, can be accessed from the available reaction families and helper molecules. **Figure 2** shows the distribution of carbon atoms in the target pool of molecules (i.e., potential corrosion inhibitors); note that the source molecules (i.e., bioprivileged molecules), which span C5-C7, have 103 C5, 103 C6, and 100 C7 members, roughly an equal number near 100 for each carbon number. Approximately 96% of the target molecules have more than seven carbon atoms. It is thus clear from this straightforward analysis that to access most of the target molecules from the specified candidate bioprivileged molecules, carbon atoms need to be introduced and molecular weight growth needs to be possible, and potentially even preferenced given the limited number of steps allowed, among the reaction families and helper molecules included.

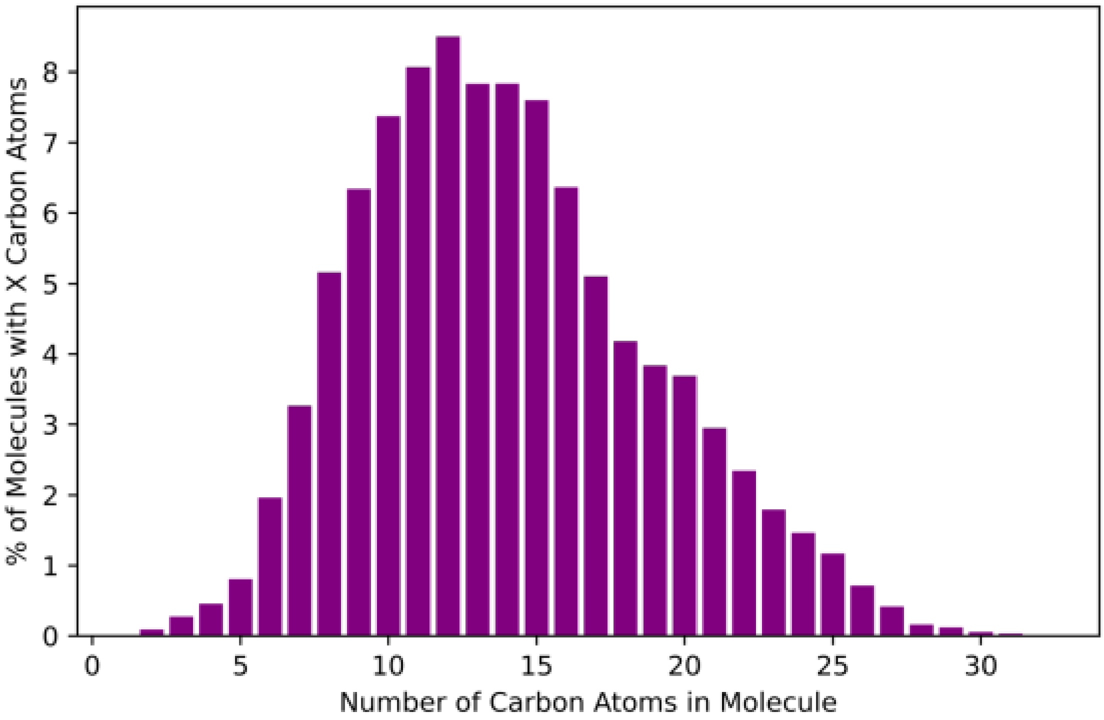


**Figure 2** Percentage of target molecules in the pool of potential corrosion inhibitors that contains a given number of carbon atoms. Approximately 96% of corrosion inhibitor targets have more carbon atoms than the source pool of bioprivileged molecules, which span C5-C7.

To understand the diversity of chemistries that might need to be introduced during reaction network generation, the presence of heteroatoms in the pool of target corrosion inhibitors was tallied as tabulated in **Table 5**. Literature shows that heteroatoms such as oxygen, nitrogen, sulfur, and phosphorous contribute to improving corrosion-inhibiting properties of organic molecules.[30,52] Some heteroatoms, such as phosphorous, were not found in a significant number of molecules in the target pool, which appeared in only 10 out of 44,003 molecules. To the contrary, nitrogen and oxygen are present in more than 66% of the molecules in the target pool. Additionally, sulfur atoms appear in 19% of species, and fluorine and bromine atoms are present in 9% and 2% of species, respectively. Thus, to focus our reaction network generation efforts, we concentrated on diversification of oxygen moieties, which are found in both the source bioprivileged molecules and target corrosion inhibitors, and chemistries that introduce nitrogen, sulfur, and halogens as the most fruitful functionalities to emphasize in specifying the reaction families and defining the helper molecules.

**Table 5** Number of corrosion inhibitors in the target pool out of a total of 44,003 molecules that contain a given heteroatom atom. Oxygen and nitrogen are the most common heteroatoms, occurring in approximately 84% and 69% of molecules, respectively. The next most common heteroatom, sulfur, occurs in approximately 19% of the target pool.

| Heteroatom | # of Targets with Heteroatom |
|---|---|
| Oxygen | 36751 |
| Nitrogen | 30295 |
| Sulfur | 8570 |
| Fluorine | 3941 |
| Bromine | 959 |
| Phosphorous | 10 |

The final analysis of the characteristics of the molecule pools was to tally the distribution of Tanimoto similarity scores between all possible pairs of source and target molecules. The Tanimoto similarity score provides a measure of the “distance” between molecules in chemical space; lower Tanimoto scores indicate a larger distance in chemical space, implying that more reactions would be required to transform a source molecule into the target corrosion inhibitor. Thus, each bioprivileged candidate was compared pairwise to each predicted corrosion inhibitor (for a total of 13,068,891 comparisons), with the distribution of Tanimoto similarity scores summarized in the histogram in **Figure 3**. Nearly all similarity scores, approximately 99.99%, fall below a Tanimoto similarity score of 0.5. Efforts in drug development using the Tanimoto similarity score as a measure of synthetic accessibility suggest that molecules with Tanimoto similarity scores above 0.7 are accessible after two synthesis steps, with molecules having values above 0.85 being “similar.”[53,54] If we apply a cutoff of 0.7 to the results in **Figure 3**, 34 target molecules would be deemed to be accessible via reaction networks with two generations.

However, molecules of interest in drug development are typically of higher molecular weight than the target corrosion inhibitors in this work, and efforts with objectives similar to ours have used lower cutoffs, particularly when retrosynthesis of small molecules is of interest.[34,55] When we apply a lower cutoff of 0.55 to the results in **Figure 3**, there are 678 instances of a bioprivileged molecule and a candidate corrosion inhibitor having a similarity score that meets or exceeds this threshold, with 326 unique corrosion inhibitors involved in these 678 pairings. That is, there should be at least 678 pathways to 326 products within one or two synthesis steps. Importantly, the scope of accessible molecular space without the introduction of new functionalities via helper molecules is limited, providing guidance for the choice of chemistries and justifying the introduction of helper molecules. While the use of the Tanimoto similarity score does not offer the potential efficiency of more customizable similarity searches, it provides a tractable and computationally inexpensive approach for approximating accessibility.[56–59]

