Machine learning guided optimization of an artificial carbon dioxide fixation cycle and its extension towards value-added compounds

Dissertation

kumulativ

zur Erlangung des Grades eines

Doktor der Naturwissenschaften

(Dr. rer. Nat.)

des Fachbereichs Biologie der Philipps-Universität Marburg

Vorgelegt von

Christoph Diehl

Aus Siegen

Marburg, 2022

Die vorliegende Arbeit mit dem Titel **„Machine learning guided optimization of an artificial carbon dioxide fixation cycle and its extension towards value-added compounds“** wurde von 10.2017 bis 03.2022 unter der Betreuung von Herrn Prof. Dr. Tobias J. Erb am Max-Planck-Institut für terrestrische Mikrobiologie in Marburg in der Arbeitsgruppe Biochemie und Synthetischer Metabolismus erstellt.

Die Arbeit wurde vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) am 29.03.2022 als Dissertation angenommen.

Erstgutachter: Prof. Dr. Tobias J. Erb

Zweitgutachter: Prof. Dr. Hans-Ulrich Mösch

Datum der Disputation: 10.06.2022

Erklärung

Ich versichere, dass ich die Dissertation mit dem Titel **„Machine learning guided optimization of an artificial carbon dioxide fixation cycle and its extension towards value-added compounds“** selbständig und ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfsmittel bedient habe.

Diese Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den

Christoph Diehl

"To Margulis, the great Oxidation Event had lessons for today. The first was that people who thought that living creatures couldn't affect the climate had no idea of the power of life. The second was that the onset of climate change meant that Homo sapiens was getting into the biological big leagues—we were tiptoeing into the terrain of bacteria, algae, and other truly important creatures. The third was that species, like sullen teenagers, don't pick up after themselves. Cyanobacteria sprayed their oxygen garbage all over Earth without concern for the consequences—littering on an epic scale. People were doing the same with carbon dioxide."

From "The Wizard and the Prophet" by Charles C. Mann

Table of Contents

Abstract

The global carbon cycle is a highly balanced exchange system of carbon between the geo-, hydro- and atmosphere. Since the industrial revolution, the combustion of fossil fuels is one of the main reasons for the shift of carbon levels towards higher concentrations in the hydro- and atmosphere. The amount of carbon dioxide (CO_2) in the air is comparibly low but it is a highly potent greenhouse gas due to its structural properties. Its capability of absorbing infrared light, which would otherwise escape into space, leads to an increase in atmospheric temperatures. Because CO_2 molecules are very stable, the conversion into multi-carbon-molecules is energy demanding and mainly done by organisms which use light as an energy source. Therefore, the main workhorses of capturing CO_2 from the atmosphere are plants and algae, which incorporate the carbon for the production of biomass. In these complex organisms the Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) is responsible for the carboxylation reaction. RuBisCO is considered a slow catalyst with a high error rate in accepting oxygen (O_2) instead of CO_2.

In 2016, Schwander *et al.* published a new-to-nature pathway for the fixation of CO_2. The so-called **c**rotonyl-coenzyme A (CoA) /**et**hylmalonyl-**C**oA/**h**ydroxybutyryl-CoA (CETCH) cycle was designed to circumvent RubisCO. In contrast to the Calvin cycle, the CETCH cycle is based on a highly efficient crotonyl-CoA carboxylase/reductase (Ccr), which does not display any side reaction with oxygen such as RuBisCO. Due to the overall pathway design, including Ccr as the key catalyst, the CETCH cycle has a higher net efficiency than natural aerobic CO_2-fixation pathways and thus harbors the potential to play an important role in the reduction of atmospheric CO_2 levels.

To gain insights into this complex *in vitro* assay consisting of more than 25 components, we established a high-throughput workflow that enabled us to test hundreds of reaction conditions simultaneously. Prerequisite was the implementation of an acoustic liquid handling robot with a minimal pipetting volume of 25 nl. This enabled a fast reaction assembly and a drastic reduction in assay volume while maintaining a high pipetting accuracy. The acquired data was used to subsequently train an XGBoost-based machine learning algorithm aiming to optimize the CETCH cycle reaction parameters. After five rounds of optimization, the final model predicts reaction parameters for assay conditions with a ten-fold improvement on the manually optimized pathway version published in 2016. In addition, the algorithm identified the important components of the pathway and revealed one enzyme as a potential bottleneck of the current assay. Follow up experiments showed that the loss of intermediates by side reactions or hydrolysis is the limiting factor of the assay.

Furthermore, we sought to extend the product portfolio of the CETCH cycle beyond its primary product glyoxylate. To this end, we first coupled the CETCH cycle to the β-hydroxyaspartate cycle. This enabled the production of oxaloacetate from two molecules glyoxylate. Adding only three additional enzymes from the serine cycle leads to the formation of acetyl-CoA, which we used to produce different terpenes via the mevalonate pathway. Despite the wide range of products, their synthesis has remained limited to the use of molecules produced downstream of the CETCH cycles primary product glyoxylate. Intermediates of the cycle were inaccessible, as their removal would lead to stalling of the pathway and a pre-mature arrest of CO_2–fixation. To enable the utilization of CETCH cycle intermediates, we implemented anaplerotic routes that use the fixed CO_2 to replenish drained cycle intermediates. We successfully reconstituted three anaplerotic pathways that enabled the production of the polyketide 6-deoxyerythronolide B (6-dEB). The biosynthesis of 6-dEB (C21) requires one molecule of propionyl-CoA and six molecules methylmalonyl-CoA, both intermediates of the CETCH cycle. The biosynthesis of complex molecules from CO_2 in context of different highly convoluted pathways with up to 50 reactions highlights the robustness and versatility of the CETCH cycle.

Zusammenfassung

Der globale Kohlenstoffkreislauf ist ein hochgradig ausgewogenes System welches den Kohlenstoffaustausch zwischen Geo-, Hydro- und Atmosphäre umfasst. Seit der industriellen Revolution ist die Verbrennung fossiler Brennstoffe der Hauptgrund für den Anstieg des Kohlenstoffgehalts in der Atmosphäre, sowie im Wasser. Zwar ist die Konzentration an Kohlenstoffdioxid (CO_2) in der Luft nach wie vor sehr gering, da es jedoch aufgrund seiner strukturellen Eigenschaften ein hochwirksames Treibhausgas ist, hätte eine Verdopplung der Menge katastrophale Auswirkungen auf das Klima. Seine Fähigkeit Infrarotlicht zu absorbieren, welches sonst ins Weltall entweichen würde, führt zu einem Anstieg der atmosphärischen Temperaturen. Da CO_2-Moleküle sehr stabil sind, ist die Umwandlung in langkettige Kohlenstoff-Moleküle sehr energieaufwändig und erfolgt hauptsächlich durch Organismen die Licht als Energiequelle nutzen. Die Hauptakteure bei der Aufnahme von CO_2 aus der Atmosphäre sind daher Pflanzen und Algen, die den Kohlenstoff zur Erzeugung von Biomasse aufnehmen. In diesen komplexen Organismen ist die Ribulose-1,5-Bisphosphat Carboxylase/Oxygenase (RuBisCO) für die Carboxylierung verantwortlich. RuBisCO gilt als langsamer Katalysator mit einer hohen Fehlerquote bei der Aufnahme von Sauerstoff (O_2) anstelle von CO_2.

2016 wurde eine Studie von Schwander et al. veröffentlicht, in welcher ein neuartiger Stoffwechselweg zur Fixierung von CO_2 vorgestellt wurde. Der sogenannte Crotonyl-Coenzym A (CoA)/Ethylmalonyl-CoA/Hydroxybutyryl-CoA (CETCH)-Zyklus wurde entwickelt, um RubisCO zu umgehen. Im Gegensatz zum Calvin-Zyklus basiert der CETCH-Zyklus auf einer hocheffizienten Crotonyl-CoA Carboxylase/Reduktase (Ccr), bei der keine Nebenreaktionen mit Sauerstoff auftreten. Aufgrund der Nutzung von Ccr und dem restlichen Aufbau des Stoffwechselweges hat der CETCH-Zyklus eine höhere Nettowirksamkeit als natürliche aerobe CO_2-Fixierungswege und hat somit das Potential, eine wichtige Rolle bei der Reduzierung des atmosphärischen CO_2-Gehalts zu spielen.

Um Einblicke in dieses komplexe *in vitro* Reaktionsnetzwerk aus mehr als 25 Komponenten zu gewinnen, haben wir eine Hochdurchsatzverfahren etabliert um Hunderte von Reaktionsbedingungen gleichzeitig zu testen. Voraussetzung war die Implementierung eines akustischen Pipettier-Roboters mit einem minimalen Pipettiervolumen von 25 nl. Dies ermöglichte einen schnelleren Durchsatz und eine drastische Verringerung des Reaktionsvolumens ohne Verluste bei der Pipettiergenauigkeit. Die gewonnenen Daten wurden verwendet um einen XGBoost-basierten Machine Learning Algorithmus zu trainieren, um die Reaktionsparameter des CETCH-Zyklus zu optimieren. Durch fünf Runden Optimierung konnten die Reaktionsparameter so verändert werden, dass der CETCH-Zyklus im Vergleich zur publizierten Version von 2016 um den Faktor zehn verbessert werden konnte. Darüber

hinaus konnten durch den Algorithmus die wichtigsten Komponenten des Stoffwechselwegs identifiziert werden und ein Enzym als potentieller Engpass ausgemacht werden. Des weiteren konnte durch anschließende Experimente gezeigt werden, dass der Verlust von Metaboliten durch Nebenreaktionen oder Hydrolyse der begrenzende Faktor ist.

Neben der Optimierung des CETCH-Zyklus erweiterten wir dessen Produktportfolio um CO_2 direkt in höherwertige Chemikalien umzuwandeln. Zu diesem Zweck koppelten wir zunächst den CETCH-Zyklus mit dem β-Hydroxyaspartat-Zyklus. Dies ermöglichte die Herstellung von Oxaloacetat aus zwei Molekülen Glyoxylat. Durch drei weitere Enzyme aus dem Serin-Zyklus konnten wir Acetyl-CoA synthetisieren, welches wir zur Herstellung verschiedener Terpene verwendeten. Zwar konnten wir die Produktion von verschiedenen Stoffen demonstrieren, die Synthese beschränkte sich jedoch auf Produkte welche aus dem Primärprodukt Glyoxylat gewonnen werden. Die Nutzung von CETCH Metaboliten war nicht möglich, da dies zu einem vorzeitigen Stillstand der CO_2-Fixierung führen würde. Um die Konzentrationen von CETCH-Metaboliten zu erhöhen, implementierten wir von der Natur inspirierte, anaplerotische Sequenzen, welche das fixierte CO_2 in den Zyklus zurückführen. Dazu rekonstituierten wir erfolgreich vier anaplerotische Routen, von welcher drei die Produktion des Polyketids 6-Desoxyerythronolid B (6-dEB) ermöglichten. Die Biosynthese von 6-dEB (C21) erfordert ein Molekül Propionyl-CoA und sechs Moleküle Methylmalonyl-CoA, beides Zwischenprodukte des CETCH-Zyklus. Durch die Biosynthese komplexer Moleküle aus CO_2 in hochgradig komplexen Stoffwechselwegen mit über 50 Reaktionen konnten wir die Robustheit und Vielseitigkeit des CETCH-Zyklus demonstrieren.

List of manuscripts and publications

Following publications were prepared during the research on the optimization and extension of the CETCH cycle. **The first three manuscripts are included as chapters of this thesis**. The publications four to six are not included as chapters due to the minor experimental work contributed from my side. However, they are mentioned in the introduction and/or discussion.

***** These authors contributed equally to the publication

1) A versatile active learning workflow for optimization of genetic and metabolic networks.
   PANDI, A.*, **DIEHL, C.***, YAZDIZADEH KHARRAZI, A., FAURE, L., SCHOLZ, S. A., NATTERMANN, M., ADAM, D., CHAPIN, N., FOROUGHIJABBARI, Y., MORITZ, C., PACZIA, N., CORTINA, N. S., FAULON, J. L. & ERB, T. J. 2022. *In revision. Submitted to Nature Communications (15.02.2022). Preprint available on bioRxiv (https://doi.org/10.1101/2021.12.28.474323).*

2) A modular in vitro platform for the production of terpenes and polyketides from CO_2
   SUNDARAM, S., **DIEHL, C.**, CORTINA, N. S., BAMBERGER, J., PACZIA, N. & ERB, T. J. 2021. Angewandte Chemie.

3) Enhancing the synthetic capabilities of a complex *in vitro* metabolic network through anaplerotic reaction modules
   DIEHL, C.*, GERLINGER, P. D.*, PACZIA, N. & ERB, T. J. 2021. *In revision. Submitted to Nature Chemical Biology (21.12.2021).*

4) Light-powered CO2 fixation in a chloroplast mimic with natural and synthetic parts
   MILLER, T. E., BENEYTON, T., SCHWANDER, T., **DIEHL, C.**, GIRAULT, M., MCLEAN, R., CHOTEL, T., CLAUS, P., CORTINA, N. S., BARET, J. C. & ERB, T. J. 2020. Science

5) A new-to-nature carboxylation module to improve natural and synthetic CO_2-fixation
   SCHEFFEN, M., MARCHAL, D. G., BENEYTON, T., SCHULLER, S. K., KLOSE, M., **DIEHL, C.**, LEHMANN, J., PFISTER, P., CARRILLO, M., HE, H., ASLAN, S., CORTINA, N. S., CLAUS, P., BOLLSCHWEILER, D., BARET, J.-C., SCHULLER, J. M., ZARZYCKI, J., BAR-EVEN, A. & ERB, T. J. 2021. Nature Catalysis

6) Advances and applications of cell-free systems for metabolic production
   MORITZ, C., SUNDARAM, S., **DIEHL, C.**, ADAM, D., BORKOWSKI, O. & PANDI, A. 2021. Microbial Cell Factories Engineering for Production of Biomolecules.

1. Introduction

1.1. A short introduction to carbon dioxide

Life on earth as we know it today is based on organic carbon. The combination of several characteristics and conditions led to the outstanding role of this element as the backbone of most biological relevant molecules. First, carbon is one of the most abundant elements. Since the emergence of our planet it was present in different compounds, which varied ever since and can be found in every layer of earth's crust[1]. Secondly, carbon is a very versatile element, which can form up to four stable bonds with a broad variety of other elements and itself. The ability of carbon-carbon bond formation under physiological conditions allows the construction of long chains as backbones for molecules. Furthermore, the capability to form four bonds allows branching of carbon chains or attaching other elements. If life could be based on other elements such as silicon is still discussed[2], but the success story of carbon based life is undisputed. There are millions of different carbon based molecules. Regarding the last century carbon based compounds like plastic, charcoal or other fossil fuels changed the world dramatically. First seen as panaceas and the driving forces of industrialization and wealth, the view on them has changed as the downsides occurred. Burning charcoal led to drastic air pollution in industrial countries, the combustion of fossil fuels released stored carbon mainly as carbon dioxide and plastic waste can be found almost everywhere on the planet. As the first sentence of this paragraph is true for carbon itself, it is also true for carbon dioxide: The fate of life on earth became dependent on carbon dioxide. The interplay between life and the levels of carbon dioxide reaches billions of years back in earth's history and accelerated with the evolution of photosynthesis. To understand why such a simple molecule has tremendous effects on the climate, we have to take a closer look at its physical properties. The climate on earth is the outcome of many interconnected factors and allows us humans and all other species to live. The atmosphere around the earth is responsible for the mild conditions to which we are adapted. In brief, the sun emits light, which is absorbed by the earth. Afterwards the heated surface emits energy as light in the infrared spectrum. Predominantly absorbed by water molecules, this energy is further transferred to oxygen and nitrogen molecules, thereby heating up the atmosphere. Due to the physical properties (more precisely the absorption spectrum) of water vapor, some of the light is not absorbed, but can escape from the atmosphere through gaps in the absorption spectrum, therefore preventing the atmosphere from heating up to intolerable temperatures. In 1938 Guy Stewart Callendar published his work where he showed that carbon dioxide has a similar absorption spectrum to atmospheric water. Nevertheless, two larger gaps in the spectrum of water vapor around 4 and 10 µm are complemented by the absorption spectrum of carbon dioxide. Therefore the more carbon dioxide is in the air, the more energy remains trapped and heats up the atmosphere[3]. Beside carbon dioxide, there are other

greenhouse gases (mainly methane and nitrous oxide) trapping infrared light by closing gaps in the absorbance spectrum of water vapor. Nevertheless, there are two reasons why carbon dioxide is considered the main driving force in increasing global temperatures. First, the amounts released into the atmosphere exceed the amounts of other greenhouse gases drastically[4]. Ever since the industrial revolution, the generated surplus of carbon dioxide release exceeds the assimilation capacity of the global carbon cycle. Although the oceans buffer a huge fraction of the additional released carbon dioxide as the main carbon sinks, they could not prevent the increase in atmospheric levels[5]. Secondly, carbon dioxide is a very stable compound, which barely reacts with other molecules, can stay in the atmosphere for thousands of years, while sequestration is very energy demanding. Nevertheless, there are several enzymes, which manage to incorporate carbon dioxide under physiological conditions, the most abundant of which is the Ribulose-1,5-bisphosphat-carboxylase/-oxygenase (RuBisCO), the key enzyme of the Calvin-Benson-Bassham Cycle (CBB).

1.2. RuBisCO: Earth´s most abundant enzyme

The release of carbon dioxide by combustion of fossil fuels and limiting the uptake of CO_2 by deforestation are two main reasons for the drastic increase in atmospheric CO_2 concentrations. Deforestation results in fewer plants, which are together with other phototropic organisms the main players in carbon fixation. Fewer plants therefore result in less RuBisCO, the enzyme that is used by plants to assimilate biomass by incorporating carbon dioxide. As one of the key players in photosynthesis, RuBisCO caught the attention of many researchers and was the subject of more than 5000 scientific publications[6]. For more than 45 years RuBisCO has been called the most abundant protein on earth, which was the result of some back-of-the-envelope calculations due to its abundance of up to 50% of a leafs protein content[7,8]. A recently published and more sophisticated calculation comes to the same conclusion, but additionally states that the abundance of RuBisCO is even an order of magnitude higher[9]. Although RuBisCOs abundance indicates an extremely successful path, it has been shown to be surprisingly inefficient. Under *in vitro* conditions, the average catalytic rate of RuBisCO is around three molecules per second, which is slow compared to other enzymes and even carboxylases (Figure 1.). Under physiological conditions this already slow rate is even further decreased to ~0.03 s^{-1} in land plants and ~0.6 s^{-1} in marine photosynthetic organisms[9]. As mentioned before, plants equip their leafs with RuBisCO to levels reaching up to 50% of the leafs total protein content to partially compensate the low activity. However, even at these levels the carboxylation reaction is still the bottleneck under high light conditions. Beside the low activity, the side reactivity with oxygen is another major flaw of the RuBisCO enzyme. In roughly every fifth reaction oxygen is incorporated instead of carbon dioxide[10]. In this case, the enzyme forms only one molecule of 3-

phosphoglycerate (3-PGA) and one molecule of 2-phosphoglycolate (2-PG) instead of two molecules of 3-PGA. 2-PG needs to be detoxified in a wasteful, yet essential, process called photorespiration.

To understand why RuBisCO became the enzyme it is today, we have to go back in history and realize that the earth has changed dramatically since the appearance of the first RuBisCO ancestors. While RuBisCOs provenance and why evolution chose it as the primary carbon dioxide fixing enzyme remains unclear, it is undoubted that its rise and properties are in close relation with the carbon dioxide and oxygen levels during the last three billion years[10]. At the beginning of this period, emerging phototrophic organisms adapted to use the abundance of water to satisfy their demand for hydrogen. As a result, they produced oxygen as a side product during water splitting. As far as we know today, there were no or almost no free oxygen molecules and life was generally adapted to anoxic conditions, which included a higher atmospheric carbon dioxide concentration than nowadays. To emphasize the impact of the production of elemental oxygen on most of the organisms, scientists named it the "Oxygen Catastrophe", the "Oxygen Crisis" or even the "Oxygen Holocaust"[11]. A more neutral and appropriate term would be "The Great Oxidation Event". Since most organisms were not metabolically equipped to accommodate the presence of the strong oxidizing power of free oxygen, the new trick these phototropic organisms learned led to an extinction of a vast majority of organisms. However, the adapted organisms could proliferate wherever they found light, water and carbon dioxide. Despite the Great Oxidation Event leading to vast amounts of oxygen, the "Second Great Oxidation Event" finally led to oxygen levels as they are present today[12].

1.3. Escaping the fate of RuBisCO

To increase the rate of natural carbon dioxide fixation and circumvent RuBisCO as the bottleneck of photosynthesis, scientists envisioned three main routes: Engineering RuBisCO towards a faster and more accurate carboxylase, engineering photorespiratory bypasses to compensate for RuBisCOs mistakes or replacing the CBB cycle with new-to-nature pathways for carbon fixation. Although the first option seems to be the most obvious one, so far no substantial parallel improvement of turnover rate and specificity could be achieved. As reviewed by Erb and Zarzycki, most evidence points towards a reciprocally linked specificity and activity that seems to be already close to the optimum[10,13]. However, an improved RuBisCO would still be the most elegant and simple way to be implemented in a broad range of crops and other photosynthetic organisms and could, once established, instantly play a crucial role in the reduction of atmospheric CO_2 levels. For the latter options, entire pathways need to be implemented, which is therefore even in model organisms more complex to achieve. Nonetheless, several new-to-nature solutions for photorespiratory bypasses and CO_2-fixation cycles

were published and partially realized[14-18] One of the cited examples, the crotonyl-coenzyme A (CoA)/ethylmalonyl-CoA/hydroxybutyryl-CoA (CETCH) cycle, is based on reductive carboxylation by the crotonyl-CoA carboxylase/reductase (Ccr). This enzyme belongs to the group of the well-studied enoyl-CoA carboxylases/reductases (Ecrs) and is substantially faster than RuBisCOs (Figure 1.). In comparison to RuBisCOs, Ccrs have no side-reactivity with O_2 and prevent reduction of crotonyl-CoA by shielding the active side from water molecules[19,20].

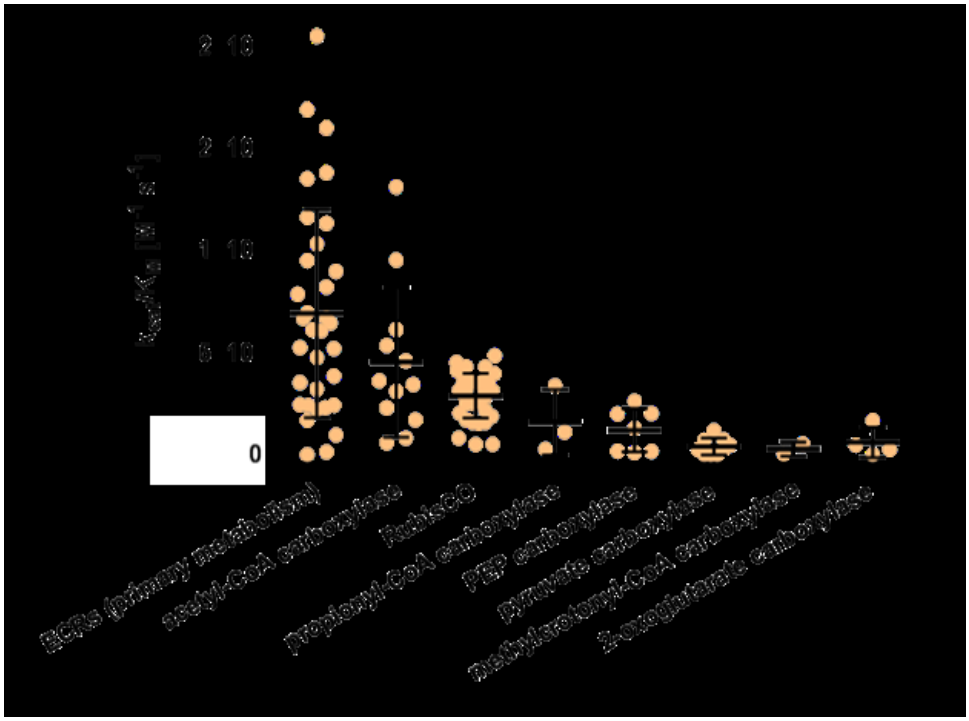


Figure 1. Catalytic efficiencies of carboxylases. Shown are catalytic efficiencies of different carboxylases (yellow dots) grouped in their respective classes. Figure adapted from Schwander et al., 2016[16].

1.4. The CETCH cycle, a man-made CO_2-fixation pathway

Published in 2016, the CETCH cycle was the first synthetic CO_2-fixation cycle realized *in vitro*[16]. Considering the design principles suggested by Bar-Even et al., the pathway was designed to have superior kinetics, is thermodynamically favored and is more efficient than natural CO_2-fixation cycles[14]. Additionally, only enzymes that work under aerobic conditions were used and not only existing, but also theoretically possible reactions were taken into account while designing the pathways. Several rounds of optimizations were performed (including enzyme engineering) to create the version CETCH 5.4. In this version the whole assays consists of 29 components, 17 of which are enzymes (12 core and 5 auxiliary enzymes) from nine different organisms covering all domains of life and including three engineered enzymes[16]. A major part of the cycle (from crotonyl-CoA to succinyl-

CoA) consists of reactions from the ethylmalonyl-CoA pathway[21]. The reaction sequence from succinyl-CoA to crotonyl-CoA resembles a part of the hydroxypropionate/hydroxybutyrate cycle.

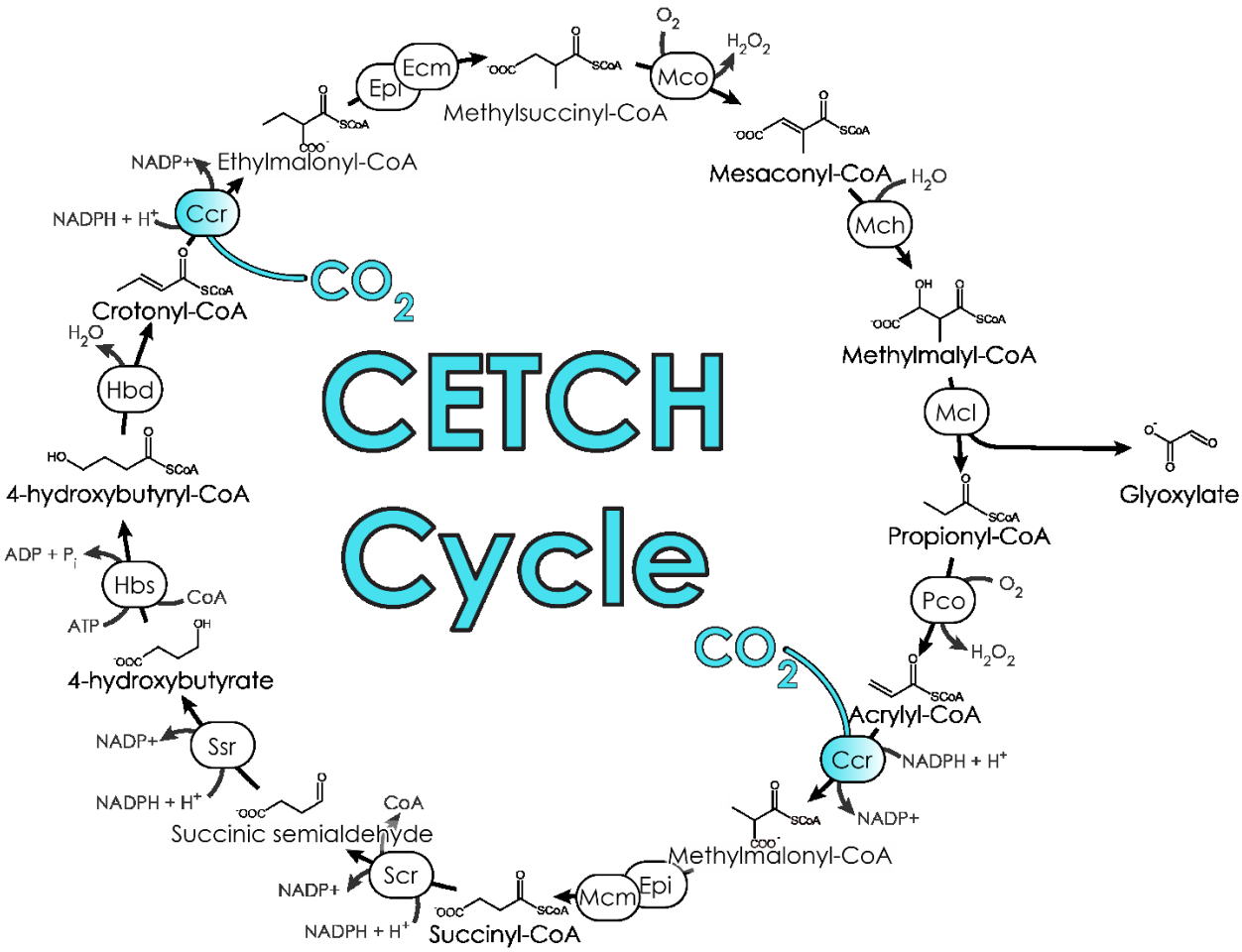


Figure 2. The Crotonyl-CoA-EThylmalonyl-CoA-4-Hydroxybutyryl-CoA Cycle. The CETCH core cycle as the version 5.4[16]. Shown are the 12 core enzymes (round boxes, full names see chapter 3.4.). The accessory enzymes catalase (hydrogen peroxide removal), Carbonic anhydrase (to provide carbon dioxide from bicarbonate), Formate dehydrogenase (for NADPH regeneration) and Creatine phosphokinase (for ATP regeneration) are not displayed.

Starting from propionyl-CoA, the first reaction of the CETCH cycle is catalyzed by an engineered a Flavin Adenosine Dinucleotide (FAD)-dependent short-chain acyl-CoA oxidase from *Arabidopsis thaliana* (Pco). The oxidation of propionyl-CoA to acrylyl-CoA enabled the use of Ccr as it accepts both crotonyl-CoA and acrylyl-CoA[22]. This made the use of the propionyl-CoA carboxylase (Pcc) from the initial draft of the cycle obsolete. The advantages were the consumption of one less ATP and rendering the pathway biotin independent. Because the original enzyme (Acx4) accepts 4-hydroxybutyryl-CoA (one of the intermediates) as a substrate, it was engineered towards a higher specificity for propionyl-CoA[16]. After the carboxylation of acrylyl-CoA by Ccr, the product (2*S*)-methylmalonyl-CoA is first

isomerized by the Ethylmalonyl-CoA/methylmalonyl-CoA epimerase (Epi) to the (2*R*)-methylmalonyl-CoA conformation. This is necessary since the Methylmalonyl-CoA mutase (Mcm) only accepts the (2*R*)-conformation to rearrange it to succinyl-CoA. Under release of CoA, succinyl-CoA is reduced to succinic semialdehyde, which is further reduced to 4-hydroxybutyrate. 4-hydroxybutyrate is activated to 4-hydroxybutyryl-CoA. This reaction is catalyzed by the 4-hydroxybutyryl-CoA synthetase from *Nitrosopumulus maritimus.* In CETCH version 5.4 this is the only ATP-dependent reaction. In contrast to other 4-hydroxybutyryl-CoA synthetases, this is an ADP-forming enzyme[23]. Another enzyme from *N. maritimus* catalyzes the next reaction, the dehydration from 4-hydroxybutyryl-CoA to crotonyl-CoA. The 4-hydroxybutyryl-CoA dehydratase has an iron-sulfur cluster and a homolog of *N. maritimus* was chosen because it was much more oxygen tolerant than other isoenzymes catalyzing this reaction[23]. In the next step, crotonyl-CoA is carboxylated by Ccr to ethylmalonyl-CoA. Similar to the transformation of methylmalonyl-CoA to succinyl-CoA, the product of Ccr, (2*S*)-ethylmalonyl-CoA, is converted to (2*R*)-ethylmalonyl-CoA by the bi-functional Epimerase and is further converted to methylsuccinyl-CoA by the Ethylmalonyl-CoA mutase (Ecm). In contrast to the Epimerase, each mutase (Mcm and Ecm) is highly specific for its respective substrate. It was speculated that the radical mechanism of these Coenzyme B_{12}-dependent reactions requires a tightly adapted active site to prevent the deactivation of the enzyme by the radical[22]. In nature, the oxidation of methylsuccinyl-CoA to mesaconyl-CoA is carried out by a FAD-dependent Methylsuccinyl-CoA dehydrogenase (Mcd), which shuttles the electrons via an Electron Transfer Flavoprotein (ETF) to the respiratory chain in the membrane[24]. To overcome the necessity of this system and build a pathway, which is feasibly *in vitro*, the design of the CETCH cycle included a Methylsuccinyl-CoA oxidase (Mco) to transfer the electrons directly onto molecular oxygen (O_2). As no Mco was known, the Mcd from *Rhodobacter sphaeroides* was engineered towards becoming a direct oxidase[24]. The product, mesaconyl-CoA, is then converted by the Mesaconyl-CoA hydratase (Mch) to β-methylmalyl-CoA, which is split by the corresponding lyase (Mcl1) into the starting substrate propionyl-CoA and glyoxylate. The former is thus, recycled and can be used for another round of the cycle, whereas glyoxylate represents the primary CO_2 fixation product of CETCH. Beside the enzymes of the core sequence, accessory enzymes were added: Carbonic anhydrase for faster equilibration of bicarbonate and carbon dioxide, Formate dehydrogenase for NADPH recycling, Polyphosphate kinase for ATP regeneration and Catalase for H_2O_2 detoxification[16].

In 2020 and 2021 even more complex versions of the CETCH cycle were published: First with another artificial CO_2-fixation module, the tatronyl-CoA (TaCo) pathway, for the production of the C3-compound glycerate[18]. Secondly, in combination with extracted thylakoid membranes from spinach to harvest the energy from light and use it to power the ATP- and NADPH-dependent reactions of the cycle[25]. These proof-of-principle studies show how natural and synthetic parts can be mixed and

matched to create novel reaction networks and inspired us to extend the product portfolio of the CETCH cycle towards more complex molecules.

1.5. Terpenes and polyketides, natures multipurpose tools

Two large groups of secondary metabolites with a very broad spectrum of functions are terpenes and polyketides. Compared to primary metabolites, secondary metabolites are often more complex and not needed for self-preservation. However, they play crucial roles for organisms in terms of self-defense, communication or as building blocks for hormones and occur among all domains of life[26-28]. The vast diversity of those compounds is derived from their modular synthesis. Terpenes are synthesized by the iterative usage of the isoprenoid precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP). Those building blocks are either derived from the mevalonate pathway, which converts three acetyl-CoA molecules into one IPP or DMAPP, or via the 2-*C*-methyl-D-erythritol
4-phosphate (MEP) pathway, using pyruvate and glyceraldehyde 3-phosphate[29-31]. Starting from IPP and DMAPP, terpenes can consist of a single isoprene unit such as hemiterpene but are usually iteratively added to form monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20), et cetera. While a multiple of five is the common form, deviations are possible. Additionally, natural and synthetic modifications diversify the range of terpenes with bioactive properties[32,33]. In the second chapter of this thesis, "A modular in vitro platform for the production of terpenes and polyketides from CO_2", we could show how these compounds are directly produced from CO_2.

While terpenes are synthesized from isoprenoid precursors, polyketide synthases (PKSs) use CoA-thioesters as their building blocks. These PKSs are large reaction complexes with either multiple domains or multiple proteins working together. One of the best-studied examples is the 6-Deoxyerythronolide B Synthase (DEBS), which belongs to the class I PKSs (of three). The sheer numbers highlight the complexity of this biological machinery: DEBS consists of three proteins of which each has different modules and a total of 28 active sites[34]. Using one molecule of propionyl-CoA as a starter unit and six molecules of methylmalonyl-CoA as extender units the final product 6-deoxyerythronolide B (6-dEB) consists of 21 carbons. During this process, the six methylmalonyl-CoAs are attached successively via decarboxylative claisen condensation, resulting in the release of the CoA moieties, six CO_2 molecules and the consumption of six reducing equivalents NADPH. Although in this reaction sequence propionyl-CoA and methylmalonyl-CoA are used, other common starter and extender units are for example acetyl- and malonyl-CoA[35]. While complexity hampers the complete understanding of those convoluted molecular machines, it also harbors great potential for engineering and the

subsequent creation of new-to-nature molecules by integration of non-natural starter and extender units. The existence of several CoA-thioesters as intermediates of the CETCH cycle sparked the question whether those could be harnessed for the production of polyketides.

1.6. Anaplerosis, in-flight refueling for cells

The drainage of intermediates from pathways for several purposes is a common scheme in metabolic networks. The interconnection of different pathways into a greater system is necessary to guarantee the flexibility and additivity of cells. Primary metabolism and especially some key nodes like the tricarboxylic acid (TCA) cycle however, need special protection to assure the integrity. More than 50 years ago, the term anaplerosis was introduced to describe one or more reactions that supply missing intermediates to pathways which stall upon a lack of them[36]. One of the most prominent examples is the carboxylation of pyruvate or phosphoenolpyruvate to oxaloacetate to prevent stalling the TCA cycle prior to the condensation yielding citrate. The accumulation of acetyl-CoA is an indicator for the absence of oxaloacetate and activates the pyruvate- and phosphoenolpyruvate carboxylase. Beside single reactions like the carboxylation of (phosphoenol)pyruvate or the dehydrogenation of glutamate to oxoglutarate, there are several anaplerotic reaction sequences, such as the ethylmalonyl-CoA, the glyoxylate or the methylaspartate pathway[21,37,38]. Despite the relevance of anaplerosis for the flexibility and additivity, so far no synthetic approach of this scheme was applied for *in vitro* networks. In chapter three, “Enhancing the synthetic capabilities of a complex in vitro metabolic network through anaplerotic reaction modules”, the principle of anaplerosis was employed to grant access to CETCH intermediates and use them for the production of the C21-polyketide 6-dEB.

1.7. Optimization of biological systems with machine learning: The machines are taking over!

The increasing complexity of man-made reaction networks often goes beyond the scope of traceability and is in contradiction to the initial concept of a simple and easily controlled environment. Machine learning algorithms are therefore used to optimize, predict and understand complex interactions of biological systems like cell-free expression systems, *in vivo* gene expression, drug discovery or protein engineering [39-42]. The history of machine learning traces back to the early days of modern computers in the 1950s. Interestingly, some of today’s terms like “machine learning”, “artificial neuron”, “neural network” or “perceptrons” were already defined and partially applied back then[43-45]. In 1959, Arthur

L. Samuel published his work at IBM where he programmed a computer to train its ability to play the game checkers. He summarized his work in the abstract as follows:

"Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkable short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations." [45]

Subsequently, the work on machine learning continued and several key components paved the way for its success; the democratization of computers and the emergence of the internet were the main reasons for the development of more and more algorithms. Nowadays artificial intelligence and deep learning are employed to solve some of the biggest questions of the 21st century, such as the protein structure prediction targeted by AlphaFold[46]. Despite of numerous publications on diverse topics and problems, the usage in laboratories with no background in bioinformatics is very limited. This might be due to either the complexity of such applications or their high pricing if commercialized. Since the prices for DNA synthesis and other techniques are dropping and more and more affordable screening methods become available, the need for easy to use optimization tools will grow as an increasing number of research groups can afford the generation of large data sets.

A promising machine learning tool for various applications is the eXtreme Gradient Boosting (XGBoost) package, which was used in several scripts that won recent coding challenges and showed superior performance in comparison to other algorithms (Figure 3.)[47,48]. XGBoost is a gradient boosting algorithm, which was developed for the optimized (extreme) use of resources (hardware). Gradient boosting is mainly done with decision trees as weak learners, which are subsequently ensembled to minimize the loss function and can be used for regression and classification problems. It features several characteristics, which make it a favorable application as a general optimization tool for biological systems. It works well with smaller datasets due to its sparsity awareness but can handle large datasets with considerably low hardware specifications due to the optimized use of them. Furthermore, it can be used with automated hyperparameter tuning, which finds the best parameters to fit the data. The optimal use of resources together with its scalability makes it currently one of the most popular machine learning algorithms. In chapter one, "A versatile active learning workflow for

optimization of genetic and metabolic networks", we show how the XGBoost algorithm was integrated in a user-friendly layout and used to optimize the parameters of the CETCH cycle.

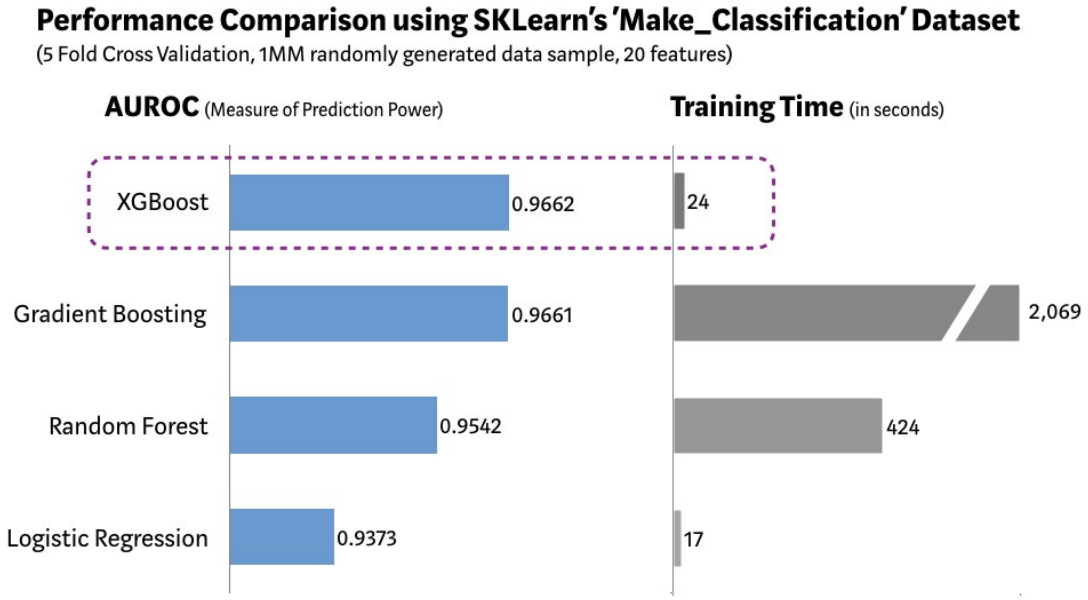


Figure 3. Performance of XGBoost gradient boosting in comparison with other popular algorithms. When benchmarked with SKLearn's "Make_Classification" dataset, XGBoost showed a superior performance in classification, while being significantly faster than similar performing algorithms. Figure from Vishal Morde[48]

1.8. References

1. Kelemen, P.B. & Manning, C.E. Reevaluating carbon fluxes in subduction zones, what goes down, mostly comes up. *Proc Natl Acad Sci U S A* **112**, E3997-4006 (2015).
2. Petkowski, J.J., Bains, W. & Seager, S. On the Potential of Silicon as a Building Block for Life. *Life (Basel)* **10**(2020).
3. Callendar, G.S. The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society* **64**, 223-240 (1938).
4. Edenhofer, O. et al. IPCC 2014; Climate Change 2014: Mitigation of Climate Change. (2014).
5. Solomon, S. et al. Fourth Assessment Report of the Intergovernmental Panel on Climate Change. in *Cambridge University Press* (2007).
6. Portis, A.R., Jr. & Parry, M.A. Discoveries in Rubisco (Ribulose 1,5-bisphosphate carboxylase/oxygenase): a historical perspective. *Photosynth Res* **94**, 121-43 (2007).
7. Ellis, R.J. The most abundant protein in the world. *Trends in Biochemical Sciences* **4**, 241-244 (1979).
8. Kung, S. & Marsho, T. Regulation of RuDP carboxylase/oxygenase activity and its relationship to plant photorespiration. *Nature* **259**, 325-326 (1976).
9. Bar-On, Y.M. & Milo, R. The global mass and average rate of rubisco. *Proc Natl Acad Sci U S A* **116**, 4738-4743 (2019).
10. Erb, T.J. & Zarzycki, J. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Curr Opin Biotechnol* **49**, 100-107 (2018).
11. Margulis, L. & Sagan, D. Chapter 6. The Oxygen Holocaust. in *Microcosmos: Four Billion Years of Evolution from Our Microbial Ancestors* 99-115 (University of California Press, 1986).
12. Holland, H.D. The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B Biol Sci* **361**, 903-15 (2006).
13. Tcherkez, G.G., Farquhar, G.D. & Andrews, T.J. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences* **103**, 7246-7251 (2006).
14. Bar-Even, A., Noor, E., Lewis, N.E. & Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences* **107**, 8889-8894 (2010).
15. Shih, P.M., Zarzycki, J., Niyogi, K.K. & Kerfeld, C.A. Introduction of a synthetic CO(2)-fixing photorespiratory bypass into a cyanobacterium. *J Biol Chem* **289**, 9493-500 (2014).
16. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
17. Trudeau, D.L. et al. Design and in vitro realization of carbon-conserving photorespiration. *Proc Natl Acad Sci U S A* **115**, E11455-E11464 (2018).
18. Scheffen, M. et al. A new-to-nature carboxylation module to improve natural and synthetic CO2 fixation. *Nature Catalysis* (2021).
19. Peter, D.M. et al. Screening and engineering the synthetic potential of carboxylating reductases from central metabolism and polyketide biosynthesis. *Angewandte Chemie International Edition* **54**, 13457-13461 (2015).
20. Stoffel, G.M. et al. Four amino acids define the CO2 binding pocket of enoyl-CoA carboxylases/reductases. *Proceedings of the National Academy of Sciences* **116**, 13964-13969 (2019).
21. Erb, T.J. et al. Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway. *Proceedings of the National Academy of Sciences* **104**, 10631-10636 (2007).
22. Erb, T.J., Retey, J., Fuchs, G. & Alber, B.E. Ethylmalonyl-CoA mutase from Rhodobacter sphaeroides defines a new subclade of coenzyme B12-dependent acyl-CoA mutases. *J Biol Chem* **283**, 32283-93 (2008).
23. Konneke, M. et al. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proc Natl Acad Sci U S A* **111**, 8239-44 (2014).

24. Burgener, S., Schwander, T., Romero, E., Fraaije, M.W. & Erb, T.J. Molecular Basis for Converting (2S)-Methylsuccinyl-CoA Dehydrogenase into an Oxidase. *Molecules* **23**(2017).
25. Miller, T.E. et al. Light-powered CO2 fixation in a chloroplast mimic with natural and synthetic parts. *Science* **368**, 649-654 (2020).
26. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl* **48**, 4688-716 (2009).
27. Matsumi, R., Atomi, H., Driessen, A.J. & van der Oost, J. Isoprenoid biosynthesis in Archaea--biochemical and evolutionary implications. *Res Microbiol* **162**, 39-52 (2011).
28. Oldfield, E. & Lin, F.Y. Terpene biosynthesis: modularity rules. *Angew Chem Int Ed Engl* **51**, 1124-37 (2012).
29. Bloch, K. The biological synthesis of cholesterol. *Science* **150**, 19-28 (1965).
30. Arigoni, D. et al. Terpenoid biosynthesis from 1-deoxy-D-xylulose in higher plants by intramolecular skeletal rearrangement. *Proc Natl Acad Sci U S A* **94**, 10600-5 (1997).
31. Lichtenthaler, H.K. The 1-Deoxy-D-Xylulose-5-Phosphate Pathway of Isoprenoid Biosynthesis in Plants. *Annu Rev Plant Physiol Plant Mol Biol* **50**, 47-65 (1999).
32. Rivas, F., Parra, A., Martinez, A. & Garcia-Granados, A. Enzymatic glycosylation of terpenoids. *Phytochemistry Reviews* **12**, 327-339 (2013).
33. Eiben, C.B. et al. Mevalonate Pathway Promiscuity Enables Noncanonical Terpene Production. *ACS Synth Biol* **8**, 2238-2247 (2019).
34. Khosla, C., Tang, Y., Chen, A.Y., Schnarr, N.A. & Cane, D.E. Structure and mechanism of the 6-deoxyerythronolide B synthase. *Annu Rev Biochem* **76**, 195-221 (2007).
35. Staunton, J. & Weissman, K.J. Polyketide biosynthesis: a millennium review. *Natural product reports* **18**, 380-416 (2001).
36. Kornberg, H. Anaplerotic sequences in microbial metabolism. *Angewandte Chemie International Edition in English* **4**, 558-565 (1965).
37. Kornberg, H. The role and control of the glyoxylate cycle in Escherichia coli. *Biochemical Journal* **99**, 1 (1966).
38. Khomyakova, M., Bukmez, O., Thomas, L.K., Erb, T.J. & Berg, I.A. A methylaspartate cycle in haloarchaea. *Science* **331**, 334-7 (2011).
39. Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687-694 (2019).
40. Zrimec, J. et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun* **11**, 6141 (2020).
41. Borkowski, O. et al. Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun* **11**, 1872 (2020).
42. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2**, 573-584 (2020).
43. McCulloch, W.S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. 1943. *Bull Math Biol* **52**, 99-115; discussion 73-97 (1943).
44. Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, (Cornell Aeronautical Laboratory, 1957).
45. Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **44**, 206-226 (1959).
46. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
47. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).
48. Morde, V. & Setty, V.A. XGBoost Algorithm: Long May She Reign! (2019).

2. A versatile active learning workflow for optimization of genetic and metabolic networks

Amir Pandi[1]*, Christoph Diehl[1]*, Ali Yazdizadeh Kharrazi[2], Léon Faure[3], Scott A. Scholz[1], Maren Nattermann[1], David Adam[1], Nils Chapin[1], Yeganeh Foroughijabbari[1], Charles Moritz[1], Nicole Paczia[4], Niña Socorro Cortina[1,5], Jean-Loup Faulon[3,6,7], and Tobias J. Erb[1,8]

[1] Department of Biochemistry & Synthetic Metabolism, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[2] DataChef, Amsterdam, The Netherlands

[3] Micalis Institute, INRAE, AgroParisTech, University of Paris-Saclay, Jouy-en-Josas, France

[4] Core Facility for Metabolomics and Small Molecule Mass Spectrometry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[5] LiVeritas Biosciences, Inc., South San Francisco, USA

[6] Genomique Metabolique, Genoscope, Institut Francois Jacob, CEA, CNRS, Univ Evry, University of Paris-Saclay, Evry, France

[7] Manchester Institute of Biotechnology, SYNBIOCHEM center, School of Chemistry, The University of Manchester, Manchester, UK

[8] SYNMIKRO Center of Synthetic Microbiology, Marburg, Germany

* Amir Pandi and Christoph Diehl contributed equally to this work

Author contributions

T.J.E. and A.P. conceived the work. **A.P. and C.D. designed the general workflow and setup of the optimization tool.** A.P. assessed the performance of the different algorithms, established the workflow for the optimization of the *LacI* gene circuit and the optimization of a transcription & translation unit. **C.D. designed the workflow for the optimization of the CETCH cycle and performed all related experiments.** A.Y.K. wrote all Python notebooks and simulations. L.F. reviewed the workflow's code and consistency and provided constructive inputs. S.A.S. performed western blot and RT-qPCR experiments. M.N. prepared the enzyme mutants dataset. D.A, N.C., Y.F., and C.M. assisted with the experimental work. N.P. developed the LC-MS methods for the analysis of glycolate and CoA esters. N.S.C. wrote the code for the conversion of the ECHO® exceptions lists into new worklists. J-L.F. and T.J.E. supervised and directed the work. **A.P., C.D., A.Y.K. and T.J.E. wrote the manuscript with input from all other authors.**

2.1. Abstract

The study, engineering and application of biological networks require practical and efficient approaches. Current optimization efforts of these networks are often limited by wet lab labor and cost, as well as the lack of convenient, easily adoptable computational tools. Aimed at democratization and standardization, we describe METIS, a modular and versatile active machine learning workflow with a simple online interface for the optimization of biological target functions with minimal experimental datasets. We demonstrate our workflow for various applications, from simple to complex gene circuits and metabolic networks, including several cell-free transcription and translation systems, a *LacI*-based multi-level controller and a 27-variable synthetic CO_2-fixation cycle (CETCH cycle). Using METIS, we could improve above systems between one and two orders of magnitude compared to their original setup with minimal experimental efforts. For the CETCH cycle, we explored the combinatorial space of ~10^{25} conditions with only 1,000 experiments to yield the most efficient CO_2-fixation cascade described to date. Beyond optimization, our workflow also quantifies the relative importance of individual factors to the performance of a system. This allows to identify so far unknown interactions and bottlenecks in complex systems, which paves the way for their hypothesis-driven improvement, which we demonstrate for the *LacI* multi-level controller that we were able to improve by more than 30-fold after having identified resource competition as limiting factor. Overall, our workflow opens the way for convenient optimization and prototyping of genetic and metabolic networks with customizable adjustments according to user experience, experimental setup, and laboratory facilities.

2.2. Introduction

The understanding and engineering of biological systems require practical and efficient experimental approaches[1-5]. Machine learning algorithms hold a big promise for the study, design, and optimization of different biological systems[6-9], including genomics studies[10-12], protein, enzyme and metabolic engineering[4,13,14], prediction and optimization of CRISPR sequences and proteins[15-18], as well as complex genetic circuits design and optimization[19-21]. Yet, applying machine learning is limited by the need for informatics expertise and large user-labeled datasets, which are typically time-, labor- and cost-intense.

Active learning, sometimes called optimal experimental design[22,23], is a type of machine learning that interactively suggests a next set of experiments after being trained on previous results[24]. This makes active learning valuable for wet-lab scientists, especially when dealing with a limited number of user-labeled

data[25]. Active learning approaches reduce experimental time, labor and cost and have been used in cellular imaging[26], systems biology[27], biochemistry[28-30], and synthetic biology[31]. Despite these examples, a challenge in applying active learning methods for experimental biologists is the lack of customizable programs and workflows.

Here, we describe METIS (Machine learning guided Experimental Trials for Improvement of Systems, named after the ancient goddess of wisdom and crafts Μῆτις, *lit.* "wise counsel"), a modular and versatile active machine learning workflow for optimization of a biological objective function (an output/target that depends on multiple factors) with minimal datasets. We created METIS for experimentalists with no experience in programming, who can use the entire process of personalized active learning, experimental setup, data analysis and visualization without any advanced computational skills. METIS runs on Google Colab, a free online platform to write and execute Python codes developed for education, data science, and machine learning purposes[32]. The open platform does not need any installation and registration and can be simply used via a personal copy of the respective notebook.

To establish the workflow, we first assessed the performance of different machine learning algorithms on a minimal training dataset and experimentally validated the best performing algorithm (XGBoost) by optimization of an *in vitro* cell-free transcription-translation (TXTL) system of *Escherichia coli* that is commonly used in cell-free synthetic biology for a variety of applications[33], including biosensor development[34], metabolic pathway prototyping[35], and gene circuit design[36]. We then developed the modular architecture of METIS for user-defined applications through the customization of different parameters and factors.

We showcase the versatility of METIS on various biological systems, starting with an *in vitro* gene circuit. Gene circuits have recently received attention (e.g., as biosensors), but are still limited in their applicability due to their poor *in vitro* performance[35,37]. Applying our workflow, we could improve the activity of a recently reported *LacI*-based multi-level controller[38] by two orders of magnitude, notably by identifying and overcoming a fundamental bottleneck (i.e., resource competition) in the design of the system. We further demonstrate ten-fold improved protein production from an optimized transcription & translation unit, demonstrating that our workflow can be used for biological sequences based on categorical factors (i.e., combinatorial variants of a T7 promoter, ribosome binding site (RBS), N- and C-terminal amino acids). Finally, we use METIS to improve a complex metabolic network, the so-called crotonyl-CoA/ethylmalonyl-CoA/hydroxybutyryl-CoA (CETCH)[39] cycle, a new-to-nature synthetic CO_2-fixation cycle, comprising 17 different enzymes plus 10 different cofactors and components, which was shown to be

(thermodynamically) more efficient compared to natural photosynthesis. Yet, the network's full kinetic potential had not been exploited so far, as efficient strategies to explore its combinatorial space had been lacking so far. Using METIS allowed us to improve productivity of the CETCH cycle by ten-fold with (only) 1,000 experiments, resulting in the most efficient CO_2-fixing *in vitro* system described to date. Overall, these results demonstrate the ability of our workflow for the optimization of various complex biological networks with minimal experimental efforts, providing multiple opportunities for the study and engineering of different biological systems in the future.

2.3. Results

Assessing the performance of different algorithms for our workflow

We first tested which machine learning algorithm would perform best with a limited number of experimental data typical for a standard research lab setup. To that end, we took advantage of an existing dataset from a recent optimization of an *E. coli* extract-based *in vitro* TXTL system[31]. In their study, Borkowski *et al*. optimized cell-free protein production in *E. coli* lysate by varying 12 different factors including salts, energy mix, amino acids, and tRNAs, and measuring production yield of Gfp (produced by a plasmid expressing *Gfp*) as output. Altogether, the dataset encompassed around 1000 data points. We fitted the dataset to obtain a standard as a gold regressor and divided it further into test and training sets, with 20% and 80% of data, respectively. While the latter set was used to train the model, the test set was used to validate the regressor (**Methods**).

We used the gold regressor to assess the performance of four different machine learning algorithms over 10 rounds of active learning (**Fig. 1a**). The tested algorithms included deep neural networks (DNN), multilayer perceptrons (MLP), linear regressors, and XGBoost gradient boosting, which all show different capabilities for a given problem set and its data sample size. Over 10 rounds of active learning with 100 data points in each round, XGBoost and linear regressors showed better performance (**Fig. 1b**) compared to DNN and MLP, which generally need larger datasets to outperform other models[40].

For our workflow, we selected XGBoost, which is an improved random forest-type algorithm, working through gradient boosted decision trees[41] by aggregating and compiling sets of models. This makes XGBoost a fast and powerful algorithm that performs efficiently even with small datasets as shown recently for different biological applications[18,42,43]. To determine the minimum dataset required for

optimization, we compared active learning rounds with 5, 10, 25, and 100 data points. Notably, a sample size as low as 10 data points still allowed sufficient yield optimization (i.e., in the scale of the original study[31]) within 10 learning cycles (**Fig. 1b**).

Testing the workflow with minimal experimental work

Having validated the workflow with an existing data set, we next sought to test it in a real-world experimental setup, simulating a situation in which the number of combinations that can be tested is limited by available equipment, readout and experimental cost. We chose (again) to optimize relative Gfp production (yield) in an *E. coli* lysate TXTL system (**Fig. 1c**) that consists of 13 variable factors (components).

To optimize composition of the TXTL system, we defined a concentration range for each of the 13 factors (**Code availability**), and performed an active learning process over 10 rounds with 20 experiments per round (**Fig. 1d**, see **Supplementary Note 1** for details) quantifying Gfp yield (i.e., Gfp fluorescence reported from each composition normalized by the Gfp fluorescence of the standard composition[33]), as objective function. Over 10 rounds of active learning, the relative yield increased up to 20 and the median increased from zero to over 10 in the 9th round (**Fig. 1e**). Note that low-yield data points (even those observed in the late learning cycles) are equally informative as high-yield ones, because they allow to explore the landscape around and beyond local maxima, as defined by the exploration to exploitation ratio of our workflow that we fully discuss in **Supplementary Note 2**. Using a standard curve for purified Gfp, we calculated that for the highest yields (reaching 15 in **Fig. 1e**) the production of Gfp was 30-fold higher than in previous optimizations[31] and the myTXTL® commercial kit (**Supplementary Fig. 1**).

Beyond the simple optimization of a given system, our workflow can also quantify the contribution of different factors during optimization. **Fig. 1f** represents feature importance, i.e., the effect of each individual factor on the objective function. The importance is given as a relative fraction (or percentage) in the prediction of the values of the objective function by the model, with the sum of all factors set to 100%. Our analysis showed that tRNA mix and Mg-glutamate were the most important components in optimizing Gfp yield, while cAMP and NAD were the least important contributors. **Fig. 1g** shows the distribution of Gfp yield at different concentrations of individual factors. Decreasing concentrations of tRNA and NTP mixes correlated with high yield, while PEG 8000, Mg-glutamate, 3-PGA, folinic acid, and spermidine showed similar effects at increasing concentrations. Together, these data did not only result in an optimized TXTL system but also allowed to identify the most crucial components during system

optimization, providing the basis for a deeper understanding of the system itself. All combinations and yields are provided as results files for each experimental round (**Data availability**), and the Google Colab notebook with all analyses and visualization modules are also accessible (**Code availability**).

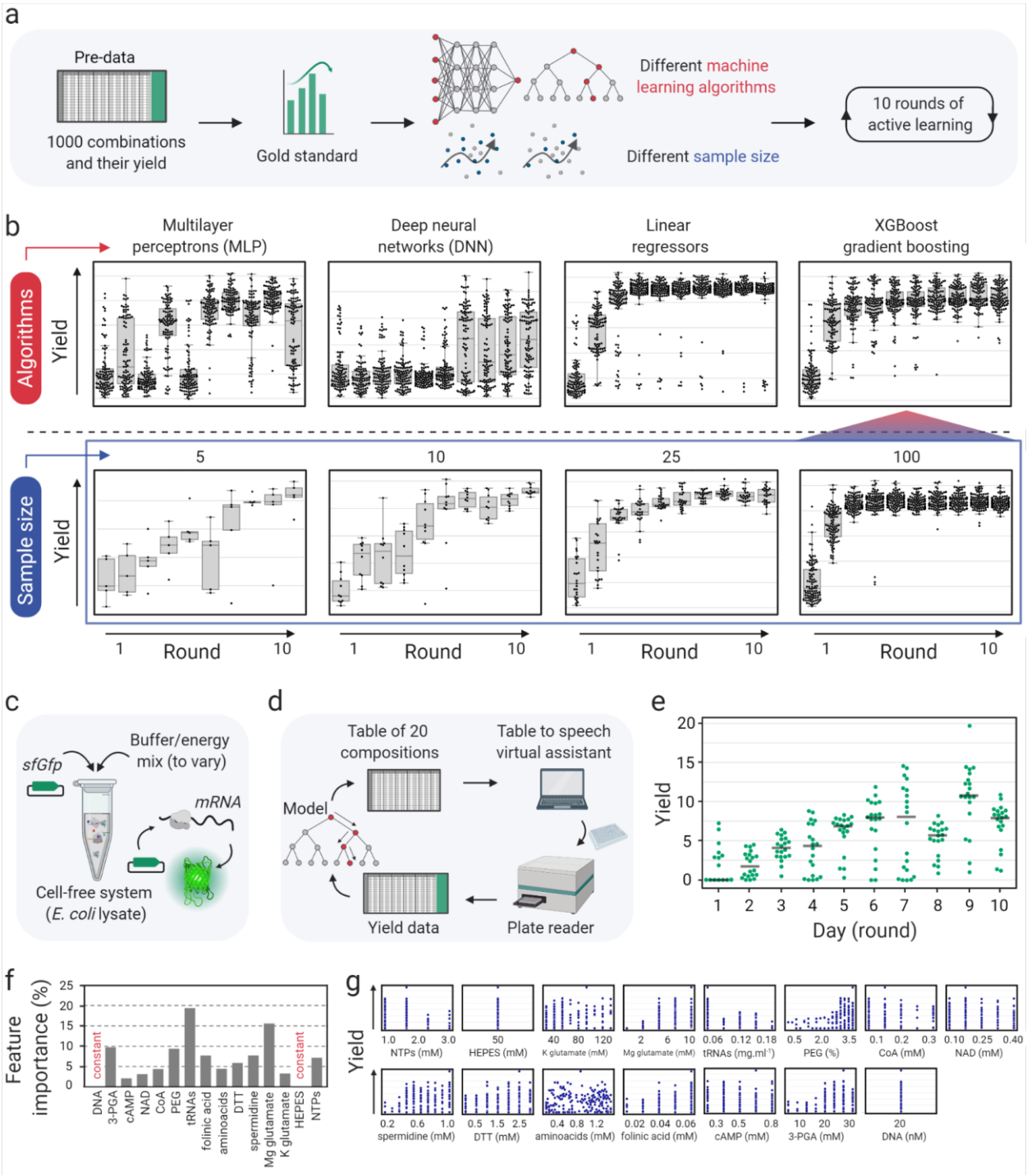


Fig. 1: Assessing the performance of different algorithms and testing the active learning workflow with minimal data points. a) An existing dataset of cell-free gene expression compositions composed of 1000 data points was used to build a gold standard regressor and assess the performance of different machine learning algorithms in 10 rounds of active learning. Different

algorithms suggest combinations and predict their associated yields and the gold standard regressor assigns the yields to evaluate the prediction. **b)** Top panel: performance of 4 algorithms, multilayer perceptrons (MLP), deep neural networks (DNN), linear regressors, and XGBoost gradient boosting in 10 rounds of active learning (100 data points per round). Bottom panel: performance of the XGBoost gradient boosting algorithm as the selected algorithm with different sample sizes, 5, 10, 25, and 100 per round. **c)** An *in vitro* or cell-free transcription-translation (TXTL) system (based on *E. coli* lysate) to test the workflow with 20 data points per round. 20 nM plasmid expressing *sfGfp* (super-folder *Gfp*) with promoter J23101 and RBS B0032 were added to the cell-free reaction mix along with 13 components of reaction buffer and energy mix. **d)** Overview of the active learning cycle. 13 components are varied starting with random compositions and over 10 rounds of results are imported to the model, which learns and suggests new compositions for improvement of the objective function. **e)** The plot presenting the average of triplicates of the objective function (yield) for compositions in 10 rounds (days) of active learning. **f)** Feature importance percentages show the effect of each factor on the model's decision to calculate yields for the suggested compositions. **g)** Distribution of different concentrations of each factor within the measured yields. See also **Supplementary Fig. 2, 3** for variation of concentration of each factor from round 1 to 10, and mutual interactions between factors, respectively.

