Biosynth Pipeline: An integrated platform for merging the computational design of chimeric type I polyketide synthases with enzymatic pathways for chemical biosynthesis

Yash Chainani[1,2,3,4], Jacob Diaz[3,4], Margaret Guilarte-Silva[1,2], Vincent Blay[3,4], Keith E. J. Tyo[1,2], Linda J. Broadbelt[1,2,3,4], Aindrila Mukhopadhyay[3,4], Jay D. Keasling[3,4,5,6,7,8], Hector Garcia Martin[3,4,9,10], and Tyler W. H. Backman[3,4]

1. Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA
2. Center for Synthetic Biology, Evanston, IL, USA
3. Joint BioEnergy Institute, Emeryville, CA, USA
4. Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
5. QB3, University of California, Berkeley, Berkeley, CA, USA
6. Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA, USA
7. Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA
8. Center for Biosustainability, Danish Technical University, Lynby, Denmark
9. BCAM, Basque Center for Applied Mathematics, Bilbao, Spain,
10. DOE Agile BioFoundry, Emeryville, California

Abstract

Synthetic biology offers the promise of manufacturing chemicals more sustainably than petrochemistry. Yet, both the rate at which biomanufacturing can synthesize these molecules and the net chemical accessible space are limited by existing pathway discovery methods which rely on arduous literature searches. Here, we present an automated retrobiosynthesis tool, Biosynth Pipeline, that simultaneously tackles both problems by integrating multifunctional type I polyketide synthases (PKSs) with monofunctional enzymes to propose the synthesis of desired target chemicals via two new tools: DORAnet and RetroTide. While monofunctional enzymes are valuable for carefully decorating a substrate's carbon backbone, they typically cannot expand the backbone itself. PKSs can, instead, predictably do this through their unique ability to catalyze carbon-carbon bond formation reactions iteratively. We evaluated the performance of Biosynth Pipeline against a previously published set of 155 molecules of interest for biomanufacturing, and report that Biosynth Pipeline could produce exact designs for 93 of them, as well as pipelines to a chemically similar product for most of the remaining molecules. Furthermore, Biosynth Pipeline successfully proposes biosynthetic routes for complex therapeutic natural products (cryptofolione and basidalin) for which no known biosynthetic pathway currently exists.

INTRODUCTION

An increasing concern regarding climate change has led to growing public and private interest in synthetic biology to enable the biomanufacturing of a wide array of commodity chemicals[1–5]. In 2022, the United States' government signed an executive order to invest 2 billion dollars into the U.S. bioeconomy to divert chemical manufacturing away from pollutive fossil-fuel based processes and towards environmentally benign bioengineered ones[6]. Further, the overall demand for renewable products has also been solidifying, driven by the need to achieve sustainability in manufacturing while reducing carbon emissions to mitigate rising global temperatures [7,8].

Despite this increasing interest, the design of biosynthetic pathways to produce key commodity chemicals without known biosynthesis pathways remains a bottleneck[9–11]. Long and cumbersome literature searches are the status quo to piece together known or potentially promiscuous enzymatic reactions that may eventually lead to a target product from simple precursors and intermediates. Such manual pathway discovery methods diminish the rate at which biomanufacturing can replace the ~3500 or so high-volume, petroleum-derived chemicals currently in circulation[12,13].

In addition to this slow rate of access, the net chemical space that existing methods can reach is also limited by the customarily exclusive use of monofunctional enzymes. Monofunctional enzymes catalyze individual reactions only and, to date, comprise a majority of engineered pathways[4,14,15]. While such enzymes are valuable for the precise modification of functional groups on a substrate's carbon backbone, they offer few chances to increase the length of the backbone itself. Consequently, relying exclusively on monofunctional enzymes restricts the achievable carbon chain length of target products to about that of the original carbon source. Metabolic maps of bio-based chemicals in the literature highlight that pathways beginning from pyruvate and other central carbon metabolites yield C3-C8 products only[16,17]. Such short

backbones render it difficult for synthetic biology to compete with the longer hydrocarbons (>C10) present in petrochemical feedstocks, especially when making products more complicated than linear carbon chains[18].

In fact, attempting to synthesize molecules solely using monofunctional enzymes overlooks how nature synthesizes many valuable and complex natural products: by combining multifunctional and monofunctional enzymes for complementary purposes. Multifunctional enzymes (which catalyze multiple reactions) are used by these natural pathways, genetically encoded in the biosynthetic gene clusters (BGCs) of several microbes, plants, and fungi, to first construct elongated carbon scaffolds. Only after the construction of such scaffolds are monofunctional enzymes used for structural fine-tuning[19–22]. A particularly interesting and valuable class of multifunctional enzymes are polyketide synthases (PKSs), which can catalyze the expansion of carbon backbones[23–25]. PKSs are responsible for the synthesis of many medicinally relevant and structurally diverse molecules, from macrocyclic antibiotics (e.g., erythromycin[26]) to long aliphatic anticancer therapeutics (e.g., gephyronic acid[27,28]). PKSs broadly fall into one of three categories: type I, II, and III[23]. Type II and III PKSs consist of discrete, unlinked enzymes operating independently[29,30] while type I PKSs comprise several covalently linked enzymatic domains working together as an assembly-line to synthesize large polyketides from simple building blocks[31]. These building blocks are typically acyl-coenzyme A (acyl-CoA) derivatives, such as acetyl-CoA or malonyl-CoA[23,25]. Through decarboxylative Claisen condensation reactions performed by their ketosynthase (KS) domain, PKSs are able to iteratively add 2 to 3 carbon atoms from such acyl-CoA units onto a previously loaded acyl-CoA substrate[23,25]. Over the course of several modules, each of which minimally contains a KS, an acyltransferase (AT), and an acyl carrier protein (ACP) domain, PKSs can extend the initial substrate from just 2 to 6 carbons to over 20 in length[23,25]. This unique ability of PKSs to catalyze C-C bond formation

reactions is immensely beneficial for biomanufacturing and could be repurposed towards the synthesis of chemicals, such as biofuels, for which carbon chain length correlates positively with fuel efficiency[32].

Beyond constructing elongated backbones, PKSs allow for the incorporation of desired functional groups into the synthesized molecule[33–38]. This incorporation results from the diversity of starter and extender acyl-CoA units that can be acylated onto AT domains[38–40]. For instance, when designing monomers, terminal alkene groups (useful for polymerization) may be included through unsaturated units such as allylmalonyl-CoA[41,42]. Meanwhile, aromatic rings, a key structural component of many monomers, can be added via aromatic units such as cinnamoyl-CoA[43]. Moreover, since a PKS substrate is always tethered to its synthase, the likelihood of intermediate loss can be lower with PKSs than in engineered metabolic pathways wherein intermediates may be funneled into different pathways altogether while being shuttled around individual enzymes in the cell[44]. Such derailments can lead to unexpected byproducts. Due to the predictable deterministic logic of Type I PKSs, however, one can expect the final product to be directly correlated with the domains and acyl-CoA units involved [26].

This modular and predictable, yet tunable, nature of multifunctional PKSs makes them an attractive platform for biomanufacturing. While PKSs have historically been challenging to engineer and are characterized by low turnover rates[33,45], ongoing research has only further confirmed their capacity for structural derivatization: recently, the deoxyerythronolide synthase (DEBS) responsible for erythromycin's biosynthesis was engineered to accept fluorinated acyl-CoA units so as to regioselectively incorporate fluorine into its products[46]. This is crucial since an increasing number of approved drugs today comprise fluorine atoms[47,48]. Because PKSs follow a modular deterministic logic, an engineered PKS provides a straightforward path to

create a large library of chemical analogues that can be explored for improved properties via domain and module[49]. We believe that combining the versatility and scaffolding ability of PKSs with the precision of monofunctional enzymes, thereby mimicking the biosynthesis of natural products, can ultimately unlock a wider space of chemicals for biomanufacturing than would be accessible with either type of enzyme alone.

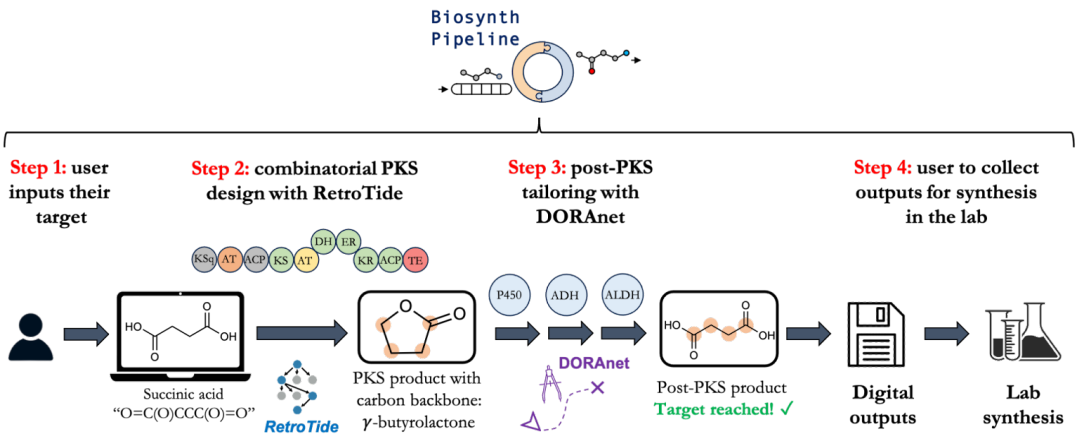


Fig. 1: Biosynth Pipeline reaches target molecules *in silico* by first constructing their carbon backbone and then performing structural refinements.

Biosynth Pipeline uses both multifunctional type I polyketide synthases (PKSs) and monofunctional enzymes for chemical biosynthesis. Our platform is built upon two key rule-based components – RetroTide and DORAnet. When a user inputs a target structure, RetroTide first suggests chimeric PKS designs to synthesize the carbon backbone of this target. DORAnet then performs enzymatic post-PKS decorations of the PKS product to reach the target chemical. If an integrated PKS and post-PKS pathway design is found, the user can collect this design from Biosynth Pipeline for experimental synthesis.

Here, we present Biosynth Pipeline, a novel and automated retrobiosynthesis software that seamlessly integrates the design of multifunctional chimeric type I PKSs with that of monofunctional enzymatic pathways into a single platform (Fig. 1). Given the vast design space that is already inherent to both chimeric PKSs and metabolic pathways comprising monofunctional enzymes, the combination of these unlocks a design space that is considerably larger than either route alone. To efficiently navigate this wider space, Biosynth Pipeline first uses RetroTide, a chimeric PKS tool presented here to construct the carbon scaffold of a target

molecule, and then DORAnet to subsequently tailor the PKS product towards the final target through monofunctional enzymes. While several *in silico* retrobiosynthesis tools already exist for monofunctional enzymes[10,11,50–58], to our knowledge, none exist for designing chimeric multifunctional enzymes, and especially not for the combination of the two. Existing tools[59,60] and databases[61–63] for multifunctional enzymes are valuable in matching known natural product structures to known biosynthetic gene clusters, but do not suggest novel PKS chimeras nor post-PKS modification pathways for a given target structure (that may or may not classify as a natural product). If an integrated pathway to a desired target exists, Biosynth Pipeline provides: (1) a chimeric PKS design to synthesize a PKS intermediate product as structurally similar to the target as possible, and (2) a monofunctional enzymatic pathway design to transform this PKS intermediate into the final target. Additional capabilities of utility to the synthetic biologist are included within Biosynth Pipeline for in-depth post-PKS pathway analyses. These include reaction feasibility prediction and thermodynamics calculations [64].