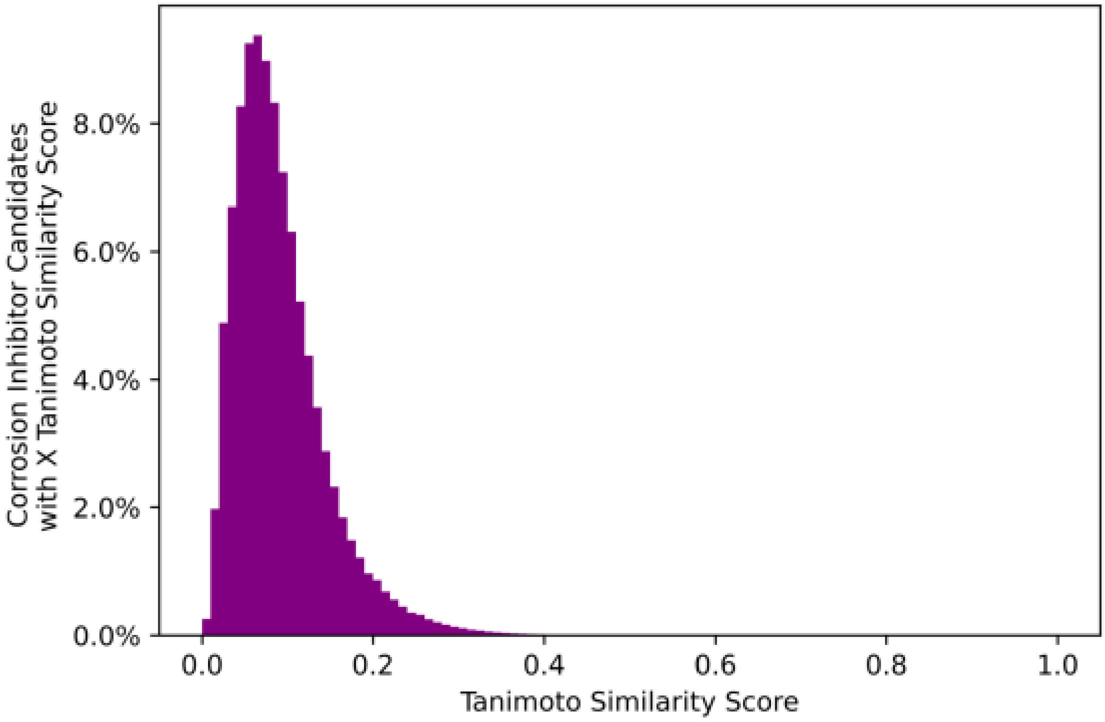
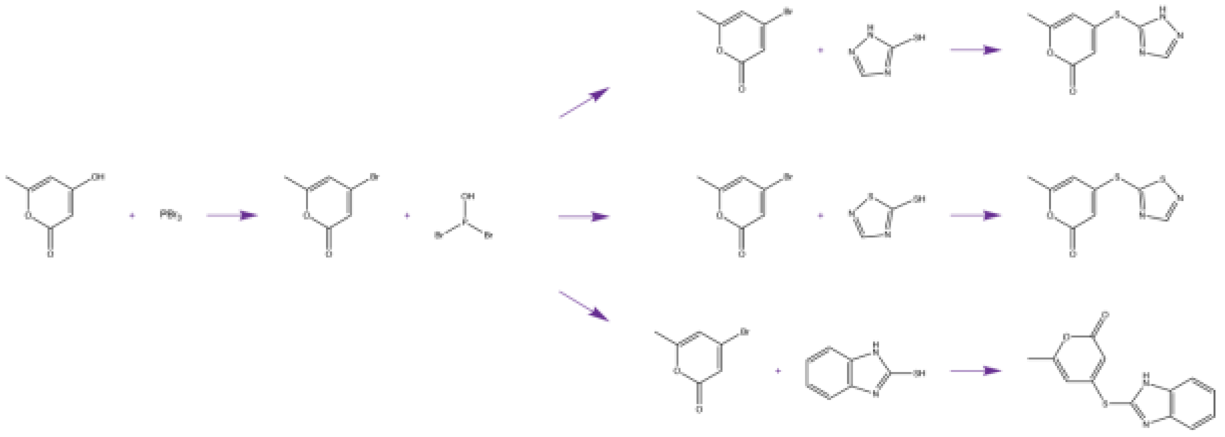


**Figure 3** Histogram showing the percentage of Tanimoto similarity scores that fall within a given range for 13,068,891 pairs of bioprivileged molecules and candidate corrosion inhibitors. Bin size = 100. Over 99.99% of matches fall below a Tanimoto similarity score of 0.5.

### 1.5.2 Calibration of Reaction Network Generation Against Known System

To calibrate the proposed framework, we applied our approach to replicate known pathways from experimental results (**Figure 4**). Specifically, pathways to make corrosion inhibitors by adding heterocyclic structures to triacetic acid lactone (TAL) were replicated, as reported by Huo et al.[60] Specifically, we recreated the pathways to Species #1, #2, and #9 in Scheme 1 of Huo et al., which are shown in **Figure 4** below. To represent these pathways more adequately, $PBr_3$ was used to introduce bromine atoms as opposed to HBr. The structures from Huo et al. not only provide a test of our computational framework, but they were also seeds in the generation of novel corrosion inhibitors using VAEs.



[C,c;+0:1][O+0H:2].[Br;+0:3][P+0:4]>>[*:1][*:3].[*:2][*:4]

(a)

[C,c;+0:1][F,Cl,Br,I;+0:2].[S!H0X2+0:3]>>[*:1][*:3].[*:2]

(b)

**Figure 4** Pathways to corrosion inhibitors as synthesized in Huo et al. as recreated in Pickaxe. The brominated complex is formed in (a) generation 1 and the final product is formed in (b) generation 2.

While the successful production of the pathways to the three corrosion inhibitors that were tested experimentally using automated network generation serves as a proof-of-concept for utilizing for using this tool as a retrosynthesis planner, it does not provide novel pathways. Thus, the work continued with the application of the reaction families used to reproduce the pathways found in Huo et al. as well as various expanded lists to all bioprivileged molecules.

### 1.5.3 Application of Limited Reaction Rules and Small Set of Helper Molecules to All Bioprivileged Molecules

The first step in the creation of novel pathways was to deploy the reaction families in **Table 1** and the helper molecules outlined in **Table 2**. After two generations, over 3,083,285 total molecules were produced across all networks, 3,051,655 of which were unique. Of these molecules, 64 were exact matches for predicted corrosion inhibitors (SMILES provided in the SI). Their accessibility from the candidate bioprivileged sources is shown in **Figure 5**. Approximately 40% of the predicted corrosion inhibitors were formed by more than one candidate bioprivileged molecule; one molecule, valeric acid, was made by 13 sources.

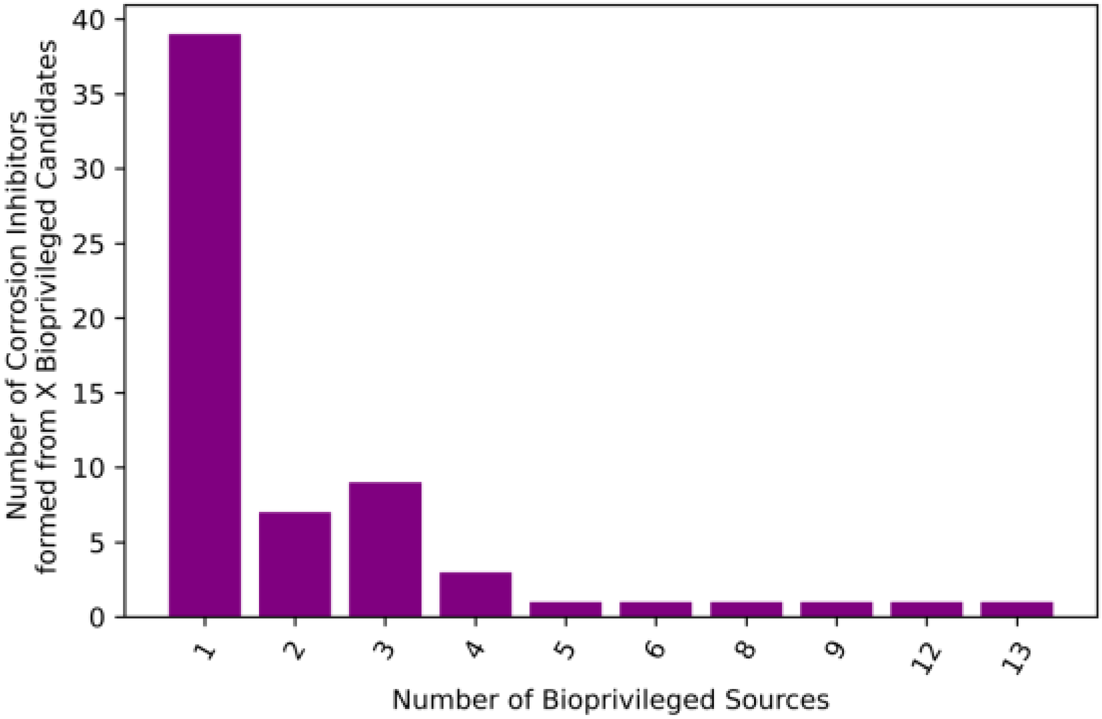


**Figure 5** Accessibility of targets from candidate bioprivileged sources. 48% of targets were formed by only one bioprivileged source. No target was formed by all bioprivileged molecules.

In addition to the distribution of sources, the prolificness of the bioprivileged candidates is also of interest. For those bioprivileged molecules that formed at least one corrosion inhibitor, the number of corrosion inhibitors these bioprivileged molecules individually formed is shown in **Figure 6**. Notably, only 106 of the bioprivileged candidates, or approximately 36%, formed corrosion inhibitors in their networks. Of the 106 that formed corrosion inhibitors, approximately 72% formed only one corrosion inhibitor.

