Natural Product Reports

Engineering biosynthetic enzymes for industrial natural product synthesis

Journal:	*Natural Product Reports*
Manuscript ID	NP-REV-12-2019-000071.R1
Article Type:	Review Article
Date Submitted by the Author:	01-Apr-2020
Complete List of Authors:	Galanie, Stephanie; Oak Ridge National Laboratory, Biosciences Division Entwistle, David; Codexis, Inc., Process chemistry Lalonde, James; Inscripta, Inc., Microbial Digital Genome Engineering

ARTICLE

Engineering biosynthetic enzymes for industrial natural product synthesis

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Stephanie Galanie,*[a] David Entwistle [b] and James Lalonde [c]

Natural products and their derivatives are commercially important medicines, agrochemicals, flavors, fragrances, and food ingredients. Industrial strategies to produce these structurally complex molecules encompass varied combinations of chemical synthesis, biocatalysis, and extraction from natural sources. Interest in engineering natural product biosynthesis began with the advent of genetic tools for pathway discovery. Genes and strains can now readily be synthesized, mutated, recombined, and sequenced. Enzyme engineering has succeeded commercially due to the development of genetic methods, analytical technologies, and machine learning algorithms. Today, engineered biosynthetic enzymes from organisms spanning the tree of life are used industrially to produce diverse molecules. These biocatalytic processes include single enzymatic steps, multi-enzyme cascades, and engineered native and heterologous microbial strains. This review will describe how biosynthetic enzymes have been engineered to enable commercial and near-commercial syntheses of natural products and their analogs.

1 Introduction

There are two main goals in industrial engineering of biosynthetic enzymes: (1) produce valuable products, often inaccessible by other synthetic methods, and (2) decrease cost by increasing process productivity, titer, and yield.[1] While all enzymes that catalyze bond-forming reactions are considered biosynthetic, we have primarily restricted this discussion to enzymes involved in secondary metabolism. These specialized enzymes catalyze reactions that challenge or even evade synthetic chemists and often exhibit exquisite chemo- and regio-selectivity. However, natural wild-type enzymes are not always well suited to the industrial conditions of expression in standard microbial host organisms, low or high pH, elevated temperature, or high organic co-solvent.[2] Additionally, substrate scope, cofactor recycling, catalytic rate, and tolerance to high substrate and product concentrations may need to be altered or enhanced.[2] To use these enzymes in an industrial process, they must be engineered to mitigate these limitations and meet the productivity targets required for commercial viability. To our knowledge, this is the first review to focus on the industrial development and use of engineered secondary metabolism enzymes to produce natural products and their analogs at or near commercial scale. Excellent reviews and perspectives have been compiled on engineering biocatalysts by academic and industrial leaders,[3-5] on biocatalysis for natural product synthesis by Classen and Pietruszka,[6] Friedrich and Hahn,[7] and Tibrewal and Tang,[8] and on synthetic biology

[a] *Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. E-mail: galaniess@ornl.gov*
[b] *Process Chemistry, Codexis, Inc., Redwood City, California, USA.*
[c] *Microbial Digital Genome Engineering, Inscripta, Inc., Pleasanton, California, USA.*

approaches to combinatorial biosynthesis by Kim, Moore, and Yoon.[9] Reviews have also been published on approaches to engineer microorganisms as cell factories,[10, 11] approaches to supply plant-derived natural products,[12, 13] and opportunities to discover and engineer enzymes from natural product biosynthesis.[14-16] Here, we present a non-exhaustive overview of companies and their products realized through engineering of biosynthetic enzymes with examples from the last two decades.

2 From classical strain engineering to directed evolution

After decades of engineering microbial strains to produce antibiotics and other natural products by classical non-targeted mutagenesis and screening, advances in genetic engineering technology significantly shortened development times. Researchers could now manipulate individual genes in native or recombinant hosts and make directed mutations, facilitating discovery and engineering. Reports of isolating and genetically disrupting large multi-functional polypeptides, including polyketide synthases (PKS)[17, 18] and non-ribosomal peptide synthases (NRPS)[19, 20] inspired efforts to generate new molecules via combinatorial biosynthesis. Combinatorial biosynthesis included alternate precursor feeding and rational protein engineering in the form of domain inactivation and swapping, primarily in the native producing host or a closely related organism.[21] During the same time period, directed evolution was developed. Directed evolution achieves desired performance by mutating a gene or genome, screening and accumulating beneficial improvements over multiple generations.[22] Adding genetic recombination at the single gene or whole genome level accelerated both classical strain development and directed evolution. This enabled commercial-scale production of natural products using engineered biosynthetic enzymes. Examples of engineered enzymes mentioned in this review are summarized in Table 1.

2.1 Commercializing combinatorial biosynthesis

Multiple companies in the 1990s began to commercialize combinatorial biosynthesis. In 1995, Khosla's laboratory spun out Kosan Biosciences to produce drug candidates, including analogs of erythromycin, geldanamycin, and epothilone, using combinatorial biosynthesis with PKS modules. Example analogs are shown in Fig. 1. Kosan successfully produced a library of macrolides by genetic substitution of modules in the 6-deoxyerythronolide B PKS (DEBS) with those from the rapamycin PKS.[23]

Table 1 Examples of engineered biosynthetic enzymes developed industrially or in commercial processes for natural product and analog synthesis.

Engineered biosynthetic enzyme	Source organism and natural product class	Commercialized by
6-deoxyerythronolide B,[23] geldanamycin,[24] and epothilone[25] PKS	*Saccharopolyspora erythraea, Streptomyces hygroscopicus, Myxococcus xanthus* macrolide polyketides	Kosan Biosciences/Bristol-Myers Squibb
spinosyn PKS[26]	*Saccharopolyspora spinosa* macrolide polyketide	Biotica Technology (Isomerase Therapeutics)/Dow Agrosciences (Corteva™ Agriscience)
rapamycin PKS[27]	*Streptomyces hygroscopicus* macrolide polyketide	Biotica Technology (Isomerase Therapeutics)/ Wyeth Laboratories (Pfizer)
daptomycin and A54145 NRPS[28, 29]	*Streptomyces roseosporus* and *Streptomyces fradiae*cyclic lipopeptide	Cubist Pharmaceuticals
β-carotene ketolase[30]	*Sphingomonas* sp. DC18 carotenoid (tetraterpenoid)	DuPont
Genome of tylosin-producer[31]	*Streptomyces fradiae* macrolide polyketide	Maxygen (Codexis)/Eli Lilly
α-ketoglutarate-dependent-dioxygenase deacetoxycephalosporin C synthase (DAOCS, expandase)[32-38]	*Streptomyces clavuligerus* cephalosporin β-lactam peptide	DSM, Synmax Biochemical collaboration
AveC avermectin intermediate spirocyclase and dehydratase[39-42]	*Streptomyces avermitilis* macrolide polyketide	Pfizer, Biotica Technology, and Maxygen (Codexis)
salutaridine reductase[43, 44]	*Papaver somniferum* isoquinoline alkaloid	Arzeda Corp.
LovD lovastatin intermediate transesterase[45-47]	*Aspergillus terreus* polyketide	Codexis/Arch Pharmalabs Ltd.
tryptophan halogenase[48-50]	*Lechevalieria aerocolonigenes* indole alkaloid	Novartis collaboration
steviol glycoside glycosyltransferases and sucrose synthase[51-53]	*Stevia rebaudiana* and others terpenoid glycoside	Amyris/DSM/Raízen/ASR Group, Codexis/Tate & Lyle, Cargill, Conagen, Evolva, Manus
prenyltransferase NphB[54-56]	*Streptomyces sp.* CL190 hybrid isoprenoid-polyketide	Invizyne Technologies, funded by Bioactive Ingredients Corporation

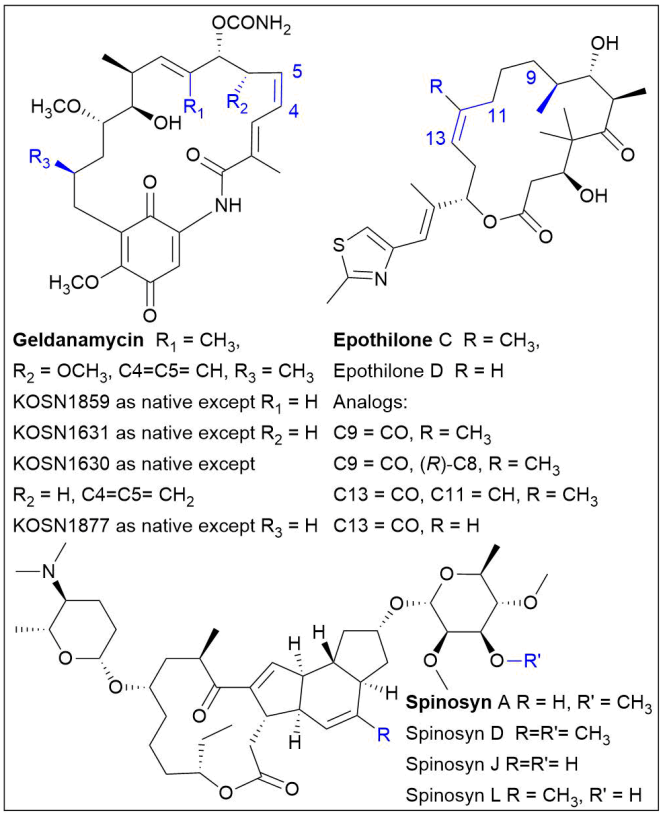


Fig. 1 Type I PKS-derived natural product analogs produced by combinatorial biosynthesis. Blue indicates modified portions of the molecule.

Manipulations of the PKS modules were performed in *E. coli* and molecules were produced in *Streptomyces spp*. Kosan developed geldanamycin analogs that were not accessible through chemical synthesis and that had enhanced affinity for the cancer target Hsp90 through engineered biosynthesis in the native producer *Streptomyces hygroscopicus*.[24] The researchers developed three different genetic methods to swap seven different geldanamycin PKS acyltransferase domains for two unique rapamycin acyltransferase domains with alternate specificities. They found that although complementation required initial investment to develop strains and vectors, it provided greater speed and flexibility in facilitating combinatorial biosynthesis. Kosan scientists also produced multiple unexpected oxo-derivatives of microtubule-targeting epothilones by using homologous recombination to inactivate two of the *Myxococcus xanthus* PKS ketoreductase domains.[25] Bristol-Myers Squibb acquired Kosan in 2008 for its PKS-derived drug candidate microtubule stabilizers and Hsp90 inhibitors.[57] Importantly, these molecules included both known and new-to-nature compounds, and the next aim for combinatorial biosynthesis was to increase titers of desired products.

In 1996, Leadlay and Staunton spun-out Biotica Technology (now Isomerase Therapeutics) to commercialize their platform for producing novel macrolide analogs by engineering the N-terminal loading module, which controls the substrate scope of chain initiation.[58, 59] Biotica collaborated with Dow Agrosciences (now Corteva™ Agriscience) on the macrocyclic *Saccharopolyspora spinosa* natural product insecticide Spinosad, composed of spinosyn A and D (Fig. 1), which had received the Green Chemistry award in 1999. Biotica produced novel spinosyns by using homologous recombination to replace the loading module in the spinosyn PKS with those from the avermectin and DEBS PKS assemblies and feeding alternative carboxylic acid precursors.[26] One of the derivatives, which was not accessible through semisynthesis alone, was chemically hydrogenated, producing a compound that had lower lethal concentrations for five insect pests than spinosyn A/D. However, titers were decreased relative to wild-type PKS. Biotica also applied their starter unit modification strategy to rapamycin by feeding alternative carboxylic acids to a *rapK* deleted mutant of the native producing strain, thereby producing new anticancer candidate mTOR and kinase inhibitors that were licensed to Wyeth Laboratories (now Pfizer).[27]

The spinosyn efforts continued at Dow, where artificial neural network modelling was combined with a synthetic modification program, resulting in the identification of a hydrogenated and 3'-O-ethylated spinosyn analog with broad improvement in insecticidal activity.[60] This required starting from a 3'-O-desmethyl rhamnose spinosyn (J and L, Fig. 1), and strains with mutations that eliminated this methylation were identified through an extensive classical strain improvement program conducted at Lilly and Dow. The resulting product, Spinectoram, a semi-synthetic derivative of spinosyns J and L, received the Green Chemistry award in 2008.[61, 62]

Cubist Pharmaceuticals' scientists produced novel daptomycin derivatives (Fig. 2) by exchanging modules to build 30 combinatorial biosynthetic pathways.[28] Daptomycin is a cyclic lipopeptide produced by an NRPS in *Streptomyces roseosporus* and is bactericidal to methicillin-resistant *Staphylococcus aureous* (MRSA).[63] Daptomycin was discovered by Eli Lilly, licensed to Cubist Pharmaceuticals in 1997, and received FDA approval in 2003.[64] Fermentations of the engineered *S. roseosporus* strains produced derivatives at titers of 1-100 mg/L.[28] Cubist was acquired by Merck in 2014.[65] Excellent reviews of PKS and NRPS engineering by module inactivation

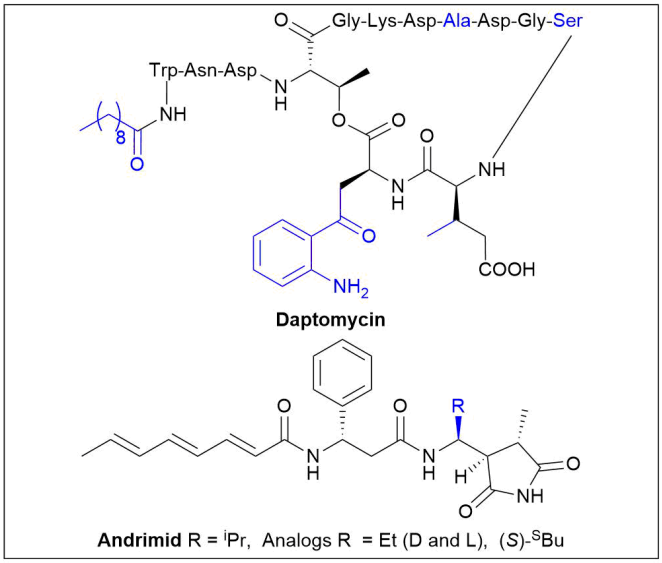


Fig. 2 NRPS- and NRPS-PKS derived natural product analogs produced by combinatorial biosynthesis.

and swapping, including recent synthetic biology approaches, are available.[9, 66, 67]

2.2 Applying directed evolution to natural product biosynthesis

Combinatorial biosynthesis successfully produced a range of new bioactive molecules, primarily from microbes such as *Streptomyces*, but obtaining commercially relevant productivities remained challenging. New protein and genetic engineering methods were needed to overcome this obstacle. Directed evolution was initially developed in the 1990s and recognized with the 2011 Draper prize to Arnold and Stemmer and more recently with part of the shared 2018 Nobel prize in Chemistry to Arnold. A round of directed evolution consists of three phases – library construction, library screening, and selection of a parent or parents and/or diversity to carry forward for subsequent rounds (Fig. 3).