Development of a user-friendly, versatile modular architecture for our workflow

After demonstrating that our workflow is capable of working efficiently with minimal datasets, we sought to build a modular architecture that can be easily applied for the optimization of different biological objective functions. We implemented our workflow in Google Colab Python notebooks that can be accessed by the user –without installation or registration– simply through a personal copy of the notebook from a web browser. Defining the objective function and the variable factors (**Fig. 2a**), the user can simply open the link of Google Colab notebook and directly use the workflow as shown in **Fig. 2a, b**, **Supplementary Fig. 4-6**.

In **Supplementary Note 2,** we provide a detailed description of all features of METIS. The modular workflow enables the use of factors with numerical values (examples in **Fig. 1, 3, 5**), categories (examples in **Fig. 4, Supplementary Fig. 19**), or both (example in **Fig. 3, 5**). Active learning can be initialized by random combinations generated by the workflow in the first round (example in **Fig. 1, 3, 4**). Alternatively, pre-existing datasets can be imported and used for optimization or simulations (examples in **Supplementary Fig. 18, 19**). Although our workflow is designed as an active learning approach over iterative experimental rounds, it can be also used in a classical machine learning setup, when only using one round of experiments. Multiple data analysis and visualization modules are available that can be used in each round of active learning as shown in example applications (**Fig. 2b**, **Supplementary Note 2**). The workflow is able to generate a pipetting table output (exemplified for the experiments in **Fig, 1, 3**), which alongside our table-to-speech virtual assistant tool, improves the speed and accuracy of manual pipetting (**Supplementary Note 1, 2**). For more complex experiments where multiple components in different volumes are required, the workflow can be interfaced with lab automation (e.g., an Echo® acoustic liquid handling robot, see optimization of the CETCH cycle in **Fig. 5**).

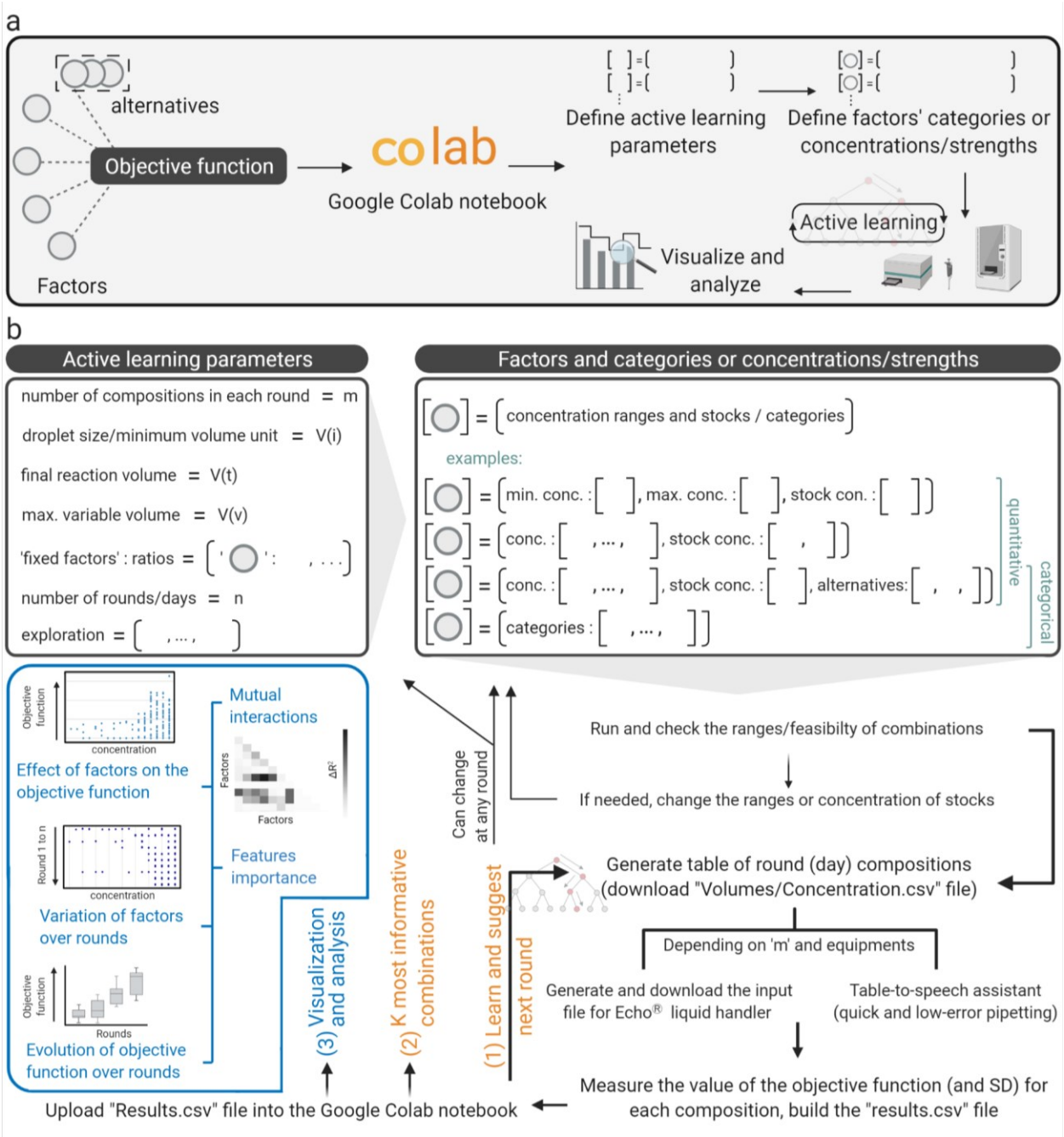


Fig. 2: A representation of METIS (active machine learning for biological systems) modular workflow. a) The first step is choosing an objective function (an output/target that depends on multiple factors), then continuing with the Google Colab Python notebook, performing experiments, and visualizing and analyzing results. **b)** Users should define active learning parameters depending on the application, equipment, and the size of the combinatorial space. Factors' ranges/categories are conditions that are varied to explore the behavior of the objective function. In each round of active learning, while the users perform experiments and label the suggested combinations with measured objective function values (parameters and factors' conditions can be readjusted at any round), the data can be analyzed and visualized using the workflow's modules. See **Supplementary Note 2** for a detailed explanation and guide for each step and also **Supplementary Fig. 4-6**.

Application of the modular workflow for optimization of a *LacI* gene circuit

Next, we aimed to apply METIS for optimization of *LacI*-based gene circuits that were described recently[43]. Greco *et al*. developed a strategy for stringent gene expression by engineering transcriptional and/or translational small RNA inhibitors upstream of a *Gfp* reporter gene under the control of the pTAC promoter (**Fig. 3a**). Starting from a standard pTAC architecture, a so-called single-level controller (SLC), Greco *et al.* constructed three different multi-level controllers (MLC): pTHS (toehold switch; translational control), pSTAR (small transcription activating RNA; transcriptional control), and pDC (double controller; transcriptional and translational control)[38]. Notably, the authors could improve the rate of *in vitro* protein production by 35-fold with different MLC designs. Yet in these efforts, the fold-change in total protein production remained low (**Supplementary Fig. 7**), which was likely the result of leaky repressor-regulated promoters in the OFF state, as noted earlier[34,37]. A high fold-change in protein production, however, would be strongly desired for application of gene circuits, e.g., as diagnostic sensors, where a high signal-to-noise ratio is important.

Here, we aimed at using our workflow to increase both the dynamic range and fold-change of *in vitro* protein production for the SLC and MLC circuits (**Fig. 3b**). To improve fold-change and suppress leaky protein production, we supplied an additional plasmid expressing *lacI* under the control of a T7 promoter (transcribed by purified T7 RNA polymerase). The active learning cycle received input from several factors in the *E. coli* cell-free system; amino acids and tRNAs, which are important when extra DNA is added, DTT as reducing reagent, spermidine for DNA-protein binding, and PEG 8000 as crowding agent. We performed 10 rounds of active learning with the objective function of fold-change (FC) $\times$ dynamic range (DR) of Gfp (**Fig. 3a**), to score those compositions that result not only in a high fold-change but also total Gfp produced. While the objective function (fold-change (FC) $\times$ dynamic range (DR), bottom plot in **Fig. 3c**) improved during the active learning cycle, we did not observe a substantial improvement in fold-change of Gfp production alone (upper plot in **Fig. 3c**). Feature importance analysis identified the concentration of the P_{T7}-*LacI* plasmid as strong contributor (**Fig. 3d, Fig. 3e, Supplementary Fig. 8**), indicating deleterious LacI-protein/DNA interactions or resource limitation of the TXTL system through production of the lacI protein[44].

Performing a titration experiment with P_{T7}-*LacI*, we could show that addition of the *LacI* plasmid has indeed a strong negative effect on Gfp production (**Fig. 3f, g**, see also **Supplementary Note 3** for details of the active learning cycle and titration experiments). To further investigate this effect, we titrated the

LacI plasmid with either T7 or a constitutive promoter against a fixed concentration of the *Gfp* expression plasmid under control of either T7 or a constitutive promoter. While increasing concentrations of the plasmid with constitutive *LacI* expression did only slightly affect *Gfp* expression from the T7 promoter, increasing concentrations of *LacI* plasmid under T7 control strongly affected Gfp production, especially when Gfp was expressed from the constitutive promoter (**Fig. 3h**). These results indicated a resource competition between the two plasmids, according to which the T7 promoter wins competition at the transcriptional and consequently the translational level. Quantifying the levels of *Gfp* and *LacI* mRNA by qPCR confirmed a direct correlation between mRNA and Gfp production levels, further supporting the resource competition hypothesis (**Fig. 3i**).

To overcome resource competition, we tested purified LacI protein instead of the *LacI* plasmid in the TXTL system, which resulted in improved Gfp productivity (**Supplementary Fig. 9**). Thus, we sought to optimize Gfp fold-change with using purified LacI protein instead of a *LacI* expression plasmid. Using a module of METIS called "K most informative combinations" (with the number K to be defined by the user), we extracted the 20 most informative combinations of the active learning cycle, and repeated these 20 setup by replacing P_{T7}-*LacI* plasmid with purified LacI protein (**Fig. 3j**), resulting in a strong improvement in the objective function, and in particular Gfp fold-change. Continuing with only one additional round of active learning using this dataset, we were able to improve the fold-change to up to 123 (**Fig. 3k**), which is 15-fold improvement compared to that of 10 rounds of active learning with the P_{T7}-*LacI* plasmid and 34-fold improvement compared to the initial setup.

Overall, these experiments demonstrated how our workflow can be used to improve the signal-to-noise-ratio of an existing *in vitro* gene circuit by two orders of magnitude. Notably, the feature importance module of METIS, which identified apparent bottlenecks (i.e., resource competition by the *LacI* plasmid) and the K most informative combinations module of the workflow were crucial for success. A Google Colab notebook and all combinations and results are provided through **Code and Data availability**.

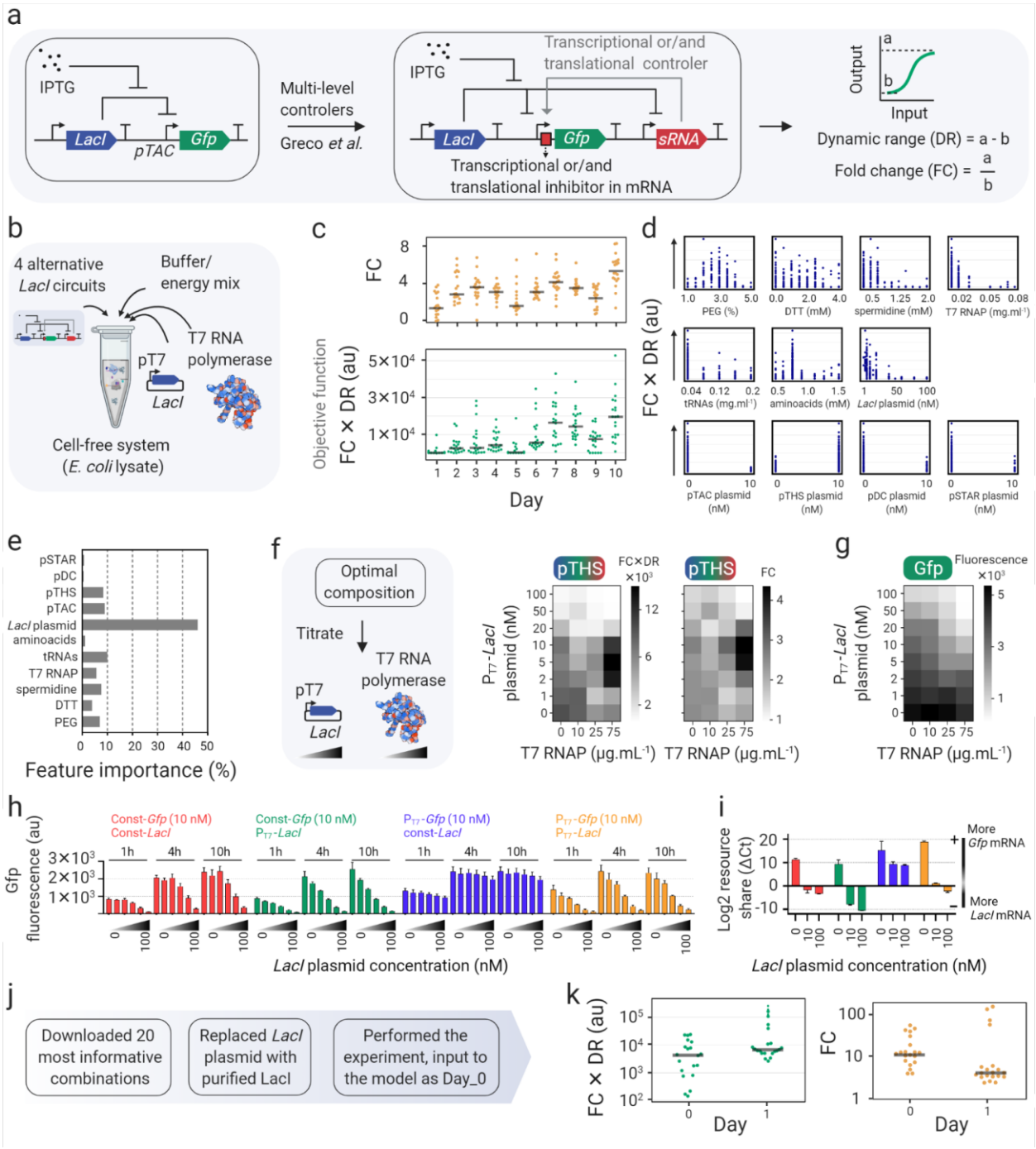


Fig. 3: Application of METIS for optimization of a *LacI* gene circuit. a) Single and multi-level controller *LacI* gene circuits from Greco *et al.*[43] Characterization of these circuits through dynamic range (DR) and/or fold-change (FC) of the output (Gfp fluorescence) between 0 and 10 mM input (concentration of IPTG). **b)** Imported in the active learning notebook, the varied components of the reactions included 4 *lacI* circuits as alternatives, some factors of buffer and energy mix of *E. coli* cell-free system along with the lysate, as well as T7 RNA polymerase and a second plasmid expressing *lacI* under a T7 promoter. **c)** The average of triplicates as the result of 10 rounds of active learning as plots for the objective function (FC × DR) and fold change (FC) values. **d)** Plots showing the distribution of measured yield values within the ranges of each factor. **e)** Feature importance percentages showing the effect of each factor on decision-making by the model to predict objective function values. **f)** Titration of P_{T7}-*LacI* plasmid and T7 RNA polymerase with the optimal composition (from 10 rounds of active learning that achieved with pTHS *LacI* circuit, the toehold switch as the second level controller of gene expression through translation). The heatmaps show

FC × DR (left) and FC (right) values (average of triplicates) of the titration. **g)** The same titration experiment as in **(f)** but instead of the pTHS circuit, *Gfp* was expressed under a constitutive promoter (independent from the P_{T7}-*LacI* plasmid and T7 RNA polymerase on the protein production). **h)** Titration (0, 1, 3, 10, 30, and 100 nM) of *LacI* plasmids with either constitutive or T7 promoter in combination with 10 nM of a *Gfp* plasmid (with T7 or constitutive promoter). **i)** The RT-qPCR results of the relative level of *LacI* and *Gfp* mRNAs for a similar experiment in **(h)** (0, 10, 100 nM *LacI* plasmids, and 10 nM *Gfp* plasmids) after 10 hours. Relative log2 resource share between *LacI and Gfp* mRNA in each sample is reported in order to account for RNA purification efficiency variability **j)** The 20 most informative combinations were downloaded after the 10-round active learning and the P_{T7}-*LacI* plasmid with purified LacI were replaced. After performing the experiments and measuring the objective function, we imported them as Day 0 and continued with experiments of the next round's predictions (Day 1). **k)** Plots of the objective function FC × DR (left) and FC (right) values (average of triplicates) of 20 most informative combinations with purified LacI followed by Day 1 experiments suggested by the workflow. See **Data availability** for combinations and objective function values.

Application of the workflow for optimization of a transcription & translation unit

To demonstrate that our workflow can also be used with categorical factors such as biological sequences, we tested METIS for the optimization of a transcription & translation unit. This unit is composed of six variants of a T7 promoter[45], six ribosome binding sites (RBS)[46], as well as 15 variations of N-terminal amino acids 3 to 5[47], and 20 variations of the last two C-terminal amino acids[48], which is in line with two recent studies that reported the importance of N- and C-terminal amino acids on mRNA translation[48,49].

To establish a convenient cell-free screening system, we sought to use linear DNA (i.e., a PCR product)[49] as template in combination with GamS, a small, 136 amino acid-long nuclease inhibitor from phage λ[50] that binds and protects linear DNA from degradation. First, we validated that addition of linear DNA with GamS resulted in gene expression levels comparable to that of plasmid DNA (**Fig. 4a**), which allowed the fast and efficient assembly of DNA templates through PCR primers without extensive cloning, transformation, and plasmid preparation steps (**Fig. 4b**).

We then optimized the transcription & translation unit that theoretically consists of 6 (P_{T7}) × 6 (RBS) × 15 (N-terminal) × 20 (C-terminal) = 10,800 potential conditions (i.e., combinations) through screening of only 200 combinations in 4 rounds of active learning (**Fig. 4b**). As the objective function, we defined the yield of the Gfp fluorescence readout of each transcription & translation unit normalized by a construct comprising wild-type T7 promoter, B0032 RBS and *sfGfp*. Yields were quantified after 6 hours of incubation of the different transcription & translation units at 30 °C in the *E. coli* cell-free system supplemented with purified GamS and T7 polymerase. Over 4 rounds of active learning, yield of the transcription & translation unit improved up to 12-fold on Day 3. Using a high exploration rate on Day 3 resulted in a wide distribution of yields, but no further improvement, indicating that an optimum had been reached (**Fig. 4c**). The distribution of alternative factors within the yield of 200 combinations and a representation of the feature importance are shown in **Supplementary Fig. 10**. Altogether, our

experiments demonstrated again, how METIS can be used to improve a described genetic unit by more than an order of magnitude with minimal experimental efforts.

After having rapidly explored the combinatorial space of the sequence controlling the transcription & translation unit in a cell-free setup, we additionally investigated the effect of the 20 most informative combinations *in vivo* (**Fig. 4d**). Surprisingly, however, the cell-free and *in vivo* yields for the 20 combinations showed a relatively low correlation of 0.41 (Day 0, **Fig. 4e, Supplementary Fig. 11**). This indicated that although cell-free systems offer rapid prototyping solutions, the optimal candidates are not necessarily directly transferable *in vivo*. To investigate whether we can further improve the performance of the transcription & translation unit *in vivo*, we used the data from Day 0 and continued with one more round of experiments guided by our workflow (Day 1, **Fig. 4f**). This resulted in an improvement by 130% for the highest yield *in vivo*.

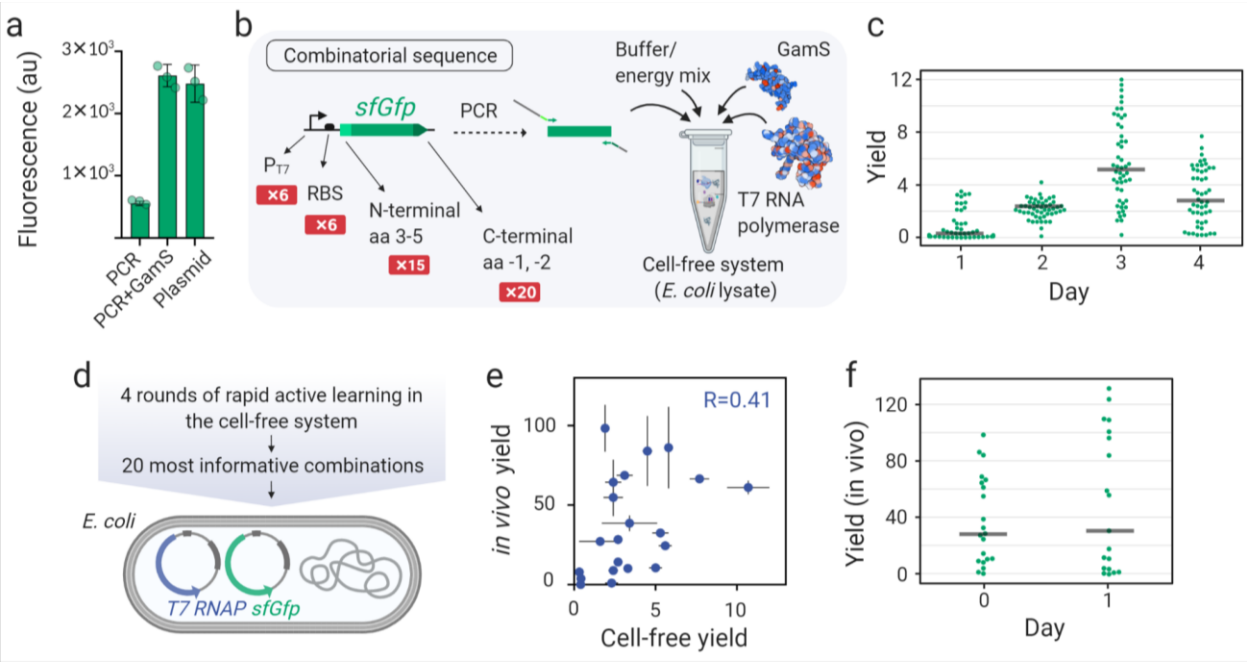


Fig. 4: Application of METIS for optimization of a transcription & translation unit. a) The cell-free expression of *sfGfp* (super-folder *Gfp*) using plasmid, linear DNA (PCR) and linear DNA plus GamS protein, a nuclease inhibitor that protects linear DNA from degradation. The bars and the error bars are the average and standard deviation of triplicates, respectively. **b)** Design of a transcription & translation unit controlled by variants of a T7 promoter, ribosome binding site (RBS), N-terminal amino acids 3, 4, and 5, and the last two C-terminal amino acids. The combinatorial transcription & translation units are expressed from linear DNA in the TXTL system consisting of the *E. coli* lysate, buffer and energy mix, as well as purified GamS and T7 RNA polymerase. **c)** The plot representing the average of triplicates as the result of 4 rounds of active learning, with 50 transcription & translation units tested per round. The yield is the Gfp fluorescence readout after 6 hours at 30 °C normalized by the same value from the reference constructs commonly used in the lab (**Methods**). **d)** A list of 20 most informative combinations of 4-day active learning performed in the cell-free system (c) was downloaded and the combinations were cloned in a vector and transformed into *E. coli* DH10β

harboring a plasmid expressing auto-regulated T7 RNA polymerase (**Methods**). **e)** Cell-free versus *in vivo* yields (average and standard deviation of triplicates) for the 20 most informative combinations. **f)** *In vivo* yield results (average of triplicates) of Day 0 (20 most informative combinations) and Day 1 (suggested by the workflow). See **Data availability** for combinations and yield values.

Application of the workflow for optimization of an *in vitro* CO_2-fixation pathway (CETCH cycle)

Finally, we aimed at assessing the performance of METIS for the optimization of complex metabolic networks. The collection of thousands of different enzymes and recent progress in enzyme engineering has opened the way for the design and construction of synthetic metabolic networks with new-to-nature properties[35,51,52]. One recent example is the CETCH cycle (**Fig. 5a**), a synthetic *in vitro* metabolic network consisting of 17 different enzymes that was built around a highly efficient CO_2-fixing enzyme, Crotonyl-CoA carboxylase/reductase (Ccr), converting CO_2 into the C2-compound glyoxylate[39] and/or glycolate[53]. Notably, the CETCH cycle is more efficient than natural occurring CO_2-fixing pathways like the Calvin-Benson-Bassham (CBB) cycle[39]. However, since the enzymes used for its construction derive from different organisms and thus metabolic backgrounds, several rounds of rational optimization were needed to harmonize the enzyme reactions and cofactors used in the cycle; and even though the kinetic parameters of the individual enzymes are known, their interactions in such a complex setup are non-linear, hardly predictable and basically impossible to disentangle with pure rational approaches. Hence, we sought to use our active learning workflow to improve the CETCH cycle's productivity further.

The setup of the CETCH cycle consists of 26 components encompassing 13 core enzymes, as well as four accessory enzymes, and nine other components such as magnesium chloride, CoA, NADPH, ATP and the starting substrate propionyl-CoA (see all components in **Fig. 5** and their concentration range in the **Code availability**). To minimize handling errors and automate the experimental setup of individual CETCH assays, we used an ECHO® 525 acoustic liquid handler with a minimal pipetting volume of 25 nL. Miniaturizing the assay to 10 µl of total volume allowed us to work with 384-well plates and assay 125 different conditions in triplicates per active learning round (**Fig. 5b**). To determine the CETCH cycle's productivity (i.e. formation of glycolate from CO_2), we developed an LC-MS (liquid chromatography-mass spectrometry) method using $^{13}C_2$-glycolic acid as an internal standard (See **Methods**).

For the first five rounds of optimization, we used product yield (glycolate) as objective function (for a description of the used parameters see **Supplementary Note 4**). After four iterative rounds, we reached a final concentration of 2.87 ± 0.09 mM glycolate in the best performing condition starting from 100 µM propionyl-CoA (**Fig 5c**). This yield translates into 57.4 fixed CO_2-equivalents per acceptor (propionyl-CoA)

and is >10 times more productive compared to the originally reported, rationally optimized version 5.4 of the CETCH cycle[38].

As we had not restricted the component resources during optimization, most of the superior conditions used more enzymes (compared to CETCH 5.4) to increase glycolate production (**Supplementary Fig. 12**). Next, we aimed at increasing specific productivity of the CETCH cycle. To that end, we took the data from the initial five rounds of unrestricted optimization and divided the glycolate yield values by the total concentration of enzymes used for each combination. This data was fed back to METIS and three additional rounds of active learning were performed with the new objective function, called "efficiency" (**Fig. 5d**). Optimization of efficiency identified one condition in round seven that is about six times more efficient than CETCH 5.4 and 14% more efficient than the best condition from the unrestricted optimization achieved in round four (**Fig. 5e**, see also **Supplementary Fig. 12, 13**).

To learn more about the possible bottlenecks of the CETCH cycle, we used the feature importance module of the METIS workflow along with plots visualizing the yield distribution over the range of each factor (**Supplementary Fig. 14, 15**). One of the most important contributors for both optimization efforts is the enzyme Methylsuccinyl-CoA dehydrogenase (Mco) (**Fig. 5f, g**). The enzyme's low activity of 0.1 U/mg and its unstable substrate methylsuccinyl-CoA, which is prone to spontaneous hydrolysis, likely require large amounts of Mco to preserve flux through the cycle[54]. During efficiency optimization, the two most important components were 4-hydroxybutyryl-CoA synthetase (Hbs) and coenzyme B_{12} (**Fig. 5g**). Analysis of the top 10% best performing conditions (**Supplementary Fig. 12, 13**) revealed that the concentrations of Hbs and B_{12} were significantly lower compared to the control (CETCH 5.4). To verify that high concentrations of Hbs have a negative impact on the cycle, we tested our control assay with ten times less and with five times more of the enzyme. Indeed, increasing Hbs concentration in the original assay decreased yield by 40%, while decreasing Hbs by one order of magnitude did not lower glycolate yield (**Supplementary Fig. 16**). Regarding the negative impact of higher concentrations of B_{12}, we reasoned that cobalt released from damaged cofactor could inhibit enzymes. Similar to high concentrations of Hbs, addition of cobalt to the original assay led to a decrease in glycolate yield (**Supplementary Fig. 16**).

To understand the dynamic behavior of the different CETCH cycle variants, we manually repeated the top three conditions (highest glycolate yields), a control (see **Supplementary Note 4**) and three underperforming conditions, taking time point samples for eight hours. The yield from this manual approach reflected the yield from the previous automated, miniaturized experiments, validating the

results of our optimization efforts (**Supplementary Table 3**). Interestingly, the final glycolate yield after eight hours (**Fig. 5h**) and the initial glycolate formation rates of these conditions over the first 15 minutes (**Fig. 5i**) were highly correlated (**Fig. 5j**), indicating that total flux and not improved enzyme/cofactor stability (or life-time) was responsible for the observed increased productivity of the system. This trend was further confirmed by a detailed analysis of 9 CoA-ester intermediates at different time points (**Fig. 5k, l**). Quantification of the CoA-ester intermediates did not show accumulation of single metabolites in the underperforming conditions or the control, indicative of specific bottlenecks (**Fig. 5l, Supplementary Fig. 16**). Instead, the underperforming conditions showed overall a faster depletion of intermediates, in line with the hypothesis that high flux through the cycle is important to prevent the loss of intermediates towards side reactions or hydrolysis.

In summary, our optimization efforts of the CETCH cycle resulted in variants that showed more than ten-fold productivity and almost six-fold improved efficiency, representing the most efficient *in vitro* CO_2-fixing system described to date.

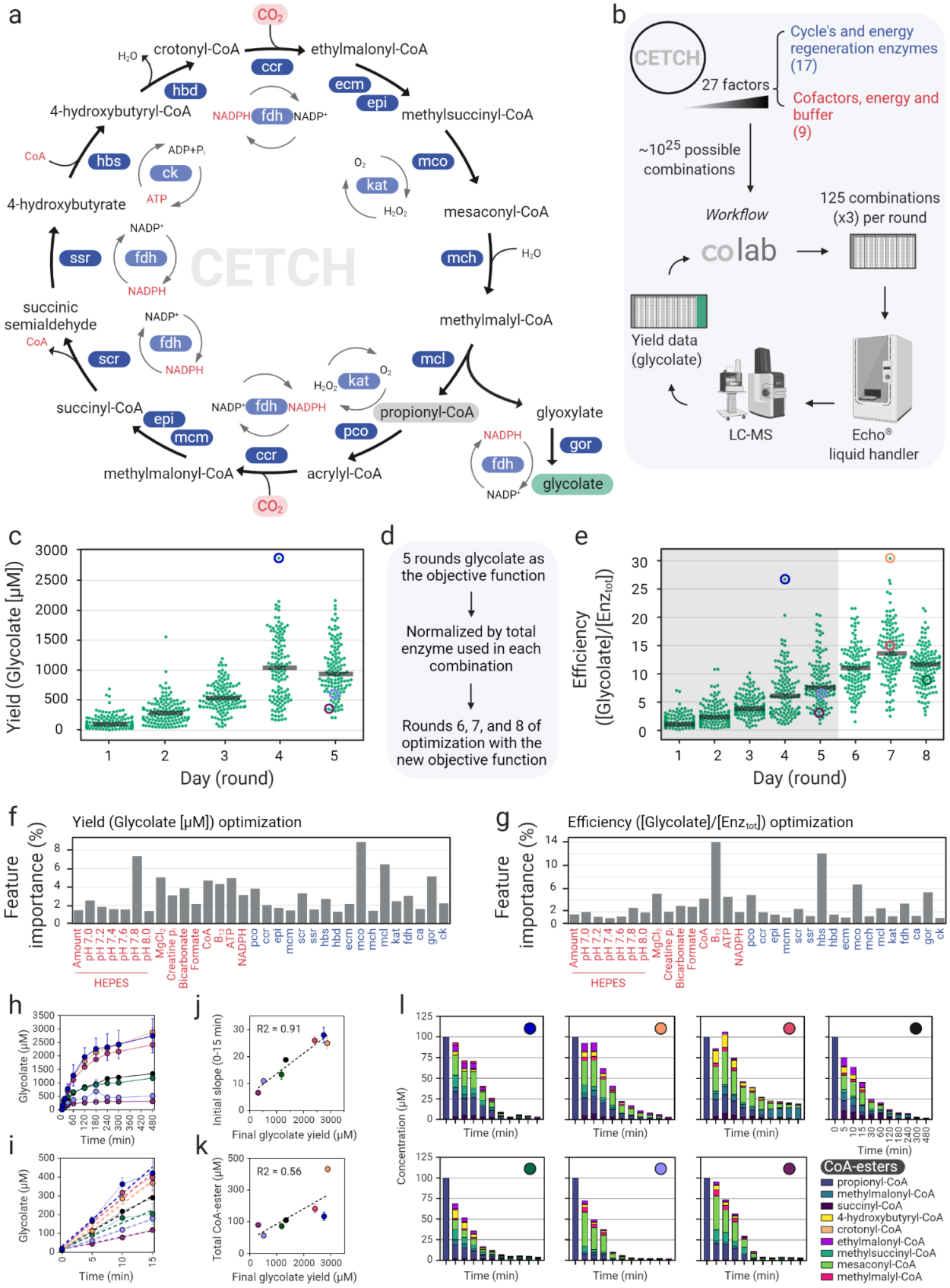

a
CETCH
crotonyl-CoA
ethylmalonyl-CoA
4-hydroxybutyryl-CoA
methylsuccinyl-CoA
4-hydroxybutyrate
mesaconyl-CoA
succinic semialdehyde
methylmalyl-CoA
succinyl-CoA
propionyl-CoA
glyoxylate
methylmalonyl-CoA
acrylyl-CoA
glycolate
b
CETCH
27 factors
Cycle's and energy regeneration enzymes (17)
Cofactors, energy and buffer (9)
~10^25 possible combinations
Workflow
colab
125 combinations (x3) per round
Yield data (glycolate)
LC-MS
Echo® liquid handler
c
Yield (Glycolate [µM])
Day (round)
d
5 rounds glycolate as the objective function
Normalized by total enzyme used in each combination
Rounds 6, 7, and 8 of optimization with the new objective function
e
Efficiency ([Glycolate]/[Enz_tot])
Day (round)
f
Yield (Glycolate [µM]) optimization
Feature importance (%)
HEPES
g
Efficiency ([Glycolate]/[Enz_tot]) optimization
Feature importance (%)
HEPES
h
Glycolate (µM)
Time (min)
i
Glycolate (µM)
Time (min)
j
R2 = 0.91
Initial slope (0-15 min)
Final glycolate yield (µM)
k
R2 = 0.56
Total CoA-ester (µM)
Final glycolate yield (µM)
l
Concentration (µM)
Time (min)
CoA-esters
propionyl-CoA
methylmalonyl-CoA
succinyl-CoA
4-hydroxybutyryl-CoA
crotonyl-CoA
ethylmalonyl-CoA
methylsuccinyl-CoA
mesaconyl-CoA
methylmalyl-CoA

Fig. 5: Application of METIS for optimization of an *in vitro* CO_2-fixation pathway (CETCH cycle). a) Reaction sequence of the CETCH cycle (pco: propionyl-CoA oxidase, ccr: crotonyl-CoA carboxylase/reductase, epi: ethylmalonyl-CoA/methylmalonyl-CoA epimerase, mcm: methylmalonyl-CoA mutase, scr: succinyl-CoA reductase, ssr: succinic semialdehyde reductase, hbs: 4-hydroxybutyryl-CoA synthetase, hbd: 4-hydroxybutyryl-CoA dehydratase, ecm: ethylmalonyl-CoA mutase, mco: methylsuccinyl-CoA oxidase, mch: mesaconyl-CoA hydratase, mcl: β-methylmalyl-CoA lyase, gor: glyoxylate reductase, kat: catalase, fdh: formate dehydrogenase, ck: creatine phosphokinase). For source of enzymes and kinetic parameters see Schwander *et al*.[38]. **b)** Workflow for the iterative optimization of the CETCH cycle (for details on the Google Colab tool see **Fig. 2** and **Supplementary Note 2**). 125 conditions were tested in each round. The generated worklist was fed to an ECHO® liquid handler which pipetted the assays. The reactions were started with 100 µM propionyl-CoA and stopped after 3h. The glycolate content was measured by LC-MS and used to train the model to predict more efficient combinations. **c)** Optimization of the CETCH cycle with glycolate yield as the objective function. Each dot represents the mean of one CETCH assay as a triplicate. **d)** Summary of the optimization and the switch of the objective function. **e)** Transformed data of **(c)** (glycolate yield divided by the total amount of enzymes = efficiency) for rounds 1-5, shaded region, and the data of three additional rounds of optimization with efficiency as the objective function (rounds 6-8). **f, g)** Calculated feature importance of the varied factors after five rounds of yield optimization **(c)** and three additional rounds of efficiency optimization **(e)**. **h)** Glycolate production over 8 h of the three best performing conditions (highest glycolate yields) (blue, orange and red), a control (black) and three randomly picked conditions which performed worse (green, lavender, burgundy). These assays were manually pipetted. These conditions are shown in **(c)** and/or **(e)**. The plotted values are the means of three replicates and the error bars represent the standard deviations. **i)** Magnified sector of the first 15 min from **(h)** with their cognate trendlines to calculate the slope. **j)** Plotted values of the initial production rate versus the final glycolate yield. **k)** Plotted values of the total amount of measured CoA esters within the eight hours versus the final glycolate yield. **l)** Quantified CoA esters of the seven assays over 8 h. The color code in the top-right corner of each plot resembles **(h, j, i, and k)** and the timepoints are shown in the plot of the control (black dot). The amount of propionyl-CoA within the zero samples is the added amount (100 µM) to start the reaction and was not measured by LC-MS. Each compound is plotted with error bars in (**Supplementary Fig. 17**).

2.4. Discussion

In this work, we describe METIS, a versatile, modular active learning workflow for the optimization of various biological objective functions, such as genetic and metabolic networks. This study democratizes machine learning applications for experimentalists without any programming skills or sophisticated lab equipment. We provide Google Colab notebooks (see **Code availability**) that can be adapted to different optimization applications and even used for data-driven predictions (for use of the latter see **Supplementary Table 1, Supplementary Note 5, Supplementary Figures 18)**.

For tailoring the workflow, the number of rounds and experiments per round need to be defined, which should take into account the number of different factors and their conditions, complexity of the objective function, as well as experimental throughput. For applications with a larger combinatorial space, more combinations need to be tested (**Fig. 5**). However, if the number of experiments is limited by cost, effort, or lab equipment, performing active learning in more rounds can be used to compensate for a lower number of total combinations tested. To explore a system beyond a local optimum, it is advised to adapt the exploration to exploitation ratio for each round individually (fully discussed in **Supplementary Note 2**). Users should apply their knowledge on the system and implicitly check whether the value of a given factor is fixed too early, probably indicating a low exploration to exploitation ratio. On the other hand, a high exploration to exploitation ratio might push the model towards random combinations, asking for a

proper balance to enable explorative as well as exploitation sampling. In our empirical experience, the exploration to exploitation ratio should gradually decrease towards the late rounds of active learning to enable more explorative combinations in early rounds and more exploitation in late rounds for efficient optimization (**Supplementary Note 2**).

Workflows can be started either from scratch (random combination as initialization) or using existing datasets (then performing active learning). Although our workflow is designed as an active learning approach (over multiple rounds of experiments), it can also be used as a classical machine learning with only one round of experiments. Factors of a given objective function can be numerical and/or categorical. Active learning parameters can be further customized using a detailed explanation in **Supplementary Note 2**.

METIS provides a variety of choices for visualization and analysis of results. Most importantly, our workflow can quantify importance of individual features and provide a number of most informative combinations, which has both proven particularly useful during *LacI* gene circuit optimization (**Fig. 3**). Using these features of the workflow allowed us to not only to improve the fold-change of the circuit, but also spot and, using additional experiments, verify a major bottleneck in the further optimization of the system (i.e., the *LacI* expression plasmid). After replacing the *LacI* expression plasmid with purified LacI protein, we were able to improve the circuit by more than two orders of magnitude compared to the original system. Notably, we did not have to re-perform active learning when switching to purified LacI instead of the *LacI* plasmid. The 20 most informative combinations generated through our workflow offered a short and quick path toward optimization.

Applying METIS onto different biological systems, we demonstrate that our workflow is able to optimize several complex genetic and metabolic networks of medium to large combinatorial space with minimal experimental efforts. As example, we improved the CETCH cycle a system of 27 variable factors including enzymes, cofactors, and buffer composition, spanning a theoretical combinatorial space of ~10^{25} different conditions. Performing only 1,000 (triplicate) assays over 8 rounds of active learning yielded a system with ten-fold improved productivity and six-fold increased efficiency, representing the most efficient *in vitro* CO_2-fixation system described to date.

The development and application of complex genetic and metabolic networks in synthetic biology is strongly increasing which requires new tools for their data-driven analysis. Efficient explorative approaches are needed not only for the optimization of existing biological networks, but also for the

design and realization of new-to-nature genetic and metabolic networks for which sampling the entire combinatorial space becomes practically impossible. Apart from network optimization with minimal experimental datasets, METIS can simultaneously help to discover so far unknown interactions and bottlenecks in these networks, which paves the way for their hypothesis-driven improvement. In the *LacI* circuit optimization, we showed how a bottleneck (i.e., ressource competition) can be identified, targeted, and finally overcome, which allowed us to improve the system by 34-fold. Similarly, during optimization of the CETCH cycle, we identified Mco, Hbs and B_{12} as limiting factors.

Numerous applications of the METIS workflow can be envisioned in the future, including the optimization of growth media and/or biochemical assays, genetic circuits, from simple transcription & translation units to more complex designs, or the guided engineering of proteins, enzymes, and metabolic pathways *in vivo* and *in vitro*. With its convenience and easy access, METIS opens the door for the study, prototyping, (combinatorial) engineering, and optimization of these systems in an efficient, standardized, and systematic manner.

2.5. Methods

<u>Gold regressor and analyzing different machine learning algorithms</u>

To find out which machine learning algorithm and sample size are suitable for our workflow, we conducted the following simulation:

- 1017 data points (compositions and yields) were collected from a recent study[31].
- An XGBRegressor model (gold regressor) was trained on 80% of the dataset and 20% of the dataset was used for validation and to avoid overfitting via early stopping.
- 100 combinations produced randomly within the range of each factor for Day_1.
- Instead of doing experiments in the laboratory to determine the yield of each combination, yield values were assigned by the gold regressor.

  Note that, in this phase the test model predicts the yields and ranks them to suggest for the experiments of the next day, and the gold regressor (trained on pre-data) is used to assign yield values by prediction instead of performing the experiments in the laboratory.

- For each machine learning model (MLP, DNN, linear regressors, XGBoost) an ensemble of 20 models with different hyperparameters was produced.

Note that the linear regressors is a deterministic approach so we just duplicated a model 20 times for which all predictions are the same.

- Each ensemble was trained on Day_1 data.
- 100000 random combinations were generated, and their corresponding yield was predicted by the ensemble of models and ranked by UCB score (see method section for the core algorithm of active learning), top 100 combinations were suggested for the next day. Yields were assigned by the gold regressor.
- The last two steps were repeated for other days, and on each day the model was trained on all the previous days' data.

Note that, in **Fig. 1b** for different sample sizes with XGBoost, 5, 10, 25, or 100 combinations were suggested for the next day.

General description of METIS notebook

All scripts used in this study were written in Python 3. Our modular tool, METIS, runs on Google Colab working through web browsers with a link without users needing to install Python or any packages.

Packages used in the development of METIS:

- Data processing: pandas (1.1.4) and numpy (1.18.5)
- Data visualization: matplotlib (3.2.2) and seaborn (0.11.0)
- Machine learning and deep learning: scikit-learn (0.22.2.post1), xgboost (0.90), and Keras (2.3.1) using TensorFlow backend.

The core algorithm of active learning

After measuring the value of the objective function (yield) for random combinations of Day_1, we continued with the following algorithm:

- RandomSearchCV is used to find the optimal 20 hyperparameters for the XGBoost model.
- The ensemble of 20 models is trained with the hyperparameters on data from all previous days (Day_1 to present day).
- 100000 combinations out of possible combinations are randomly selected.
- The mean and standard deviation of ensemble predictions are calculated.
- The combinations are sorted based on Upper Confidence Bound (UCB) score[31]: exploitation * (average of predictions) + exploration * (standard deviation of predictions).

Note that, for both Gfp production and *lacI* circuits study, exploitation was equal to 1 and exploration changed for each day (day 2, 3 = 1.41, day 4, 5, 6, 7 = 1, day 8, 9, 10 = 0.5)

- To perform experiments of the next day, the combinations with the highest UCB values are suggested.

The high standard deviation represents the uncertainty and improves the prediction power of models, whereas a high average value weighs favorable combinations leading to higher yields. Hence a coupled score taking into account these two factors ranks the most promising combinations[31].

Note that the active learning for optimization of objective functions is sometimes called Bayesian optimization[55].

Finding K most informative combinations

The K most informative combinations are calculated using the following algorithm:

- RandomSearchCV is used to find the optimal 20 hyperparameters for the XGBoost model.
- 2000 subsets of length K are selected from the tested combinations. The total number of possible subsets is:

$$\binom{N}{K} = \frac{N!}{K! \times (N-K)!}$$

- Then a new XGBoost with the optimal hyperparameter is trained on each subset. The model performance is then validated on unseen combinations using the Spearman correlation coefficient.
- All subsets are sorted based on their Spearman correlation coefficient, the top 5 are then chosen. Each of these 5 could be used.

  Note that increasing the number of subsets leads to a longer training time.

Finding feature importance

Feature importance values have been calculated with the following algorithm:

- RandomSearchCV is used to find the optimal hyperparameter for the XGBoost model.

- The model is trained using the selected hyperparameter. Using the built-in "feature_importances_" property of the XGBoost package, the ratio of feature importance is calculated throughout the training process for each day cumulatively.

Finding nonlinear (mutual) interactions

In complex systems, factors usually interact with each other and epistatically affect the output. These interactions can be among many factors, however, the most relevant is the mutual or double interaction between factors[56]. This analysis can be a hint to discover biological phenomena's behavior. The mutual interactions were calculated through the following algorithm[57]:

- A linear regression model is fitted on the dataset and its performance is evaluated based on the R squared of predicted and actual values. This performance is considered as the baseline.
- Iteratively, a new feature is added to the temporary dataset that equals $F_i \times F_j$ for i and j in the list of factors.
- The linear regression is fitted on the temporary dataset (which now has one more feature, $F_i \times F_j$) its performance is measured similarly to the baseline.
- The difference between each performance and the baseline, j, is calculated and visualized.

METIS prediction

In contrast to METIS optimization that tries to find the most promising combinations through maximizing the objective function, METIS prediction aims to maximize the model performance on the prediction of the objective function for unseen combinations. We modified the core active learning algorithm:

- Instead of UCB (exploitation $\times$ mean + exploration $\times$ std), combinations are sorted based on only their std value and set exploitation to zero. This enables picking the most uncertain combination for the next round.
- At the end of each round, it returns a trained model instead of promising combinations, and the R squared of prediction is improved over rounds.

Performance analysis using cross-validation

To evaluate the model performance of the enzyme engineering notebook, we used k-fold cross-validation. In each round, all the tested combinations are divided into k subsets (k=5 for **Supplementary Fig. 18, 19**), then in five steps we trained the model on 4 and evaluated its performance (R^2 Pearson) on the other

subset. This process was repeated for all 5 subsets. In the end, the average performance on all subsets was reported as the model's performance. We used sklearn built-in function for cross-validation.

Table-to-speech virtual assistant

This tool helps molecular biologists to boost their manual liquid handling through reading volume and destination well in ascending order, therefore minimizes the need for changing the pipetting volume. We used the Google Text2Speech python package to transform the text into a voice file. There are two ways to interact with this notebook to continue with the next pipetting volume. The first is to do it manually with your keyboard (what we did), the second is using the voice assistant. For transforming voice to text (specific commands like 'next', 'repeat', etc.). We used the SpeechRecognition (3.8.1) python package. The code is available on https://github.com/amirpandi/Liquid-Handling-Assistant.

Plasmid and DNA preparation

The constitutive *Gfp* under the control of J23101 promoter and B0032 RBS was built in a recent study (pBEAST-J23101-B0032-*sfGfp*)[58]. Using golden gate cloning (BsaI-HF®v2 NEB #R3733L, T4 DNA ligase NEB #M0202T), in this plasmid, the super folder *Gfp* gene was replaced by *LacI* for constitutive-*LacI*, then the promoter was replaced by a T7 promoter (gaatttaatacgactcactatagggaga) to construct P_{T7}-*LacI* plasmid. Since we used T7 promoters, a T7 terminator from Temme *et al*.[59] (tactcgaacccctagcccgctcttatcgggcggctaggggtttttgt) was cloned downstream. The version of *LacI* gene is similar to those in *LacI* circuits built by Greco *et al*.[38] Plasmids for the cell-free gene expression were purified using the Machery-Nagel NucleoBond Xtra Maxi kit. For protein purification using His tag, *sfGfp* and *LacI* genes were cloned with an N-terminal His tag under IPTG-inducible T7 promoter.

For cell-free experiments for optimization of the transcription & translation unit (**Fig. 4b**), PCRs were performed using Q5® High-Fidelity 2X Master Mix (NEB #M0492L), *sfGfp* as the template, and primers with overhangs harboring P_{T7}, RBS, and N-terminal sequence (forward primer) and C-terminal (reverse primer) at the final volume of 50 µL. After verification of PCRs using agarose gel, Monarch PCR & DNA Cleanup Kit (NEB #T1030L) was used to purify the fragments and they were all adjusted to the concentration of 100 nM to use for active learning experiments.

For *in vivo* experiments of the transcription & translation unit (**Fig. 4d**) PCRs were done similar to the cell-free experiment. Restriction sites for BsaI enzyme were designed on either side of PCR fragments enabling for goldengate assembly into a pSEVA224 vector (a low copy plasmid with kanamycin marker) from the

SEVA collection[60,61]. Since we used T7 promoters, a T7 terminator from Temme *et al*.[59] (tactcgaacccctagcccgctcttatcgggcggctaggggtttttgt) was cloned downstream.

Protein purification

For all enzymes involved in the CETCH cycle, expression and purification were performed as previously described[62]. Other proteins, T7 RNA polymerase (addgene #124138), GamS (addgene #45833), sfGfp, and LacI were His-tag purified using Protino® gravity columns (Machery-Nagel #745250) and Protino® Ni-NTA Agarose (Machery-Nagel #745400). 1 L cultures in LB media supplement with appropriate antibiotic were subcultured (1:100) from overnight precultures. Cultures were grown at 37 °C for two hours, then induced by 0.1 mM IPTG, incubated for 3 more hours at 37 °C to produce proteins. Cells were harvested at 8000g for 10 min, pellets were resuspended with 5 mL NPI-10 buffer, and sonicated. Samples were centrifuged at 18000g for 1 hour at 4 °C. The equilibration, wash, and elution steps were done according to the manufacturer's protocol. Next, imidazole desalting was performed using PD-10 desalting columns (GE Healthcare #17085101) according to the manufacturer's protocol. The purification was verified using the SDS page and the protein concentrations were determined using the Bradford assay. Glycerol was added to the protein samples to a final percentage of 10%, then they were aliquoted and after flash-freezing in liquid nitrogen, stored at -80 °C.

Lysate Preparation

E. coli lysate was prepared using an autolysis strategy[63]. Freeze-thawing *E. coli* BL21-Gold (DE3) cells with a pAS-LyseR plasmid produce a high-quality extract. Overnight precultures in LB-ampicillin media at 37 °C were subcultured in 5x 2 L 2xYTPG medium supplemented with ampicillin and grown at 37 °C to the OD=1.5. Cells were harvested (2000g, 15 min, room temperature) in 10 centrifuge bottles and 90 mL of cold S30A buffer (50 mM Tris-HCl at pH 7.7, 60 mM K-glutamate, 14 mM Mg-glutamate, to the final pH of 7.7) was added to each. After vigorous vortexing, each was divided into two preweighed 50 mL falcons and centrifuged (2000g, 15 min, room temperature). The supernatants were removed carefully and after weighing falcons with pellets, the net weights were calculated. Two volumes of cold S30A with 2 mM DTT, were used to resuspend each pellet (2.8 mL for 1.4 g pellet), which were then vortex-mixed, and stored at -80 °C. The next day, frozen cells were thawed in a water bath at room temperature, vigorously vortex-mixed, and incubated at 37 °C shaking for 45 min. The vortexing and 45 min incubation steps were repeated. Finally, the samples were centrifuged (30000g, 60 min, 4 °C) to obtain the cell extract. The supernatants were gently pipetted out in 1.5 tubes, recentrifuged (20000g in a tabletop centrifuge, 5 min,

4 °C) to remove all the remaining cell debris aliquoted, and after freezing in liquid nitrogen stored at -80. For the composition of the cell-free reaction buffer and energy mix, all chemicals were used as by Sun *et al*.[33] except for amino acids (L-amino acids set, Sigma #LAA21-1KT).

Cell-free reactions

To perform the active learning experiments in **Fig. 1, 3**, Table2Seech_Volume.csv file of each round was downloaded from the notebook and uploaded to the table-to-speech virtual assistant notebook. Before starting the pipetting, we arranged all pipette tips with numbers written on one side of tip boxes (two boxes side by side) from 1 to 20 (for 20 data points). PCR tubes in which the compositions were going to be mixed also were numbered on racks from 1 to 20. The numbering increases the accuracy of the manual pipetting. Next, the table-to-speech assistant was run on a laptop on the bench and the space key was set in the Google Colab settings to run the code. After pipetting each factor into the corresponding destination, while the right hand was replacing the tip, the left hand pressed the space key to hear the next pipetting step in a headphone as well as to see the action appearing on the screen. The table-to-speech assistant goes line by line for each factor and ranks the pipetting values from minimum to maximum, hence, minimizes changes in the pipette volume. For fixed elements such as HEPES and lysate, a master mix was made and after finishing pipetting all combinations, the master mix was added to each. All the steps were performed on ice. At the end, samples were gently mixed (not to generate bubbles) using a multichannel pipette and 10 µL of each was transferred into a 384-well plate (Greiner Bio-One #784076). Note that the volume of mixtures should be at least 20% in excess in PCR tubes not to face difficulties in the final pipetting step into the 384-well plate. The Gfp fluorescence was monitored (excitation: 485, emission: 528 nM, gain: 80) every 10 min in a plate reader (Tecan Infinite 200 PRO).

The yield (objective function) in **Fig. 1e**, as provided in the **Data availability**, is the Gfp fluorescence (after 6 hours incubation at 30 °C) of each composition normalized by a composition in which the concentration of all variable factors is at mid-range. However, the plotted yields are those values divided by 0.33, the average ratio of Gfp fluorescence between the active learning reference and a commonly used composition[33]. The objective function of the *LacI* circuit active learning in **Fig. 3c** is fold-change (FC) $\times$ dynamic range (DR) of the output (Gfp fluorescence) between 0 and 10 mM input (concentration of IPTG). For cell-free reactions in **Fig. 4c**, the final volume of 5 µL was prepared directly in a 384-well plate, 10 nM final concentration of each linear DNA was transferred and the mix of other components of the cell-free lysate plus T7 polymerase (40 $\mu g.mL^{-1}$) and GamS (2 µM) was added while gently mixing. The yield (objective function) in **Fig. 4c** is the Gfp fluorescence readout (after 6 hours of incubation at 30 °C) of each

transcription & translation unit normalized by the Gfp fluorescence of a commonly used sequence in our lab, wild-type T7 promoter, B0032 RBS, and *sfGfp* sequence. For all cell-free reactions, the Gfp fluorescence readout of the extract with no DNA was subtracted before yield calculations. All compositions and concentrations used in cell-free reactions in **Fig. 1, 3, 4** are accessible next to the corresponding active learning notebook at https://github.com/amirpandi/METIS.

RT-qPCR experiment

Total RNA was extracted from cell-free expression reactions with a kit (NEB #T2010), following the manufacturer's instructions. Initial qPCR analysis indicated that a substantial amount of plasmid DNA remained in control reactions, which did not include reverse transcriptase to synthesize cDNA. Therefore, samples were subsequently treated to an additional DNase treatment by TURBO DNA-free™ Kit (Invitrogen™ #AM1907) according to the manufacturer's instructions. The resulting RNA produced a substantial qPCR signal (iTaq Universal SYBR Green Supermix Bio-Rad #1725120) when converted to cDNA by ProtoScript® II Reverse Transcriptase (NEB #M0368) using the standard protocol and random hexamer primers (ThermoFisher #SO142), but not in control reactions lacking reverse transcriptase. In order to account for potential sample-to-sample variability in extraction efficiency, all data presented herein is represented as a relative difference in cycle threshold (Ct) between *Gfp* and *LacI* cDNA within each sample. Standard curves with known concentrations of plasmid DNA were analyzed in parallel for *Gfp* and *LacI* primer sets, indicating comparable qPCR efficiencies and template specificity. No further normalization was required.

Western blot

Cell-free expression reactions and LacI-6xHis purified protein dilutions were mixed with 4 µL of non-reducing sample loading buffer (Thermo Scientific #39001) and incubated at 90 °C for 5 minutes. The samples were then loaded into pre-cast SDS-PAGE gels (Bio-Rad #4561095) and separated by electrophoresis. The gel was then immediately placed into a Bio-Rad TransBlot® Turbo apparatus for protein transfer onto a nitrocellulose membrane (Bio-Rad #1704158). Since all samples were produced from the same batch of cell-free expression reaction mix or were of known concentration, total protein concentration was not assessed. Western blot analysis was performed using a monoclonal antibody against LacI (Sigma-Aldrich #05-503-I) and an anti-mouse HRP-conjugated secondary antibody (Invitrogen #31430). After dispensing the detection reagent as indicated by the manufacturer (Neogen #324175), the blot was immediately imaged on a Bio-Rad ChemiDoc. A single clear band corresponding to the molecular

weight of LacI was detected in lanes containing purified LacI or expression from a *LacI*-containing plasmid (See inset, **Supplementary Fig. 9**). Band intensity was quantified using ImageJ and reported in parallel to the chemiluminescence image (**Supplementary Fig. 9**).

In vivo experiment of transcription & translation units

After cloning transcription & translation units into the pSEVA224 vector (plasmid and DNA preparation section), they were transformed into *E. coli* DH10β harboring an autoregulated T7 RNA polymerase circuit (addgene #71428)[64]. 3 colonies of each were cultured in LB with 30 µg.mL^{-1} ampicillin + 30 µg.mL^{-1} kanamycin in a 96 deep well plate. After 10 hours of cultivation at 37 °C, 10 µL of each was added to 190 µL LB with 30 µg.mL^{-1} ampicillin + 30 µg.mL^{-1} kanamycin in a 96-well plate (Thermo Scientific #137101). The Gfp fluorescence was monitored (excitation: 485, emission: 528 nM, gain: 80) every 30 min in a plate reader (Tecan Infinite 200 PRO) shaking at 37 °C. The *in vivo* yield in **Fig. 4e, f** is the Gfp fluorescence readout (after 6 hours) of each transcription & translation unit normalized by the Gfp fluorescence of a commonly used sequence in our lab, wild-type T7 promoter, B0032 RBS, and *sfGfp* sequence. The Gfp fluorescence readout of cells with no *sfGfp* gene was subtracted before yield calculations.

Workflow for CETCH assays in 384-well plates

The worklist generated by the METIS script was dissected into 5 worklists: dH_2O, Buffers and Cofactors, Enzymes, Carbonic Anhydrase, and Substrate. In cases where pipetting errors occurred, we used our Exceptions_to_Worklist script for correction of failed transfers (provided in **Code availability**). This script generates a new worklist out of the exception file generated by the ECHO® and provides a list with how much volume needs to be added into which well. Dissecting the worklists guarantees for example that all buffers are transferred before enzymes are added. Note that we used fresh enzyme stocks in each round to prevent loss of activity due to repetitive freeze-thaw cycles. As source plates we used ECHO® qualified 384-Well PP 2.0 Plus Microplates from Labcyte and used AQ_GP as the liquid class (AQueous solution; Glycerol/Protein). This liquid class was tested previously with the stocks of our assay components.

We also added a control condition with composition derived from the published assay of CETCH 5.4 (composition see Assays for determination of new enzyme stocks after round two). Controls can be added in the code as specials. The yield of this condition increased by a factor of 3 after round two (data not shown), where new enzyme batches of four enzymes were used. To identify the enzyme that was the reason for that, we tested the control assay with each of the four old enzymes separately (**Supplementary Fig. 16b**). Despite being important in the control (~280 µM in round one and two), catalase did not seem

important in each condition, since we reached yields of <1500 µM already in round two with the old stock (**Fig. 5c**).