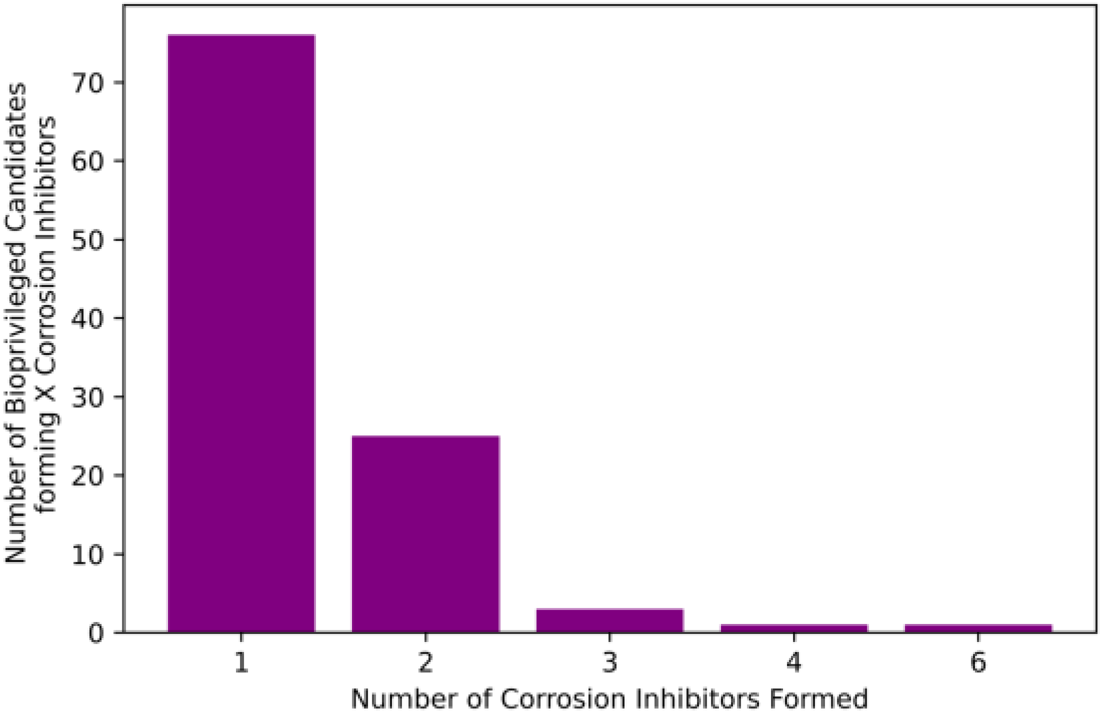
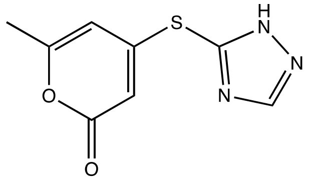
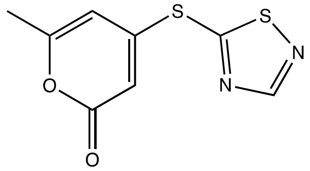
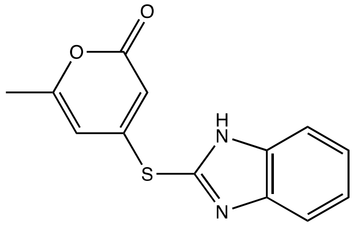


**Figure 6** For those bioprivileged molecules that formed at least one corrosion inhibitor, the number of corrosion inhibitors these bioprivileged molecules individually formed is shown. Most (72%) bioprivileged molecules that formed corrosion inhibitors formed only one corrosion inhibitor.

Drawing on guidance from the analysis of the Tanimoto similarity scores, 64 matches are an underprediction from what is expected from the results in **Figure 3**. This raises the question of whether a similarity score based on the bioprivileged molecule alone, without contributions from the helper molecules, is the source of this discrepancy. To examine this, we analyzed the three corrosion inhibitors that were synthesized experimentally for which pathways using "complex helper molecules" were reproduced using Pickaxe in terms of their similarity scores. The similarity score between the final products, their reactants individually, and their reactants as a composite were calculated and tabulated in **Table 6**. Although there are only two synthesis steps to make these heterocyclic corrosion inhibitors, the Tanimoto similarity score falls below 0.55 for all but one of the nine comparisons. Additionally, the level of specificity used to generate these molecules by tailoring the helper molecules to be identical to the reagents used experimentally is an infeasible exercise for matching 44,003 molecules. Thus, it was next examined whether incorporation of a more extensive list of reaction families but the use of simple helper molecules propagated for two generations could increase the number of pairs of bioprivileged molecules and corrosion inhibitors; the results are tabulated in the next section.

**Table 6** The similarity score between the corrosion inhibitors synthesized by Huo et al. via the pathways in **Figure 4** and either each of their reactants individually or their reactants as a composite. Although the molecules are synthesized in two steps, suggesting that they would be described by Tanimoto similarity scores higher than 0.5, the Tanimoto similarity score falls below 0.5 for all but one of the comparisons.

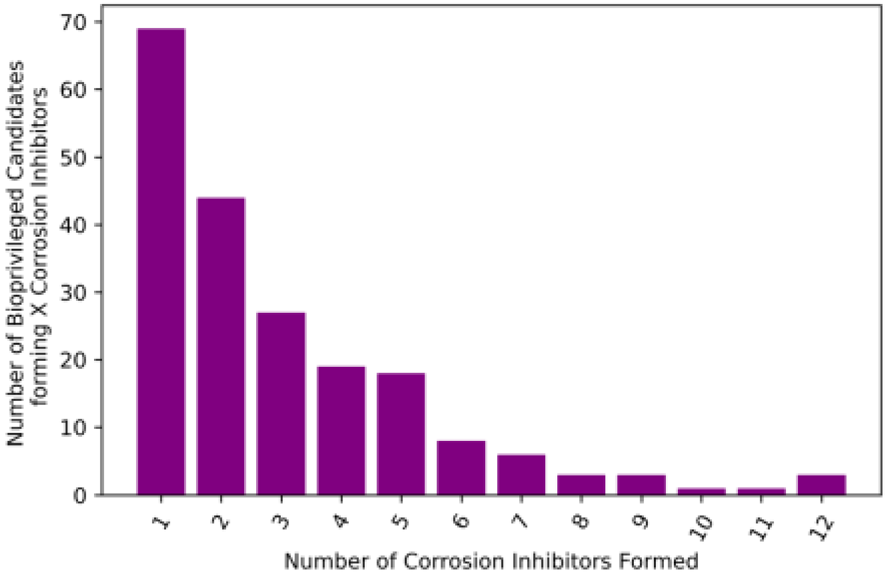
| **Corrosion Inhibitor Target** | **Tanimoto Similarity Score to Reactants** | | |
|---|---|---|---|
| | **TAL** | **Heterocyclic Helper** | **Composite of Reactants** |
|  | 0.33 | 0.13 | 0.42 |
|  | 0.31 | 0.16 | 0.42 |

| | | | |
|---|---|---|---|
|  | 0.29 | 0.38 | 0.58 |

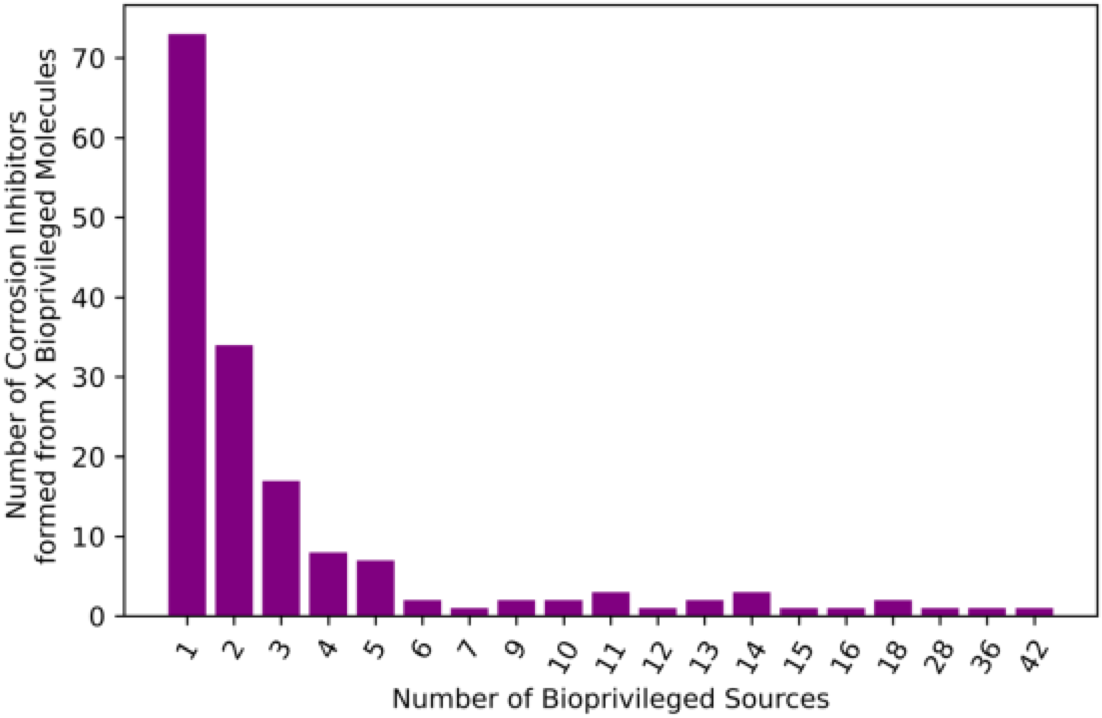
### 1.5.4 Finding Targets and Pathways in Expanded Networks: Extensive Set of Reaction Operators with Simple Helper Molecules

Using the set of simple helper molecules in **Table 3** and the expansive set of reaction operators summarized in Table S3, 19,072,011 total products were formed in the union of the 297 networks via a total of 32,649,547 reactions, with 18,841,363 of those products being unique. The intersection of the products formed from the 297 networks was small, with approximately only 2% of the products being formed by more than one expansion, and thus, each of the 297 expansions contributed a large number of distinct molecules. Of the 18,841,363 unique products, 162 were exact matches for the potential corrosion inhibitors. This increased number is closer in line with predictions based on the analysis of the Tanimoto similarity scores shown in **Figure 3** and confirm that the use of an expanded set of reaction operators was productive. Note, however, that a larger set of corrosion inhibitor “hits” comes with computational challenges, as the number of pathways to produce them is vast. Thus, first statistics related to the corrosion inhibitors that were formed and the bioprivileged candidates to produce them were tallied before examining any pathways in detail.