While decades of mutagenesis and screening had been applied to antibiotic-producing microbes, this strategy was not translated to the molecular scale and applied to biosynthetic enzymes and pathways until the 2000s. Initially, libraries were constructed randomly by error-prone polymerase chain reaction (epPCR) targeting a specific number of mutations per sequence or rationally by site-directed mutagenesis with defined or degenerate DNA oligonucleotides (saturation mutagenesis). DuPont used directed evolution to produce astaxanthin (Fig. 4), a carotenoid used as a pigment for poultry and fish feed.[30] They first identified a β-carotene ketolase gene by screening an orange-pigmented environmental isolate genomic insert library in a zeaxanthin-producing *E. coli* strain. Next, they subjected the best hit to random mutagenesis by epPCR, and isolated orange and red-orange colonies with astaxanthin increased from 14% to 83%.[30] The beneficial mutations included coding mutations, silent mutations in the coding region, and the ribosome binding site, affecting both enzyme activity and translation rate.

Academic researchers applied directed evolution to chimeric combinatorial biosynthetic pathways to improve product titers. For this project, they used the *E. coli* enterobactin and *Pantoea agglomerans* antibiotic andrimid NRPS-PKS (Fig. 2) with an adenylation (A) domain swapped with one from a different NRPS-PKS.[68] A zone-of-inhibition assay was used to screen epPCR libraries of A domain mutants in an *E.coli* strain expressing a chimeric NRPS-PKS with an inactivated A domain. The best variants were re-assayed and a parent was identified for the next round of evolution. After three rounds of mutagenesis and screening, activity was recovered from 3% to 33% of wild-type and clones were identified that produced new andrimid derivatives from fed precursors. Notably, mutations that improved antibiotic yield or activity were distributed throughout the A domain, demonstrating the power of directed evolution to provide difficult to rationalize protein engineering solutions.

2.3 Accelerating directed evolution with recombination

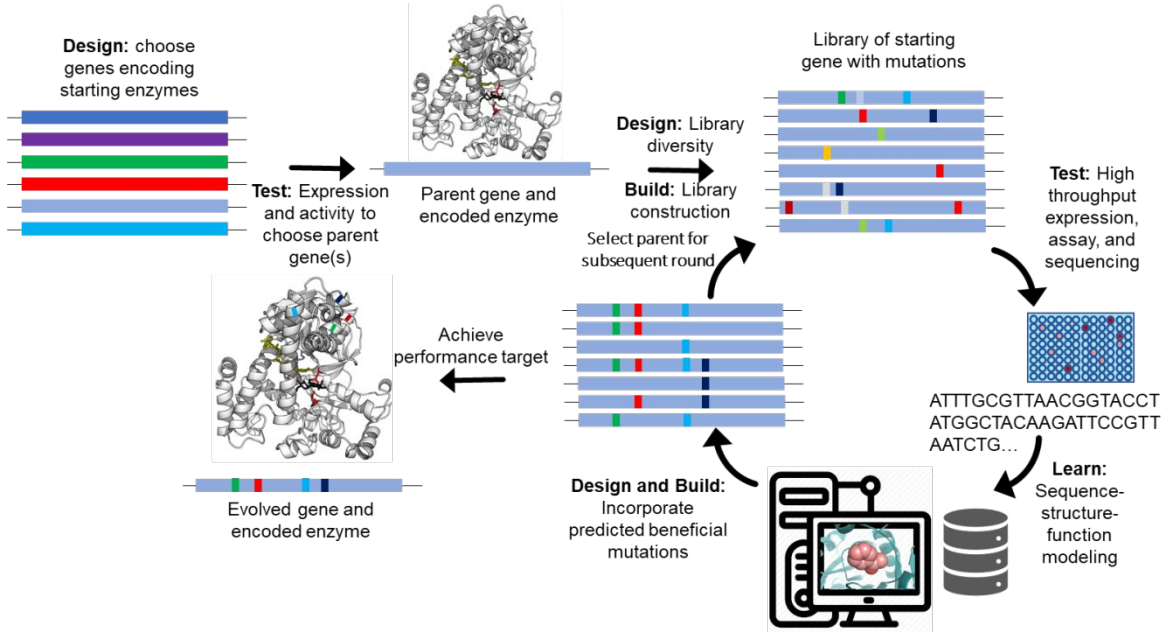
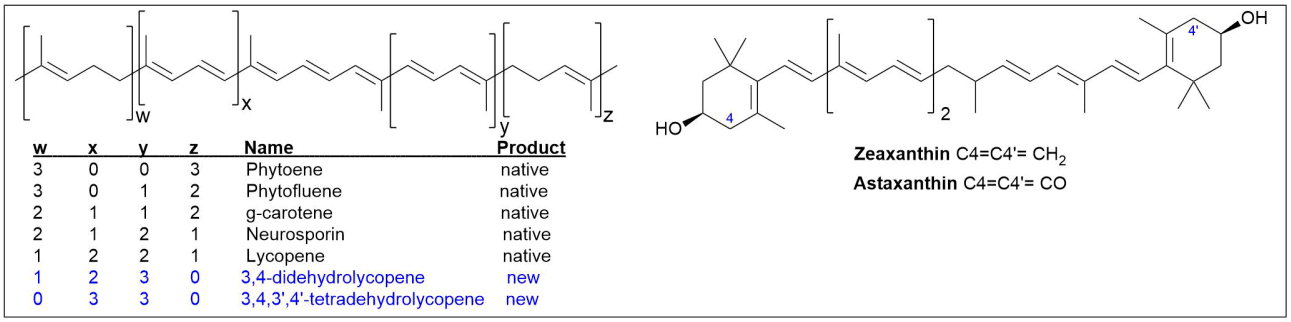


Fig. 3 Enzyme directed evolution workflow. Directed evolution follows the design, build, test, learn cycle. Starting with identifying potential starting genes from literature and databases, which are then expressed and assayed for activity. The gene that produces the most protein with the highest desired activity is used as the template for library construction. Libraries are designed using a variety of strategies (see 3.1), and then built, transformed, expressed, assayed, and sequenced to generate data from which sequence-structure-function models are built. These models then learn from the data to predict which mutations are beneficial and which sequence is the most fit, and these mutations are incorporated into the subsequent library. Once the performance targets are met, evolution is complete.



w	x	y	z	Name	Product
3	0	0	3	Phytoene	native
3	0	1	2	Phytofluene	native
2	1	1	2	g-carotene	native
2	1	2	1	Neurosporin	native
1	2	2	1	Lycopene	native
1	2	3	0	3,4-didehydrolycopene	new
0	3	3	0	3,4,3',4'-tetradehydrolycopene	new

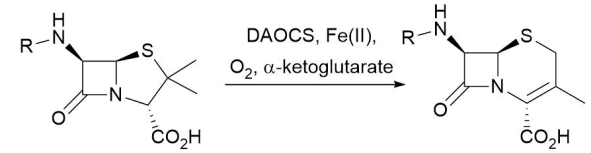
Fig. 4 Carotenoid substrate and products of engineered β-carotene ketolase and shuffled carotenoid biosynthetic pathway. Blue indicates new analogs or enzyme catalyzed modifications.

Stemmer introduced DNA shuffling, also known as molecular breeding or *in vitro* homologous recombination, which recombines diversity in a mode analogous to that effected by sexual reproduction.[69] To our knowledge, the first published application of molecular breeding to a biosynthetic enzyme was the *in vitro* evolution of phytoene desaturase and lycopene cyclase for the production of modified carotenoids.[70] Stemmer's gene shuffling method was used to generate desaturase and cyclase enzyme libraries in *E. coli* starting from parent genes from the Gram-negative bacterial genus *Erwinia*. Mutants of interest were identified by visually screening colonies for varied carotenoid color. These mutants were expressed in combination with two additional carotenoid biosynthesis enzymes, phytoene synthase and geranylgeranyldiphosphate synthase (GGDPS), to produce a range of carotenoids (Fig. 4), including a molecule not produced by either of the organisms from which the parent genes were obtained. This provided another example of biosynthesis of new-to-nature compounds.

To produce natural products, directed evolution has been applied to individual genes encoding biosynthetic enzymes, multi-gene pathways, and entire genomes. In a collaboration between Maxygen (now Codexis) and Eli Lilly, whole genome shuffling was used to increase titers of the macrolide antibiotic tylosin in the commercial producing organism *Streptomyces fradiae.*[31] In one year, one round of random chemical mutagenesis followed by two rounds of whole genome shuffling reached tylosin titers exceeding those achieved in 20 years with 20 rounds of classical strain improvement (random chemical or UV-induced mutagenesis and screening without recombination). This early industrial example of directed evolution via whole genome shuffling was accomplished by recursive protoplast fusion, a technique published in the 1970s.[71, 72] Importantly, this study demonstrated the higher efficiency of "sexual recursive recombination," in which 7-11 improved variants were selected and recombined each round relative to "asexual recursive mutagenesis," in which an individual best parent was selected and mutagenized each round.

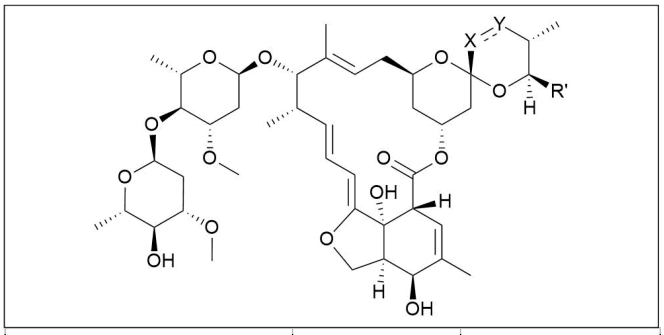
Directed evolution has also been used to produce precursors for semi-synthetic β-lactam antibiotics. Penicillins, produced by fermenting *Penicillium chrysogenum*, can be obtained at higher volumetric productivity than cephalosporins, produced by fermenting *Acremonium chrysogenum*.[73] Therefore, one strategy for decreasing the cost of cephalosporins is to produce them from penicillins. DSM N.V. disclosed a method to produce the semi-synthetic cephalosporin precursors 7-amino-(deacetoxy)cephalosporanic acid (7-ACA and 7-ADCA) from penicillins related to that published by Crawford.[33, 38, 74] In this method, acyl side chain precursors are fed to produce acyl-6-aminopenicillanic acid (acyl-6-APA), which an expandase expands from the 5-membered penicillin ring to the 6-membered cephalosporin ring (acyl-7-ADCA or acyl-7-ACA),[32, 37, 38] and finally the acyl side chain is removed with penicillin G acylase.[33, 74] DSM partnered with Maxygen to evolve enzymes for penicillin intermediates, which were commercialized.[75-78] Expandase, discovered in cephalosporin producers, has narrow specificity for penicillin N, while *P.* chrysogenum produces primarily penicillins G and V.[32, 37] The directed evolution of expandase (Fig. 5) was performed by several research groups, including an academic collaboration with Synmax Biochemical.[34-36] This project first identified new expandase genes by screening isolates for β-lactam antibiotic activity, using Southern blots to identify strains with expandase homologs, isolating genes from an insert library, and performing family shuffling to generate diversity. Random and rational structure-guided mutagenesis were used to identify diversity, and DNA shuffling was used to recombine diversity to achieve improved affinity for and activity on penicillin G.

Industrial engineering of the avermectin pathway is another example of developing and applying these emerging approaches. The commercial veterinary antiparasitic doramectin is an analog of the *Streptomyces avermitilis* macrocyclic lactone avermectin and was produced along with



R group	Substrate	Product	Comment
D-aminoadipoyl	Penicilin N	DAOC	native reaction
Phenacyl	Penicillin G	G-7-ADCA	>100 fold improved k_{cat}/K_M

Fig. 5 Expandase (deacetoxycephalosporin C synthase, DAOCS) natively catalyzes the expansion of penicillin N to DAOC and was engineered to act on the non-native substrate penicillin G to produce G-7-aminodeacetocephalosporanic acid (G-7-ADCA) which can be deacylated to provide the semi-synthetic precursor 7-ADCA.



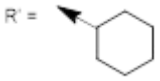
Conditions	Main product X-Y = CH=CH	By-product X-Y = CH2-CH(OH)
Wild type *S. avermitilis* strain Native feeding regime	Avermectin B1a	Avermectin B2
	R' = 	
Mutant *S. avermitilis* strain Non-native feeding regime	Doramectin 50%	CHC-B2 50%
	R' = 	
Mutant *S. avermitilis* strain + mutant AveC Non-native feeding regime 20% increased doramectin titer	Doramectin 93%	CHC-B2 7%

Fig. 6 Engineering of the avermectin biosynthetic gene *aveC* resulted in increased doramectin-to-byproduct ratio and increased titer.

an undesirable analog byproduct by a mutant strain fermented with cyclohexanecarboxylic acid in a process developed by Pfizer and Biotica (Fig. 6).[39] Substitutions in the gene *aveC* were known to affect the ratio of doramectin to byproduct, but the function of the AveC protein was unknown. In partnership with Pfizer, Maxygen performed four rounds of directed evolution targeting this gene and reduced the byproduct-to-doramectin ratio from 1:1 to 0.07:1 while increasing doramectin titer ~20%.[40, 41]

The team developed semi-synthetic gene shuffling, in which a parent gene is used as the template and the DNA shuffling reaction is spiked with oligonucleotides (oligos) encoding beneficial mutations from previous libraries. Semi-synthetic DNA shuffling incorporates the advantages of sexual recursive recombination and results in smaller theoretical library sizes by controlling the number of possible mutations per library, the mutation rate per construct, and the occurrence of frameshifts. Protein and strain engineers developed a variety of techniques to leverage sexual recursive recombination and to minimize library and screen sizes. These techniques use incorporation of oligos and recombination to generate diversity that ranges from completely random to site-specific and are reviewed elsewhere.[4, 79, 80]

These approaches marked a shift from "blind" directed evolution, in which limited sequence information was needed or used and few oligos were required, to a more sequence-activity focused era. As summarized in Table 2, by the mid-2000s, academic and industrial researchers had established the power of engineering biosynthetic enzymes to increase productivity and make new molecules, of recombination to accelerate directed evolution, and of efficient diversity generation to reduce the size of screens needed to identify improved variants.

3 Shifting enzyme engineering from a project to a platform technology

Major technological advances in the mid-2000s resulted in disruptive innovation in bioengineering, changing how enzyme libraries were designed, built, screened, sequenced, analyzed, and recombined. These changes further accelerated the pace of directed evolution and the multifactorial improvements that could be achieved in enzyme performance, enabling the evolution of increasingly diverse enzyme classes. In this section, we provide examples of how key developments (Table 2) from 2004-present have improved the design, build, test, learn cycle and enabled engineering of biosynthetic enzymes.