After starting the assays with 100 µM propionyl-CoA we used an Axygen® Breathable Sealing Film (BF-400-S) to cover the 384-well PCR Plate (AB-1384) to allow the transfer of oxygen. The reaction (10 µL volume) was carried out at 30 °C and mild shaking at 160 rpm in an Infors HT Ecotron shaker. The reactions were stopped after 3 h with 1.25 µL of 500 mM polyphosphate and 1.25 µL of 50% formic acid. While the formic acid quenches the reaction, the polyphosphate was used for enhanced precipitation of the proteins. The plate was spun for 1 h at 2272g and 4 °C to pellet the proteins.

For analysis by LC-MS, we used a multichannel pipette to transfer 1 µL of the supernatant into 9 µL of precooled dH_2O in a new 384-Well Thermo-Fast® plate. Afterward, we added 10 µL of 10 µM $^{13}C_2$ labeled glycolic acid as an internal standard. The plate was sealed with a Corning™ Microplate Aluminum Sealing Tape (6570). The assay plate with the quenched reactions was sealed with a Corning™ Microplate Aluminum Sealing Tape too and stored at -80 °C.

Timepoint assays of 7 selected conditions

The assays were done in triplicates containing 150 µL volume each and were carried out in a 1.5 mL reaction tube (at 30 °C, 500 rpm). The reactions were started with 100 µM propionyl-CoA. 12 µL samples were taken and quenched in 1.5 µL 50% formic acid and 1.5 µL 500 mM sodium polyphosphate (emplura®) at 5, 10, 15, 30, 60, 120, 180, 240, 300 and 480 min. The samples were spun for 20 min at 4 °C and 20.000g, before the supernatant was transferred into Thermo Scientific™ Abgene 96 Well Polypropylene Storage Microplates (AB-1058) and sealed with Corning™ Microplate Aluminum Sealing Tape. While 2 µL were used to prepare a 1:10 dilution in water for the measurement via LC-MS, the remaining samples were stored at -80°C. The concentrations for the assays are shown in the table below (Buffers and cofactors in mM, enzymes in µM). See **Supplementary Table 3** for the details of these conditions.

LC-MS analysis of CoA esters

All CoA esters were measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LC/MS) equipped with a UHPLC (Agilent Technologies 1290 Infinity II) using a 150 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 2 µL of the diluted samples (1:10 in water). The flow was set to 0.400 mL/min and the separation was performed using 50 mM ammonium formate pH 8.1 (buffer A) and acetonitrile (buffer B). We quantified the CoAs using external

standard curves prepared in water with formic acid at pH 3. The standard curves were measured before and after the samples. Except for methylsuccinyl-CoA, all compounds were stable. For methysuccinyl-CoA we calculated the concentration as an average of the two standard curves at the time point the sample was measured. The parameters for the multiple reaction monitoring (MRMs) and the gradient are shown in the tables below. The data analysis was done with Agilent MassHunter Quantitative Analysis (for QQQ). See **Supplementary Table 4** (Gradient for the separation of CoA esters) and **Supplementary Table 5** (MRM transitions).

LC-MS analysis of glycolate

Glycolate was measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LC/MS) equipped with a UHPLC (Agilent Technologies 1290 Infinity II) using a 150 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 0.5 µL. The diluted samples (1:10 in water), as well as the external standard curve, were diluted 1:2 with 10 µM $^{13}C_2$-labeled glycolic acid as internal standard. The flow was set to 0.100 mL/min and the separation was performed using dH_2O with 0.1% formic acid (buffer A) and methanol with 0.1% formic acid (buffer B). The parameters for the multiple reaction monitoring (MRMs) and the gradient are displayed below. Data analysis was done using the Agilent Mass Hunter Workstation Software. See **Supplementary Table 6** (Gradient for the separation of CoA esters) and **Supplementary Table 7** (MRM transitions).

Data availability

All active learning data along with corresponding notebooks are provided at https://github.com/amirpandi/METIS.

Code availability

All notebooks run on Google Colab notebooks and are accessible on https://github.com/amirpandi/METIS. All scripts used in this study were written in Python 3. Packages used in the development of the workflow are pandas (1.1.4) and numpy (1.18.5), matplotlib (3.2.2) and seaborn (0.11.0), scikit-learn (0.22.2.post1), xgboost (0.90), and Keras (2.3.1) using TensorFlow backend.

Acknowledgments

We wish to thank V. Gureghian for stimulating discussions during conception of the work, T. E. Gorochowski (University of Bristol, UK) for providing SLC and MLC constructs, and M. Kushwaha (INRAE, Jouy en Josas, France) for *E. coli* DH10β, harboring the autoregulated T7 polymerase construct. We thank S. Burgener, S. Luo, A. Sánchez-Pascuala Jerez, N. Odermatt, and A. Schrodt for graphical, technical and experimental support, as well as fruitful discussion. This work was supported by a European Molecular Biology Organization (EMBO) long-term postdoctoral fellowship (A.P. ALTG 165-2020), the Gordon and Betty Moore Foundation, GBMF10652, grant DOI https://doi.org/10.37807/GBMF10652 (T.J.E.), the Max Planck Research Network in Synthetic Biology (MaxSynBio) of the Max Planck Society and the Federal Ministry of Education and Research (BMBF; T.J.E.), the UDOPIA PhD program and the French National Research Institute for Agriculture, Food, and Environment (INRAE) through the Métaprogramme BIOLPREDICT program (L.F.), BMBF Grant, MetAFor, No. 031B0850B (T.J.E.), ANR iCFree grant (ANR-20-BiopNSE) (J-L.F.) and the Max-Planck Society. Figures were created with Biorender.com.

2.6. References

1. Purnick, P.E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* **10**, 410-22 (2009).
2. Smanski, M.J. et al. Functional optimization of gene clusters by combinatorial design and assembly. *Nat Biotechnol* **32**, 1241-9 (2014).
3. Dolberg, T.B. et al. Computation-guided optimization of split protein systems. *Nat Chem Biol* **17**, 531-539 (2021).
4. Radivojevic, T., Costello, Z., Workman, K. & Garcia Martin, H. A machine learning Automated Recommendation Tool for synthetic biology. *Nat Commun* **11**, 4879 (2020).
5. Naseri, G. & Koffas, M.A.G. Application of combinatorial optimization strategies in synthetic biology. *Nat Commun* **11**, 2446 (2020).
6. Carbonell, P., Radivojevic, T. & García Martín, H. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synthetic Biology* **8**, 1474-1477 (2019).
7. Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C. & Collins, J.J. Next-Generation Machine Learning for Biological Networks. *Cell* **173**, 1581-1592 (2018).
8. Gilliot, P.A. & Gorochowski, T.E. Sequencing enabling design and learning in synthetic biology. *Curr Opin Chem Biol* **58**, 54-62 (2020).
9. Volk, M.J. et al. Biosystems Design by Machine Learning. *ACS Synth Biol* **9**, 1514-1533 (2020).
10. Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F.J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389-403 (2019).
11. Libbrecht, M.W. & Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-32 (2015).
12. Liu, J., Li, J., Wang, H. & Yan, J. Application of deep learning in genomics. *Sci China Life Sci* **63**, 1860-1878 (2020).
13. Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687-694 (2019).
14. Wittmann, B.J., Johnston, K.E., Wu, Z. & Arnold, F.H. Advances in machine learning for directed evolution. *Curr Opin Struct Biol* **69**, 11-18 (2021).
15. Gussow, A.B. et al. Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat Commun* **11**, 3784 (2020).
16. Kim, H.K. et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat Biotechnol* **36**, 239-241 (2018).
17. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat Commun* **10**, 4284 (2019).
18. Eitzinger, S. et al. Machine learning predicts new anti-CRISPR proteins. *Nucleic Acids Res* **48**, 4698-4708 (2020).
19. Hiscock, T.W. Adapting machine-learning algorithms to design gene circuits. *BMC Bioinformatics* **20**, 214 (2019).
20. Saltepe, B., Bozkurt, E.U., Gungen, M.A., Cicek, A.E. & Seker, U.O.S. Genetic circuits combined with machine learning provides fast responding living sensors. *Biosens Bioelectron* **178**, 113028 (2021).
21. Racovita, A. & Jaramillo, A. Reinforcement learning in synthetic gene circuits. *Biochem Soc Trans* **48**, 1637-1643 (2020).
22. Gazut, S., Martinez, J.M., Dreyfus, G. & Oussar, Y. Towards the optimal design of numerical experiments. *IEEE Trans Neural Netw* **19**, 874-82 (2008).
23. Yu, K., Bi, J. & Tresp, V. Active learning via transductive experimental design. in *Proceedings of the 23rd international conference on Machine learning* 1081–1088 (Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006).

24. Olsson, F. A literature survey of active machine learning in the context of natural language processing. (2009).
25. Sommer, C. & Gerlich, D.W. Machine learning in cell biology - teaching computers to recognize phenotypes. *J Cell Sci* **126**, 5529-39 (2013).
26. Jones, T.R. et al. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proc Natl Acad Sci U S A* **106**, 1826-31 (2009).
27. Pournara, I. & Wernisch, L. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics* **20 17**, 2934-42 (2004).
28. Naik, A.W., Kangas, J.D., Sullivan, D.P. & Murphy, R.F. Active machine learning-driven experimentation to determine compound effects on protein patterns. *Elife* **5**, e10047 (2016).
29. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov Today* **20**, 458-65 (2015).
30. Osmanbeyoglu, H.U., Wehner, J.A., Carbonell, J.G. & Ganapathiraju, M.K. Active machine learning for transmembrane helix prediction. *BMC Bioinformatics* **11 Suppl 1**, S58 (2010).
31. Borkowski, O. et al. Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun* **11**, 1872 (2020).
32. GoogleColaboratory. https://colab.research.google.com/.
33. Sun, Z.Z. et al. Protocols for implementing an Escherichia coli based TX-TL cell-free expression system for synthetic biology. *J Vis Exp*, e50762 (2013).
34. Pandi, A., Grigoras, I., Borkowski, O. & Faulon, J.L. Optimizing Cell-Free Biosensors to Monitor Enzymatic Production. *ACS Synth Biol* **8**, 1952-1957 (2019).
35. Karim, A.S. et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat Chem Biol* **16**, 912-919 (2020).
36. Pandi, A. et al. Metabolic perceptrons for neural computing in biological systems. *Nat Commun* **10**, 3880 (2019).
37. Swank, Z., Laohakunakorn, N. & Maerkl, S.J. Cell-free gene-regulatory network engineering with synthetic transcription factors. *Proc Natl Acad Sci U S A* **116**, 5892-5901 (2019).
38. Greco, F.V., Pandi, A., Erb, T.J., Grierson, C.S. & Gorochowski, T.E. Harnessing the central dogma for stringent multi-level control of gene expression. *Nat Commun* **12**, 1738 (2021).
39. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
40. Najafabadi, M.M. et al. Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2**, 1 (2015).
41. Brownlee, J. XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. (Machine Learning Mastery). *https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/* (2016).
42. Li, W., Yin, Y., Quan, X. & Zhang, H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet* **10**, 1077 (2019).
43. Yu, B. et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* **36**, 1074-1081 (2020).
44. Borkowski, O. et al. Cell-free prediction of protein expression costs for growing cells. *Nat Commun* **9**, 1457 (2018).
45. Galiñanes Reyes, S., Kuruma, Y. & Tsuda, S. Uncovering cell-free protein expression dynamics by a promoter library with diverse strengths. *bioRxiv*, 214593 (2017).
46. Ribosome Binding Sites/Prokaryotic/Constitutive/Community Collection. https://parts.igem.org/Ribosome_Binding_Sites/Prokaryotic/Constitutive/Community_Collection.

47. Verma, M. et al. A short translational ramp determines the efficiency of protein synthesis. *Nat Commun* **10**, 5774 (2019).
48. Weber, M. et al. Impact of C-terminal amino acid composition on protein expression in bacteria. *Mol Syst Biol* **16**, e9208 (2020).
49. Yim, S.S., Johns, N.I., Noireaux, V. & Wang, H.H. Protecting Linear DNA Templates in Cell-Free Expression Systems from Diverse Bacteria. *ACS Synth Biol* **9**, 2851-2855 (2020).
50. Murphy, K.C. Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of Escherichia coli RecBCD enzyme. *J Bacteriol* **173**, 5808-21 (1991).
51. Erb, T.J., Jones, P.R. & Bar-Even, A. Synthetic metabolism: metabolic engineering meets enzyme design. *Current opinion in chemical biology* **37**, 56-62 (2017).
52. Bowie, J.U. et al. Synthetic Biochemistry: The Bio-inspired Cell-Free Approach to Commodity Chemical Production. *Trends Biotechnol* **38**, 766-778 (2020).
53. Miller, T.E. et al. Light-powered CO2 fixation in a chloroplast mimic with natural and synthetic parts. *Science* **368**, 649-654 (2020).
54. Burgener, S., Schwander, T., Romero, E., Fraaije, M.W. & Erb, T.J. Molecular Basis for Converting (2S)-Methylsuccinyl-CoA Dehydrogenase into an Oxidase. *Molecules* **23**(2017).
55. Archetti, F. & Candelieri, A. *Bayesian Optimization and Data Science*, (2019).
56. Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Molecular systems biology* **5**, 297-297 (2009).
57. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, (Springer, 2013).
58. Voyvodic, P.L. et al. Plug-and-play metabolic transducers expand the chemical detection space of cell-free biosensors. *Nat Commun* **10**, 1697 (2019).
59. Temme, K., Hill, R., Segall-Shapiro, T.H., Moser, F. & Voigt, C.A. Modular control of multiple pathways using engineered orthogonal T7 polymerases. *Nucleic acids research* **40**, 8773-8781 (2012).
60. Martínez-García, E. et al. SEVA 3.0: an update of the Standard European Vector Architecture for enabling portability of genetic constructs among diverse bacterial hosts. *Nucleic Acids Research* **48**, D1164 - D1170 (2020).
61. D3veloperSCS_SEVA. http://seva-plasmids.com/.
62. Sundaram, S. et al. A modular in vitro platform for the production of terpenes and polyketides from CO2. *Angewandte Chemie International Edition* (2021).
63. Didovyk, A., Tonooka, T., Tsimring, L. & Hasty, J. Rapid and Scalable Preparation of Bacterial Lysates for Cell-Free Gene Expression. *ACS Synth Biol* **6**, 2198-2208 (2017).
64. Kushwaha, M. & Salis, H.M. A portable expression resource for engineering cross-species genetic circuits and pathways. *Nat Commun* **6**, 7832 (2015).

2.7. Supplementary Information

2.7.1. Supplementary Notes

Supplementary Note 1: Active learning process for optimization of cell-free Gfp production

We started the active learning cycle by generating 20 random compositions for the first round, pipetting them, measuring the Gfp fluorescence in a plate reader, then we collected the results after 6 hours of incubation at 30 °C. The mean and standard deviation of each composition was imported to the model, and after training, it suggested a new set of 20 compositions. The yield was calculated from the measured Gfp fluorescence of every composition normalized by a composition in which all variable factors are at their mid-range. This normalization was used for active learning (**Data availability**). In the end, as plotted in **Fig. 1e**, all values were normalized by composition from a common protocol in the field of cell-free synthetic biology to be comparable with other studies (**Methods**). The cycle (**Fig. 1d**) was repeated for 10 rounds. In each round, the model became more predictive in generating new compositions and ranking the best 20 suggestions for the next round. Since pipetting 13 elements in 20 samples is error-prone and needs a substantial amount of time and effort, we developed a table-to-speech virtual assistant that is run on Google Colab and can be easily used on a computer or smartphone. This tool takes as input the suggested table of volumes to pipette, goes through it line by line (factors), ranks them from minimum to maximum volume, and reads and graphically shows them on the screen. With organized pipette tip sets and destination tubes (**Methods**), this considerably improved the speed, accuracy, and comfort of pipetting of such complex compositions.

Supplementary Note 2: User guide and description of features of the modular workflow

Fig. 2b shows the workflow of using METIS from the adjustment of the parameters to data visualization and analysis. The user input consists of two sections for i) active learning parameters and ii) factors with a range and/or category. In the active learning parameters section, the user should define the number of combinations willing to perform in each round depending on the number of factors and their conditions, also considering the equipment for i) pipetting compositions (or cloning genetic constructs in case of biological sequences) and ii) measuring the objective function. The number of rounds of active learning also should be defined as any arbitrary number. It is important to note that the number of total experiments is recommended to be divided into more rounds as this improves the power of active learning[1]. Other parameters are the total volume of each composition, the portion of the volume that

could be varied excluding and listing the constant factors. This is important to specify the total volume and the volume of fixed components subtracted from it because the model needs to convert concentrations to volumes and vice versa. The minimum droplet size defines the minimum volume unit of liquid handling robots or the minimum unit that one can pipette, usually 0.2 or 0.5 µL. This means all the volumes proposed by the model will be a factor of 0.2/0.5 µL or, for example, 2.5/25 nL with Echo® acoustic robot.

The exploration/exploitation ratio referred to as "exploration" in the parameters section is one of the most important elements to thoughtfully define, especially when the user applies the workflow for a different application than those presented in this work. For applications with numerical factors (i.e., our examples in **Fig. 1, 3**), the exploration rate should be >1 for the early rounds of active learning, and toward the end of the cycle, it should tend to values <1. Importantly, these values depend on the number, importance, and type (numerical or categorical) of factors, therefore, for tailoring the workflow, it should be taken into account. A high exploration indicates that the model will take more risks in suggesting combinations (exploration) rather than being efficient (exploitation). High exploration ratios are needed in the early rounds to explore the space and escape from local optimal combinations. However, in the later rounds, more efficient suggestions are required, which corresponds to focusing on achieving higher yields using what the model has learned. The ratio should be very carefully assigned, and the users should rely on their knowledge of the system or check a few of the suggestions, especially for the use with categorical factors like for our examples in **Fig. 4, Supplementary Fig. 19**. Note that, even if an exploration ratio is assigned for Day 1 (in the absence of Day 0), the first step is fully randomized, and the ratio has no effect on it. The exploration ratio for the examples reported in this work is as follows. 10 rounds of Gfp production in the cell-free system (**Fig. 1e**): Not defined, 1.41, 1, 1, 1, 1, 0.5, 0.5, 0.5, for 10 rounds of *LacI* gene circuit optimization (**Fig. 3c**): Not defined, 1.41, 1.0, 1.0, 1.41, 1.41, 1.41, 1.41, 1.0, 0.5, for 2 rounds of 20 most informative combinations of *LacI* gene circuits with purified LacI (**Fig. 3i**): Not defined for Day 0, 0.5 for Day 1, for 4 rounds for the transcription and translation unit (**Fig. 4c**): Not defined, 1.41, 4.0, 1.0, for 2 rounds of 20 most informative combinations of transcription translation unit *in vivo* (**Fig. 4f**): Not defined for Day 0, 2 for Day 1, and for the single round of enzyme engineering suggestions for the hypothetical next day (**Supplementary Fig. 20**): 2.0. It must be noted that, since in METIS optimization combinations are sorted based on only their std value and exploitation is set to zero (this enables picking the most uncertain combination for the next round (**Methods**)), the exploration ratio is not defined.

In the factors section (**Fig. 2b**), the user should list all the elements participating in the objective function, that can be categorical and/or numerical features. Numerical features are defined by values, such as the concentration of factors in an *in vitro* system or in a growth medium, or strength of regulatory sequences such as promoters, RBSs (ribosome binding site), or CRISPR RNAs. Categorical features are not defined by values but characters/names such as regulatory sequences without numerical scores, or when for a gene there are multiple candidates from different organisms. Combined features are those with categories (alternative sequences/genes/constructs) and their level through concentration or strength (*LacI* gene circuits in **Fig. 3**). The user can range numerical values either by giving the range boundaries (minimum and maximum) or by specifying all the values. After running this section, for factors with range boundaries, it might be the case that the number of conditions is high, or their distribution lacks the desired coverage within the range. In this situation, we suggest picking 5 to 10 values from the proposed conditions from the output code of this section and manually specifying those factors' concentrations. The stocks' concentration is an important parameter since it must support the volume of the minimum final concentration in the defined range dependent on the minimum volume unit. One highlighted feature of our workflow that empowers its performance is letting the range of factors have more variable values than only 4 conditions in Borkowski *et al*[2]. The more randomized space allows the model to sample more informatively within the ranges, and this flexibility improves the performance. After the user has defined the active learning parameters and factors, the model shows all ranges taking into account the final volume of the mixture to give a percentage of possible compositions. A percentage of 100 means it is possible to compose a condition in which all factors are at their highest volume (concentration) and lower percentages indicate compositions are limited. When needed, the stocks' concentration can be changed, and if on the other hand it limits the lowest volume, two stocks with different concentrations can be provided. This is one of the features of the workflow that can solve issues when it comes to practice.

If a pre-existing dataset is used to train the model prior to starting active learning, they can be imported (compositions and yields) to initiate the workflow on Day 0 instead of Day 1 (used examples for 20 most informative combinations in **Fig. 3i, 4f** and the enzyme engineering application in **Supplementary Note 7, Supplementary Fig. 19, 20**). The dataset should be provided in the format of the results.csv files (**Data availability**).

When the parameters and factors are all set up, the model can generate the first round of experiments from the file Volumes_1.csv (round 1). The model generates Volumes.csv and Concentration.csv files in each round. The user should perform the suggested experiments, measure the objective function, and insert the results (mean and standard deviations) into new columns in the Concentrations_1.csv file generated by the model for that round and rename it to Results_1.scv. To continue with the next day (round), the user should upload the Results_1.csv file into the Google Colab notebook, run all sections, and the model executes the next day's compositions. In this step, the model trains itself on the results and suggests new compositions predicted to give higher values for the objective function (or more predictive combinations in METIS prediction to build a more predictive model and not necessarily optimizing the objective function). This cycle is repeated for n that is defined in the parameters section. The file used for pipetting is Volume.csv generated in each round. If the laboratory accesses Echo® acoustic liquid handler, our workflow can generate an input file corresponding to the suggested compositions. In this case, the source and destination plate wells should be specified by the user after downloading the Echo® file. Otherwise, compositions can be pipetted by other liquid handlers or manually using the Volume.csv file. In examples in **Fig. 1, 3**, for 20 compositions in each round, it took us around three hours to pipette all compositions using our table-to-speech virtual assistant. The table-to-speech assistant increases the speed, accuracy, and ease of pipetting that should be used through a separate Google Colab notebook that we also provided in this work. The notebook takes the Table2Speech_Volume.csv file as input, ranks, and reads them throughout all compositions for each factor. By adjusting the parameters in Google Colab the assistant will go to the next well with any click specified by the user (**Supplementary Note 1, Code availability**). If the objective function is as in our examples in **Fig. 4, Supplementary Fig. 20**, the combinations should be constructed or cloned, the Volume.csv lists the categories of factors that should be constructed and not real volumes to pipette.

During active learning, parameters or ranges can be altered at any round. This feature enables the user to make readjustments for the next rounds depending on the model's performance, the evolution of the objective function over rounds, and how factors behave within their ranges. Additionally, in the section named "specials", customized compositions can be manually added to any round and these will be subtracted from the number of suggestions in that round. If a user is willing to readjust or import special compositions, we recommend doing this in the later rounds of active learning and let the model perform its task in early rounds. For the *LacI* circuit optimization cycle, we show examples where customized

compositions have been manually added. Controls (i.e., negative control or reference combination) could be part of specials, in particular when an Echo® file is used.

The workflow generates cumulative outputs in each round (see **Methods** for how they are mathematically/statistically generated). The first set of outputs are compositions as Concentrations.csv and Volumes.csv. The second set is the list of K most informative combinations. This list can be used after active learning, to facilitate optimization of similar systems with the same type of factors and objective function such that one does not need to redo the whole active learning. By doing experiments for K (user-defined) combinations and measuring their objective function, results are imported as Day 0 to optionally continue with one or more rounds of active learning (see *LacI* circuit section). The last set of outputs are several types of analysis/visualization and the respective raw data, a box plot for the evolution of objective function over rounds, two groups of individual plots for each factor one showing the daily variation of factors within their ranges, and the other group is about all measured yields within ranges. The other analysis/visualization outputs are a plot and list of feature importance values, contribution of each factor in the prediction of yield by the model. These provide information on factors playing more important roles and which have less or no effect on the objective function. The last module is a heatmap of mutual interactions between every two factors providing useful results to optimize and study the system as well as to spotlight unknown interactions.

Supplementary Note 3: Complementary note on the experiment and discussion for optimization of *LacI* gene circuits

The median of objective function increased especially from Day 6, and the fold-change raised from less than 4 to over 8 on Day 10 (**Fig. 3c**). The median on Day 9 dropped because we included 5 of 20 as special compositions with high concentrations of P_{T7}-*LacI* plasmid and purified T7 RNA polymerase that led to very low objective function values. We did so since the model was trending to suggest very low concentrations of these two factors (**Supplementary Fig. 8**) which we first assumed would improve the fold change and objective function. Although this manipulation in the active learning process reduced the objective function values on Day 9, we gained a more profound insight into the system's behavior. Additionally, this helped the model to improve the objective function values at the highest on Day 10. We hypothesize that the low protein production is because of resource competition or inhibitory protein-

protein interactions especially at high concentrations of the plasmid that higher amounts of amino acids and tRNAs could not compensate for it. This was noticed by the model such that during the 10-day period it was tending to suggest lower concentrations of P_{T7}-*LacI* plasmid and T7 RNA polymerase (**Fig. 3d, Supplementary Fig. 8**). **Fig. 3d, e** represent yield data points within each factor's range and the features' importance, respectively. The model calculated that P_{T7}-*LacI* plasmid is by far the most important feature because of its hugely negative effect on protein production (**Fig. 3e**).

To test the hypothesis of resource competition or inhibitory protein-protein interaction, we designed titration experiments in which with the optimal composition from active learning (which was with the pTHS circuit) we varied the concentration of P_{T7}-*LacI* plasmid and T7 RNA polymerase (**Fig. 3f**). Increasing the concentration of P_{T7}-*LacI* plasmid or T7 RNA polymerase diminishes the fold-change × dynamic range (FC × DR) and fold change (FC). The maximum of these values was achieved at low concentrations of P_{T7}-*LacI* plasmid and the highest amount of T7 RNA polymerase. To further support this hypothesis, instead of pTHS circuit, we added a *Gfp* expressing plasmid with a constitutive promoter that has no interaction with the P_{T7}-*LacI* plasmid and T7 RNA polymerase but shares the same resources for transcription and translation. Increasing the concentration of either P_{T7}-*LacI* or T7 RNA polymerase depleted resources for expression of the constitutive *Gfp* (**Fig. 3g**). Since the addition of the second plasmid prevented reaching a sufficient production-repression balance, we sought to test the system with His-tag purified LacI. We extracted the 20 most informative combinations from the above active learning cycle. We removed T7 RNA polymerase from the factors and replaced the P_{T7}-*LacI* plasmid with His-tag purified LacI and gave random values to it (**Fig. 3h**). These 20 combinations were performed, the results were collected and imported as Day 0 to the workflow. We already could see a huge improvement in the objective function and fold-change on Day 0 (**Fig. 3i**) due to using purified LacI. As high yield values were associated with the lower concentrations of purified LacI, we widened the range of purified LacI toward lower concentrations for the next round (Day 1). The Day 1 experiment resulted in 4 data points with a fold change of around 100 (right plot in **Fig. 3i**).

Supplementary Note 4: <u>Setup for the METIS based optimization of the CETCH cycle</u>

For the optimization of the CETCH cycle, we did 125 conditions in triplicates per round. In each condition, all components were variable except for the substrate propionyl-CoA, which was fixed to 100 µM. For the

buffer (Hepes), we not only wanted to optimize the concentration, but also test different pH. Therefore, we gave the possibility to choose between 6 different pHs (7.0, 7.2, 7.4, 7.6, 7.8 and 8.0). For the five rounds of unrestricted optimization, we used exploration values of 1.41, 1.41, 1.0, 1.0 and 0.5. After transforming the data for the efficiency optimization ([Glycolate]/[Enz_{tot}]= Gycolate yield in µM and total concentration of enzymes in µM), we did three additional rounds of efficiency optimization where we used exploration values of 1.0, 0.5 and 0.5.

After the first two rounds, we purified new batches of mco, hbd, cat and ssr. In the third round, using the new enzyme stocks, we noticed that our positive control (see material and methods) had a roughly three times higher product yield compared to the first two rounds. Therefore, we tested our control manually with the new enzymes and 4 control setups where only the old enzyme stock of either mco, hbd, cat or ssr was used. We could identify the old catalase stock as the reason for the lower yields in the first two rounds (**Supplementary Fig. 16b**). We stored the new catalase stock in liquid nitrogen upon use. Additionally, we removed unnecessary values after the first three rounds to reduce the combinatorial space (see **Supplementary Table 2**). All other values are in the GitHub repository.

Assays for determination of new enzyme stocks after round two (**Supplementary Fig. 16b**)

The assays were done in triplicates containing 30 µL volume each and were carried out in a 1.5 mL reaction tube (at 30 °C, 500 rpm). The reactions were started with 100 µM propionyl-CoA. 8 µL samples were taken and quenched in 1 µL 50% formic acid and 1 µL 500 mM sodium polyphosphate (emplura®) at 1, 2, and 3 h. The samples were spun for 20 min at 4 °C and 20.000g, before the supernatant was transferred into new tubes. As described earlier, the samples were diluted and mixed with the internal standard for LC-MS measurement.

The assays contained:

100 mM HEPES pH 7.6, 5 mM $MgCl_2$, 50 mM sodium bicarbonate, 20 mM sodium formate, 20 mM creatine phosphate, 0.5 mM CoA, 0.1 mM CoB_{12}, 2 mM ATP, 5 mM NADPH, 3.06 µM pco, 0.62 µM ccr, 0.74 µM epi, 0.30 µM mcm, 2.62 µM scr, 0.53 µM ssr, 5.34 µM hbs, 0.73 µM hbd, 0.58 µM ecm, 21.90 µM mco, 0.28 µM mch, 2.79 µM mcl, 1.64 µM cat, 14.56 fdh, 0.02 µM ca, 1.10 µM gor, 0.39 µM ck.

The experiments labeled mco, hbd, cat and ssr were done with the old stocks of the enzymes used in the first two rounds. The controls contained enzymes from the four new enzymes batches.

Assays for different concentrations of hbs and cobalt (**Supplementary Fig. 16a**)

The assays were done in triplicates containing 30 µL volume each and were carried out in a 1.5 mL reaction tube (at 30 °C, 500 rpm). The reactions were started with 100 µM propionyl-CoA. 8 µL samples were taken and quenched in 1 µL 50% formic acid and 1 µL 500 mM sodium polyphosphate (emplura®) at 1, 2 and 3 h. The samples were spun for 20 min at 4 °C and 20.000g, before the supernatant was transferred into new tubes. As described earlier, the samples were diluted and mixed with the internal standard for LC-MS measurement.

The hbs assays contained:

100 mM HEPES pH 7.6, 5 mM $MgCl_2$, 50 mM sodium bicarbonate, 20 mM creatine phosphate, 0.5 mM CoA, 0.1 mM CoB_{12}, 2 mM ATP, 10 mM NADPH, 10 mM NADH, 3.06 µM pco, 0.62 µM ccr, 0.74 µM epi, 0.30 µM mcm, 2.62 µM scr, 0.53 µM ssr, 0.53 µM hbs (10%), 5.34 µM hbs (control), 26.7 µM hbs (500%), 0.73 µM hbd, 0.58 µM ecm, 21.90 µM mco, 0.28 µM mch, 2.79 µM mcl, 1.64 µM cat, 0.02 µM ca, 1.10 µM gor, 0.39 µM ck.

The cobalt assays contained:

100 mM HEPES pH 7.6, 5 mM $MgCl_2$, 50 mM sodium bicarbonate, 20 mM sodium formate, 20 mM creatine phosphate, 0.5 mM CoA, 1.0 mM cobalt (1 mM), 0.0 mM cobalt (0 mM), 2 mM ATP, 5 mM NADPH, 3.06 µM pco, 0.62 µM ccr, 0.74 µM epi, 0.30 µM mcm, 2.62 µM scr, 0.53 µM ssr, 5.34 µM hbs, 0.73 µM hbd, 0.58 µM ecm, 21.90 µM mco, 0.28 µM mch, 2.79 µM mcl, 1.64 µM cat, 14.56 fdh, 0.02 µM ca, 1.10 µM gor, 0.39 µM ck.

Supplementary Note 5: Optimization versus prediction purpose

So far, we demonstrated the use of our active learning workflow for the experimentally guided optimization of cell-free protein expression, *LacI* gene circuit, transcription and translation unit, and the CETCH cycle. Machine learning algorithms learn patterns in data and predict unseen cases which can be either applied for optimization (as demonstrated with our workflow above) or prediction purposes. Note that while both approaches employ a dedicated prediction step, their overall aim is different (**Supplementary Table 1**). To further modularize our workflow, we created an additional METIS package for prediction (not optimization) of an objective function of given combinations. **Supplementary Table 1** summarizes the features of two Google Colab packages, METIS optimization and METIS prediction, which

differ in goal and application, query strategy, and outputs. The query strategy or the approach of METIS optimization is maximizing the objective function as well as the standard deviation of the ensemble regressor (i.e., the result of multiple regressors that individually operate on the data) whereas METIS prediction only maximizes the standard deviation of the ensemble regressor. These two also differ in the output types; METIS optimization suggests combinations with high ranked objective function whereas METIS prediction generates a trained model that computes the objective function for input combinations.

Each round's suggestions in METIS prediction aim to build a more predictive model by maximizing the correlation between predicted and measured objective functions (and not necessarily maximizing the objective function as in METIS optimization), hence improve the R^2 of prediction over rounds of learning. Beyond these differences, the two notebooks share the same features/modules as described in **Fig. 2** and **Supplementary Note 2**. To test the prediction notebook, we ran a simulation on a dataset of 1094 data points[3] from the PURE (purified recombinant elements) cell-free protein expression system in which recombinant proteins and the buffer composition were varied (**Supplementary Note 6, Supplementary Fig. xx**). To demonstrate the performance of the model, we split the dataset into train and validation sets and used the validation set to measure (using R^2) the accuracy of the prediction on the test set. Because of the random nature of splitting, the simulation was repeated 5 times. (**Supplementary Fig. 18**). Overall, the workflow was able to improve the prediction of the objective function over rounds of active learning.

Supplementary Note 6: <u>Application of METIS prediction for simulation on a PURE cell-free system dataset</u>

To show METIS's potential in predicting an objective function, we performed an active learning simulation on an available PURE (purified recombinant elements) cell-free system dataset[3] (**Supplementary Fig. 18a, b**). We also assessed the predictability of the dataset and our model's general performance, we calculated 5-fold cross-validation on the whole dataset (1098 data points). The result of 5-fold cross-validation and the prediction of a single sample test set is shown in **Supplementary Fig. 18c, d**, respectively.

We simulated an active learning cycle as presented in **Supplementary Fig. 18a**:

- 20% of the dataset was separated as the test set to be used for validation, which was not taken into account when the model is trained on the training set (the other 80%).
- In the first round, 80 data points from the training dataset were selected randomly and the model was trained on them.

- In the next round, using what the model has learned, the rest of the training set (80 subtracted) was sorted based on the uncertainty value of combinations (**Methods**) and the top 80 combinations were picked.
- The yield of the picked combinations was assigned from the dataset and the model was trained on them.
- This process was repeated 10 times, hence, at the end the model was trained on 800 data points.
- In each step performance of the model was evaluated on the test set.

This process includes multiple random steps which is why we repeated the whole process five times as shown in **Supplementary Fig 18a**.

Supplementary Note 7: Application of the workflow for combinatorial enzyme engineering

Machine learning is a powerful tool to address complicated biological questions such as prediction/engineering of the structure/activity of proteins/enzymes[4,5]. Challenges in engineering a protein using machine learning tools are the dependency on large datasets and difficulties to make genotype-phenotype links for hundreds or thousands of variants[4]. Except for phenotypes easy to measure such as those related to regulatory sequences (transcription factors), there is a lack of modular characterization methods for engineering proteins. Enzymes are of difficult proteins to engineer because each enzyme requires a different characterization method which mainly allows for low to medium throughput experiments. Moreover, traditional enzyme engineering approaches which mostly rely on altering one amino acid at a time, are likely to be trapped into the local optima of the enzyme activity[6]. Since our tool is able to work with minimal datasets and the gradient boosting algorithm can capture mutual interactions, it can also be used for engineering enzymes. From a recent study in our lab, we took a dataset of mutants of oxalyl-CoA decarboxylase from *Methylorubrum extorquens*[7] to run simulations and give an example of how to use the workflow for such applications. **Supplementary Fig. 19a** shows the active site of the enzyme and surrounding amino acid residues.

We imported the dataset comprising 847 combinations of mutations in the active site of the enzyme (**Data availability**). As provided in the modular workflow, we first extracted and plotted the feature importance values (**Supplementary Fig. 19b**). E135G and Y497F were calculated as conditions with the highest effect

on the yield. Although such conclusions could be made by experimental biochemists analyzing the mutants, our tool can quantify the effect of various amino acids at each position on the activity of the enzyme. The tailored workflow that we created for protein/enzyme engineering is provided with extra modules for the analysis and visualization of a mutant dataset. One of these provides 5-fold cross-validation on the whole dataset (**Supplementary Fig. 19c**). K-fold cross-validation assesses the predictivity of machine learning models on an independent dataset (**Methods**). An average of 0.65 (Pearson R^2) for 5-fold cross-validation was achieved on this dataset. The other module plots the performance of the model on a single test set, 20% of the whole data, after being trained on the other 80%. **Supplementary Fig. 19d** shows predicted values of the test set versus their measured values for the mutants dataset which supports an existing correlation between predicted and actual values. Our enzyme engineering notebook also enables continuing with active learning cycles from an imported dataset (**Supplementary Fig. 20**). One can also start from scratch to engineer an enzyme by following a procedure similar to the examples shown in this study.

Preparation of the dataset for enzyme mutants

Data for the iterative saturation mutagenesis of MeOXC was acquired as described previously[7]. Briefly, a three-enzyme cascade was employed to turn over formaldehyde and formyl-CoA to glyoxylate. The last step in the cascade formed hydrogen peroxide, which was used to convert Ampliflu Red to resorufin, allowing fluorometric detection. Mutants were evaluated by two parameters. First, the maximal rate of signal production was determined (parameter A), and second, the final amount of signal produced after 2 hours of runtime was measured (parameter B). The assigned objective function to each mutant is the product of parameters A and B. All values were normalized to the wildtype. If possible, the specific mutations were matched to the dataset. Where no sequencing data was available, the average signal was determined and assigned to the missing mutants.

2.7.2. Supplementary Tables

Supplementary Table 1. Summary of METIS optimization and prediction Google Colab packages. In this study, we provided two modular Google Colab notebooks that can be adjusted for different applications as shown in the study. These two notebooks differ in applications, query strategy, inputs and outputs.

Notebook	**Application**	**Query Strategy**	**Input/start**	**Output**
METIS optimization	-Optimization of an objective function, i.e., a composition, pathway, genetic construct/circuit	-Maximize the objective function as well as the standard deviation of ensemble regressor	-Existing data OR -Start with randomly generated compositions	-Compositions of categorical and/or quantitative factors that lead to higher yields
METIS prediction	-Prediction of an objective function, i.e., a composition, pathway, genetic construct/circuit	-Maximize only the standard deviation of ensemble regressor	-Existing data OR -Start with randomly generated compositions	-A trained model that predicts the objective function of given compositions

Supplementary Table 2: The removed conditions after the first 3 rounds of yield active learning.

Compounds	Removed values
HEPES	25
MgCl2	22.5, 25
Creatine P	80, 100
Bicarbonate	100
Formate	100
B12	0.6, 0.8, 1.0
pco	0.1914, 0.3827, 0.669725, 1.243775
hbs	1.06785, 2.1357, 3.737475, 6.941025
mco	1.36885, 2.7377, 4.10655, 5.4754, 6.84425, 9.58195, 13.6885, 19.1639

fdh	1.456, 2.912, 5.824, 10.192
gor	0.55235, 1.1047, 1.65705, 2.2094

Supplementary Table 3: Timepoint assays of 7 selected conditions with their active learning yield (AL yield) and efficiency (AL efficiency) values. Blue, day 4, condition 29, pH 7.8; orange, day 7, condition 15, pH 7.8; red, day 7, condition 76, pH 7.2; black, control, pH 7.6. green, day 8, condition 17, pH 7.2; lavender, day 5, condition 17, pH 7.4; burgundy, day 5, condition 55, pH 7.8.

	HEPES	MgCl2	CP	Bicarb.	Form.	CoA	B12	ATP	NADPH	pco	ccr	epi	mcm	scr	ssr	hbs	hbd	ecm	mco	mch	mcl	kat	fdh	ca	gor	ck	**Manual Yield (3h)**	**AL yield**	**AL Efficiency**
blue	75	12.5	5	10.0	40	0.5	0.1	10	3.75	2.3	2.2	1.5	1.5	7.0	4.4	0.5	1.5	2.9	46.5	0.3	14.7	1.6	13.1	0.0	4.4	2.7	2262.5	2869.6	26.8
orange	75	12.5	60	2.5	20	0.4	0.0	3	3.75	3.1	1.9	0.7	2.9	3.5	1.7	0.5	0.7	1.4	26.0	0.3	3.6	3.3	30.6	0.1	5.0	0.8	2126.8	2617.7	30.5
red	150	7.5	20	20.0	40	0.4	0.0	9	2.50	2.3	2.2	3.7	0.3	1.7	4.4	1.6	0.4	2.6	113.6	2.3	8.4	4.9	23.3	0.1	3.9	2.7	1872.5	2586.8	14.5
black	100	5.0	20	50.0	20	0.5	0.1	2	5.00	3.1	0.6	0.7	0.3	2.6	0.6	5.3	0.7	0.6	21.9	0.3	2.8	1.6	14.6	0.0	1.1	0.4	959.6		
green	175	10.0	10	5.0	80	0.4	0.1	9	2.50	2.3	0.6	3.0	0.3	5.2	1.1	1.6	0.4	2.3	26.0	1.4	14.7	1.6	40.8	0.1	2.8	2.0	920.2	919.9	8.7
lavender	50	2.5	5	2.5	5	0.4	0.1	7	7.50	4.0	2.8	0.7	1.2	9.6	2.2	1.6	1.1	2.6	46.5	0.6	1.2	8.2	7.3	0.1	4.7	3.1	682.7	627.9	6.4
burgundy	50	17.5	10	5.0	20	4.0	0.1	5	5.00	3.1	1.5	3.7	0.9	13.1	4.4	12.3	1.8	2.9	34.2	2.0	1.6	4.9	23.3	0.0	3.6	0.8	296.6	354	3.1

Supplementary Table 4: Gradient for the separation of CoA esters.

Time [min]	**A [%]**	**B [%]**
0.0	100	0
0.5	100	0
6.0	96	4
10.0	77	23
11.0	20	80

12.0	20	80
12.1	100	0
15.0	100	0

Supplementary Table 5: Multiple reaction monitoring (MRM) transitions for measurement of CoA esters.

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
Ethylmalonyl-CoA (Quantifier)	882.1	331.2	25	380	41	5	Positive
Ethylmalonyl-CoA (Qualifier)	882.1	428	25	380	29	5	Positive
Methylsuccinyl-CoA (Quantifier)	882	375.1	25	380	33	5	Positive
Methylsuccinyl-CoA (Qualifier)	882	428	25	380	29	5	Positive
Mesaconyl-CoA (Quantifier)	880.1	375.1	25	380	25	5	Positive
Mesaconyl-CoA (Qualifier)	880.1	428	25	380	35	5	Positive
Succinyl-CoA (Quantifier)	868.1	361.1	25	380	35	5	Positive
Succinyl-CoA (Qualifier)	868.1	428.1	25	380	35	5	Positive
Methylmalonyl-CoA (Quantifier)	868.1	317.1	25	380	41	5	Positive
Methylmalonyl-CoA (Qualifier)	868.1	428	25	380	31	5	Positive
4-hydroxybutyryl-CoA (Quantifier)	854.1	347.1	25	380	37	5	Positive

4-hydroxybutyryl-CoA (Qualifier)	854.1	428	25	380	30	5	Positive
Crotonyl-CoA (Quantifier)	836.1	329	25	380	33	5	Positive
Crotonyl-CoA (Qualifier)	836.1	428	25	380	26	5	Positive
Propionyl-CoA (Quantifier)	824.1	317.1	25	380	31	5	Positive
Propionyl-CoA (Qualifier)	824.1	428	25	380	28	5	Positive
B-methylmalyl-CoA (Quantifier)	898.1	391.1	25	380	39	5	Positive
B-methylmalyl-CoA (Qualifier)	898.1	428.1	25	380	33	5	Positive

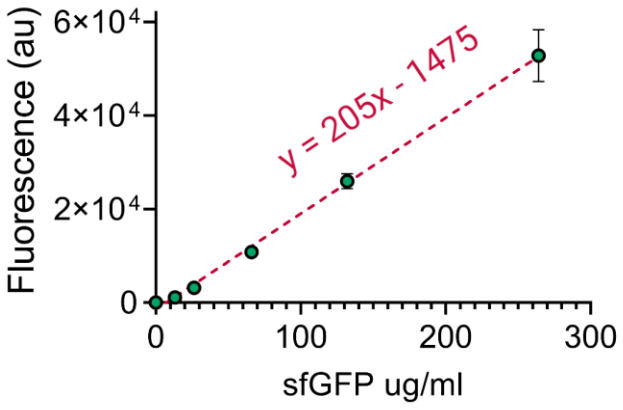
Supplementary Table 6: Gradient for the separation of glycolate.

Time [min]	A [%]	B [%]
0	100	0
4	100	0
6	0	100
7	0	100
7.1	100	0
12	100	0

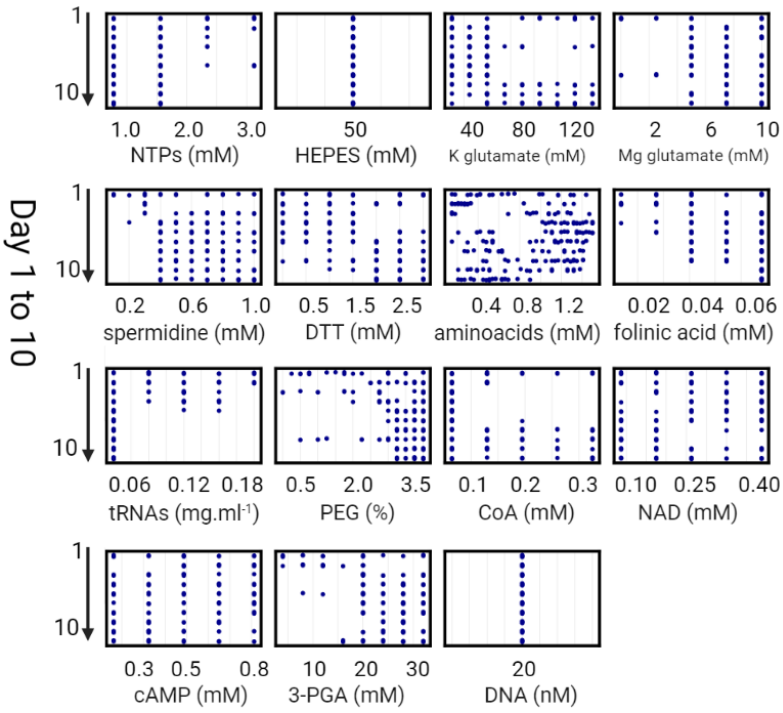
Supplementary Table 7: Multiple reaction monitoring (MRM) transitions for measurement of glycolate.

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
^{12}C-Glycolate (Quantifier)	75	47	150	380	9	5	Negative
^{12}C-Glycolate (Qualifier)	75	75	150	380	0	5	Negative
^{13}C-Glycolate (Quantifier)	77	48	150	380	9	5	Negative
^{13}C-Glycolate (Qualifier)	77	77	150	380	0	5	Negative

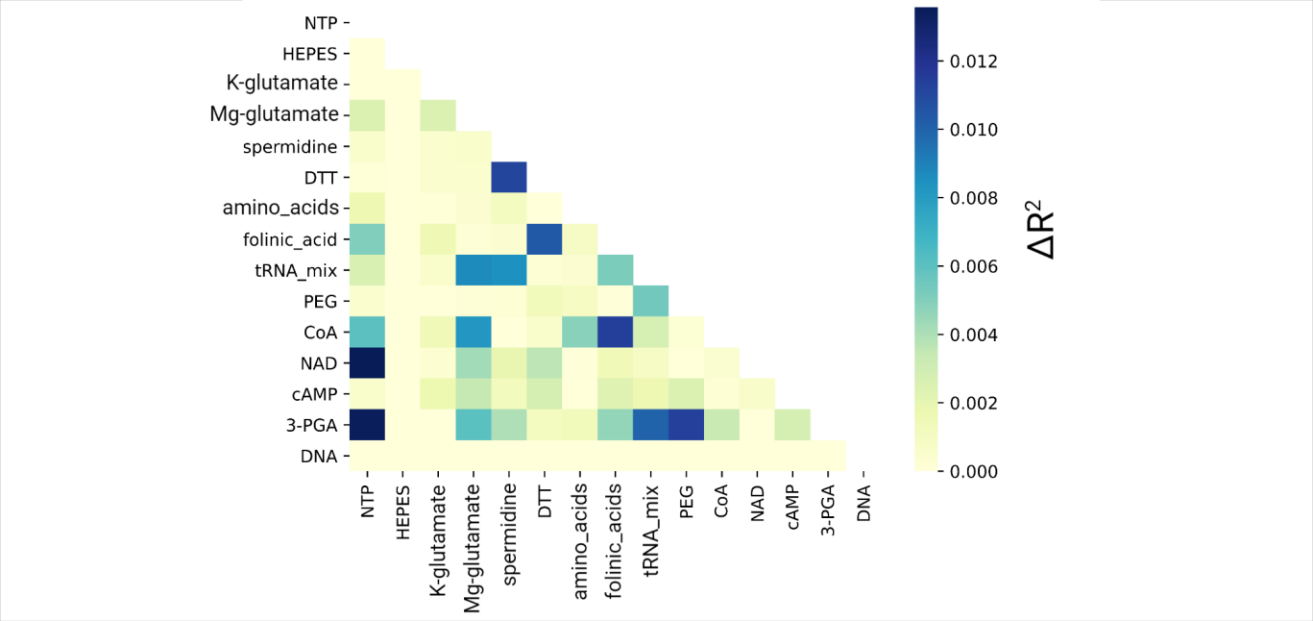
2.7.3. Supplementary Figures



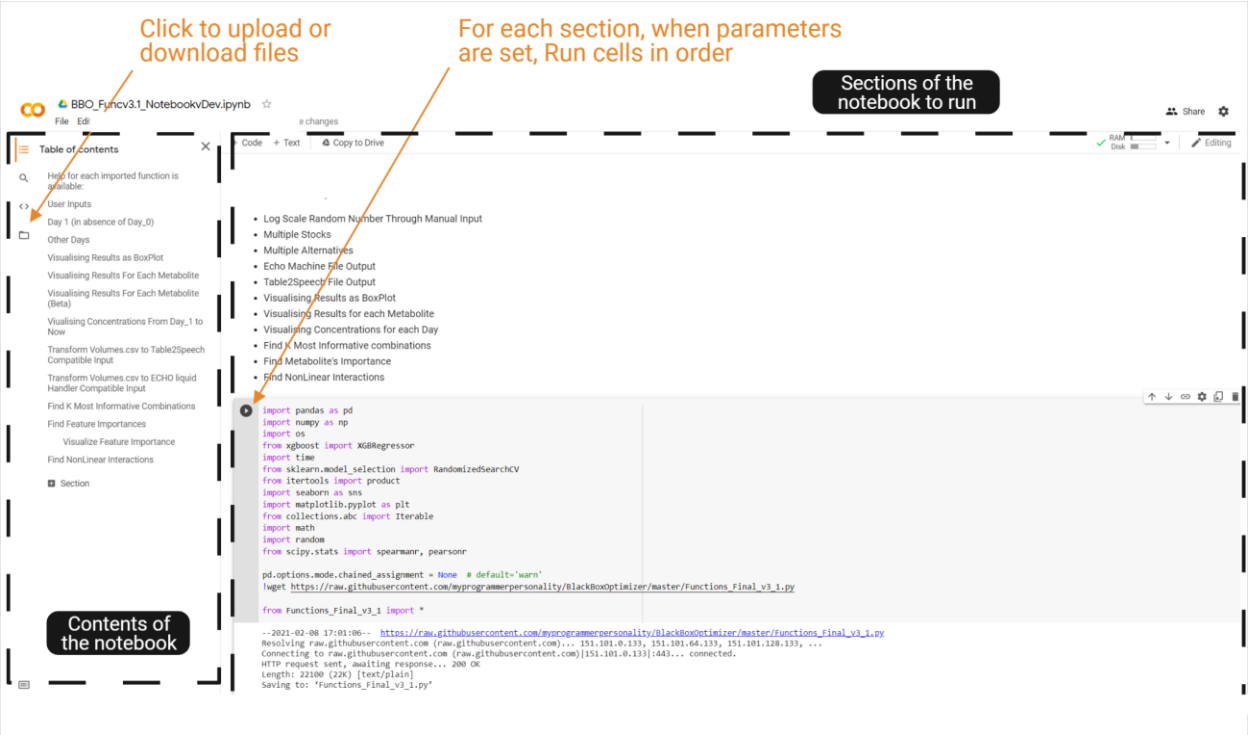
Supplementary Fig. 1: Standard curve for fluorescence readout of purified sfGfp (26806 Da) at the condition of the reference cell-free reaction. The experiment was performed in the same plate reader (Tecan Infinite 200 PRO, excitation/emission wavelengths of 485/528 nM and gain = 80). The average of measured Gfp fluorescence for the reference composition[8] is 3849.13 which corresponds to a concentration of 25.97 ug/ml (0.97 µM). This is the amount of sfGfp for the reference that we used to normalize all the data points in **Fig. 1e**. Quick math shows that the points with a yield of ~15 in **Fig. 1e** have a GPF production of ~0.97 $\times$ 15 = 14.5 µM. Compared to a previous optimization and a commercial kit (See the supplementary information of Borkowski *et al.* Figure 8b[2]) we achieved more than 30-fold higher sfGfp production. It should be noted that we used autolysate preparation protocol whereas in the other work, a sonication protocol was used. Moreover, it is important to note that the concentration of plasmid DNA used in this study is 20 nM compared to the 10 nM in the compared work, however, they used a stronger RBS, B0034. The strength of RBS (B0032) in our plasmid is 30% of B0034[9].



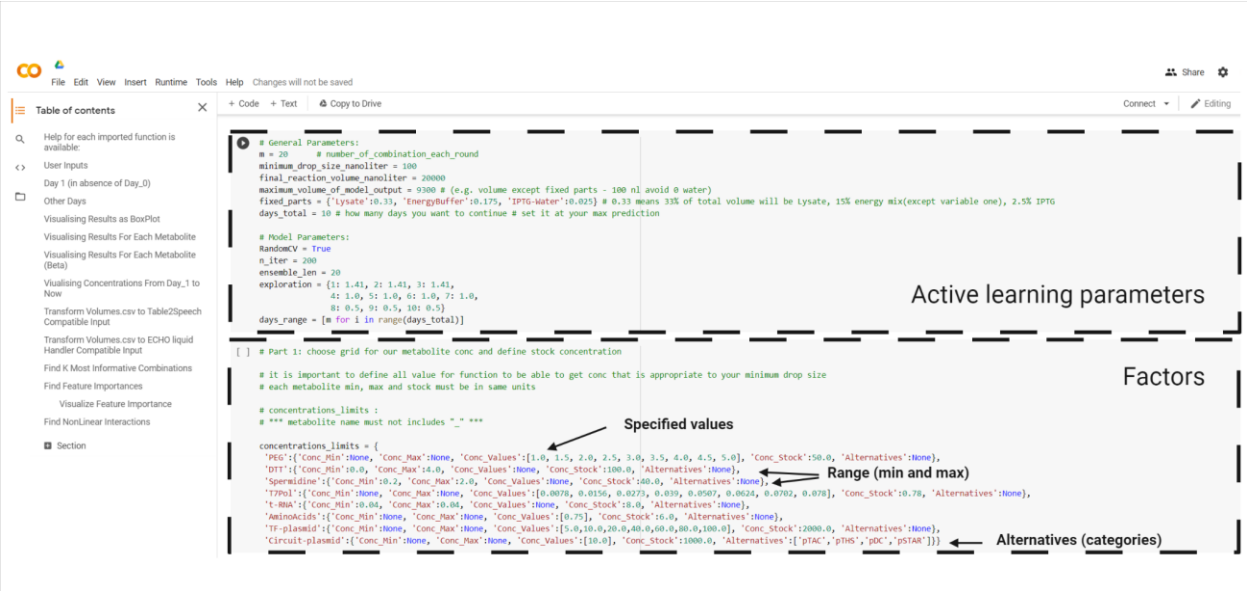
Supplementary Fig. 2: Concentration variations of each of the *E. coli* cell-free system factors from rounds 1 to 10 of active learning. These plots show how the concentration of each factor varied from day 1 to 10 through suggestions of the model aiming to increase the yield. The features' importance (**Fig. 1f**), and yield dependencies (**Fig. 1g**) could be analyzed altogether with these plots.



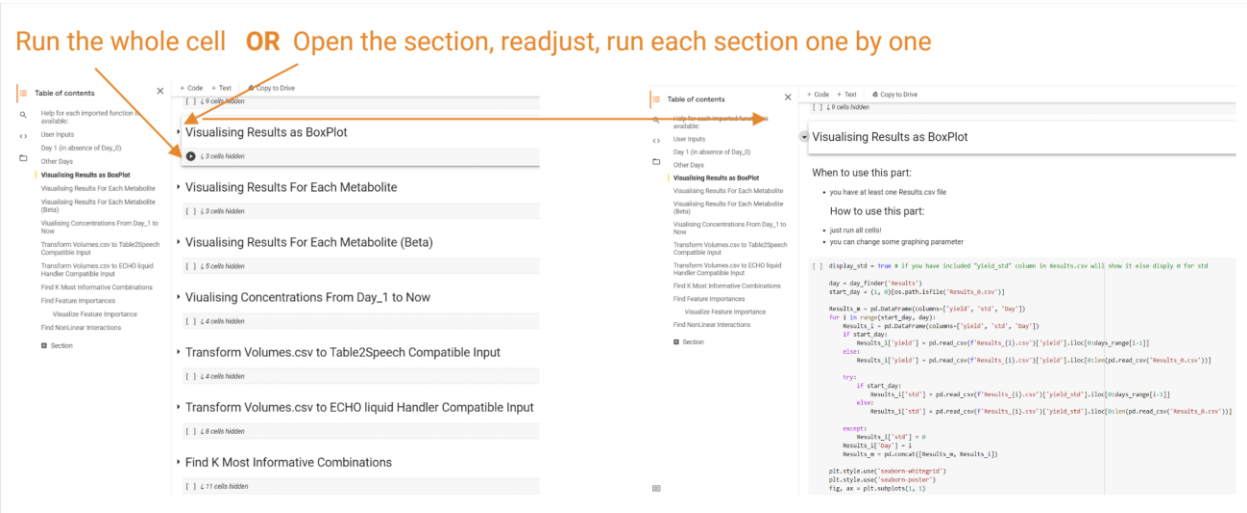
Supplementary Fig. 3: Mutual interactions between every two factors for active learning of the *E. coli* cell-free system. This plot shows the calculated mutual interactions (**Methods**) between every two factors, a useful analysis of the active learning dataset helping to study the system. This module is also integrated into the modular workflow.



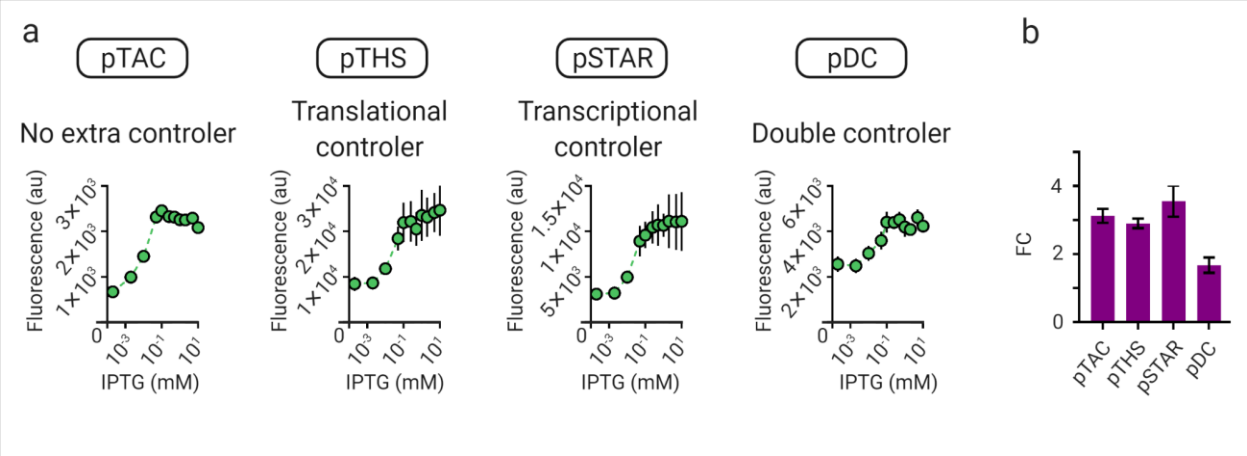
Supplementary Fig. 4: A screenshot of the Google Colab notebook for interacting with the workflow. At the left side of the page, the folder icon (shown by the arrow) is where files should be uploaded (Results) or downloaded (those that the user generated during the usage such as Volume, Concentration, Features' importance, K most informative, as well as figures as .png or .svg). When the run time is over, the user should reupload the files. The main body of the tool is where different cells (modules) of the code are accessible that should be run in order by clicking on the run icon (shown by the arrow). Except for the first round for which there is no results or input file, the files should be uploaded before starting to run the cells. For more details see **Code availability** to open (after making a personal copy) any of provided notebooks in your browser.