**Figure 7** examines the two dimensionalities of the networks: how many potential corrosion inhibitors are formed by each bioprivileged candidate (**Figure 7**a) and via how many bioprivileged candidate sources each corrosion inhibitor is formed (**Figure 7**b). To provide more insight into these measures which were introduced in the last section, the first criterion acknowledges the need of a bioprivileged molecule to be diversifiable – the candidate that can form the most potential corrosion inhibitors is the most valuable in this paradigm. The second criterion acknowledges accessibility of the corrosion inhibitor candidates as predicted by reaction network generation. Those with more sources and pathways are more likely to have experimentally viable pathways among those predicted. Related to the first criterion, i.e., diversifiability, there were some excellent candidates, as 43 bioprivileged molecules formed five or more corrosion inhibitors. Related to the second criterion, three potential corrosion inhibitors emerged as highly attractive, as they could be formed by more than 20 bioprivileged molecules. Notably, with the expanded reaction family operators available, 202 candidate bioprivileged molecules, or 68% of the starting molecules, formed one or more predicted corrosion inhibitor.



(a)



(b)

**Figure 7** For expanded list of reaction operators and list of simple helper molecules, summary of (a) diversifiability of bioprivileged sources as measured by how many potential corrosion inihibitors can be formed from them in two generations and (b) accessibility of potential corrosion inhibitors formed from bioprivileged candidates as measured by how many bioprivileged molecules can form them in two generations.

To narrow the list further, within the paradigm of bioprivileged molecules, those candidates which form the most potential corrosion inhibitors are of the most interest. Four candidate bioprivileged molecules have the potential of forming more than ten corrosion inhibitors; these promiscuous bioprivileged candidates are shown in **Figure 8**. Notably, all of these are C5 bioprivileged candidates. The most promiscuous C6 and C7 candidate bioprivileged molecules produce “only” six and eight target corrosion inhibitors, respectively. Another notable feature is the lack of a ring structure in any of the most prolific bioprivileged molecules; while ring structures are present both in the proposed corrosion inhibitor list and amongst the candidate bioprivileged molecules, the complexity of the ring structures within the proposed corrosion inhibitors makes them inaccessible within two generations, and thus, linear structures emerge as the most viable pairs from the present analysis.

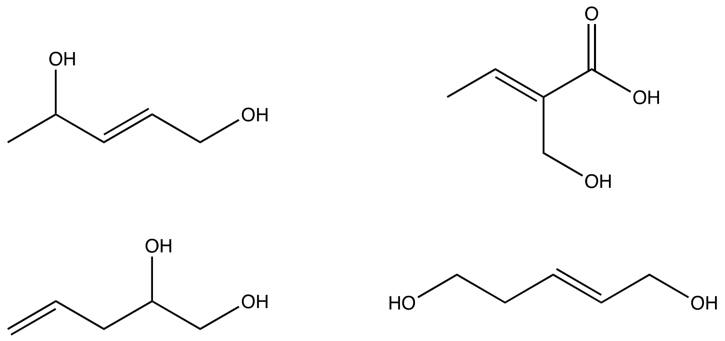
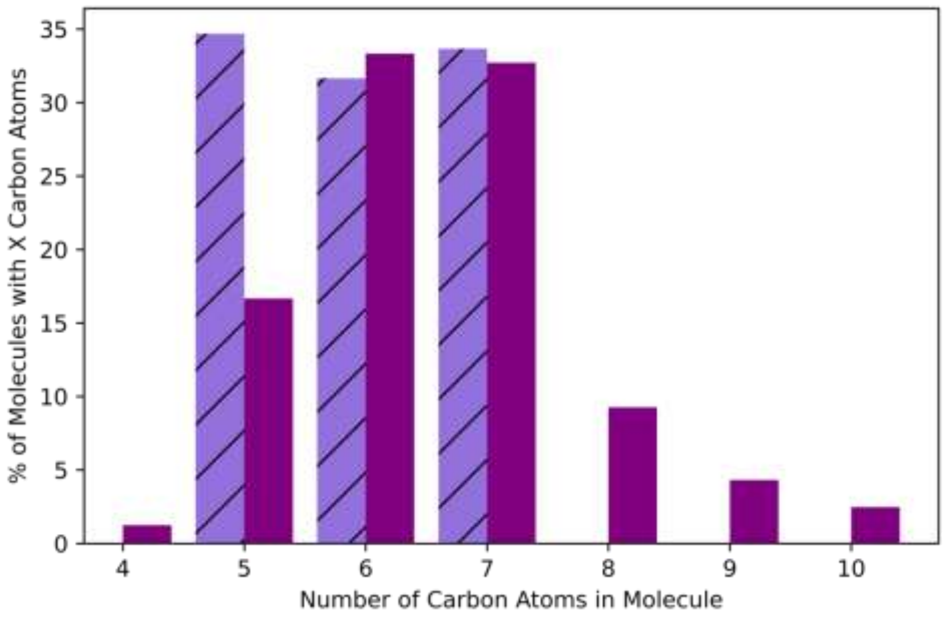
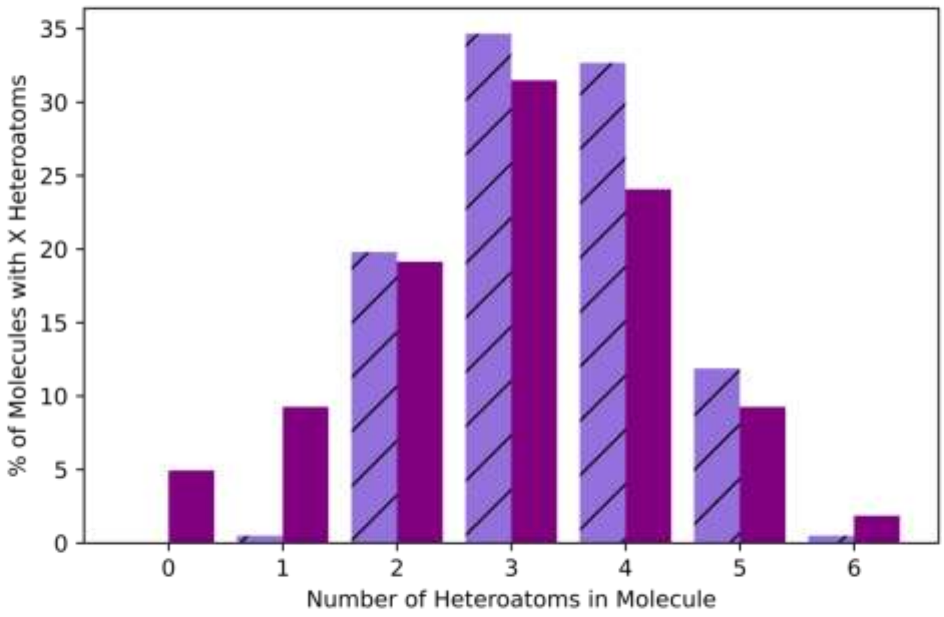


**Figure 8** Using the expanded list of reaction operators on Table S3 and the list of simple helper molecules in **Table 3**, four starting molecules emerged as the most promiscuous bioprivileged molecules, as they each form more than 10 potential corrosion inhibitors.

The 162 potential corrosion inhibitors that were formed were examined to assess how many carbon atoms and heteroatoms they contained, and it was seen that a bias towards small molecules, as expected from **Figure 2**, is exhibited in the targets that were formed. From **Figure 9**a, although 96% of the corrosion inhibitor targets have a carbon atom count greater than seven, only 21% of the targets generated have a carbon atom count greater than seven. This is consistent with the observation noted earlier that the promiscuous bioprivileged molecules came from the C5 pool and have limited opportunities to grow more than two or three carbon atoms beyond their initial carbon backbone. To the contrary, in **Figure 9**b, the heteroatom distribution more closely resembled that seen across the entire target pool, with 66% of the targets that were formed containing between three and five heteroatoms. It is important to note that the number of heteroatoms was not normalized by the number of carbon atoms, and the mean number of heteroatoms per carbon atom could differ.