3.1 Designing and building better libraries

Library design refers to the selection of the starting gene or genes and diversity to be incorporated into the library. As mentioned in section 2, the starting genes for the earliest enzyme libraries were individual characterized genes. Later, homologs were identified for shuffling using molecular biology techniques, for example cloned from previously sequenced bacterial gene clusters.[70] Genome mining, the functional or oligo probe-based screening of genomic insert library, was used to find a starting enzyme, as in the β-carotene ketolase example.[30] Environmental mining was also used to identify enzymes in cultured and uncultured organisms by functional and sequence-based screening. In the expandase example, functional bioassays and sequence-based Southern blots were used to identify cultured soil isolates that produced β-lactam antibiotics and contained expandase homologs, which were then cloned for homolog shuffling.[35] In another example, Ecopia BioSciences sequenced genomic DNA fragments from 70 actinomycete strains in plasmids, identified natural product biosynthetic gene fragments, and then used those fragments to probe a cosmid library containing full length biosynthetic gene clusters (BGCs).[81] They identified 11 new BGCs encoding unknown enediyne PKS natural products that were only detected by bioassay under specific fermentation conditions, and therefore would have been missed if not screened using a sequence-based method. By the mid-2000s, as the Human Genome Project drove down the cost of sequencing at centers such as the US DOE Joint Genome Institute, it became possible for companies to sequence metagenomic insert libraries built directly from environmental DNA and mine the sequence data for homologs to clone and screen. For example, the biotechnology company Diversa used combined functional and sequence-based discovery approaches to identify pectinolytic

enzymes from microbial metagenomic libraries from tropical soil samples and nitrilases from over 600 unique environmental samples.[82-84] These approaches identified novel enzyme encoding genes directly from uncultured organisms found in the environment. While functional screening approaches require that the genes in the insert library be expressed, sequence-based approaches can identify even unexpressed genes.

Table 2 Summary of key technology developments and their application in enzyme engineering.

	Design Methods to select starting gene(s) and diversity	**Build** Methods to introduce and recombine diversity	**Test** Methods to screen and sequence variants	**Learn** Methods to develop predictive sequence-function models
1999-2001	Single, double, and triple PKS module substitutions (domain swapping)[23, 85]	Blue Heron offers commercial gene synthesis[86] Method to recombine genes with low homology[87] Module substitution accomplished by inserting restriction sites by PCR mutagenesis and digesting and ligating, then expressing the mutant PKS in a PKS deleted strain[23] Gene site saturation mutagenesis by PCR with degenerate oligos for each codon[88, 89]	Modified PKS products detected by LC-MS.[23]	Phylogenetic mapping to identify functional regions of proteins[90]
2002-2004	Sexual recursive recombination[31] Superimposition of structures to identify active site residues to target[34] >100 potential starting enzymes discovered from by functional screening of environmental DNA libraries from uncultured organisms[84]	Whole genome shuffling[31] Synthetic shuffling to create combinatorial libraries from oligos[91] Multi-site-directed mutagenesis kit available[92] epPCR and site directed mutagenesis, family shuffling of genes and semi-synthetic family shuffling[34, 35, 40, 41] Gene assembly on a microfluidic chip from pooled oligonucleotides[93] Codon optimization of synthetic genes demonstrated[94]	Genome shuffled variants fermented in 96-well plates and products detected spectrophotometrically[31] Mutants assayed by visual clearing zone of antibiotic[34] or by flow injection MS/MS of 96-well fermentations[40] Best variants sequenced to select mutations to combine[34, 40] MALDI-TOF method for enzyme screening[95] U.S. standards set for microplate dimensions[96]	Partial least squares regression used to determine which mutations are beneficial in combinatorial library[97]
2005-2008	30 hybrid NRPS pathways[28] First examples of computational enzyme design[98, 99]	Hybrid NRPS pathways built using λ-Red-mediated recombination[28] Circular permutation[100]	Next generation sequencing commercialized[101] Ultra-high performance liquid chromatography with <2 μm particles[102] RapidFire solid phase extraction-MS commercialized[103]	Support vector machines used to predict stability[104, 105] Multiple linear regression algorithms used to guide protein engineering[106]
2009-2012	Genes selected from database and synthesized[107] Target surface exposed cysteines based on homology modeling[108]	Multiplex automated genomeengineering[109] Phage assisted continuous evolution[110]	Fluorescence-activated droplet sorting[111]	Support vector machine used to predict protein solubility[112]
2013-2016	Many user-friendly tools become available for library design[113]	CRISPR/Cas9 gRNA guided nicking and recombination[114-116]	Optical single-cell analysis and microcapillary arrays[117-119]	Gaussian process model guided p450 chimera evolution[120]
2017-2020	Autodock and Rosetta used commercially to examine substrate binding and design focused libraries of mutants[55]	CRISPR-enabled trackable genome engineering and base editors[121, 122]	Sciex launched Echo acoustic mass spectrometer[123] Mass-activated droplet sorting[124]	Co-evolution and deep learning strategies prevailed in structure prediction competition[125] Application of machine learning methods to 200-10000 variants[126]

In addition to Diversa, several companies, including a few mentioned here, have specialized in compiling and leveraging sequenced (meta)genomic insert libraries or strain collections to provide new secondary metabolites and enzymes that could be the starting point for directed evolution. Radiant Genomics, founded in 2012, developed a natural product producing clone library from uncultivated organisms and was acquired by Zymergen in 2018.[127, 128] Warp Drive Bio was founded in 2012 and backed by Sanofi to find new drugs from soil metagenomes and specialized in finding molecules to disrupt protein-protein interactions.[129, 130] Lodo Therapeutics, spun-out of Sean Brady's laboratory in 2016, has a collaboration with Genentech to discover small molecule drugs from microbial metagenomics.[131, 132] Varigen Biosciences, founded in 2017 with academic roots dating to 2000, develops cloning products for sequence-specific capture, sequencing, and expression of 20-150 kilobase environmental DNA directly from soil and sequenced metagenomic libraries that contain full BGCs.[133] Metagenomic mining strategies are also being expanded to eukaryotic microbes by LifeMine.[134]

In 2005-2007, the first three commercial instruments for next generation sequencing (NGS) were launched and greatly enhanced sequencing throughput by massively parallelizing sequencing reactions in flow cells.[101] NGS also involves ligating sheared DNA to oligo adapters, so it is no longer necessary to clone environmental or genomic DNA into vectors in order to sequence it. This resulted in an explosion of genetic sequence databases with new genes, genomes, and metagenomes. Researchers could now mine public databases to select unexpressed genes from uncultured organisms from samples from all over the world. Commercial gene synthesis to directly obtain genes from databases was already available, launched by Blue Heron in 1999, but was expensive with long turnaround times.[135, 136] As documented in the "Carlson curves," the prices per base pair for oligo and gene synthesis have fallen continuously over the last two decades due to advances in array-based synthesis and gene assembly, making synthesis of a starting gene an increasingly viable option and alternative to molecular cloning.[93, 137, 138] Gene synthesis enables researchers to reduce codon bias and alter regulatory and structural elements when expressing a gene heterologously, often resulting in improved expression.[94]

In 2002, Maxygen combined gene assembly from synthetic "backbone" oligos and "spiking" oligos (to introduce mutations) with fragmentation-based shuffling, a method they called synthetic shuffling.[91] In 2004, Kosan researchers assembled a 32-kb polyketide synthase gene cluster from short oligos in-house.[139] In 2009, researchers in the Voigt laboratory at UCSF had 89 gene homologs from the NCBI sequence database synthesized by DNA 2.0, screened them for activity with *E. coli* expression, and then incorporated them into a yeast pathway, an approach they called "synthetic metagenomics."[107] In 2015, Ginkgo Bioworks and Twist Bioscience announced a 100 million base pair gene synthesis deal, and in 2017, a gigabase deal.[140] In a large collaborative effort to produce 10 molecules in 3 months, including the natural products pyrrolnitrin, carvone, barbamide, rebeccamycin, pacidamycin D, vincristine, C-1027, and epicolactone, researchers used five expression hosts and built 215 strains using genes primarily identified bioinformatically.[141] The decreasing cost and increasing throughput of DNA sequencing and synthesis have had many impacts, one of which is ability of a researcher to choose a sequence from a database, have it synthesized, and use it as the starting point of a directed evolution campaign. In addition to mining natural diversity for starting enzymes, gene synthesis has enabled reconstruction of ancestral enzymes,[142-144] which are often thermostable and promiscuous, and construction of designed or re-designed enzymes.[5, 145-147]

Once a starting gene has been selected, researchers must choose what diversity to incorporate into the library and how to incorporate it. With 20 possible amino acids at every position, even a protein with only 100 residues has 10^{130} possible unique sequences. Therefore, researchers must have a strategy to test a reasonable subset of the enormous possible diversity. Library diversity can be completely irrational, incorporating only random mutations through a strategy such as error-prone PCR[85, 148-152] or an *E. coli* mutator strain,[153, 154] or completely rational, using site-directed mutagenesis[155] to mutate specific amino acid positions to desired new amino acids based on structural analysis of the protein.

Rational library design strategies leverage phylogenetic, structural, and/or biophysical properties of residues to select which positions to target with which residues. Rational approaches include swapping and/or deletion of domains at defined locations, as described in several of the examples in section 2. In a structure-based academic examination of terpene synthase specificity, contact mapping was used to identify residues within van der Walls radii of the substrate in a crystal structure (first-tier) and residues in contact with those (second-tier), out to 12.5 Å from the center of the active site.[156] The authors then modelled the structure of a terpene synthase with alternate specificity onto the structure to identify residues within the contact mapped set that varied and then targeted these for replacement. They identified 9 amino acid differences between the enzymes that mediate product specificity. Rational engineering has also been used to improve stability and solubility. The Tang laboratory increased simvastatin synthase protein solubility ~50% by mutating surface-exposed cysteines based on a homology structural model.[108]

While X-ray crystal structures are certainly enabling for rational approaches, they are not a requirement. From 1994-present, the Critical Assessment of Protein Structure Prediction (CASP) results have tracked the maturation of protein structure prediction algorithms. In 1997, protein homology modelling and site-directed mutagenesis were already widely used in the pharmaceutical industry to determine substrate selectivity.[157] Algorithms now mine sequence and structure databases to provide informative structures for library design from gene sequence. Evolutionary conservation has been used to determine residues that are structurally and functionally important.[90] The CASP11 and 12 competitions found that combining machine learning techniques with co-evolution analysis facilitates *ab initio* folding.[125, 158] Co-evolution analysis predicts physical contacts between residues by identifying pairs

of amino acids that co-evolve in a phylogeny. Phylogeny-leveraging structure prediction methods require many homologous DNA sequences and have therefore greatly benefited from growth in sequence databases due to NGS.[159]

In addition to rational site-directed mutagenesis based on the inspection of solved or predicted protein structures, computational enzyme (re-)design has begun to produce enzymes with new or altered activities. This field has been recently reviewed, and the most successful examples combine *de novo* design with directed evolution.[160, 161] Arzeda, a 2010 spin-out from the Baker laboratory, disclosed their approach to enzyme engineering, which takes into account active site constraints, including residue identity and geometry, and protein structure, either from crystal structure or modeled by a homology search, then applies computational docking and side-chain redesign or repacking, and finally ranks the resulting constructs by predicted energy of ligand binding.[44] In one example, the biosynthetic enzyme salutaridine reductase, natively involved in opiate biosynthesis, was redesigned to catalyze the oxidization of 4-hydroxy-2-oxo-pentanoic acid as part of a pathway to produce valerolactone and other C5 feedstock chemicals.[43, 44]

Many library designs are semi-rational, incorporating aspects of targeted changes and aspects of random diversity. Decreasing oligo costs has greatly enabled semi-rational approaches. A 2013 review describes "focused mutagenesis" and "DNA recombination" strategies, both of which are semi-rational.[162] In focused mutagenesis strategies, specific residues are targeted for mutation based on their expected impact on function, stability, or structural flexibility.[113, 163-165] DNA shuffling of homologs is one recombination-based strategy that enables incorporation of diversity from multiple parent genes at non-conserved sites.[69] By 2003, researchers had developed a variety of strategies for producing combinatorial libraries of chimeric genes from multiple templates with relaxed homology requirements.[166] One such method was incremental truncation for the creation of hybrid enzymes, in which aliquots of different sized 5' and 3' gene fragments are taken over time from a slow exonuclease reaction and then ligated to form fusions.[87] Circular permutation, in which the sequence is reordered such that the original N- and C-termini of a protein are linked and a library of new N- and C-termini are created, was published in 2005.[100] Focused mutagenesis techniques include site-saturation mutagenesis, replacing a single position with all possible amino acids; scanning mutagenesis, replacing all positions individual with a specific amino acid; and saturation mutagenesis, in which all positions are replaced individually with all possible amino acids. In 1989, Genentech scientists published two scanning mutagenesis methods, homolog and alanine, where sites within a protein were replaced systematically with either the sites present in homologs or with alanine to identify sites of protein-protein interaction.[167, 168]

In 2001, Diversa was the first to publish the systematic replacement of every position in the entire protein coding sequence with all 20 amino acids with a full or reduced codon set, which they called gene site-saturation mutagenesis.[88, 89] A year later, the commercial QuikChange® Multi site-directed mutagenesis kit became available, enabling combinatorial libraries incorporating 1-20 amino acids at 1-5 sites.[92] Alternatively, the combinatorial library can be generated without a template gene directly by gene assembly from synthetic oligos with desired degeneracy.[91] Today, gene synthesis vendors (for example, IDT, Blue Heron/Eurofins, Genscript, Twist, GeneArt/ThermoFisher) offer several gene library products, including single mixed base, multiple mixed bases, scanning libraries, and truncation libraries. However, cost and turnaround times are typically higher than producing the library using molecular biology techniques, for which recent reviews are available.[162, 169, 170] As the price of oligos falls, the price of gene synthesis and variant libraries falls as well.

In the last decade, extensive efforts, of which we will highlight just a few, focused on multiplexed *in vivo*, continuous, and genome-scale methods to generate and recombine diversity in bacteria and eukaryotes. In 2009, researchers in the Church laboratory published an *in vivo* genome-scale multiplex automated genome engineering (MAGE) technique, developed based on previous recombineering approaches,[171] that they applied to optimize 1-deoxy-D-xylulose-5-phosphate biosynthesis for lycopene production in *E. coli*.[109] The automated set up grows cells, induces expression of λ-Red recombination proteins, prepares cells for electroporation of degenerate oligos with 5' and 3' homology to target genes, electroporates, and then recovers cells for the next cycle. A 5-fold increase in lycopene titers was achieved over 15 automated generation cycles targeting 24 genes in 3 days.[109] The Liu laboratory performed *in vivo* mutagenesis by inducing proofreading suppression and lesion bypass in a host cell infected with bacteriophage carrying the starting gene, and the desired evolved gene activity was linked to phage infectivity such that phage with the desired activity would reproduce.[110] The infected *E. coli* population was maintained in a turbidostat such that the rate of mutation introduction into the phage would greatly exceed the rate of mutation of the *E. coli* host cells. In these experiments, 45-200 rounds of evolution were completed in 1.5-8 days, and evolved an RNA polymerase with significantly altered promoter specificity.