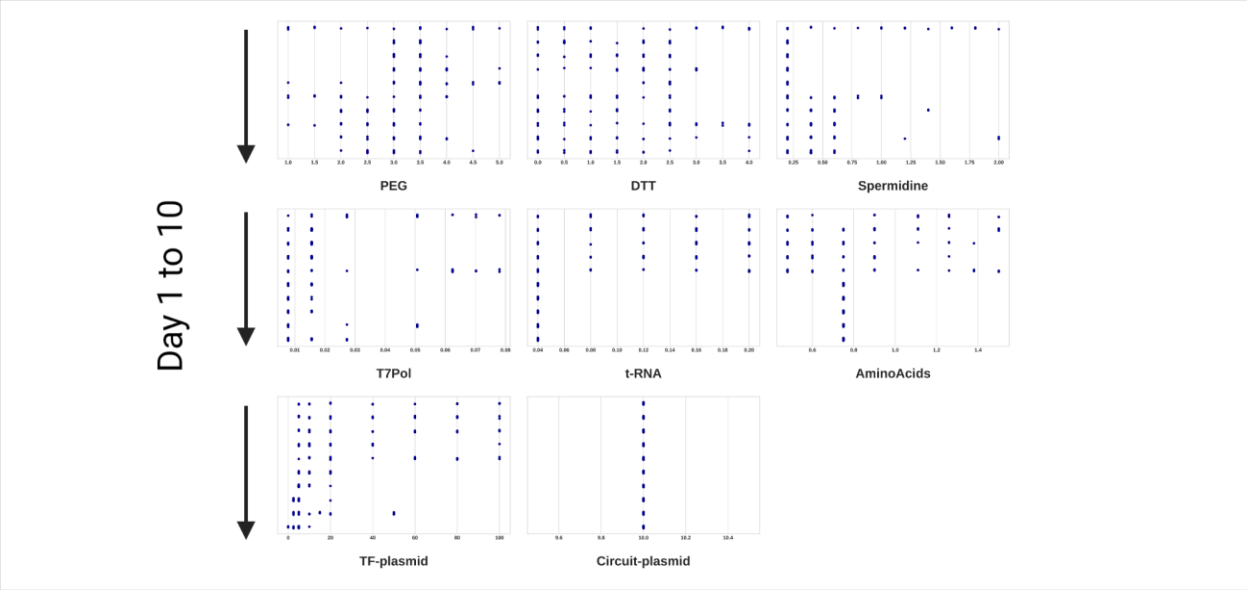
Supplementary Fig. 5: A screenshot of the sections of active learning parameters and factors. These two sections appear at the top of the notebook. These cells come after the sections for python classes/modules that should run beforehand. See **Supplementary Note 2** for detailed explanations.



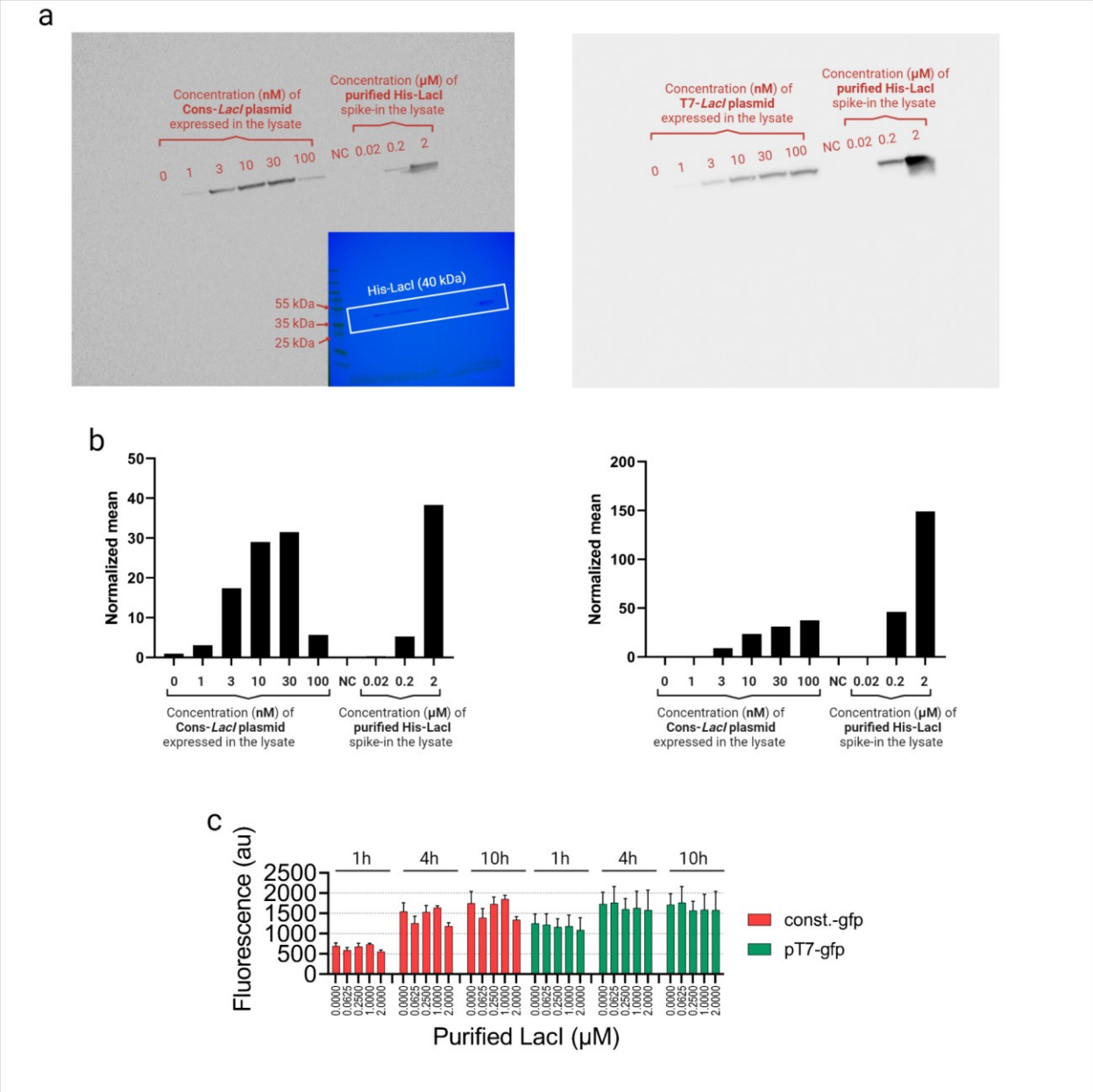
Supplementary Fig. 6: A screenshot of different modules of the workflow. The modules that a user aims to use can be run all at once or the user can open individual code sections/tabs and make changes for example in the plot size, color, etc.



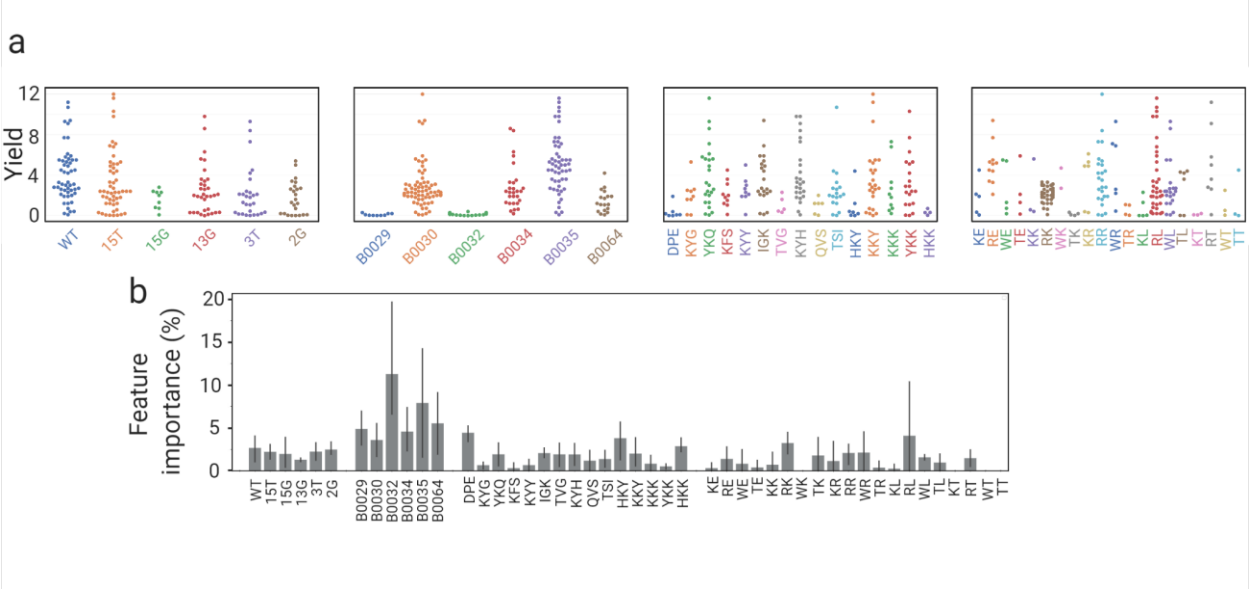
Supplementary Fig. 7: The behavior of *LacI* gene circuits in the cell-free protein expression system. These results presented in Greco *et al*.[10] as the "rate" in the protein production; however, here, we plotted the "amount" of protein produced after 6 h of incubation at 30 °C. **(a)** IPTG dose-response curve for SLC and MLC constructs. **(b)** The fold-change (FC) value of the plots in **(a)** between 0 and 10 mM of IPTG.



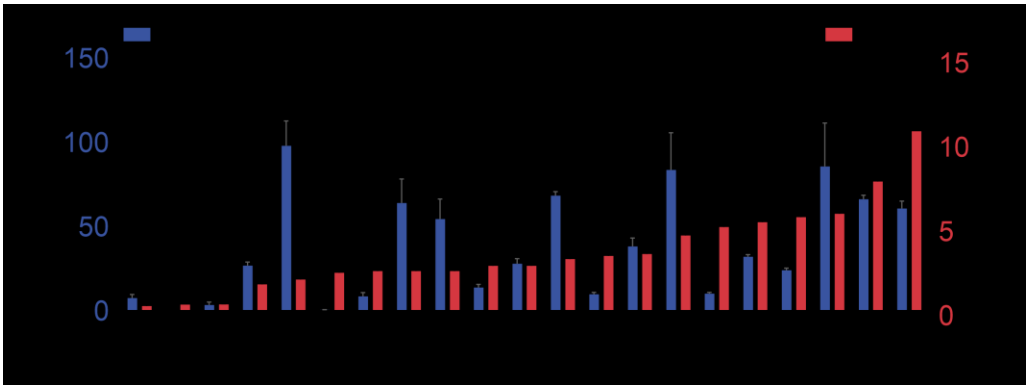
Supplementary Fig. 8: Variation of the concentration of factors in 10 rounds of active learning for *LacI* gene circuits optimization. These plots show suggestions of different concentrations of each factor by the model aiming to improve the objective function.



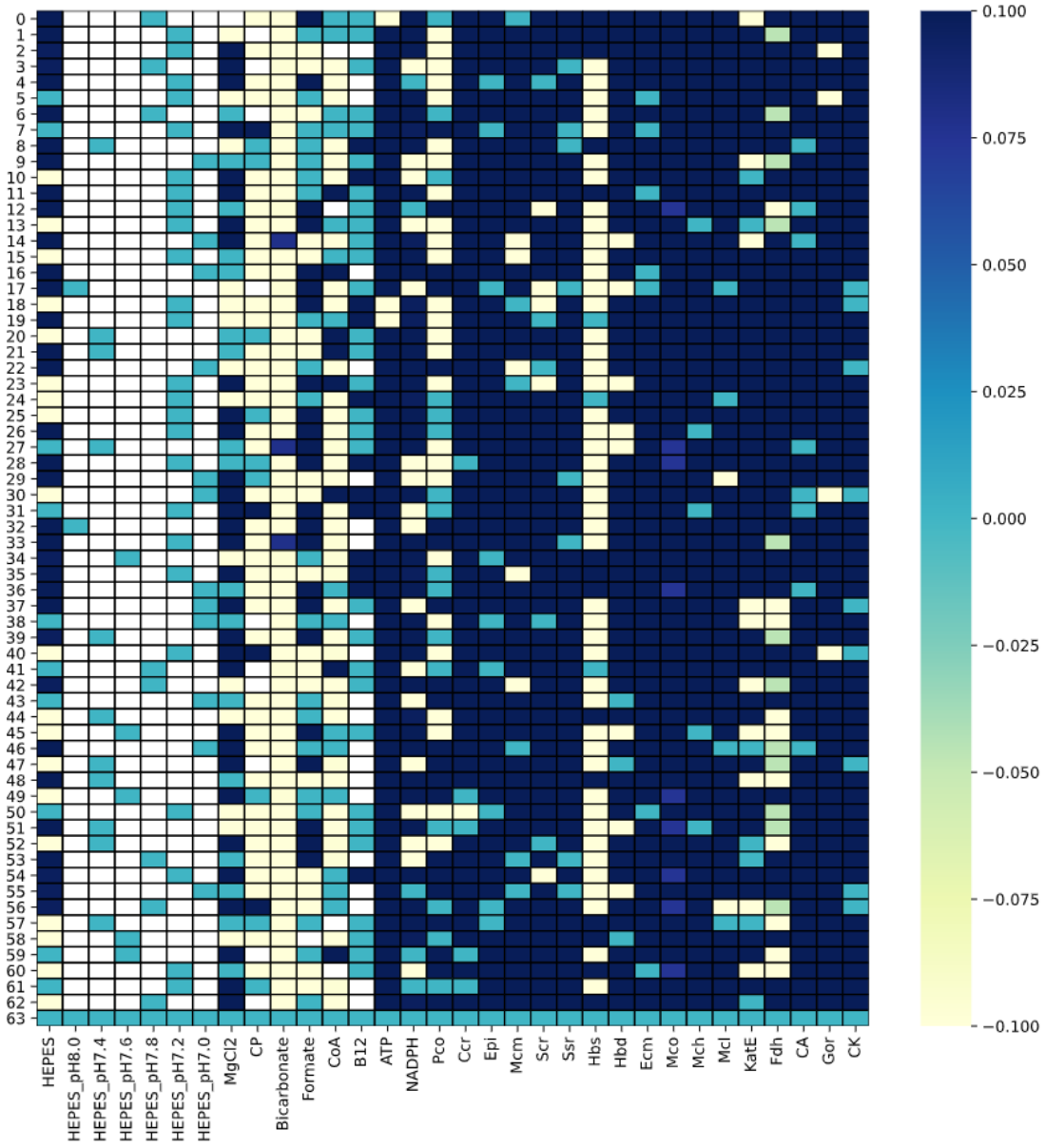
Supplementary Fig. 9: Gfp fluorescence level is affected by expression competition with *LacI*, not by interference by the LacI protein itself. a) Western chemiluminescence detection of LacI protein level produced in cell-free reactions starting with different concentrations of the *LacI* plasmid. The inset shows an overlay of the chemiluminescent signal and a bright image of the membrane, where the ladder is visible. **b)** Quantification of the mean intensity of LacI bands in **(a)** using imageJ software. These data show that all cell-free reactions produce less than 2 µM of the LacI protein. **c)** Gfp fluorescence from cell-free reactions with different concentrations of purified LacI-6xHis protein included indicates that the LacI protein itself does not affect *Gfp* expression level at 2 µM and below.



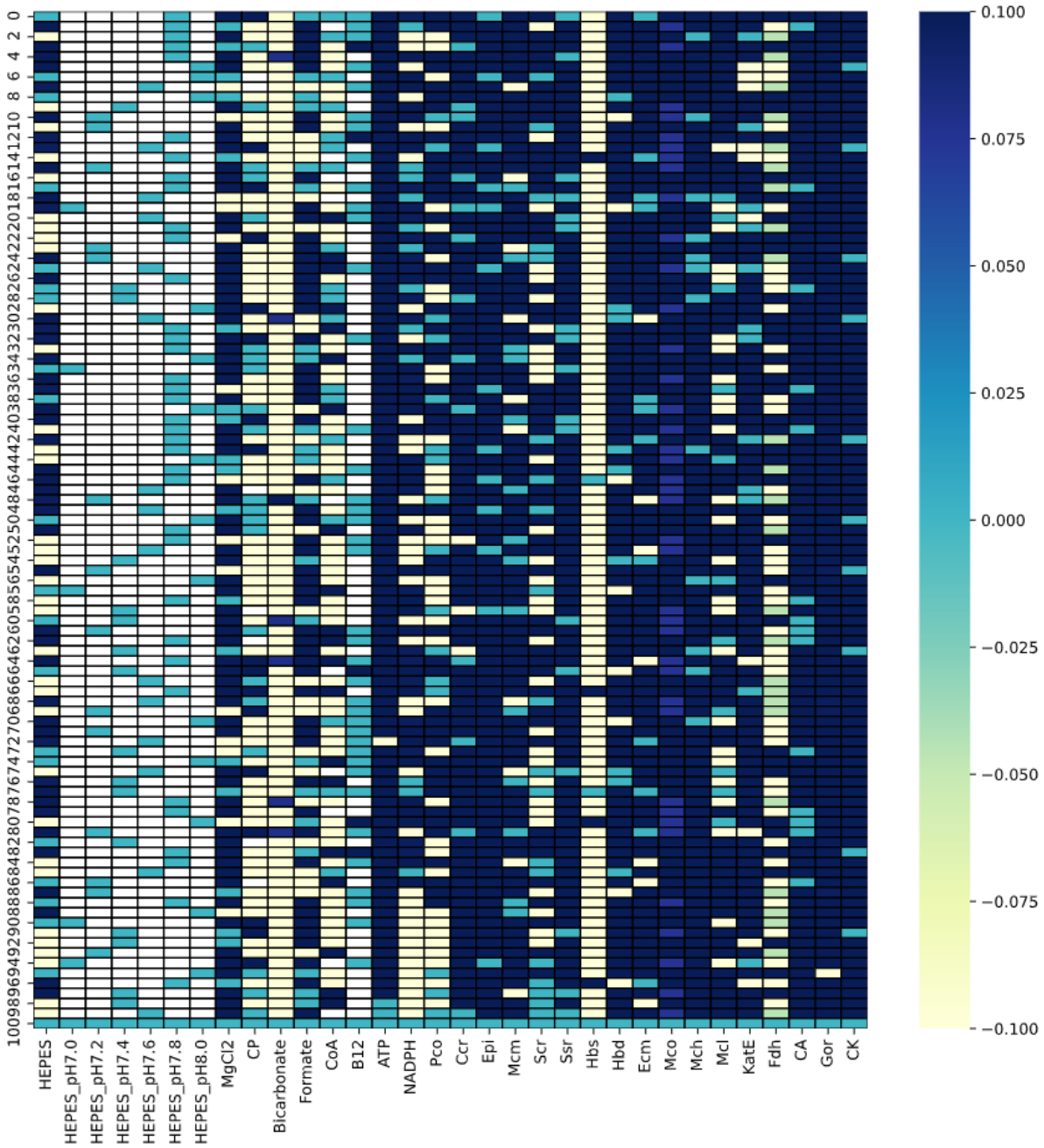
Supplementary Fig. 10: Analysis of the active learning data for the transcription and translation unit. (a) Distribution of alternative factors within the yield of 200 combinations. **(b)** The feature importance percentage of each condition for the model to assign predicted yield values. The bars and error bars are the average and standard deviation of the importances in 4 days



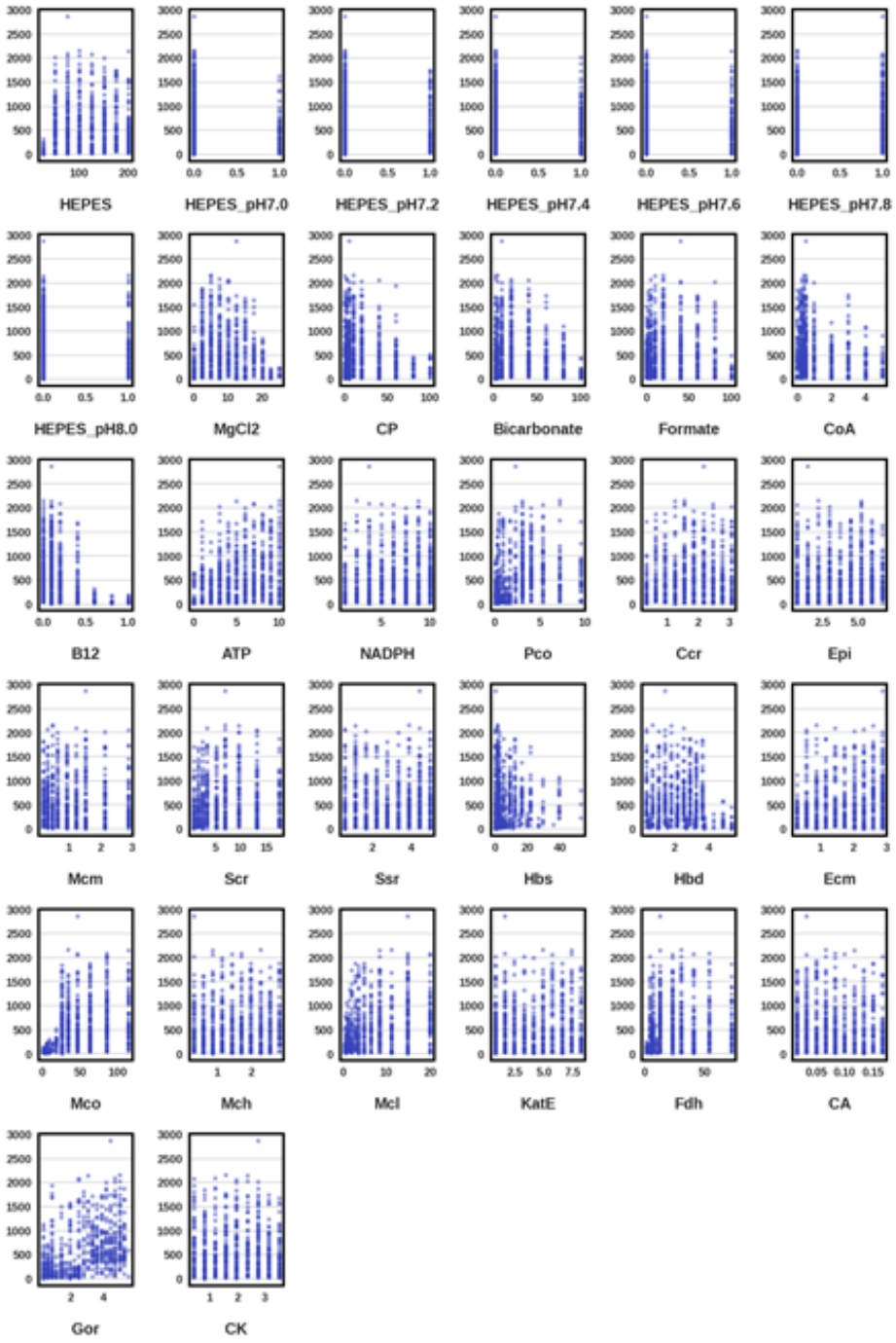
Supplementary Fig. 11: Cell-free versus *in vivo* yields of the 20 most informative combinations for the transcription & translation unit. The bar plot representation of results (average and standard deviation of triplicates) shown in **Fig. 4e**.



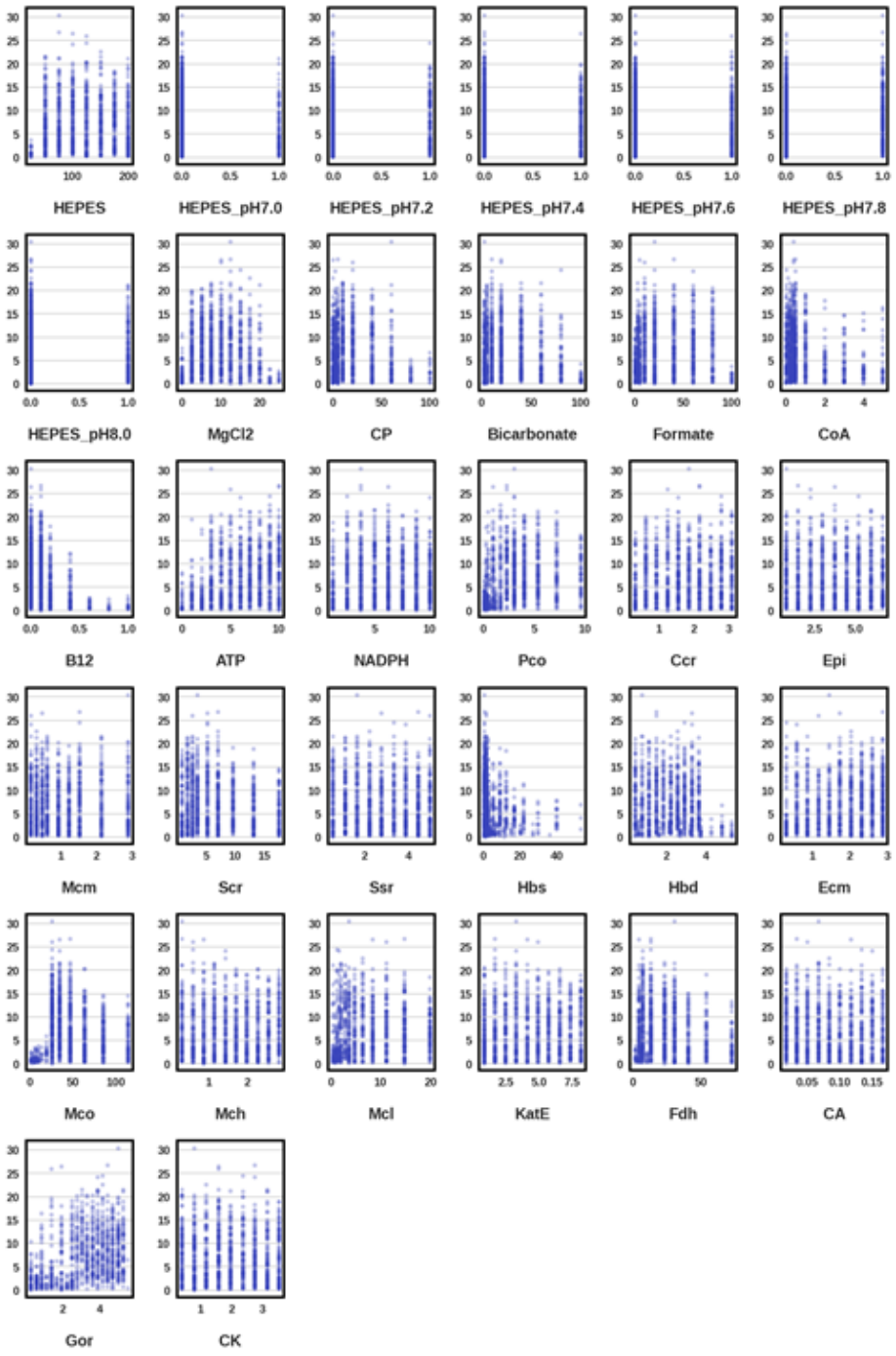
Supplementary Fig. 12: Scaled factors heatmap of top 10% yields in CETCH yield (glycolate, round 1-5) active learning. The concentration range for factors was scaled to the same range.



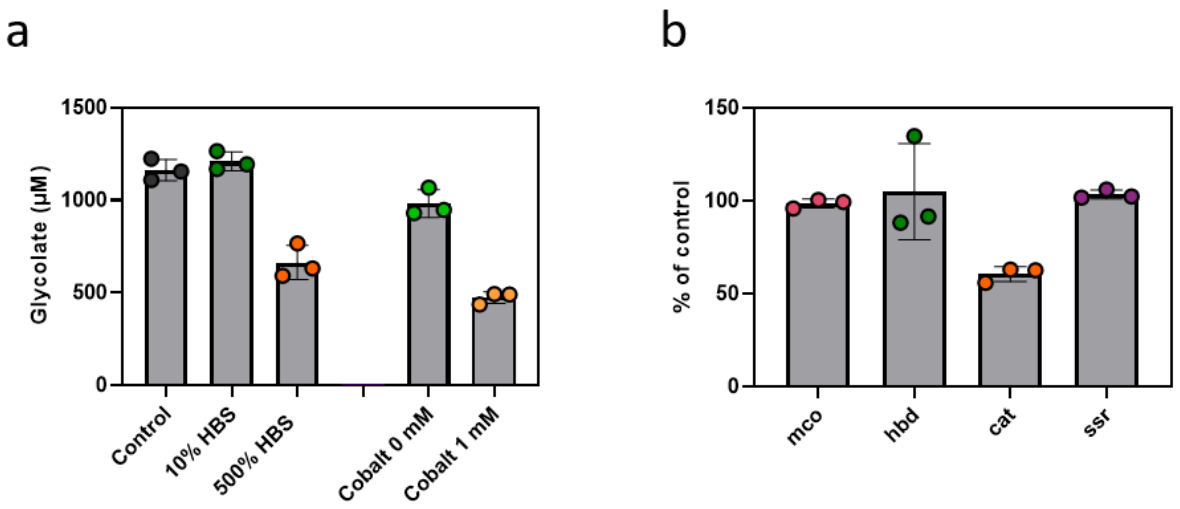
Supplementary Fig. 13: Scaled factors heatmap of top 10% yields in CETCH efficiency (glycolate concentration divided by the total enzyme concentration, round 1-8) active learning. The concentration range for factors was scaled to the same range.



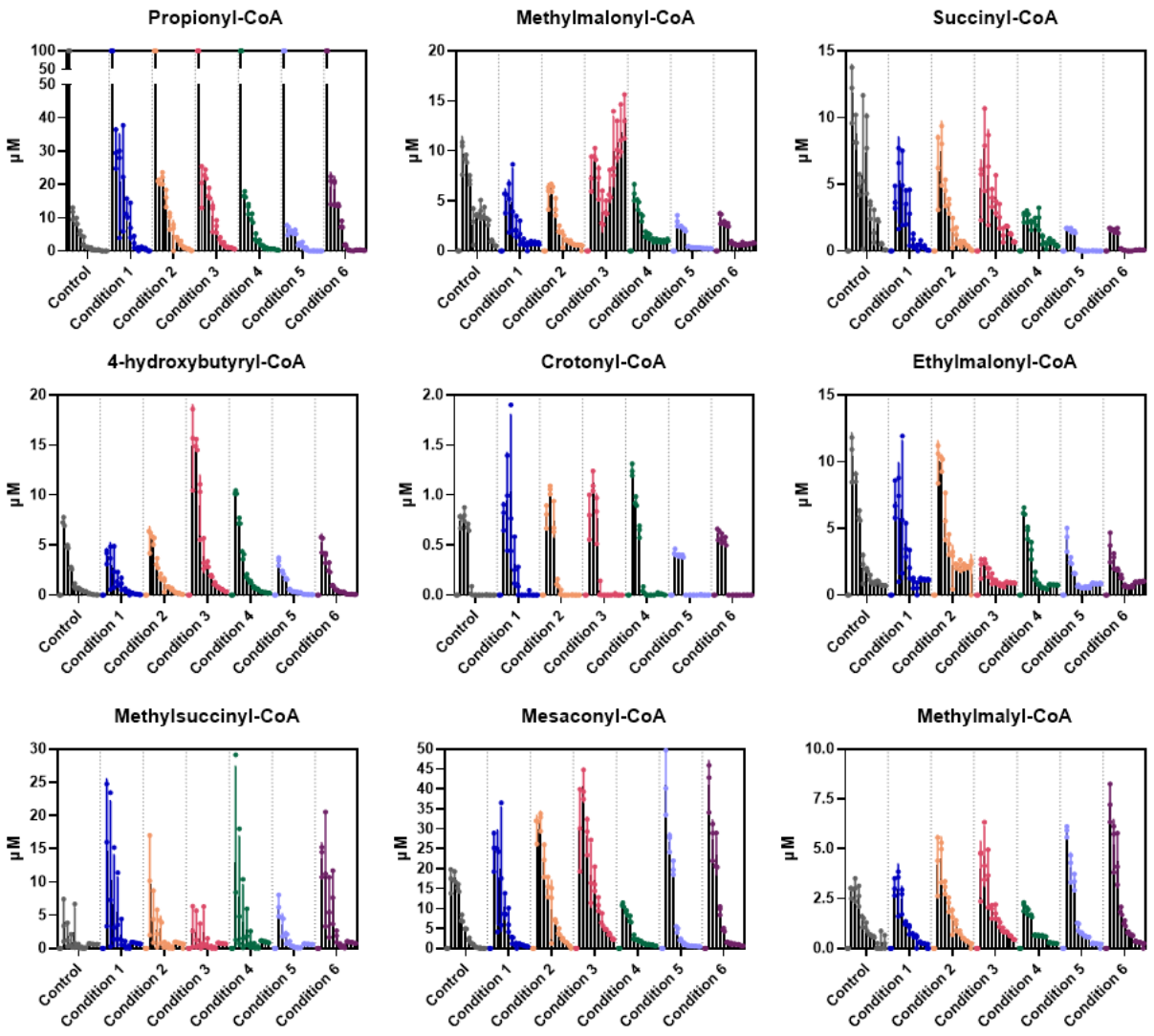
Supplementary Fig. 14: Distribution of measured yield (glycolate, round 1-5) values within the ranges of each factor. Distribution of all factors within the yield of 5 rounds.



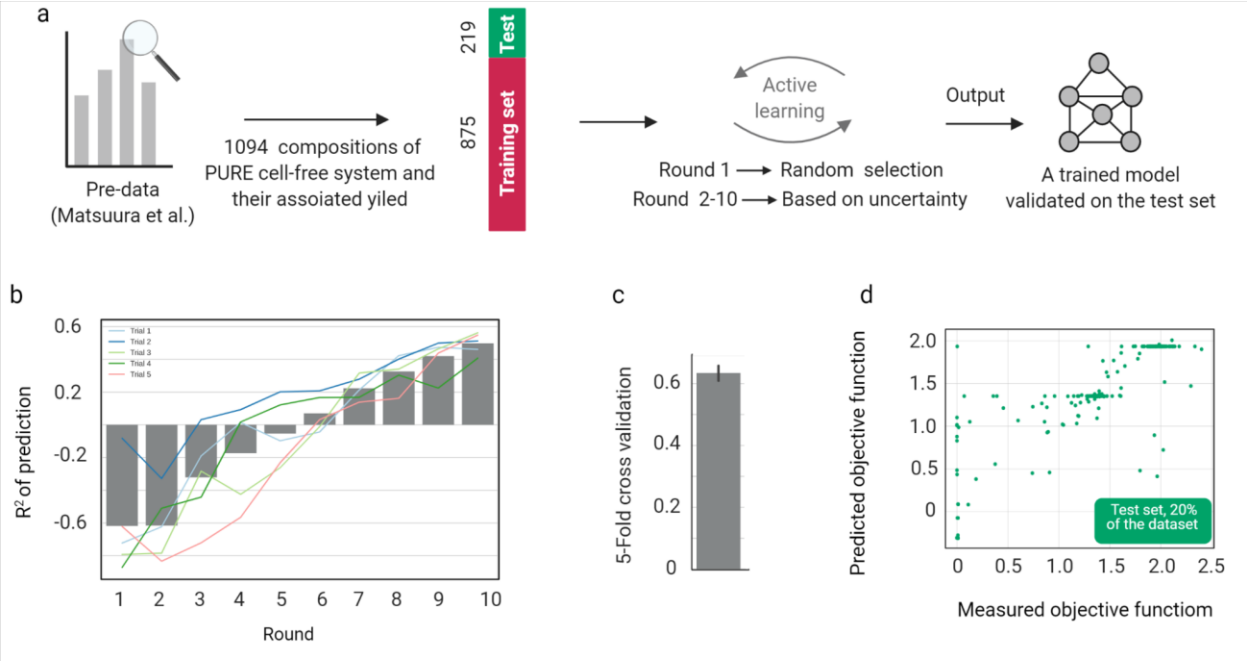
Supplementary Fig 15: Distribution of measured efficiency (glycolate concentration divided by the total enzyme concentration, round 1-8) values within the ranges of each factor. Distribution of all factors within the efficiency of 8 rounds.



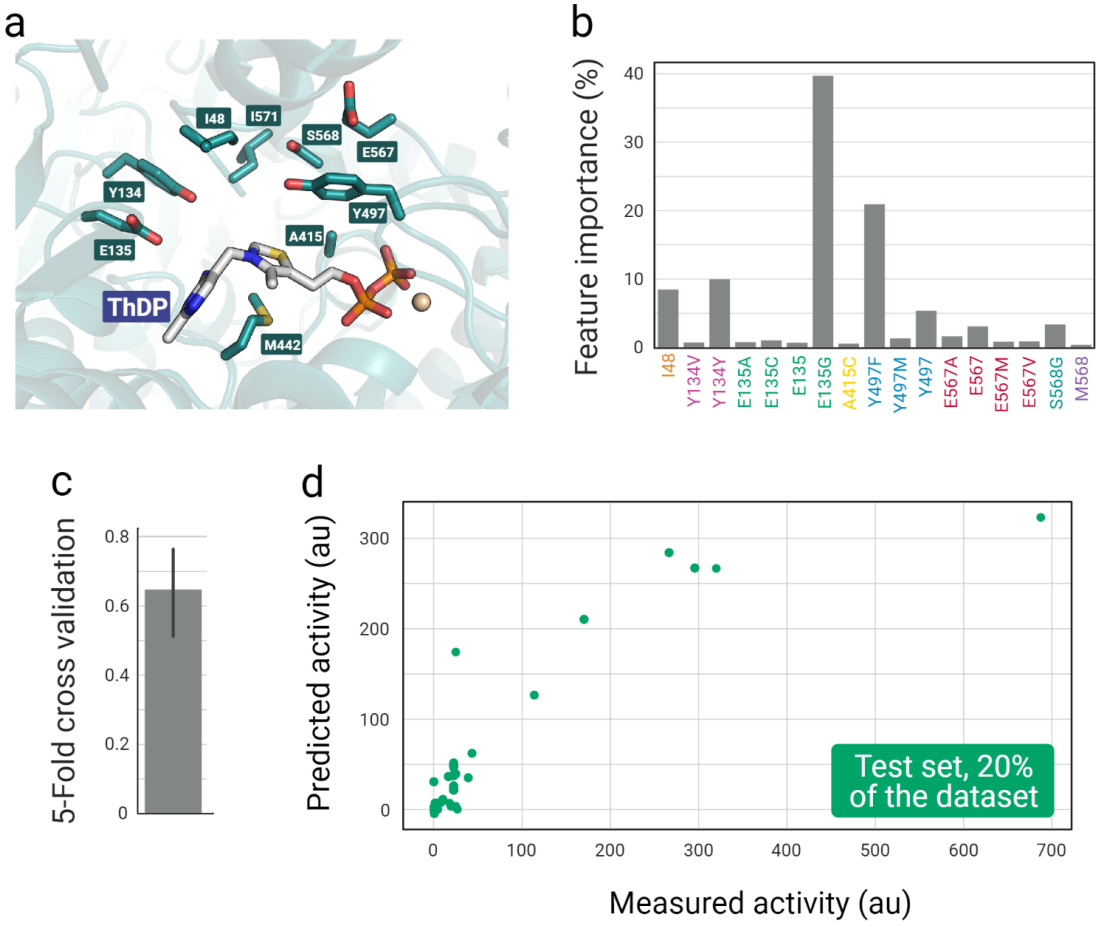
Supplementary Fig. 16: Assays for testing newly purified enzyme batches after round 2 (b) and for testing different concentrations of hbs and cobalt (a). (a) The assays were done as described in material and methods (Assays for different concentrations of hbs and cobalt). The shown glycolate concentrations are from the sample taken after 180 min. The control contained 0.1 mM Coenzyme B_{12}, whereas the assays with 0 mM and 1 mM cobalt did not contain any Coenzyme B_{12}. **(b)** Test of old batches of mco, hbd, cat and ssr. After recognizing an increase in the product yield of our control after round two, we tested the old stocks of the enzymes which were purified freshly for round three in combination with the new enzymes. We could identify the old batch of catalase as the reason for the lower yield. The values of the grey bars are the means of three replicates (dots) and the error bars represent the standard deviations.



Supplementary Fig. 17: Quantified CoA-esters from selected assay conditions. The values for time point 0 are not measured, the reactions were started with 100 µM propionyl-CoA. The values of the black bars are the means of three replicates (dots) and the error bars represent the standard deviations. For details see LC-MS analysis of CoA esters.



Supplementary Fig. 18: Application of METIS for the prediction of a PURE cell-free system dataset. (a) We took this dataset from a recent study in which they varied recombinant proteins and buffer compositions, adapted it as a Results.csv file. The dataset was divided into two subsets 80% for training and 20% for testing. (b) During 10 rounds of active learning, the model was assessed by the initial test set and the R^2 of the prediction improved over rounds. See **Supplementary Note 6** for a detailed explanation of the process. **(c)** 5-fold cross-validation to evaluate the average performance of the model on the whole dataset. **(d)** a single round of validation of a model on the test set.



Supplementary Fig. 19: Simulations on a dataset of 847 mutants of oxalyl-CoA decarboxylase enzyme to showcase the application of METIS workflow for combinatorial enzyme engineering. (a) The active site of oxalyl-CoA decarboxylase with surrounding residues annotated and the covalently bound cofactor thiamine diphosphate (ThDP). **(b)** Calculated features' importance by the model on the imported dataset that define how big the effect of each condition is if the model suggests new amino acid combinations. **(c)** 5-fold cross-validation, a common approach to assess the prediction power of machine learning models. **(d)** Predicted values on a single unseen test set (20% of the whole dataset) versus their measured values. (see also **Supplementary Note 7**)

Supplementary Fig. 20: Distribution of amino acids in each position of the oxalyl-CoA decarboxylase active site predicted by the model out of 50 suggestions. This shows an example in case a user aims to synthesize or clone new enzyme sequences and test a round of combinatorial mutants.

2.7.4. Supplementary References

1. Sommer, C. & Gerlich, D.W. Machine learning in cell biology - teaching computers to recognize phenotypes. *J Cell Sci* **126**, 5529-39 (2013).
2. Borkowski, O. et al. Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun* **11**, 1872 (2020).
3. Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Molecular systems biology* **5**, 297-297 (2009).
4. Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687-694 (2019).
5. Wittmann, B.J., Johnston, K.E., Wu, Z. & Arnold, F.H. Advances in machine learning for directed evolution. *Curr Opin Struct Biol* **69**, 11-18 (2021).
6. Wittmann, B.J., Yue, Y. & Arnold, F.H. Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden. *bioRxiv*, 2020.12.04.408955 (2020).
7. Nattermann, M. et al. Engineering a Highly Efficient Carboligase for Synthetic One-Carbon Metabolism. *ACS Catal* **11**, 5396-5404 (2021).
8. Sun, Z.Z. et al. Protocols for implementing an Escherichia coli based TX-TL cell-free expression system for synthetic biology. *J Vis Exp*, e50762 (2013).
9. Ribosome Binding Sites/Prokaryotic/Constitutive/Community Collection. https://parts.igem.org/Ribosome_Binding_Sites/Prokaryotic/Constitutive/Community_Collection.
10. Greco, F.V., Pandi, A., Erb, T.J., Grierson, C.S. & Gorochowski, T.E. Harnessing the central dogma for stringent multi-level control of gene expression. *Nat Commun* **12**, 1738 (2021).

3. A modular in vitro platform for the production of terpenes and polyketides from CO_2

Srividhya Sundaram[1], Christoph Diehl[1], Niña Socorro Cortina[1,2], Nicole Paczia[3], Jan Bamberger[4] and Tobias J. Erb[1,5]

[1] Department of Biochemistry & Synthetic Metabolism, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[2] LiVeritas Biosciences, Inc., South San Francisco, USA

[3] Core Facility for Metabolomics and Small Molecule Mass Spectrometry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[4] Equipment Center for Mass Spectrometry and Elemental Analysis, Department of Chemistry, Philipps-University Marburg, Germany

[5] SYNMIKRO Center of Synthetic Microbiology, Marburg, Germany

Author contributions

T.J.E., S.S. and C.D. conceived the work. S.S. and C.D. designed and performed the experiments. N.S.C., N.P. and C.D. developed the LC-MS methods for the analysis of glycolate, malate and CoA esters. J.B. developed the GC-MS methods for the analysis of the terpenes and polyketides. T.J.E. supervised and directed the work. **S.S., C.D. and T.J.E. wrote the manuscript with input from all other authors.**

3.1. Abstract

A long-term goal in realizing a sustainable biocatalysis and organic synthesis is the direct use of the greenhouse gas CO_2 as feedstock for the production of bulk and fine chemicals, such as pharmaceuticals, fragrances and food-additives. Here we developed a modular in vitro platform for the continuous conversion of CO_2 into complex multi-carbon compounds, such as monoterpenes (C10), sesquiterpenes (C15) and polyketides. Combining natural and synthetic metabolic pathway modules, we established a route from CO_2 into the key intermediates acetyl- and malonyl-CoA, which can be subsequently diversified through the action of different terpene and polyketide synthases. Our proof-of-principle study demonstrates the simultaneous operation of different metabolic modules comprising of up to 29 enzymes in one pot, which opens the way for developing and optimizing synthesis routes for the generation of complex CO_2-based chemicals in the future.

3.2. Introduction

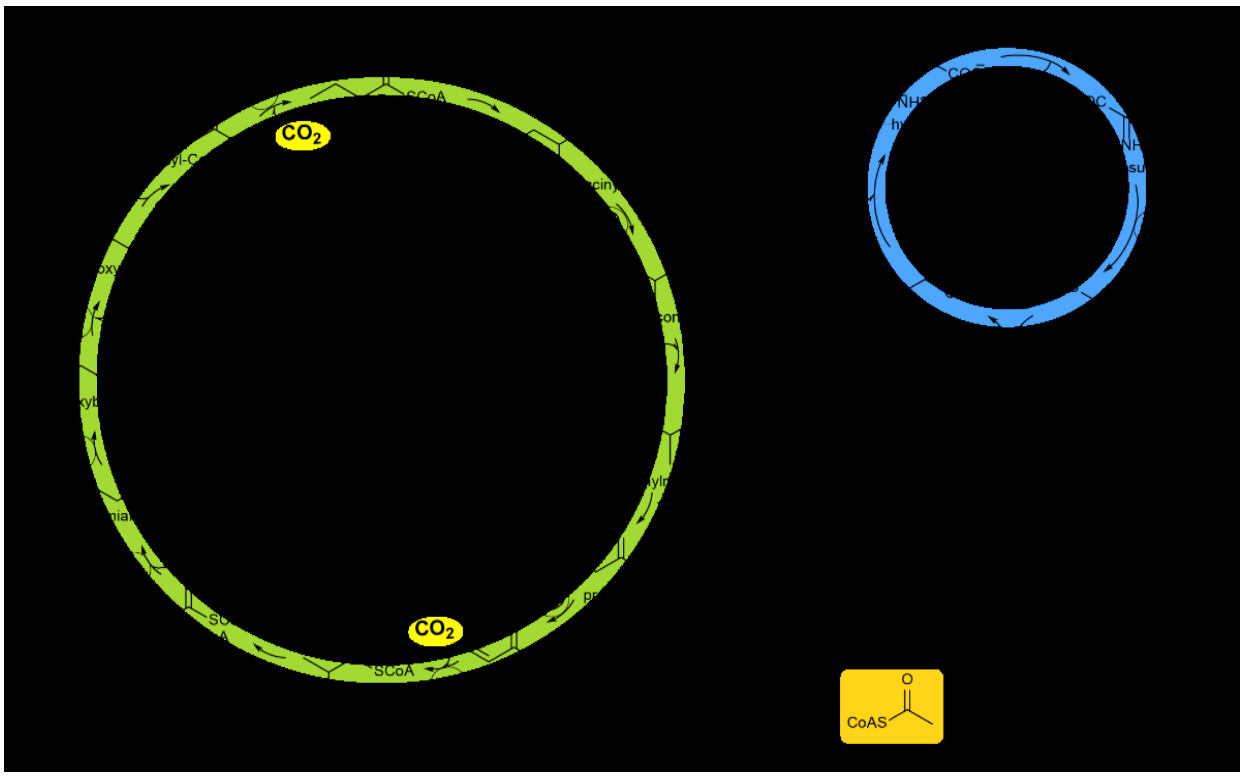
Cell-free synthetic biology involves the *in vitro* assembly of multiple purified and semi-purified enzymes into metabolic cascades to generate natural and new-to-nature high value chemicals. The *in vitro* reconstitution of metabolic pathways allows the convenient manipulation and optimization of reaction conditions, enzyme concentrations, cofactor supply, energy flux and yield, which are difficult to control in living microorganisms[1]. Glucose is frequently used as an inexpensive feedstock for these complex systems. Over the recent years, the use of other carbon sources, such as sucrose, cellulose, glycerol, xylose and starch was demonstrated[2]. However, the direct conversion of atmospheric CO_2 or other C1 precursors into value-added compounds has proven a major challenge for *in vitro* systems.

To address this challenge, the synthetic crotonyl-coenzyme (CoA)/ethylmalonyl-CoA/hydroxybutyryl-CoA cycle (CETCH) was developed. This *in vitro* pathway combines 18 enzymes from nine different organisms to generate the C2-compound glyoxylate from CO_2 at a rate of 5 nmoles per minute per mg protein[3]. Very recently, CETCH was successfully coupled to photosynthetic membranes for the light-driven conversion of CO_2 into glycolate[4]. However, a successful coupling of CETCH to downstream anabolic pathway modules that would allow to extend its product spectrum beyond glyoxylate or glycolate is still lacking. One interesting set of target molecules are natural products, in particular terpenes and polyketides, which are used as flavours, pharmaceuticals, biofuels and commodity chemicals. These complex compounds are synthesized *in vivo* from the simple C2-building blocks like acetyl-CoA by individual enzymes (terpene synthases) or multi-enzyme complexes (polyketide synthases, PKSs) respectively[5-7].

Here we developed a multi-modular *in vitro* platform to access different terpenes and polyketides directly from CO_2. To that end, we first coupled the synthetic CETCH with a natural glyoxylate assimilation module to convert CO_2 into acetyl-CoA. We further demonstrate how acetyl-CoA can be diversified into an array of terpenes and polyketides through downstream processing by different terpene and PKS biosynthetic modules. Overall, this proof-of-principle study might pave the way towards realizing modular, multi-enzyme reaction cascades for the sustainable synthesis of complex chemicals from simple C1 building blocks, such as CO_2 in the future.

3.3. Results

To capture CO_2 into glyoxylate, we first established CETCH and determined its productivity in our experimental setup. To that end, we run CETCH version 5.4 (**Supplementary text**) and quantified glycolate production by adding glyoxylate reductase to the CETCH core cycle. Starting from 100 µM propionyl-CoA, CETCH produced ~730 µM glycolate within 3 hours under the chosen conditions (**Figure 1A**).



Scheme 1. Coupling of CETCH-BHAC modules for acetyl-CoA formation

Next, we aimed at establishing a coupling module for the further conversion of glyoxylate into acetyl-CoA, which would allow to couple CETCH with downstream terpene/polyketide-producing pathways. We sought to employ the β-hydroxyaspartate cycle (BHAC), a pathway used by marine proteobacteria for glyoxlate assimilation[8,9]. The BHAC converts two molecules of glyoxylate into oxaloacetate via four enzymes, requiring only one molecule of NADH and one amino group that is constantly recycled during the process, making the BHAC the most efficient reaction sequence for the conversion of C2 molecules into C4 compounds described to date. Oxaloacetate can then be further converted into acetyl-CoA via malate dehydrogenase, malate thiokinase and malyl-CoA lyase (**Scheme 1**). We reconstituted the BHAC *in vitro* using N-terminal His-tagged proteins, produced in *Escherichia coli* (**Table S5**). To optimize BHAC productivity, we tested different concentrations of transaminase BhcA and co-substrate glycine, using malate dehydrogenase as readout (see **Supplementary text**). However, starting from 500 µM glyoxylate, malate yields were comparable between the different conditions tested (~70%; **Figure S1A**), indicating that the BHAC was operating robustly *in vitro* (**Figure 1B**). Next, we coupled the BHAC with CETCH. When we added the enzymes of the BHAC after 60 min to the CETCH assay, CETCH plus BHAC yielded ~200 µM acetyl-CoA, corresponding to a conversion of glyoxylate into acetyl-CoA at 30% yield (**Figure S1B**).

Notably, we achieved similar acetyl-CoA yields, when we coupled the CETCH with BHAC directly from the beginning (**Figure 1C and S1B**), indicating that the 18 enzymes of the CETCH and BHAC can be operated simultaneously in one pot. CO_2 fixation was also confirmed by isotopic labeling as before. Using ^{13}C-labeled bicarbonate and ^{13}C-formate (released as $^{13}CO_2$ during cofactor recycling), fully labeled malate was observed after two hours, proving that CETCH turned multiple times (**Figure 1D**). Considering that already single ^{13}C-labeled malate is stoichiometrically exclusively derived from fixed CO_2 (**Figure S2**), these experiments demonstrated that CO_2 can be continuously converted into acetyl-CoA by directly coupling CETCH and BHAC.

In the last step in the CETCH-BHAC cascade, the cleavage of malyl-CoA into glyoxylate and acetyl-CoA by malyl-CoA lyase, is reversible with a $\Delta G^{0'}$ of -3 ± 5.8 kJ mol^{-1}. To test, whether this reaction runs into an equilibrium, we determined malyl-CoA concentrations after 90 min. Much to our surprise, the concentration of malyl-CoA was below 1 µM (**Figure S3A**), indicating that this compound is specifically degraded over time in the reaction mixture. Indeed, when we incubated both acetyl-CoA and malyl-CoA in the assay matrix in the absence and presence of all CETCH and BHAC enzymes (with exception of malyl-CoA lyase Mcl), acetyl-CoA appeared stable, while malyl-CoA was consumed by one or more of the enzymes (**Figure S3B and S3C**). Thus, when coupling CETCH and BHAC, the last reaction in the cascade

reaches an equilibrium between acetyl-CoA plus glyoxylate and malyl-CoA. Malyl-CoA will be degraded over time, thereby limiting total yield, unless the flux is further pulled into downstream reactions that consume acetyl-CoA (see below).

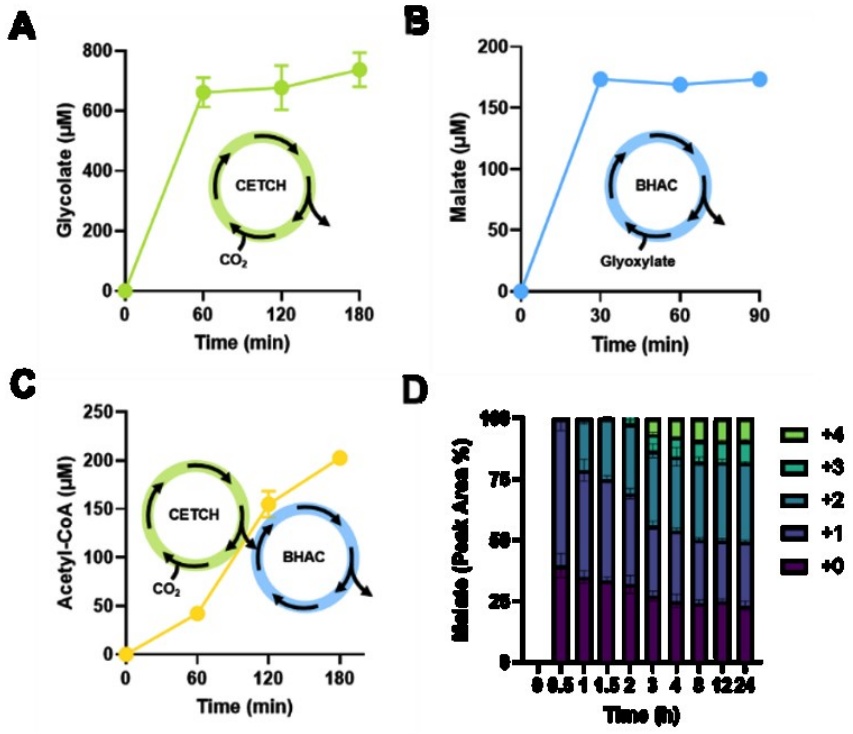


Figure 1. Coupling of CETCH with the BHAC for acetyl-CoA production. **A)** Glycolate production by CETCH from 100 µM propionyl-CoA. Glyoxylate reductase was used to convert the primary product glyoxylate to glycolate. **B)** Malate production by BHAC from 500 µM glyoxylate. Malate dehydrogenase was used to convert oxaloacetate to malate. C) Coupling of CETCH with the BHAC to produce acetyl-CoA (see **Scheme 1**), from 100 µM propionyl-CoA. **D)** Fractional labeling of malate from CETCH-BHAC coupling and Mdh using 13C-labeled bicarbonate and 100 µM propionyl-CoA. The 13C is incorporated as CO2 by the Ccr as shown in Scheme 1. +0, +1, +2, +3, +4 indicates the number of carbons of the malate (C4) derived from 13CO2 incorporation. The reactions were performed in triplicates and the mean ± S.D. are plotted.

For the further conversion of acetyl-CoA into terpenes, we aimed at coupling the 18 enzymes of the CETCH-BHAC cascade with different terpene biosynthetic modules, comprising of the nine enzymes of the mevalonate biosynthetic pathway and various terpene synthases (**Figure 2A, Table S5**). We established five different terpene biosynthetic modules, by prototyping the production of monoterpenes (C10) limonene (**1**), sabinene (**2**) and α-pinene (**3**), as well as sesquiterpenes (C15) α-bisabolene **4**) and β-farnesene (**5**) from acetyl-CoA in the CETCH-BHAC assay matrix (**Table S5**). To constantly supply cofactors, we employed the regeneration systems used in CETCH. To regenerate the NAD(P)H pool, we used an engineered formate dehydrogenase (Fdh) that accepts both, NADPH and NADH[10]; to maintain the ATP pool, we used a polyphosphate transferase system[11]. The extraction of terpenes was optimized by testing different solvents (**Figure S4A & S4B**). Production of **1-5** from optimized by testing different terpene

synthase concentrations (**Figure S5**) acetyl-CoA was validated with authentic standards and further optimized by testing different terpene synthase concentrations (**Figure S5**).

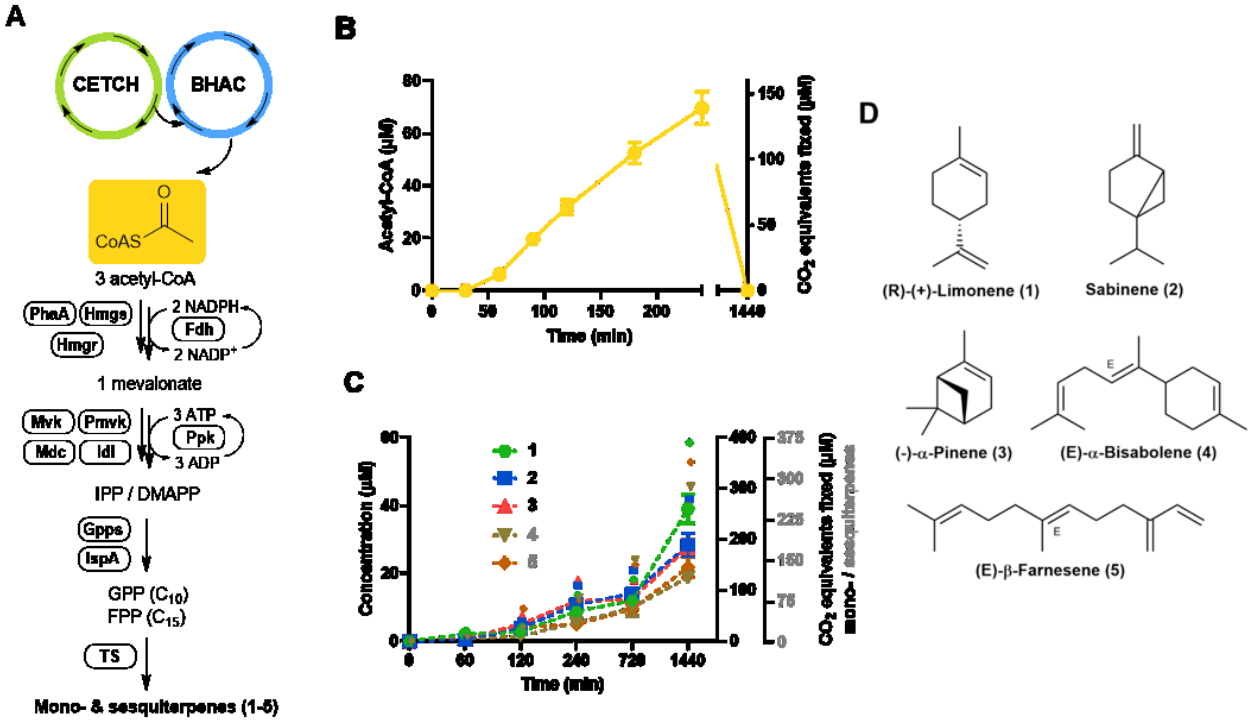


Figure 2. Coupling of CETCH-BHAC cycles for terpene biosynthesis. **A)** General scheme of the mevalonate pathway. The cofactors (NADPH and ATP) are fluxed from the CETCH/BHAC cycles and are constantly recycled. IPP: Isopentenyl pyrophosphate; DMAPP: Dimethylallyl pyrophosphate; TS: terpene synthase. **B)** Formation of acetyl-CoA from CETCH over 24 h. The reaction is started with 100 µM propionyl-CoA and analysed as described in the methods. **C)** Time course of terpenes production from 100 µM propionyl-CoA. Colored labels correspond to the amount of CO2 fixed over time. The analytes are measured by GCMS as descried in the methods. The reactions were performed in triplicates and the mean ± S.D. are plotted. **D)** Structures of the mono- and sesquiterpenes produced.

When we pre-produced acetyl-CoA with the CETCH-BHAC cascade for 4 h, before adding the different terpene biosynthetic modules, monoterpenes **1-3** and sesquiterpenes **4-5** were produced at concentrations of ~10 µM and ~5 µM, respectively (**Figure S6B**). However, when operating the CETCH-BHAC cascade with the different terpene biosynthetic modules simultaneously in one pot, product yields were increased three- to four-fold (**Figure S6C**). The CETCH-BHAC cascade alone produced about 70 µM acetyl-CoA within 4 h (**Figure 2B**). We obtained monoterpenes **1-3** between concentrations of 30 to 40 µM, and sesquiterpenes **4** and **5** at 20 µM (**Figure 2C, 2D and Table 1**), supporting the hypothesis that the direct downstream conversion of acetyl-CoA is crucial to improve product yield. LC-MS analysis of the reaction mixtures confirmed presence of different mevalonate pathway intermediates, (**Figure S7B**), but only trace amounts of residual acetyl-CoA (**Figure S7C**), indicating that acetyl-CoA is efficiently fed into

the different downstream terpene biosynthetic modules. As a positive control, the production of **1-5** was also validated from 0.5 mM acetyl-CoA (**Figure S7D**).

Compounds	Yield (µM)	CO_2 fixed (µM)	Productivity (mg l^{-1} h^{-1})
1	39 ± 4	390	0.22
2	28 ± 3	283	0.16
3	27 ± 2	279	0.16
4	19 ± 2	283	0.19
5	22 ± 3	327	0.22

Table 1. Net productivity of terpenes from 100 µM propionyl-CoA in 24 h. Yield was determined by GCMS using authentic standards (**Figure S8**). Data represent n=3 ± S.D. The assay contained 4.6 mg/ml of total enzymes.

To test whether the CETCH-BHAC cascade would also fuel polyketide biosynthesis, we attempted to couple it with the iterative PKS C-1027 (PKS_{SgcE})[12]. PKS_{SgcE} has been reported to form 1,3,5,7,9,11,13-pentadecaheptaene (PDH, **7**), an all-*trans* polyene, which upon chemical hydrogenation leads to pentadecane (PD), a prime component of diesel fuel. PKS_{SgcE} uses one acetyl-CoA, eight malonyl-CoA and seven NADPH to generate a nine-membered enediyne precursor (**Figure 3A**). The PKS undergoes eight iterative cycles, during which the dehydratase (DH) domain remains inactive in the last two cycles and the keto-reductase (KR) domain in the ultimate cycle.