(a)



(b)

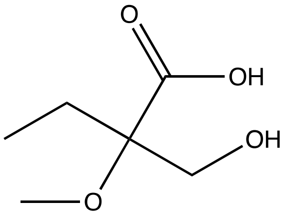
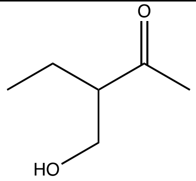
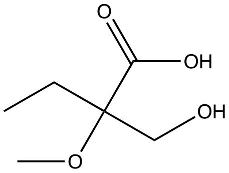
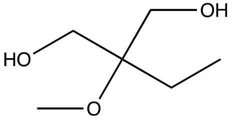
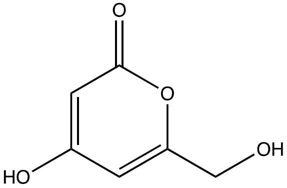
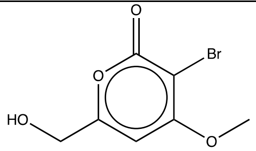
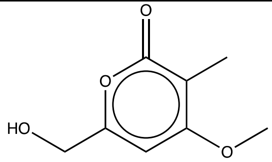
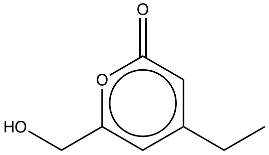
**Figure 9** Analysis of the 162 potential corrosion inhibitors that were formed from 202 bioprivileged molecules as starting molecules according to the number of carbon atoms and heteroatoms they contain: (a) carbon atom distribution of the 162 potential corrosion inhibitors that were formed (solid bar) compared to the carbon atom distribution of the 202 bioprivileged

candidates as starting molecules (hashed bar) and (b) heteroatom distribution of the 162 potential corrosion inhibitors that were formed (solid bar) compared to the heteroatom distribution of the 202 bioprivileged candidates as starting molecules (hashed bar).

The next feature of note to explore was the extent of connectivity, i.e., how many pathways result in one of the corrosion inhibitors as a product, and the pathways were characterized by how many generations they spanned and how many reactions comprised them. Note that due to the fact that two first-generation products may combine to form a product in a second-generation reaction, two generations of reaction can result in pathways comprised of three reactions. These statistics for exemplar corrosion inhibitor/bioprivileged molecule pairs are summarized in **Table 7**. The statistics of individual pathways connecting corrosion inhibitor/bioprivileged molecule pairs are derived from searches of very large reaction networks. The smallest individual network had 2,471 reactions, while the largest individual network had 4,343,820 reactions. The second largest network in terms of reactions was 1,195,823, or a 72% reduction from the largest network. The average number of reactions in each individual network was approximately 110,000, with a standard deviation of approximately 282,000, exemplifying the broad distribution of reaction network sizes and skewness. The examples shown in **Table 7** are typical of the diversity of the pathways to potential corrosion inhibitors, as most products have multiple pathways.

**Table 7.** Statistics for pathways connecting two exemplar corrosion inhibitor/bioprivileged molecule pairs that summarize how the pathways were characterized, i.e., by how many generations they spanned and how many reactions comprised them. The first candidate bioprivileged molecule listed is also a corrosion inhibitor target, **Figure 1**.

| **Candidate Bioprivileged Starting Molecule** | **Potential Corrosion Inhibitor Formed** | **# of Pathways of X Generations** | | **# of Pathways with X Reactions** | | |
|---|---|---|---|---|---|---|
| | | **1** | **2** | **1** | **2** | **3** |

| | | | | | | |
|---|---|---|---|---|---|---|
|  |  | 0 | 4 | 0 | 4 | 0 |
| |  | | | | | |
| |  | 1 | 7 | 1 | 7 | 0 |
|  |  | 0 | 4 | 0 | 4 | 0 |
| |  | 0 | 4 | 0 | 4 | 0 |
| |  | 0 | 1 | 0 | 1 | 0 |

Interestingly, the total network size did not necessarily correlate with most corrosion inhibitors formed, as seen in **Figure 10**. One of the best performing bioprivileged molecules as measured by the degree of accessibility to predicted corrosion inhibitors led to the generation of a network in the lower half of the network size distribution at 60,805 total reactions. All of the top four bioprivileged molecules had networks comprised of fewer than 150,000 reactions. Likewise, the largest network, with over four million reactions, only produced one corrosion inhibitor. Recall, as well, that the networks only had about 2% overlap. This highlights that the number of reactions does not necessarily dictate the successful formation of target products, while the expansion of the number of reaction operators, and thus reaction types, does have a positive correlation.

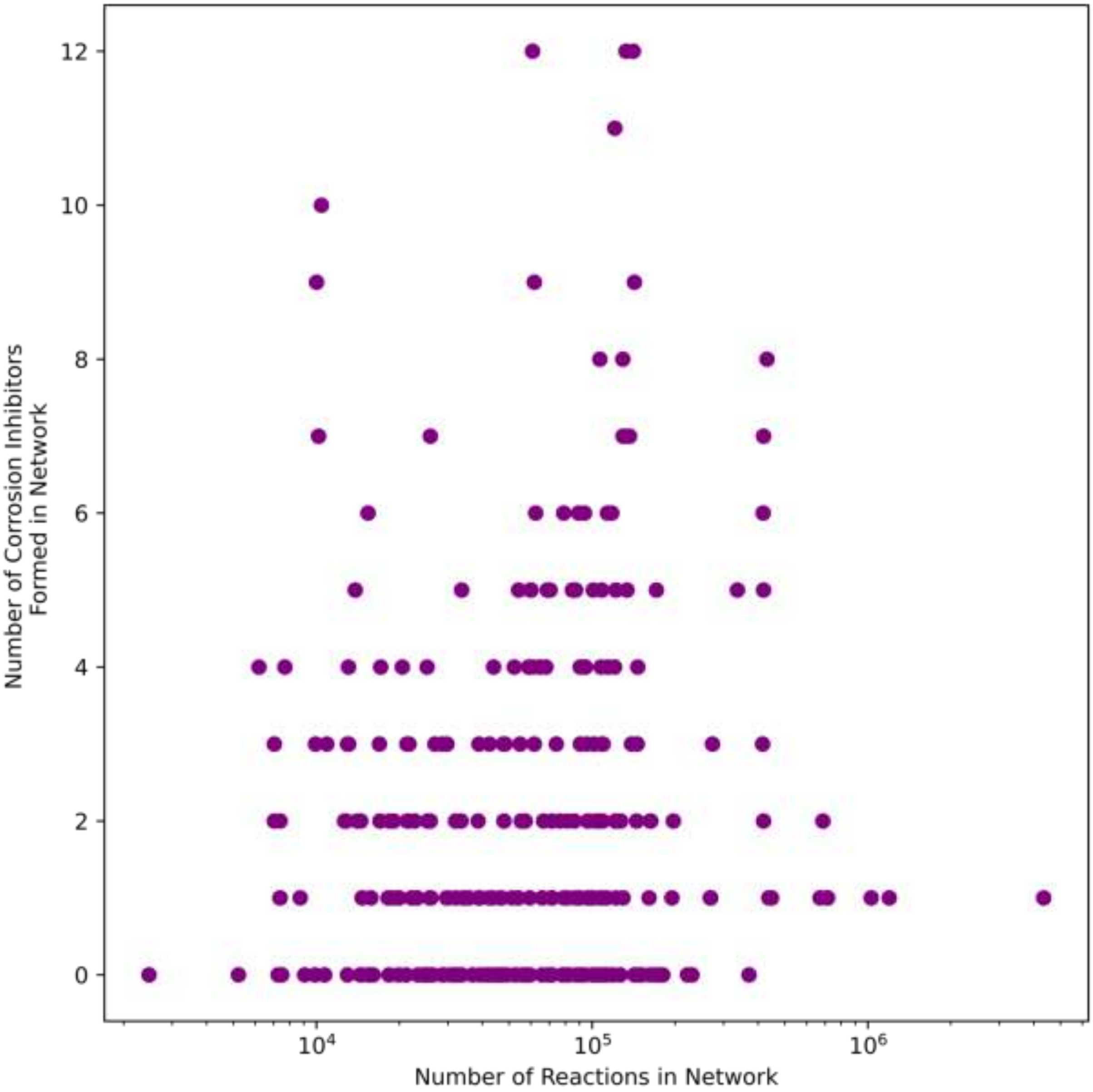


**Figure 10** Comparison of the number of corrosion inhibitors formed as targets versus the number of total reactions in each individual reaction network for the case of the expanded list of reaction operators and small helper molecules.