In 2013, Cas9 nucleases were engineered to induce nicking in the genome of cultured human cells at 2 sites simultaneously, directed by a single guide RNA[114] or two guide RNAs, enabling deletions or base changes.[115, 116] In 2017, the Gill laboratory published CRISPR-enabled trackable genome engineering (CREATE), which was used for *in vivo* scanning saturation mutagenesis and licensed to Muse Bio, now Inscripta.[121] The technique involves synthesized oligo library pools (ordered from Agilent Technologies at <$0.001/base pair) with a modular design of a guide RNA and a homology arm that recombines to mutate the target site and introduce a synonymous mutation into the cleavage site to prevent further cleavage. This library is cloned to produce a replicating plasmid library which is transformed into *E. coli* for genome editing of ~50,000 loci with ~75% efficiency. Plasmids can then be sequenced from genome-edited colonies with the desired phenotype and mapped to the genome, providing a phenotype-genotype linkage. Proof of concept was also provided for yeast, and a modified design with

a similar workflow (CRISPR-Cas9- and homology-directed-repair-assisted genome-scale engineering, CHAnGE) was demonstrated by generating an *S. cerevisiae* genome-wide disruption strain collection.[172] To improve the efficiency and precision of CRISPR/Cas9-based genome editing, the Liu laboratory engineered a Cas9 nickase-cytidine deaminase fusion to direct cytidine to thymine changes, and later evolved a dead Cas9-adenine deaminase fusion to direct adenine to guanine changes, both of which were demonstrated in human cell lines.[122, 173] These enzymes are the foundational technology for multiple start-up companies, and modified versions of them have been used for *in vivo* screening of genomic saturation mutagenesis libraries in rice (400 amino acids)[174, 175] and mice (77 amino acids).[176]

3.2 Testing libraries with new technologies and learning for more rapid improvement

Because "you get what you screen for," the best variants in each round of directed evolution of an enzyme are most ideally chosen from screens under conditions identical to the intended process.[2] However, with the many diversity generation techniques available, library sizes vary from less than 100 variants to extremely large (there are over 10 billion combinations of 3 amino acid changes in a 400 amino acid protein). There are many ways to screen or select improved enzyme variants, so we will simply point out a few technological advances over the past two decades with relevance to engineering biosynthetic enzymes for industrial production of natural products, for more examples and detail, see the recent review by Markel, *et. al*.[177]

The first high throughput (HTP) screens used in directed evolution were selections, in which survival of a colony is linked to the desired output, for example antibiotic resistance, and optical readouts (e.g., fluorescent, luminescent, colorimetric), either directly from colonies or from plate-based activity assays. Automation has enabled industrial-scale screening, and, in 2004, U.S. standards were set for the dimensions of 96-, 384-, and 1536-well plates.[96] Direct label-free analytical methods have advanced significantly and have been applied to enzyme screening. In 2004, a matrix-assisted laser-desorption ionization time-of-flight mass spectrometry (MS) method for screening an enzyme against ligands presented on the MALDI surface was published.[95] In 2005, Waters published ultra-high performance liquid chromatography (LC) with <2 µm porous particles, reducing typical LC and LC-MS separation times by ~5-fold without loss of resolution.[102] In 2006, BioTrove (acquired by Agilent in 2011) commercialized the RapidFire MS based on automated microscale solid-phase-extraction.[103, 178] In 2019, Sciex's Echo MS was announced, which is based on nanoliter acoustic droplet ejection combined with an open port interface and was developed as a collaboration between Sciex, Labcyte, and the US DOE.[123, 179] The RapidFire and Echo enable sensitive, fast (0.3-10 seconds/sample), direct detection of analytes from microtiter plates at the scale of ~10^{4-6} samples per run, bringing MS-based detection up to the speed of optical plate readers.[180] AstraZeneca, Labcyte, and Waters have published the use of an acoustic mist ionization interface to MS-based to screen 10^5 samples per day.[181]

Sorting methods directly isolate variants of interest, eliminating the need cherry pick variants that performed well in assays from stored stocks, and reducing the number of steps involving liquid handling with well-plates. Fluorescence activated cell sorting (FACS) via flow cytometry is able to screen 10^5 cells per second and has been used extensively for protein engineering via yeast display.[182] For enzymes, FACS has been used to sort cells expressing fluorescent protein fusions, gel microdroplets co-encapsulating fluorescent substrates and library clones,[183] double emulsion droplets[184, 185] and gel-shell beads[186] containing enzymatic reactions.[187, 188] Fluorescence activated droplet sorting, FADS, was published in 2009 and later used in 2015 to sort 10^7 assays of 10^6 metagenomic hydrolases in 2 hours (10^3 samples per second).[111, 187] Microfluidic droplet sorting has also been published using other detection modalities.[187] Platforms for FADS have been commercialized (e.g., by RainDance Technologies), and Berkeley Lights has developed a massively parallel optical single-cell analysis instrument to sort and measure individual cells which can then be released and cultured.[189] This year, Merck and the Kennedy laboratory published the integration of droplet sorting with MS for enzyme evolution (mass activated droplet sorting, MADS), sorting 10^4 *in vitro* transcription-translation samples in 6 h (0.7 seconds per sample).[124] Kits for coupled *in vitro* transcription-translation have been available for over 20 years.[190] This approach replaces cell culturing, expression induction, and cell lysis with expression from linear or circular DNA in lysates (commonly *E. coli*, wheat germ, or rabbit reticulocytes) shortening the time required to express an enzyme, reducing cellular degradation, and eliminating toxicity to the host cell, but typically yielding smaller amounts and concentrations of protein. MADS is limited to molecules that do not diffuse between droplets and are ionized by electrospray and the number of droplets per second is limited by the scan rate of the mass spectrometer (MS). Overall, current screening methods offer a range of throughputs and information density, with selection being the fastest while giving the least information, followed by fluorescence/optical based sorting methods, followed by direct label-free analytical methods, which are the slowest but most information dense. Even for direct methods, the time required for detection of enzyme activity is often shorter than the time required for reaction incubation.

Today, researchers typically obtain both sequence and activity data for every screened enzyme variant, making it possible to analyze sequence-activity relationships (SAR) for every screened library. Machine learning is applied to develop quantitative SAR models that incorporate data from each library screen into the design of subsequent libraries. An early academic example of a statistical machine learning method was the application of principle component analysis to determine quantitative structure-function relationships from just 15 enzyme mutants.[191] Codexis' ProSAR performs a partial least squares linear regression analysis on the activity and sequences of a library with multiple mutations per sequence to determine the contribution that each makes to the activity and to predict

the activity of potential combinations.[97, 192] The next library is then designed to incorporate selected residues at the most influential positions and variable residues at the other positions. This approach is robust to assay noise and vastly reduces the amount of screening needed – for 40 variable amino acid positions, screening 120 systematically varied proteins is enough to predict activities of the entire theoretical 10^{12}-member library.[97] ProSAR was used for the directed evolution of an *Arthrobacter* ω-transaminase to catalyze a reaction for which it had no initial activity.[193] The enzyme binding pocket was modeled based on a homolog with only 28% sequence identity and used to inform selection of residues to mutate and screening substrates that would be iteratively closer to the structure of the desired substrate. Over 11 rounds of evolution, libraries were built based on residue variation from homologs, random mutagenesis, site saturation mutagenesis, homology structure-informed rational design, and combinatorial incorporation of variants deemed beneficial by ProSAR. The final commercial enzyme converts 25-48 g substrate/g enzyme-day in 50% DMSO at 45 ˚C with complete stereospecificity to the limit of detection.

As described in recent reviews of machine learning in enzyme engineering, linear regression is the simplest model to predict a desired property from a data set of sequenced variants, while kernel methods, such as support vector machines, and decision trees and their ensembles can provide more accurate predictions for $<10^5$ samples.[126, 194, 195] Eight different linear regression algorithms were used to guide DNA 2.0's gene synthesis-based engineering of proteinase K, resulting in 20-fold improved activity after testing <100 variants in two rounds.[106] Support vector machines were used to predict the effects of mutations on protein stability from protein sequence and trained and validated with 77-84% accuracy.[104, 105] A support vector machine was also used to predict whether a protein was highly soluble or aggregation prone with 80% accuracy after being trained on experimental cell-free soluble expression data for a subset of the entire *E. coli* proteome.[112, 196] For larger data sets, artificial neural networks provide better predictions. Many of the methods described in these reviews have been cross-validated against publicly available data to model the effects of mutations on activity, solubility, and stability, but so far few have been implemented in literature to guide directed enzyme evolution at the bench. In a 2013 academic example, chimeras of 3 parent cytochrome P450 enzymes were engineered to improve thermostability by 14˚ C in 9 rounds of optimization testing 65 designed sequences total. The model was a Gaussian process model to predict thermostability and associated uncertainty from sequence with a structural distance-based kernel function, trained on 242 preexisting thermostability and carbon monoxide binding measurements.[120] The first round predicted 20 sequences to minimize uncertainty in the fitness landscape, 17 of which were functional, and the second predicted 10 additional sequences, 9 of which were functional. After these two exploratory rounds, 6 rounds of 5 variants each were designed and tested using an algorithm to trade-off exploitation (most thermostable sequence based on the model) and exploration (further minimize uncertainty), and a final round using an algorithm to choose 5 variants to maximize thermostability. Strain and protein engineering companies are developing in-house SAR databases and applying machine learning algorithms to accelerate directed evolution. Amyris presented machine learning strategies to promote strain candidates and predict performance in scaled-up fermentations at BioDisrupt 2019,[197] Zymergen discussed using their tool Prospector to analyze gene-phenotype relationships in a recent press release,[198] Ginkgo Bioworks platform is based on their "foundry-codebase feedback loop,"[199] and Codexis uses artificial intelligence to predict function from structure and guide directed evolution.[200, 201] Large pharmaceutical companies are using similar strategies for protein, gene, and cell biotherapeutics. For example, Novartis recently announced a partnership with Microsoft for applying artificial intelligence in several areas of drug discovery and development.[202] These predictive models allow researchers to reach targets more quickly with more limited sampling of sequence space.

3.3 New classes of biosynthetic enzymes engineered in the age of sequencing and synthesis

Early directed evolution programs focused on tractable enzymes – easy to express, stable proteins (often bacterial) with easy to screen activities (often antibiotic selections). Conversely, enzymes less favored as starting points for evolution campaigns were large, multimeric, unstable, eukaryotic, membrane-bound, post-translationally modified, and/or had uncommon or expensive cofactors or chemistries. The overall decreased cost and increased speed of the design-build-test-learn cycle for protein engineering means that engineers can take more shots on goal, employing a diversity of library designs for more rounds of evolution. As mentioned earlier, machine learning approaches enable these faster cycles to also be smarter, taking enzymes from wild-type to process relevant at lower cost using semi-rational strategies. Smarter,

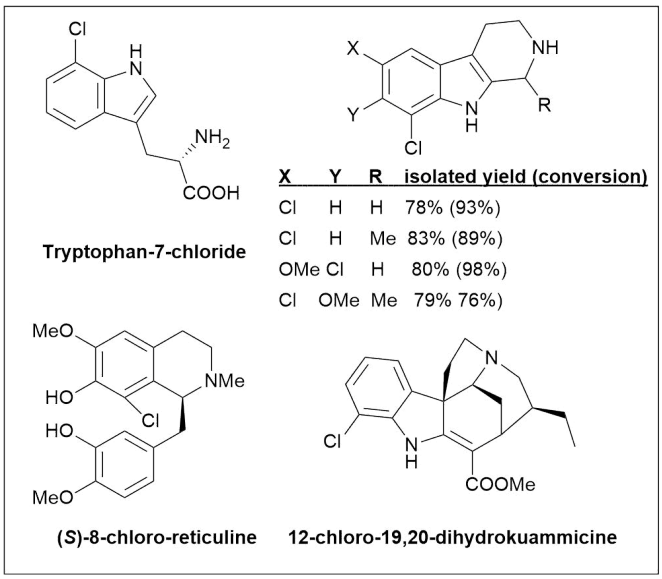


Fig. 7 Haloalkaloids natively produced by RebH (tryptophan-7-Cl), produced by engineered RebH (right), and produced by feeding tryptophan analogs to an engineered strain (8-Cl-reticuline).

faster enzyme engineering approaches mean that less tractable enzymes can now be evolved, and increasingly large SAR datasets are produced. Here, we will present three brief examples of engineered biosynthetic enzymes – (1) two different soil bacteria halogenases, which catalyze the rare chemistry of carbon-halogen bond formation and have been explored for industrial use, (2) the fungal enzyme LovD, evolved to use a small acyl carrier rather than an acyl carrier domain and to produce the natural product analog and blockbuster drug simvastatin, and (3) the plant enzyme PyKS, evolved for expression and thermostability.

Enzymes catalyzing the rare chemistry of halogenation have now been evolved by several research groups. Halogenases are mechanistically and structurally diverse enzymes that catalyze the formation of carbon-halogen bonds.[203, 204] The flavin-dependent tryptophan-7-halogenase RebH from soil bacteria catalyzes the first step in the biosynthesis of rebbecamycin, an indolocarbazole alkaloid (Fig. 7).[205] For biocatalytic purposes, the RebH enzyme was combined with its flavin reductase partner RebF and glucose and glucose dehydrogenase as a source of reducing equivalents.[49, 206] RebH and a single mutant of RebH have been co-expressed with RebF to generate chlorinated monoterpene indole alkaloids found in periwinkle, such as 12-chloro-19,20-dihydrokuammicine (Fig. 7).[207, 208] Three rounds of random mutagenesis by epPCR and screening against progressively larger substrates resulted in an enzyme variant that was able to chlorinate the non-halogenated analog of deformylflustrabromine, a marine natural product that inhibits biofilm formation, and eight other alkaloids at 10 mg scale, despite no initial activity on the substrate (Fig. 7).[206] Novartis and the Lewis laboratory collaborated to examine the substrate specificity of several wild-type and engineered flavin-dependent halogenases, signifying the reaction's industrial relevance.[50] As part of organofluorine natural product biosynthesis in *Streptomyces*, the hexameric 5'-fluoro-5'-deoxyadenosine synthase (FDAS) catalyzes the reaction between a fluoride ion and SAM and releases methionine.[209] FDAS has been evolved through a single round of mutagenesis and screening to improve its activity on 5'-chloro-5'-deoxyadenosine to install a radiolabeled fluorine for positron emission tomography (PET) imaging.[210] Halogens have been incorporated into complex natural products by precursor feeding, including a methyltriketide lactone in *E. coli*[211] and benzylisoquinoline alkaloids such as 8-chloro-reticuline (Fig. 7) in an *S. cerevisiae* strain with 30 introduced genes[212] and could be incorporated without feeding with an appropriately engineered halogenase.

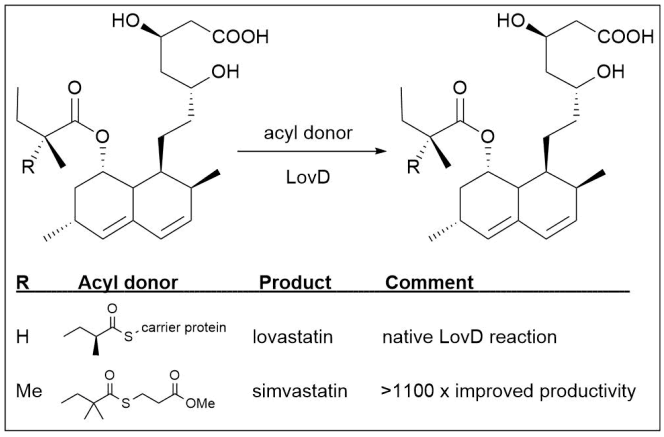


Fig. 8 Native and engineered activities of the fungal polyketide transesterase LovD.

An industrial example of an engineered eukaryotic enzyme is the evolution of *Aspergillus terreus* transesterase LovD (Fig. 8). LovD natively interacts with the lovastatin PKS acyl carrier protein domain and catalyzes the transfer of a 2-methylbutyrate side chain to itself and then to monacolin J acid to produce lovastatin. By using homology modeling and machine learning (ProSAR, described earlier) to design libraries, LovD was engineered to instead accept 2,2-dimethylbutyrate from a small acyl carrier and commercially produce the drug simvastatin.[45, 46] This demonstrated the disruption of a protein-protein interaction and a change in substrate specificity.