Products are released from PKS_{SgcE} either via spontaneous lactonisation yielding **6,** or through hydrolysis by the type II-standalone thioesterase [12,13] TE_{SgcE}, yielding **7**. A recent study reported that production of **7** depends on the ratio of PKS_{SgcE}:TE_{SgcE}[14]. We protoyped the *in vitro* production of **6** and **7** by mixing various concentrations of PKS_{SgcE} and TE_{SgcE} with acetyl-CoA, malonyl-CoA and NADPH at 30 ˚C. Production of **6** was directly confirmed from the reaction mixture with high resolution LCMS (*m/z* $[M + H]^+$ = 285.1485). Production of **7** was confirmed in ethyl acetate extracts of the reaction mixture by UV-Vis spectroscopy (absorption maxima at 336, 355, 373, 395 nm) and high resolution LCMS (*m/z* $[M + H]^+$ = 199.1476) (**Figure S10A**). The amount of **7** increased with increasing TE_{SgcE} concentrations with a maximum production at 2.5 µM PKS_{SgcE} and 40 µM TE_{SgcE} (**Figure S10B**). Neither **6** nor **7** were detected, when we tested PKS_{SgcE} mutant C171A, in which the KS was inactivated (**Figure S10C**), demonstrating successful reconstitution of PKS_{SgcE} *in vitro.*

Note that CETCH features methylmalonyl-CoA and ethylmalonyl-CoA as intermediates, which serve as extender units in the biosynthesis of several polyketides and might pose a problem when directly coupling CETCH with PKS. To study whether PKS_{SgcE} would accept methyl- and ethylmalonyl-CoA besides malonyl-CoA we tested these compounds separately and in combination with purified PKS_{SgcE} and

analysed the reaction mixture for the production methyl- (**8**) and ethyl- (**9**) substituted heptaenes with high resolution LCMS. Indeed, **8** and **9** were produced in a TE-dependent fashion when the corresponding precursors were available (**Figure S10D, I, II, III**). However, suggesting that operating CETCH and PKS_{SgcE} simultaneously does not pose a challenge, as long as a sufficient pool of malonyl-CoA is present.

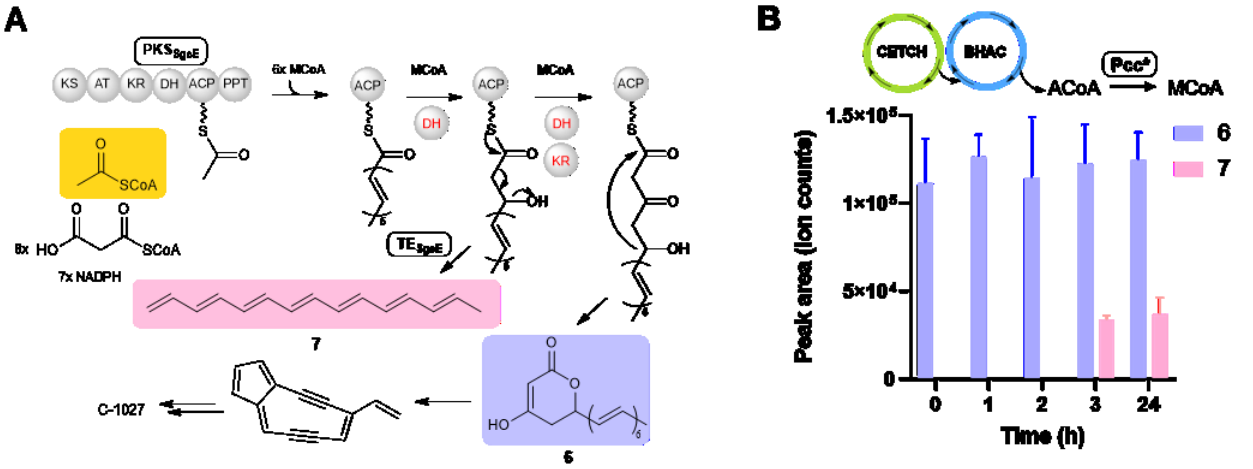


Figure 3. Coupling of CETCH-BHAC cycles for pentadecaheptaene biosynthesis. **A)** Proposed catalysis by PKSSgcE and TESgcE in the biosynthesis of enediyne antibiotic C-1027. KS: ketosynthase; AT: acyltransferase; KR: ketoreductase; DH: dehydratase; ACP: acyl carrier protein; PPT: phosphopantethenyl transferase. **B)** Formation of 7 via CETCH-BHAC from 100µM propionyl-CoA at different time points. The production of 7 stalls at 3 h and remains the same until 24 h. All the assays were performed in triplicates and the mean ± S.D. are plotted.

Having this information at hand, we finally directly coupled the CETCH-BHAC cascade with 2.5 µM PKS_{SgcE} and 40 µM TE_{SgcE}, and added 2 µM propionyl-CoA carboxylase variant D407I (Pcc*) that shows 10% activity with acetyl-CoA to provide the extender unit malonyl-CoA from acetyl-CoA. In the coupled system, **6** was produced in relatively high amounts independent of TE_{SgcE}, while TE_{SgcE}-dependent production of **7** reached a maximum around 3 h, demonstrating the successful biosynthesis of complex polyketides with our coupled system (**Figure 3B**).

3.4. Discussion

In conclusion, we show that more complex molecules such as terpenes and polyketides can be built exclusively from CO_2 combining the synthetic CO_2-fixing CETCH with different biosynthetic modules. Although cofactor regeneration in our modular platform was based on formate, we note that this C1 compound can be regenerated electrochemically and/or enzymatically from CO_2 and thus provide a carbon neutral energy (and carbon) source[15,16]. Moreover, CETCH was recently coupled with chloroplast extracts[4]. Energizing our modular platform with photosynthetic membranes could make our multi-enzyme system completely independent of chemical energy in the future.

The net productivity of terpenes from CO_2 reached with our modular platform is currently ~0.2 mg l^{-1} h^{-1}. This is lower compared to glucose-based cell-free protein synthesis (CFPS) and other *in vitro* production systems, which range between 2-100 mg l^{-1} h^{-1} [17-19]. *In vivo*, **5** has been produced up to 2 g l^{-1} h^{-1} in *S. cerevisiae* by combining an artificial acetyl-coA biosynthetic pathway with the NADH-preferring Hmgr[20]. The maximum titres reported for **2** and **3** are 15 mg l^{-1} h^{-1}, 100 mg l^{-1} h^{-1}, and 1.2 mg l^{-1} h^{-1} in *E.coli* under fed-batch or shake flask fermentations[21-23], whereas the production of **4** reached 13 mg l^{-1} h^{-1} in both *E.coli* and *S. cerevisiae*[24]. However, it should be noted that these production rates are based on the direct supply of multi-carbon compounds and were achieved only after rigorous optimization of the different pathways both *in vivo* and *in vitro*.

It is conceivable that the productivity of our *in vitro* system can also be enhanced further by optimising enzyme concentrations, activity and stability (e.g., through immobilisation or the use of thermostable enzyme variants). For example, the mevalonate module itself has already been demonstrated to be self-sustaining over a long period of time (~7 days). Using modelling approaches or computer-aided design-build-test cycles focusing on identifying optimal enzyme stochiometries, inter-mediate concentrations and critical reaction parameters could further increase production rates of our *in vitro* system. Moreover, the product portfolio of our platform can be further expanded. Using natural, engineered and chimeric terpene synthases or PKSs will allow to access compounds that are not known from to traditional synthetic chemistry or biology so far. As a case example we demonstrated that PKS_{SgcE} can be used to produce the natural polyene PDH (**7**), but that the enzyme is in principle also able to synthesize so far unknown multi-branched polyenes (**8**, **9**), if our modular platform was expanded to provide methyl-, and or ethylmalonyl-CoA precursors (instead of malonyl-CoA) from CO_2. Finally, protoyping and optimizing complex reaction networks *in vitro* might provide important information for the successful implementation of these pathways *in vivo* to create novel production strains for the synthesis of complex multi-carbon compounds from CO_2 in the future[25,26].

3.5. References

1. Bowie, J.U. et al. Synthetic Biochemistry: The Bio-inspired Cell-Free Approach to Commodity Chemical Production. *Trends Biotechnol* **38**, 766-778 (2020).
2. Taniguchi, H., Okano, K. & Honda, K. Modules for *in vitro* metabolic engineering: pathway assembly for bio-based production of value-added chemicals. *Synth. Syst. Biotechnol.* **2**, 65-74 (2017).
3. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
4. Miller, T.E. et al. Light-powered CO2 fixation in a chloroplast mimic with natural and synthetic parts. *Science* **368**, 649-654 (2020).
5. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl* **48**, 4688-716 (2009).
6. Weissman, K.J. Introduction to polyketide biosynthesis. *Methods Enzymol.* **459**, 3-16 (2009).
7. Oldfield, E. & Lin, F.Y. Terpene biosynthesis: modularity rules. *Angew Chem Int Ed Engl* **51**, 1124-37 (2012).
8. Schada von Borzyskowski, L. et al. Marine Proteobacteria metabolize glycolate via the beta-hydroxyaspartate cycle. *Nature* **575**, 500-504 (2019).
9. Kornberg, H. & Morris, J. β-hydroxyaspartate pathway: A new route for biosyntheses from glyoxylate. *Nature* **197**, 456-457 (1963).
10. Galkin, A., Kulakova, L., Tishkov, V., Esaki, N. & Soda, K. Cloning of formate dehydrogenase gene from a methanol-utilizing bacterium *Mycobacterium vaccae* N10. *Appl. Microbiol. Biotechnol.* **44**, 479-483 (1995).
11. Nocek, B. et al. Polyphosphate-dependent synthesis of ATP and ADP by the family-2 polyphosphate kinases in bacteria. *Proc Natl Acad Sci U S A* **105**, 17730-5 (2008).
12. Zhang, J. et al. A phosphopantetheinylating polyketide synthase producing a linear polyene to initiate enediyne antitumor antibiotic biosynthesis. *Proc. Natl. Acad. Sci. USA* **105**, 1460-1465 (2008).
13. WaiáLiew, C. Products of the iterative polyketide synthases in 9-and 10-membered enediyne biosynthesis. *Chem. Comm.*, 7399-7401 (2009).
14. Liu, Q. et al. Engineering an iterative polyketide pathway in *Escherichia coli* results in single-form alkene and alkane overproduction. *Met. Engg.* **28**, 82-90 (2015).
15. Schuchmann, K. & Müller, V. Direct and reversible hydrogenation of CO_2 to formate by a bacterial carbon dioxide reductase. *Science* **342**, 1382-1385 (2013).
16. Philips, M.F., Gruter, G.-J.M., Koper, M.T. & Schouten, K.J.P. Optimizing the electrochemical reduction of CO_2 to formate: A state-of-the-art analysis. *ACS Sustain. Chem. Eng.* **8**, 15430-15444 (2020).
17. Korman, T.P., Opgenorth, P.H. & Bowie, J.U. A synthetic biochemistry platform for cell free production of monoterpenes from glucose. *Nat Commun* **8**, 15526 (2017).
18. Dudley, Q.M., Nash, C.J. & Jewett, M.C. Cell-free biosynthesis of limonene using enzyme-enriched *Escherichia coli* lysates. *Synth. Biol.* **4**, ysz003 (2019).
19. Dudley, Q.M., Karim, A.S., Nash, C.J. & Jewett, M.C. Cell-free prototyping of limonene biosynthesis using cell-free protein synthesis. *Met. Engg.* (2020).
20. Meadows, A.L. et al. Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **537**, 694-697 (2016).
21. Zhang, H. et al. Microbial production of sabinene—a new terpene-based precursor of advanced biofuel. *Microb. Cell Fact.* **13**, 20 (2014).

22. Tashiro, M. et al. Bacterial production of pinene by a laboratory-evolved pinene-synthase. *ACS Synth. Biol.* **5**, 1011-1020 (2016).
23. Wu, J., Cheng, S., Cao, J., Qiao, J. & Zhao, G.-R. Systematic optimization of limonene production in engineered *Escherichia coli*. *J. Agric. Food. Chem.* **67**, 7087-7097 (2019).
24. Peralta-Yahya, P.P. et al. Identification and microbial production of a terpene-based advanced biofuel. *Nat. Comm.* **2**, 1-8 (2011).
25. Zhang, J. et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Comm.* **11**, 1-13 (2020).
26. Gleizer, S. et al. Conversion of *Escherichia coli* to generate all biomass carbon from CO2. *Cell* **179**, 1255-1263. e12 (2019).

3.6. Supplementary Information

3.6.1. General Materials and Equipment

Limonene, α-pinene and β-farnesene were obtained from Sigma Aldrich (Munich, Germany). Sabinene was obtained from Santa Cruz Biotechnology Inc. (Dallas, USA) and α-bisabolene was obtained from Alfa Aesar (Haverhill, USA). Chemicals and materials for cloning and protein expression were purchased from New England Biolabs GmbH (Frankfurt am Main, Germany) and Macharey-Nagel GmbH (Düren, Germany). Synthesis of optimized genes was done by Baseclear AG (Leiden, Netherlands). High resolution MS measurements were performed using an IDX Orbitrap High Performance Benchtop HRMS with an electrospray ion source and an Integrion HPLC system (Thermo Scientific). Mass spectroscopic data are reported as mass per charge ratio (*m/z*). Site-directed mutagenesis was performed using Quikchange II XL kit (Agilent). Identity of all recombinant proteins was confirmed using SDS-PAGE.

3.6.2. Experimental Procedures

Chemical synthesis of CoA esters

The synthesis of CoA esters and their analysis by LCMS was performed based on a previously published study by Peter *et al*.[1],

Analysis of CoA esters

Malyl-CoA and acetyl-CoA were measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LS/MS) equipped with an UHPLC (Agilent Technologies 1290 Infinity II) using a 50 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 2 µl of the diluted samples (1:10 in water). The flow was set to 0.250 ml/min and the separation was performed using 50 mM ammonium formate pH 8.1 (buffer A) and acetonitrile (B). We quantified the CoAs using external standard curves prepared in 1:10 diluted (water) sample matrix. The parameters for the multiple reaction monitoring (MRMs) are displayed in table S1 and the gradient in table S2. Data analysis was done using the Agilent Mass Hunter Workstation Software.

Table S1 MRM transitions for malyl- and acetyl-CoA

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
Malyl-CoA (Quantifier)	884.1	377.1	30	380	37	5	Positive
Malyl-CoA (Qualifier)	884.1	428	30	380	29	5	Positive
Acetyl-CoA (Quantifier)	810.1	302.2	30	380	35	5	Positive
Acetyl-CoA (Qualifier)	810.1	428	30	380	35	5	Positive

Table S2 Gradient used for the separation of CoAs

Time [min]	A [%]	B [%]
0	100	0
2	100	0
5	94	6
8	77	23
10	20	80
11	20	80
12	100	0
12.5	100	0

Plasmids, cloning and mutagenesis

The plasmids generated by Schwander *et al*.[2], and Borzyskowski *et al*.[3], were used to produce the enzymes to reconstitute CETCH and BHAC. The plasmid for the malate dehydrogenase is derived from Kitagawa *et al*.[4], The plasmid for the production of MtkAB is a gift from Thomas Schwander. The optimized genes corresponding to the terpene and PKS pathways were either obtained from a previous study[5] or synthesized as from Baseclear (Leiden, Netherlands). The list of all the plasmids with the details of vector and purification tags is listed in table S5. Primers and protocol for the point mutation of the PKS_{SgcE} KS domain were designed based on the Quikchange II XL mutagenesis kit manual (Agilent). The plasmid that expresses PKS_{SgcE} under the control of T7 promoter was used as template to make the mutation. The primers used for generating the point mutation are PKS_{SgcE}_KS_{Cys}_fw: CTACACGGTTGATGGCGCGGCTTCCTCTAGCTTGCTGAG and PKS_{SgcE}_KS_{Cys}_rv: CTCAGCAAGCTAGAGGAAGCCGCGCCATCAACCGTGTAG. After PCR amplification and inactivation of any

template DNA by DpnI, 1/10th the sample volume of 3 M sodium acetate (pH 5.2) was added to precipitate the amplified product. The precipitate was washed with 2.5 volumes of 100% ice-cold ethanol. After a brief centrifugation, the DNA pellet was further washed with 70% ethanol. The enriched plasmid was then directly used for transformation of *E. coli* XL-10 gold ultra-competent cells by electroporation. After confirming the mutant by DNA sequencing (Microsynth), the plasmid was introduced into *E. coli* BL21 (DE3) (New England Biolabs).

Protein production and purification

The plasmids to reconstitute the terpene and PKS pathway were expressed in *E. coli* BL21 (DE3). E. coli transformants were cultivated in LB medium at 37 °C. After A_{600nm} reached ~ 0.4 - 0.5, the cells were induced with 0.1 mM IPTG at 18 °C for 16 - 20 h. The cell pellet was dissolved (10 ml buffer/g pellet) in 150 mM Tris buffer pH 7.5 containing 0.2 M NaCl. After disrupting the cells by sonication, the cells were centrifuged at 20,000 g at 4 °C for 30 min. The lysed supernatant was then loaded onto a Ni-NTA column (Macherey Nagel) connected to a FPLC machine. Proteins were eluted using the same buffer with 0.25 or 0.5 M imidazole. For Idi, PKS_{SgcE} and TE_{SgcE}, the buffers also contained 1 mM DTT to avoid protein precipitation. The fraction containing the target protein from Ni-NTA column was diluted twice with 100 mM Tris (pH 7.5) and purified further by an ion-exchange column (5 mL HiTrap Q HP, GE Healthcare). Proteins were eluted over 20 column volumes of 100 mM Tris (pH 7.5) and 1 M NaCl, and the target proteins were concentrated using Amicon columns (MWCO 10, 30 and 100 kDa – Millipore). All the purified proteins were stored in 50 mM Tris buffer pH 7.5 containing 20 mM NaCl and 10% glycerol at -80 °C until further analysis. Except Hmgr, Idi and PKS_{SgcE}, the proteins were stable and active up to a period of 6 months under this storage condition.

Proteins to reconstitute the CO_2 to acetyl-CoA conversion (CETCH, BHAC and additional enzymes) were produced in *E.coli* BL21 (DE3) or Rosetta (DE3) pLysS (methylsuccinyl-CoA and propionyl-CoA oxidases (Mco, Pco)). For expression of 4-hydroxybutyryl-CoA synthase (Hbs) we co-expressed the 60 kDa chaperoin (groESL) for the correct folding of the protein. After transformation in the expression strains, the cultures were grown overnight on LB-agar plates containing the selection antibiotics. 2 l of salt buffered TB medium was directly inoculated with colonies from the selection plates and grown on 37 °C and 90 rpm till A_{600nm} 0.5-1.0. In general, the cultures were cooled down to 21 °C and induced with 0.25 mM IPTG. For 4-hydroxybutyryl-CoA hydratase (Hbd) 100 µM of Fe(II)SO4, 100 µM Fe(III)citrate and 20

mM fumarate were added along with IPTG. The Hbd-expressing culture was grown until A_{600nm} 4.0 and cooled down in a closed sterile Schott bottle to express the protein under microaerobic conditions. Except Pco, the expression of the proteins was done overnight. Pco was expressed at 25 °C for 4 h. The cells were harvested by centrifugation (15 min, 4 °C, 6000 g). Afterwards the cells were resuspended (2 ml buffer/g pellet) lysis buffer (500 mM NaCl, 50 mM HEPES, 10% glycerol, pH 7.8 at RT). 5 mM $MgCl_2$, 10 µg/ml DNAse and one tablet of SigmaFAST Protease Inhibitor Cocktail (Sigma-Aldrich) were added. The cells were lysed by micro fluidizer (twice at 16.000 psi). Afterwards the cell debris was spun down at 50,000 g for 1 h at 4 °C. The supernatant was filtered through a 0.45 µm membrane. Except for the glyoxylate reductase, the lysate was mixed with 3 ml Protino Ni-NTA agarose beads (Macherey-Nagel) and incubated on ice for 30-45 min (70 rpm). Afterwards the beads were collected in a gravity column and washed with three column volumes (cv) of lysis buffer. For the removal of unspecific bound proteins, the beads were washed with three cv of lysis buffer containing an additional 50 mM of imidazole and three cv with 75 mM imidazole. The elution was done with two cv of lysis buffer containing 500 mM imidazole. Since the glyoxylate reductase has a streptavidin (Strep) tag, the lysate was loaded on a Cytiva StrepTrap™ HP prepacked column attached to an Äkta start FPLC adjusted to a flow rate of 1 ml/min. The desalting/storage buffer was used for lysis and purification. The elution was done using the desalting/storage buffer with 5 mM d-Desthiobiotin. The collected recombinant proteins were concentrated using Amicon Ultra 15 mL Centrifugal Filters (Merck) accordingly to the protocol provided by the supplier. For desalting the protein solution was loaded on a HiLoad 16/600 Superdex 200 pg column (GE Healthcare). The desalting/storage buffer contained 200 mM NaCl, 50 mM HEPES and 10% glycerol and was adjusted to pH 7.8 at room temperature (22 ˚C). For Hbs and Hbd a concentration of 500 mM NaCl was used. The collected fractions were pooled and concentrated again. FAD was added to Pco and Mco depending on the concentration of protein. Enzymes requiring metal ions and cofactors were stored in 5 mM $MgCl_2$ and 2 mM Coenzyme B12 respectively. For the final storage, glycerol was added to a final concentration of 20 %. The proteins were flash frozen in liquid nitrogen and stored at -80 °C until further analysis.

In vitro reconstitution of BHAC

We tested four different setups for the reconstruction of the whole BHAC (**Figure S1A**). The general assay mix contained 100 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 20 mM sodium formate, 5 mM NADH, 5 mM NADPH, 0.1 mM pyridoxalphosphate, 14.4 µM (0.67 mg/ml) formate dehydrogenase, 0.33 µM (0.0115 mg/ml) malate dehydrogenase, 2.26 µM (0.099 mg/ml) BhcB, 1.37 µM (0.049 mg/ml) BhcC, 14.84 µM

(0.508 mg/ml) BhcD and 0.5 mM glyoxylate as substrate. The four setups contained additionally: 1) 0.5 mM glycine and 0.79 µM (0.043 mg/ml) BhcA, 2) 5.0 mM glycine and 0.79 µM (0.043 mg/ml) BhcA, 3) 0.5 mM glycine and 19.83 µM (0.890 mg/ml) BhcA and 4) 5.0 mM glycine and 19.83 µM (0.890 mg/ml) BhcA. The reactions were carried out at 30 °C in duplicates and in 50 µl reaction volume. 12 µl samples were withdrawn at 30, 60 and 90 min and quenched with 1.5 µl of 50 % formic acid and 1.5 µl of 500 mM polyphosphate for protein precipitation. The quenched samples were kept on ice until the end of the experiment and spun down at 20,000 g for 20 min at 4 °C. The supernatant was transferred into fresh tubes and stored at -20 °C until measurement.

***In vitro* reconstitution of the terpene biosynthesis modules**

The reactions for the *in vitro* production of **1-5** were performed in a sealed glass vial. Briefly, in a 100 µL reaction, 1 mM NADPH, 20 mM formate, 3 mM ATP, 3 mM PEP, 2 mM NADH, 1 mM DTT, 5 mM $MgCl_2$, 10 mM KCl were added in 50 mM Tris buffer pH 8.0. The list of enzymes and their amounts are listed in table S5. The reaction was initiated by adding 0.5 mM or 1 mM acetyl-CoA. To trap the volatile monoterpenes (**1-3**), the assay mix was overlayed with 30 µl of isopropylmyristate. The samples were incubated at 30 °C with shaking at 400 rpm up to 24 h. At specified intervals, the organic layer is withdrawn and diluted with hexane. The volume of isopropylmyristate withdrawn was simultaneously added to the reaction mix during the course of the assay. For the samples assaying the production of **4** and **5**, at these intervals, the workup of the samples was done by extracting twice the volume with ethyl acetate. The mix was then spun at 20,000 g for 15 min at 4 °C. The aqueous phase was mixed with equal volume of methanol and centrifuged to precipitate the proteins. Both the organic and aqueous phase were saved at -80 °C until further analysis. All the reactions were set up in triplicates.

***In vitro* reconstitution of the PDH production**

The *in vitro* assay for the reconstitution of PKS pathway to produce pentadecaheptaene (PDH) was performed as described previously[6,7] with minor modifications. In 100 mM phosphate buffer pH 8.0, 0.2 mM or 1.2 mM acetyl-CoA, 1.2 mM malonyl CoA, 1.2 mM NADPH, 1 mM DTT, 8 mM $MgCl_2$, 40 mM $KHCO_3$, 1 mM ATP, 0 to 10 µM PKS_{SgcE}, 0 to 50 µM TE_{sgcE} and 2 µM Pcc* were added to a total of 200 µl. The list of enzymes and their amounts are listed in table S5. The assay was performed at 30 °C with shaking at 400 rpm up to 24 h. At specified intervals, the sample was withdrawn and the polyketides are extracted twice

the volume with ethyl acetate. After evaporating the organic layer with an upstream flow of nitrogen, the residue was dissolved in 100 µl ethyl acetate. For UV-Vis analysis, the extract was measured at 395 nm to detect the formation of pentadecaheptaene **7**. After adding equal volume of methanol to the aqueous layer to precipitate proteins, the mix was spun down at 20,000 g at 4 °C for 10 min. The samples were stored at -80 °C until further analysis. All the reactions were set up in triplicates.

Coupling of CETCH and BHAC

For the coupling of the CETCH with the BHAC we used the same enzyme concentrations of the CETCH core cycle as in For the BHAC enzymes and the Mdh we used the amounts as described in ***In vitro* reconstitution of BHAC** I) above, except for the BhcD where the amount was increased by a factor of five. MtkAB was added at a concentration of 13.54 µM (1 mg/ml) and the glyoxylate reductase at a concentration of 0.62 µM (0.020 mg/ml). Other components were added in the following concentrations: 5 mM $MgCl_2$, 20 mM polyphosphate, 50 mM sodium bicarbonate, 20 mM sodium formate, 1 mM coenzyme A, 0.1 mM coenzyme B_{12}, 5 mM ATP, 5 mM NADPH, 5 mM NADH, 1 mM glycine, 0.1 mM pyridoxalphosphate and 100 µM propionyl-CoA as substrate. For the positive control (**Figure S1B (a)**), the enzymes of the CETCH core cycle were used together with the glyoxylate reductase to produce glycolate. To produce acetyl-CoA the enzymes of the CETCH core cycle were combined with BHAC enzymes, and Mdh plus MtkAB (**Figure S1B (c)**). The assays were done in triplicates at 50 µl each. 13.5 µl samples were taken at 60, 120 and 180 min and quenched in 1.5 µl 50% formic acid to stop the reaction. For the split assay, the CETCH core enzymes were used to produce glyoxylate for 60 min. The assay was done in a single assay, split into two batches after 60 min where either the glyoxylate reductase (**Figure S1B (b)**) or the BHAC enzymes, Mdh and MtkAB were added (**Figure S1B (d)**) and then further divided in triplicates to 50 µl. Samples were taken as described before at 120 and 180 min. The quenched samples were kept on ice until the end of the experiment and spun down at 20,000 g for 20 min at 4 °C. The supernatant was transferred into fresh tubes and stored at -20 °C until measurement.

Coupling of CETCH, BHAC and terpene biosynthetic modules

The coupling assay was performed in 2 steps by preparing the CETCH-BHAC and the terpene assay mix separately and then mixing equal volume (50 µl) of both in one-pot. In 100 mM Hepes buffer pH 7.5, the CETCH-BHAC assay mix contained 5 mM ATP, 5 mM NADPH, 5 mM NADH, 5 mM $MgCl_2$, 20 mM

polyphosphate, 50 mM bicarbonate, 20 mM formate, 1 mM CoA, 0.1 mM vitamin B_{12}, 1 mM glycine, 0.1 mM PLP, and the enzymes with amounts specified in table S5. In 50 mM Tris pH 8.0, the terpenoid assay mix contained 20 mM formate, 3 mM PEP, 10 mM KCl together with the enzymes listed in table S5 including 80 µg limonene synthase, 60 µg sabinene synthase, 80 µg α-pinene synthase, 40 µg α-bisabolene synthase, 40 µg β-farnesene synthase. After mixing both the mixes to 100 µl, the reaction was started with 0.1 mM propionyl-CoA and were incubated at 30 °C with shaking at 400 rpm up to 24 h. As positive controls, the CETCH-BHAC and the terpene assays were performed in parallel by adding 0.5 mM acetyl CoA to the latter. At regular intervals, samples were withdrawn from both the positive controls and the tests. Work-up of the samples to detect **1-5** was performed as described in ***In vitro* reconstitution of the terpene biosynthesis modules**. All the reactions were set up in triplicates.

Analysis of CO_2 incorporation using ^{13}C-labeled sodium bicarbonate and sodium formate

To verify the incorporation of CO_2 by the CETCH cycle as described in Schwander *et al*., we performed the CETCH-BHAC coupling (**Figure 1C**) with 50 mM ^{13}C-labeled sodium bicarbonate (and carbonic anhydrase) and 20 mM ^{13}C-labeled sodium formate. ^{13}C-labeled sodium formate was used to derive $^{13}CO_2$ released by the formate dehydrogenase for NADPH regeneration. All the other components that were present are described in **Coupling of CETCH and BHAC** and the sampling procedure remained the same**.** Malate-CoA ligase was ommited to produce malate as the final readout. The reaction was started with either 100 µM propionyl-CoA (positive control) or ddH_2O (negative control). For the evaluation by LC-MS, we used a targeted method to quantify the decarboxylated fragment of malate.

UPLC-MS analysis of malate

The different fragments of ^{13}C-labeled malate were measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LS/MS) equipped with an UHPLC (Agilent Technologies 1290 Infinity II) using a 150 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 1 µl of the diluted samples (1:25 in water). The flow was set to 0.100 ml/min and the separation was performed using dH_2O with 0.1% formic acid (buffer A) and methanol with 0.1% formic acid (B). Since malate is a dicarboxylic acid and it was unclear which carboxylic group leaves the molecule, we measured all the possible transitions The parameters for the multiple reaction monitoring (MRMs) are displayed in

table S3 and the gradient in table S4. Data analysis was done using the Agilent Mass Hunter Workstation Software.

Table S3 MRM transitions for decarboxylated fragment of malate

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
Malate +0 (Quant.)	133	89	35	380	11	5	Negative
Malate +0 (Qual.)	133	133	35	380	0	5	Negative
Malate +1 (Quant.)	134	89	35	380	11	5	Negative
	134	90	35	380	11	5	Negative
Malate +1 (Qual.)	134	134	35	380	0	5	Negative
Malate +2 (Quant.)	135	90	35	380	11	5	Negative
		91	35	380	11	5	Negative
Malate +2 (Qual.)	135	135	35	380	0	5	Negative
Malate +3 (Quant.)	136	91	35	380	11	5	Negative
	136	92	35	380	11	5	Negative
Malate +3 (Qual.)	136	136	35	380	0	5	Negative
Malate +4 (Quant.)	137	92	35	380	11	5	Negative
Malate +4 (Qual.)	137	137	35	380	0	5	Negative

Table S4 Gradient used for the measurement of the decarboxylated fragments of malate

Time [min]	A [%]	B [%]
0	100	0
4	100	0
6	0	100
7	0	100
7.1	100	0
12	100	0

Coupling of CETCH, BHAC and PDH production

In 100 mM phosphate buffer pH 8.0, 40 mM $KHCO_3$, 2.5 µM PKS_{SgcE}, 40 µM TE_{SgcE} and 2 µM Pcc* were added to a total of 50 µl. The list of enzymes and their amounts are listed in table S5. The reaction was initiated by adding equal volume of CETCH-BHAC mix (test) or 1.2 mM acetyl CoA (positive control). The assay was performed at 30 °C with shaking at 400 rpm up to 24 h. At specified intervals, the sample was withdrawn and the polyketides were extracted twice the volume with ethyl acetate. After evaporating the organic layer with an upstream flow of nitrogen, the residue was dissolved in 100 µl ethyl acetate. After adding equal volume of methanol to the bottom aqueous layer to precipitate proteins, the mix was spun down at 20,000 g at 4 °C for 10 min. All the reactions were set up in triplicates.

UPLC-MS analysis of terpene and polyketide intermediates

Analysis of the all the terpenes was done in GCMS (Agilent 5973N/6890N single quadrapole) by measuring 1 µL of the samples. An OPTIMA 5 column (30 m long, 0.32 mm inner diameter, 0.25 µm thick) was used for the separation with an initial temperature of 60 °C (2 min hold) followed by a gradient from 20 °C (1 min) to 150 °C then from 40 °C (1 min) to 320 °C. A constant flow rate of 1 ml/min was used. The injector had a temperature of 210 °C and was set for a 1:25 split. The MS had a mass range from 34 to 550 Da covered. The aqueous phase from the independent and coupled terpene experiments was directly analysed for the isoprenoid intermediates using UPLC-high resolution mass spectrometer (Orbitrap IDX™) set to negative ionisation mode. SeQuant ZIC-pHILIC (150 x 4.6 mm) was used for separating the isoprenoid intermediates. UPLC conditions: isocratic elution (10 mM ammonium carbonate and 118 mM ammonium hydroxide in acetonitrile:water (60.1:39.8)) for 10 min at a flow rate of 0.45 ml/min; injection volume: 3 µL; mass range: 65 – 1100 *m/z*). For the analysis of polyketides, both the organic and aqueous phased were analysed directly using UPLC-high resolution mass spectrometer (Orbitrap IDX™) set to positive ionisation mode. Kinetic EVO C18 column (50 x 2.1 mm) was used for the separation of the polyketide intermediates. UPLC conditions: 95 % of 0.1 % formic acid in water (Solvent A) for 2 min; 5 – 95 % 0.1 % formic acid in acetonitrile (Solvent B) for 2 – 11 min; 95 % B at 12 min; 95 % until 14 min. flow rate: 0.25 ml/min; injection volume: 5 µL; mass range: 100 – 1100 *m/z*).

3.6.3. Supplementary Text

Optimization of CETCH, summarized from earlier publications

CETCH had been optimized earlier in several rounds (CETCH 1.0 to CETCH 5.4)[2]. All versions of CETCH were tested in buffer containing 50 mM sodium bicarbonate which was further equilibrated with carbonic anhydrase to provide CO_2 in a dissolved form to the assay. In CETCH 2.0, methylsuccinyl-CoA dehydrogenase was engineered into a methylsuccinyl-CoA oxidase (Mco) to catalyze the oxidation of methylsuccinyl-CoA with molecular oxygen, which allowed CETCH to turn multiple times as shown by ^{13}C-labeling experiments. In CETCH 3.0, a read-out module was introduced to convert glyoxylate to malate that allowed a better quantification of CETCH assays. Also, an engineered formate dehydrogenase was used to regenerate the cofactor NADPH and (simultaneously) CO_2 in the assay. In CETCH 4.0, to protect the cofactors and intermediates and to prevent the oxidative damage from H_2O_2 (produced by Mco),

catalase (KatE) was added which resulted in increased efficiency in CO_2 fixation. In CETCH 5.0, to maintain a stable ATP pool and to regenerate ATP for Hbs, a polyphosphate transferase (Ppk) was included that increased the efficiency of the cycle to fix 4.3 CO_2-equivalents per acceptor molecule in 90 min. Finally, after further improvements to the cycle (optimizing Ccr), the efficiency of CETCH 5.4 reached a maximum of 5.4 fixed CO_2-equivalents per acceptor in 90 min. The CETCH cycle reached a plateau after 90 min and malate production could not be increased beyond 540 µM, indicating that malate inhibits CETCH cycle enzymes.

Optimization of BHAC, this work

To establish and optimize the BHAC, we reconstituted the BHAC *in vitro* using N-terminal His-tagged proteins, produced in *E. coli*. To test the functioning of the BHAC cycle, we started the reaction with 500 µM glyoxylate and monitored the formation of malate from oxaloacetate, using malate dehydrogenase (Mdh) over time. Note that β-hydroxyaspartate aldolase (BhcC), the first enzyme reaction of the BHAC that catalyzes the aldol condensation of glyoxylate with glycine, has an apparent Km of 4.3 ± 0.3 mM for glycine[3]. Thus, while providing high glycine concentrations might facilitate the first reaction, it might also lead to a faster depletion of glyoxylate, which is required in the last step as acceptor for aspartate-glyoxylate aminotransferase (BhcA), eventually creating a bottleneck. To optimize BHAC productivity, we initially tested two different glycine (0.5 mM and 5 mM), as well as two different BhcA (0.79 µM and 19.83 µM) concentrations. However, over the course of 90 min, total malate yields were comparable between the different conditions tested (~70%), indicating that the BhcC was operating robustly across a wide range of co-substrate and BhcA concentrations *in vitro* (**Figure S1A**).

3.6.4. Supplemetary Figures and Tables

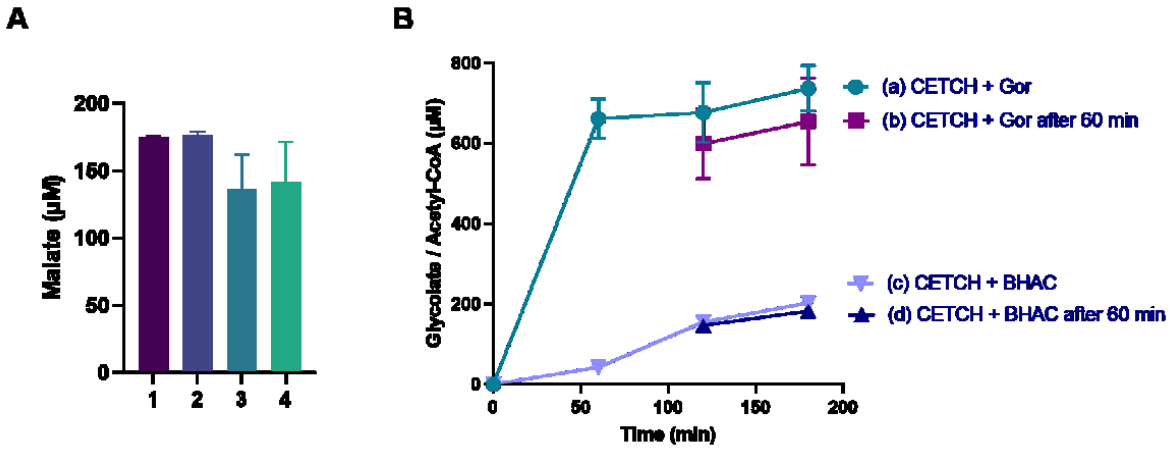


Figure S1 BHAC reconstitution and coupling to CETCH cycle for acetyl-CoA production. **A)** Malate production by the BHAC after 90 min. The general setup is described in: ***In vitro* reconstitution of BHAC.** All the reactions were started with 500 µM glyoxylate. With this setup, we tested different BhcA and glycine concentrations: 1) 0.5 mM glycine and 0.79 µM (0.043 mg/ml) BhcA, 2) 5.0 mM glycine and 0.79 µM (0.043 mg/ml) BhcA, 3) 0.5 mM glycine and 19.83 µM (0.890 mg/ml) BhcA and 4) 5.0 mM glycine and 19.83 µM (0.890 mg/ml) BhcA. The data represent n=2 ± standard deviation. **B)** Acetyl-CoA vs. glycolate production by the CETCH (+BHAC). The general setup is described in: **Coupling of CETCH and BHAC.** All the reactions were started with 100 µM propionyl-CoA. In this setup we tested whether the low acetyl-CoA yield is due to interference of the BHAC enzymes with the core cycle or due to intermediate drainage. (a) CETCH core cycle with Glyoxylate reductase (Gor). (b) Glyxoylate reductase added after 60 min. (c) BHAC enzymes, Mdh and MtkAB added after 60 min. (d) CETCH with BHAC enzymes, Mdh and MtkAB. The data represent n=3 ± standard deviation.

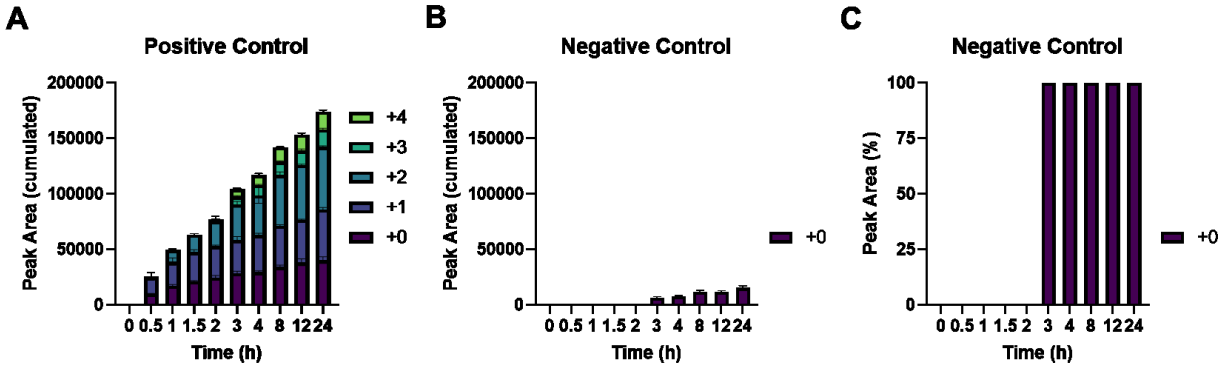


Figure S2 Fractional labeling of malate by incorporation of $^{13}CO_2$. +0, +1, +2, +3, +4 indicates the number of ^{13}C are incorporated into malate. To verify the incorporation of CO_2 by the CETCH cycle we repeated the CETCH-BHAC coupling (**Figure 1C**) with 50 mM ^{13}C-labeled sodium bicarbonate (and carbonic anhydrase) and 20 mM ^{13}C-labeled sodium formate. ^{13}C-labeled sodium formate was used to derive $^{13}CO_2$ released by the formate dehydrogenase for NADPH regeneration. In the first three turns of the CETCH cycle only single labeled glyoxylate is produced while the second $^{13}CO_2$ derived carbon is incorporated into CETCH cycle intermediates. For the formation of oxaloacetate and therefore malate by the BHAC, initially added glycine is used. Since the last reaction in the BHAC for the production of oxaloacetate requires another molecule of glyoxylate generated from fixed CO_2, a single labeled molecule of malate is stoichiometrically completely build from fixed CO_2. **A**) Total level of malate dissected into the labeled fractions. For fractional labeling of the positive control see **Figure 1D**. **B**) Fractional labeling in percentage of the negative control containing buffer. **C**) Total malate in the negative control. The data represent n=3 ± standard deviation.

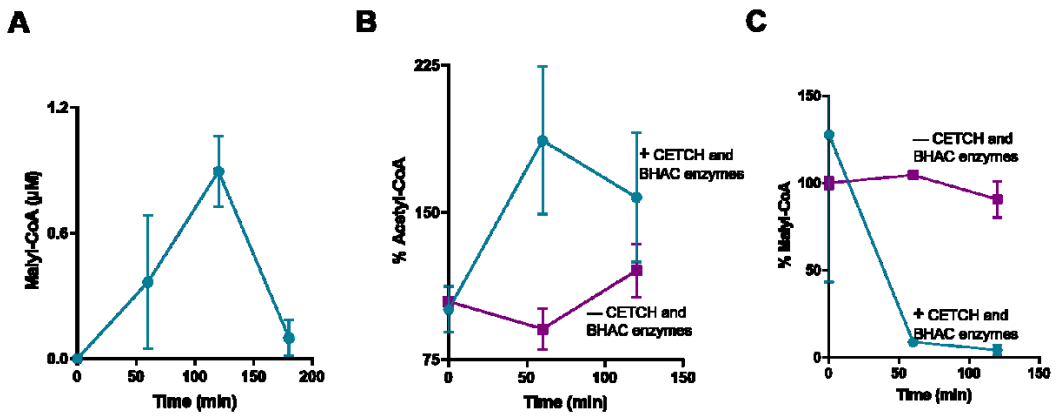


Figure S3 Profile of CoA esters in the CETCH-BHAC coupled assay. **A**) Residual concentration of malyl-CoA in the assay shown in Figure S1B (c). Stability of **B**) acetyl-CoA and **C**) malyl-CoA under assay conditions described in **Coupling of CETCH and BHAC**. All enzymes and cofactors except Mcl (to avoid the cleavage of malyl-CoA) was added to the positive control (**+**). Only the cofactors are added to the negative control (—). 100% corresponds to 300 µM of acetyl- and malyl-CoA. While acetyl-CoA was stable, malyl-CoA was depleted in less than 60 min. The data represent n=2 ± standard deviation.

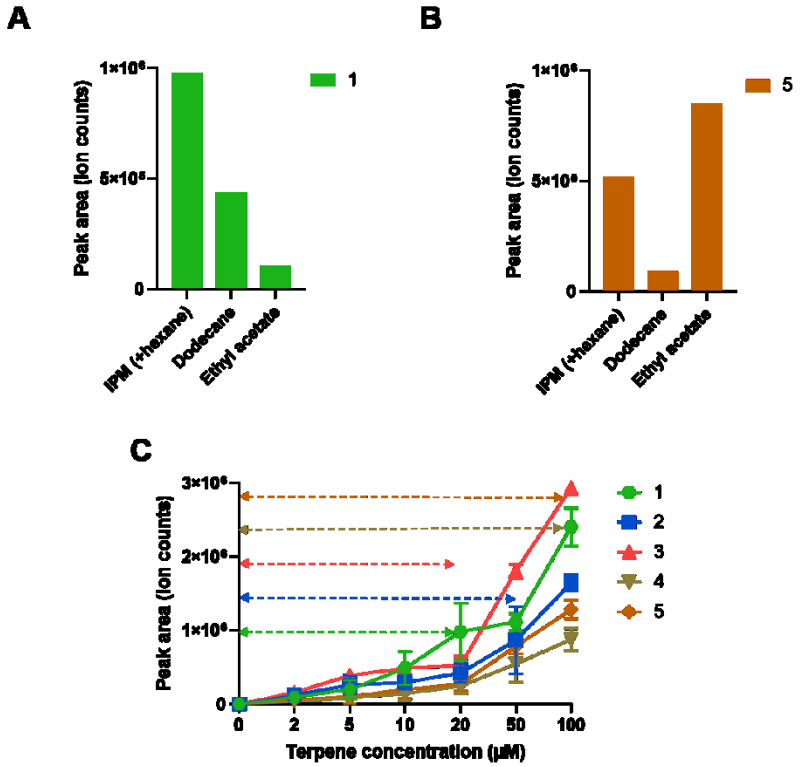


Figure S4 Optimization of terpene extraction in different solvents as measured by GCMS. Comparison of yield of **A**) 50 µM limonene (**1**) (representation) and **B**) 50 µM β-farnesene (**5**) (representation) with different organic solvents. IPM: isopropylmyristate. 30 µl IPM was added as an overlay to a 100µl standard + buffer mix. After a brief incubation, the IPM layer was carefully withdrawn and diluted with hexane before measurement using GCMS. Similarly, 10 % dodecane was also tested as an overlay to trap the terpenes. The withdrawn dodecane layer was further diluted with ethyl acetate for GCMS measurement. As a third solvent, 2x volumes of ethyl acetate (200 µl to a 100 µl standard + buffer mix) was tested. Followed by centrifugation at 20,000 g for 15 min at 4 °C, the organic phase was directly used for GCMS measurement. IPM resulted in maximum trapping

of monoterpenes **1-3** which was routinely used for subsequent measurements. For sesquiterpenes **4-5**, ethyl acetate was the best solvent. **C**) Measurement of terpene standards by GCMS. The linear range used for quantification of the corresponding terpenes is shown as dashed double-arrowed lines.

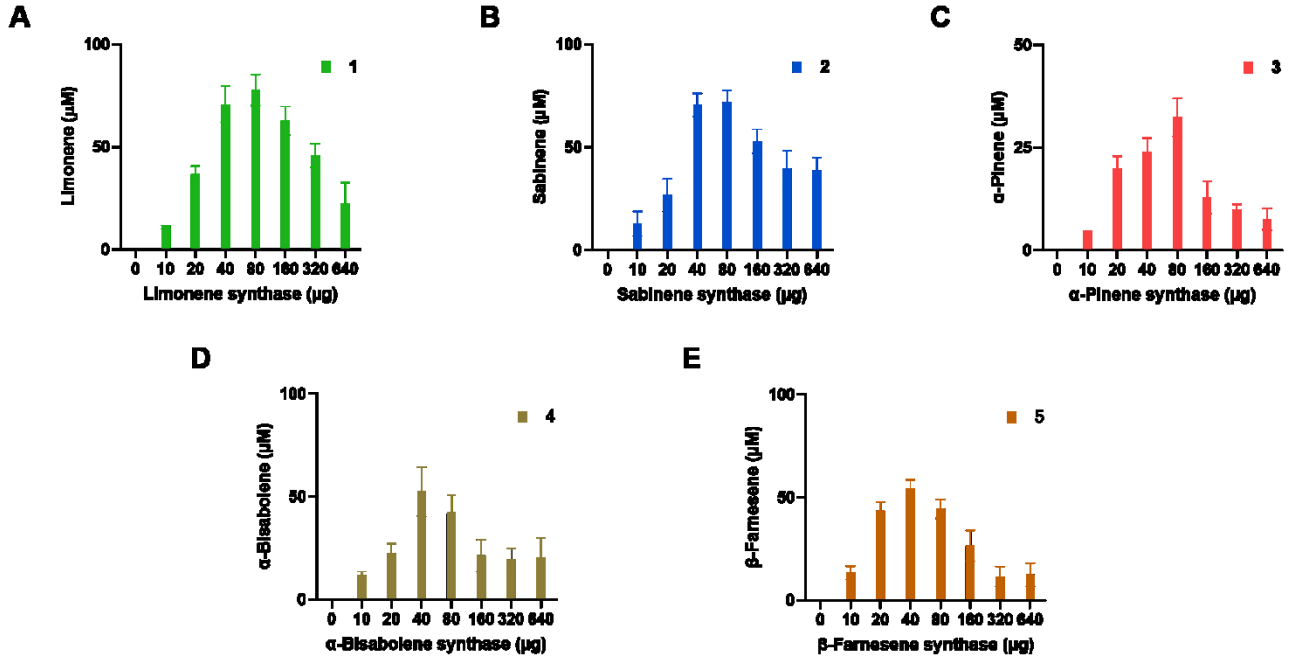


Figure S5 Comparison of different terpene synthase concentration (0 to 640 µg) for the production of terpenes and measurement by GCMS. The reaction is started with 0.5 mM acetyl CoA and run at 30 °C for 24 h. The extraction of monoterpenes **1-3** was done using isopropylmyristate overlay followed by dilution with hexane while sesquiterpenes **4-5** were extracted using 2x volumes of ethylacetate. The concentration of individual terpenes was quantified using the standard graph (**Figure S4C** linear range). The concentration at which maximum terpene production was observed has been chosen for the subsequent analysis.

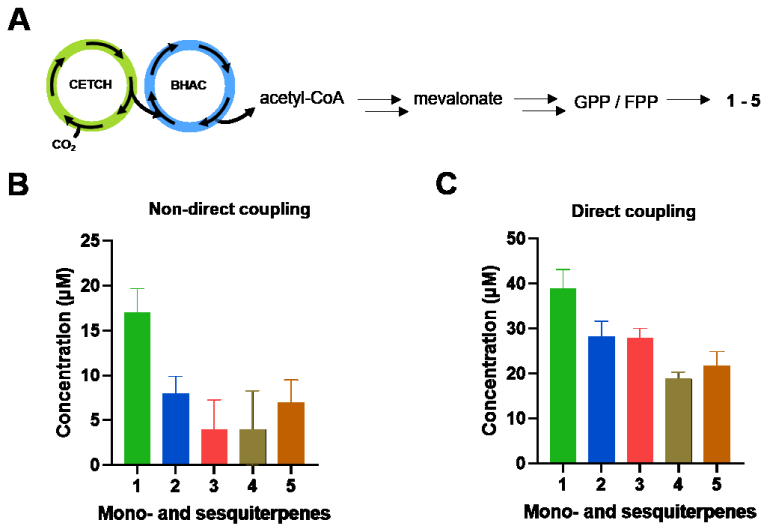


Figure S6 A) General scheme of the assay as described in methods section '**Coupling of CETCH, BHAC and terpene biosynthetic modules'**. Net production of terpenes in non-direct vs. direct coupling assay in 24 h. **B**) The CETCH-BHAC assay is first run independently for 4 h to which the terpene assay mix was subsequently added. In this non-direct coupling assay, the overall yield of monoterpenes **1-3** and sesquiterpenes **4-5** were below 20 µM. **C**) CETCH-BHAC cascade and terpene biosynthesis modules were operated in a single pot continuously. In this direct coupling approach, the net yield of terpenes improved 3- to 4-fold.

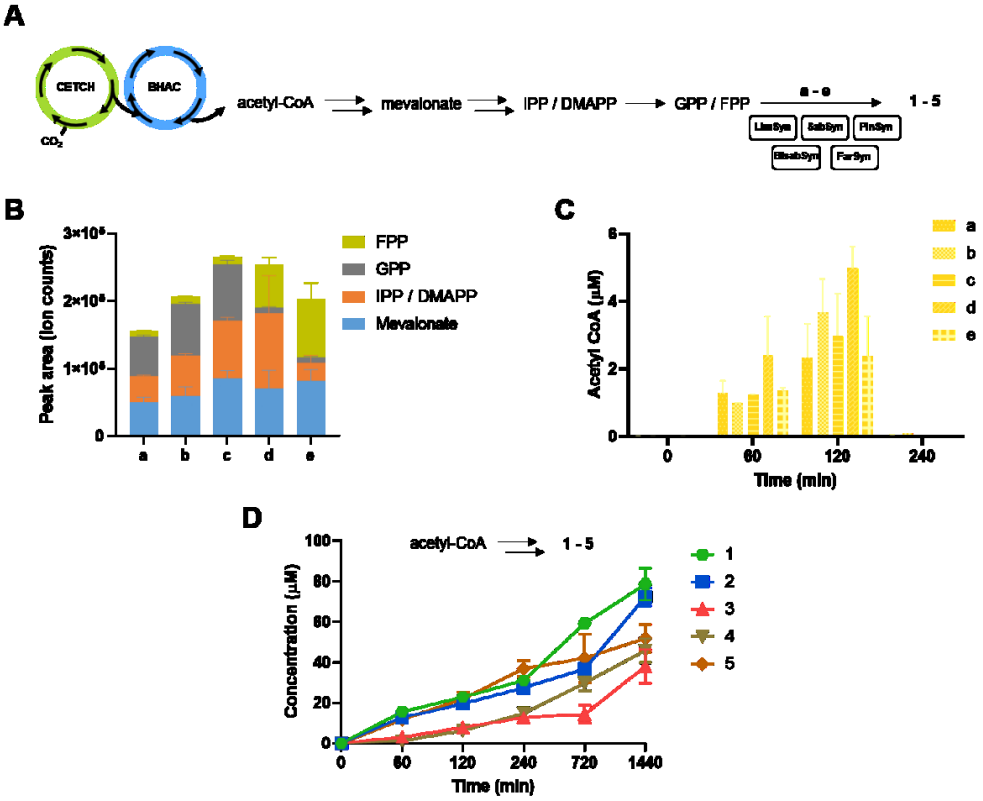


Figure S7 A) General scheme of the direct coupling assay as described in methods section '**Coupling of CETCH, BHAC and terpene biosynthetic modules'**. a, b, c, d, e refers to individual reactions with limonene synthase, sabinene synthase, α-pinene synthase,

α-bisabolene synthase and β-farnesene synthase respectively. **B**) Accumulation of mevalonate pathway intermediates measured by LCMS at 24 h. Equal volume of methanol was added to the final assay mix to stop the reaction and to precipitate the proteins. After centrifugation at 20,000 g for 15 min at 4°C, the supernatant was used directly to measure by LCMS. **C**) Concentration of residual acetyl-CoA measured over 4 h. From 4 h, only negligible amounts of acetyl-CoA could be detected. **D**) Acetyl CoA to terpenes as a positive control. The reaction is started with 0.5 mM acetyl-CoA.

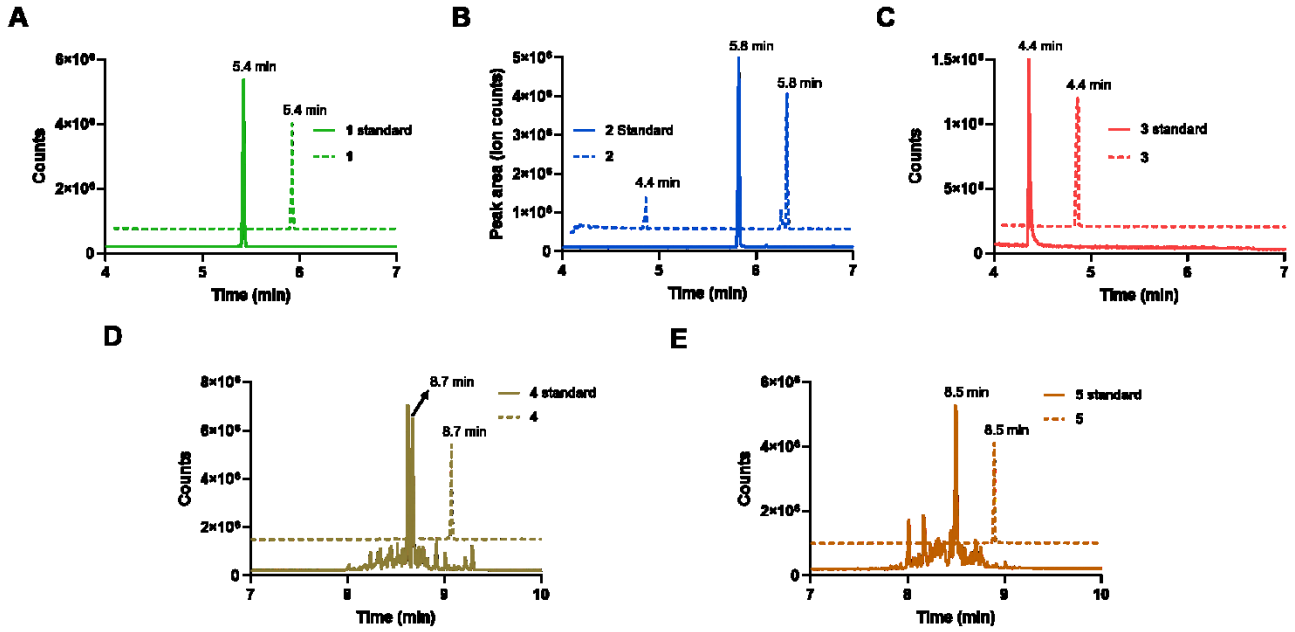


Figure S8 GCMS analysis of the production of mono- and sesquiterpenes from the CETCH-BHAC-terpene coupled assay (refer figure 2C) and comparison with authentic standards. The assay is performed as described in the methods section '**Coupling of CETCH, BHAC and terpene biosynthetic modules'**. For clarity, only the traces at 24 h time point are shown. **A**) Representative trace of limonene (**1**) from 100 µM propionyl-CoA. **B**) Representative trace of sabinene (**2**). A fraction of α-pinene (**3**) was also observed at a retention time of 4.4 min. **C**) Representative trace of α-pinene (**3**). **D**) Representative trace of α-bisabolene (**4**). The bisabolene standard come as a mixture of isomers however, exclusively α-bisabolene is observed in the GCMS trace in the assay sample. **E**) Representative trace of β-farnesene (**5**).

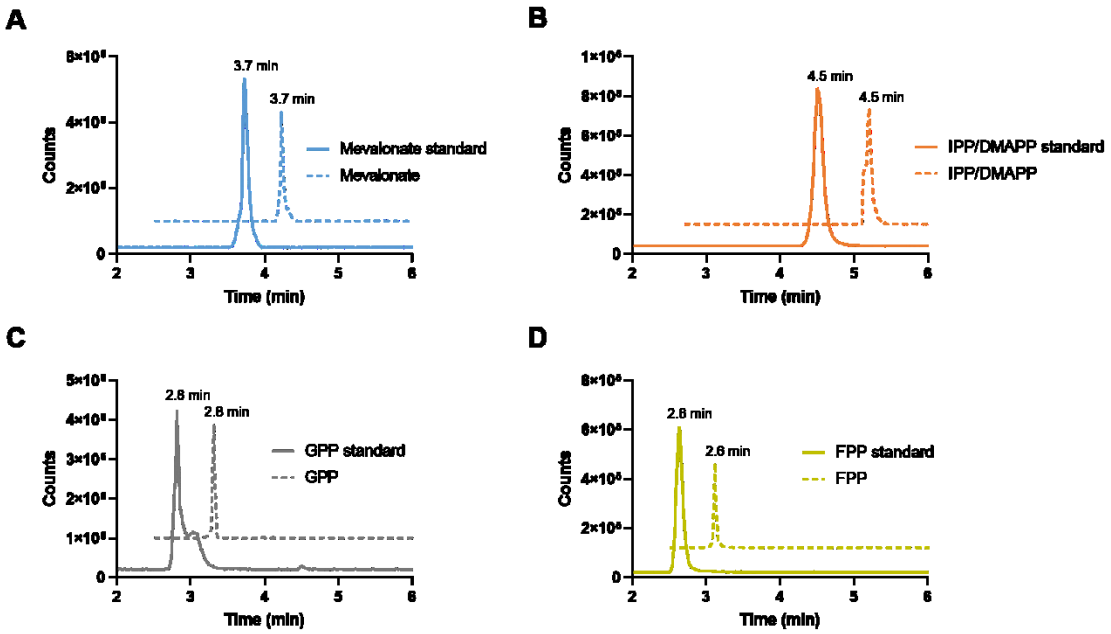


Figure S9 LCMS analysis of the terpene intermediates from the CETCH-BHAC-terpene coupled assay (refer figure 2C) and comparison with authentic standards. Representative traces of **A**) mevalonate, **B**) IPP/DMAPP, **C**) GPP and **D**) FPP. IPP and DMAPP could not be separated even after optimizing the chromatographic method. Equal volume of methanol was added to the final assay mix (and the standards, as a positive control) to stop the reaction and to precipitate the proteins. After centrifugation at 20,000 g for 15 min at 4°C, the supernatant was used directly to measure by LCMS.