We have demonstrated the capabilities of Biosynth Pipeline to recommend biosynthetic pathways across three distinct scenarios: a structurally diverse set of 15 molecules, 155 targets of relevant biomanufacturing candidates, and two natural therapeutic products for which no biosynthetic pathway is currently known. First, we deployed our tool to suggest biosynthetic pathways to 15 structurally diverse molecules of varying carbon chain lengths (from 3 to 18) that we curated. In doing so, only a single post-PKS modification step was allowed and, for all 15 molecules, Biosynth Pipeline was able to construct the requisite carbon backbone. Subsequently we validated Biosynth Pipeline against a larger set of 155 molecules obtained from a previously published list of potential biomanufacturing candidates[13], allowing for up to 2 post-PKS modification steps this time. Against this set, Biosynth Pipeline was able to suggest pathways to completely synthesize 60% of molecules. As a final test on the general capabilities of our tool,

we prompted Biosynth Pipeline to synthesize the therapeutic natural products cryptofolione[65] and basidalin[66], for which no known biosynthetic pathways currently exist. Remarkably, Biosynth Pipeline succeeded in suggesting PKS designs and post-PKS pathways to synthesize these products completely.

Altogether, our work represents one of the first *in silico* demonstrations of combining the power of PKSs to expand carbon backbones with the capability of monofunctional enzymes to engineer a broad range of molecules - from simple commodities to complex natural products.

RESULTS

Architecture of Biosynth Pipeline.

Biosynth Pipeline is a retrobiosynthesis tool with two key rule-based components (Fig. 2): RetroTide (presented here) and DORAnet (publicly available at https://github.com/wsprague-nu/doranet/tree/main). When a target chemical is input, Biosynth Pipeline first attempts to synthesize this molecule and its carbon skeleton using only PKSs via RetroTide. PKS designs ranked by the chemical similarity of the corresponding PKS product to the target are then provided by RetroTide (see Methods for the different similarity metrics available). If a target can be synthesized by PKSs alone, Biosynth Pipeline terminates and users can search the ClusterCAD database[67] (https://clustercad.jbei.org/pks/) for natural PKS designs closely matched with the predicted chimera. This search will indicate the genetic edits, such as domain swaps, insertions, and/or truncations needed to produce the chimera from the closest matched natural PKS and the BGCs encoding for such a PKS. If the desired target cannot be produced by PKSs alone, however, then enzymatic post-PKS modifications via DORAnet are attempted on the PKS product that is most chemically similar to the target molecule.

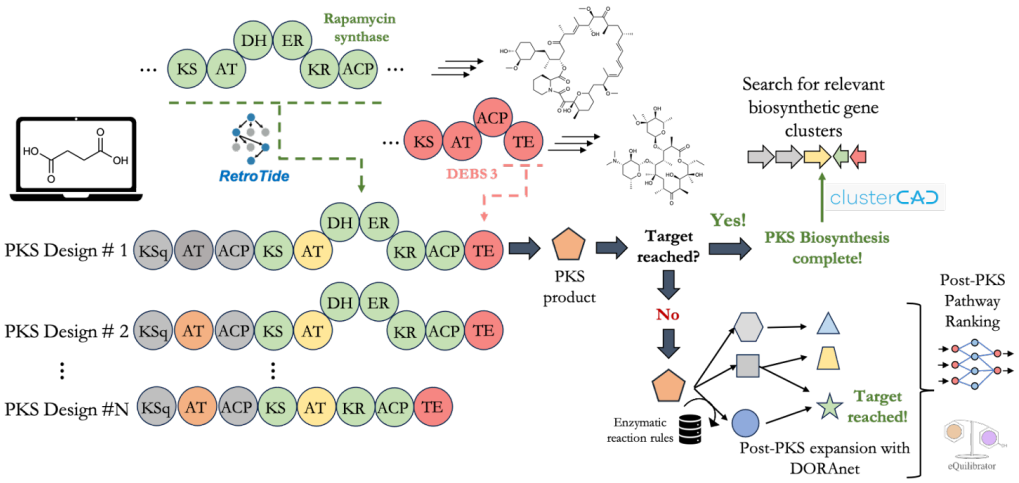


Fig. 2: Biosynth Pipeline combinatorially designs novel, chimeric PKSs from existing PKSs to produce intermediates structurally similar to the target molecule.

When a target molecule is input, Biosynth Pipeline first calls upon RetroTide to suggest chimeric PKS designs by combinatorially mixing domains from existing PKSs, such as rapamycin synthase or 6-deoxyerythronolide B synthase (DEBS). Here, each color represents a domain from a distinct known PKS. RetroTide's PKS designs are output in descending order of chemical similarity as measured between the corresponding PKS product and the target chemical. If the target can be synthesized solely with PKSs, Biosynth Pipeline stops and users can search the ClusterCAD database for closely matched PKSs. If the target cannot be synthesized by PKSs, then Biosynth Pipeline calls upon DORAnet to modify the top-ranked PKS product with monofunctional enzymes, whose activities are digitally encoded as reaction rules. The post-PKS pathways generated can then be ranked with thermodynamics and/ or pathway feasibility analyses. If the top-ranked PKS product does not lead to the final target through post-PKS modifications, then alternate PKS designs are considered.

DORAnet performs these post-PKS modifications by recursively expanding upon the PKS product from RetroTide using our previously published reaction rules that encode for the potential promiscuities of known enzymes[68]. Depending on the number of post-PKS modification steps specified, DORAnet may enumerate far more post-PKS pathways than can practically be manually analyzed. Thus, to elucidate the most feasible pathway chemistries for experimentation, our previously released DORA-XGB reaction feasibility classifier has been incorporated into Biosynth Pipeline to rank post-PKS reactions by their feasibility. DORA-XGB is publicly available at https://github.com/tyo-nu/DORA_XGB, and scores the feasibility of a given enzymatic reaction between 0 and 1, wherein a higher score represents a more feasible reaction. This score reflects both the thermodynamic feasibility of a reaction under typical physiological

conditions in the cell as well as the generalized likelihood that a designed functional group on a substrate would undergo catalysis.

In the event that post-PKS modifications of the top PKS product cannot yield the final target molecule, Biosynth Pipeline pursues two alternatives to further attempt to synthesize this product. Firstly, alternate, lower-ranked PKS designs and their corresponding PKS products from RetroTide are considered (Fig. 2). This is because a structurally dissimilar PKS intermediate may actually prove to be more amenable to the enzymatic modifications required to reach the final target molecule. The maximum number of alternate PKS designs considered is a hyperparameter that can be tuned by users, but is set to 5 by default to limit computational expense. The second alternative involves working backwards from the desired chemical target with DORAnet, to identify alternative precursors that can lead to the desired target. These intermediates can then be tested with RetroTide to see if they can be synthesized exactly by a PKS. This alternative approach can identify PKS precursors initially rejected by RetroTide, due to low overall chemical similarity to the final product. These contingencies and our primary algorithm driving Biosynth Pipeline ensure that the space of combined pathways has been adequately explored. The default configurations for Biosynth Pipeline produce results in just minutes on standard personal computers, but can be scaled to explore a larger chemical space on more powerful computer systems.

Selection of starter and extender acyl-CoA derivatives by RetroTide

Acyl-CoA derivatives are the key building blocks through which PKSs synthesize polyketide products (Fig. 3). All type I PKSs begin their assembly line synthesis of polyketides by first loading an acyl-CoA or acyl-ACP derivative via their loading module. In this work, we refer to this initial acyl-CoA unit as a starter unit while an acyl-CoA used by a downstream extension module is referred to as an extender unit. The loading module of a given PKS can often contain a ketosynthase-like (KSq) domain, which if present, loads a dicarboxylic acid wherein one of the acid groups has been acylated by CoA (e.g., malonyl-CoA). If this KSq domain is absent in the loading module, then PKSs load a monocarboxylic acid wherein the single acid group has already been CoA-acylated (e.g., propionyl-CoA or acetyl-CoA).

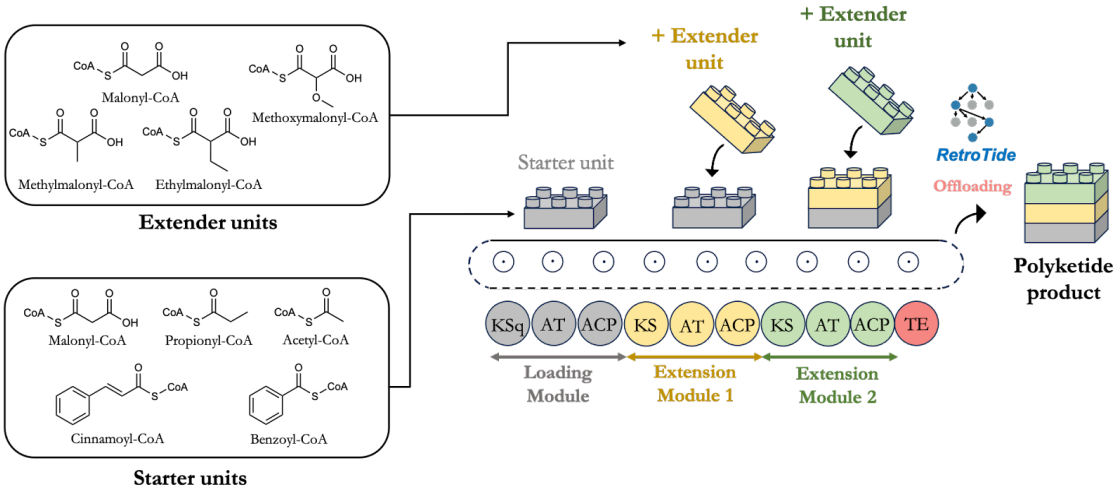


Fig. 3: RetroTide builds polyketide products *in silico* using a large and customizable set of starter and extender acyl-CoA derivatives.

PKSs use acyl-CoA derivatives as building blocks for synthesizing complex and functionally diverse polyketides. The loading module of PKSs can often contain a ketosynthase-like (KSq) domain, which if present, loads dicarboxylic acids wherein one of the acid groups has been acylated by CoA (e.g., malonyl-CoA). If this KSq domain is absent, however, then PKSs load monocarboxylic acids wherein the single acid group has already been acylated by CoA (e.g., propionyl-CoA or acetyl-CoA). Meanwhile, the ketosynthase (KS) domains in downstream extension modules use only CoA-acylated dicarboxylic acids as extension units. By considering a large and customizable set of acyl-CoA derivatives as starter and extender units within these constraints, RetroTide can build polyketide products that are as structurally similar to the final target product as possible.