A recent academic example of evolving a plant biosynthetic enzyme using modern DNA synthesis, HTP sequencing, and computationally informed semi-rational strategies is the engineering of the plant type III PKS pyrrolidine ketide synthase (PyKS).[213] PyKS is involved in the biosynthesis of tropane alkaloids, which include the anti-spasmodic hyoscamine, anti-nausea scopolamine, and the narcotic stimulant cocaine.[214] The researchers generated a scanning saturation mutagenesis (each position mutated to all possible residues individually) library of a PyKS-green fluorescent protein fusion with a theoretical size of >8500, screened the library expressed in *E. coli* by FACS and sequenced the top 5-10% of mutants to determine which mutations resulted in increased fluorescence, a proxy for stability. Based on a protein structure homology model, they then filtered these 1115 stabilizing mutations to 116 unlikely to affect activity using a computational filter for evolutionary conservation, distance to active site, degree of burial in protein core, and that disallowed proline substitution.[215] Of these, 21 were incorporated into a combinatorial library with a theoretical size of $2x10^6$, which was sorted by FACS, ultimately obtaining a PyKS with >10-fold improved expression and >10 ˚C improvement in melting temperature and comparable activity and affinity to wild-type.[213] This enzyme could be incorporated into an engineered microbial biosynthesis of tropane alkaloids.[216] Enzyme engineering is now able to improve difficult-to-work with enzymes, using machine learning guided semi-rational strategies, at an industrially relevant pace,[217] to produce fit-for-process biocatalysts.

4 Engineering biosynthetic enzymes in context

4.1 Considerations for choosing a biocatalytic context

Today, industrial R&D at many companies focuses on incorporating synthetic chemistry, biocatalysis, and metabolic engineering approaches to obtain the most economical route to a given product at the appropriate scale. Natural products are produced commercially by methods that encompass fermentation of native and non-native strains, extraction from

agricultural feedstocks, synthetic and/or biocatalytic derivatization of fermented or extracted material, *de novo* total synthesis, and chemo-enzymatic synthesis. Thus, engineered biosynthetic enzymes are produced in native and non-native hosts and used *in vitro* in single steps, with cofactor recycling, in chemo-enzymatic routes, and in one-pot cascade reactions. As always, process chemistry requires careful cost-benefit calculations to determine the best route to a given target amidst an ever-increasing landscape of possibilities. Other reviews have covered the tradeoffs between chemical and biocatalysis and summarized advances in biocatalytic cascades.[218-220]

Even after deciding to perform a given reaction enzymatically, many options exist for how to produce and use the enzyme, and the tradeoffs involved in these options are summarized in Table 3. Single enzyme steps are quick to develop, relatively cheap, and can be engineered to reach high productivities under harsh conditions. However, they can only catalyze one chemical transformation, so the product must be significantly more valuable than the substrate. On the other extreme, fermentation has lower tolerance for harsh conditions and typically has lower productivities, titers, and yields, but can use very inexpensive starting materials. Multi-enzyme cascades span from three enzymes to perform multiple linear steps and/or recycle a cofactor to entire synthetic metabolisms built to use cheap carbon sources.[221]

.

Table 3 Considerations for selecting biocatalyst context.

	Single step	Multi-enzyme cascade	Whole cell biocatalysis	Cell free metabolism	Fermentation[a]
Number of genes expressed heterologously	1-2 (may include cofactor recycle)	1-10	0-30		
Number of genes to be engineered	1	1-10	1-genome		
Process condition limitations	<95 ˚C, <50% organic cosolvent		Within tolerance of organism's ability to regenerate cofactors	Within tolerance of cell free metabolism enzymes	Within tolerance of organism's ability to grow
Recyclability	Low-high	Low-medium	Low-medium	Low	None
Starting material	Chemical intermediate, co-substrate, cofactor	Chemical intermediate, co-substrate, catalytic amounts of cofactor(s)	Chemical intermediate, co-substrate	Simple carbon source/biomass, may include co-substrate, catalytic amounts of cofactor(s)	Simple carbon source/biomass, media may be supplemented
Time to develop	Fastest	Medium, depends on number of new/engineered enzymes	Depends on number of genes targeted and mutation rate	Depends on number of new/engineered enzymes	Depends on number of genes targeted and mutation rate
Factors contributing to enzyme/strain production cost	Volumetric productivity of enzyme in fermentation, enzyme lifetime, turnover number, and any purification steps needed	Same as single step for each enzyme	Depends on volumetric productivity of cells in fermentation, cell lifetime, and turnover number	Same as single step for each enzyme produced by fermentation or cell free protein synthesis	Depends on product titer, yield on feedstock, and volumetric productivity
Substrate load and product titer	High (10s-100s g/L)		Medium (1-100 g/L)	Low-medium (1-10 g/L)	Low (mgs - 10 g/L)
Yield	High (80-100%)		High (80-100%)	Medium (60-95%)	Low (theoretical maximum yield varies, typically <50%)
Productivity	1-30 g/L/h		1-10 g/L/h	<1 g/L/h	
Factors contributing to product purification cost	Removal of buffer salts, byproducts, and catalytic amounts of protein(s), cofactor, cosubstrate		Removal of buffer salts, byproducts, and catalytic amounts of cells, cofactor,	Same as single step and multi-enzyme cascade, with complexity	Removal of media, byproducts, cells, extracellular protein

		cosubstrate	increasing with pathway length	

[a]Fermentation necessarily implies a microbial heterologous host. Other hosts, such as plants, could also be a context for (an) engineered enzyme(s).

As cascades and engineered strains become more common, retrobiosynthesis is emerging to design new-to-nature biosynthetic pathways to achieve atom economy (minimize waste by maximizing mass efficiency and theoretical yield) and step economy (minimize the number of synthetic steps).[222-224] As mentioned in section 3, directed evolution of an enzyme or strain is most ideally performed under conditions identical to those of the intended process.[2] In this section, our goal is to highlight examples of biosynthetic enzymes engineered for a specific *in vitro* or *in vivo* process context to produce commercial or near commercial natural molecules and analogs.

4.2 Engineering enzymes in heterologous and native hosts

Targeted directed evolution of biosynthetic pathways has been performed in the context of native producing microorganisms. For example, in an extension of efforts to produce novel andrimid-related antibiotics (Fig. 2) by directed evolution of the NRPS-PKS,[68] the AdmK domain, which incorporates the valyl subunit, was engineered.[225] The three least evolutionarily conserved sites were targeted with saturation mutagenesis and colonies were grown, multiplexed, and screened by liquid chromatography-tandem MS, which improved screening throughput and ability to identify novel chemical structures relative to zone-of-inhibition assays. Novel andrimid derivatives with increased potency against Gram-negative and Gram-positive bacterial strains were identified, and titers comparable to wild-type were produced by supplementing the media with amino acids.

In one of the first published examples of combining protein and metabolic engineering in a heterologous host, researchers in the Prather and Stephanopoulos groups applied directed evolution to two plant-derived enzymes in an *E. coli* strain engineered to produce the diterpenoid precursor of ginkgolides from *Ginkgo biloba*.[226] Based on a homology structural model of levopimaradiene synthase (LPS), 15 mutations from paralogous enzymes were introduced into the substrate binding pocket. The mutants were screened in the context of the engineered strain, and positions with mutations that resulted in higher total or altered distribution of diterpenoids were targeted by saturation mutagenesis. GGDPS was also evolved in the strain context with the LPS replaced by a two-enzyme lycopene-producing pathway so that an epPCR-generated random library could be screened by eye for lycopene formation. The engineered strain with double mutants of both LPS and GGDPS resulted in 700 mg/L levopimaradiene, compared to 0.15 mg/L before enzyme engineering.

As described in section 3, there has been enormous progress in multiplex genome editing tools to make both targeted and untargeted changes, and these have seen increasing application in natural product discovery.[227] In an application to natural product biosynthesis, a team led by Keasling and Jensen developed Cas9-mediated protein evolution reaction (CasPER) to evolve enzymes in the host strain genomic context, elevating isoprenoid production >10-fold.[228] Massively parallel editing with markerless *in vivo* mutagenesis is possible, as are genome-wide knockouts or promoter swaps, to identify hundreds of beneficial genomic changes.[229, 230] In one project, Inscripta identified and combined three such variants, yielding >10,000-fold increases in lysine titers. *In vivo* mutagenesis has also accelerated development of continuous evolution strategies, in which evolution is conducted by *in vivo* mutagenesis, selection, and amplification, rather than round-by-round. These approaches have been recently reviewed.[231-233] We anticipate near-term publication of studies producing natural products and analogs with native and heterologous hosts engineered by multiple rounds of machine learning-guided multiplexed targeted genome-scale engineering with development times that vastly outpace semi-synthetic genome shuffling.

4.3 Engineering commercial enzymes in multiple contexts

In this section, we will present three brief case studies of enzyme engineering in the context of *in vitro* cascades and, for two examples, competing *in vivo* processes – (1) a 9-enzyme cascade reaction to produce the HIV inhibitor islatravir, (2) multiple industrial approaches to produce the natural stevia sweetener rebaudioside M (RebM), and (3) multiple pre-commercial approaches to produce cannabinoids.

Recently, Merck and Codexis reported a nine-enzyme cascade reaction with five evolved enzymes to produce the HIV inhibitor islatravir, a nucleoside analog, in three linear steps.[234] For the first two linear steps, galactose oxidase and pantothenate kinase were each evolved in the presence of their auxiliary enzymes, and for the third step purine nucleoside phosphorylase was evolved in the presence of previously evolved phosphopentomutase. To enable separation of the protein catalysts, 3 of the enzymes were immobilized. Overall, 5 enzymes were evolved for stereoselectivity, activity, or by-product tolerance, and 51% yield was achieved in a one-pot sequential process.

Synthesis of the natural stevia sweetener rebaudioside M (RebM, Fig. 9) has been achieved through multiple examples of modern industrial biocatalysis. FDA generally regarded as safe (GRAS) notifications for RebM include processes to produce the molecule by extraction of specially bred stevia plant,[235] whole cell bioconversion from stevia plant extract with engineered *Pichia pastoris*,[236] production from fermentation of sugar using engineered *Saccharomyces cerevisiae* or *Yarrowia lipolytica*,[237-239] and *in vitro* enzyme cascade conversions from stevia extract using enzymes expressed in *E. coli*.[240-242] To compare the fermentation and cell-free cascade approaches, consider the processes developed by Amyris and Codexis. Amyris, has developed an *S. cerevisiae* strain for the production of RebM,[239] and the resulting food ingredient Purecane™ was commercialized in late 2018 with Raízen and ASR Group. As

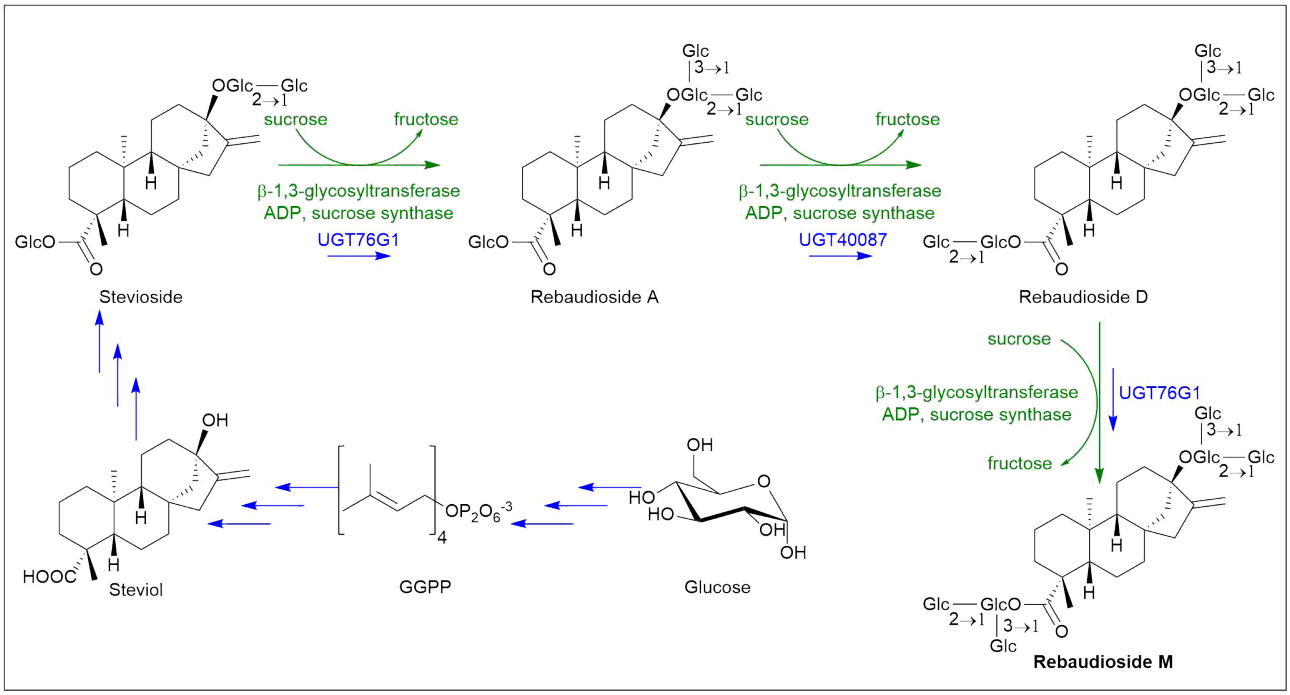


Fig. 9 Amyris *in vivo* route to rebaudioside M, in blue, and Codexis *in vitro* route, in green. GGPP, geranyl-geranyl pyrophosphate; ADP, adenine dinucleoside phosphate. Although the route is shown as linear, it is a metabolic matrix, in which the specificities of the enzymes are engineered such that rebaudioside M is the final product.

disclosed in patent literature, the Amyris process leverages their experience developing yeast strains to produce terpenoids. The major technical challenge was conversion of rebaudioside A (RebA) to rebaudioside D (RebD), a 1,2'-glucosylation. They screened over 100 UDP-dependent glycosyltransferase genes, codon optimized for yeast, in a RebA producing strain. Four were selected as parents for constructing N-terminal domain and loop chimeras, a rational strategy to influence substrate specificity.[52]

The Codexis-Tate & Lyle partnership has pursued a 3-enzyme cascade reaction to produce RebM starting from stevia plant extract containing a mixture of less glycosylated steviol glycosides. Codexis provided proof of concept of the enzymatic synthesis in 2015, and by 2018 Tate & Lyle commercialized the sweetener ingredient. The three enzymes were highly engineered in context of the cascade, with all assays including at least two enzymes, and were evolved toward multiple criteria in tandem (e.g., thermostability, activity, increasing substrate concentration), over the course of 23 disclosed rounds of evolution of a β-1,3-glycosyltransferase, 19 rounds of a β-1,2-glycosyltransferase, and 16 rounds of a sucrose synthase.[51] Instead of using an organism's metabolism to provide the sugar donor, the cascade reaction uses sucrose as an inexpensive sugar donor and sucrose synthase to recycle the cofactor. Since cost is proprietary, it remains unclear which process will be more successful in the marketplace.