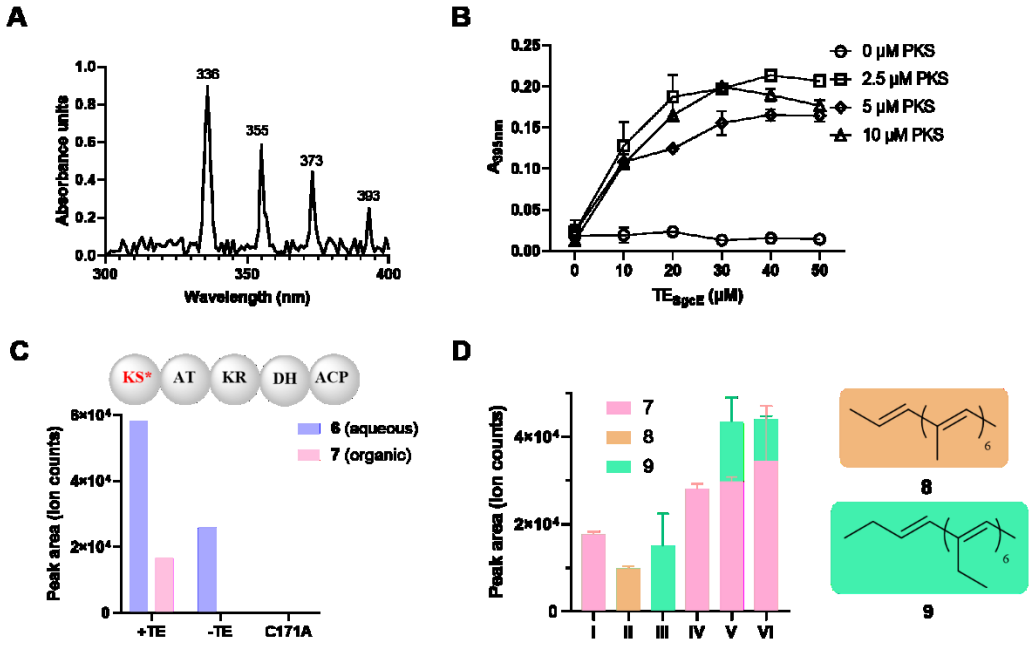


Figure S10 A) UV-Vis profile of pentadecaheptaene (**7**). **7** exhibited a spectrum typical of a polyene with multiple absorption maxima between 300 and 400 nm. **B**) Absorbance of the ethyl acetate extracts from the PKS enzymatic assay at 395 nm. The

production of **7** at various PKS_{SgcE} and TE_{SgcE} concentrations is shown. 2.5 µM PKS_{SgcE} and 40 µM PKS_{TE} was used for the subsequent analysis. **C**) Analysis of the KS_{C171A} mutant. Compared to the positive control (+TE), neither the production of **6** nor **7** was observed. **D**) Formation of substituted heptaenes (**8** and **9**) from 100 µM propionyl-CoA in the CETCH-BHAC-PKS coupled assay, using different extender units. I: malonyl-CoA (positive control); II: methylmalonyl-CoA; III: ethylmalonyl-CoA; IV: malonyl- + methylmalonyl-CoA; V: malonyl- + ethylmalonyl-CoA; VI: malonyl- + methylmalonyl- + ethylmalonyl-CoA. All the assays were performed in triplicates and the mean ± S.D. are plotted.

Table S5. List of enzymes used in the study

Cycle/pathway	Abbreviation	Full name	Source	Vector	Tag	Origin Reference
CETCH	Pco	Propionyl-CoA oxidase	*A. thaliana*	pET16b	His	[2]
CETCH	Ccr	Crotnonyl-CoA carboxylase/reductase	*M. extorquens*	pET16b	His	[2]
CETCH	Epi	Epimerase	*R. sphaeroides*	pET16b	His	[8]
CETCH	Mcm	Methylmalonyl-CoA mutase	*R. sphaeroides*	pET16b	His	[9]
CETCH	Scr	Succinyl-CoA reductase	*C. kluyveri*	pCDF-Duet-1	His	[2]
CETCH	Ssr	Succinic semialdehyde reductase	*H. sapiens*	p2BP1	His	[2]
CETCH	Hbs	4-hydroxybutyryl-CoA synthetase	*N. maritimus*	pET16b	His	[10]
CETCH	Hbd	4-hydroxybutyryl-CoA dehydratase	*N. maritimus*	pRSET-B	His	[2]
CETCH	Ecm	Ethylmalonyl-CoA mutase	*R. sphaeroides*	pET16b	His	[8]
CETCH	Mco	Methylsuccinyl-CoA oxidase	*R. sphaeroides*	pET16b	His	[2]
CETCH	Mch	Mesaconyl-CoA hydratase	*R. sphaeroides*	pET16b	His	[11]
CETCH	Mcl	Malyl-CoA/citramalyl-CoA lyase	*R. sphaeroides*	pET16b	His	[12]
CETCH	KatE	Catalase	*E. coli*	pCAN24N (ASKA JW1721)	His	[4]
CETCH	Fdh	Formate dehydrogenase (D221A)	*M. vaccae*	pET21a	His	[13]
CETCH	smPPK2-I	Polyphosphate kinase ADP - ATP	*S. meliloti*	pET28a	His	[2]
CETCH	Gor	Glyoxylate/sucinnic semialdehyde reductase	*G. oxidans*	pTE1125	Strep	Gift from Martina Carrillo Camacho
BHAC	BhcA	Aspartate glycine aminotransferase	*P. denitrificans*	pET16b	His	[3]
BHAC	BhcD	Iminsuccinate reductase	*P. denitrificans*	pET16b	His	[3]
BHAC	BhcB	Beta-hydroxyaspartate dehydratase	*P. denitrificans*	pET16b	His	[3]
BHAC	BhcC	Beta-hydroxyaspartate aldolase	*P. denitrificans*	pET16b	His	[3]
BHAC	Mdh	Malate dehydrogenase	*E. coli*	pCAN24N (ASKA JW3205)	His	[4]
BHAC	MtkAB	Malate thiokinase	*M. extorquens*	pET28b	His	Gift from Thomas Schwander
Terpene	PhaA	Acetyl-CoA acetyltransferase	*C. necator*	pET28a	His	[5]
Terpene	Hmgs	HMG-CoA synthase (A110G)	*E. faecalis*	pET28a	His	[5]
Terpene	Hmgr	HMG-CoA reductase	*E. faecalis*	pET28a	His	[5]
Terpene	Mvk	Mevalonate kinase	*M. mazei*	pET28a	His	[5]
Terpene	Pmvk	Phosphomevalonate kinase	*S. pneumoniae*	pET28a	His	[5]
Terpene	Mdc	Mevalonate-PP decarboxylase	*S. pneumoniae*	pET28a	His	[5]

Terpene	Idi	Isopentenyl-PP isomerase	*E. coli*	ASKA JW2857	His	[4]
Terpene	Gpps	Farnesyl-PP synthase (S82F)	*G. stearothermophilus*	pET28a	His	[5]
Terpene	IspA	Farnesyl-PP synthase	*E.coli*	pET28a	His	[4]
Terpene	LimSyn	(+)-Limonene synthase	*M. spicata*	pET28a	His	[5]
Terpene	SabSyn	Limonene synthase (N345A)	*M. spicata*	pET28a	His	[5]
Terpene	PinSyn	α-Pinene synthase	*P. sitchensis*	pET28a	His	[5]
Terpene	BisabSyn	α-Bisabolene synthase	*A.grandis*	pET28a	His	Synthesized gene
Terpene	FarSyn	β-Farnesene synthase	*M.piperita*	pET28a	His	Synthesized gene
Terpene	PK/LDH	Pyruvate kinase/lactate dehydrogenase	*Sigma*			
PKS	PKS_{SgcE}	C-1027 polyketide synthase	*S.globisporus*	pET28a	His	Synthesized gene
PKS	TE_{SgcE}	C-1027 thioesterase	*S.globisporus*	pET28a	His	Synthesized gene
PKS	Pcc*	Propionyl-CoA carboxylase (D407I)	*M. extorquens*	JZ105	His	Gift from Jan Zarzycki

Table S6. Kinetic data of enzymes

Cycle/pathway	Abbreviation[a]	mg mL^{-1}	mg in assay[b]	*Vmax* (U mg^{-1})	K_M (mM)	U ml^{-1} assay	Reference
CETCH	Pco	2.5	0.007	12	0.044	0.8	[2]
CETCH	Ccr	6.1	0.001	110	0.17	1	[2]
CETCH	Epi	5.0	0.0005	440	0.08	2	[8]
CETCH	Mcm	4.8	0.001	20	0.14	0.2	[14]
CETCH	Scr	15.1	0.007	29	0.003	2	[15]
CETCH	Ssr	8.3	0.001	3.9	0.013	0.04	[2]
CETCH	Hbs	14.1	0.02	2	0.19	0.4	[10]
CETCH	Hbd	8.8	0.002	26	0.06	0.5	[10]
CETCH	Ecm	8.4	0.002	7	0.06	0.1	[8]
CETCH	Mco	35.0	0.07	0.1	0.03	0.07	[2]
CETCH	Mch	4.5	0.005	1500	n.d.	75	[11]
CETCH	Mcl	5.9	0.025	5	0.01	1	[12]
CETCH	KatE	27.6	0.006	11740	86.5	704	[16]
CETCH	Fdh	27.0	0.03	1.4	0.37	0.4	[13]
CETCH	smPPK2-I	7.5	0.004	12	0.032	0.5	[17]
CETCH	Gor	3.65	0.001	n.d.	n.d.		unpublished
BHAC	BhcA	2.7	0.005	116	0.23	6	[3]
BHAC	BhcD	5.8	0.003	57	0.2	2	[3]
BHAC	BhcB	36.3	0.025	358	0.09	90	[3]
BHAC	BhcC	43.4	0.004	1	2.9	0.04	[3]

BHAC	Mdh	23	0.002	1611	0.04	32	[18]
BHAC	MtkAB	10.2	0.05	n.d.	n.d.		unpublished
Terpene	PhaA	19	0.002	81	0.4	2	[5]
Terpene	Hmgs	10	0.005	1.5	0.01	0.08	[5]
Terpene	Hmgr	3.6	0.03	4	0.02	1.2	[5]
Terpene	Mvk	17.9	0.005	8	0.07	0.4	[5]
Terpene	Pmvk	18.6	0.005	15	0.008	0.8	[5]
Terpene	Mdc	11.8	0.03	4	0.1	1.2	[5]
Terpene	Idi	4.6	0.025	2.1	0.0035	0.5	[19]
Terpene	Gpps	10	0.005	7	0.005	0.4	[5]
Terpene	IspA	9.5	0.005	n.d.	n.d.		
Terpene	LimSyn	34	0.08	n.d.	n.d.		[5]
Terpene	SabSyn	29	0.06	n.d	n.d.		[5]
Terpene	PinSyn	19	0.08	n.d.	n.d.		[5]
Terpene	BisabSyn	29	0.04	n.d.	n.d.		
Terpene	FarSyn	9	0.04	n.d	n.d.		
Terpene	PK/LDH	1U/µl	0.0005	n.d.	n.d.		
PKS	PKS_{SgcE}	20.2	0.03	n.d.	n.d.		
PKS	TE_{SgcE}	15	0.06	n.d.	n.d.		
PKS	Pcc*	3.1	0.03	n.d.	n.d.		

[a] Refer to Table S3 for enzyme name and source [b] Amount corresponds to 100 µl assay volume

Table S7 Synthesized genes

β-Farnesene synthase

GCTAGCATGGCGACTAACGGCGTAGTGATTAGCTGCCTGCGTGAAGTGAGACCGCCGATGACTAAACATGCGCCTAGCATGTGGACTGACA
CCTTCTCTAACTTTAGTCTTGATGATAAAGAGCAACAGAAATGTAGTGAAACCATTGAAGCACTGAAACAGGAAGCGCGCGGTATGCTGATG
GCTGCTACCACTCCACTGCAGCAGATGACCCTGATCGACACCCTGGAACGTCTGGGTCTGTCCTTCCATTTCGAAACCGAAATCGAATATAAA
ATTGAACTGATCAACGCTGCTGAAGACGACGGTTTCGACCTGTTTGCGACCGCTCTGCGTTTCCGTCTGCTGCGTCAGCACCAGCGTCATGTTT
CTTGTGACGTTTTCGATAAATTCATCGATAAAGATGGTAAATTCGAAGAATCTCTGTCTAACAACGTTGGTGGCCTGCTGTCCCTGTACGAAGT
TGCGCACGTGGGTTTCCGTGAAGAACGCATCCTGCAGGAAGCTGTGAACTTCACCCGTCACCACCTGGAAGGTGCTGAACTGGACCAGAGCC
CGCTGCTGATCCGTGAAAAAGTTAAACGTGCGCTGGAACACCCGCTGCACCGTGACTTCCCGATAGTCTACGCACGTCTGTTCATCTCTATTTA
TGAAAAGGACGACTCGCGCGATGAACTGTTGCTGAAACTCAGTAAGGTGAACTTTAAATTTATGCAGAACCTGTATAAAGAAGAACTGTCTC
AGCTGTCTCGTTGGTGGAACACCTGGAACCTGAAATCTAAACTGCCGTATGCACGTGATCGTGTTGTTGAAGCATACGTTTGGGGCGTTGGTT
ACCACTACGAACCGCAGTACTCCTATGTTCGTATGGGTCTGGCTAAAGGTGTTCTGATCTGCGGTATTATGGATGACACCTATGACAACTACG
CTACCCTGAACGAAGCACAGCTGTTCACCCAGGTTCTGGATAAATGGGATCGTGACGAAGCGGAACGTCTGCCGGAATACATGAAAATCGTT
TACCGTTTCATCCTGTCTATCTACGAAAACTACGAACGTGATGCTGCGAAACTGGGTAAATCCTTCGCTGCTCCGTACTTCAAAGAAACCGTGA
AACAGCTGGCGCGTGCATTCAACGAAGAACAGAAATGGGTAATGGAACGTCAGCTGCCGTCCTTCCAGGACTACGTGAAAAACAGTGAAAA
AACCTCCTGCATCTACACCATGTTCGCGAGCATCATCCCAGGCCTGAAATCCGTTACCCAGGAAACCATCGACTGGATCAAATCTGAACCGAC
CCTGGCAACCTCTACCGCGATGATCGGTCGCTACTGGAACGATGCGAGCTCTCAGCTGCGTGAATCTAAAGGCGGTGAAATGCTGACCGCTC
TGGACTTCCACATGAAAGAATACGGTCTGACCAAAGAAGAAGCTGCGTCTAAATTCGAAGGCCTGGTGGAAGAAACTTGGAAAGATATCAA
CAAAGAATTCATCGCGACCACTAACTACAACGTTGGCCGTGAAATCGCTATCACCTTCCTGAACTACGCGCGTATCTGCGAAGCAAGCTATAG
CAAAACCGACGGTGACGCGTACAGCGATCCGAACGTTGCGAAAGCGAACGTTGTGGCGCTGTTCGTTGATGCTATCGTTTTCTAAGTCGAC

α-Bisabolene synthase

CAGTTCCCGCAGACCGTTGATTGGATTCTTAAAAACCAGTTGAAAGATGGCAGCTGGGGCATTCAGTCCCACTTCCTGCTGTCTGACCGTCTC
CTGGCCACTCTGTCTTGCGTGCTGGTTCTGCTGAAATGGAACGTTGGCGACCTGCAAGTTGAACAGGGCATCGAATTCATTAAATCTAACCTG
GAACTGGTTAAAGATGAAACTGATCAGGATTCTCTTGTAACTGACTTCGAAATCATCTTCCCGTCCCTGCTGCGTGAAGCGCAGTCTCTGCGTC
TCGGCCTGCCTTATGACCTGCCGTACATCCATCTGCTGCAGACTAAACGCCAGGAACGTCTGGCAAAACTGAGCCGTGAAGAAATTTACGCG

GTTCCGAGCCCGCTGCTGTACTCTCTGGAAGGTATCCAGGACATCGTAGAATGGGAACGCATCATGGAAGTTCAATCCCAGGATGGTAGCTT
CCTGAGCTCTCCGGCTAGCACTGCATGCGTCTTTATGCACACCGGTGACGCTAAATGCCTGGAATTCCTGAACTCCGTAATGATCAAATTTGGT
AACTTCGTTCCGTGCCTGTACCCGGTAGATCTCCTGGAACGTCTGCTGATTGTTGACAACATCGTGCGCCTGGGTATCTACCGTCATTTTGAAA
AAGAAATCAAAGAAGCACTGGACTACGTTTACCGTCACTGGAACGAACGTGGCATCGGCTGGGGCCGCCTGAATCCGATTGCAGACCTGGA
AACCACCGCTCTGGGCTTTCGCCTGCTGCGCCTGCACCGCTACAACGTTTCCCCGGCAATCTTCGACAACTTCAAAGATGCTAATGGTAAATTC
ATCTGCTCCACCGGTCAGTTTAACAAAGATGTAGCGTCCATGCTGAACCTCTACCGCGCTTCCCAGCTGGCGTTTCCGGGTGAAAACATCCTC
GATGAAGCTAAATCCTTCGCGACCAAATATCTGCGTGAAGCCTTGGAAAAATCTGAAACCAGCAGCGCTTGGAACAACAAGCAAAACCTGTC
CCAGGAGATCAAATACGCGCTTAAAACTTCCTGGCACGCTTCAGTGCCGCGCGTTGAAGCGAAACGTTACTGCCAGGTTTACCGTCCAGACTA
TGCACGTATCGCAAAGTGCGTGTACAAACTGCCGTACGTAAACAACGAAAAATTCCTGGAGCTGGGTAAACTGGATTTCAACATTATTCAGTC
AATCCACCAGGAAGAAATGAAAAACGTGACCTCTTGGTTCCGTGATTCTGGTCTGCCGTTGTTCACCTTCGCGCGTGAACGTCCTCTGGAATT
CTACTTCCTGGTTGCCGCAGGTACCTACGAACCGCAGTATGCAAAATGTCGTTTCCTGTTCACTAAAGTTGCGTGCCTGCAGACTGTTCTGGAC
GACATGTATGATACCTACGGCACTCTGGACGAACTGAAACTGTTCACTGAGGCTGTGCGTCGTTGGGATCTGTCTTTCACCGAAAACCTGCCG
GATTACATGAAACTTTGTTACCAGATCTACTATGACATTGTCCACGAAGTGGCGTGGGAAGCTGAAAAAGAACAGGGTCGTGAACTCGTTTC
GTTCTTCCGTAAAGGTTGGGAAGATTACCTGCTTGGTTATTATGAAGAAGCAGAATGGCTGGCTGCTGAATACGTGCCGACCCTGGATGAGT
ACATCAAAAACGGTATCACCTCCATCGGTCAACGCATCCTGCTGCTGAGCGGTGTGCTGATCATGGACGGCCAGTTGCTGAGCCAGGAAGCA
CTGGAAAAAGTTGATTACCCAGGTCGTCGTGTACTGACCGAACTGAATTCTCTGATCAGCCGTCTGGCGGATGACACCAAAACTTATAAAGCG
GAAAAAGCACGTGGTGAACTGGCTTCCTCTATTGAATGCTATATGAAAGATCACCCGGAATGTACCGAAGAAGAGGCACTGGATCACATTTA
CTCCATCCTCGAACCGGCGGTTAAAGAACTGACCCGTGAGTTCCTGAAACCGGATGATGTTCCGTTCGCGTGTAAGAAAATGCTGTTCGAAG
AAACTCGTGTGACCATGGTTATCTTCAAAGATGGTGACGGTTTCGGTGTTTCTAAACTGGAAGTTAAAGACCACATCAAAGAATGCCTGATCG
AACCGCTGCCGCTGTAACTCGAG

PKS$_{SgcE}$

CATATGAGCCGTATCGCTATCGTTGGTGTTGCATGCACCTATCCGGACGCAACCACCCCGCGTGAACTGTGGGAAAACGCAGTAGCAGGCCG
TCGTGCCTTTCGTCGTCTGCCGGACGTGCGTATGCGTCTGGACGATTACTGGAACCCGGACCCGACCGTTCCGGACACCTTCTACGCGCGTAA
CGCGGCTGTTCTGGAAGGCTGGGAATTCGATCGTGTAGCTCATCGTATTGCTGGTTCCACCTTCCGTAGCACTGATCTGACCCACTGGCTGGC
TCTGGACACCGCGACGCGTGCGCTGGCTGACGCAGGTTTCCCGGCAGGTGAAGGTCTGCCGACTGAACGTACCGGCGTTGTAGTTGGTAAC
ACCCTGACGGGTGAGTTTAGCCGCGCTAACGGTCTGCGCCTGCGTTGGCCGTACGTTCGCCGCATCCTGGCTGACGCACTGCAGGAACAGGA
ATGGGACGACGACCGCCTGGGCGCCTTCCTGCGCGGCGTGGAAGAAGCCTACAAGAAACCGTTCCCGGCTGTCGACGAAGACACGCTGGCC
GGCGGCTTGAGCAACACCATTGCTGGTCGCATCTGTAACCACTTTGACCTGAACGGTGGCGGCTACACGGTTGATGGCGCGTGCTCCTCTAG
CTTGCTGAGCATCACCACCGCGGCTACCTCTCTGCAGTCCGGTGACCTGGACGTCGCAGTGGCGGGTGGCGTTGATCTGAGCATTGACCCGT
TTGAAATCATCGGCTTCGCCAAAACTGGGGCGCTGGCGCGTAAAGAAATGCGTCTGTACGATCGTGGCTCCAACGGTTTCTGGCCGGGCGAG
GGTTGCGGCATGGTTGTTCTGATGCGTGAAGAAGACGCCGTTGCGTCCGGCCACCGCATCTATGCATCTATCGCGGGTTGGGGCATTAGCTC
TGACGGTCAGGGCGGCATTACTCGCCCGGAAGTATCCGGCTACCAGCTGGCACTGTCCCGCGCTTATGACCGTGCCGGTTTCGGTATTGAAA
CCGTGCCGCTGTTCGAGGGTCATGGTACCGGCACCGCGGTAGGGGACGCAACGGAACTGCGTGCGATCATGAGCGCACGTGCGGCGGCGG
ATCCGCACGCGCCGTCTGCTGTGATCACCTCTATCAAAGGCATGATCGGTCACACCAAAGCCGCTGCAGGCATCGCTGGCCTGATTAAGGCTG
TAATGGCGCTGGACAGCGGTGTGCTGCCGCCGGCTATTGGTTGTGTTGATCCGCATGACCTGCTCACTGACGAATCGGCGAACCTGCGTGTT
CTGCGTAAGGCGGAAAGCTGGCCGGAAAACGCGCCGTTGCGTGCGGGCATCACCGCGATGGGTTTCGGCGGTATCAACACCCATGTTGTCC
TGGATCGCTCCGACGCCTCCGGTCGTCGTCCGGCGGTTAACCGTCGTACTACTCTGCTGGCGAACTCTCTCCAGGATTCTGAGCTGCTCCTGCT
TGACGGTGAGTCCCCGGCGGCGCTGGCGCGTCGTCTGACCCAGGTGGCGGATTTCGCCGCACAGGTATCCTATGCGCAGCTGGGTGACCTG
GCAGCCACGCTCCAGCGTGAACTGCGTGATCTGCCTCACCGCGCCGCCGTAGTGGCTACCTCTCCAGAAGATGCGGAACTCCGTCTTCGTGGC
CTGGCGGAAACCGCCGGCGGTCGTGCACCTGATGATGGTCCGGTATTCAGTCAAGATGGCCGCGCGTTCCTGGGTACCGCTGCTGAAGGTG
CACGTGTAGGCTTCCTGTTTCCGGGTCAGGGCTCCGGTACCTCCACCGCTGGCGGCGCTCTGGCACGTCGCTTTACTGAAGCAGCAGAAGTG
TATGCACGTGCTGGTTTACCTACTGCAGGTGACATGGTTGCTACCCATGTTGCTCAGCCACGTATCGTTACCGGTTCGACCGCTGGTTTGCGC
GTGCTGGAAGCGCTCGGCATCGAAGCTGATATCGCGCTGGGTCATTCCCTGGGCGAACTGTCTGCGCTGCACTGGGCCGGTGCACTGGATGA
AACTACCCTCCTGGAAGCGGCCCGCACGCGCGGCGCGGCTATGGCGGCACACTCTGCGTCTGGTACCATGGCTTCCCTGACTGCCACTCCAG
AGGAAGCTGTGCGCTTAGTGGAAGGTCTGCCGGTGGTGATCTCGGGCTACAACGGCCCGCGTCAGACCGTAGTAGCCGGGACTGTGGAAGC
GGTTGAATCCGTTGGCGAGCGCGCGGCGGCCGCTGAGATTGCGTTCACCCGTTTAGCGGTTAGCCACGCGTTCCATAGCCCGCTGGTAGCTC
CGGCTGCCGAATCCTTTGGTGACTGGCTCGCGAAAGCACCGCTGGGTGGTCTGGGCCGTCGCGTAGTTTCCACCGTGACGGGCGCTGAACTG
GAGCGTGACACAGATCTGGCTAAACTCCTTCGTCAACAGATTACCGACCCGGTCTTATTTACCCAGGCGGTTCGTGCGGCTGCCGCGGAAGTA
GACCTGTTCGTTGAAGTTGGCCCAGGTCGTGTCCTGAGCGTTCTGGCTGCAGAAACCGCGGGTAAACCGGCGGTTGCGTTGAATACTGACGA
TGAATCTCTGCGCGGTCTGCTGCAGGTTGTTGGCGCTGCGTTCGTAATCGGCGCCCCGATCATTCACGAGCGTCTGTTCAATGATCGCCTGAC
TCGCCCGTTAGAAGTAGGCAAAGAATTCCTGTTTCTGTCAAGCCCGTGTGAACAGGCGCCGGAATTTACCCTGCCGGCAGCGGCTCGCGAAC
CGCTGGTGCAGGAGCATGACGCTCCCACCACCGCTGGCGCTGGTGACACTGCTGAAGAATCTGCCCTGGACGTCCTGCGTGCGCTGGTTGCG
GAGCGTGCTGAACTGCCGTCTGAGCTGATCGATGAAAATTCCTCCCTGCTGGACGATCTGCACATGTCGTCTATCACTGTTGGCCAGATTGTT
AACCAGACCGCAGTGCGTCTGGGTCTGGCACCGTCCAGCATCCCGACCAACTTCGCTACCGCGACCCTGGCTGAACTGGCGTCCGCGCTGAC
TACTTTGGTCGAAACCGGCGCGGATCCGACTGCTGCTCCGGTTGTAACGGGTTCCGCGGCGTGGGCCCGTCCTTTCTCTGTCGATCTGGACGA
ATTACCACTGCCGCCGGCGGTGGCTGATGAAAAGGACGGCACTTGGGAATTGTTTACCTCTGCTGATCACCCGTTCGCTGAAGAAGTTCGTC
GTGCTCTGCAGGACGCGGCTGTAGGTTCTGGTGTTCTAGTCTGCCTGCCGGCTGGCTGCTCTCCGGACCAGCTGGAACTGGCTTTAGATGGC
GCACGTAGCGCACTCGCGGGTTCTCAGGAAGGCCGTTTCGTGCTGGTTCAGCATGATCGTGGTGCTGCTGGTCTGGCAAAAACTCTGCACCT
GGAAGCCCCGCACCTCCGCACTACCGTGGTTCACACCCCGGTAGCTGACGGTGCTGCTGACCGTGTTGCCGCGGAAGTGGCGGCGACTACCC
ACTTTTCTGAAGTTCACCTGGACCGTGATGGTACCCGTCGTGTTCCGGTGCTGCGTGCGCTGCCGTTCGCACCGGACCGTACTGACCAGGTTC
TGGGTCCGGATGACGTTCTGCTGGTTACCGGTGGTGGTAAAGGCATCACTGCTGAATGTGCTCTGGCAGTTGCTGAACGTACCGGTGCGGCT
CTGGCGGTGCTGGGCCGTTCTGATCCGGGCTCTGACCAGGATCTGGCTGCGAACCTGGGTCGTATGCGTGAGTCCGGTATTCGCGTTGCGTA
CGCGGCCGCTGATGTGACCGATCCGGTCCGTGTTGCGGGCGCCGTTGCTGAACTGACTGGTGCACTGGGCAGCGTTACCGCTGTTCTGCACG
GCGCAGGTCGTAACGAACCGACCGCGCTGGGTGGCCTGGATATGGCAGCGGTGCGCTCGACTCTGGCACCGAAAGTTGATGGCCTGCGTCA
CGTGCTCGACGTTGTAGGTGAACAGAACCTGCGTCTGCTTGTTACCTTCGGTTCTATCATTGGTCGCGCTGGCCTCCGTGGCGAAGCGCACTA
CGCTACCGCTAACGAATGGCTGGCAGGCCTGACCGAGGATGTTGCACGTCGTAACCCGGACTGTCGTGCACTGTGCATGGAATGGAGCGTG
TGGTCTGGTGTTGGTATGGGTGAAAAACTGTCCGTAGTTGAATCTTTGTCCCGTGAGGGTATCGTTCCGGTTTCTCCGGATCAGGGTATCGAA

ATCCTGCTGCGCCTGATCTCCGACCCGGACGCTCCAGTAGTGACCGTTATCAGCGGTCGTACCGAAGGTATCGGTACTGTTCGTCGTGAGCAG
CCGCCGCTCCCGCTGCTGCGCTTCACCGGTGAACCGCTGGTTCGCTACCACGGTGTTGAACTGGTTACCGAAGCGGAACTGAACGCAGGCAC
TGATCTGTATCTGACCGACCACATGCTTGATGGCAACCTGCTCCTGCCGGCAGTGATTGGTATGGAAGCTATGGTTCAGGTTGGCTCTGCGGT
TACTGGCCGTCGTGACGTACCGGTCATTGAAGACGCTCGCTTCCTGCGTCCGATTGTTGTGCCACCGGGCGGTACCACTCGTATCCGTATCGC
CGCCACCGTTACCGGTACCGATCGTGTTGACGTTGCGGTTCACGCCCAGGACACCGGTTTTGCGGCTGAACACTTCCGCGCTCGTCTGGTATA
CGGCGGCGCAGCGATCCCGGATGGTGCGCCGGACCAGGTGGGCCCGAAAGTACCGACCGCACCGCTGGATCCGGCGACTGATCTGTATGGT
GGTGTGCTGTTCCAGGGGGAACGCTTCCAGCGTCTGCGTCGTTTCCATCGTGCTGCGGCACGTCACGTGGATGCGGAAGTCGCACTGGACAC
CGCTAGCGGCTGGTTCGCGGGTTTTCTGCCGGGCACTCTGCTGCTCTCTGATCCGGGTATGCGTGACGCTCTGATGCACGGGAACCAGGTTT
GCGTGCCGGACGCAACCCTGCTGCCAAGCGGCATCGAACGTCTGTACCCGATGGCGGCTGGCGAAGATCTGCCGGAACTGGTTCGCTATTGC
GCAACTGAACGTCATCGCGACGGCGACACCTACGTGTACGACATCGCGGTTCGTACCCCGGACGGTTCTGTAGTTGAACGTTGGGACGGTCT
GACGCTGCACGCTGTACGTAAAAGCGACGGTTCCGGCCCATGGGTGGCTCCGCTGCTGGGTTCCTACCTGGAACGTACTCTGGAAGAAGTTC
TGGGCACCCACGTTGATGTTGCAGTGGAGCCGGTTCCGGCTGATAGCGGTGGTAGCGTTGCTGACCGTCGTAAAGCGACCGCCCGTGCAGTT
CAGCGTGCGCTGGGTGAATCCGTTAAAGTGCGTTATCGCCCGGATGGTCGTCCGGAACTGGACGGCGTTCGTCGTCTGAGCGCGGCGCACG
GCCCAGGCGTGACCCTGGGCGTAGTCGGTACTACCACTGTAGCATGCGACATCGAAGCCGTGACCGCACGTGGCGCTCAGGAATGGGAAGG
TCTGCTGGGTGAACACGGTAACCTGGCAGCTCTGGTAGCGAAAGAAACCGGTGAAACCCCAGATCACGCGGCGACCCGTGTTTGGACGGCG
GTTGAATGTCTGAAAAAAGCTGGTCTGCCGGCGGGCGCACCGCTGACCCTGGAACCGCAGGTACGCTCCGGTTGGATCGTGCTGACCGCGG
GTGGTCTGCGCATCGCCACCTTTGCGACCACCCTGCGTCACGTTGAAGAACCGGTAGTGCTGGCGTTCCTGACCGCAGGCACTGATGATGCT
GCGCCGGGCTCTGCTCGTGCTTAAAAGCTT

TE_{SgcE}

CATATGACCGCGACCAACCCGGATTACTTCGAACTGCGTCACACCGTTGGTTTCGAAGAAACCAACCTGGTTGGTAACGTTTACTACGTTAAC
TACCTGCGTTGGCAGGGTCGTTGCCGTGAACTGTTCCTGAAAGAACGTGCGCCGTCTGTTCTGGCTGAAGTTCAGGAAGACCTGAAACTGTTC
ACCCTGAAAGTTGACTGCGAATTCTTCGCTGAAATCACCGCGTTCGACGAACTGAGCATCCGTATGCGTCTGTCTGAACTGCGTCAGACCCAG
CTGGAATTCACCTTCGACTACATCAAACTGGGTGACGACGGTGGTGAAACCCTGGTTGCGCGTGGTCGTCAGCGTATCGCGTGCATGCGTGG
TCCGAACACCGCGACCGTTCCGACCCTGATCCCGGAAGCGCTGGCGGAAGCTCTGGCGCCGTACTCTGACCGTGCGGGTTCTTACGCGGGTC
GTGCGGCTTAAAAGCTT

3.6.5. Supplementary References

1. Peter, D.M., Vogeli, B., Cortina, N.S. & Erb, T.J. A Chemo-Enzymatic Road Map to the Synthesis of CoA Esters. *Molecules* **21**, 517 (2016).
2. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
3. Schada von Borzyskowski, L. et al. Marine Proteobacteria metabolize glycolate via the beta-hydroxyaspartate cycle. *Nature* **575**, 500-504 (2019).
4. Kitagawa, M. et al. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**, 291-9 (2005).
5. Korman, T.P., Opgenorth, P.H. & Bowie, J.U. A synthetic biochemistry platform for cell free production of monoterpenes from glucose. *Nat Commun* **8**, 15526 (2017).
6. Chen, X., Guo, Z.F., Lai, P.M., Sze, K.H. & Guo, Z. Identification of a nonaketide product for the iterative polyketide synthase in biosynthesis of the nine-membered enediyne C-1027. *Angew. Chem. Int. Ed.* **122**, 8098-8100 (2010).
7. Liu, Q. et al. Engineering an iterative polyketide pathway in *Escherichia coli* results in single-form alkene and alkane overproduction. *Met. Engg.* **28**, 82-90 (2015).
8. Erb, T.J., Retey, J., Fuchs, G. & Alber, B.E. Ethylmalonyl-CoA mutase from Rhodobacter sphaeroides defines a new subclade of coenzyme B12-dependent acyl-CoA mutases. *J Biol Chem* **283**, 32283-93 (2008).
9. Erb, T.J. et al. Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway. *Proceedings of the National Academy of Sciences* **104**, 10631-10636 (2007).
10. Konneke, M. et al. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proc Natl Acad Sci U S A* **111**, 8239-44 (2014).
11. Zarzycki, J. et al. Mesaconyl-coenzyme A hydratase, a new enzyme of two central carbon metabolic pathways in bacteria. *J Bacteriol* **190**, 1366-74 (2008).
12. Erb, T.J., Frerichs-Revermann, L., Fuchs, G. & Alber, B.E. The apparent malate synthase activity of Rhodobacter sphaeroides is due to two paralogous enzymes, (3S)-Malyl-coenzyme A (CoA)/{beta}-methylmalyl-CoA lyase and (3S)- Malyl-CoA thioesterase. *J Bacteriol* **192**, 1249-58 (2010).
13. Hoelsch, K., Suhrer, I., Heusel, M. & Weuster-Botz, D. Engineering of formate dehydrogenase: synergistic effect of mutations affecting cofactor specificity and chemical stability. *Appl Microbiol Biotechnol* **97**, 2473-81 (2013).
14. Erb, T.J. PhD Thesis. *University of Freiburg* (2009).
15. Söhling, B. & Gottschalk, G. Purification and characterization of a coenzyme-A-dependent succinate-semialdehyde dehydrogenase from *Clostridium kluyveri*. *Eur. J. Biochem.* **212**, 121-127 (1993).
16. Sevinc, M.S., Ens, W. & Loewen, P.C. The cysteines of catalase HPII of Escherichia coli, including Cys438 which is blocked, do not have a catalytic role. *Eur J Biochem* **230**, 127-32 (1995).
17. Nocek, B. et al. Polyphosphate-dependent synthesis of ATP and ADP by the family-2 polyphosphate kinases in bacteria. *Proc Natl Acad Sci U S A* **105**, 17730-5 (2008).
18. Nicholls, D.J. et al. The importance of arginine 102 for the substrate specificity of Escherichia coli malate dehydrogenase. *Biochem Biophys Res Commun* **189**, 1057-62 (1992).
19. Durbecq, V. et al. Crystal structure of isopentenyl diphosphate: dimethylallyl diphosphate isomerase. *EMBO J.* **20**, 1530-1537 (2001).

4. Enhancing the synthetic capabilities of a complex in vitro metabolic network through anaplerotic reaction modules

Christoph Diehl[1*], Patrick D. Gerlinger[1*], Nicole Paczia[2], Tobias J. Erb[1,3]

[1] Department of Biochemistry & Synthetic Metabolism, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[2] Core Facility for Metabolomics and Small Molecule Mass Spectrometry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

[3]SYNMIKRO Center of Synthetic Microbiology, Marburg, Germany

[*]Christoph Diehl and Patrick D. Gerlinger contributed equally to this work

Author contributions

P.D.G. and C.D. conceived the work. P.D.G. and C.D. designed and performed the experiments. N.P. and C.D. developed the LC-MS methods for the analysis of glycolate, malate and CoA esters. N.P. and P.D.G. developed the LC-MS methods for PKS analysis. T.J.E. supervised and directed the work**. P.D.G., C.D. and T.J.E. wrote the manuscript with contributions from all authors.**

4.1. Abstract

Anaplerosis is an essential feature of metabolism that allows the robust operation of central metabolic pathways (e.g. the citric acid cycle) by constantly replenishing drained intermediates. However, this fundamental concept has not been applied to the construction of synthetic metabolic networks so far. Here, we sought to employ this strategy to the CETCH cycle, a new-to-nature *in vitro* CO_2-fixation pathway that features several C3-C5 biosynthetic precursors. We drafted four different anaplerotic reaction modules, which allow the direct use of CO_2 to replenish the cycle's intermediates and demonstrate functionality of our designs by producing 6-deoxyerythronolide B (6-dEB), the C21 macrolide backbone of erythromycin. Our best design allowed the carbon-positive synthesis of 6-dEB via 54 reactions catalyzed by more than 30 enzymes *in vitro*, notably at yields comparable to the isolated polyketide synthase. Overall, this work showcases how anaplerotic modules can be tailored to enhance and expand the synthetic capabilities of complex catalytic reaction networks.

4.2. Introduction

Synthetic biology aims at creating biological parts and systems that do not exist in nature. This includes the design and realization of new-to-nature enzymes and metabolic networks, which allow to expand the biochemical capabilities of metabolism beyond those developed by natural evolution[1,2]. Design and realization of synthetic pathways for the capture and conversion of carbon dioxide (CO_2) that are more efficient than natural photosynthesis are of particular interest[3,4]. A prominent example is the CETCH cycle, a synthetic CO_2 fixing *in vitro* reaction network. It requires 20% less energy than the Calvin cycle[5] and features the reductive carboxylation of enoyl-CoA esters, a recently discovered CO_2 fixation principle that is an order of magnitude more efficient than to RubisCO[6,7].

While new-to-nature pathways offer multiple opportunities to access novel products as well as more efficient biosynthetic routes[8-10], the properties and biosynthetic capabilities of such designer networks are still lacking behind those of naturally evolved metabolic networks. Natural pathways operate robustly in the context of living cells and enable the flexible (re-)distribution of metabolic flux depending on the biosynthetic needs of the cell. This is in stark contrast to their synthetic counterparts that are typically limited in metabolic flexibility and adaptability, especially in an *in vitro* setup[11].

In case of the CETCH cycle, one of the shortcomings is that this synthetic pathway is restricted to only one dedicated output reaction that yields the C2 compound glyoxylate as primary product of two CO_2 fixations

per turn of the cycle. It has been shown recently that glyoxylate can be further converted into acetyl-CoA and fuel the biosynthesis of different high value products, including several mono- and sesquiterpenes *in vitro*[12]. Yet, to harness the full potential of the CETCH cycle, it would be necessary to access also the cycle's core sequence intermediates. Being able to use different C3-, C4-, and C5-Coenzyme A thioesters for biosynthetic purposes would allow to turn the CETCH cycle into a versatile biosynthetic platform that could feed into various biosynthetic routes. However, removing core intermediates without re-filling them would quickly drain the pool of acceptor molecules required to keep CO_2 conversion running and inevitably lead to a stalling of the CETCH cycle.

One fundamental building principle of naturally evolved metabolic networks is anaplerosis, i.e., reactions or reaction sequences that continuously replenish those intermediates of central carbon metabolism that are directed away into different biosynthetic routes, thereby allowing for a robust and dynamic operation of metabolic networks[13-16]. The defining example is the citric acid cycle, which acts as turntable of cellular metabolism and is constantly refilled by multiple reaction sequences, such as (phosphoenol)pyruvate carboxylase, malic enzyme, and the glyoxylate cycle[17-22]. Consequently, to build synthetic (*in vitro*) metabolic networks and complex biocatalytic reaction cascades that match the flexibility and adaptability of natural metabolism, it will be essential to include anaplerosis as a fundamental design principle into the design of new-to-nature metabolic systems.

Here we sought to expand the biosynthetic capabilities of the CETCH cycle beyond its output molecule glyoxylate by developing anaplerotic reaction sequences to use otherwise non-accessible intermediates from the cycle, in particular propionyl- and methylmalonyl-CoA, which serve as extender units in the biosynthesis of natural products, such as polyketides[23]. Inspired by natural metabolic routes, we designed four anaplerotic reaction sequences for the carbon-neutral and carbon-positive conversion of glyoxylate into different intermediates of the CETCH cycle. We reconstructed the different pathways, optimized their performance and tested their ability to support the biosynthesis of the polyketide 6-deoxyerythronolide B (6-dEB)[24,25], the macrolide backbone of erythromycin directly from CO_2 via the CETCH cycle.

Overall, implementing the concept of anaplerosis into a complex *in vitro* metabolic network of more than 50 different reactions, enabled us to operate this system without the need to provide additional substrates to directly synthesize complex molecules from CO_2. Our work represents a stepping-stone towards the realization of biocatalytic cascades mimicking the properties and intricacies of the natural metabolic networks of living cells.

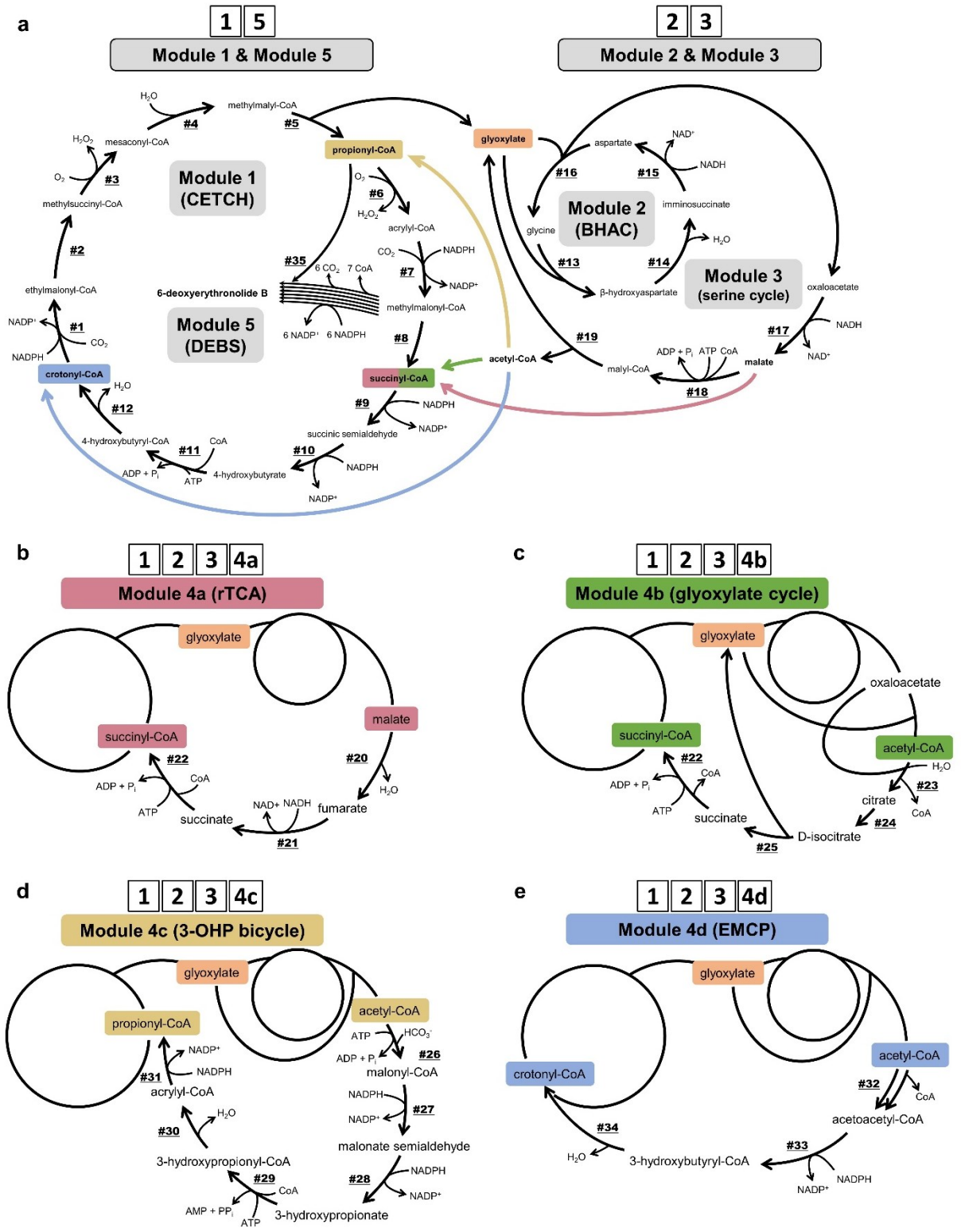


Figure 1. Overview scheme for all reaction cascades presented in this study. ***a*** *The core reactions from CETCH (module 1), BHAC (module 2) and partial serine cycle (module 3) for the conversion of glyoxylate as CETCH output molecule into acetyl-CoA. Additionally shown are DEBS (module 5) and the anaplerotic reaction sequences (colored arrows), which are elaborated in detail* ***b-e****, indicating the point of re-entry into module 1 as well as other bifurcations in module 4b.*

4.3. Results

Reconstitution of the CETCH cycle (module 1)

The CETCH cycle is revolving around two reductive carboxylation reactions (#1, #7) that catalyze the two carbon extension steps of the cycle. Starting from the C3-compound propionyl-CoA, the C4-metabolite methylmalonyl-CoA is formed (#6-7) and further converted into the C5-molecule ethylmalonyl-CoA through a series of reactions (#8-12, #1). Ethylmalonyl-CoA is subsequently transformed into methylmalyl-CoA that is cleaved into glyoxylate and propionyl-CoA (#2-5), the latter can enter another round of the CETCH cycle, while the former remains as the primary CO_2 fixation product and output molecule of the CETCH cycle.

To establish the CETCH cycle *in vitro* (module 1, Figure 1) we modified its setup compared to its initial description (CETCH version 5.4)[5]. To circumvent the use of externally added acetyl-CoA, we replaced the malate-readout of the original cycle by a glycolate-based readout, using glyoxylate reductase, which irreversibly converts glyoxylate into glycolate. In addition, we used a creatine phosphokinase-based ATP regeneration system instead of polyphosphate kinase. When starting the reaction with 100 µM propionyl-CoA, the optimized cycle produced 381 ± 6.6 µM glycolate within 90 minutes. This translates into 7.6 ± 0.2 fixed CO_2 molecules per starting acceptor molecule (propionyl-CoA) and 3.8 cycle turnovers, which is 1.4-fold better compared to the recently published setup (Figure 2a)[5]

Establishing modules 2 & 3 for glyoxylate conversion

To convert glyoxylate – the output molecule of module 1 (i.e. CETCH cycle) – back into intermediates of the core cycle, we sought to establish downstream reaction modules, which would transform glyoxylate into oxaloacetate, malate and/or acetyl-CoA. These compounds could then serve as starting points for different anaplerotic reaction sequences. To that end, we aimed at employing the beta-hydroxyaspartate cycle (BHAC) from *Paracoccus denitrificans* that transforms two molecules glyoxylate through amino-group cycling and consumption of one NADH into oxaloacetate (module 2, Figure 1a)[12,26,27]. Extending the BHAC by an additional reaction sequence, involving malate dehydrogenase (Mdh, #17), malate:thiokinase (Mtk, #18) and malyl-CoA lyase (Mcl, #5, #19) would allow the further conversion of oxaloacetate into glyoxylate, which could reenter the BHAC, and acetyl-CoA (module 3, Figure 1a)[12] .

However, one crucial reaction in module 2 (i.e. BHAC) is the immediate reduction of the labile intermediate imminosuccinate (#14) to prevent its spontaneous hydrolysis into oxaloacatetate and concomitant loss of the amino group for recycling[27]. To verify that the concentration of imminosuccinate

reductase (Isr, #15) would be sufficient to drive full reduction of iminosuccinate to aspartate, we tested different glyoxylate concentrations ranging from 0 to 1000 µM, while fixing the concentration of the amino-donor molecule glycine at 250 µM (Figure 2b). As readout, we used malate dehydrogenase (Mdh)[28], which converts oxaloacetate into malate. Reaching conversions close to the expected theoretical yields in all assays showed that module 2 was indeed fully active without significant hydrolysis (Figure 2b, Figure S2).

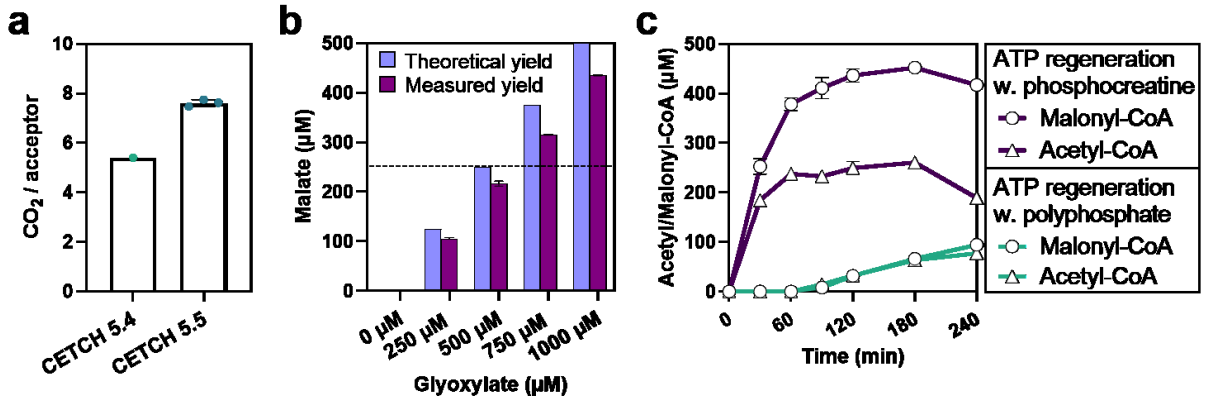


Figure 2. Optimization of module 1-3. a *CETCH efficiency. CO_2/acceptor shows how many CO_2 molecules per molecule of initially added substrate (propionyl-CoA) were fixed. Data for CETCH 5.4 is taken from Schwander et al.. Compared to the publication the CETCH 5.5 contains several minor adaptations (see reconstitution of the CETCH cycle).* ***b*** *Substrate concentrations to determine BHA cycle efficiency and amino group recycling. The reactions were started with either 0, 250, 500 750 or 1000 µM glyoxylate while the glycine concentration is fixed at 250 µM, stopped after 60 minutes and measured by LC-MS. The theoretical maximum yield of full glyoxylate conversion is displayed in light blue, the measured values as dark purple (86 % ± 1 % conversion). The dashed line indicates the threshold above which the amino group is obligatorily recycled (see Figure S2).* ***c*** *Coupling of the CETCH module with feedback module 2 and 3. All reactions were started with 100 µM propionyl-CoA. Experiments utilizing phosphocreatine for ATP regeneration included 2 U/ml Creatine phosphokinase, the experiments using polyphosphate 0.5 U/ml Polyphosphate kinase. Conversion of acetyl-CoA to malonyl-CoA was ensured by addition of 100mU/ml propionyl-CoA carboxylase D407I (Pcc*, #26). All experiments were performed in technical triplicates.*

Optimizing the interplay of modules 1-3

Coupling modules 1, 2 and 3 yielded approximately 100 µM acetyl-CoA (Figure 2c), when starting from 100 µM propionyl-CoA, indicating one complete turnover through the combined modules 1-3. As already observed during optimization of module 1, switching to a creatine phosphokinase-based ATP regeneration significantly improved acetyl-CoA yield of modules 1-3 to 260 µM within 3 hours. However, after 30 minutes acetyl-CoA production had already stalled and even started to decrease after 3 hours (Figure 2c). This is explained through the condensation of acetyl-CoA with glyoxylate back into malyl-CoA, when reaching a certain acetyl-CoA threshold concentration due to the reversibility of Mcl (#5)[5,12,29]. We aimed at further improving productivity of the coupled system by constantly withdrawing acetyl-CoA through

carboxylation into malonyl-CoA, using a propionyl-CoA carboxylase variant Pcc* (#26)[30], as reported recently[12]. Indeed, the yield of malonyl-CoA almost doubled when using Pcc* together with the phosphocreatine ATP regeneration system, while a similar effect of Pcc* was not observed in the polyphosphate setup (Figure 2c). Reaching a yield of more than 450 µM of malonyl-CoA, the coupled modules 1-3 were even more productive than module 1 alone: Starting from 100 µM propionyl-CoA, we observed more than 4.5 cycle turnovers of coupled modules 1-3 within 3 hours, and notably 4 cycle turnovers after 90 min, compared to only 3.8 for optimized module 1 alone (Figure 2a). Having established the basic reaction network to convert glyoxylate into oxaloacetate, malate or acetyl-CoA, we moved on to the design and implementation of different anaplerotic feedback modules.

Design of the different anaplerotic modules 4a-d

In the next step, we drafted different anaplerotic pathways to transform oxaloacetate, malate or acetyl-CoA into intermediates of module 1. To that end, we searched for known pathway segments that would allow to regenerate above starting molecules into different C3- or C4-CoA esters. Starting from malate, we identified a reaction series from the reductive TCA cycle[31] that produces succinyl-CoA as re-entry point into module 1 (module 4a, Figure 1b). Starting from acetyl-CoA (and oxaloacetate), we sought to utilize reactions of the glyoxylate cycle[14] that yield succinyl-CoA (module 4b, Figure 1b), reactions of the 3-hydroxypropionate (3-OHP)[32,33] cycle regenerating propionyl-CoA (module 4c, Figure 1b), and reactions of the ethylmalonyl-CoA (EMC) pathway providing crotonyl-CoA (module 4d, Figure 1b)[15,34].

We analyzed the thermodynamic profile of anaplerotic modules 4a-d, to test for thermodynamic feasibility (Figure S1), and established the different reaction sequences in the following. To assess and optimize the performance of the different feedback modules, we decided to quantify the production of methylmalonyl-CoA by modules 1-4[a/b/c/d] starting from glyoxylate. Even though we were still working in a closed-loop system with no direct output molecule, we reasoned that high methylmalonyl-CoA levels could serve as proxy for production yields of 6-dEB, our final benchmark molecule (see below).

Realization of anaplerotic module 4a (reductive TCA)

Feedback module 4a (Figure 1b) is the only C4-conserving pathway, which does not start from acetyl-CoA. This is achieved by branching off module 3 after reduction of oxaloacetate to malate (Figure 1b, #17). Malate is subsequently converted via a fumarate hydratase (Fum, #20) into fumarate, reduced to succinate (Frd, #21) and finally converted into the core cycle intermediate, succinyl-CoA, by a succinate:CoA ligase (Scs, #22).

While two of the three additional enzymes required (Fum and Scs) could be directly deployed from *E.coli*[35,36], finding a suitable candidate for reduction of fumarate proved to be more difficult. Most Frd are quinone-dependent, multi-subunit membrane-bound enzymes, which are oxygen sensitive, excluding their application in an aerobic *in vitro* setup. However, we identified a suitable candidate from *Trypanosoma brucei* mitochondria that is NADH dependent and composed of a single, soluble subunit. Unfortunately, in the absence of fumarate Frd reduces molecular oxygen, thereby generating H_2O_2 and consuming NAD(P)H[37]. While H_2O_2 can be detoxified by catalase (Cat), which is present in the assay at 1.5 U/ml, this substrate-independent side reaction of Frd with oxygen strongly affects the NAD(P)H pool. When testing different concentrations of Frd in the context of the full pathway, starting from 250 µM glyoxylate, we found that 10 mU Frd resulted in the highest yields of methylmalonyl-CoA (49 ± 18.9 µM after 90 minutes). Additional formate to increase NAD(P)H regeneration by Fdh did not improve the yield, even when using 100 mU Frd (Figure 3a).

Developing anaplerotic module 4b (glyoxylate part cycle)

The design of feedback module 4b (1b) resembles the glyoxylate cycle[14]. To synthesize citrate from acetyl-CoA and oxaloacetate (#23), we decided to use citrate synthase from *Synechocystis* sp. PCC 6803 (Cit, #23)[38]. In contrast to citrate synthases from heterotrophic bacteria, *Synechocystis* Cit has an 'inverted' responsiveness towards allosteric effectors, i.e., shows no inhibition by $MgCl_2$, ATP or NADH, and becomes activated by ADP. For the generation of isocitrate from citrate (Acn, #24), we employed AcnA from *E. coli*, which compared to its homologue AcnB, has a lower catalytic efficiency, but is more oxygen tolerant[39]. We selected *E. coli* isocitrate lyase (Icl) to convert isocitrate into glyoxylate and succinate (#25), which can re-enter CETCH cycle as succinyl-CoA via succinate:CoA ligase (#22), as in module 4a.

Note that module 4b uses oxaloacetate (produced by module 2) for the condensation reaction with acetyl-CoA by Cit (#23), as well as for the formation of acetyl-CoA itself via module 3 (#17-19), which leads to a direct competition between Cit and Mdh for oxaloacetate. Due to the more favorable kinetic parameters of Mdh towards oxaloacetate ($K_M \approx 40$ µM, $k_{cat} \approx 930$ s^{-1})[38] compared to Cit ($K_M \approx$ 90 µM, $k_{cat} \approx 2.8$ s^{-1}, $K_{M,\text{acetyl-CoA}}$ = 220 µM)[38], we hypothesized that a successful implementation of module 4b would directly depend on the concentration of Mdh. Indeed, the amount of acetyl-CoA produced when combining modules 1- 4b showed an inverse correlation with the amount of Mdh used (Figure S4a). Indeed, the setup with 7.9 U/ml Mdh led to an accumulation of acetyl-CoA and produced only 6 µM methylmalonyl-CoA, indicating that the oxaloacetate pool was completely drained by Mdh (Figure 3, Figure S4Aa). In contrast, the setup containing 0.8 U/ml Mdh was able to produce almost 60 µM methylmalonyl-CoA, indicating that

sufficient oxaloacetate was available for the Cit reaction to proceed. Therefore, we chose this Mdh concentration for the implementation of module 4b.

Establishing anaplerotic module 4c (3-OHP part cycle)

Feedback module 4c (Figure 1b) is based on reactions of the 3-OHP bi-cycle. We used the aforementioned engineered propionyl-CoA carboxylase (Pcc*, #26)[30] to produce malonyl-CoA from acetyl-CoA. Malonyl-CoA is further reduced via a bifunctional malonyl-CoA/malonate semialdehyde reductase (#27, #28)[40,41], resulting in 3-OHP. Finally, for the reaction of 3-OHP to propionyl-CoA (#29-31), we used propionyl-CoA synthetase (Pcs), a multicatalytic nanocompartment which was characterized recently[42]. The whole module 4c is comprised of only three enzymes catalyzing six reactions, consuming three NADPH and three ATP equivalents, resulting in the highest change of Gibbs free energy ($\Delta_r G'$) compared to the other feedback modules (Figure S1). Initial experiments resulted in the formation of up to 490 µM methylmalonyl-CoA after two hours starting from 250µM glyoxylate (Figure 3c). Addition of adenylate kinase to regenerate the AMP produced by Pcs increased the yield further to more than 700 µM after four hours (Figure 3c).

Realization of anaplerotic module 4d (EMC pathway)

Feedback module 4d comprises a partial EMC pathway (Figure 1b). This module requires only three reactions, yielding crotonyl-CoA, which allows re-entering the CETCH cycle prior to one of the carboxylation steps (#1). Module 4d is by far the most energetically efficient sequence based on the calculated Gibbs free energy change (Figure S1).

To establish the pathway, we employed the acetyl-CoA-acetyltransferase Aat (#32), its cognate reductase Aar (#33)[43] and a putative β-hydroxybutyryl-CoA dehydratase Bbd (#34). Functionality and kinetic parameters of Bbd were determined beforehand (Figure S8a). While Aat showed some promiscuity with propionyl-CoA *in vitro* (Figure S3b), only acetoacetyl-CoA formation was observed in context of the CETCH cycle (Figure S3b). Note that the condensation reaction of Aat is inhibited by free CoA[44] (Figure S3a). To assess the effects of CoA, we systematically varied the initial amount of free CoA up to 1 mM in the assay, when testing the full feedback module starting from 250 µM glyoxylate. Interestingly, and contrary to earlier observations with the isolated enzyme, where 1 mM of CoA inhibited the condensation reaction of Aat almost completely (Figure S3a), 1 mM CoA showed the highest yields of methylmalonyl-CoA when introduced into the whole pathway (Figure S4b), but also caused the highest accumulation of acetyl-CoA (Figure S4c). To test whether higher CoA concentrations further increase yield, we repeated the

experiment with 1 mM or 2.5 mM CoA. Using 2.5 mM CoA reduced the production of methylmalonyl-CoA, indicating that the level of free CoA in module 4d is optimal around 1 mM (Figure 3d).

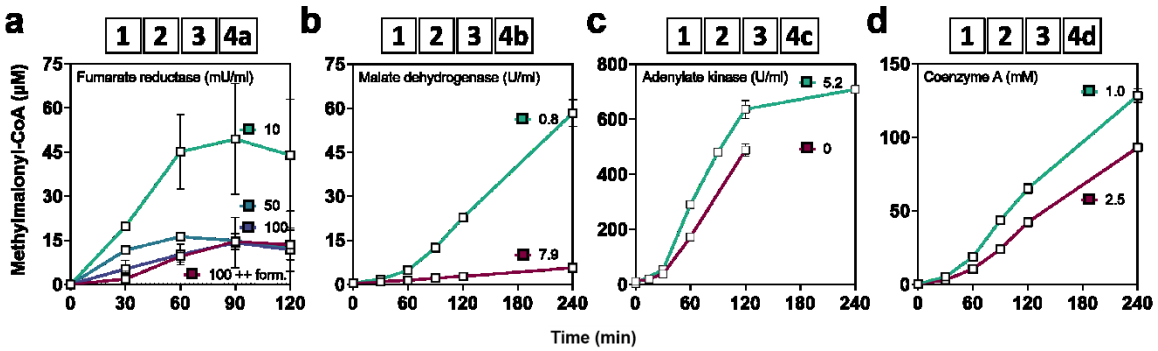


Figure 3. Methylmalonyl-CoA accumulation enabled by anaplerotic feedback pathways with varying conditions. *All conditions contained Modules 1, 2 & 3.* ***a*** *Module 4a with different amounts of Frd.. +form.indicates usage of 100 mM formate, whereas all other experiments were carried out with 50 mM formate.* ***b*** *Module 4b with different amounts of Mdh.* ***c*** *Module 4c with or without Adk.* ***d*** *Module 4d with different amounts of CoA. All experiments were performed in technical triplicates.*

Anaplerotic feedbacks allow efficient 6-dEB production from CO_2

To test our anaplerotic modules in a complex biosynthetic scenario, we next aimed at assessing the performance of the different pathway versions for the biosynthesis of 6-dEB. 6-dEB is synthesized by the 6-deoxyerythronolide B synthase (DEBS), a type I polyketide synthase (PKS) that uses propionyl-CoA as a starter unit and six methylmalonyl-CoA as extender units per molecule of 6-dEB (module 5, Figure 1a, Figure S5). The polyketide is synthesized via six subsequent decarboxylative Claisen condensations, accompanied by release of CoA and the consumption of six NADPH reducing equivalents[45]. To quantify 6-dEB production, we correlated the linear range of polyketide production from a positive control with measured reduction rates[25] (Figure S6).