Among the 202 candidate bioprivileged molecules that produced corrosion inhibitors in their reaction networks, for a total of 602 matches, there were a total of 4,512 pathways. A breakdown of the pathways is in **Table 8**. Most of the products, 99.2%, were formed in the second generation. Approximately 62% of the products formed in the second generation were comprised of reactions from two products from the first generations. Three pathways were comprised of zero reactions, which are the three molecules in **Figure 1**. The maximum number of potential pathways to one of the targets was 332 pathways. The most desirable pathways would be the shortest, as theoretically those would be the most accessible in an experimental setting. Thus, the pathways to targets in the first generation and pathways only comprised of two reactions would be the highest priority to consider pursuing experimentally.

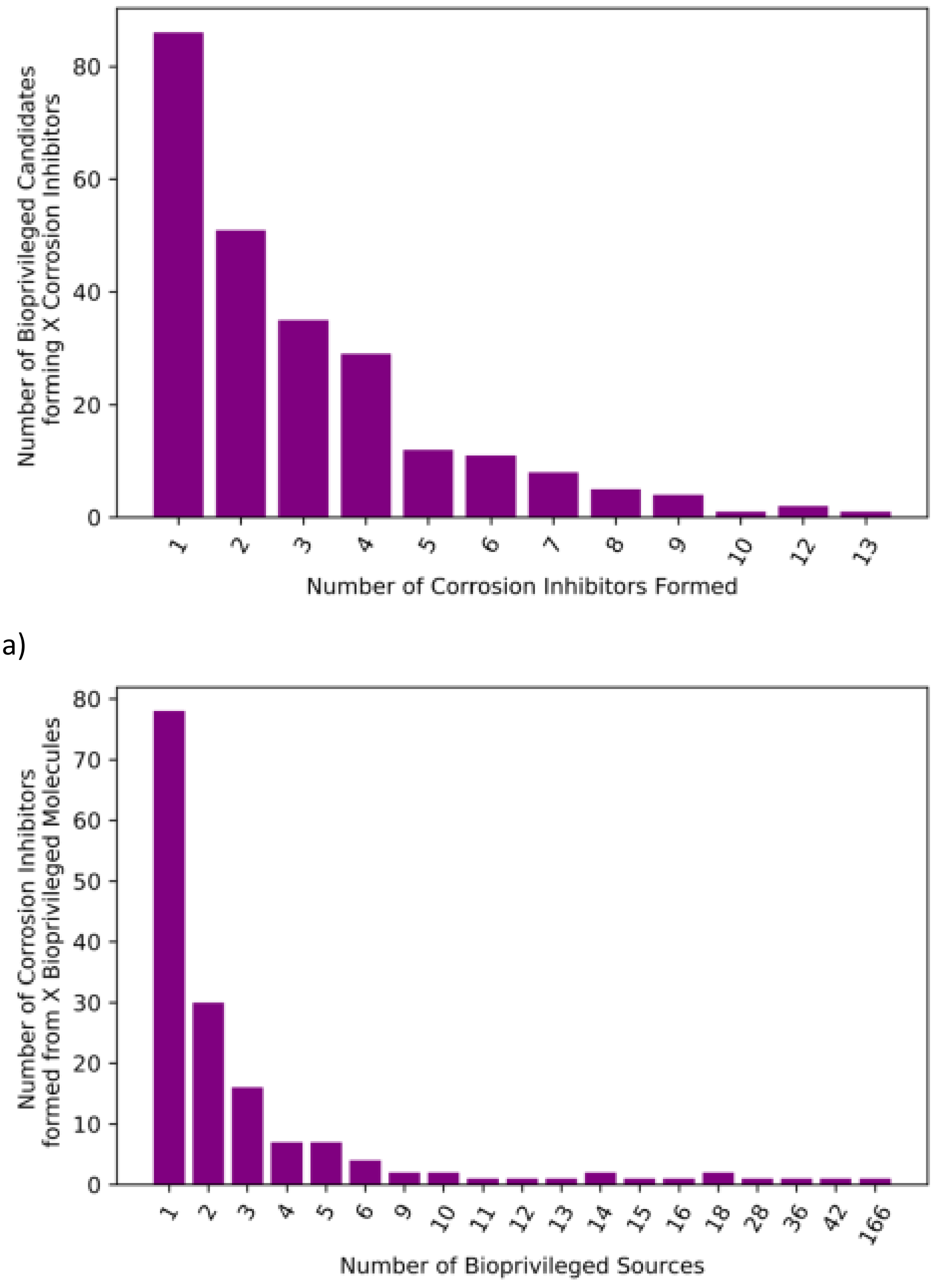
**Table 8** Analysis of pathways to make targets, including generation the product was formed and number of reactions for product formation. Three bioprivileged molecules were corrosion inhibitor targets, which would make them a 0th generation product.

| Network Attributes of Corrosion Inhibitors | # of Pathways |
|---|---|
| 1st Generation Product | 52 |
| 2nd Generation Product | 4,457 |
| Product from two reactions | 2,279 |
| Product from three reactions | 2,159 |

### 1.5.5 Retrosynthesis in Reaction Networks with Alternate Helper Molecules

As helper molecules largely contribute to the breadth of reaction space exploration, an alternative set of helper molecules (**Table 4**) was introduced to the more extensive set of reaction families (Table S3). Using each of the 297 bioprivileged molecules as a starting point, the union of the reaction networks resulted in 35,592,071 total products from a total of 61,210,576 reactions, with approximately 35,310,000 of those products being unique. The intersection of the products formed from the 297 networks was small, with approximately only 1% of the products being formed by more than one expansion. Of these products, 159 were exact matches for the potential corrosion inhibitors, which is even smaller than what was expected from **Figure 3** and achieved using the set of helper molecules in **Table 3**. Notably, there was an 87% increase in the total number of products and reactions, but a 2% decrease in exact matches compared to the previous case when helper molecules with smaller carbon backbones were used.

As was done earlier, **Figure 11** examines two dimensionalities of the networks: how many potential corrosion inhibitors are formed by each bioprivileged candidate (**Figure 11**a) and via how many bioprivileged candidate sources each corrosion inhibitor is formed (**Figure 11**b). Four predicted corrosion inhibitors could be formed by more than 20 bioprivileged molecules, and 44 bioprivileged molecules formed five or more corrosion inhibitors. Thus, the diversifiability and accessibility of corrosion inhibitors did not have a notable improvement over the previous run as evaluated by these two metrics. However, overall there were more bioprivileged molecules and potential corrosion inhibitors that were connected by putative pathways with the expanded set of helper molecules. Specifically, 245 candidate bioprivileged molecules, or 82% of the starting molecules, formed one or more predicted corrosion inhibitor, and the number of total corrosion inhibitors that could be reached increased by 19% to 714, far exceeding the prediction from **Figure 3**. This is an interesting result that demonstrates that the choice of both reaction families and helper molecules drive the pathway discovery process. While fewer unique corrosion inhibitors were formed than might be expected given the scale of the number of molecules and reactions, but those that were formed were more accessible from the bioprivileged molecules.



(a)

(b)

**Figure 11** For expanded list of reaction operators and list of more complex helper molecules, summary of (a) diversifiability of bioprivileged sources as measured by how many potential corrosion inhibitors can be formed from them in two generations and (b) accessibility of potential corrosion inhibitors formed from bioprivileged candidates as measured by how many bioprivileged molecules can form them in two generations.

As shown in **Figure 12**, similar results were found for a lack of correlation between network size and number of corrosion inhibitors found as was observed for the previous run. The largest network was once again from the expansion of furylpyruvic acid, C1=COC(=C1)CC(=O)C(=O)O, which had 7,649,792 total reactions and formed only one potential corrosion inhibitor, 2-furoic acid, O=C(O)c1ccco1. This product was the same corrosion inhibitor target formed when simple helper molecules were used. The minimum network size increased to 6,238 reactions, and the average number of reactions doubled to approximately 206,100. The standard deviation increased to approximately 528,000, demonstrating that the variance of network size was also large in this case. Because there were even fewer target corrosion inhibitors reached for this case, we did not analyze the individual pathways in detail.  However, the specific corrosion inhibitors that were connected to bioprivileged molecules are summarized in the Supplementary Information in Table S5, and pathways to them can be reproduced using Pickaxe by an interested reader.