225 In order to compile a list of starter units for RetroTide, starter units from canonical type I PKSs
226 documented within the ClusterCAD[67] database were first considered. This list was then expanded
227 further with more unusual starter units that have either previously been shown to be type I PKS
228 extenders that could be converted into starters via domain or module[69]. Certain starter units from
229 type II or type III PKSs were also included so as to expand the chemical space of molecules that
230 can be reached. The complete list of 30 starter units available within RetroTide and their
231 SMILES strings are reported in SI Table 1. We aimed to include an expansive set of rare and
232 hypothetical starters and extenders that can be paired down by users as appropriate, rather than
233 limiting the chemical space by only including the most common PKS substrates. Note that in
234 these starter and extender chemical structures, [S] is used as a placeholder for the substrates
235 attachment to a CoA or ACP. Many substrates may exist in both ACP and CoA forms, and are
236 included only once in RetroTide, but the resulting designs can be interpreted as suitable for either
237 substrate. RetroTide also does not simulate KSq decarboxylation, so to include designs which use
238 both carboxylated and decarboxylated starters, both variants must be included in the starter
239 database.

240 We emphasize that users also have the option to either add to or trim the list of starters and
241 extenders that Retrotide has access to depending on the experimental feasibility of working with
242 a given starter or extender in the laboratory. Constraining the list of starters and extenders can
243 enable Biosynth Pipeline to converge to a pathway design faster for a given target but may be
244 accompanied by the tradeoff of reaching a smaller chemical space.

245 **Showcasing Biosynth Pipeline through the biosynthesis of two simple molecules.**

246 We report here two use-cases that exemplify, with simple molecules, the two complementary
247 approaches used in Biosynth Pipeline (Fig. 6). The first involves the biosynthesis of

248 4-hydroxybutyric acid, a key intermediate in the downstream production of various C4 chemicals
249 and for which bioengineered pathways currently exist[70–72].

250 When 4-hydroxybutyric acid is input as a target into Biosynth Pipeline, the top PKS design from
251 RetroTide synthesizes 4-hydroxybutyric acid completely (Fig. 6A). This predicted chimera
252 features one loading module and one extension module. RetroTide suggests that the loading
253 module should load hydroxymalonyl-CoA as a starter unit to account for the terminal hydroxyl
254 group in 4-hydroxybutyric acid while the extension module should load malonyl-CoA as an
255 extender unit to complete the C4 backbone of 4-hydroxybutyric acid. Note that this may not be a
256 naturally occurring PKS starter, please see the section "Selection of starter and extender
257 acyl-CoA derivatives by RetroTide" for our rationale in including it anyways. The domain
258 architecture of this extension module features a fully reducing
259 dehydratase-enoylreductase-ketoreductase (DH-ER-KR) loop that reduces the acetyl group in
260 hydroxyacetyl-CoA to produce the fully reduced C4 backbone of 4-hydroxybutyric acid. Since
261 the target molecule is reached here with the top PKS design, Biosynth Pipeline terminates and
262 does not need to attempt any post-PKS modifications.

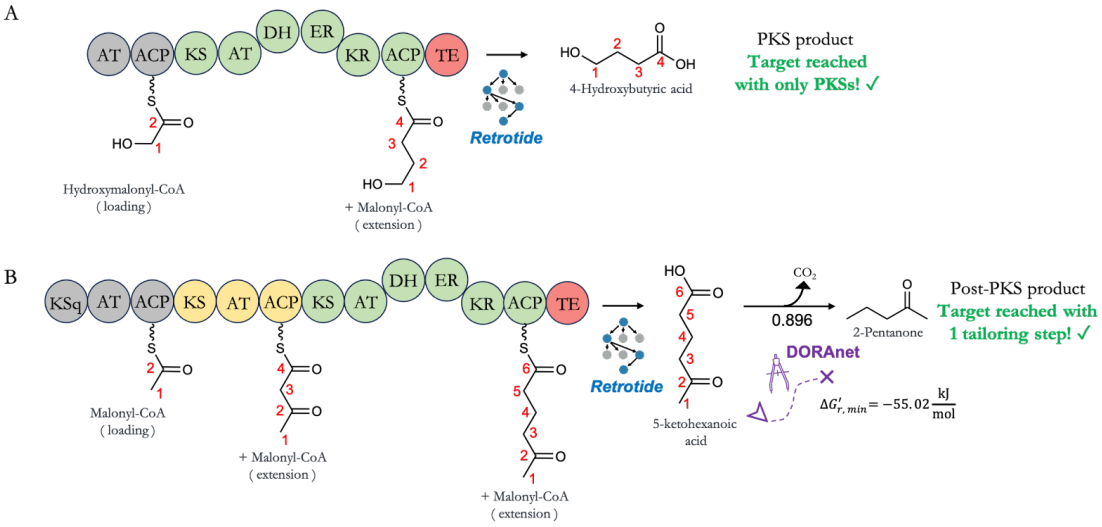


Fig. 4: Demonstrated biosynthesis of (A) 4-hydroxybutyric acid and (B) 2-pentanone with Biosynth Pipeline.

(A) When 4-hydroxybutyric acid is input as a target structure into Biosynth Pipeline, the top PKS design from RetroTide synthesizes 4-hydroxybutyric acid completely in just two modules. Hydroxymalonyl-CoA is recommended as a starter unit while malonyl-CoA is recommended as an extender unit. Since any additional post-PKS modifications are not required, Biosynth Pipeline terminates. **(B)** Conversely, when 2-pentanone is input as a target structure, the top PKS design from RetroTide reaches only as far as 5-ketohexanoic acid. Pickaxe subsequently takes over and suggests a decarboxylation reaction to synthesize 2-pentanone. This decarboxylation step is predicted to be highly feasible with an enzymatic reaction feasibility score of 0.896.

When 2-pentanone is input as a target into Biosynth Pipeline, however, the top-ranked PKS design only takes the user as far as 5-ketohexanoic acid, and needs completion through the use of DORAnet (Fig. 6B). The PKS design involves two extension modules in addition to the loading module, and malonyl-CoA is used throughout as both a starter and an extender unit. Since the final 2-pentanone target has not been reached, Biosynth Pipeline attempts to modify the 5-ketohexanoic acid PKS product by calling on DORAnet. A decarboxylation reaction is subsequently suggested with a high predicted feasibility score of 0.896 for transforming 5-ketohexanoic acid into 2-pentanone. Our predicted feasibility scores range from 0 to 1 with 0 being completely infeasible and 1 being completely feasible. Searching our generalized reaction rules for promiscuous enzymes that could potentially catalyze this reaction reveals that acetoacetate decarboxylase is a good candidate because it has previously shown to be sufficiently

274 promiscuous to decarboxylate levulinic acid to produce 2-butanone[73]. Levulinic acid shares a
275 similar structure to 5-ketohexanoic acid with just one less carbon atom present in its backbone
276 between the ketone and acid groups. Alternatively, another PKS design suggested by RetroTide
277 produces 3-ketohexanoic acid. In this PKS design, the fully reducing DH-ER-KR loop appears in
278 the first extension module while the second extension module features only the KS, AT, and ACP
279 domains. Given that methyl ketone synthases have previously demonstrated decarboxylation
280 activity on keto-acids[74,75], 3-ketohexanoic acid may also be a useful intermediate in the synthesis
281 of 2-pentanone.

282 In these two simple use cases, PKSs are shown to be valuable in synthesizing the carbon
283 backbones of user-specified targets from simple acyl-CoA derivatives while post-PKS
284 modifications are useful for structural refinements of the PKS product to reach the final target.
285 For more complex targets, users may wish to allow longer post-PKS pathways, though this may
286 sometimes be accompanied with the tradeoff of potentially more infeasible or thermodynamically
287 bottlenecked post-PKS reactions.

288 **Testing Biosynth Pipeline against a curated set of fifteen structurally diverse molecules.**
289 In order to ascertain the general utility of PKSs in synthesizing a variety of carbon backbones
290 and to investigate the utility of considering lower-ranked PKS designs, we prototyped Biosynth
291 Pipeline against 15 molecules that we curated for their structural diversity (Fig. 6). These consist
292 of 2 cyclic (caprolactone and γ-butyrolactone) and 13 aliphatic (tiglic acid, acrylic acid, levulinic
293 acid, adipic acid, heptane, aminobutyric acid, 1,4-butanediol, butanone, maleic acid,
294 2-hexenedinitrile, 2-butenenitrile, nonadecene, and sphingosine) molecules with various
295 functional groups. While several of these compounds have known biosynthetic pathways, we
296 selected them as a test to evaluate our systems ability to both create new PKS based pathways,
297 and reproduce known PKS based pathways. For each of these 15 molecules, 5 alternate PKS

298 designs were considered beyond the top-ranked PKS design. For testing we limited the number
299 of alternate PKS designs to 5 to limit the computational complexity of networks that may be
300 generated with each post-PKS product, however this is a user selectable parameter that can be
301 increased to potentially improve designs, while also increasing computational demands. Further,
302 only a single post-PKS modification step was used for two reasons. Firstly, to confirm that
303 chimeric PKSs designed by RetroTide can reach the desired number of carbon atoms in
304 commodity chemicals (from as little as 3 in acrylic acid to as many as 19 nonadecene from just
305 simple acyl-CoA building blocks, and without having to provide carbon atoms through
306 monofunctional enzymes. Secondly, from an experimental perspective, fewer post-PKS
307 modification steps minimize the likelihood of intermediate loss: while the enzymatic domains of
308 type I PKSs are covalently connected and the growing carbon chain substrate is always tethered
309 to a given module's acyl carrier protein (ACP) domain through the ACP's phosphopantetheinyl
310 arm[25], regular monofunctional enzymes do not enjoy such physical proximity in the cell[44,76]. In
311 fact, it is common for engineered metabolic pathways to become disrupted when their
312 intermediates are siphoned off into other pathways valuable for primary metabolism, ultimately
313 resulting in low final product titres. If successfully engineered, chimeric PKSs are less
314 susceptible to such derailments. Recently, the amino acid linkers that covalently bind the
315 catalytic domains of type I PKSs together have even been used to fuse together monofunctional
316 enzymes in engineered metabolic pathways, leading to more than double the product titre in
317 some cases[77]. Given such high substrate concentrations localized around individual catalytic
318 domains rather than globally dispersed around the cell, PKSs also have the potential to force
319 through reactions that may be thermodynamically uphill and would otherwise remain
320 uncatalyzed by regular, monofunctional enzymes.