Before quantifying 6-dEB production, we verified and compared methylmalonyl-CoA production of our four optimized pathways without module 5 (i.e. DEBS). When testing modules 1-4[a/b/c/d], all four feedback routes accumulated methylmalonyl-CoA at comparable levels observed earlier (Figure 3 and 4a). The different pathway networks were able to convert 250 µM glyoxlyate into methylmalonyl-CoA with an effective carbon conversion of 104% (modules 1-4a), 88% (modules 1-4b), 398% (modules 1-4c) and 55% (modules 1-4d), respectively (Figure 4a, Table S7). Thus, (at least) two out of four anaplerotic pathways were carbon-positive in respect to methylmalonyl-CoA yield under these conditions.

When testing the different pathways together with DEBS (modules 1-4[a/b/c/d]-5), three out of the four anaplerotic modules yielded detectable amounts of 6-dEB, in contrast to the control, i.e. module 1+5 without any anaplerotic reaction sequence (Figure 4b). Interestingly, only the setup with module 4d did not show detectable polyketide production, even though modules 1-4d had produced methylmalonyl-CoA at relevant concentrations in the absence of module 5 before (see above, Figure 4a)[46]. Although we did not detect 6-dEB for modules 1+5 and modules 1-4d-5, we noticed that the profiles of CoA esters shifted between the setups with and without module 5, indicating that the presence of DEBS had influenced carbon flux in the systems, probably by depleting the system quickly from methylmalonyl-CoA (Figure 4 c-e).

The setups with feedback module 4a and 4b yielded 6.0 ± 0.2 µM and 2.8 ± 1.0 µM 6-dEB, respectively, while the setup with module 4c produced 31.9 ± 1.6 µM 6-dEB (Figure 4b). Synthesis of one molecule of 6-dEB requires one C3-CoA (propionyl-CoA) and six C4-CoAs (methylmalonyl-CoA). Thus, the production of 32 µM 6-dEB through modules 1-4c-5 was carbon-positive, with an effective carbon conversion of 172% (Table S7), indicating that this system had successfully captured CO_2 into 6-dEB. Notably, the yields of modules 1-4c-5 were comparable to module 5 provided directly with propionyl- and methylmalonyl-CoA (39.1 ± 0.3 µM from 0.8 mM propionyl- and 1 mM methylmalonyl-CoA, respectively, Figure 4b), demonstrating that the anaplerotic feedback module allowed our system consisting of 54 reactions to operate similarly efficient compared to the isolated DEBS alone.

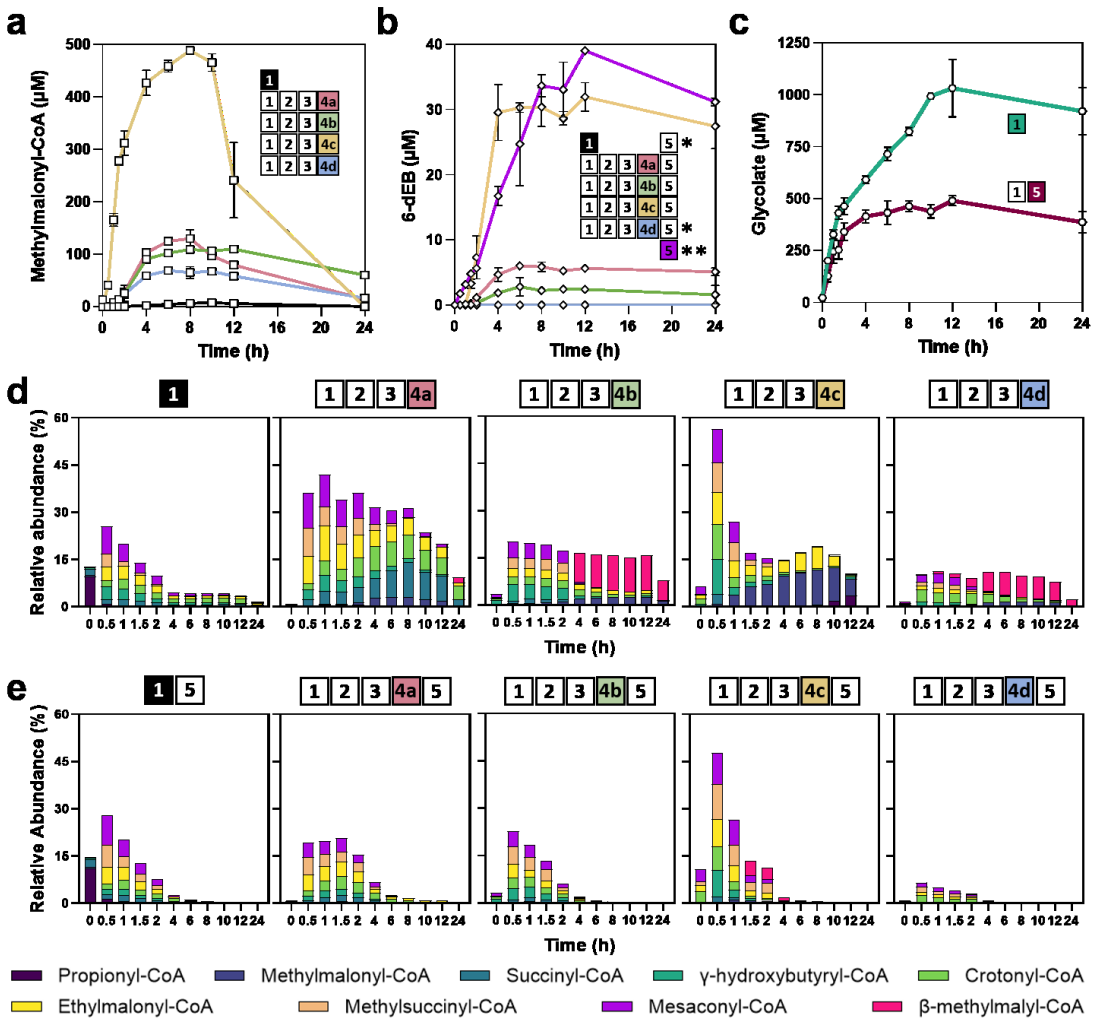


Figure 4. Substrate accumulation and polyketide production by the different anaplerotic pathways. ***a*** *Methylmalonyl-CoA (MM-CoA) accumulation in the different feedback pathways and the CETCH core cycle.* ***b*** *6-dEB production by the different pathways and a control *not detectable **positive control was started with 0.8 mM propionyl-CoA and 1.0 mM methylmalonyl-CoA.* ***c*** *Glycolate production of module 1 (CETCH core cycle) with and without module 5 (DEBS).* ***d*** *and* ***e*** *show relative CETCH core cycle intermediate formation with (E) or without (D) DEBS. The highest EIC value (from D and E) of each displayed compound was set to 11.11% (9 compounds = 100%) and all other extracted ion counts (EIC) values of that compound were set relative to that. For absolute values of the EIC including the standard error, see Figure S7. All experiments were performed in technical duplicates.*

4.4. Discussion

One of the biggest challenges in contemporary biology and chemistry is to construct synthetic systems that exhibit the complexity and characteristics of naturally existing biological systems. Here we aimed at creating an *in vitro* catalytic network that is able to produce a chemically challenging molecule, 6-dEB, directly from CO_2. This was achieved by developing and subsequently coupling different reaction modules, and in particular different anaplerotic reaction sequences to replenish core metabolites of the network

by CO_2. While the core network alone (CETCH cycle, module 1) failed to synthesize 6-dEB, our anaplerotic feedback modules enabled it to produce the polyketide with up to 172% carbon conversion and notably at yields compared to the isolated polyketide machinery *in vitro*. Our results highlight the importance of anaplerotic reaction sequences, not only for the robust operation of natural, but also synthetic catalytic networks. While we focused in our work on 6-dEB as model product, we note that many other complex molecules, such as other polyketides and polymers (polyhydroxyalkanoates), could in principle be derived from the core cycle augmented with our anaplerotic modules.

What determines productivity of anaplerotic reaction modules? One important aspect is their energetic requirements, i.e. their Gibbs free energy profile (Figure S1, module 4c > 4b > 4a > 4d), but probably also their re-entry point into the core network (CETCH cycle, module 1). Module 4c, which is the most efficient (but also energetically most expensive) reaction sequence, directly yields propionyl-CoA, the starter unit of DEBS, that can be converted by just two additional steps into methylmalonyl-CoA, the extender unit of DEBS. Modules 4a and b, on the other hand, yield succinyl-CoA, and therefore require another 11 and 13 enzymatic reactions, respectively, to arrive at the same metabolites, potentially subjected to more kinetic and enzymatic bottlenecks. As an example, Frd in module 4a has a high side reactivity with NAD(P)H and oxygen as an alternative electron acceptor, likely depleting the pools of NAD(P)H, affecting the activity of all NAD(P)H dependent enzymes of the whole reaction network. This is indicated by accumulation of reaction substrates of NAD(P)H-dependent enzymes (e.g. succinyl-CoA and crotonyl-CoA, Figure 4d). Further protein engineering efforts or screening of alternative homologs could yield variants overcoming the deficiencies of Frd or any other enzymes constraining the reaction network.

Overall, the successful coupling and simultaneous operation of up to 54 reactions provides a first step towards the creation of dynamic, yet robust *in vitro* catalytic networks. Considering future approaches of building complex catalytic systems, we note that while using anaplerotic reaction sequences provides more robustness and flexibility to catalytic networks, more and additional layers of regulation will be required to achieve the intricate design of natural metabolic networks. This includes the allosteric control and/or compartmentalization of reactions, as well as layers of translational regulation to dynamically regulate catalytic networks. Approaches using cell-free transcription-translation systems and recent efforts to couple synthetic metabolism to light-controlled energy modules might provide the requirements to establish such exquisite control in the future, paving the way for further efforts that make use of complex enzymatic cascades in biology and chemistry[47].

4.5. References

1. Bilgin, T. & Wagner, A. Design constraints on a synthetic metabolism. *PloS one* **7**, e39903 (2012).
2. Erb, T.J., Jones, P.R. & Bar-Even, A. Synthetic metabolism: metabolic engineering meets enzyme design. *Current opinion in chemical biology* **37**, 56-62 (2017).
3. Bar-Even, A., Noor, E., Lewis, N.E. & Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences* **107**, 8889-8894 (2010).
4. Gong, F., Zhu, H., Zhang, Y. & Li, Y. Biological carbon fixation: from natural to synthetic. *Journal of CO2 Utilization* **28**, 221-227 (2018).
5. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
6. Erb, T.J., Brecht, V., Fuchs, G., Müller, M. & Alber, B.E. Carboxylation mechanism and stereochemistry of crotonyl-CoA carboxylase/reductase, a carboxylating enoyl-thioester reductase. *Proceedings of the National Academy of Sciences* **106**, 8871-8876 (2009).
7. Peter, D.M. et al. Screening and engineering the synthetic potential of carboxylating reductases from central metabolism and polyketide biosynthesis. *Angewandte Chemie International Edition* **54**, 13457-13461 (2015).
8. Valliere, M.A., Korman, T.P., Arbing, M.A. & Bowie, J.U. A bio-inspired cell-free system for cannabinoid production from inexpensive inputs. *Nature Chemical Biology* **16**, 1427-1433 (2020).
9. Des Soye, B.J., Gerbasi, V.R., Thomas, P.M., Kelleher, N.L. & Jewett, M.C. A highly productive, one-pot cell-free protein synthesis platform based on genomically recoded Escherichia coli. *Cell chemical biology* **26**, 1743-1754. e9 (2019).
10. Bogorad, I.W., Lin, T.-S. & Liao, J.C. Synthetic non-oxidative glycolysis enables complete carbon conservation. *Nature* **502**, 693-697 (2013).
11. Claassens, N.J., Burgener, S., Vögeli, B., Erb, T.J. & Bar-Even, A. A critical comparison of cellular and cell-free bioproduction systems. *Current opinion in biotechnology* **60**, 221-229 (2019).
12. Sundaram, S. et al. A modular in vitro platform for the production of terpenes and polyketides from CO2. *Angewandte Chemie International Edition* (2021).
13. Kornberg, H. Anaplerotic sequences in microbial metabolism. *Angewandte Chemie International Edition in English* **4**, 558-565 (1965).
14. Kornberg, H. The role and control of the glyoxylate cycle in Escherichia coli. *Biochemical Journal* **99**, 1 (1966).
15. Erb, T.J. et al. Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway. *Proceedings of the National Academy of Sciences* **104**, 10631-10636 (2007).
16. Borjian, F., Han, J., Hou, J., Xiang, H. & Berg, I.A. The methylaspartate cycle in haloarchaea and its possible role in carbon metabolism. *The ISME journal* **10**, 546-557 (2016).
17. Kornberg, H. & Krebs, e.H. Synthesis of cell constituents from C 2-units by a modified tricarboxylic acid cycle. *Nature* **179**, 988-991 (1957).
18. Cronan Jr, J.E. & Laporte, D. Tricarboxylic acid cycle and glyoxylate bypass. *EcoSal Plus* **1**(2005).
19. Jitrapakdee, S. et al. Structure, mechanism and regulation of pyruvate carboxylase. *Biochemical journal* **413**, 369-387 (2008).
20. Gokarn, R., Eiteman, M. & Altman, E. Metabolic Analysis of Escherichia coliin the Presence and Absence of the Carboxylating Enzymes Phosphoenolpyruvate Carboxylase and Pyruvate Carboxylase. *Applied and Environmental Microbiology* **66**, 1844-1850 (2000).
21. Stols, L. & Donnelly, M.I. Production of succinic acid through overexpression of NAD (+)-dependent malic enzyme in an Escherichia coli mutant. *Applied and Environmental Microbiology* **63**, 2695-2701 (1997).

22. Kwon, Y.D., Kwon, O.H., Lee, H.S. & Kim, P. The effect of NADP-dependent malic enzyme expression and anaerobic C4 metabolism in Escherichia coli compared with other anaplerotic enzymes. *Journal of applied microbiology* **103**, 2340-2345 (2007).
23. Chan, Y.A., Podevels, A.M., Kevany, B.M. & Thomas, M.G. Biosynthesis of polyketide synthase extender units. *Natural product reports* **26**, 90-114 (2009).
24. Caffrey, P., Bevitt, D.J., Staunton, J. & Leadlay, P.F. Identification of DEBS 1, DEBS 2 and DEBS 3, the multienzyme polypeptides of the erythromycin-producing polyketide synthase from Saccharopolyspora erythraea. *FEBS letters* **304**, 225-228 (1992).
25. Lowry, B. et al. In vitro reconstitution and analysis of the 6-deoxyerythronolide B synthase. *Journal of the American Chemical Society* **135**, 16809-16812 (2013).
26. Kornberg, H. & Morris, J. The utilization of glycollate by Micrococcus denitrificans: the β-hydroxyaspartate pathway. *Biochemical Journal* **95**, 577 (1965).
27. Schada von Borzyskowski, L. et al. Marine Proteobacteria metabolize glycolate via the beta-hydroxyaspartate cycle. *Nature* **575**, 500-504 (2019).
28. Nicholls, D.J. et al. The importance of arginine 102 for the substrate specificity of Escherichia coli malate dehydrogenase. *Biochem Biophys Res Commun* **189**, 1057-62 (1992).
29. Erb, T.J., Frerichs-Revermann, L., Fuchs, G. & Alber, B.E. The apparent malate synthase activity of Rhodobacter sphaeroides is due to two paralogous enzymes, (3S)-Malyl-coenzyme A (CoA)/{beta}-methylmalyl-CoA lyase and (3S)- Malyl-CoA thioesterase. *J Bacteriol* **192**, 1249-58 (2010).
30. Schwander, T. Philipps-Universität Marburg (2018).
31. Evans, M., Buchanan, B.B. & Arnon, D.I. A new ferredoxin-dependent carbon reduction cycle in a photosynthetic bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **55**, 928 (1966).
32. Alber, B.E. & Fuchs, G. Propionyl-coenzyme A synthase from Chloroflexus aurantiacus, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO2 fixation. *Journal of Biological Chemistry* **277**, 12137-12143 (2002).
33. Zarzycki, J., Brecht, V., Müller, M. & Fuchs, G. Identifying the missing steps of the autotrophic 3-hydroxypropionate CO2 fixation cycle in Chloroflexus aurantiacus. *Proceedings of the National Academy of Sciences* **106**, 21317-21322 (2009).
34. Alber, B.E., Spanheimer, R., Ebenau-Jehle, C. & Fuchs, G. Study of an alternate glyoxylate cycle for acetate assimilation by Rhodobacter sphaeroides. *Molecular microbiology* **61**, 297-309 (2006).
35. Woods, S.A., Schwartzbach, S.D. & Guest, J.R. Two biochemically distinct classes of fumarase in Escherichia coli. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* **954**, 14-26 (1988).
36. Nolte, J.C. et al. Novel characteristics of succinate coenzyme A (Succinate-CoA) ligases: conversion of malate to malyl-CoA and CoA-thioester formation of succinate analogues in vitro. *Appl Environ Microbiol* **80**, 166-76 (2014).
37. Mracek, J., Snyder, S.J., Chavez, U.B. & Turrens, J.F. A soluble fumarate reductase in Trypanosoma brucei procyclic trypomastigotes. *J Protozool* **38**, 554-8 (1991).
38. Ito, S., Koyama, N. & Osanai, T. Citrate synthase from Synechocystis is a distinct class of bacterial citrate synthase. *Sci Rep* **9**, 6038 (2019).
39. Varghese, S., Tang, Y. & Imlay, J.A. Contrasting sensitivities of Escherichia coli aconitases A and B to oxidation and iron depletion. *Journal of bacteriology* **185**, 221-230 (2003).
40. Hugler, M., Menendez, C., Schagger, H. & Fuchs, G. Malonyl-coenzyme A reductase from Chloroflexus aurantiacus, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO(2) fixation. *J Bacteriol* **184**, 2404-10 (2002).

41. Son, H.F. et al. Structural insight into bi-functional malonyl-CoA reductase. *Environmental Microbiology* **22**, 752-765 (2020).
42. Bernhardsgrutter, I. et al. The multicatalytic compartment of propionyl-CoA synthase sequesters a toxic metabolite. *Nat Chem Biol* **14**, 1127-1132 (2018).
43. Peoples, O.P. & Sinskey, A.J. Poly-β-hydroxybutyrate biosynthesis in Alcaligenes eutrophus H16: characterization of the genes encoding β-ketothiolase and acetoacetyl-CoA reductase. *Journal of Biological Chemistry* **264**, 15293-15297 (1989).
44. Oeding, V. & Schlegel, H.G. Beta-ketothiolase from Hydrogenomonas eutropha H16 and its significance in the regulation of poly-beta-hydroxybutyrate metabolism. *Biochem J* **134**, 239-48 (1973).
45. BEVITT, D.J., CORTES, J., HAYDOCK, S.F. & LEADLAY, P.F. 6-Deoxyerythronolide-B synthase 2 from Saccharopolyspora erythraea: Cloning of the structural gene, sequence analysis and inferred domain structure of the multifunctional enzyme. *European journal of biochemistry* **204**, 39-49 (1992).
46. Dunn, B.J., Cane, D.E. & Khosla, C. Mechanism and specificity of an acyltransferase domain from a modular polyketide synthase. *Biochemistry* **52**, 1839-1841 (2013).
47. Miller, T.E. et al. Light-powered CO2 fixation in a chloroplast mimic with natural and synthetic parts. *Science* **368**, 649-654 (2020).

4.6. Supplementary Information

4.6.1. Materials & Methods

Plasmid construction

Except for Cit, all used plasmids were constructed previously (see Table S9). For construction of a Cit expression plasmid, previously described methods were adapted[1]. In brief, Synechocystis PCC 6803 was grown in BG-11, harvested via centrifugation, lysed through sonication and the cell debris used as a template for PCR amplification of its citrate synthase with the following primers: forward – CAAGGTACCGACTGATAACGAAGTGTTTAAAG, reverse – CTGCGGCCGCTTAAATAATCGCATTGGGGTC. The corresponding product was purified and, together with the target vector pET-51b, cut with FastDigest restriction enzymes (Thermo Scientific) KpnI and NdeI (underscored). Following DNA purification, vector and insert were ligated with a T4 ligase according to protocol (NEB) and the ligation mix transformed into electrocompetent E. coli DH5-α. The final construct (N-Strep Cit) was verified via sequencing (Microsynth).

Production and purification of recombinant proteins

Unless otherwise denoted, all proteins were purified alike. Upon transformation into the *E. coli* expression strain BL21(DE3) (Thermo Scientific) (carrying an additional plasmid for co-expression of the chaperones GroEL-GroES for Hbs), *E. coli* BL21(DE3) Rosetta (Novagen) for Mco and Pco and *E. coli* BAP1[2] for all DEBS plasmids, 2 L of salt buffered TB medium were directly inoculated with colonies from the selection plates and grown on 37°C and 90 rpm till OD_{600} 0.5-1.0. Subsequently, cultures were cooled down to 21°C, induced with 25 µM IPTG and grown overnight. Cultures producing Hbd, where 100 µM of Fe(II)SO_4, 100 µM Fe(III)citrate and 20 mM fumarate were added at induction, grown to an OD_{600} 4 and cooled down in a sterile Schott bottle for protein production under microaerobic conditions. Furthermore, production of Pco was done at 25°C for 4 h. Following cell pellet collection by centrifugation (15 min, 4°C, 6000x g), the cells were resuspended in two parts (w/v) lysis buffer (buffer A, 500 mM NaCl, 50 mM HEPES, 10% glycerol, pH 7.8) and 5 mM $MgCl_2$, 10 µg/ml DNAse and one tablet of Sigma*FAST* Protease Inhibitor Cocktail (Sigma-Aldrich) added. Cell lysis was performed using a microfluidizer (two iterations at 16.000 psi), followed by centrifugation at 50.000 xg for 1 h at 4°C. The supernatant was filtered through a 0.45 µm membrane, mixed with 3 ml preequilibrated (Buffer A) Protino Ni-NTA agarose beads (Macherey-Nagel) and incubated on ice for 30-45 min (70 rpm). Afterwards the beads were collected in a 14 ml gravity column and washed with three column volumes (cv) of lysis buffer, followed by two washing steps with three cv of lysis buffer

containing additional 50 mM of imidazole and three cv with 75 mM imidazole. The elution was done with two cv of lysis buffer containing 500 mM imidazole (buffer B). The elution fractions were concentrated with an Amicon Ultra 15 mL Centrifugal Filters (Merck), possessing an adequate molecular weight cutoff.

All CETCH core enzymes, as well as the propionyl-CoA synthase were desalted on a HiLoad 16/600 Superdex 200 pg column (GE Healthcare). Downstream enzymes for the feedback modules 1, 2a 2b and 2c were desalted with 2 x 5 ml HiTrap® desalting columns (GE Healthcare). For both steps a desalting/storage buffer containing 200 mM NaCl, 50 mM HEPES and 10% glycerol, pH 7.8 (buffer C) was used. For Hbs and Hbd, buffer C contained 500 mM NaCl. All DEBS proteins, except the strep-tagged LD(4) (see below), got additionally separated using a 5 ml Q-Sepharose HiTrap® anion exchange column (GE healthcare), with an 80 ml gradient from buffer D (50 mM HEPES, 10% glycerol, pH 7.8) to buffer E (500 mM NaCl, 20% glycerol, pH 7.8). The collected fractions were pooled and concentrated again. FAD was added to Pco and Mco equivalent to the protein concentration. Enzymes requiring $MgCl_2$ or Coenzyme B_{12} were stored in buffer C containing 2 mM of the respective cofactor. If not already included in storage buffer, glycerol was added to a final concentration of 20% (v/v) and the proteins flash-frozen in liquid nitrogen and stored at -80 °C.

Production of proteins containing a Strep-Tag (LD(4) and Cit) was as stated above. For all following steps, buffer C was used. After lysis and centrifugation, the supernatant was loaded onto a preequilibrated 1 ml StrepTrap column (GE healthcare) and ultimately eluted using buffer C containing 2 mM d-desthiobiotin. Concentration and storage did not differ from the steps described above.

Enzyme Kinetics

Activity Assay of β-hydroxybutyryl-CoA dehydrogenase (Bbd) (Figure S1A)

The activity of Bbd was measured in a coupled assay using Etr1p. The assay was done in 150 µL volume in a high precision quartz cuvette (10 mm, Hellma Analytics) at 30°C and contained 200 mM HEPES-KOH pH 7.5, 0.5 mM NADPH, 0.009 µg Bbd and 3 µg Etr1p. The reactions were started with 5, 10, 25, 50, 100, 200 and 300 µM (*R*)-3-hydroxybutyryl-CoA. The consumption of NADPH was observed at 365 nm ($\Delta\varepsilon_{365}$ = 3.3 mM^{-1} cm^{-1}) using a UV-Vis spectrophotometer (Cary 60, Agilent Technologies). The V_{max} and K_m were determined by fitting the values for U/mg using the Michaelis-Menten equation.

Activity Assay of Malate:thiokinase (MtkAB) (Figure S1B)

MtkAB activity was determined by coupling the reaction to Mcl, producing acetyl-CoA and glyoxylate. The latter was further reduced to glycolate using a corresponding reductase (Gox) and monitored following NADPH consumption at 360 nm ($\Delta\varepsilon_{360}$ = 3.4 mM^{-1} cm^{-1}) on a on a Cary-60 UV/Vis spectrometer (Agilent) using high precision quartz cuvettes (10 mm, Hellma Analytics). The assay contained 200 mM HEPES-KOH pH 7.5, 100 mM $MgCl_2$, 5 mM ATP, 2 mM CoA and, 0.8 mM NADPH, 14.73 µg MtkAB (equivalent amounts of both subunits), 3 µg Mcl, 2.1 µg Gox and varying amounts of L-malate in a final volume of 100 µl.

Enzyme assays

All samples were processed as follows: The samples were taken and quenched in a solution containing 10% v/v 50% formic acid and 10% v/v 500 mM sodiumpolyphosphate to promote protein precipitation (except for the coupling experiments (Figure 4) where no polyphosphate was used). The samples were centrifuged at 11.000 xg for 20 min at 4°C to pellet the proteins. The supernatant was transferred into a new tube and stored at -80°C until measurement.

CETCH 5.5 (Figure 2A)

The assay for the CETCH 5.5 was done in triplicates in 30 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 1 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 2 mM ATP and 5 mM NADPH. The enzyme concentrations were used according to table S1. The reaction was started with 100 µM propionyl-CoA and the tubes were shaken at 450 rpm at 30°C. The samples (8 µL) were taken after 90 min and quenched.

BHAC validation (Figure 2B)

The assay for the BHAC validation was done in triplicates in 50 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 10 mM NADPH, 10 mM NADH, 0.25 mM glycine and 0.1 mM pyridoxalphosphat. The enzymes for the BHAC and Mdh were used according to the concentrations shown in table S1. The reactions were started with 0, 0.25, 0.5, 0.75 or 1 mM of glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (12 µL) were taken after 60 min and quenched.

Malonyl-CoA stability test (Figure S3A)

The assay was done in duplicates in 15 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 5 mM

$MgCl_2$, 20 mM sodiumpolyphoshate, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 1 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. One setup was done with all enzymes needed for the production of acetyl-CoA except Mcl. One setup included only the matrix without enzymes. The reactions were started with 300 µM malonyl-CoA and the tubes were shaken at 450 rpm at 30°C. The samples (4.5 µL) were taken at 0, 60 and 120 min and quenched.

CETCH to acetyl-CoA optimization (Figure 2C)

The assay was done in triplicates in 100 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 20 mM phosphocreatine or sodiumpolyphoshate, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 1 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. One setup was done with and one without Pcc*. The reactions were started with 100 µM propionyl-CoA and the tubes were shaken at 450 rpm at 30°C. The samples (12 µL) were taken at 0, 30, 60, 90, 120, 180 and 240 min and quenched.

CETCH with Module 4a (Figure 3)

The assay was done in triplicates in 50 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 50 or 100 mM formate (HCOONa), 2 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1, whereas the conditions with 50 and 100 mU Frd included 5, respectively 10 times more Frd as stated. The reactions were started with 250 µM glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (8 µL) were taken at 30, 60, 90 and 120 min and quenched.

CETCH with Module 4b (Figure 3 and S3B)

The assay was done in triplicates in 75 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 2 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. Mdh was either used at 7.9 U/ml or 0.8 U/ml. The reactions were started with 250 µM glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (8 µL) were taken at 0, 30, 60, 90, 120 and 240 min and quenched.

CETCH with Module 4c (Figure 3)

The assay was done in triplicates in 75 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 2 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1, one assay with and one without Adk. The reactions were started with 250 µM glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (8 µL) were taken at 0, 30, 60, 90 and 120 min (+240 min sample for the experiment with Adk) and quenched.

CETCH with Module 4d (Figure S3C-D)

The assay was done in triplicates in 90 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 0.25, 0.5 or 1 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. The reactions were started with 500 µM glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (12 µL) were taken at 0, 15, 30, 60, 120 and 240 min and quenched.

CETCH with Module 4d (Figure 3)

The assay was done in triplicates in 65 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 1 or 2.5 mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. The reactions were started with 500 µM glyoxylate and the tubes were shaken at 450 rpm at 30°C. The samples (8 µL) were taken at 0, 30, 60, 90, 120 and 240 min and quenched.

CETCH with all modules with and without DEBS (Figure 4)

The assays were done in duplicates in 150 µl assay volume and included 100 mM HEPES-KOH pH 7.5, 10 mM $MgCl_2$, 20 mM phosphocreatine, 50 mM bicarbonate ($NaHCO_3$), 20 mM formate (HCOONa), 2 or 1 (Module 4d) mM Coenzyme A, 0.1 mM Coenzyme B_{12}, 5 mM ATP, 5 mM NADH, 5 mM NADPH. 1 mM glycine and 0.1 mM pyridoxalphosphat. The enzyme concentrations were used accordingly to table S1. For the assay with module 4a the 4Fe-4S cluster of Acn was reconstituted before the assay. Therefore, the purified enzyme was incubated with 5 mM dithiothreitol and 15 mM ammonium iron II sulfate. After 30

min on ice the buffer was exchanged with the storage buffer using a Zeba™ Micro Spin Desalting Column, 7K MWCO, 0.5 mL (ThermoScientific) accordingly to the provided protocol. To avoid NADPH oxidation by Frd during preparation of the assay, Frd was added as the last enzyme before starting the reaction in the assay with module 4a. The reactions were started with 250 µM glyoxylate or 125 µM propionyl-CoA (CETCH controls) and the tubes were shaken at 450 rpm at 30°C. The samples (13 µL) were taken at 0, 0.5, 1, 1.5, 2, 4, 6, 8, 10, 12 and 24 h and quenched.

DEBS assays

All DEBS assay were carried out in duplicates and contained 100 mM HEPES-KOH pH 7.5, 200 mM NaCl, 4 mM NADPH (0.7 mM for spectrophotometric assays), 4 µM Epi and 2 µM of each DEBS protein (see Figure S5). The reactions were started upon addition of 0.8 mM propionyl-CoA and 1 mM methylmalonyl-CoA (which were omitted for the negative control). For quantification, samples were taken after 0, 20, 40, 60, 80, 100, 120, 150 and 180 minutes, quenched with a final concentration of 5 % (v/v) formic acid and stored at -80°C until measurement.

Reduction rates were measured on a Cary-60 UV/Vis spectrometer (Agilent) using 10 mm quartz cuvettes (Hellma) following NADPH absorption at 360 nm ($\Delta\varepsilon_{360}$ = 3.4 $mM^{-1}cm^{-1}$).

Table S1. Concentrations of enzymes used in the assays. * Wilbur-Anderson unit.** Cell extract All commercially available enzymes were purchased from Sigma-Aldrich.

	CETCH	Module 4a	Module 4c	Module 4b	Module 4d			
	µM	µM	µM	µM	µM		U/mg	Source
Abb								
Pco	3.00	3.00	3.00	3.00	3.00		12	[3]
Ccr	0.58	0.58	0.58	0.58	0.58		110	[3]
Epi	0.60	0.60	0.60	0.60	0.60		440	[4]
Mcm	0.36	0.36	0.36	0.36	0.36		20	[5]
Scr	2.68	2.68	2.68	2.68	2.68		29	[6]
Ssr	0.76	0.76	0.76	0.76	0.76		4	[3]
Hbs	5.12	5.12	5.12	5.12	5.12		2	[7]
Gbd	0.56	0.56	0.56	0.56	0.56		26	[7]
Ecm	0.55	0.55	0.55	0.55	0.55		7	[4]
Mco	21.4	21.4	21.4	21.4	21.4		0.1	[3]
Mch	1.26	1.26	1.26	1.26	1.26		1500	[8]
Mcl	13.6	13.6	13.6	13.6	13.6		5	[9]
Cat	1.37	1.37	1.37	1.37	1.37		11740	[10]
Fdh	14.4	14.4	14.4	14.4	14.4		1	[11]
Cpk	0.39	0.39	0.39	0.39	0.39		150	commercial
Ppk								
CA	0.02	0.02	0.02	0.02	0.02		*2000	commercial
Gor	1.10						-	[12]
Bha		2.26	2.26	2.26	2.26		116	[13]
Bhd		1.37	1.37	1.37	1.37		92	[13]
Isr		14.8	14.8	14.8	14.8		358	[13]
Agt		1.93	1.93	1.93	1.93		77	[13]
Mdh		1.33	1.33	0.13	1.33		1611	[14]
Mtk							1.5	This Work
Fum		0.66					340	[15]
Frd		1.05					0.007	[16]
Scs		1.00		1.00			19	[17]
Pcc*			7.69				1	[18]
Mcr			0.28				10	[19]
Pcs			3.09				1	[20]
Adk			0.18				1247	[21]
Cit				8.98			4	[1]
Acn				15.88			6	[22]
Icl				12.54			38	[23]
Aat					17.74		5	[24]
Aar					28.27		**0.6	[25]
Bbd					52.47		965	This work
DEBS (each protein)	2							-

4.6.2. LC-MS Measurements

Analysis of CoA esters

All CoA esters were measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LS/MS) equipped with an UHPLC (Agilent Technologies 1290 Infinity II) using a 50 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 2 µl of the diluted samples (1:10 in water). The flow was set to 0.250 ml/min and the separation was performed using 50 mM ammonium formate pH 8.1 (buffer A) and acetonitrile (B). We quantified the CoAs using external standard curves prepared in 1:10 diluted (water) sample matrix. The parameters for the multiple reaction monitoring (MRMs) are displayed in table S3 and the gradient in table S2. Data analysis was done using the Agilent Mass Hunter Workstation Software.

Table S2. Gradient for the separation of CoA esters

Time [min]	A [%]	B [%]
0	100	0
2	100	0
5	94	6
8	77	23
10	20	80
11	20	80
12	100	0
12.5	100	0

Table S3. MRM transitions

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
Malyl-CoA (Quantifier)	884.1	377.1	25	380	37	5	Positive
Malyl-CoA (Qualifier)	884.1	428	25	380	29	5	Positive
Acetyl-CoA (Quantifier)	810.1	302.2	25	380	35	5	Positive
Acetyl-CoA (Qualifier)	810.1	428	25	380	35	5	Positive
Ethylmalonyl-CoA (Quantifier)	882.1	331.2	25	380	41	5	Positive
Ethylmalonyl-CoA (Qualifier)	882.1	428	25	380	29	5	Positive
Methylsuccinyl-CoA (Quantifier)	882	375.1	25	380	33	5	Positive
Methylsuccinyl-CoA (Qualifier)	882	428	25	380	29	5	Positive
Mesaconyl-CoA	880.1	375.1	25	380	25	5	Positive

(Quantifier)							
Mesaconyl-CoA (Qualifier)	880.1	428	25	380	35	5	Positive
Succinyl-CoA (Quantifier)	868.1	361.1	25	380	35	5	Positive
Succinyl-CoA (Qualifier)	868.1	428.1	25	380	35	5	Positive
Methylmalonyl-CoA (Quantifier)	868.1	317.1	25	380	41	5	Positive
Methylmalonyl-CoA (Qualifier)	868.1	428	25	380	31	5	Positive
Malonyl-CoA (Quantifier)	854.1	245	25	380	32	5	Positive
Malonyl-CoA (Qualifier)	854.1	428	25	380	28	5	Positive
Γ-hydroxybutyryl-CoA (Quantifier)	854.1	347.1	25	380	37	5	Positive
Γ-hydroxybutyryl-CoA (Qualifier)	854.1	428	25	380	30	5	Positive
Crotonyl-CoA (Quantifier)	836.1	329	25	380	33	5	Positive
Crotonyl-CoA (Qualifier)	836.1	428	25	380	26	5	Positive
Propionyl-CoA (Quantifier)	824.1	317.1	25	380	31	5	Positive
Propionyl-CoA (Qualifier)	824.1	428	25	380	28	5	Positive
Methylsuccinyl-CoA (Quantifier)	824.1	317.1	25	380	31	5	Positive
Methylsuccinyl-CoA (Qualifier)	824.1	428	25	380	28	5	Positive
B-methylmalyl-CoA (Quantifier)	898.1	391.1	25	380	39	5	Positive
B-methylmalyl-CoA (Qualifier)	898.1	428.1	25	380	33	5	Positive

Glycolate quantification

Glycolate was measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LS/MS) equipped with an UHPLC (Agilent Technologies 1290 Infinity II) using a 150 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 1 µl. The diluted samples (1:10 in water) as well as the external standard curve were diluted 1:2 with 10 µM ^{13}C-labeled glycolate as internal standard. The flow was set to 0.100 ml/min and the separation was performed using dH_2O with 0.1% formic acid (buffer A) and methanol with 0.1% formic acid (B). The parameters for the multiple reaction monitoring (MRMs) are displayed in table S5 and the gradient in table S4. Data analysis was done using the Agilent Mass Hunter Workstation Software.

Table S4. Gradient for the analysis of glycolate

Time [min]	A [%]	B [%]
0	100	0
4	100	0

6	0	100
7	0	100
7.1	100	0
12	100	0

Table S5. MRM transitions

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
^{12}C-Glycolate (Quantifier)	75	47	150	380	9	5	Negative
^{12}C-Glycolate (Qualifier)	75	75	150	380	0	5	Negative
^{13}C-Glycolate (Quantifier)	77	48	150	380	9	5	Negative
^{13}C-Glycolate (Qualifier)	77	77	150	380	0	5	Negative

Malate quantification

Malate was measured on a triple quadrupole mass spectrometer (Agilent Technologies 6495 Triple Quad LS/MS) equipped with an UHPLC (Agilent Technologies 1290 Infinity II) using a 150 x 2.1 mm C18 column (Kinetex 1.7 µm EVO C18 100 Å) at 25 °C. The injection volume was 5 µl. The diluted samples (1:10 in water) as well as the external standard curve were diluted 1:2 with 10 µM ^{13}C-labeled malate as internal standard. The flow was set to 0.150 ml/min and the separation was performed using dH_2O with 0.1% formic acid (buffer A) and methanol with 0.1% formic acid (B). The parameters for the multiple reaction monitoring (MRMs) are displayed in table S5 and the gradient in table S6. Data analysis was done using the Agilent Mass Hunter Workstation Software.

Table S6. Gradient for the analysis of malate

Time [min]	A [%]	B [%]
0.0	85	15
7.0	0	100
9.0	0	100
9.1	85	15
15.0	85	15

Table S7. MRM transitions

Compound	Precursor Ion	Product Ion	Dwell	Fragmentor	Collision Energy	Cell Accelerator Volt.	Polarity
^{12}C-Malate (Quantifier)	133	115	150	80	11	5	Negative

^{12}C-Malate (Qualifier)	133	133	150	80	0	5	Negative
^{13}C-Malate (Quantifier)	137	119	150	80	11	5	Negative

HPLC-MS analysis of 6-dEB

5 µl of the quenched assays were analyzed via HPLC-ESI-TOF on a 6550 iFunnel Q-TOF LC-MS (Agilent) with a 1.8 µm Zorbax SB-C18 column, 50 x 2.1 mm (Agilent) and using H_2O (A) and acetonitrile (B) both containing 0.1% formic acid. The gradient condition were as follows: 0 min 5 % B, 1 min 5 % B, 6 min 95 % B, 6.5 min 95 % B, 7 min 5 % B with a flow rate of 250 µl/min.. Capillary voltage was set at 3.5 kV and nitrogen gas was used as nebulizing (20 psig), drying (13 l/min, 225 °C) and sheath gas (12 l/min, 400°C). MS data were acquired with a scan range of 50-1200 m/z. Data were analyzed using the MassHunter Analysis software (Agilent). Evaluated 6-dEB (RT 3.26 min) adducts are shown in table S6.

Table S8. Analyzed 6-dEB adducts

Adduct	[M+H]+	[M+Na]+	[M-H_2O+H]+
m/z	387.274116	409.256058	369.263551

4.6.3. CoA Standards Synthesis

All CoA thioesters were synthesized and purified according to previously established protocols[26,27]

4.6.4. Supplementary Figures and Tables

Table S9. List of enzymes used in this work

#	Name	Full name	Catalyzed reaction	Origin	Comment	Source
1	Ccr	crotonyl-CoA carboxylase/reductase	Crotonyl-CoA + NADPH + CO2 ⇌ Ethylmalonyl-CoA + NADP+	*M. extorquens*		[3]
2	Epi	methylmalonyl-/ethylmalonyl-CoA epimerase	(2S)-Ethylmalonyl-CoA ⇌ (2R)-Ethylmalonyl-CoA	*R. sphaeroides*		[3]
	Ecm	ethylmalonyl-CoA mutase	(2R)-Ethylmalonyl-CoA ⇌ Methylsuccinyl-CoA	*R. sphaeroides*		[3]
3	Mco	methylsuccinyl-CoA oxidase	Methylsuccinyl-CoA + O2 ⇌ Mesaconyl-CoA + H2O2	*R. sphaeroides*		[3]
4	Mch	mesacony-CoA hydratase	Mesaconyl-CoA + H2O(l) ⇌ beta-Methylmalyl-CoA	*R. sphaeroides*		[3]
5	Mcl	β-methylmalyl-CoA lyase	beta-Methylmalyl-CoA ⇌ Glyoxylate + Propionyl-CoA	*R. sphaeroides*		[3]
6	Pco	propionyl-CoA oxidase	Propionyl-CoA + O2 ⇌ Acrylyl-CoA + H2O2	*A. thaliana*	A. thaliana short chain acyl-CoA oxidase 4 T134L	[3]
7	Ccr	Crotonyl-CoA carboxylase/reductase	Acrylyl-CoA + NADPH + CO2 ⇌ Methylmalonyl-CoA + NADP+	*M. extorquens*		[3]
8	Epi	methylmalonyl-/ethylmalonyl-CoA epimerase	(2S)-Ethylmalonyl-CoA ⇌ (2R)-Ethylmalonyl-CoA	*R. sphaeroides*		[3]
	Mcm	methylmalonyl-CoA mutase	Methylmalonyl-CoA ⇌ Succinyl-CoA	*R. sphaeroides*		[3]
9	Scr	succinyl-CoA reductase	Succinyl-CoA + NADPH ⇌ Succinic semialdehyde + NADP+ + CoA	*C. kluyveri*		[3]
10	Ssr	succinic semialdehyde reductase	Succinic semialdehyde + NADPH ⇌ 4-Hydroxybutyric acid + NADP+	*H. sapiens*		[3]
11	Hbs	4-hydroxybutyryl-CoA synthetase	4-Hydroxybutyric acid + ATP + CoA ⇌ 4-Hydroxybutyryl-CoA + ADP + Pi	*N. maritimus*		[3]
12	Hbd	4-hydroxybutyryl-CoA dehydratase	4-Hydroxybutyryl-CoA ⇌ Crotonyl-CoA + H2O(l)	*N. maritimus*		[3]
13	Bha	β-hydroxyaspartate aldolase	3-hydroxyaspartate ⇌ Iminosuccinate + H2O	*P. denitrificans*	bhcC	[28]
14	Bhd	β-hydroxyaspartate dehydratase	Glyoxylate + Glycine ⇌ 3-hydroxyaspartate	*P. denitrificans*	bhcB	[28]
15	Isr	imminosuccinate reductase	Iminosuccinate + NADPH ⇌ Aspartate + NADP+	*P. denitrificans*	bhcD	[28]
16	Agt	aspartate-glyoxylate aminotransferase	Aspartate + Glyoxylate ⇌ Oxaloacetate + Glycine	*P. denitrificans*	bhcA	[28]
17	Mdh	malate dehydrogenase	Oxaloacetate + NADH ⇌ Malate + NAD+	*E. coli*		[29]
18	Mtk	malyl-CoA synthetase	Malate + ATP + CoA ⇌ Malyl-CoA + ADP + Pi	*M. extorquens*	mtkAB; subunit beta A, subunit alpha B	[29]
19	Mcl	β-methylmalyl-CoA lyase	Malyl-CoA ⇌ Glyoxylate + Acetyl-CoA	*R. sphaeroides*		[3]

20	Fum	fumarate hydratase	Malate ⇌ Fumarate + H2O	*E. coli*		[30]
21	Frd	fumarate reductase	Fumarate + NADH ⇌ Succinate + NAD+	*T. brucei*		[16]
22	Scs	succinyl-CoA synthetase	Succinate + ATP + CoA ⇌ Succinyl-CoA + ADP + Pi	*E. coli*	sucC subunit beta, sucD subunit alpha	[30]
23	Cit	citrate synthase	Acetyl-CoA + Oxaloacetate + H2O(l) ⇌ Citrate + CoA	*Synechocystis sp.6803*		This work
24	Acn	aconitate hydratase A	Citrate ⇌ Isocitrate	*E. coli*		[30]
25	Icl	isocitrate lyase	Isocitrate ⇌ Glyoxylate + Succinate	*E. coli*		[30]
26	Pcc*	propionyl-CoA carboxylase	Acetyl-CoA + HCO3- + ATP ⇌ Malonyl-CoA + AMP + PPi	*M. extorquens*	M. extorquens propionyl-CoA carboxylase D407I	[31]
27	Mcr	malonyl-CoA reductase	Malonyl-CoA + NADPH ⇌ Malonate semialdehyde + NADP+ + CoA	*C.aurantiacus*		[19]
28	Mcr	malonyl-CoA reductase	Malonate semialdehyde + NADPH ⇌ 3-Hydroxypropionate + NADP+	*C.aurantiacus*		[19]
29	Pcs	propionyl-CoA synthase	3-Hydroxypropionate + ATP + CoA ⇌ 3-Hydroxypropionyl-CoA + AMP + PPi	*Erythrobacter NAP1*		[32]
30	Pcs	propionyl-CoA synthase	3-Hydroxypropionyl-CoA ⇌ Acrylyl-CoA + H2O(l)	*Erythrobacter NAP1*		[32]
31	Pcs	propionyl-CoA synthase	Acrylyl-CoA + NADPH ⇌ Propionyl-CoA + NADP+	*Erythrobacter NAP1*		[32]
32	pha	acetoacetyl-CoA thiolase	2 Acetyl-CoA ⇌ Acetoacetyl-CoA + CoA	*C.necator*		[33]
33	phb	acetoacetyl-CoA reductase	Acetoacetyl-CoA + NADPH ⇌ (S)-3-Hydroxybutyryl-CoA + NADP+	*C.necator*		[33]
34	phj	enoyl-CoA hydratase	(S)-3-Hydroxybutyryl-CoA ⇌ Crotonyl-CoA + H2O(l)	*P.aeruginosa*		[34]
35	DEBS	6-deoxyerythronolide B synthase	Propionyl-CoA + 6 NADPH + 6 Methylmalonyl-CoA ⇌ 6-Deoxyerythronolide B + 6 CO2 + 7 CoA + 6 NADP+ + H2O	*S. erythrea*	summarized reaction sequence	[35]

Table S10. Yield comparison of the different modules. mm-CoA: methylmalonyl-CoA. Carbons/mm-CoA and Total carbons in product are calculated for the methylmalonyl backbone (C4).

	Substrate (glyoxylate) [µM]	Initial carbons [µM]		Product (mm-CoA) [µM]	Carbons/ mm-CoA	Total carbons in product [µM]	Yield [%]
4a	250	500		130	4	520	104
4b	250	500		110	4	441	88
4c	250	500		498	4	1992	398
4d	250	500		69	4	276	55

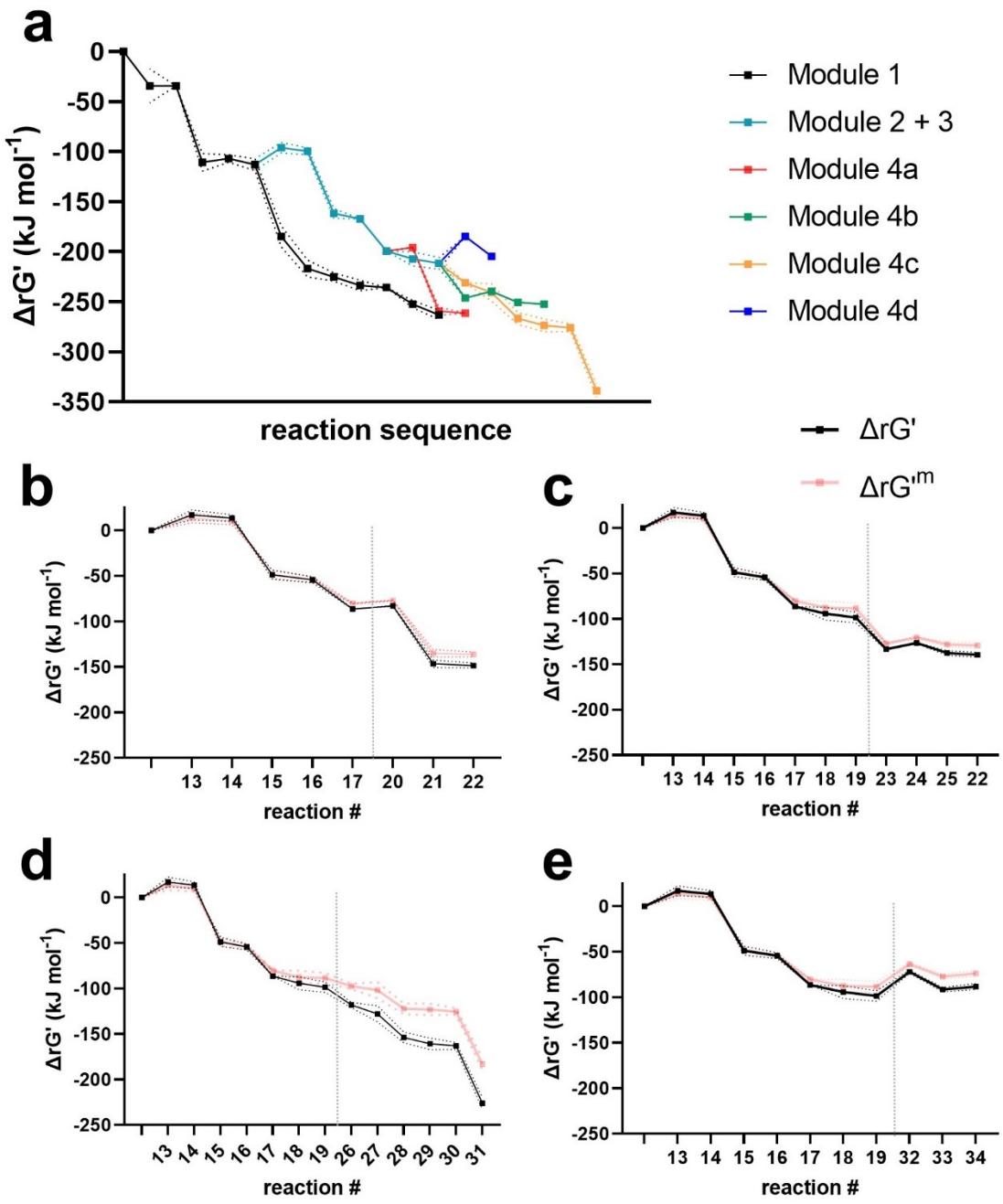
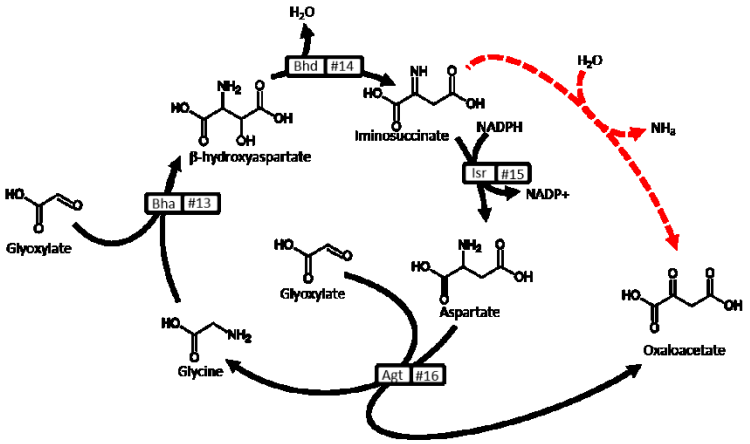


Figure S1. Gibbs free energy profiles. (A) Overview of all reactions. (B) Modules 2+3+4a, (C) Modules 2+3+4b, (D) Modules 2+3+4c, (E) Module 2+3+4d. All reactions start from glyoxylate, branching points between modules are indicated by grey dashed lines. The ΔrG' values were estimated using the eQuilibrator v3.0 tool[36] at pH 7.5, I = 0.25 and pMg = 3. All substrate and product (CoA, acids, aldehydes) concentrations were assumed to be 250 µM, with the following exceptions: #23 200 µM acetyl-CoA and 50 µM oxaloacetate, #24-25 all reactants 50 µM (and 250 µM glyoxylate), #32-34 125 mM acetoacetyl-CoA and every following reactant. Concentrations of other metabolites were estimated as follows: NADPH = 4.5 mM; NADP+ = 0.5 mM; NADH = 4.5 mM, NAD+ = 0.5 mM; ATP = 3 mM; ADP

= 1.4 mM; AMP = 0.5 mM; CoA = 1 mM; CO_2(g) = 3.31 mM; CO_2(total) = 50 mM.



	Glycine	Glyoxylate	Theoretical yield hydrolysis	Theoretical yield enzymatically	Measured yield
1)	250 µM	0 µM	0 µM	0 µM	1.1 ± 0.0 µM
2)	250 µM	250 µM	250 µM	125 µM	106.1 ± 1.6 µM
3)	250 µM	500 µM	250 µM	250 µM	216.6 ± 5.0 µM
4)	250 µM	750 µM	250 µM	375 µM	315.2 ± 0.7 µM
5)	250 µM	1000 µM	250 µM	500 µM	435.0 ± 1.0 µM

Figure S2. BHAC performance at different glyoxylate concentrations. To assess whether the whole BHAC is active, we tested different concentrations of glyoxylate while the amount of glycine was fixed at 250 µM. An incomplete reaction sequence could lead to non-enzymatic hydrolysis (red arrow) of iminosuccinate to oxaloacetate and therefore to a usage of carbons from the initially added glycine. All values of the measured yield, which exceed 250 µM, indicate a functional cycle because only the recycling of the amino group from the aspartate (#16), and therefore the usage of another molecule of glyoxylate, allows for these concentrations. This implies the usage of two glyoxylate molecules (C2) for the formation of one molecule of oxaloacetate (C4).

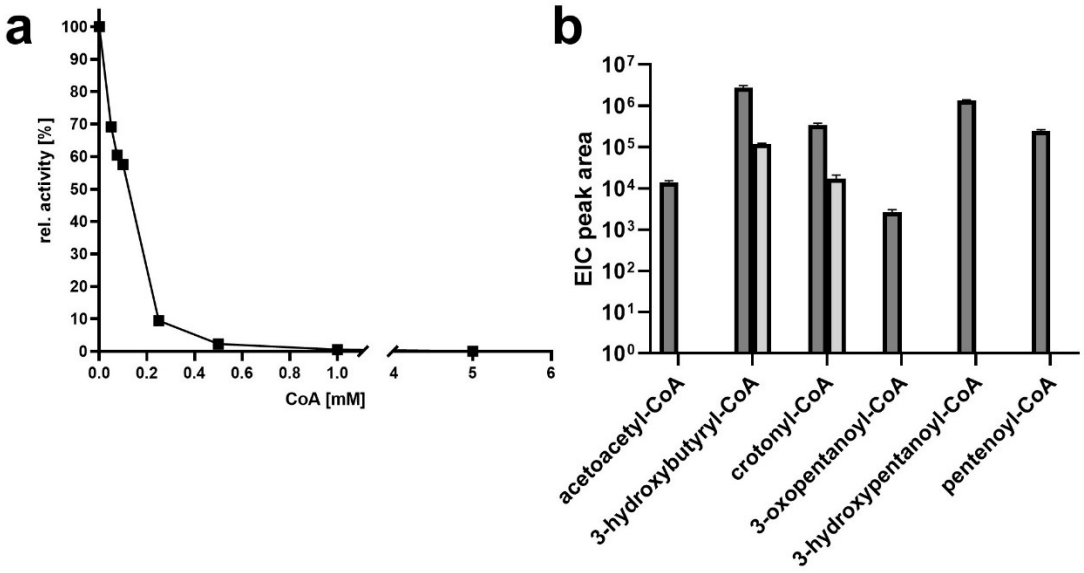


Figure S3. Inhibition of the condensation reaction of Aat by free Coenzyme A. (A) Promiscuity of Aat assayed with acetyl-CoA and propionyl-CoA. (B) Dark grey are EIC values obtained for each compound in an assay containing all three pathway enzymes (Aat, Aar, Bbd), acetyl-CoA and propionyl-CoA. In the final feedback assay, only 3-hydroxybutyryl-CoA and crotonyl-CoA were detected (light grey).

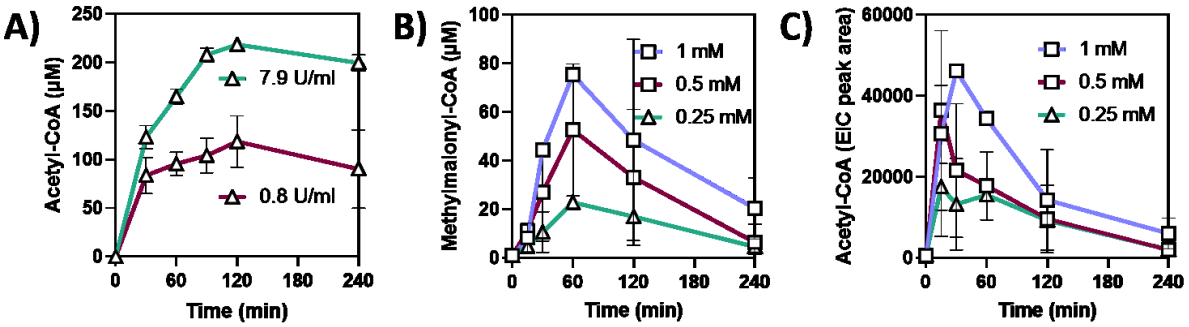


Figure S4. Enzyme assays for pathway optimization. A) Acetyl-CoA accumulation when using module 4b with different amounts of Mdh. More Mdh leads to accumulation of acetyl-CoA and a decrease in methylmalonyl-CoA (Figure 3). Setup see 1.3.6. B) and C) Methylmalonyl-CoA and acetyl-CoA accumulation when using module 4d with different amounts of CoA (1, 0.5 and 0.25 mM). Setup see 1.3.8.

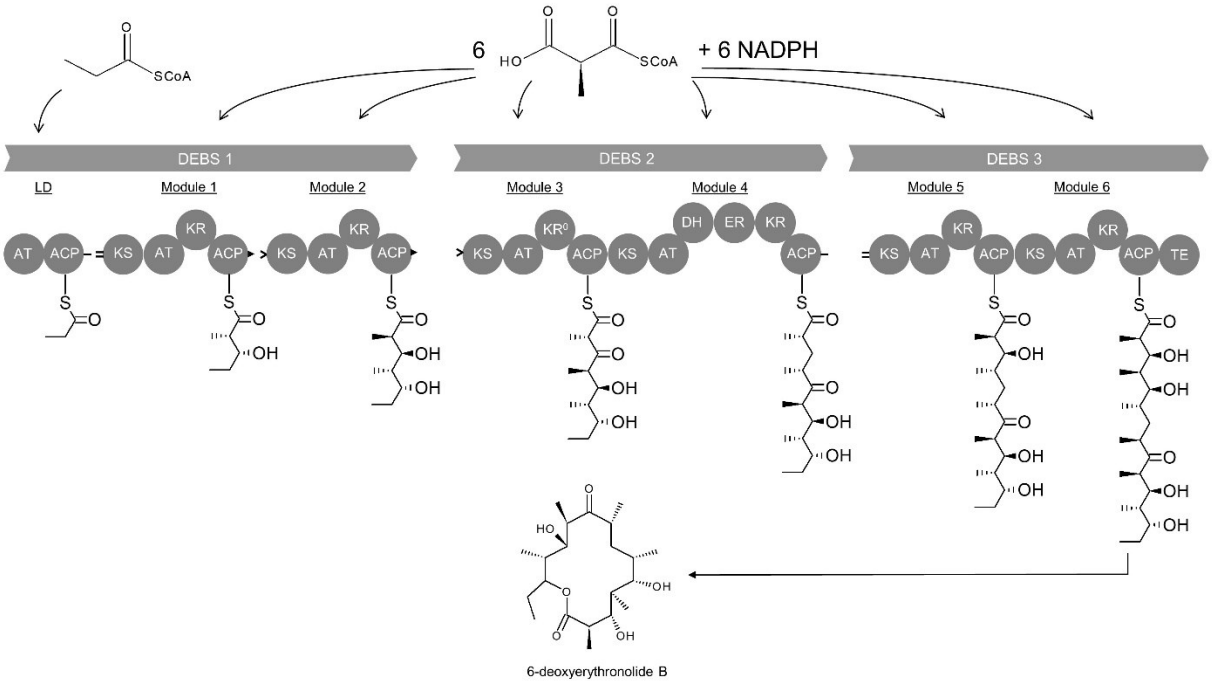


Figure S5. 6-deoxyerythronolide B synthase (DEBS). Displayed is the genetic architecture (DEBS 1-3), as well as the dissected in vitro assembly line (DEBS 1 into loading domain (LD), module 1 and module 2), indicated by the linker domains[35]. Each molecule 6-dEB is derived from one molecule propionyl-CoA and six molecules (2S)-methylmalonyl-CoA, under the consumption of six reducing equivalents NADPH.