Currently, there is a race to produce cannabinoids, polyketide-terpenoid natural products formed from fatty acid and isoprenoid precursors. At the end of 2019, Amyris shipped the first fermentation-derived cannabinoid, produced from genetically engineered yeast.[243] Researchers at UC Berkeley, LBNL, ETH, and Demetrix published an engineered *S. cerevisiae* pathway for the synthesis of cannabinoids with milligrams per liter titers.[244] In an effort supported by Bioactive Ingredients Corporation, the Gonzalez group produced the cannabinoid precursor olivetolic acid at 80 mg/L in engineered *E. coli* with glycerol and added hexanoate[245] and has since identified a prenyltransferase (PT) mutant that enables milligrams per liter production of prenylated cannabinoids.[55] The Bowie group achieved 67-92% prenylation of two polyketide cannabinoid precursors from glucose at titers up to 1.74 g/L, which exceed cell toxicity limits. Their production platform consists of 27 purified enzymes and the authors formed Invizyne Technologies to commercialize cell free processes.[246] The cell free prenylation was achieved with an engineered bacterial PT, rather than the integral membrane PT from *Cannabis sativa*, and by performing additional optimization of the cell free pathway and process conditions.[56] The *Streptomyces* aromatic PT binding site was engineered by first screening a 22 construct library designed using Rosetta to predict mutations that would improve substrate binding and then constructing 7 additional constructs based on the results, increasing activity more than 100-fold. Among other competitors, Ginkgo Bioworks and Intrexon are targeting a production cost of <$1/g.[247-250] Amyris, Ginkgo, and Intrexon all have active partnerships to develop and commercialize cannabinoid-producing microbial strains.

5 Conclusions and outlook

Natural products have a critical role in the growing bioeconomy. As society searches for molecules that are environmentally sustainable, performance enhanced, and cost effective, enzymes, including those from secondary metabolism, will play a role in many synthetic solutions. The development of enzyme

engineering platforms that leverage advances in DNA synthesis and sequencing, HTP analytical methods, and machine learning has enabled commercial success of engineered biosynthetic enzymes. We expect to see these platforms become increasingly predictive, faster, and able to target improvements at pathway and genome scales, as well as increasingly economic due to miniaturization and multiplexing. We have presented case studies of engineered biosynthetic enzymes used to produce natural products and analogs in a variety of contexts, as well as a framework for considering the tradeoffs between single enzyme steps, multi-enzyme cascades, and fermentation routes. As different bioengineering approaches compete in the marketplace over a larger range of molecular targets, the conditions for viability of each will be established.

6 Conflicts of interest

David Entwistle is an employee of Codexis, Inc., and Stephanie Galanie and James Lalonde are former employees of Codexis. James Lalonde is a current employee of Inscripta, Inc. All authors are inventors on patents related to engineering enzymes and strains.

7 Acknowledgements

Stephanie Galanie is funded by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy (DOE), and by the Center for Bioenergy Innovation (CBI, FWP ERKP886), a U.S. DOE Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. The authors thank Timothy J. Tschaplinski, Mircea Podar, Ryan Salvador, and the reviewers for their comments on this manuscript.

8 References

1. A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts and B. Witholt, *Nature*, 2001, **409**, 258-268.
2. C. Schmidt-Dannert and F. H. Arnold, *Trends Biotechnol.*, 1999, **17**, 135-136.
3. U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore and K. Robins, *Nature*, 2012, **485**, 185-194.
4. M. S. Packer and D. R. Liu, *Nat. Rev. Genet.*, 2015, **16**, 379-394.
5. C. Zeymer and D. Hilvert, *Annu. Rev. Biochem.*, 2018, **87**, 131-157.
6. T. Classen and J. Pietruszka, *Bioorg. Med. Chem.*, 2018, **26**, 1285-1303.
7. S. Friedrich and F. Hahn, *Tetrahedron*, 2015, **71**, 1473-1508.
8. N. Tibrewal and Y. Tang, *Annu. Rev. Chem. Biomol. Eng.*, 2014, **5**, 347-366.
9. E. Kim, B. S. Moore and Y. J. Yoon, *Nat. Chem. Biol.*, 2015, **11**, 649-659.
10. J. V. Pham, M. A. Yilma, A. Feliz, M. T. Majid, N. Maffetone, J. R. Walker, E. Kim, H. J. Cho, J. M. Reynolds, M. C. Song, S. R. Park and Y. J. Yoon, *Front Microbiol*, 2019, **10**, 1404-1404.
11. J. W. Lee, D. Na, J. M. Park, J. Lee, S. Choi and S. Y. Lee, *Nat. Chem. Biol.*, 2012, **8**, 536-546.
12. A. G. Atanasov, B. Waltenberger, E.-M. Pferschy-Wenzig, T. Linder, C. Wawrosch, P. Uhrin, V. Temml, L. Wang, S. Schwaiger, E. H. Heiss, J. M. Rollinger, D. Schuster, J. M. Breuss, V. Bochkov, M. D. Mihovilovic, B. Kopp, R. Bauer, V. M. Dirsch and H. Stuppner, *Biotechnol Adv*, 2015, **33**, 1582-1614.
13. A. Cravens, J. Payne and C. D. Smolke, *Nat. Comm.*, 2019, **10**, 2142.
14. P. Bernhardt and S. E. O'Connor, *Curr. Opin. Chem. Biol.*, 2009, **13**, 35-42.
15. S. E. O'Connor, *Annual Review of Genetics*, 2015, **49**, 71-94.
16. C. Li, R. Zhang, J. Wang, L. M. Wilson and Y. Yan, *Trends Biotechnol.*, 2020, DOI: 10.1016/j.tibtech.2019.12.008.
17. J. Cortes, S. F. Haydock, G. A. Roberts, D. J. Bevitt and P. F. Leadlay, *Nature*, 1990, **348**, 176-178.
18. S. Donadio, M. Staver, J. McAlpine, S. Swanson and L. Katz, *Science*, 1991, **252**, 675-679.
19. A. Lawen and R. Zocher, *J. Biol. Chem.*, 1990, **265**, 11355-11360.
20. G. Weber and E. Leitner, *Curr. Genet.*, 1994, **26**, 461-467.
21. D. E. Cane, C. T. Walsh and C. Khosla, *Science*, 1998, **282**, 63-68.
22. M. T. Reetz and K.-E. Jaeger, in *Topics in Current Chemistry*, ed. W.-D. Fessner, Springer Verlag, Berlin, 1999, vol. 200, pp. 31-57.
23. R. McDaniel, A. Thamchaipenet, C. Gustafsson, H. Fu, M. Betlach and G. Ashley, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 1846-1851.
24. K. Patel, M. Piagentini, A. Rascher, Z. Q. Tian, G. O. Buchanan, R. Regentin, Z. Hu, C. R. Hutchinson and R. McDaniel, *Chem. Biol.*, 2004, **11**, 1625-1633.
25. L. Tang, L. Chung, J. R. Carney, C. M. Starks, P. Licari and L. Katz, *J. Antibiot.*, 2005, **58**, 178-184.
26. L. S. Sheehan, R. E. Lill, B. Wilkinson, R. M. Sheridan, W. A. Vousden, A. L. Kaja, G. D. Crouse, J. Gifford, P. R. Graupner, L. Karr, P. Lewer, T. C. Sparks, P. F. Leadlay, C. Waldron and C. J. Martin, *J. Nat. Prod.*, 2006, **69**, 1702-1710.
27. M. A. Gregory, H. Petkovic, R. E. Lill, S. J. Moss, B. Wilkinson, S. Gaisser, P. F. Leadlay and R. M. Sheridan, *Angew. Chem., Int. Ed.*, 2005, **44**, 4757-4760.
28. K. T. Nguyen, D. Ritz, J. Q. Gu, D. Alexander, M. Chu, V. Miao, P. Brian and R. H. Baltz, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17462-17467.
29. K. T. Nguyen, X. He, D. C. Alexander, C. Li, J. Q. Gu, C. Mascio, A. Van Praagh, L. Mortin, M. Chu, J. A. Silverman, P. Brian and R. H. Baltz, *Antimicrob. Agents Chemother.*, 2010, **54**, 1404-1413.
30. L. Tao, J. Wilczek, J. M. Odom and Q. Cheng, *Metab. Eng.*, 2006, **8**, 523-531.
31. Y.-X. Zhang, K. Perry, V. A. Vinci, K. Powell, W. P. C. Stemmer and S. B. del Cardayré, *Nature*, 2002, **415**, 644-646.
32. K. Maeda, J. M. Luengo, O. Ferrero, S. Wolfe, M. Y. Lebedev, A. Fang and A. L. Demain, *Enzyme Microb. Technol.*, 1995, **17**, 231-234.
33. DSM N.V., US Pat., 6,020,151, 1996.
34. C. L. Wei, Y. B. Yang, W. C. Wang, W. C. Liu, J. S. Hsu and Y. C. Tsai, *Appl. Environ. Microbiol.*, 2003, **69**, 2306-2312.

35. J. S. Hsu, Y. B. Yang, C. H. Deng, C. L. Wei, S. H. Liaw and Y. C. Tsai, *Appl. Environ. Microbiol.*, 2004, **70**, 6257-6263.
36. C. L. Wei, Y. B. Yang, C. H. Deng, W. C. Liu, J. S. Hsu, Y. C. Lin, S. H. Liaw and Y. C. Tsai, *Appl. Environ. Microbiol.*, 2005, **71**, 8873-8880.
37. C. A. Cantwell, R. J. Beckmann, J. E. Dotzlaf, D. L. Fisher, P. L. Skatrud, W.-K. Yeh and S. W. Queener, *Curr. Gen.*, 1990, **17**, 213-221.
38. L. Crawford, A. M. Stepan, P. C. McAda, J. A. Rambosek, M. J. Confer, V. A. Vinci and C. D. Reeves, *Bio/Technology*, 1995, **13**, 58-62.
39. S. Gaisser, L. Kellenberger, A. L. Kaja, A. J. Westoa, R. E. Lill, G. Wirtz, S. G. Kendrew, L. Low, R. M. Sheridan, B. Wilkinson, I. S. Galloway, K. Stutzman-Engwall, H. A. I. McArthur, J. Staunton and P. F. Leadlay, *Org. Biomol. Chem.*, 2003, **1**, 2840-2847.
40. K. Stutzman-Engwall, S. Conlon, R. Fedechko, F. Kaczmarek, H. McArthur, A. Krebber, Y. Chen, J. Minshull, S. A. Raillard and C. Gustafsson, *Biotechnol. Bioeng.*, 2003, **82**, 359-369.
41. K. Stutzman-Engwall, S. Conlon, R. Fedechko, H. McArthur, K. Pekrun, Y. Chen, S. Jenne, C. La, N. Trinh, S. Kim, Y. X. Zhang, R. Fox, C. Gustafsson and A. Krebber, *Metab. Eng.*, 2005, **7**, 27-37.
42. P. Sun, Q. F. Zhao, F. T. Yu, H. Zhang, Z. H. Wu, Y. Y. Wang, Y. Wang, Q. L. Zhang and W. Liu, *J. Am. Chem. Soc.*, 2013, **135**, 1540-1548.
43. Arzeda Corp., US Pat., 9,523,105, 2013.
44. Arzeda Corp., US Pat., 10,025,900, 2016.
45. X. Gao, X. Xie, I. Pashkov, M. R. Sawaya, J. Laidman, W. Zhang, R. Cacho, T. O. Yeates and Y. Tang, *Chem. Biol.*, 2009, **16**, 1064-1074.
46. G. Jiménez-Osés, S. Osuna, X. Gao, M. R. Sawaya, L. Gilson, S. J. Collier, G. W. Huisman, T. O. Yeates, Y. Tang and K. N. Houk, *Nat. Chem. Biol.*, 2014, **10**, 431-436.
47. Codexis Selected for Third Presidential Green Chemistry Award in Seven Years, https://www.businesswire.com/news/home/20120618005503/en/Codexis-Selected-Presidential-Green-Chemistry-Award-Years, (accessed December 2019).
48. E. Yeh, S. Garneau and C. T. Walsh, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 3960-3965.
49. J. T. Payne, M. C. Andorfer and J. C. Lewis, *Angew. Chem., Int. Ed.*, 2013, **52**, 5271-5274.
50. M. C. Andorfer, J. E. Grob, C. E. Hajdin, J. R. Chael, P. Siuti, J. Lilly, K. L. Tan and J. C. Lewis, *ACS Catal.*, 2017, **7**, 1897-1904.
51. Codexis, Inc., US Pat., 20180223264A1, 2018.
52. Amyris, Inc., US Pat., 20190185826A1, 2019.
53. Codexis, Inc., Tate & Lyle Ingredients Americas LLC, US Pat., 20200032227A1, 2020.
54. T. Kuzuyama, J. P. Noel and S. B. Richard, *Nature*, 2005, **435**, 983-987.
55. S. Qian, J. M. Clomburg and R. Gonzalez, *Biotechnol. Bioeng.*, 2019, **116**, 1116-1127.
56. M. A. Valliere, T. P. Korman, N. B. Woodall, G. A. Khitrov, R. E. Taylor, D. Baker and J. U. Bowie, *Nat. Commun.*, 2019, **10**, DOI: 10.1038/s41467-41019-08448-y.
57. Bristol-Myers Squibb to Acquire Kosan Biosciences, https://news.bms.com/press-release/financial-news/bristol-myers-squibb-acquire-kosan-biosciences, (accessed November 2019).
58. A. F. A. Marsden, B. Wilkinson, J. Cortés, N. J. Dunster, J. Staunton and P. F. Leadlay, *Science*, 1998, **279**, 199-202.
59. P. F. Long, C. J. Wilkinson, C. P. Bisang, J. Cortés, N. Dunster, M. Oliynyk, E. McCormick, H. McArthur, C. Mendez, J. A. Salas, J. Staunton and P. F. Leadlay, *Mol. Microbiol.*, 2002, **43**, 1215-1225.
60. T. C. Sparks, G. D. Crouse, J. E. Dripps, P. Anzeveno, J. Martynow, C. V. DeAmicis and J. Gifford, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 393-401.
61. U. Galm and T. C. Sparks, *J. Ind. Microbiol. Biotechnol.*, 2016, **43**, 185-193.
62. Dow Agrosciences LLC, US Pat., 20110281359A1, 2011.
63. S. J. Rehm, H. Boucher, D. Levine, M. Campion, B. I. Eisenstein, G. A. Vigliani, G. R. Corey and E. Abrutyn, *J. Antimicrob. Chemother.*, 2008, **62**, 1413-1421.
64. R. H. Baltz, *Nat. Biotechnol.*, 2006, **24**, 1533-1540.
65. Merck to Acquire Cubist Pharaceuticals for $102 Per Sahre in Cash, https://www.mrknewsroom.com/news-release/corporate-news/merck-acquire-cubist-pharmaceuticals-102-share-cash, (accessed March 2020).
66. B. S. Moore and C. Hertweck, *Nat. Prod. Rep.*, 2002, **19**, 70-99.
67. R. H. Baltz, *J. Ind. Microbiol. Biotechnol.*, 2018, **45**, 635-649.
68. M. A. Fischbach, J. R. Lai, E. D. Roche, C. T. Walsh and D. R. Liu, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 11951-11956.
69. W. P. C. Stemmer, *Nature*, 1994, **370**, 389-391.
70. C. Schmidt-Dannert, D. Umeno and F. H. Arnold, *Nat. Biotechnol.*, 2000, **18**, 750-753.
71. R. H. Baltz, *J. Gen. Microbiol.*, 1978, **107**, 93-102.
72. D. A. Hopwood and H. M. Wright, *Mol. Gen. Genet.*, 1978, **162**, 307-317.
73. R. P. Elander, *Appl. Microbiol. Biotechnol.*, 2003, **61**, 385-392.
74. DSM N.V., US Pat., 6,518,039, 1998.
75. K. Brown, *Technol. Rev.*, 2000, **103**, 84-89.
76. Maxygen, Inc., US Pat., 6,579,678, 2000.
77. Maxygen, Inc., US Pat., 6,716,631, 2000.
78. Maxygen, Inc., US Pat., 7,795,030, 2008.
79. S. Lutz and W. M. Patrick, *Curr. Opin. Biotechnol.*, 2004, **15**, 291-297.
80. M. Goldsmith and D. S. Tawfik, in *Methods in Protein Design*, ed. A. E. Keating, Academic Press, San Diego, CA, First edn., 2013, vol. 523, ch. 12, pp. 257-283.
81. E. Zazopoulos, K. Huang, A. Staffa, W. Liu, B. O. Bachmann, K. Nonaka, J. Ahlert, J. S. Thorson, B. Shen and C. M. Farnet, *Nat. Biotechnol.*, 2003, **21**, 187-190.
82. E. Mathur, G. Toledo, B. Green, M. Podar, T. Richardson, M. Kulwiec and H. Chang, *Ind. Biotechnol.*, 2005, **1**, 283-287.
83. A. I. Solbak, T. H. Richardson, R. T. McCann, K. A. Kline, F. Bartnek, G. Tomlinson, X. Tan, L. Parra-Gessert, G. J. Frey, M. Podar, P. Luginbühl, K. A. Gray, E. J. Mathur, D. E. Robertson, M. J. Burk, G. P. Hazlewood, J. M. Short and J. Kerovuo, *J. Biol. Chem.*, 2005, **280**, 9431-9438.
84. D. E. Robertson, J. A. Chaplin, G. DeSantis, M. Podar, M. Madden, E. Chi, T. Richardson, A. Milan, M. Miller, D. P. Weiner, K. Wong, J. McQuaid, B. Farwell, L. A. Preston, X. Tan, M. A. Snead, M. Keller, E. Mathur, P. L. Kretz, M. J. Burk and J. M. Short, *Appl. Environ. Microbiol.*, 2004, **70**, 2429-2436.
85. K. Chen and F. H. Arnold, *Bio/Technology*, 1991, **9**, 1073-1077.