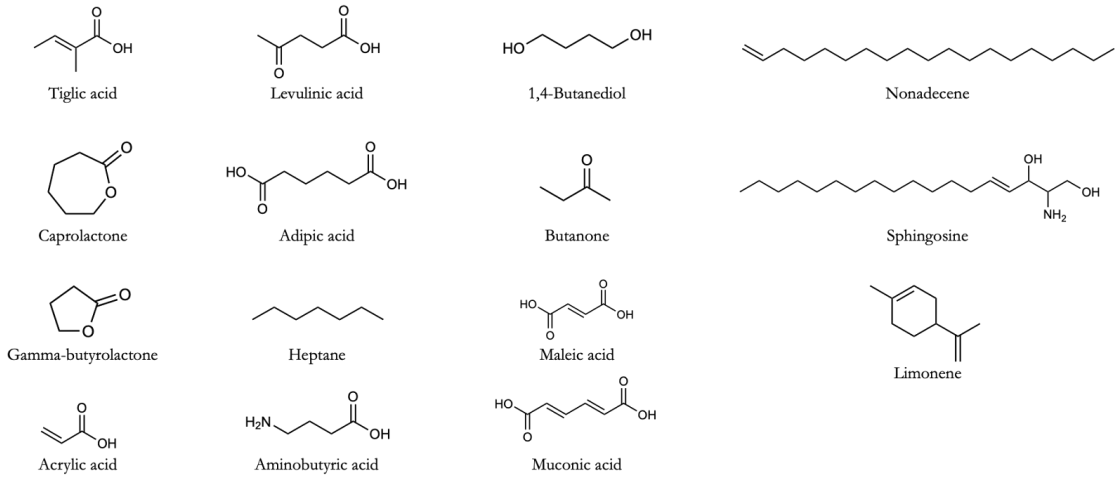


Fig. 5: A structurally diverse set of 15 molecules were curated for the prototyping of Biosynth Pipeline.

Fifteen molecules of varying carbon chain lengths and complexity were chosen as an initial prototyping set for Biosynth Pipeline to investigate if chimeric PKSs suggested by RetroTide would be able to construct the carbon backbones for each of these molecules. Both cyclic and aliphatic molecules are included in our set and amongst the aliphatic molecules, carbon backbones with as little as 3 (acrylic acid) and as many as 19 (nonadecene) carbon atoms are chosen to test the versatility of Biosynth Pipeline in constructing scaffolds of different lengths.

RetroTide was able to synthesize a carbon scaffold with at least as many carbon atoms as the final target for all 15 targets. Interestingly, for tiglic acid, caprolactone, and γ-butyrolactone, the top chimeric PKS designs were not only able to build the carbon backbones of these products but also accomplish the structural decorations necessary to synthesize all three targets completely (Fig. 5, SI Fig. 1). For the remaining 10 molecules, since PKSs did not synthesize the final target, a single-step post-PKS modification reaction was used with DORAnet. Amongst these 10 molecules, for acrylic acid, levulinic acid, adipic acid, 1,4-butanediol, and maleic acid, the top-PKS product serves as an excellent precursor for post-PKS modifications and these targets can be synthesized completely in a single post-PKS tailoring step (Fig. 5). For butanone and heptane, however, the top PKS product does not lead to the target molecule after a single post-PKS modification step. Surprisingly, in this case a lower-ranked PKS design (the rank 4 PKS design) leading to a more structurally dissimilar intermediate actually serves as a better launchpad for the post-PKS modifications required to transform the corresponding PKS products

into the final targets (Fig. 5). Here, we calculate chemical similarity by first computing the maximum common substructure (MCS) between a given (either PKS or post-PKS) product and the target chemical and then calculating the ratio of atoms in the MCS to the total number of unique atoms between both molecules, e.g. the MCS Tanimoto similarity. Finally, for limonene, aminobutyric acid, and sphingosine, Biosynth Pipeline was not able to reach these targets across all 6 PKS designs considered with a single monofunctional enzyme step.

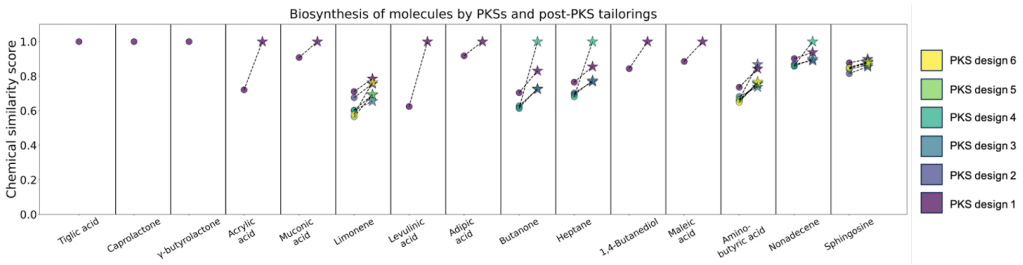


Fig. 6: Biosynth Pipeline successfully constructs the carbon backbones for all 15 curated molecules and synthesizes 11 molecules completely.

For 3 molecules out of the total 15 in our curated prototyping set (i.e. tiglic acid, caprolactone, γ-butyrolactone), Biosynth Pipeline synthesizes these molecules completely with the top-ranked type I polyketide synthase (PKS) design and does not require any further post-PKS modifications. For 8 molecules (acrylic acid, muconic acid, levulinic acid, adipic acid, butanone, heptane, 1,4-butanediol, and maleic acid), Biosynth Pipeline synthesizes these molecules completely using a single post-PKS modification step. Out of these 8 molecules, acrylic acid, muconic acid, levulinic acid, adipic acid, 1,4-butanediol, and maleic acid can be reached by a single-step modification of the top-ranked PKS design's product while butanone, heptane, and nonadecene can only be reached by starting with a PKS product that is more structurally dissimilar with respect to the target molecule than the top-ranked PKS product. There exist 3 molecules out of the 15 total (limonene, aminobutyric acid, and sphingosine) that cannot be synthesized at all by Biosynth Pipeline within just one post-PKS step even after considering five alternate PKS designs.

Validating Biosynth Pipeline against a large set of valuable biomanufacturing targets

In order to validate Biosynth Pipeline against a larger, unbiased collection of realistic biomanufacturing targets, and investigate the effect on target diversity of allowing extra post-PKS steps, we turned to a previously published set of 209 chemicals to be potentially produced via biomanufacturing[36]. Of these 209 chemicals, we filtered out C1 metabolites, cofactors, charged targets as well as any chlorinated or sulfonated species so as to yield a final

validation set of 155 molecules. Cofactors were removed as target molecules because these would be consumed by DORAnet and C1 metabolites were also removed because their carbon backbone is too short for a PKS. Our reaction rules do not work on charged ions so these were removed as well. Of these 155 targets, 94 were non-aromatics while 61 were aromatics. We then prompted Biosynth Pipeline to synthesize all 155 chemicals with up to 2 post-PKS modification steps. For the 94 non-aromatic targets, we constrained Biosynth Pipeline to only using the aliphatic units of malonyl-CoA, methylmalonyl-CoA, hydroxyacetyl-CoA, methoxymalonyl-CoA, and allylmalonyl-CoA as starters and extenders, while for the 61 aromatic targets we allowed for all starters and extenders to be used. Note that not all of these substrates are naturally occurring in the positions used (e.g. as starters or extenders), see the section "Selection of starter and extender acyl-CoA derivatives by RetroTide" for our rationale on including them. In attempting to synthesize these molecules, stereochemical information was not considered and 5 alternate PKS designs beyond the top-ranked PKS design were considered.

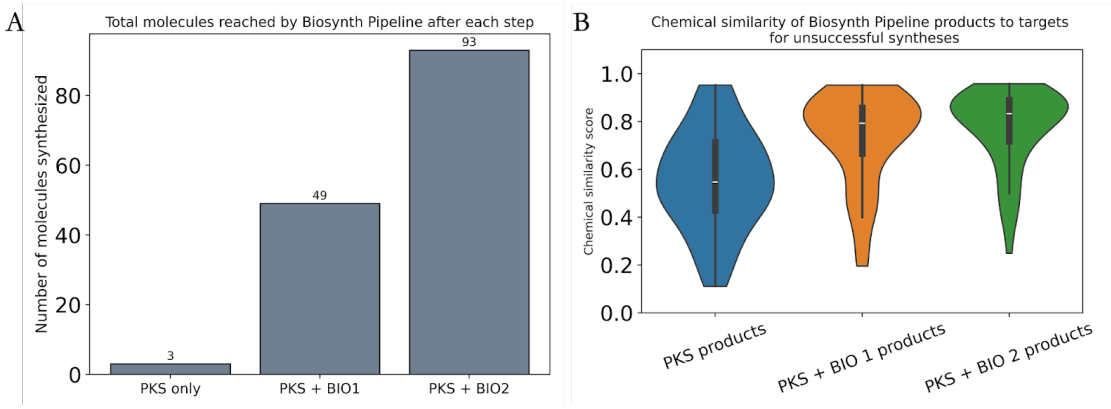


Fig. 7: Biosynth Pipeline successfully suggests integrated biosynthetic pathways to 93 biomanufacturing candidates.

a When Biosynth Pipeline is prompted to synthesize the validation set of 155 molecules that would be potential candidates for biomanufacturing, it achieves an overall hit-rate of 60% (i.e. 93/ 155). Of the 93 molecules that can be synthesized exactly, 3 are synthesized solely by PKSs while 46 more molecules can be synthesized with just one post-PKS tailoring step from the PKS product. After a second post-PKS step, 44 more molecules can be synthesized, bringing the total number of molecules synthesized by Biosynth Pipeline to 93. **b** Although the remaining 62

molecules out of 155 cannot be synthesized by Biosynth Pipeline, post-PKS modification steps are shown to improve the chemical similarity between the PKS product and the post-PKS product with respect to the target for these 62 target molecules that could not be synthesized.

Biosynth Pipeline successfully proposed designs to synthesize 93 molecules out of our total of 155, translating to a success rate of 60% (Fig 6). Of these 93 molecules, chimeric PKS-only designs from RetroTide could synthesize only 3 molecules exactly (Fig. 6A). Still, the carbon scaffolds of PKS products proved to be very valuable precursors for post-PKS tailoring. Within a single generation, DORAnet was able to synthesize 46 more molecules completely by building off of RetroTide's carbon backbones. With a second generation, DORAnet was able to synthesize 44 more molecules (Fig. 7A), thereby reaching 93 molecules in total. For the remaining 62 molecules that could not be synthesized by Biosynth Pipeline, all top-ranked PKS products still comprised at least as many carbon atoms as their corresponding final targets. Moreover, for these 78 molecules, post-PKS modifications performed on the top-PKS product are still shown to improve the structural similarity of the post-PKS product over the PKS product when both are compared to their corresponding target molecule (Fig. 6B).

We explored the possibility of attempting more than 2 post-PKS modifications on the PKS products involved in the synthesis of these 78 molecules but ultimately decided against this for two reasons. First, reaction networks grow exponentially in size with each added generation. As such, incorporating an additional post-PKS step would significantly increase computational expense. Moreover, the increase in chemical similarity with respect to the target is only marginal between the first and second post-PKS reaction as opposed to that between the PKS product and the first post-PKS modification step. This would suggest that returns in increases in chemical similarity diminish with added steps anyway.