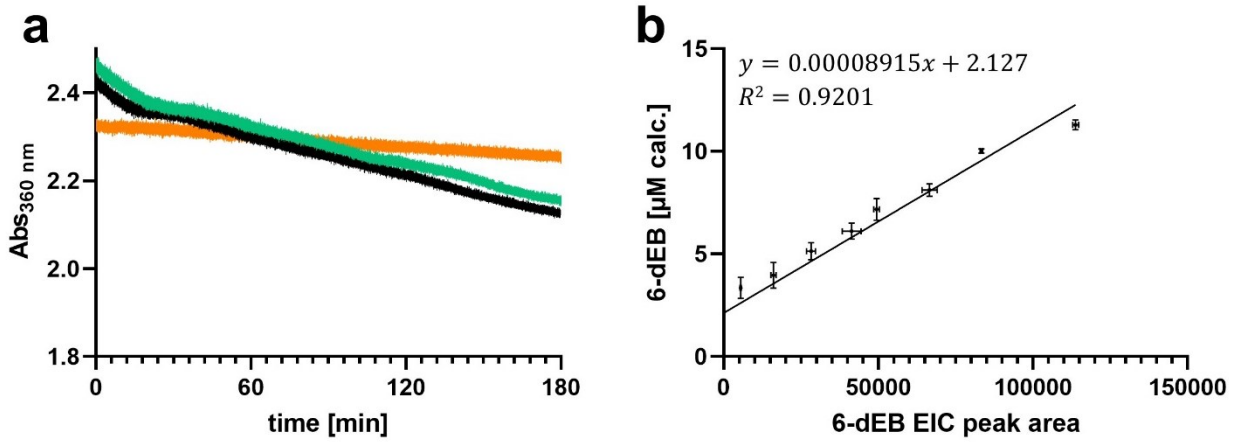
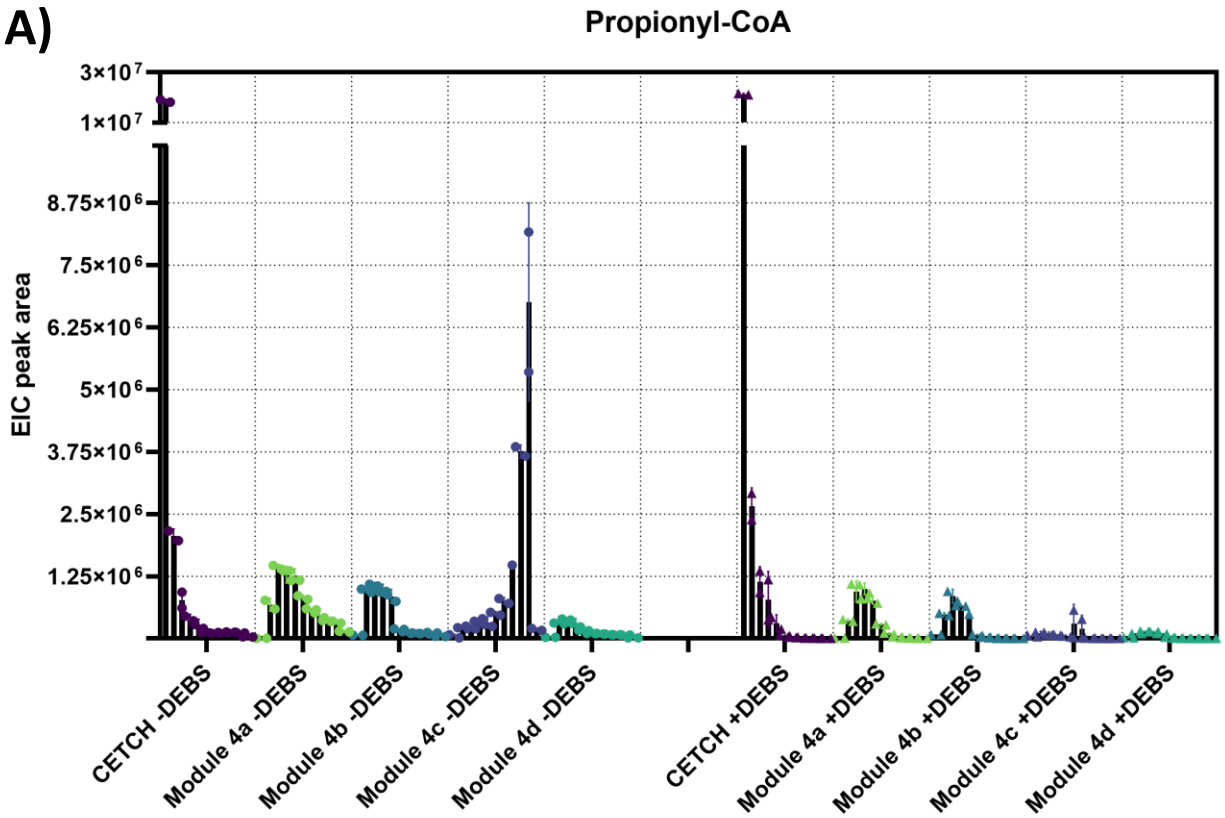
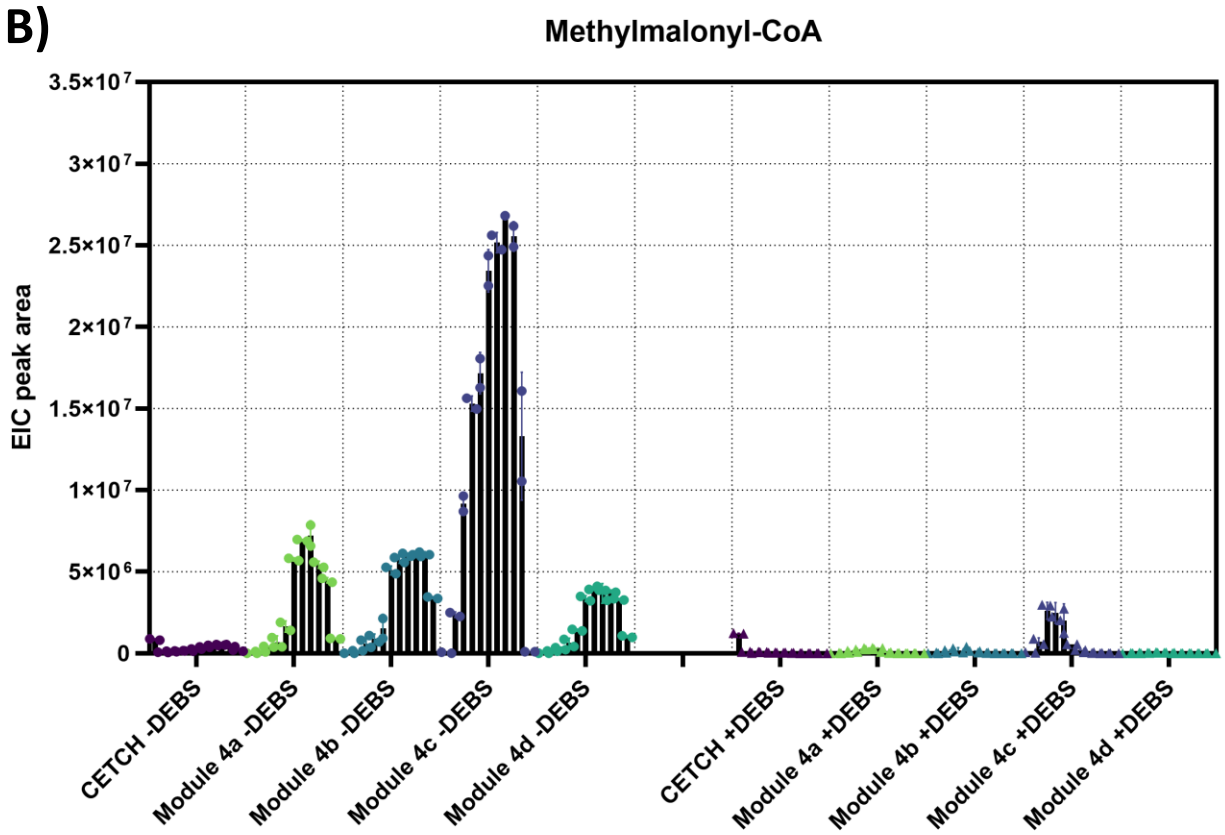
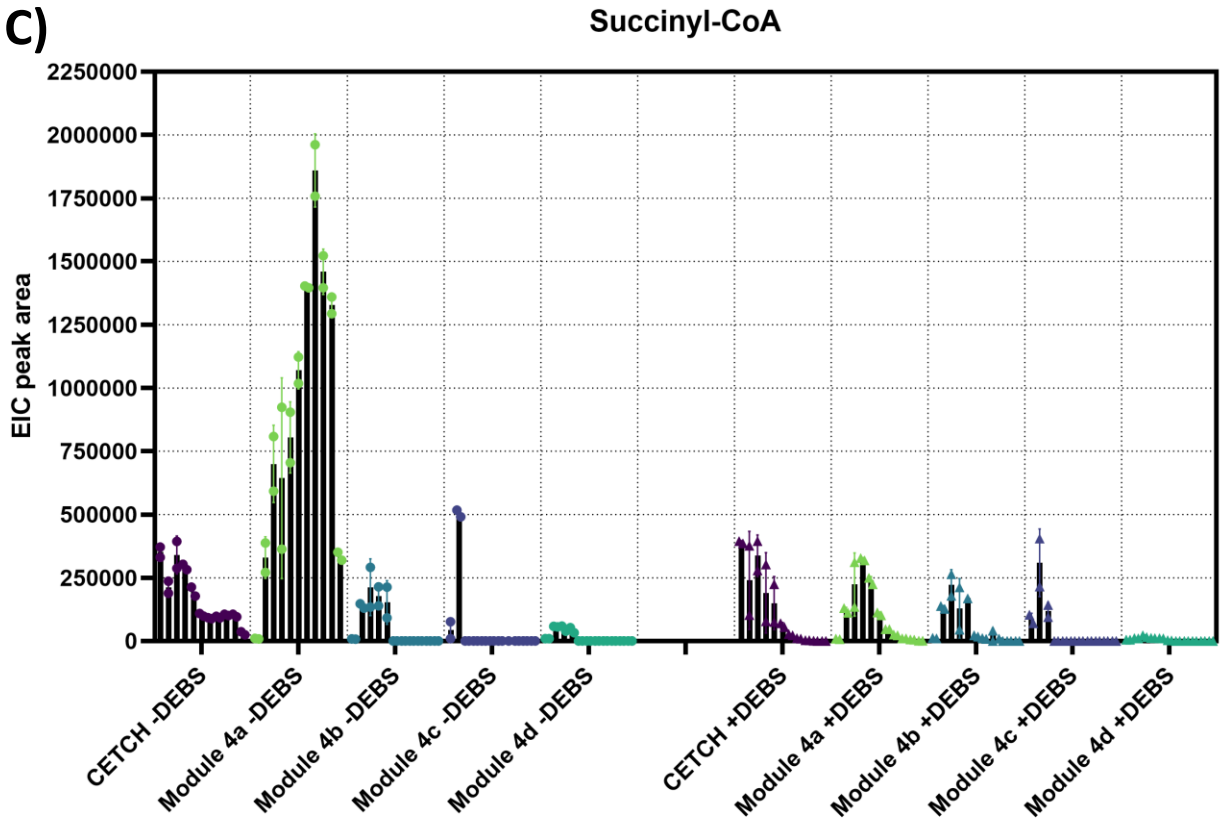
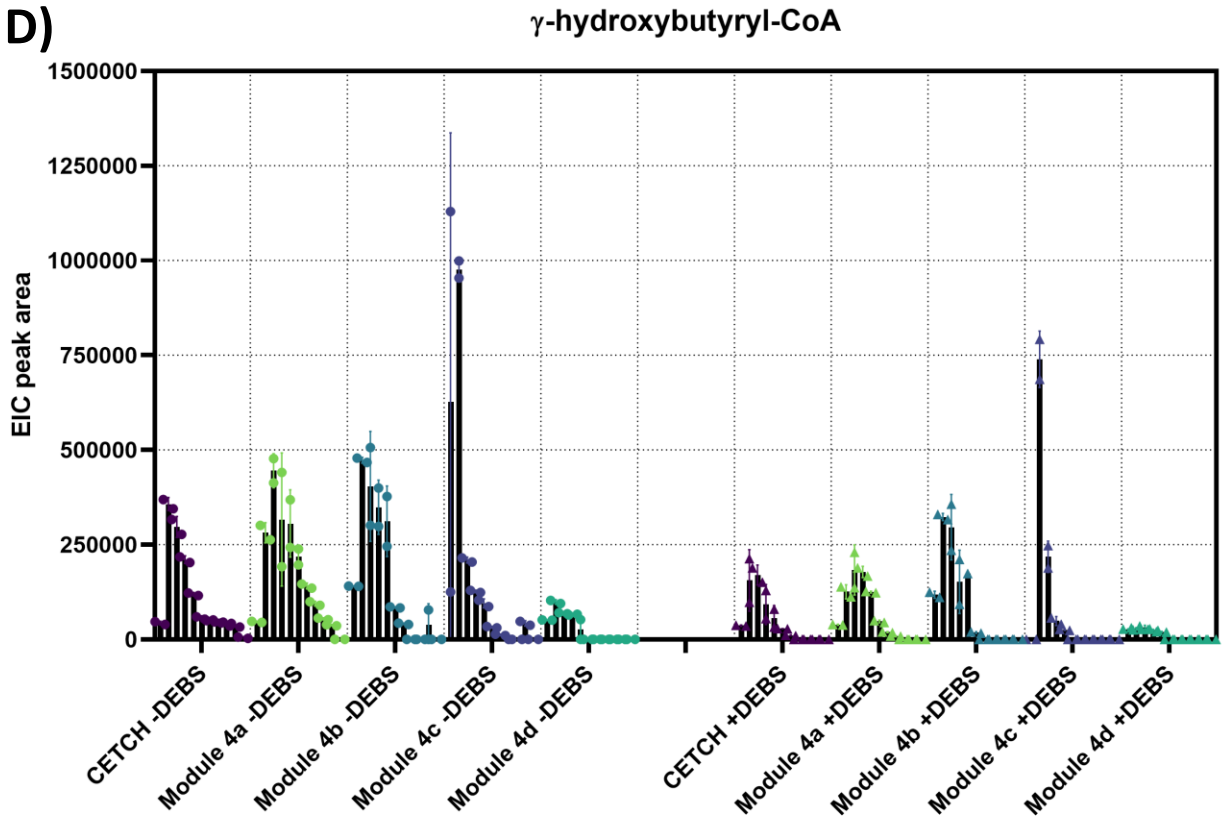


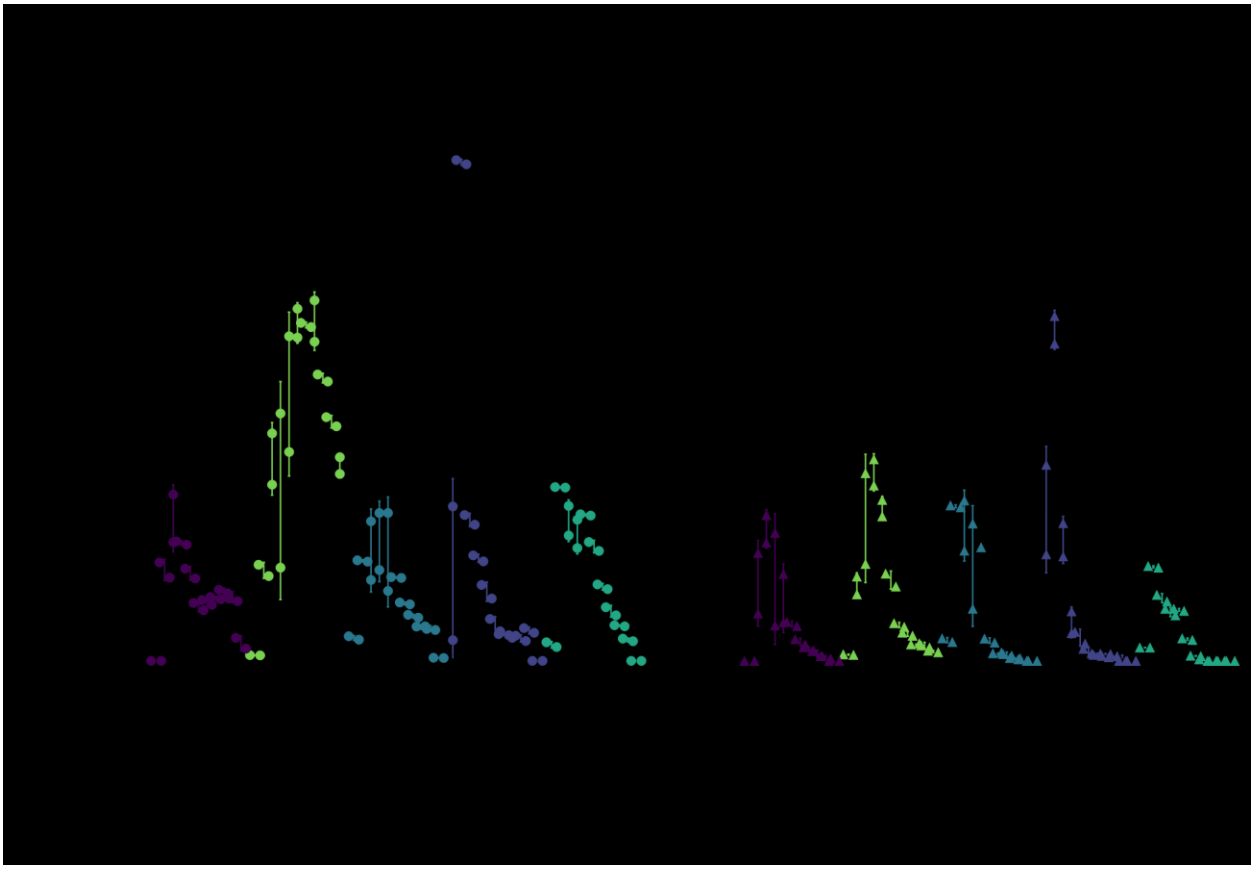
Figure S6. 6-dEB quantification. NADPH oxidation by DEBS was measured spectrophotometrically (A). After establishing stoichiometric equivalence between NADPH oxidation and 6-dEB production, the incremental reduction rates were related to the 6-dEB EIC measurements. Note that the ordinate offset has been subtracted to yield the concentrations discussed in the main text.

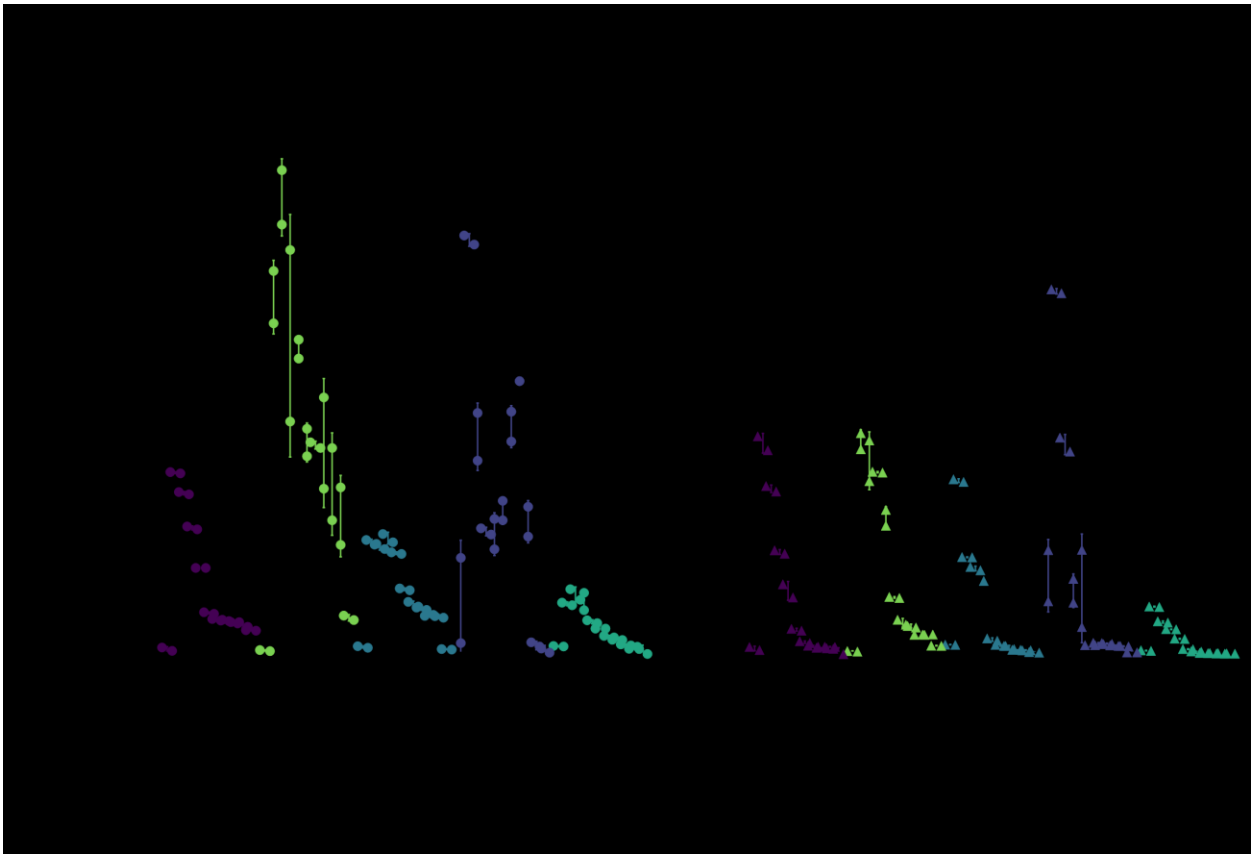

A)
Propionyl-CoA
EIC peak area
3×10⁷
1×10⁷
8.75×10⁶
7.5×10⁶
6.25×10⁶
5×10⁶
3.75×10⁶
2.5×10⁶
1.25×10⁶
CETCH -DEBS
Module 4a -DEBS
Module 4b -DEBS
Module 4c -DEBS
Module 4d -DEBS
CETCH +DEBS
Module 4a +DEBS
Module 4b +DEBS
Module 4c +DEBS
Module 4d +DEBS

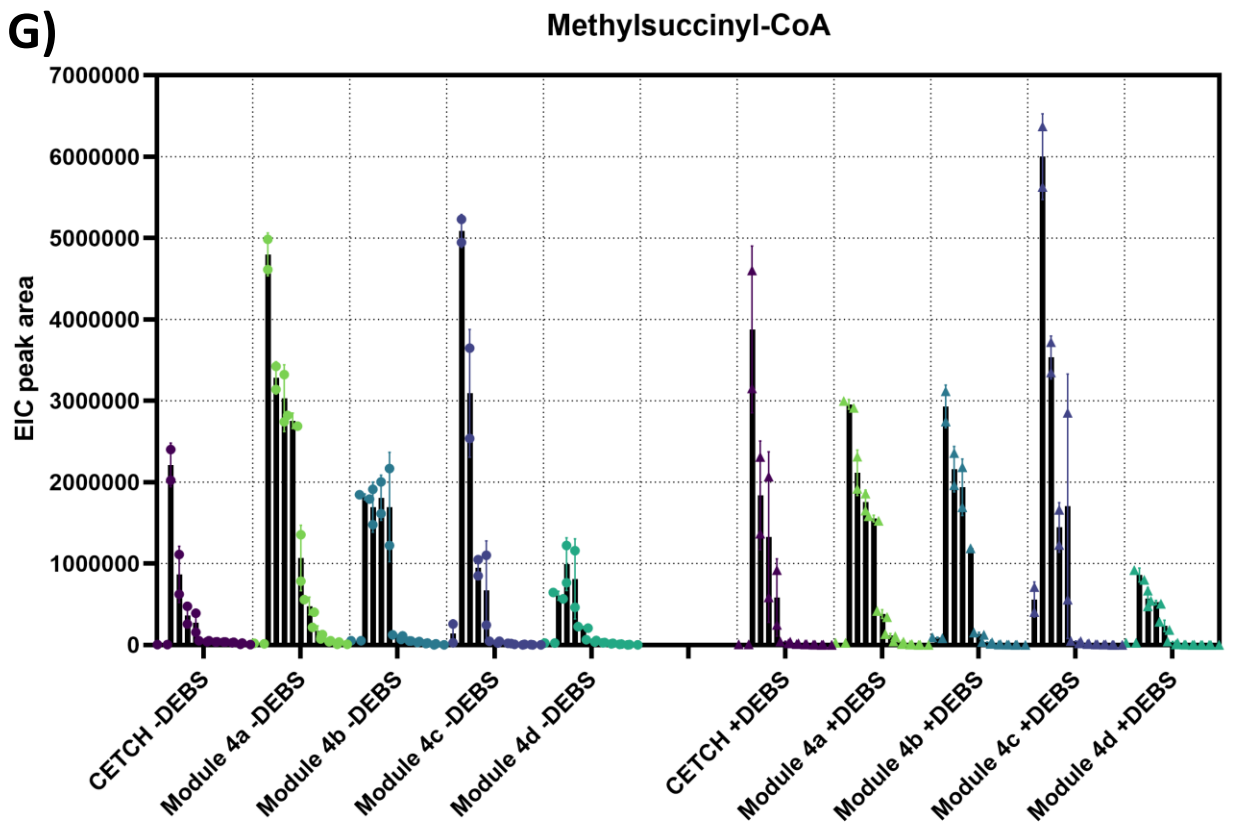

B)
Methylmalonyl-CoA
EIC peak area
3.5×10⁷
3×10⁷
2.5×10⁷
2×10⁷
1.5×10⁷
1×10⁷
5×10⁶
0
CETCH -DEBS
Module 4a -DEBS
Module 4b -DEBS
Module 4c -DEBS
Module 4d -DEBS
CETCH +DEBS
Module 4a +DEBS
Module 4b +DEBS
Module 4c +DEBS
Module 4d +DEBS

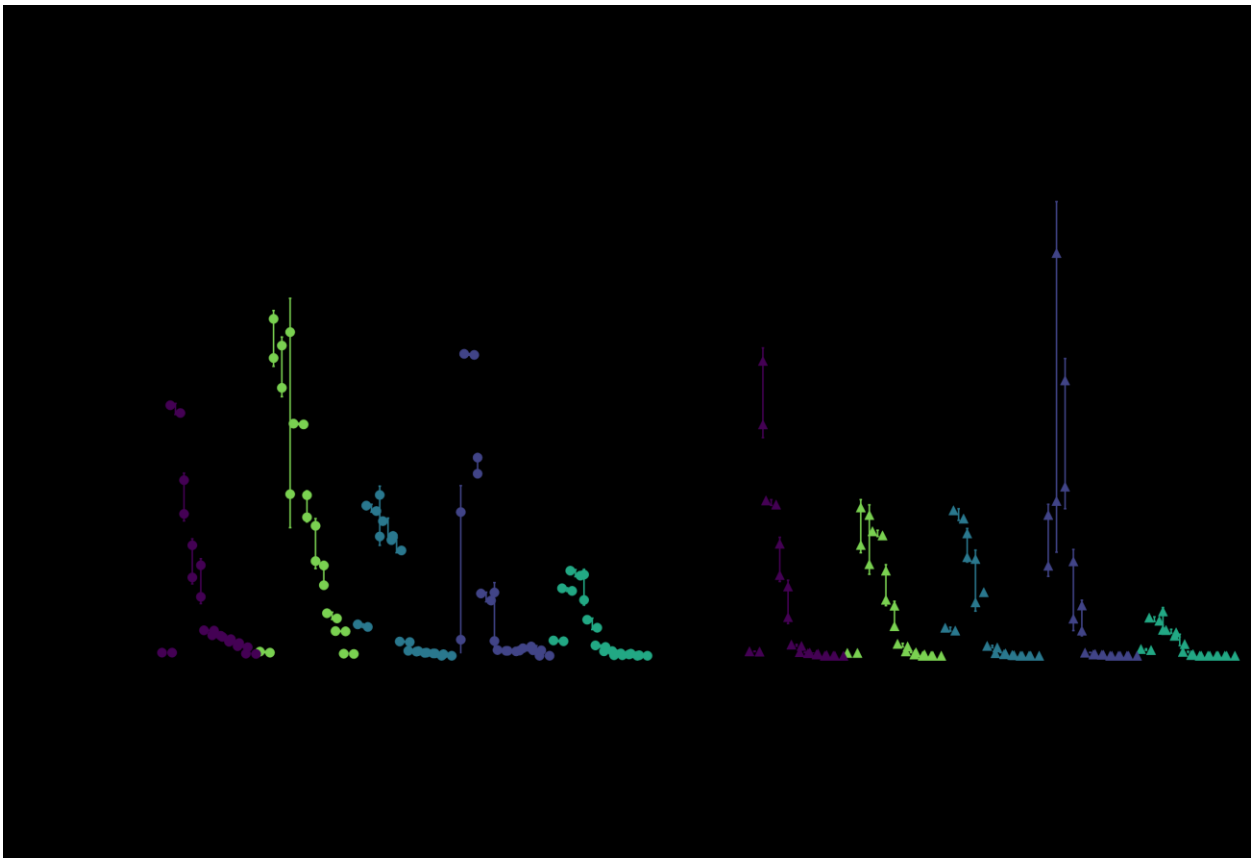

C)
Succinyl-CoA
EIC peak area
2250000
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
CETCH -DEBS
Module 4a -DEBS
Module 4b -DEBS
Module 4c -DEBS
Module 4d -DEBS
CETCH +DEBS
Module 4a +DEBS
Module 4b +DEBS
Module 4c +DEBS
Module 4d +DEBS


D)
γ-hydroxybutyryl-CoA
EIC peak area
1500000
1250000
1000000
750000
500000
250000
0
CETCH -DEBS
Module 4a -DEBS
Module 4b -DEBS
Module 4c -DEBS
Module 4d -DEBS
CETCH +DEBS
Module 4a +DEBS
Module 4b +DEBS
Module 4c +DEBS
Module 4d +DEBS






G)
Methylsuccinyl-CoA
EIC peak area
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
CETCH -DEBS
Module 4a -DEBS
Module 4b -DEBS
Module 4c -DEBS
Module 4d -DEBS
CETCH +DEBS
Module 4a +DEBS
Module 4b +DEBS
Module 4c +DEBS
Module 4d +DEBS



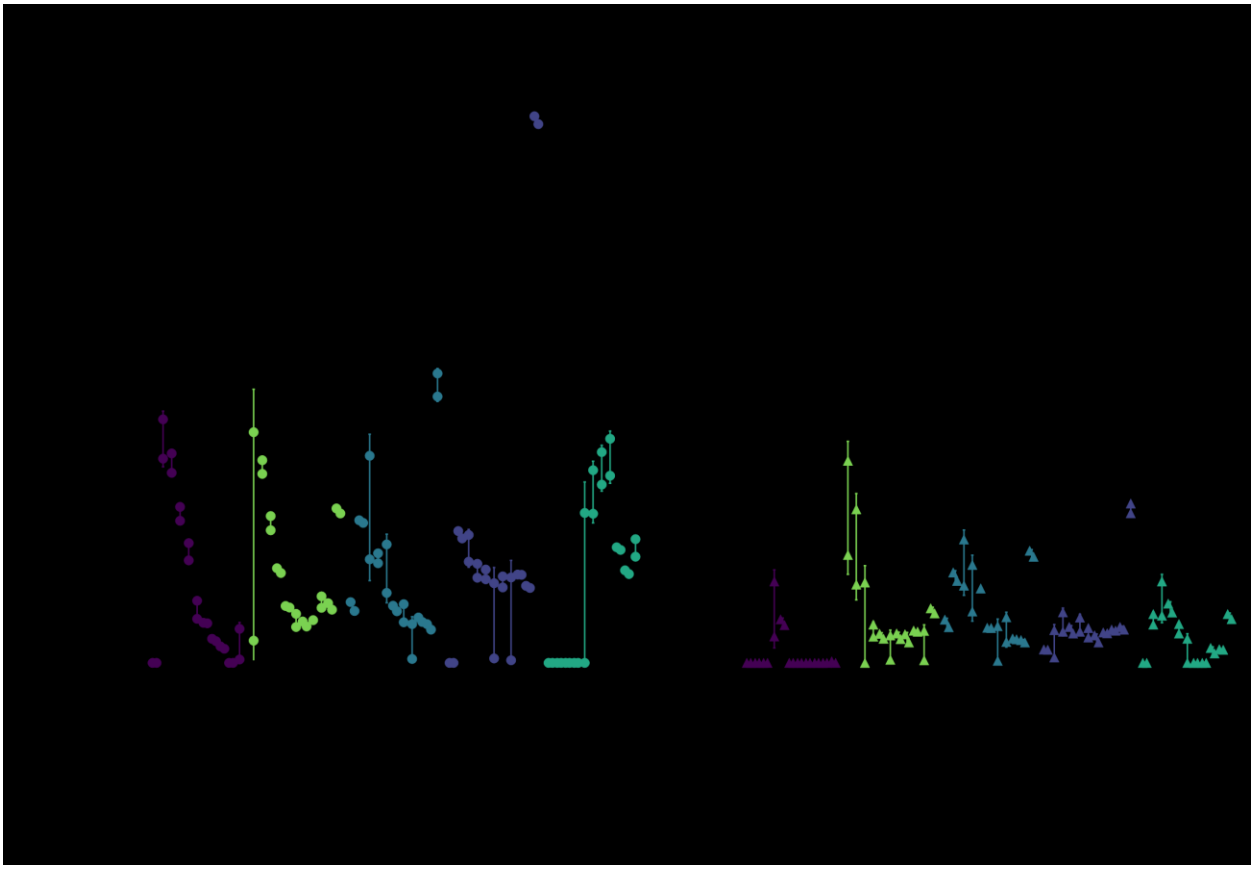


Figure S7. Extracted Ion Counts (EIC) of CoA intermediates of module 1. All shown values are the peak areas of the EICs of the quantifiers (see Table S3.). The vertical lines separate the different assays as labelled on the x-axis. The bars represent the timepoints 0, 0.5, 1, 1.5, 2, 4, 6, 8, 10, 12 and 24 h (from left to right) in each assay. A) Propionyl-CoA) B) Methylmalonyl-CoA C) Succinyl-CoA D) γ-hydroxybutyryl-CoA E) Crotonyl-CoA F) Ethylmalonyl-CoA G) Methylsuccinyl-CoA H) Mesaconyl-CoA I) β-methylmalyl-CoA

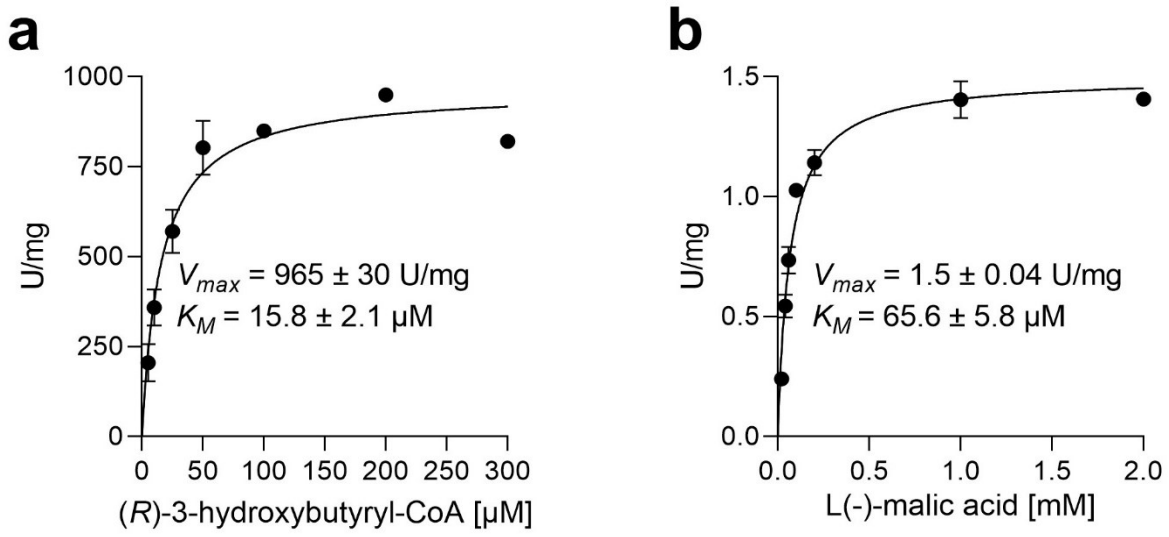


Figure S8. Enzyme Kinetics for Bbd (A) and Mtk (B)

4.6.5. Supplementary References

1. Ito, S., Koyama, N. & Osanai, T. Citrate synthase from Synechocystis is a distinct class of bacterial citrate synthase. *Sci Rep* **9**, 6038 (2019).
2. Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science* **291**, 1790-1792 (2001).
3. Schwander, T., Schada von Borzyskowski, L., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide in vitro. *Science* **354**, 900-904 (2016).
4. Erb, T.J., Retey, J., Fuchs, G. & Alber, B.E. Ethylmalonyl-CoA mutase from Rhodobacter sphaeroides defines a new subclade of coenzyme B12-dependent acyl-CoA mutases. *J Biol Chem* **283**, 32283-93 (2008).
5. Erb, T.J. University of Freiburg (2009).
6. Söhling, B. & Gottschalk, G. Purification and characterization of a coenzyme-A-dependent succinate-semialdehyde dehydrogenase from *Clostridium kluyveri*. *Eur. J. Biochem.* **212**, 121-127 (1993).
7. Konneke, M. et al. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proc Natl Acad Sci U S A* **111**, 8239-44 (2014).
8. Zarzycki, J. et al. Mesaconyl-coenzyme A hydratase, a new enzyme of two central carbon metabolic pathways in bacteria. *J Bacteriol* **190**, 1366-74 (2008).
9. Erb, T.J., Frerichs-Revermann, L., Fuchs, G. & Alber, B.E. The apparent malate synthase activity of Rhodobacter sphaeroides is due to two paralogous enzymes, (3S)-Malyl-coenzyme A (CoA)/{beta}-methylmalyl-CoA lyase and (3S)- Malyl-CoA thioesterase. *J Bacteriol* **192**, 1249-58 (2010).
10. Sevinc, M.S., Ens, W. & Loewen, P.C. The cysteines of catalase HPII of Escherichia coli, including Cys438 which is blocked, do not have a catalytic role. *Eur J Biochem* **230**, 127-32 (1995).
11. Hoelsch, K., Suhrer, I., Heusel, M. & Weuster-Botz, D. Engineering of formate dehydrogenase: synergistic effect of mutations affecting cofactor specificity and chemical stability. *Appl Microbiol Biotechnol* **97**, 2473-81 (2013).
12. Scheffen, M. et al. A new-to-nature carboxylation module to improve natural and synthetic CO2 fixation. *Nature Catalysis* **4**, 105-115 (2021).
13. Schada von Borzyskowski, L. et al. Marine Proteobacteria metabolize glycolate via the beta-hydroxyaspartate cycle. *Nature* **575**, 500-504 (2019).
14. Rommel, T.O., Hund, H.K., Speth, A.R. & Lingens, F. Purification and N-terminal amino-acid sequences of bacterial malate dehydrogenases from six actinomycetales strains and from Phenylobacterium immobile, strain E. *Biol Chem Hoppe Seyler* **370**, 763-8 (1989).
15. Ueda, Y., Yumoto, N., Tokushige, M., Fukui, K. & Ohya-Nishiguchi, H. Purification and characterization of two types of fumarase from Escherichia coli. *J Biochem* **109**, 728-33 (1991).
16. Mracek, J., Snyder, S.J., Chavez, U.B. & Turrens, J.F. A soluble fumarate reductase in Trypanosoma brucei procyclic trypomastigotes. *J Protozool* **38**, 554-8 (1991).
17. Nolte, J.C. et al. Novel characteristics of succinate coenzyme A (Succinate-CoA) ligases: conversion of malate to malyl-CoA and CoA-thioester formation of succinate analogues in vitro. *Appl Environ Microbiol* **80**, 166-76 (2014).
18. Schwander, T. Philipps University Marburg (2017).
19. Hugler, M., Menendez, C., Schagger, H. & Fuchs, G. Malonyl-coenzyme A reductase from Chloroflexus aurantiacus, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO(2) fixation. *J Bacteriol* **184**, 2404-10 (2002).
20. Bernhardsgrutter, I. et al. The multicatalytic compartment of propionyl-CoA synthase sequesters a toxic metabolite. *Nat Chem Biol* **14**, 1127-1132 (2018).

21. Saint Girons, I. et al. Structural and catalytic characteristics of Escherichia coli adenylate kinase. *J Biol Chem* **262**, 622-9 (1987).
22. Jordan, P.A., Tang, Y., Bradbury, A.J., Thomson, A.J. & Guest, J.R. Biochemical and spectroscopic characterization of Escherichia coli aconitases (AcnA and AcnB). *Biochem J* **344 Pt 3**, 739-46 (1999).
23. MacKintosh, C. & Nimmo, H.G. Purification and regulatory properties of isocitrate lyase from Escherichia coli ML308. *Biochem J* **250**, 25-31 (1988).
24. Oeding, V. & Schlegel, H.G. Beta-ketothiolase from Hydrogenomonas eutropha H16 and its significance in the regulation of poly-beta-hydroxybutyrate metabolism. *Biochem J* **134**, 239-48 (1973).
25. Liu, Q., Ouyang, S.P., Chung, A., Wu, Q. & Chen, G.Q. Microbial production of R-3-hydroxybutyric acid by recombinant E. coli harboring genes of phbA, phbB, and tesB. *Appl Microbiol Biotechnol* **76**, 811-8 (2007).
26. Peter, D.M., Vogeli, B., Cortina, N.S. & Erb, T.J. A Chemo-Enzymatic Road Map to the Synthesis of CoA Esters. *Molecules* **21**, 517 (2016).
27. Vogeli, B. et al. Combining Promiscuous Acyl-CoA Oxidase and Enoyl-CoA Carboxylase/Reductases for Atypical Polyketide Extender Unit Biosynthesis. *Cell Chem Biol* **25**, 833-839 e4 (2018).
28. von Borzyskowski, L.S. et al. Marine Proteobacteria metabolize glycolate via the β-hydroxyaspartate cycle. *Nature* **575**, 500-504 (2019).
29. Sundaram, S. et al. A modular in vitro platform for the production of terpenes and polyketides from CO2. *Angewandte Chemie International Edition* (2021).
30. Kitagawa, M. et al. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**, 291-9 (2005).
31. Schwander, T. Philipps-Universität Marburg (2018).
32. Bernhardsgrütter, I. et al. The multicatalytic compartment of propionyl-CoA synthase sequesters a toxic metabolite. *Nature chemical biology* **14**, 1127-1132 (2018).
33. Opgenorth, P.H., Korman, T.P. & Bowie, J.U. A synthetic biochemistry molecular purge valve module that maintains redox balance. *Nat Commun* **5**, 4113 (2014).
34. Egloff, K. Engineering of an artificial autotrophic CO2 fixation pathway in Methylobacterium extorquens. (Philipps-Universität Marburg, 2014).
35. Lowry, B. et al. In vitro reconstitution and analysis of the 6-deoxyerythronolide B synthase. *Journal of the American Chemical Society* **135**, 16809-16812 (2013).
36. Flamholz, A., Noor, E., Bar-Even, A. & Milo, R. eQuilibrator—the biochemical thermodynamics calculator. *Nucleic acids research* **40**, D770-D775 (2012).

5. Discussion and outlook

5.1. Machine learning guided optimization of the CETCH cycle

The publication of the CETCH cycle in 2016 was a milestone in synthetic biology. It was the first new-to-nature CO_2-fixation cycle, which was realized *in vitro* and harnessed the superior performance of Ccr. The workflow for the optimization was based on the identification of bottlenecks by measuring and identifying accumulation of intermediates of the pathway. The quantification of intermediates however, can be very time-consuming. A LC/MS method for each class of molecules is needed, which can lead to the demand of developing several methods and the measurement of each sample multiple times. Secondly, not each metabolite is stable enough to be suited for LC/MS measurements. Therefore, suitable high-throughput workflows remain a bottleneck in the realization- and omptimization-process of complex new pathways. The identification of the most important factors using METIS however, fulfills the same purpose by only measuring the product of the pathway and the analysis of the different tested system parameters. This allows accelerated optimization, reduces the requirements for expertise in analytics, and therefore makes such approaches feasible for research facilities with fewer resources. Besides overcoming major bottlenecks, the orchestration of enzymes from different metabolic backgrounds requires a compromise in assay parameters such as temperature and pH, leading to an environment in which only a few enzymes work at their maximum capability. The different metabolic backgrounds of the enzymes usually lead to side reactions. Enzymes naturally evolve within the metabolic framework of the host organism with a set of metabolites present. The presence of similar metabolites in a different metabolic context can lead to their transformation into dead-end products since the enzyme did not evolve to differentiate between the naturally occurring and the non-natural substrates. In theory, a well-tuned composition of enzyme amounts can minimize side reactions by an immediate turnover of critical metabolites as well as limiting the overabundance of critical enzymes.

The correlation of the production rate and the loss of intermediates hints towards a positive effect of higher fluxes, preventing the occurrence of side reactions and simultaneously reducing the potential of hydrolysis of and damage to intermediates by a constant turnover. A closer look at the determining factors predicted by the algorithm revealed only a few important components, while others could be present in various concentrations without having a major impact on the outcome. The most important factors predicted by the algorithm were Mco, Hbs and Coenzyme B_{12}. The need for high concentrations of Mco can be explained by its low activity compared to the other enzymes. Furthermore, its substrate methylsuccinyl-CoA has a half-life of 24 min and therefore needs to be rapidly converted to avoid

hydrolysis[1]. Mco is an engineered enzyme, which has only a fraction of its native activity[2]. Attempts to optimize the CETCH cycle *in vitro* should therefore aim at the further engineering of Mco to increase its activity or towards another way to dispose of electrons of the native Mcd reaction. While higher concentrations of Mco were beneficial, the opposite was the case for Hbs. When calculating the theoretical flux of the best conditions, Hbs was identified as the catalytic bottleneck of the whole pathway (data not shown). Why those low concentrations of Hbs are beneficial remains unclear, but they do indicate an interference with the pathway. Further testing on side reactivity, substrate independent ATP hydrolysis or contaminations from the purification are planned. Since none of the Coenzyme B_{12}-dependent enzymes (Mcm and Ecm) were highlighted as important factors, the beneficial concentrations of additional Coenzyme B_{12} (0 to 100 µM) seem to be sufficient. Coenzyme B_{12} is needed to form a radical intermediate during the rearrangement of the carbon skeleton by those mutases[3,4]. This complex reaction can lead to the inactivation of the enzyme itself and it was recently shown that a second enzyme (MeaB) protects and reactivates Mcm in both eukaryotes and prokaryotes[5,6]. Although none of the B_{12}-dependent enzymes are highlighted as potential bottlenecks, further experiments with MeaB in the context of the whole CETCH cycle are envisioned to assess whether the reactivation of Mcm can prolong pathway functionality.

Conclusively, we hope that our findings highlight the capability to identify bottlenecks and fine-tune biological networks by exclusively measuring the product of a pathway and encourages other scientists to use it for the faster prototyping of existing or new-to-nature systems. The guidance of an algorithm can help to understand non-obvious flaws of biological systems and stimulate follow-up experiments. Therefore, machine learning should be seen as a tool to explore the combinatorial space and give hints, which scientists still have to interpret and put into the context of the current research.

5.2. Extending the product portfolio of the CETCH cycle

The transformation of CO_2 into value added products is a very expensive procedure since most of those compounds are rather complex multi-carbon compounds and CO_2 sequestration is energy intensive. Therefore, most approaches in biotechnology aim to use sugars as the starting material, which are derived from plants. To harness the superior efficiency of the CETCH cycle, our goal was to showcase the direct conversion of CO_2 into complex molecules. We could demonstrate the feasibility of combining synthetic and natural pathways, such as the BHAC, into larger metabolic networks with the desired features for the production of specific molecules. While the production of terpenes was achieved by coupling enzymatic

cascades downstream of the CETCH cycle to convert glyoxylate into C5-C15 compounds, the anaplerotic feedback sequences enabled the production of molecules from CETCH cycles' intermediates. Beside the proven feasibility for the production of 6-dEB, there are many other compounds of interest which can be derived from the core cycle intermediates, such as polyhydroxybutyric acid, polylactic acid, crotonic acid or olefins. For the application in biotechnology (i.e. large-scale production) however, the complexity and the costs accompanied with protein purification and cofactor supply exceed the value of the compounds. Therefore, dramatic cost reduction by using cell lysate or the specialization on extremely expensive compounds, which are for example not producible *in vivo* due to toxicity, could close the gap between production costs and the value of the final product. Nevertheless, the integration of anaplerosis as a feature into a synthetic metabolic network does not only expand the product portfolio, but does add a layer of robustness that mimics the appearance of natural evolved pathways. Although those "complex" *in vitro* metabolic networks are orders of magnitude simpler than cells, the progress in building more complex setups will help to create even more adaptive systems to close the gap between *in vitro* and *in vivo*. For the future of synthetic metabolism, the implementation of more strategies adapted from nature to control *in vitro* metabolic networks will help to improve and understand fundamental designs of nature.

5.3. Further optimization of the CETCH cycle

Despite the improvement by a factor of >10 compared to the already manually optimized CETCH 5.4, the loss of intermediates reduces the production rates after ~60 minutes. Experiments to initialize a second production phase and revive the initial production rate by adding more substrate after 120 minutes were not successful (data not shown). Beside the most obvious explanations of enzyme wear-out or insufficient energy-supply, other factors such as inhibition by side products or damaged cofactors might contribute to reduced production rates. The character of our assays, which are usually started with energy equivalents and substrate once, is considerably passive compared to processes like continuous stirred-tank reactors or the highly adaptive metabolism in cells. The future optimization of the CETCH cycle, and *in vitro* systems in general, must therefore aim to adopt more strategies from nature to create truly adaptive environments. In contrast to many attempts of building a synthetic cell from bottom up, those strategies can be transferred to other setups to actually exceed the boundaries of nature and exploit the recent progress in engineering, like the emerging field of microfluidics. Microfluidics is mainly divided into two categories. Microfluidic chips are predominantly used to either form vesicles or droplets (often cell sized) or to build miniaturized reaction chambers with concomitant control elements, such as spatial and

temporal control, while simultaneously reducing the use of resources and therefore costs[7]. While vesicles and droplets are undoubtedly useful for studying cell-cell interactions or for screening approaches[8-10], their utilization is sometimes misused to claim a synthetic cell although adding only little to no advantage over bulk setups[11]. The complexity of already existing setups and the possibilities by using simulation guided chip designs are indicating that the aforementioned limitation of pathways by side reactions could be circumvented, leading to a higher efficiency of pathways. The division into smaller (thermodynamically favored) reaction blocks ending on stable intermediates could also mitigate the risk of loss of unstable intermediates. Beside the high level of control of substrate and energy supply through spatial and temporal regulation, the opportunity to govern additional layers of assay parameters like pH and temperature in the different chambers could lead to setups, which are actually exceeding the capabilities of nature.

Alongside the supply of substrate and energy equivalents, sufficient amounts of functional catalysts are vital for the integrity of pathways. The stability of enzymes can be highly dependent on their environment. While some of them are sensitive towards elevated temperatures or oxygen, others are working within a broad range of conditions. In the case of the CETCH cycle for example, the methylmalonyl-CoA and ethylmalonyl-CoA mutates are coenzyme B_{12} dependent enzymes. As mentioned earlier, the addition of helper enzymes to repair or reactivate enzymes such as MeaB could contribute to an extended assay. The same is true for other cofactors or molecules like NADH or NADPH, which can be damaged during the assay and where repair mechanisms exist in nature [12]. One strategy to counteract enzyme wear-out is the *in situ* production of proteins by cell free transcription-translation machinery. Some of the fixed carbon could be directly converted into amino acids to renew some of the critical enzymes in the CETCH cycle. Since two out of the four anaplerotic feedback sequences lead to TCA cycle intermediates, the de-novo biosynthesis of amino acids from α-ketoglutarate or oxaloacetate could be feasible. Beside the replenishment of proteins to endure the assays lifespan, the *in situ* production would open new possibilities for the optimization. Because the production and purification of enzymes is an expensive and laborious work, the number of mutants with potential gain of function properties that can be tested, is limited. Utilizing cell-free translation and transcription however, (un)targeted mutagenesis of genes could be employed, and those variants directly expressed together with the other enzymes to benchmark the performance within the assay. Afterwards, the information which variant outperformed the others could be retrieved from sequencing the DNA it was expressed from. Furthermore, genetic regulatory elements could be implemented to gain control over the assay. In contrast to engineering allosteric regulation, the tools for regulation of gene expression are easier to implement and faster to realize.

The shuttling of resources from inactive or damaged proteins to other parts of metabolism to scavenge carbons and energy, known as catabolism, is a common scheme in nature. Recently it was shown that proteins can be degraded and recycled into new proteins with a cell-free expression system *in vitro*[13]. Since the cell-free expression system itself consists of several proteins and numerous other components, the initial costs are high for purified components. The use of lysate however, could lower the cost to contribute to the efficiency of the assay. Besides the aforementioned protein-recycling, there are several pathways that salvage and degrade proteins into TCA cycle intermediates such as succinyl-CoA, acetyl-CoA or oxaloacetate[14]. Since a lot of carbon and energy is stored in enzymes, the conversion into a pathways' substrate could increase the productivity. The breakdown of already inactive catalysts could therefore endure the pathways lifetime and supply the pathway with energy and carbon for a final production phase to increase the yield.

5.4. Closing remarks

The progress in the field of synthetic biology led to astonishing new techniques and the realization of ideas that were thought to be science fiction a few decades ago. The combination of new screening techniques enabled by new technologies and instruments as well as decreasing prices in DNA synthesis led to the generation of enormous datasets. In parallel, the computational power and emergence of new algorithms to understand and predict biological systems enabled the use of the generated data. Artificial intelligence is already being employed for predicting protein structures or novel drugs, and is applied in almost all research areas where large (multi)omics datasets are available. While *in vitro* systems offer a great control to study their behavior, their application in real world processes for production is often limited to compounds, which cannot be synthesized chemically or are too toxic for the production *in vivo.* To unleash the potential of synthetic designer pathways it is mandatory to transfer them into organisms to lower the costs and labor. The great challenge will be the accessibility of non-model organisms with superior characteristics for tailor-made pathways, capable of accommodating the properties of such compounds. A first step will be to bridge the gap of *in vitro* and *in vivo* by engineering new screening platforms to study the interference of the host metabolism and new pathways. The 2021 iGEM team from Marburg developed such cell-free systems of various plants to tackle the problem of the slow process of plant engineering[15]. The prototyping with those approaches speeds up the transition process for the transplantation of complex pathways into production organisms. Together with recently developed CRISPR/Cas based genetic modification tools, the whole development from synthetic pathways designed

on paper to new organisms with superior characteristics will be faster than ever before and will help to tackle some of the most challenging problems humanity faces today. While the emergence of new technologies to fight climate change and world hunger indicate that it is possible to build a sustainable and fair world, the increasing gap between rich and poor mandates decisive actions of governments to promote and enforce the use of those technologies for a greater good. The ongoing corona pandemic and its consequences illustrate this quite well: While it was possible to develop a new type of vaccines within months, its production and distribution is largely limited to developed countries. Furthermore, the increasing disenchantment with politics and industry leads to an increased distrust in science. Therefore, the progress in science needs to be accompanied by sophisitcated societies, politicians, and scientists who are striving to make these new technologies accessible for everyone.

5.5. References

1. Burgener, S., Schwander, T., Romero, E., Fraaije, M.W. & Erb, T.J. Molecular Basis for Converting (2S)-Methylsuccinyl-CoA Dehydrogenase into an Oxidase. *Molecules* **23**(2017).
2. Schwander, T., von Borzyskowski, L.S., Burgener, S., Cortina, N.S. & Erb, T.J. A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900-904 (2016).
3. Erb, T.J., Retey, J., Fuchs, G. & Alber, B.E. Ethylmalonyl-CoA mutase from Rhodobacter sphaeroides defines a new subclade of coenzyme B12-dependent acyl-CoA mutases. *J Biol Chem* **283**, 32283-93 (2008).
4. Takahashi-Iniguez, T., Garcia-Hernandez, E., Arreguin-Espinosa, R. & Flores, M.E. Role of vitamin B12 on methylmalonyl-CoA mutase activity. *J Zhejiang Univ Sci B* **13**, 423-37 (2012).
5. Korotkova, N. & Lidstrom, M.E. MeaB Is a Component of the Methylmalonyl-CoA Mutase Complex Required for Protection of the Enzyme from Inactivation*. *Journal of Biological Chemistry* **279**, 13652-13658 (2004).
6. Toraya, T. G-protein signaling: A switch saves B12 radical status. *Nat Chem Biol* **9**, 530-1 (2013).
7. Convery, N. & Gadegaard, N. 30 years of microfluidics. *Micro and Nano Engineering* **2**, 76-91 (2019).
8. Sun, J., Warden, A.R. & Ding, X. Recent advances in microfluidics for drug screening. *Biomicrofluidics* **13**, 061503-061503 (2019).
9. Sharma, B., Moghimianavval, H., Hwang, S.-W. & Liu, A.P. Synthetic Cell as a Platform for Understanding Membrane-Membrane Interactions. *Membranes* **11**, 912 (2021).
10. Du, G., Fang, Q. & den Toonder, J.M.J. Microfluidics for cell-based high throughput screening platforms—A review. *Analytica Chimica Acta* **903**, 36-50 (2016).
11. Jiang, S., Caire da Silva, L., Ivanov, T., Mottola, M. & Landfester, K. Synthetic Silica Nano-Organelles for Regulation of Cascade Reactions in Multi-Compartmentalized Systems. *Angew Chem Int Ed Engl* **61**, e202113784 (2022).
12. Niehaus, T.D. et al. Plants utilize a highly conserved system for repair of NADH and NADPH hydrates. *Plant physiology* **165**, 52-61 (2014).
13. Giaveri, S. et al. Nature-Inspired Circular-Economy Recycling for Proteins: Proof of Concept. *Advanced Materials* **33**, 2104581 (2021).
14. Adeva-Andany, M.M., Lopez-Maside, L., Donapetry-Garcia, C., Fernandez-Fernandez, C. & Sixto-Leal, C. Enzymes involved in branched-chain amino acid metabolism in humans. *Amino Acids* **49**, 1005-1028 (2017).
15. iGEM_Team_Marburg_2021. OpenPlast: Establishing cell-free systems from chloroplasts as rapid prototyping platforms for plant SynBio (2021).

Danksagung

Zehn Jahre Marburg, zehn Jahre Studium und Promotion kommen nun zu einem Ende. An erster Stelle möchte ich meiner Familie und vor allem meinen Eltern danken. Neben all den schönen Erfahrungen während Studium und Promotion, war dies auch oft eine Zeit der Selbstzweifel. Ohne euer bedingungsloses Vertrauen in mich und meine Entscheidungen würde ich heute diese Arbeit nicht einreichen. Jeder Mensch braucht sein eigenes Tempo um sich zu entwickeln, und ich bin unendlich dankbar meine Zeit dafür bekommen zu haben. Auch wenn es tausend wichtigere Dinge im Leben gibt als Abschlüsse und Titel, hoffe ich doch, dass ich heute mit dieser Arbeit etwas für euer Vertrauen zurückgeben kann.

Ein großer Dank geht an meine außerwissenschaftlichen Freunde. In Siegen, in Düsseldorf, in Köln, in Lüneburg, in Hamburg oder wo auch immer ihr gerade seid. Es bedeutet mir unglaublich viel, so viele von euch schon so lange zu kennen. Es ist ein beruhigendes Gefühl zu wissen, dass es euch gibt und es noch ein Leben außerhalb des wissenschaftlichen Kosmos gibt. Dass ich immer zu euch kommen kann, wir erstmal ein Bierchen trinken und die Probleme Probleme sein lassen.

Auch bin ich unfassbar glücklich, dass ich meine Arbeit in diesem Umfeld machen durfte. Wissenschaft ist nicht immer einfach. Aber ihr alle habt es zu einer so einer wunderbaren Zeit gemacht, dass ich mit mehr als nur einem weinenden Auge gehen werde. Dabei geht es nicht nur um den Zusammenhalt und die Unterstützung im Labor, sondern auch um die wunderschöne Zeit außerhalb der Arbeit. In euch habe ich nicht nur fantastische Kollegen, sondern auch Freunde gefunden. Einen großen Anteil daran hast du, Tobi. Viel hat sich verändert in den letzten viereinhalb Jahren und nicht alles verläuft ohne Probleme. Aber dein Vertrauen in uns das Richtige zu tun, und uns die Entscheidungsfreiheit zu geben in welche Richtung unsere wissenschaftliche Entwicklung verläuft, ist einzigartig. Vielen Dank, dass du mir damals das Projekt anvertraut und mich in den letzten viereinhalb Jahren auf dieser Reise begleitet hast. Möge der Spirit der AG Erb auch über die nächsten Generationen beibehalten werden, damit eure Namen noch über Jahre hinaus im Fünf Jahreszeiten bekannt sind!

Vielen Dank an alle mit denen ich in den letzten Jahren publizieren durfte. Marieke, Tarryn, Vidhya, Amir, allen anderen beteiligten Personen und natürlich Patrick. Unsere als „Nebenprojekt“ gestartete Arbeit ist mir durch unsere gemeinsame, mit Memes gespickte, Arbeit richtig ans Herz gewachsen!

Ein besonderer Dank geht an Jan und nochmals Patrick, welche mich beim Schreiben der Arbeit unterstützt haben. Ohne Sir Patricks exquisite Formulierungen wäre diese Arbeit mit Sicherheit deutlich

schwerer zu lesen. Des weiteren ein herzliches Dankeschön an alle anderen, die während meiner Zeit am MPI so oft beim Ausarbeiten von Vorträgen und Texten geholfen haben.

Ein ganz besonderes Dankeschön möchte ich der Prüfungskommission aussprechen. Prof Dr. Hans-Ullrich Mösch, welcher sich als Zweitgutachter bereit erklärt hat. Vielen Dank für Ihre Mühe! Prof. Dr. Lars-Oliver Essen, welcher das Projekt seit Tag eins als Mitglied meines Thesis Advisory Committees begleitet hat. Prof. Dr. Hannes Link, ebenfalls als Mitglied meines Thesis Advisory Committees und als Betreuer meiner Masterarbeit. Durch die Arbeit bei dir und die fantastische Zeit damals habe ich erst so richtig den Spaß an der Wissenschaft entdeckt!

So endet nun langsam meine Zeit in Marburg. Zum Abschied möchte ich mich mit einem wednesdayesquen ahhhhhhh verabschieden und sehen wohin mich meine Beine (Vaaaaaaaaaaaater) tragen. Bis bald, Marburg!