86. Blue Heron Biotech About Us, https://www.blueheronbio.com/, (accessed March 2020).
87. M. Ostermeier, J. H. Shim and S. J. Benkovic, *Nat. Biotechnol.*, 1999, **17**, 1205-1209.
88. Kevin A. Gray, Toby H. Richardson, K. Kretz, Jay M. Short, F. Bartnek, R. Knowles, L. Kan, Paul E. Swanson and Dan E. Robertson, *Adv Synth Catal*, 2001, **343**, 607-617.
89. Diversa Corporation, US Pat., 6,171,820, 1999.
90. A. Armon, D. Graur and N. Ben-Tal, *J. Mol. Biol.*, 2001, **307**, 447-463.
91. J. E. Ness, S. Kim, A. Gottman, R. Pak, A. Krebber, T. V. Borchert, S. Govindarajan, E. C. Mundorff and J. Minshull, *Nat. Biotechnol.*, 2002, **20**, 1251-1255.
92. H. H. Hogrefe, J. Cline, G. L. Youngblood and R. M. Allen, *BioTechniques*, 2002, **33**, 1158-1165.
93. J. Tian, H. Gong, N. Sheng, X. Zhou, E. Gulari, X. Gao and G. Church, *Nature*, 2004, **432**, 1050-1054.
94. C. Gustafsson, S. Govindarajan and J. Minshull, *Trends Biotechnol*, 2004, **22**, 346-353.
95. D.-H. Min, W.-J. Tang and M. Mrksich, *Nat. Biotechnol.*, 2004, **22**, 717-723.
96. New Microplate Standards Expected to Accelerate & Streamline Industry, https://www.ansi.org/news_publications/news_story?menuid=7&articleid=4635c896-f02c-40c2-8de0-b886fdc9c54d, (accessed March 2020).
97. R. Fox, A. Roy, S. Govindarajan, J. Minshull, C. Gustafsson, J. T. Jones and R. Emig, *Protein Eng., Des. Sel.*, 2003, **16**, 589-597.
98. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard and D. Baker, *Science*, 2008, **319**, 1387-1391.
99. D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik and D. Baker, *Nature*, 2008, **453**, 190-195.
100. Z. Qian and S. Lutz, *J. Am. Chem. Soc.*, 2005, **127**, 13466-13467.
101. K. V. Voelkerding, S. A. Dames and J. D. Durtschi, *Clin. Chem.*, 2009, **55**, 641-658.
102. J. R. Mazzeo, U. D. Neue, M. Kele and R. S. Plumb, *Anal. Chem.*, 2005, **77**, 460A-467A.
103. C. D. Forbes, J. G. Toth, C. C. Ozbal, W. A. LaMarr, J. A. Pendleton, S. Rocks, R. W. Gedrich, D. G. Osterman, J. A. Landro and K. J. Lumb, *J. Biomol. Screen*, 2007, **12**, 628-634.
104. E. Capriotti, P. Fariselli and R. Casadio, *Nucleic Acids Rese.*, 2005, **33**, W306-W310.
105. J. Cheng, A. Randall and P. Baldi, *Proteins: Struct., Funct., Bioinf.*, 2006, **62**, 1125-1132.
106. J. Liao, M. K. Warmuth, S. Govindarajan, J. E. Ness, R. P. Wang, C. Gustafsson and J. Minshull, *BMC Biotechnol.*, 2007, **7**, 16.
107. T. S. Bayer, D. M. Widmaier, K. Temme, E. A. Mirsky, D. V. Santi and C. A. Voigt, *J. Am. Chem. Soc.*, 2009, **131**, 6508-6515.
108. X. Xie, I. Pashkov, X. Gao, J. L. Guerrero, T. O. Yeates and Y. Tang, *Biotechnol. Bioeng.*, 2009, **102**, 20-28.
109. H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest and G. M. Church, *Nature*, 2009, **460**, 894-898.
110. K. M. Esvelt, J. C. Carlson and D. R. Liu, *Nature*, 2011, **472**, 499-503.
111. J. C. Baret, O. J. Miller, V. Taly, M. Ryckelynck, A. El-Harrak, L. Frenz, C. Rick, M. L. Samuels, J. B. Hutchison, J. J. Agresti, D. R. Link, D. A. Weitz and A. D. Griffiths, *Lab Chip*, 2009, **9**, 1850-1858.
112. T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda and H. Taguchi, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 4201-4206.
113. M. C. Ebert and J. N. Pelletier, *Curr. Opin. Chem. Biol.*, 2017, **37**, 89-96.
114. J. A. Doudna and E. Charpentier, *Science*, 2014, **346**.
115. L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang, *Science*, 2013, **339**, 819-823.
116. P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville and G. M. Church, *Science*, 2013, **339**, 823-826.
117. Berkeley Lights, Inc., US Pat., 9,403,172, 2014.
118. Berkeley Lights, Inc., US Pat., 9,617,145, 2015.
119. B. Chen, S. Lim, A. Kannan, S. C. Alford, F. Sunden, D. Herschlag, I. K. Dimov, T. M. Baer and J. R. Cochran, *Nat. Chem. Biol.*, 2016, **12**, 76-81.
120. P. A. Romero, A. Krause and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E193-E201.
121. A. D. Garst, M. C. Bassalo, G. Pines, S. A. Lynch, A. L. Halweg-Edwards, R. Liu, L. Liang, Z. Wang, R. Zeitoun, W. G. Alexander and R. T. Gill, *Nat. Biotechnol.*, 2017, **35**, 48-55.
122. N. M. Gaudelli, A. C. Komor, H. A. Rees, M. S. Packer, A. H. Badran, D. I. Bryson and D. R. Liu, *Nature*, 2017, **551**, 464-471.
123. SCIEX Debuts Breakthrough Acoustic Ejection Mass Spectrometry Technology at ASMS 2019, https://sciex.com/about-us/press-releases/sciex-debuts-breakthrough-acoustic-ejection-mass-spectrometry-technology-at-asms-2019, (accessed October 2019).
124. D. A. Holland-Moritz, M. K. Wismer, B. F. Mann, I. Farasat, P. Devine, E. D. Guetschow, I. Mangion, C. J. Welch, J. C. Moore, S. Sun and R. T. Kennedy, *Angew. Chem., Int. Ed.*, 2020, **59**, 4470-4477.
125. J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych and A. M. J. J. Bonvin, *Proteins: Struct., Funct., Bioinf.*, 2018, **86**, 51-66.
126. S. Mazurenko, Z. Prokop and J. Damborsky, *ACS Catalysis*, 2020, **10**, 1210-1223.
127. Zymergen Acquires Metagenomics Company Radiant Genomics, https://www.businesswire.com/news/home/20180108006506/en/Zymergen-Acquires-Metagenomics-Company-Radiant-Genomics, (accessed March 2020).
128. J. Kim, A Web Enabled Database for Rapid Metagenomic Biocatalyst Discovery and Validation, https://www.sbir.gov/sbirsearch/detail/1165763, (accessed March 2020).
129. B. Adams, In conversation with: Laurence Reid, Warp Drive Bio CEO, https://www.fiercebiotech.com/biotech/conversation-laurence-reid-warp-drive-bio-ceo, (accessed March 2020).
130. Warp Drive Bio, Inc., US Pat., 9,428,845, 2011.
131. S. F. Brady, *Nat. Protoc.*, 2007, **2**, 1297-1305.
132. Genentech, Lodo Therapeutics Ink Up-to-$969M Metagenomics Drug Discovery Partnership, https://www.genengnews.com/topics/omics/genentech-lodo-therapeutics-ink-up-to-969m-metagenomics-drug-discovery-partnership/).

133. M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman and R. M. Goodman, *Appl. Environ. Microbiol.*, 2000, **66**, 2541-2547.
134. LifeMine Therapeutics, Inc., WIPO Pat., WO2019055816, 2019.
135. M. J. Czar, J. C. Anderson, J. S. Bader and J. Peccoud, *Trends Biotechnol.*, 2009, **27**, 63-72.
136. Blue Heron Biotechnology, Inc., US Pat., 6,664,112, 2001.
137. R. Carlson, On DNA and Transistors, http://www.synthesis.cc/synthesis/category/Carlson+Curves, (accessed March 2020).
138. S. Kosuri and G. M. Church, *Nat. Methods*, 2014, **11**, 499-507.
139. S. J. Kodumal, K. G. Patel, R. Reid, H. G. Menzella, M. Welch and D. V. Santi, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 15573-15578.
140. A. Esson, Twist Bioscience Provides Ginkgo Bioworks with one Billion Base Pairs of Synthetic DNA, http://www.frontlinegenomics.com/news/15286/twist-bioscience-provides-ginkgo-bioworks-one-billion-base-pairs-synthetic-dna/, (accessed March 2020).
141. A. Casini, F. Y. Chang, R. Eluere, A. M. King, E. M. Young, Q. M. Dudley, A. Karim, K. Pratt, C. Bristol, A. Forget, A. Ghodasara, R. Warden-Rothman, R. Gan, A. Cristofaro, A. E. Borujeni, M. H. Ryu, J. Li, Y. C. Kwon, H. Wang, E. Tatsis, C. Rodriguez-Lopez, S. O'Connor, M. H. Medema, M. A. Fischbach, M. C. Jewett, C. Voigt and D. B. Gordon, *J. Am. Chem. Soc.*, 2018, **140**, 4302-4316.
142. R. Perez-Jimenez, A. Inglés-Prieto, Z.-M. Zhao, I. Sanchez-Romero, J. Alegre-Cebollada, P. Kosuri, S. Garcia-Manyes, T. J. Kappock, M. Tanokura, A. Holmgren, J. M. Sanchez-Ruiz, E. A. Gaucher and J. M. Fernandez, *Nat. Struct. Mol. Biol.*, 2011, **18**, 592-596.
143. V. A. Risso, J. A. Gavira, D. F. Mejia-Carmona, E. A. Gaucher and J. M. Sanchez-Ruiz, *J. Am. Chem. Soc.*, 2013, **135**, 2899-2902.
144. U. Alcolombri, M. Elias and D. S. Tawfik, *J. Mol. Biol.*, 2011, **411**, 837-853.
145. E. A. Althoff, L. Wang, L. Jiang, L. Giger, J. K. Lassila, Z. Wang, M. Smith, S. Hari, P. Kast, D. Herschlag, D. Hilvert and D. Baker, *Protein Sci.*, 2012, **21**, 717-726.
146. O. Khersonsky, G. Kiss, D. Röthlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker and D. S. Tawfik, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 10358-10363.
147. L. Giger, S. Caner, R. Obexer, P. Kast, D. Baker, N. Ban and D. Hilvert, *Nat. Chem. Biol.*, 2013, **9**, 494-498.
148. J. D. Hermes, S. C. Blacklow and J. R. Knowles, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 696-700.
149. Y. H. Zhou, X. P. Zhang and R. H. Ebright, *Nucleic Acids Res.*, 1991, **19**, 6052-6052.
150. R. C. Cadwell and G. F. Joyce, *Genome Res.*, 1992, **2**, 28-33.
151. M. Fromant, S. Blanquet and P. Plateau, *Anal. Biochem.*, 1995, **224**, 347-353.
152. J. P. Vartanian, M. Henry and S. Wain-Hobson, *Nucleic Acids Res.*, 1996, **24**, 2627-2631.
153. G. E. Degnen and E. C. Cox, *J Bacteriol*, 1974, **117**, 477-487.
154. H. Liao, T. McKenzie and R. Hageman, *Proc. Natl. Acad. Sci. U. S. A.*, 1986, **83**, 576-580.
155. M. M. Ling and B. H. Robinson, *Anal. Biochem.*, 1997, **254**, 157-178.
156. B. T. Greenhagen, P. E. O'Maille, J. P. Noel and J. Chappell, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 9826-9831.
157. D. A. Smith, M. J. Ackland and B. C. Jones, *Drug Discovery Today*, 1997, **2**, 406-414.
158. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede and A. Tramontano, *Proteins: Struct., Funct., Bioinf.*, 2016, **84**, 4-14.
159. M. L. Metzker, *Nat. Rev. Genet.*, 2010, **11**, 31-46.
160. B. Kuhlman and P. Bradley, *Nat. Rev. Mol. Cell Biol.*, 2019, **20**, 681-697.
161. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker and K. N. Houk, *Angew. Chem., Int. Ed.*, 2013, **52**, 5700-5725.
162. A. J. Ruff, A. Dennig and U. Schwaneberg, *FEBS J.*, 2013, **280**, 2961-2978.
163. A. Pavelka, E. Chovancova and J. Damborsky, *Nucleic Acids Res.*, 2009, **37**, W376-W383.
164. J. Bendl, J. Stourac, E. Sebestova, O. Vavra, M. Musil, J. Brezovsky and J. Damborsky, *Nucleic Acids Res.*, 2016, **44**, W479-W487.
165. L. Sumbalova, J. Stourac, T. Martinek, D. Bednar and J. Damborsky, *Nucleic Acids Res.*, 2018, **46**, W356-W362.
166. F. H. Arnold and G. Georgiou, eds., *Directed Evolution Library Creation*, Humana Press, Totowa, NJ, 2003.
167. B. C. Cunningham and J. A. Wells, *Science*, 1989, **244**, 1081-1085.
168. B. C. Cunningham, P. Jhurani, P. Ng and J. A. Wells, *Science*, 1989, **243**, 1330-1336.
169. A. Currin, N. Swainston, P. J. Day and D. B. Kell, *Chem. Soc. Rev.*, 2015, **44**, 1172-1239.
170. S. Bratulic and A. H. Badran, *Curr. Opin. Chem. Biol.*, 2017, **41**, 50-60.
171. S. K. Sharan, L. C. Thomason, S. G. Kuznetsov and D. L. Court, *Nat. Protoc.*, 2009, **4**, 206-223.
172. Z. Bao, M. HamediRad, P. Xue, H. Xiao, I. Tasan, R. Chao, J. Liang and H. Zhao, *Nat. Biotechnol.*, 2018, **36**, 505-508.
173. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris and D. R. Liu, *Nature*, 2016, **533**, 420-424.
174. C. Li, R. Zhang, X. Meng, S. Chen, Y. Zong, C. Lu, J.-L. Qiu, Y.-H. Chen, J. Li and C. Gao, *Nat. Biotechnol.*, 2020, DOI: 10.1038/s41587-019-0393-7.
175. H. Butt, A. Eid, A. A. Momin, J. Bazin, M. Crespi, S. T. Arold and M. M. Mahfouz, *Genome Biol.*, 2019, **20**.
176. Q. Li, Y. Li, S. Yang, S. Huang, M. Yan, Y. Ding, W. Tang, X. Lou, Q. Yin, Z. Sun, L. Lu, H. Shi, H. Wang, Y. Chen and J. Li, *Nat. Cell Biol.*, 2018, **20**, 1315-1325.
177. U. Markel, K. D. Essani, V. Besirlioglu, J. Schiffels, W. R. Streit and U. Schwaneberg, *Chem. Soc. Rev.*, 2020, **49**, 233-262.
178. C. C. Ozbal, W. A. LaMarr, J. R. Linton, D. F. Green, A. Katz, T. B. Morrison and C. J. Brenan, *Assay Drug Dev. Technol.*, 2004, **2**, 373-381.
179. G. J. Van Berkel and V. Kertesz, *Rapid Commun. Mass Spectrom.*, 2015, **29**, 1749-1756.
180. E. E. Kempa, K. A. Hollywood, C. A. Smith and P. E. Barran, *Analyst*, 2019, **144**, 872-891.
181. I. Sinclair, M. Bachman, D. Addison, M. Rohman, D. C. Murray, G. Davies, E. Mouchet, M. E. Tonge, R. G. Stearns, L. Ghislain, S. S. Datwani, L. Majlof, E. Hall, G. R. Jones, E. Hoyes, J. Olechno, R. N. Ellson, P. E. Barran, S. D. Pringle, M. R. Morris and J. Wingfield, *Anal. Chem.*, 2019, **91**, 3790-3794.