Biosynth pipeline suggests novel pathways to molecules for which no known pathways exist

As a final stress-test on the robustness of our software, we prompted Biosynth Pipeline to suggest pathways to natural products for which no known pathways exist. We first attempted this for the antifungal cryptofolione isolated from the plant *Cryptocarya latifolia*[37]. The biosynthetic origin of cryptofolione has not been elucidated and only total synthesis pathways relying on synthetic chemistry have thus far been utilized to synthesize cryptofolione[44]. Given its therapeutic potential, cryptofolione represents a promising target for biomanufacturing with our 'PKS-first' approach presented here. When the structure of cryptofolione is input into Biosynth Pipeline without any stereochemical information, RetroTide suggested a PKS design consisting of 6 modules that could synthesize cryptofolione exactly. This predicted chimera loads an aromatic starter unit, namely cinnamoyl-CoA to account for the aromatic moiety on one end of cryptofolione's backbone. While not a known Type I PKS starter, cinnamoyl-CoA is a known type III PKS starter, and it may be possible to engineer a Type I PKS that can accept it[30]. The extension modules of this chimera then use only malonyl-CoA units to build the remainder of cryptofolione's backbone. A TE domain that catalyzes an intramolecular cyclization reaction to form the lactone ring at the other end of cryptofolione's backbone is finally used to complete cryptofolione's proposed biosynthesis. Since Biosynth Pipeline suggests a PKS design that can synthesize cryptofolione exactly, no post-PKS modification steps are needed. We note that in our proposed synthesis of cryptofolione, the alkene group adjacent to the acyl group prior to offloading is in the trans configuration while after lactonization, it changes to the cis configuration. We are uncertain if thioesterase domains that perform cyclizations would be able to invert double bond stereochemistry as, to our knowledge, no such examples have yet been observed in nature. While in this work, we focused on harnessing PKSs and enzymatic reactions to obtain the desired two-dimensional target structure, future work will incorporate more stereochemistry corrections such that even the desired three-dimensional structure is produced.

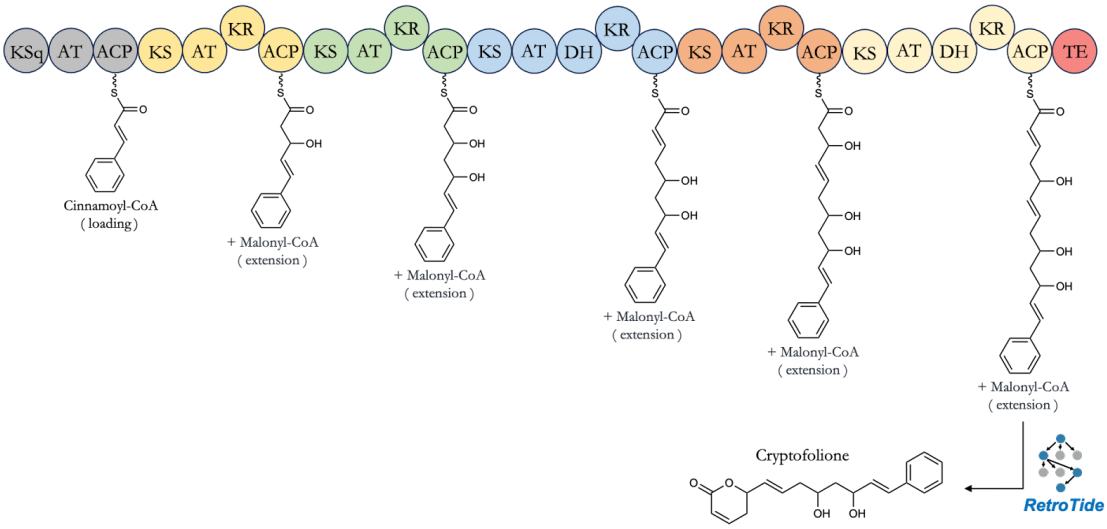


Fig. 8: Proposed biosynthesis of cryptofolione by Biosynth Pipeline.

When prompted to suggest a biosynthesis route to the natural product cryptofolione, a potential therapeutic whose biosynthetic origin remains unknown, Biosynth Pipeline calls upon RetroTide to successfully suggest a 6-module chimeric PKS that could synthesize cryptofolione exactly. This chimeric PKS selects cinnamoyl-CoA as a starter unit given the aromatic ring on one end of cryptofolione's carbon backbone and subsequently only uses malonyl-CoA as extender units. The offloading reaction for the termination domain of this chimeric PKS is set to an intramolecular cyclization reaction so that the lactone ring on the other end of cryptofolione's carbon backbone can be synthesized. Our PKS design lends insight into the possible biosynthetic origins of α-pyrones such as cryptofolione.

We then prompted Biosynth Pipeline to suggest a synthesis for the natural product basidalin. Basidalin is a smaller but potent antibiotic first extracted in 1983 from the fungus *Leucoagaricus naucina*[38]. Despite its early extraction, the first total chemical synthesis of basidalin was only achieved recently in 2016[45]. As is the case with cryptofolione, the biosynthetic origins of basidalin have yet to be uncovered. Thus, we prompted Biosynth Pipeline to synthesize basidalin and the suggested PKS design was able to synthesize basidalin's carbon backbone exactly. This predicted chimera features 3 modules in total. The loading module and first extension module both utilize hydroxymalonyl-CoA as starter and extender units respectively while the final extension module utilizes malonyl-CoA as an extender unit. Similar to the predicted chimera for the earlier proposed biosynthesis of cryptofolione, a TE domain that catalyzes an intramolecular cyclization reaction is used for the proposed biosynthesis of basidalin. Subsequently, two

post-PKS modification steps are used to synthesize basidalin exactly. The first suggested post-PKS reaction features a peroxidase while the second features an aminotransferase. The ability of Biosynth Pipeline to suggest pathways to such complex natural products could also enable users to utilize Biosynth Pipeline to annotate orphan biosynthetic gene clusters.

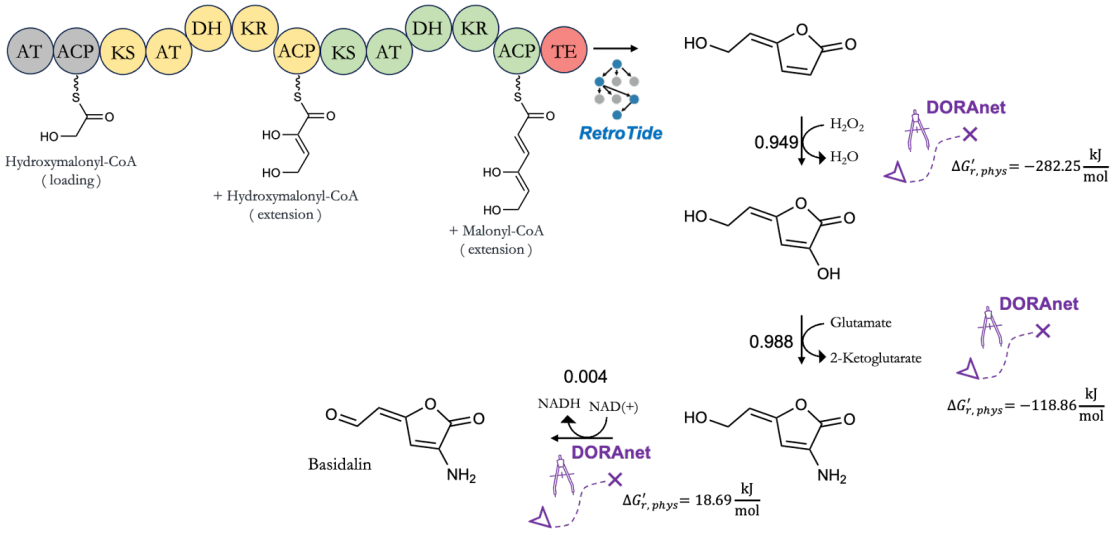


Fig. 9: Proposed biosynthesis of Basidalin by Biosynth Pipeline.

When prompted to suggest a biosynthesis route to basidalin, another therapeutic natural product whose biosynthetic origin also remains unknown, Biosynth Pipeline suggests a chimeric PKS design and two post-PKS modification steps to synthesize basidalin exactly. This chimeric PKS consists of 3 modules. The loading module and first extension module both use hydroxymalonyl-CoA as starter and extender units respectively while the final extension module uses malonyl-CoA as an extender unit. Again, an intramolecular cyclization reaction is chosen for offloading and we note that RetroTide is able to synthesize the γ-lactone backbone of basidalin exactly. Subsequently, three post-PKS modification steps involving peroxygenase and aminotransferase enzymes can produce the target basidalin product.

Discussion

Here, we present retrobiosynthesis software that harnesses both multifunctional and monofunctional enzymes to propose the *de novo* biosynthesis of a broad spectrum of chemical targets - from simple commodity chemicals to more complex natural products for which biosynthetic pathways may or may not already exist. Crucially, we have also put forth a valuable

426 paradigm for the synthesis of such molecules - that multifunctional enzymes, such as type I
427 PKSs, can be exploited for creating the carbon scaffolds of target products while regular,
428 monofunctional enzymes within biology can subsequently be brought in for structural
429 refinements that eventually transform such scaffolds into the target product exactly.

430 Our motivation for developing this 'PKS-first' algorithm to drive Biosynth Pipeline lies
431 in expanding biologists' capabilities to explore chemical space by complimenting the strengths of
432 PKSs with that of monofunctional enzymes. PKSs excel at synthesizing elongated carbon
433 backbones and/ or incorporating unique functional handles into target molecules from a common
434 pool of simple, acyl-CoA building blocks. Monofunctional enzymes, by contrast, rarely catalyze
435 such iterative C-C bond formation reactions but are adept at precise, regioselective functional
436 group transformations. Indeed, there are several reaction chemistries, such as glycosylations,
437 monooxygenations, and transaminations that PKSs are simply unable to perform and must
438 instead rely on monofunctional enzymes. Rather than attempting to manufacture chemicals using
439 only one of these routes, we propose that a wider space of molecules can be accessed by
440 integrating the two.