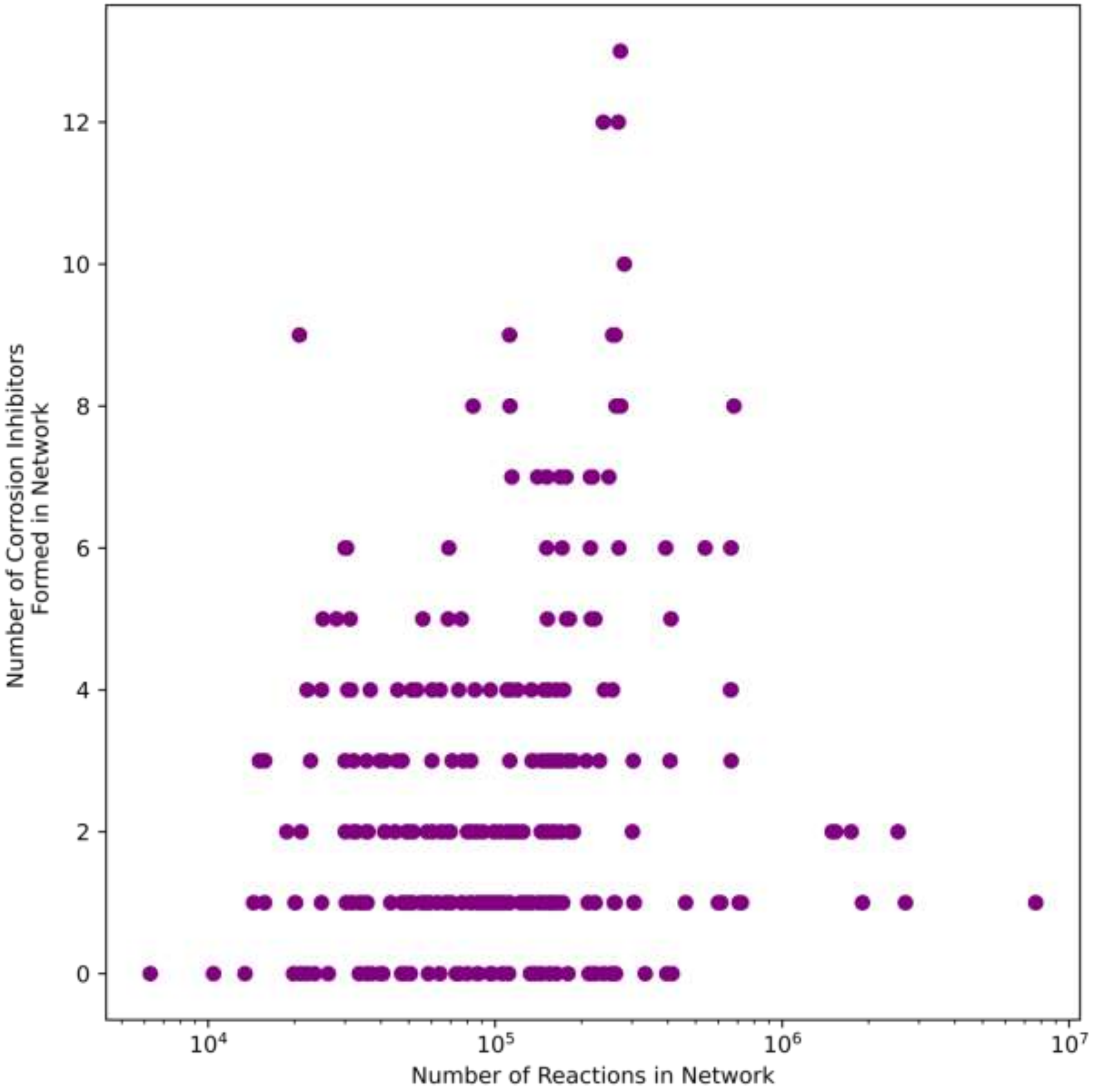


**Figure 12** Comparison of the number of corrosion inhibitors formed as targets versus the number of total reactions in each individual reaction network for the case of the expanded list of reaction operators and the larger helper molecules.

## 1.6 Conclusion

Automated network generation was used to perform computational retrosynthesis on a pool of 303 bioprivileged candidate molecules as reactants and 44,003 potential corrosion inhibitors that were generated by a variational autoencoder. While artificial intelligence was used to create the 44,003 molecules, pathways to produce them from biological starting points is appealing. However, this is too vast of a chemical space to manually design chemical pathways, especially when considering multiple starting points. Automated network generation using the Pickaxe framework can be deployed to identify possible synthesis pathways, and this was successfully demonstrated in this work. However, we also demonstrated that the choice of the reaction families and the helper molecules chosen to deploy them have a significant influence on creating pathways between bioprivileged molecules as starting points and potential corrosion inhibitors as targets. We also showed that it is useful to carry out basic analyses using chemiformatics tools, such as Tanimoto similarity indexing, to set an expectation value for the number of targets that can be achieved in a given number of steps, and prior to analyzing individual pathways, a simple summary of the number of targets that are reached from the set of starting molecules provides guidance about whether the reaction operator set and helper molecules are sufficiently rich. We also demonstrated how the use of automated network generation for retrosynthesis benefits more from the breadth of the operator set used rather than the depth of reaction networks expanded for a small operator set. However, there is little correlation between network size as measured by the total number of reactions in the union of all individual networks and the number of targets formed. Targeted inclusion of helper molecules and specific reaction types accelerates pathway discovery, especially in instances where the structural features of targets deviate greatly from the source molecules. As a final note, we point out that the most time-consuming step of the computational workflow was the identification of pathways, which contained a maximum of three reactions. To achieve more rapid analysis of putative pathways that are longer, future work would benefit from filtering on-the-fly during network expansion, such as applying thermodynamic boundaries on free energies of reaction or Tanimoto similarity filtering.

## 1.7 Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Bioenergy Technologies Office Funding Opportunity Announcement DE-FOA-0001916 BioEnergy Engineering for Products Synthesis (BEEPS)", Award Number DE-EE0008492. This work was authored in part by the Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE), under Contract No. DE-AC36-08GO28308//SUB-2021-10692 Mod 1.

## 1.8 References

1 D. J. Soeder, Fossil Fuels and Climate Change, *Fracking and the Environment*, 2021, 155–185.

2 S. Shafiee and E. Topal, When will fossil fuel reserves be diminished?, *Energy Policy*, 2009, **37**, 181–189.

3 M. Höök and X. Tang, Depletion of fossil fuels and anthropogenic climate change—A review, *Energy Policy*, 2013, **52**, 797–809.

4 D. J. Wuebbles and A. K. Jain, Concerns about climate change and the role of fossil fuel use, *Fuel Processing Technology*, 2001, **71**, 99–119.

5 K. R. Abbasi, M. Shahbaz, J. Zhang, M. Irfan and R. Alvarado, Analyze the environmental sustainability factors of China: The role of fossil fuel energy and renewable energy, *Renew Energy*, 2022, **187**, 390–402.

6 T. Werpy and G. Petersen, *Top Value Added Chemicals from Biomass: Volume I -- Results of Screening for Potential Candidates from Sugars and Synthesis Gas*, Golden, CO (United States), 2004.

7 Y. S. Jang, B. Kim, J. H. Shin, Y. J. Choi, S. Choi, C. W. Song, J. Lee, H. G. Park and S. Y. Lee, Bio-based production of C2-C6 platform chemicals, *Biotechnol Bioeng*, 2012, **109**, 2437–2459.

8 T. J. Schwartz, B. H. Shanks and J. A. Dumesic, Coupling chemical and biological catalysis: A flexible paradigm for producing biobased chemicals, *Curr Opin Biotechnol*, 2016, **38**, 54–62.

9 W. Wu, M. R. Long, X. Zhang, J. L. Reed and C. T. Maravelias, A framework for the identification of promising bio-based chemicals, *Biotechnol Bioeng*, 2018, **115**, 2328–2340.

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

11 X. Zhang, C. J. Tervo and J. L. Reed, Metabolic assessment of E. coli as a Biofactory for commercial products, *Metab Eng*, 2016, **35**, 64–74.

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

13 T. J. Schwartz, B. J. O'Neill, B. H. Shanks and J. A. Dumesic, Bridging the chemical and biological catalysis gap: Challenges and outlooks for producing sustainable chemicals, *ACS Catal*, 2014, **4**, 2060–2069.

14 C. Zhang and A. A. Lapkin, Hybridizing organic chemistry and synthetic biology reaction networks for optimizing synthesis routes, *ChemRxiv*, , DOI:10.26434/chemrxiv-2022-hh2nr.

15 B. H. Shanks and P. L. Keeling, Bioprivileged molecules: Creating value from biomass, *Green Chemistry*, 2017, **19**, 3177–3185.

16 L. M. Lopez, B. H. Shanks and L. J. Broadbelt, Identification of bioprivileged molecules: expansion of a computational approach to broader molecular space, *Mol. Syst. Des. Eng.*, 2021.

17 X. Zhou, Z. J. Brentzel, G. A. Kraus, P. L. Keeling, J. A. Dumesic, B. H. Shanks and L. J. Broadbelt, Computational Framework for the Identification of Bioprivileged Molecules, *ACS Sustain Chem Eng*, 2019, **7**, 2414–2428.

18 P. B. Raja and M. G. Sethuraman, Natural products as corrosion inhibitor for metals in corrosive media - A review, *Mater Lett*, 2008, **62**, 113–116.

19 M. Costa and C. B. Klein, Toxicity and carcinogenicity of chromium compounds in humans, *Crit Rev Toxicol*, 2006, **36**, 155–163.