182. G. Chao, W. L. Lau, B. J. Hackel, S. L. Sazinsky, S. M. Lippow and K. D. Wittrup, *Nat. Protoc.*, 2006, **1**, 755-768.
183. Diversa Corporation, US Pat., 6,174,673, 1998.
184. A. Aharoni, G. Amitai, K. Bernath, S. Magdassi and D. S. Tawfik, *Chem. Biol.*, 2005, **12**, 1281-1289.
185. E. Mastrobattista, V. Taly, E. Chanudet, P. Treacy, B. T. Kelly and A. D. Griffiths, *Chem. Biol.*, 2005, **12**, 1291-1300.
186. M. Fischlechner, Y. Schaerli, M. F. Mohamed, S. Patil, C. Abell and F. Hollfelder, *Nat. Chem.*, 2014, **6**, 791-796.
187. P. Mair, F. Gielen and F. Hollfelder, *Curr. Opin. Chem. Biol.*, 2017, **37**, 137-144.
188. S. Becker, H.-U. Schmoldt, T. M. Adams, S. Wilhelm and H. Kolmar, *Curr. Opin. Biotechnol.*, 2004, **15**, 323-329.
189. A. Mocciaro, T. L. Roth, H. M. Bennett, M. Soumillon, A. Shah, J. Hiatt, K. Chapman, A. Marson and G. Lavieu, *Comm. Biol.*, 2018, **1**, DOI: 10.1038/s42003-42018-40034-42006.
190. D. Wilkinson, *The Scientist*, 1999.
191. J. Damborský, *Quant. Struct.-Act. Relat.*, 1997, **16**, 126-135.
192. R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon and G. W. Huisman, *Nat. Biotechnol.*, 2007, **25**, 338-344.
193. C. K. Savile, J. M. Janey, E. C. Mundorff, J. C. Moore, S. Tam, W. R. Jarvis, J. C. Colbeck, A. Krebber, F. J. Fleitz, J. Brands, P. N. Devine, G. W. Huisman and G. J. Hughes, *Science*, 2010, **329**, 305-309.
194. K. K. Yang, Z. Wu and F. H. Arnold, *Nat. Methods*, 2019, **16**, 687-694.
195. G. B. Kim, W. J. Kim, H. U. Kim and S. Y. Lee, *Curr. Opin. Biotechnol.*, 2020, **64**, 1-9.
196. W. S. Noble, *Nat. Biotechnol.*, 2006, **24**, 1565-1567.
197. BioDisrupt 2019: Amyris Overview, https://investors.amyris.com/download/Amyris+Overview+John+Melo+Presentation.pdf, (accessed November 2019).
198. E. Hyde, Zymergen raises $400M+ to deliver AI-enabled biology to global bio-based industry, https://synbiobeta.com/zymergen-raises-400m-to-deliver-ai-enabled-biology-to-global-bio-based-industry/, (accessed November 2019).
199. Ginkgo Bioworks Our Platform, https://www.ginkgobioworks.com/our-platform/, (accessed November 2019).
200. Codexis, Inc., US Pat., 20150134315A1, 2015.
201. Codexis Corporate Presentation, http://ir.codexis.com/static-files/eaa02f29-85a5-4544-9da1-a15c621ba98c, (accessed November 2019).
202. P. Lee, Bringing together deep bioscience and AI to help patients worldwide: Novartis and Microsoft work to reinvent treatment discovery and development https://blogs.microsoft.com/blog/2019/10/01/bringing-together-deep-bioscience-and-ai-to-help-patients-worldwide-novartis-and-microsoft-work-to-reinvent-treatment-discovery-and-development/, (accessed November 2019).
203. A. E. Fraley and D. H. Sherman, *Bioorg. Med. Chem. Lett.*, 2018, **28**, 1992-1999.
204. D. R. M. Smith, S. Grüschow and R. J. M. Goss, *Curr. Opin. Chem. Biol.*, 2013, **17**, 276-283.
205. H. Onaka, S. Taniguchi, Y. Igarashi and T. Furumai, *Biosci., Biotechnol., Biochem.*, 2003, **67**, 127-138.
206. J. T. Payne, C. B. Poor and J. C. Lewis, *Angew. Chem., Int. Ed.*, 2015, **54**, 4226-4230.
207. W. Runguphan, X. Qu and S. E. O'Connor, *Nature*, 2010, **468**, 461-464.
208. W. S. Glenn, E. Nims and S. E. O'Connor, *J. Am. Chem. Soc.*, 2011, **133**, 19346-19349.
209. C. Dong, F. Huang, H. Deng, C. Schaffrath, J. B. Spencer, D. O'Hagan and J. H. Naismith, *Nature*, 2004, **427**, 561-565.
210. H. Sun, W. L. Yeo, Y. H. Lim, X. Chew, D. J. Smith, B. Xue, K. P. Chan, R. C. Robinson, E. G. Robins, H. Zhao and E. L. Ang, *Angew. Chem., Int. Ed.*, 2016, **55**, 14277-14280.
211. M. C. Walker, B. W. Thuronyi, L. K. Charkoudian, B. Lowry, C. Khosla and M. C. Y. Chang, *Science*, 2013, **341**, 1089-1094.
212. Y. Li, S. Li, K. Thodey, I. Trenchard, A. Cravens and C. D. Smolke, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, E3922-E3931.
213. E. E. Wrenbeck, M. A. Bedewitz, J. R. Klesmith, S. Noshin, C. S. Barry and T. A. Whitehead, *ACS Synth. Biol.*, 2019, **8**, 474-481.
214. M. A. Bedewitz, A. D. Jones, J. C. D'Auria and C. S. Barry, *Nat. Commun.*, 2018, **9**, DOI: 10.1038/s41467-41018-07671-41463.
215. J. R. Klesmith, J. P. Bacik, E. E. Wrenbeck, R. Michalczyk and T. A. Whitehead, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 2265-2270.
216. P. Srinivasan and C. D. Smolke, *Nat. Commun.*, 2019, **10**, DOI: 10.1038/s41467-41019-11588-w.
217. P. N. Devine, R. M. Howard, R. Kumar, M. P. Thompson, M. D. Truppo and N. J. Turner, *Nat. Rev. Chem.*, 2018, **2**, 409-421.
218. M. T. Reetz, *Chem. Rec.*, 2016, **16**, 2449-2459.
219. R. A. Sheldon and D. Brady, *Chem. Commun.*, 2018, **54**, 6088-6104.
220. J. H. Schrittwieser, S. Velikogne, M. Hall and W. Kroutil, *Chem. Rev.*, 2018, **118**, 270-348.
221. J. M. Sperl and V. Sieber, *ACS Catal.*, 2018, **8**, 2385-2396.
222. H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J. D. Trawick, R. E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H. B. Schreyer, S. Andrae, T. H. Yang, S. Y. Lee, M. J. Burk and S. Van Dien, *Nat. Chem. Biol.*, 2011, **7**, 445-452.
223. G. M. Lin, R. Warden-Rothman and C. A. Voigt, *Curr. Opin. Syst. Biol.*, 2019, **14**, 82-107.
224. T. Newhouse, P. S. Baran and R. W. Hoffmann, *Chem Soc Rev*, 2009, **38**, 3010-3021.
225. B. S. Evans, Y. Chen, W. W. Metcalf, H. Zhao and N. L. Kelleher, *Chem Biol*, 2011, **18**, 601-607.
226. E. Leonard, P. K. Ajikumar, K. Thayer, W. H. Xiao, J. D. Mo, B. Tidor, G. Stephanopoulos and K. L. Prather, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 13654-13659.
227. Y. Tong, T. Weber and S. Y. Lee, *Nat. Prod. Rep.*, 2019, **36**, 1262-1280.
228. T. Jakočiūnas, L. E. Pedersen, A. V. Lis, M. K. Jensen and J. D. Keasling, *Metab. Eng.*, 2018, **48**, 288-296.
229. N. Krishnamurthy, Leveraging the Power of Digital Genome Engineering - SynBioBeta, https://www.inscripta.com/resources, (accessed November 2019).
230. R. Fox, Genome Scale Mapping of Genotype to Phenotype Relationships - SynBioBeta,

https://www.inscripta.com/resources#, (accessed November 2019).
231. S. d'Oelsnitz and A. Ellington, *Curr. Opin. Biotechnol.*, 2018, **53**, 158-163.
232. Z. L. Tan, X. Zheng, Y. Wu, X. Jian, X. Xing and C. Zhang, *Microb. Cell Fact.*, 2019, **18**, DOI: 10.1186/s12934-12019-11132-y.
233. Y. Wang, X. Yu and H. Zhao, *AIChE Journal*, 2019, DOI: 10.1002/aic.16716.
234. M. A. Huffman, A. Fryszkowska, O. Alvizo, M. Borra-Garske, K. R. Campos, K. A. Canada, P. N. Devine, D. Duan, J. H. Forstater, S. T. Grosser, H. M. Halsey, G. J. Hughes, J. Jo, L. A. Joyce, J. N. Kolev, J. Liang, K. M. Maloney, B. F. Mann, N. M. Marshall, M. McLaughlin, J. C. Moore, G. S. Murphy, C. C. Nawrat, J. Nazor, S. Novick, N. R. Patel, A. Rodriguez-Granillo, S. A. Robaire, E. C. Sherer, M. D. Truppo, A. M. Whittaker, D. Verma, L. Xiao, Y. Xu and H. Yang, *Science*, 2019, **366**, 1255-1259.
235. GRAS Notice No. 512, https://wayback.archive-it.org/7993/20171031000229/https://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/NoticeInventory/ucm424538.htm, (accessed November 2019).
236. GRAS Notice No. 667, https://www.fda.gov/media/100245/download, (accessed November 2019).
237. GRAS Notice No. 744, https://www.fda.gov/media/113131/download, (accessed November 2019).
238. GRAS Notice No. 759, https://www.fda.gov/media/125423/download, (accessed November 2019).
239. GRAS Notice No. 812, https://www.fda.gov/media/130891/download, (accessed November 2019).
240. GRAS Notice No. 745, https://www.fda.gov/media/115468/download, (accessed November 2019).
241. GRAS Notice No. 780, https://www.fda.gov/media/119339/download, (accessed November 2019).
242. GRAS Notice No. 799, https://www.fda.gov/media/117109/download, (accessed November 2019).
243. Amyris announces successful shipment of first fermentation derived cannabinoid to Lavvan and provides business updates, https://investors.amyris.com/2019-12-27-Amyris-Announces-Successful-Shipment-of-First-Fermentation-Derived-Cannabinoid-to-LAVVAN-and-Provides-Business-Updates, (accessed March 2020).
244. X. Luo, M. A. Reiter, L. d'Espaux, J. Wong, C. M. Denby, A. Lechner, Y. Zhang, A. T. Grzybowski, S. Harth, W. Lin, H. Lee, C. Yu, J. Shin, K. Deng, V. T. Benites, G. Wang, E. E. K. Baidoo, Y. Chen, I. Dev, C. J. Petzold and J. D. Keasling, *Nature*, 2019, **567**, 123-126.
245. Z. Tan, J. M. Clomburg and R. Gonzalez, *ACS Synth. Biol.*, 2018, **7**, 1886-1896.
246. T. P. Korman, P. H. Opgenorth and J. U. Bowie, *Nat. Commun.*, 2017, **8**, DOI: 10.1038/ncomms15526.
247. Cronos Group and Ginkgo Bioworks Announce a Landmark Parntership to Produce Cultured Cannabinoids, https://ir.thecronosgroup.com/news-releases/news-release-details/cronos-group-and-ginkgo-bioworks-announce-landmark-partnership, (accessed November 2019).
248. Intrexon Announces Advances in Production of Medical Cannabis, https://synbiobeta.com/intrexon-announces-advances-in-production-of-medical-cannabis/, (accessed December 2019).
249. K. A. Costa, K. Vavitsas, M. Limas, B. Joseph-Nelson and J. Cumbers, Cannabinoid Fermentation: Scalability, Purity, and Sustainability for an Emerging Market, https://synbiobeta.com/wp-content/uploads/2019/07/Cannabinoid-Fermentation-SynBioBeta-Industry-Report-June-2019-v2.pdf, (accessed December 2019).
250. Intrexon Corporation, WIPO Pat., WO2019071000, 2019.