441 In this study, we demonstrated the effectiveness of our approach and of Biosynth Pipeline
442 across various test-cases of increasing complexity. First, we showcased Biosynth Pipeline's two
443 modes of operation by synthesizing 4-hydroxybutyric acid completely through PKSs, while
444 2-pentanone was reached with just one post-PKS modification step. Although metabolic
445 pathways have already been engineered for the biosynthesis of both chemicals, our proposed
446 pathways offer a compelling alternative that may minimize intermediate loss, a common
447 occurrence with classic metabolic pathways. Subsequently, we further prototyped Biosynth
448 Pipeline for the synthesis of 15 structurally diverse chemicals and proved that, for certain targets,

449 considering PKS designs beyond the optimal provides better precursors for post-PKS
450 modifications to synthesize the final target. After these initial tests, we turned to a published set
451 of molecular candidates for biomanufacturing and proved that we could synthesize 60% of them
452 (93 out of 155) with our pipeline while allowing for up to two post-PKS modification steps.
453 Biosynth Pipeline can access an even larger chemical space by considering further post-PKS
454 steps, however longer pathways will, of course, be more challenging to successfully build, so we
455 decided to limit our analysis to two to limit pathway designs to a realistic complexity. As a final
456 test of Biosynth Pipeline, we prompted it to suggest combined PKS and post-PKS pathways to
457 complex natural therapeutics for which biosynthetic gene clusters have not been elucidated. In
458 doing so, Biosynth Pipeline was able to suggest novel pathways to the natural products
459 cryptofolione (antifungal) and basidalin (antibiotic) for which no known biosynthetic pathways
460 exist.

461 One limitation of our tool is the lack of support for stereochemistry corrections. While the
462 enantioselective synthesis of compounds is a key strength of PKSs[78,79] and many monofunctional
463 enzymes[80–82], a number of retrobiosynthesis tools in the literature cannot reliably predict the
464 correct stereoisomer of both existing and novel reactions. Nonetheless, with the recent
465 development of cheminformatics tools capable of both extracting and applying
466 stereochemistry-aware reaction templates[83], we expect our future work to address such
467 corrections as well.

468 Overall, while the successful engineering and construction of chimeric PKSs remains an
469 ongoing challenge, our *in silico* pipeline highlights the potential of the field, given the diversity
470 of products that can be synthesized by integrating PKSs with post-PKS pathways. In addition to
471 suggesting pathway chemistries to target molecules using both PKSs and monofunctional

472 enzymes, Biosynth Pipeline also enables users to rank results through reaction thermodynamics
473 and feasibility calculations to observe if potentially promiscuous, wild-type enzymes would
474 catalyze predicted reactions or if such enzymes would need to be mutated further. With all of
475 these functionalities built in, Biosynth Pipeline serves as a flexible and comprehensive tool for
476 users looking to synthesize molecules via both PKSs and monofunctional enzymes. For users
477 who may be interested, we have also provided several tutorials on our Github repository that
478 show more examples of using Biosynth Pipeline in various scenarios.

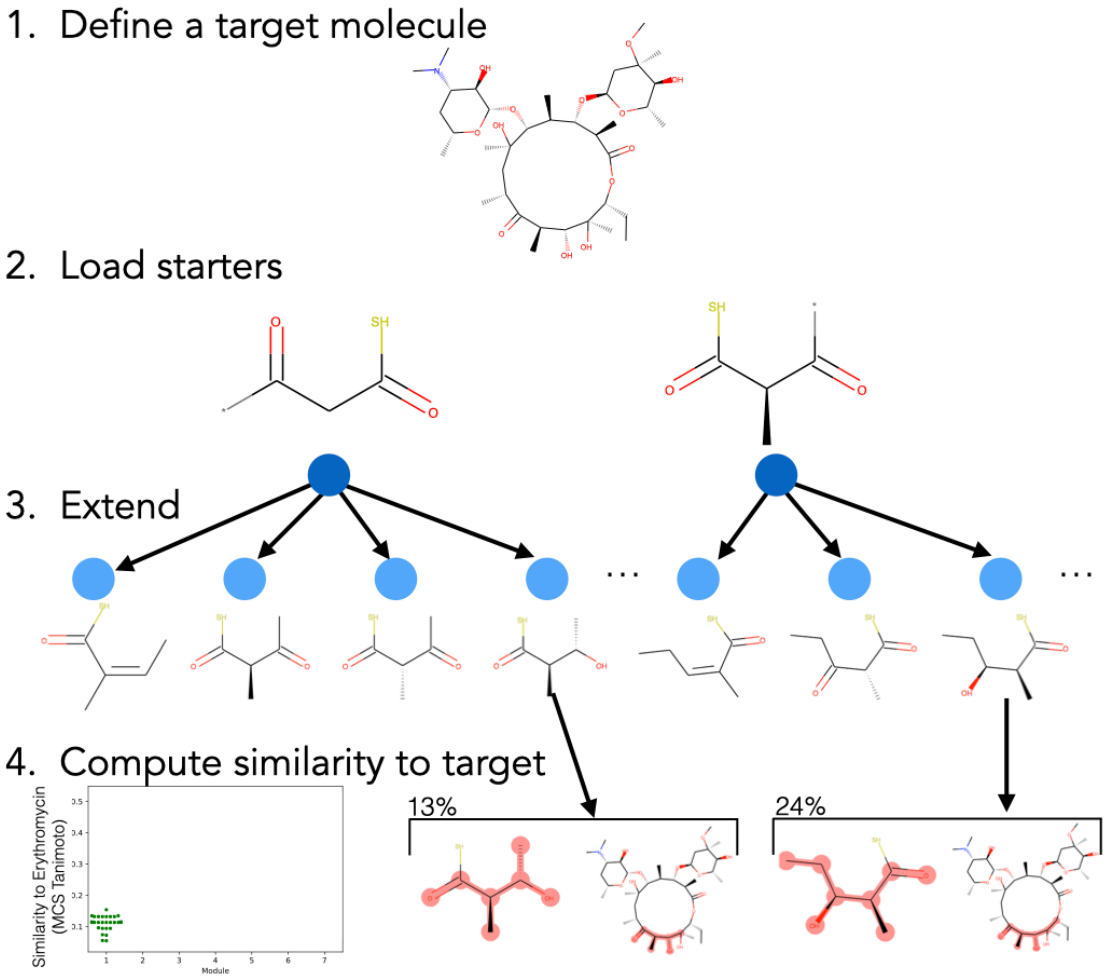
479 METHODS

480 **User-defined inputs for Biosynth Pipeline.**

481 Users can initialize a Biosynth Pipeline object in the Python programming language with the
482 simplified molecular input linear entry system (SMILES) string of their target chemical (Fig. 1).
483 Users can also specify if this target chemical should be synthesized using either only PKSs, or
484 only monofunctional enzymes, or both. In order to generate PKS designs, users can either allow
485 for all 11 starter and 31 extender acyl-CoA (SI Table 1) units available within Biosynth Pipeline
486 to be used as building blocks or choose a specific subset of these units. Biosynth Pipeline
487 supports both an intramolecular cyclization as well as a thiolysis reaction to release the bound
488 PKS product from the designed PKS's thioesterase (TE) domain and users can select from either
489 of these offloading reactions to generate their PKS product. When using combined PKS and
490 non-PKS pathways, users must specify the number of post-PKS modification steps to perform on
491 the PKS product. Currently, Biosynth Pipeline does not support any stereochemistry corrections.
492 In this work, we optimized for obtaining the correct two-dimensional structure of the target
493 molecule. Consequently, all molecules presented here are shown without their stereochemical
494 configurations.

495 **Computational design of chimeric type I PKSs with RetroTide**

Fig. 9: RetroTide algorithm for the design of PKSs from a target chemical structure.



The RetroTide algorithm has the following steps: **1)** The user provides a desired target chemical structure. **2)** RetroTide loads all possible PKS starter module final structures. **3)** The cartesian product between all possible starters, and all possible extension modules is used to compute all possible 1 module PKS products. These extension module structures are pre-computed once and reused as RetroTide runs. **4)** All structures are compared with the user defined similarity metric to the defined target, and only the N (a user selectable parameter) top scoring designs are retained. **5)** Steps 2-4 are repeated iteratively adding all possible single PKS modules to the end of each of the best scoring designs from the previous round, until scores stop increasing for each round and then the algorithm halts, providing the best scoring designs to the user. **6)** The best scoring designs are optionally 'released' as lactone/lactam, or carboxylic acid products.

Chimeric type I PKS designs are created using RetroTide, the first key component within Biosynth Pipeline. Upon providing the simplified molecular input linear entry system (SMILES)

499 string of a target molecule, RetroTide first attempts to design PKSs that can synthesize this
500 molecule using various modules, domain architectures, starter units, and extender units. Users
501 can customize which starter and extender units they would like to use in building chimeric PKSs
502 from a provided list of 11 starter and 31 extender units, or optionally expand the list themselves.
503 The complete list of starter and extender units available can be found in the SI. RetroTide builds
504 PKS designs by recursively adding modules, thereby synthesizing in the forward direction (Fig.
505 9). Within each round of adding a module, the PKS product is computed and only the top N
506 designs leading to the most chemically similar products progress onto the next recursive call.
507 Here, N is a hyperparameter that can be set by the user and has a default value of 15. The specific
508 metric of chemical similarity used by RetroTide to compute the structural similarity between the
509 current PKS product and a target product is also an option that can be selected by the user. Users
510 can choose between either computing the (1) Atom-Atom-Path similarity[47], the (2) Atom pair
511 Tanimoto similarity[48], or the (4) Tanimoto similarity using the maximum common substructure
512 between the current PKS product and the target product either with or without chirality[49]. After
513 RetroTide has completely finished designing chimeric PKSs, these designs are returned in
514 descending order of chemical similarity - as measured between each design's corresponding
515 product and the target product. Users can then decide between terminating the growing carbon
516 chain on the simulated chimeric PKS via either an intramolecular cyclization reaction, such as
517 that involved in the formation of 6-deoxyerythronolide B[25], or a thioesterase reaction, thereby
518 forming a terminal carboxylic acid group. The highest-ranked PKS design from RetroTide either
519 may or may not have produced the user's target product, depending on whether this product is
520 reachable by purely PKS chemistry.

521 With the PKS designs obtained from RetroTide, users can search the ClusterCAD PKS
522 database[39] (freely available online at https://clustercad.jbei.org/pks/) for naturally-occurring

523 designs that match closely with the suggested chimeric designs, so as to select an initial design
524 for biological engineering.

525 **Performing post-PKS modifications on RetroTide products with DORAnet**

526 Post-PKS modifications are performed using our DORAnet platform available at
527 https://github.com/wsprague-nu/doranet (previously Pickaxe V2.0[50]). If the user-defined target
528 product was not reached by RetroTide, the PKS product is funneled as a starting molecule into
529 DORAnet. DORAnet is a retrobiosynthesis tool that generates metabolic *in-silico* network
530 expansions (MINEs) by recursively expanding upon a precursor metabolite using reaction rules.
531 Users can decide between two sets of reaction rules in generating MINEs. The first set of rules to
532 choose from is our generalized and maximally promiscuous JN1224MIN rule set[51], which
533 predicts promiscuous reactions by only considering the substrate chemical moieties that can
534 directly undergo enzymatic reactions. The second, is our intermediate JN3604IMT rule set
535 (available at https://github.com/josephniwjc/enzymatic_reaction_rules), which not only considers
536 reacting moieties in predicting reactions but also the surrounding chemical neighborhoods within
537 which these moieties are present. In performing post-PKS modifications, users must specify the
538 number of post-PKS steps, i.e., the number of generations for which DORAnet would be run.
539 After post-PKS modifications, if the target product is reached, pathways between the PKS
540 product and the final, downstream target are returned. If the target product is not reached,
541 however, then pathways to the most chemically similar non-PKS product with respect to the final
542 target are returned. When running longer post-PKS modification pathways, users also have the
543 option to choose from various in-built DORAnet compound filters (such as the molecular weight
544 filter or the chemical similarity filter), so as to ensure that MINEs remain computationally
545 tractable[50].