20 R. M. Park, J. F. Bena, L. T. Stayner, R. J. Smith, H. J. Gibb and P. S. J. Lees, Hexavalent chromium and lung cancer in the chromate industry: A quantitative risk assessment, *Risk Analysis*, 2004, **24**, 1099–1108.

21 N. Strigul, A. Koutsospyros and C. Christodoulatos, Tungsten speciation and toxicity: Acute toxicity of mono- and poly-tungstates to fish, *Ecotoxicol Environ Saf*, 2010, **73**, 164–171.

22 T. C. Diamantino, L. Guilhermino, E. Almeida and A. M. V. M. Soares, Toxicity of sodium molybdate and sodium dichromate to Daphnia magna Straus evaluated in acute, chronic, and acetylcholinesterase inhibition tests, *Ecotoxicol Environ Saf*, 2000, **45**, 253–259.

23 Y. Fang, B. Suganthan and R. P. Ramasamy, Electrochemical characterization of aromatic corrosion inhibitors from plant extracts, *Journal of Electroanalytical Chemistry*, 2019, **840**, 74–83.

24 D. Sukul, A. Pal, S. K. Saha, S. Satpati, U. Adhikari and P. Banerjee, Newly synthesized quercetin derivatives as corrosion inhibitors for mild steel in 1 M HCl: Combined experimental and theoretical investigation, *Physical Chemistry Chemical Physics*, 2018, **20**, 6562–6574.

25 R. F. B. Cordeiro, A. J. S. Belati, D. Perrone and E. D'elia, Coffee Husk as Corrosion Inhibitor for Mild Steel in HCl Media, *Int. J. Electrochem. Sci*, 2018, **13**, 12188–12207.

26 H. L. Y. Sin, A. Abdul Rahim, C. Y. Gan, B. Saad, M. I. Salleh and M. Umeda, Aquilaria subintergra leaves extracts as sustainable mild steel corrosion inhibitors in HCl, *Measurement (Lond)*, 2017, **109**, 334–345.

27 U. F. Ekanem, S. A. Umoren, I. I. Udousoro and A. P. Udoh, Inhibition of mild steel corrosion in HCl using pineapple leaves (Ananas comosus L.) extract, *J Mater Sci*, 2010, **45**, 5558–5566.

28 V. V. Torres, R. S. Amado, C. F. de Sá, T. L. Fernandez, C. A. da S. Riehl, A. G. Torres and E. D'Elia, Inhibitory action of aqueous coffee ground extracts on the corrosion of carbon steel in HCl solution, *Corros Sci*, 2011, **53**, 2385–2392.

29 M. Mobin, M. Basik and J. Aslam, Pineapple stem extract (Bromelain) as an environmental friendly novel corrosion inhibitor for low carbon steel in 1 M HCl, *Measurement (Lond)*, 2019, **134**, 595–605.

30 L. Guo, I. B. Obot, X. Zheng, X. Shen, Y. Qiang, S. Kaya and C. Kaya, Theoretical insight into an empirical rule about organic corrosion inhibitors containing nitrogen, oxygen, and sulfur atoms, *Appl Surf Sci*, 2017, **406**, 301–306.

31 F. Mansfeld, in *Corrosion: Fundamentals, Testing, and Protection*, ASM International, 2003, pp. 446–462.

32 R. Baboian, *Corrosion tests and standards: application and interpretation*, ASTM international, 2005, vol. 20.

33 C. Verma, D. K. Verma, E. E. Ebenso and M. A. Quraishi, Sulfur and phosphorus heteroatom-containing compounds as corrosion inhibitors: An overview, *Heteroatom Chemistry*, 2018, **29**, e21437.

34 Y. Podolyan, M. A. Walters and G. Karypis, Assessing synthetic accessibility of chemical compounds using machine learning methods, *J Chem Inf Model*, 2010, **50**, 979–991.

35 W. D. Ihlenfeldt and J. Gasteiger, Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs, *Angewandte Chemie International Edition in English*, 1996, **34**, 2613–2633.

36 E. J. Corey and W. Todd Wipke, Computer-assisted design of complex organic syntheses, *Science (1979)*, 1969, **166**, 178–192.

37 C. W. Coley, W. H. Green and K. F. Jensen, Machine Learning in Computer-Aided Synthesis Planning, *Acc Chem Res*, 2018, **51**, 1281–1289.

38 O. Engkvist, P. O. Norrby, N. Selmi, Y. hong Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, Computational prediction of chemical reactions: current status and outlook, *Drug Discov Today*, 2018, **23**, 1203–1218.

39 S. Genheden, A. Thakkar, V. Chadimová, J. L. Reymond, O. Engkvist and E. Bjerrum, AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning, *J Cheminform*, 2020, **12**, 70.

40 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, Prediction of Organic Reaction Outcomes Using Machine Learning, *ACS Cent Sci*, 2017, **3**, 434–443.

41 J. M. Weber, P. Lió and A. A. Lapkin, Identification of strategic molecules for future circular supply chains using large reaction networks, *React Chem Eng*, 2019, **4**, 1969–1981.

42 L. J. Broadbelt, S. M. Stark and M. T. Klein, Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates, *Ind Eng Chem Res*, 1994, **33**, 790–799.

43 K. M. Shebek, J. Strutz, L. J. Broadbelt and K. E. J. Tyo, Pickaxe: a Python library for the prediction of novel metabolic reactions, *BMC Bioinformatics 2023 24:1*, 2023, **24**, 1–15.

44 S. Vernuccio and L. J. Broadbelt, Discerning complex reaction networks using automated generators, *AIChE Journal*, 2019, **65**, e16663.

45 G. Wang, L. Lopez, M. Coile, Y. Chen, J. M. Torkelson and L. J. Broadbelt, Identification of known and novel monomers for poly(hydroxyurethanes) from biobased materials, *Ind Eng Chem Res*, 2021, **60**, 6814–6825.

46 O. Dollar, N. Joshi, D. A. C. Beck and J. Pfaendtner, Attention-based generative models for de novo molecular design, *Chem Sci*, 2021, **12**, 8362–8372.

47 H. W. Tan, A. R. Abdul Aziz and M. K. Aroua, Glycerol production and its applications as a raw material: A review, *Renewable and Sustainable Energy Reviews*, 2013, **27**, 118–127.

48 C. B. Schandel, M. Høj, C. M. Osmundsen, A. D. Jensen and E. Taarning, Thermal Cracking of Sugars for the Production of Glycolaldehyde and Other Small Oxygenates, *ChemSusChem*, 2020, **13**, 688–692.

49 P. Kostetskyy, M. W. Coile, J. M. Terrian, J. W. Collins, K. J. Martin, J. F. Brazdil and L. J. Broadbelt, Selective production of glycolaldehyde via hydrothermal pyrolysis of glucose: Experiments and microkinetic modeling, *J Anal Appl Pyrolysis*, 2020, **149**, 104846.

50 D. Bajusz, A. Rácz and K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?, *J Cheminform*, 2015, **7**, 1–13.

51 A. Capecchi, D. Probst and J. L. Reymond, One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome, *J Cheminform*, 2020, **12**, 43.

52 M. Goyal, S. Kumar, I. Bahadur, C. Verma and E. E. Ebenso, Organic corrosion inhibitors for industrial cleaning of ferrous and non-ferrous metals in acidic solutions: A review, *J Mol Liq*, 2018, **256**, 565–573.

53 H. Matter, Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors, , DOI:10.1021/JM960352+.

54 D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, Neighborhood behavior: A useful concept for validation of ‘molecular diversity’ descriptors, *J Med Chem*, 1996, **39**, 3049–3059.

55 J. D. Holliday, N. Salim, M. Whittle and P. Willett, Analysis and display of the size dependence of chemical similarity coefficients, *J Chem Inf Comput Sci*, 2003, **43**, 819–828.

56 A. Cho, H. Yun, J. H. Park, S. Y. Lee and S. Park, Prediction of novel synthetic pathways for the production of desired chemicals, *BMC Syst Biol*, 2010, **4**, 1–16.

57 U. Fechner and G. Schneider, Flux (1): A virtual synthesis scheme for fragment-based de novo design, *J Chem Inf Model*, 2006, **46**, 699–707.

58 M. Garcia-Castro, S. Zimmermann, M. G. Sankar and K. Kumar, Scaffold Diversity Synthesis and Its Application in Probe and Drug Discovery, *Angewandte Chemie International Edition*, 2016, **55**, 7586–7605.

59 J. L. Baylon, N. A. Cilfone, J. R. Gulcher and T. W. Chittenden, Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification, *J Chem Inf Model*, 2019, **59**, 673–688.

60 J. Huo, W. Bradley, K. Podolak, B. J. Ryan, L. T. Roling, G. A. Kraus and B. H. Shanks, Triacetic Acid Lactone and 4-Hydroxycoumarin as Bioprivileged Molecules for the Development of Performance-Advantaged Organic Corrosion Inhibitors, *ACS Sustain Chem Eng*, 2022, **10**, 11544–11554.