Once post-PKS pathways have been elucidated, users can use the specific reaction rules that were involved in generating predicted reactions to search for UniProt identification numbers of potentially promiscuous enzymes present within either JN1224MIN or JN3604IMT, that may be able to catalyze predicted reactions.

Machine-learning guided selection of post-PKS enzymatic pathways

Depending on the number of post-PKS modification steps performed, the number of post-PKS pathways found between the PKS product and the final target may be too large to conceivably manually analyze in order to elucidate the most feasible pathway chemistries for experimentation. As such, our previously released enzymatic reaction feasibility classifier (available at https://github.com/tyo-nu/DORA_XGB) has been bundled into Biosynth Pipeline to further rank post-PKS pathways in terms of their predicted feasibility scores (Fig. 2). Feasibility scores are computed for each reaction in a pathway by considering the SMILES strings of all participating species, i.e., substrates, products, and cofactors. Our reaction feasibility model is based on the gradient boosted trees architecture and was trained using reported thermodynamically feasible reactions as positive data and thermodynamically infeasible as well as synthetically generated infeasible reactions as negative data. This allows the computed feasibility score to reflect both the thermodynamic feasibility of a reaction as well as the likelihood of an enzyme transforming a certain moiety at a given reaction site. After computing the feasibility score for all reactions in a pathway, a net feasibility score for the entire pathway is computed by taking the product of all constituent reactions' feasibility scores. Instead of using the feasibility score, which is a continuous number between 0 and 1, users can also opt for a binary label on whether a predicted reaction would be feasible or not using the built-in feasibility thresholds for the feasibility classifier.

Thermodynamic calculations of post-PKS enzymatic pathways

570 In order to further down-select predicted pathways, Biosynth Pipeline also enables users to
571 calculate the Gibbs free energies of reactions as well as the maximum/minimum driving force
572 (MDF) of multi-step pathways (Fig. 2)[52]. The open-source software eQuilibrator 3.0 was used to
573 calculate the change in Gibbs free energy due to a reaction, $\Delta_r G'$[34]. eQuilibrator 3.0 uses the
574 component contribution method for first estimating the standard Gibbs free energy change due to
575 a reaction, $\Delta_r G^\circ$, by assuming reactant concentrations of 1 M. These $\Delta_r G^\circ$ values are then
576 transformed to our prespecified conditions of temperature 298 K, ionic strength 0.25 M, pH 7.4,
577 and pMg 3.0 to calculate $\Delta_r G'$ under common cellular conditions. Users can also adjust these
578 thermodynamic parameters in their calculations as desired. In order to ensure that Biosynth
579 Pipeline is accessible to all, we did not incorporate eQuilibrator-assets into our software and only
580 incorporated the general eQuilibrator API. Thus, if there exists a new compound not present
581 within the eQuilibrator compound database, it cannot be decomposed into its groups and added to
582 the database linked below since doing so would require a ChemAxon license to calculate the pKa
583 values of new compounds. Consequently, the change in Gibbs free energy due to a reaction, $\Delta_r G'$
584 can only be computed for predicted DORAnet reactions in which all participating compounds are
585 present in the eQuilibrator compound database, which in turn can be obtained as a SQLite
586 database through Zenodo (zenodo.org/records/4128543).

587 Provided that the change in Gibbs free energy due to a reaction, $\Delta_r G'$ can be computed for all
588 reactions along a multi-step pathway, users can subsequently calculate the MDF of a post-PKS
589 pathway. This calculation seeks a set of metabolite concentrations that minimizes the $\Delta_r G'$ value
590 of the most thermodynamically uphill or bottlenecked reaction within a multi-step pathway [52].

Mining of potential bio-based chemicals set from a previously published set

In order to collect a larger set of chemicals upon which we could validate Biosynth Pipeline, we extracted 198 chemicals from a list of 209 chemicals that could hypothetically be produced by biomanufacturing as provided by Wenzhao Wu and colleagues through genome scale models of *Saccharomyces cerevisiae* and *Escherichia coli*[36]. Not all 198 chemicals extracted have existing engineered metabolic pathways for their production. Of these 198 chemicals, we filtered out C1 metabolites, common cofactors, and any sulfonated or chlorinated species, leaving us with 155 molecules in total. With these remaining targets, we queried their KEGG identifiers against the KEGG database using their application programming interface (API) and were able to receive SMILES strings for all 155 molecules. For these 155 molecules, post-PKS modifications of up to two steps were used and pathways were identified by comparing the SMILES strings of post-PKS pathways exactly with that of targets.

Code Availability

Biosynth pipeline and RetroTide are provided as free open source software at https://github.com/JBEI/BiosynthPipeline and https://github.com/JBEI/RetroTide respectively.

Competing Interests

J.D.K. has financial interests in Amyris, Ansa Biotechnologies, Apertor Pharma, Berkeley Yeast, Cyklos Materials, Demetrix, Lygos, Napigen, ResVita Bio and Zero Acre Farms. The other authors declare no competing interests.

Acknowledgements

This work was funded in part by the DOE Joint BioEnergy Institute, supported by the U. S. Department of Energy Office of Science's Biological and Environmental Research (BER)

program, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. Funding for Yash Chainani for this study was partly provided by the Northwestern University Graduate School Cluster in Biotechnology, Systems, and Synthetic Biology, which is affiliated with the Biotechnology training program. The funding for Margaret Guilarte-Silva was partly provided by the Northwestern University DeFeo Family Undergraduate Research Fellowship in Synthetic Biology and partly by the Northwestern University Summer Undergraduate Research Award. The authors would like to acknowledge Matthew Incha and Patrick Kinnuen for their reading of the manuscript and corresponding feedback. This research project was supported in part through the computational resources and staff contributions provided by the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office of Research, and Northwestern University Information Technology. This research also used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award ERCAP0028489.

Author Contributions

Yash Chainani: Methodology and Software, Data curation, Writing – original draft and subsequent edits. Visualization – generated the figures in this manuscript. **Jacob Diaz:** Methodology and Software, **Margaret Guilarte-Silva:** Methodology and Software, **Keith E. J. Tyo:** Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript. **Linda J. Broadbelt:** Conceptualization – conceptualized the approach of this study. Investigation. Formal analysis. Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript. **Aindrila Mukhopadhyay:** Conceptualization – conceptualized the approach of this study. Investigation.

Formal analysis. Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript. **Jay D. Keasling:** Conceptualization – conceptualized the approach of this study. Investigation. Formal analysis. Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript. **Hector Garcia Martin:** Conceptualization – conceptualized the approach of this study. Investigation. Formal analysis. Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript. **Tyler W. H. Backman:** Methodology and Software, Data curation, Conceptualization – conceptualized the approach of this study. Investigation. Formal analysis. Writing – review & editing, were the principal investigators who directed this project, contributed to the data analysis and interpretation, as well as edited the manuscript.

References

1. Keasling, J. D. Manufacturing Molecules Through Metabolic Engineering. *Science* **330**, 1355–1358 (2010).
2. Industrial biomanufacturing: The future of chemical production | Science. https://www.science.org/doi/full/10.1126/science.aag0804.
3. Choi, K. R. *et al.* Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. *Trends in Biotechnology* **37**, 817–837 (2019).
4. Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production - Chemical Society Reviews (RSC Publishing) DOI:10.1039/D0CS00155D. https://pubs.rsc.org/en/content/articlehtml/2020/cs/d0cs00155d?casa_token=e9TRf5

rRdtkAAAAA:A0_jcEnT_yb-9TaMUZmPg4wYizhoLfKwyZQPxyOsGQdu8KyozqAPS sIh866tAyvdTldmE-LvNnClWEs&casa_token=Z5RL6C2PpOcAAAAA:2ghaergY0xk1 x4KvwzHuzISi76wQRFEi0tm9PxZi0gjt5adz6dCWsnY9QiiTLoyvbpJ5xGh8MyhXcK0.

5. Fackler, N. *et al.* Stepping on the Gas to a Circular Economy: Accelerating Development of Carbon-Negative Chemical Production from Gas Fermentation. *Annual Review of Chemical and Biomolecular Engineering* **12**, 439–470 (2021).

6. Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy | The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.

7. Synthetic Biology Is About to Disrupt Your Industry. *BCG Global* https://www.bcg.com/publications/2022/synthetic-biology-is-about-to-disrupt-your-industry (2022).

8. Synthetic Biology Is Getting Closer to Industrial Scale. *BCG Global* https://www.bcg.com/capabilities/digital-technology-data/emerging-technologies/expert-insights/nicolas-goeldel (2024).

9. Lawson, C. E. *et al.* Machine learning for metabolic engineering: A review. *Metabolic Engineering* **63**, 34–60 (2021).

10. Carbonell, P., Radivojevic, T. & García Martín, H. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synth. Biol.* **8**, 1474–1477 (2019).

11. Pathway design using de novo steps through uncharted biochemical spaces | Nature Communications. https://www.nature.com/articles/s41467-017-02362-x.

12. Maitra, S. & Maitra, K. Chemistry of Bioproducts. in *Practices and Perspectives in Sustainable Bioenergy: A Systems Thinking Approach* (eds. Mitra, M. & Nagchaudhuri, A.) 233–267 (Springer India, New Delhi, 2020). doi:10.1007/978-81-322-3965-9_11.
13. Wu, W., Long, M. R., Zhang, X., Reed, J. L. & Maravelias, C. T. A framework for the identification of promising bio-based chemicals. *Biotech & Bioengineering* **115**, 2328–2340 (2018).
14. Yim, H. *et al.* Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* **7**, 445–452 (2011).
15. Raab, A. M., Gebhardt, G., Bolotina, N., Weuster-Botz, D. & Lang, C. Metabolic engineering of *Saccharomyces cerevisiae* for the biotechnological production of succinic acid. *Metabolic Engineering* **12**, 518–525 (2010).
16. Lee, S. Y. *et al.* A comprehensive metabolic map for production of bio-based chemicals. *Nat Catal* **2**, 18–33 (2019).
17. Jang, W. D., Kim, G. B. & Lee, S. Y. An interactive metabolic map of bio-based chemicals. *Trends in Biotechnology* S0167779922001950 (2022) doi:10.1016/j.tibtech.2022.07.013.
18. Ph.D, S. M. & Ph.D, L. F. H. *Chemistry of Petrochemical Processes*. (Elsevier, 2001).
19. Sirirungruang, S., Markel, K. & Shih, P. M. Plant-based engineering for production of high-valued natural products. *Nat. Prod. Rep.* **39**, 1492–1509 (2022).
20. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends in Microbiology* **24**, 968–977 (2016).
21. Schläpfer, P. *et al.* Genome-Wide Prediction of Metabolic Enzymes, Pathways, and

Gene Clusters in Plants. *Plant Physiology* **173**, 2041–2059 (2017).

22. Rokas, A., E. Mead, M., L. Steenwyk, J., A. Raja, H. & H. Oberlies, N. Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Natural Product Reports* **37**, 868–878 (2020).
23. Grininger, M. Enzymology of assembly line synthesis by modular polyketide synthases. *Nat Chem Biol* **19**, 401–415 (2023).
24. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. *Chem. Rev.* **119**, 12524–12547 (2019).
25. Khosla, C. Structures and Mechanisms of Polyketide Synthases. *J. Org. Chem.* **74**, 6416–6420 (2009).
26. Khosla, C., Tang, Y., Chen, A. Y., Schnarr, N. A. & Cane, D. E. Structure and Mechanism of the 6-Deoxyerythronolide B Synthase. *Annual Review of Biochemistry* **76**, 195–221 (2007).
27. T. Wagner, D. *et al.* α-Methylation follows condensation in the gephyronic acid modular polyketide synthase. *Chemical Communications* **52**, 8822–8825 (2016).
28. Young, J. *et al.* Elucidation of Gephyronic Acid Biosynthetic Pathway Revealed Unexpected SAM-Dependent Methylations. *J. Nat. Prod.* **76**, 2269–2276 (2013).
29. Hertweck, C., Luzhetskyy, A., Rebets, Y. & Bechthold, A. Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Nat. Prod. Rep.* **24**, 162–190 (2007).
30. Yu, D., Xu, F., Zeng, J. & Zhan, J. Type III polyketide synthases in natural product biosynthesis. *IUBMB Life* **64**, 285–295 (2012).
31. Keatinge-Clay, A. T. The structures of type I polyketide synthases. *Nat. Prod. Rep.* **29**, 1050 (2012).

32. Li, H., Cann, A. F. & Liao, J. C. Biofuels: Biomolecular Engineering Fundamentals and Advances. *Annu. Rev. Chem. Biomol. Eng.* **1**, 19–36 (2010).

33. Menzella, H. G. *et al.* Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nat Biotechnol* **23**, 1171–1176 (2005).

34. Engineering Broader Specificity into an Antibiotic-Producing Polyketide Synthase | Science. https://www.science.org/doi/full/10.1126/science.279.5348.199.

35. De novo biosynthesis of terminal alkyne-labeled natural products | Nature Chemical Biology. https://www.nature.com/articles/nchembio.1718.

36. Engineered Biosynthesis of Alkyne-Tagged Polyketides by Type I PKSs: iScience. https://www.cell.com/iscience/fulltext/S2589-0042(20)30122-X.

37. Polyketide Bioderivatization Using the Promiscuous Acyltransferase KirCII | ACS Synthetic Biology. https://pubs.acs.org/doi/full/10.1021/acssynbio.6b00341.

38. Englund, E. *et al.* Expanding Extender Substrate Selection for Unnatural Polyketide Biosynthesis by Acyltransferase Domain Exchange within a Modular Polyketide Synthase. *J. Am. Chem. Soc.* **145**, 8822–8832 (2023).

39. Kalkreuter, E., CroweTipton, J. M., Lowell, A. N., Sherman, D. H. & Williams, G. J. Engineering the Substrate Specificity of a Modular Polyketide Synthase for Installation of Consecutive Non-Natural Extender Units. *J. Am. Chem. Soc.* **141**, 1961–1969 (2019).

40. Malico, A. A., Nichols, L. & Williams, G. J. Synthetic biology enabling access to designer polyketides. *Current Opinion in Chemical Biology* **58**, 45–53 (2020).

41. Goranovič, D. *et al.* Origin of the Allyl Group in FK506 Biosynthesis. *Journal of Biological Chemistry* **285**, 14292–14300 (2010).

42. Mo, S. *et al.* Biosynthesis of the Allylmalonyl-CoA Extender Unit for the FK506 Polyketide Synthase Proceeds through a Dedicated Polyketide Synthase and Facilitates the Mutasynthesis of Analogues. *J. Am. Chem. Soc.* **133**, 976–985 (2011).

43. Jez, J. M., Bowman, M. E. & Noel, J. P. Expanding the biosynthetic repertoire of plant type III polyketide synthases by altering starter molecule specificity. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5319–5324 (2002).

44. Montaño López, J., Duran, L. & Avalos, J. L. Physiological limitations and opportunities in microbial metabolic engineering. *Nat Rev Microbiol* **20**, 35–48 (2022).

45. Comprehensive in Vitro Analysis of Acyltransferase Domain Exchanges in Modular Polyketide Synthases and Its Application for Short-Chain Ketone Production | ACS Synthetic Biology. https://pubs.acs.org/doi/full/10.1021/acssynbio.6b00176.

46. Sirirungruang, S. *et al.* Engineering site-selective incorporation of fluorine into polyketides. *Nat Chem Biol* **18**, 886–893 (2022).

47. Fluorine in Pharmaceuticals: Looking Beyond Intuition | Science. https://www.science.org/doi/full/10.1126/science.1131943.

48. O'Hagan, D. & Deng, H. Enzymatic Fluorination and Biotechnological Developments of the Fluorinase. *Chem. Rev.* **115**, 634–649 (2015).

49. Yuzawa, S., Backman, T. W. H., Keasling, J. D. & Katz, L. Synthetic biology of polyketide synthases. *Journal of Industrial Microbiology and Biotechnology* **45**, 621–633 (2018).

50. Carbonell, P., Parutto, P., Baudier, C., Junot, C. & Faulon, J.-L. Retropath: Automated Pipeline for Embedded Metabolic Circuits. *ACS Synth. Biol.* **3**, 565–577

(2014).

51. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metabolic Engineering* **45**, 158–170 (2018).

52. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).

53. Gricourt, G., Meyer, P., Duigou, T. & Faulon, J.-L. Artificial Intelligence Methods and Models for Retro-Biosynthesis: A Scoping Review. *ACS Synth. Biol.* (2024) doi:10.1021/acssynbio.4c00091.

54. Yu, T. *et al.* Machine learning-enabled retrobiosynthesis of molecules. *Nat Catal* **6**, 137–151 (2023).

55. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).

56. An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals | Communications Biology. https://www.nature.com/articles/s42003-018-0076-9.

57. Zheng, S. *et al.* Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat Commun* **13**, 3342 (2022).

58. Merging enzymatic and synthetic chemistry with computational synthesis planning | Nature Communications. https://www.nature.com/articles/s41467-022-35422-y.

59. Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol* **12**, 1007–1014 (2016).

60. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature communications*

11, 6058 (2020).

61. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29–W35 (2021).
62. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603–D610 (2023).
63. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/45/W1/W55/3769247.
64. Beber, M. E. *et al.* eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Research* **50**, D603–D609 (2022).
65. α-Pyrones and their derivatives from two Cryptocarya species - ScienceDirect. https://www.sciencedirect.com/science/article/pii/0031942295926966.
66. Iinuma, H. *et al.* Basidalin,a new antibiotic from basidiomycetes. *J. Antibiot.* **36**, 448–450 (1983).
67. ClusterCAD 2.0: an updated computational platform for chimeric type I polyketide synthase and nonribosomal peptide synthetase design | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/51/D1/D532/6842880.
68. Ni, Z., Stine, A. E., Tyo, K. E. J. & Broadbelt, L. J. Curating a comprehensive set of enzymatic reaction rules for efficient novel biosynthetic pathway design. *Metabolic Engineering* **65**, 79–87 (2021).
69. Robbins, T., Kapilivsky, J., Cane, D. E. & Khosla, C. Roles of Conserved Active Site Residues in the Ketosynthase Domain of an Assembly Line Polyketide Synthase. *Biochemistry* **55**, 4476–4484 (2016).

70. Choi, S. *et al.* Production of 4-hydroxybutyric acid by metabolically engineered Mannheimia succiniciproducens and its conversion to γ-butyrolactone by acid treatment. *Metabolic Engineering* **20**, 73–83 (2013).

71. Choi, S., Kim, H. U., Kim, T. Y. & Lee, S. Y. Systematic engineering of TCA cycle for optimal production of a four-carbon platform chemical 4-hydroxybutyric acid in Escherichia coli. *Metabolic Engineering* **38**, 264–273 (2016).

72. Hein, S., Söhling, B., Gottschalk, G. & Steinbüchel, A. Biosynthesis of poly(4-hydroxybutyric acid) by recombinant strains of Escherichia coli1. *FEMS Microbiology Letters* **153**, 411–418 (2006).

73. Min, K. *et al.* Conversion of levulinic acid to 2-butanone by acetoacetate decarboxylase from Clostridium acetobutylicum. *Appl Microbiol Biotechnol* **97**, 5627–5634 (2013).

74. Fadda, S., Lebert, A., Leroy-Sétrin, S. & Talon, R. Decarboxylase activity involved in methyl ketone production by Staphylococcus carnosus 833, a strain used in sausage fermentation. *FEMS Microbiology Letters* **210**, 209–214 (2002).

75. Park, J. *et al.* Synthesis of methyl ketones by metabolically engineered *Escherichia coli*. *Journal of Industrial Microbiology and Biotechnology* **39**, 1703–1712 (2012).

76. Lee, H., DeLoache, W. C. & Dueber, J. E. Spatial organization of enzymes for metabolic engineering. *Metabolic Engineering* **14**, 242–251 (2012).

77. Sun, X. *et al.* Metabolic pathway assembly using docking domains from type I cis-AT polyketide synthases. *Nat Commun* **13**, 5541 (2022).

78. T. Keatinge-Clay, A. Stereocontrol within polyketide assembly lines. *Natural Product Reports* **33**, 141–149 (2016).

79. BJOC - Polyketide stereocontrol: a study in chemical biology.

https://www.beilstein-journals.org/bjoc/articles/13/39.

80. Hall, M. Enzymatic strategies for asymmetric synthesis. *RSC Chemical Biology* **2**, 958–989 (2021).

81. Duigou, T., du Lac, M., Carbonell, P. & Faulon, J.-L. RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Research* **47**, D1229–D1235 (2019).

82. EnzymeMap: curation, validation and data-driven prediction of enzymatic reactions - Chemical Science (RSC Publishing) DOI:10.1039/D3SC02048G. https://pubs.rsc.org/en/content/articlehtml/2023/sc/d3sc02048g.

83. Wiest, O., Bauer, C., Helquist, P., Norrby, P.-O. & Genheden, S. Finding Relevant Retrosynthetic Disconnections for Stereocontrolled Reactions. *J. Chem. Inf. Model.* (2024) doi:10.1021/acs.jcim.4c00370.