ARTICLES
https://doi.org/10.1038/s41589-020-0559-0

nature chemical biology

Check for updates

In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design

Ashty S. Karim[1,2,3], Quentin M. Dudley[1,2,3], Alex Juminaga[4], Yongbo Yuan[4], Samantha A. Crowe[1,2,3], Jacob T. Heggestad[1,2,3], Shivani Garg[4], Tanus Abdalla[4], William S. Grubbe[1,2,3], Blake J. Rasor[1,2,3], David N. Coar[5], Maria Torculas[5], Michael Krein[5], FungMin (Eric) Liew[4], Amy Quattlebaum[4], Rasmus O. Jensen[4], Jeffrey A. Stuart[5], Sean D. Simpson[4], Michael Köpke[4] ✉ and Michael C. Jewett[1,2,3,6,7] ✉

The design and optimization of biosynthetic pathways for industrially relevant, non-model organisms is challenging due to transformation idiosyncrasies, reduced numbers of validated genetic parts and a lack of high-throughput workflows. Here we describe a platform for in vitro prototyping and rapid optimization of biosynthetic enzymes (iPROBE) to accelerate this process. In iPROBE, cell lysates are enriched with biosynthetic enzymes by cell-free protein synthesis and then metabolic pathways are assembled in a mix-and-match fashion to assess pathway performance. We demonstrate iPROBE by screening 54 different cell-free pathways for 3-hydroxybutyrate production and optimizing a six-step butanol pathway across 205 permutations using data-driven design. Observing a strong correlation ($r = 0.79$) between cell-free and cellular performance, we then scaled up our highest-performing pathway, which improved in vivo 3-HB production in *Clostridium* by 20-fold to 14.63 ± 0.48 g l^{-1}. We expect iPROBE to accelerate design–build–test cycles for industrial biotechnology.

For decades, scientists and engineers have turned to biological systems for making energy, medicines and materials, especially when chemical synthesis is untenable[1]. Success in these endeavors depends upon identifying sets of enzymes that can convert readily available substrate molecules (for example, glucose) to target products, with each enzyme performing one of a series of chemical modifications. Unfortunately, this is difficult because design–build–test (DBT) cycles (iterations of re-engineering organisms to test and optimize new sets of enzymes) are slow, especially with the high number of testable enzyme combinations in multistep pathways[2]. This challenge is exacerbated in industrially relevant, non-model organisms for which genetic tools are not as sophisticated, high-throughput workflows are often lacking, transformation idiosyncrasies exist and validated genetic parts are limited.

Yet, many industrial bioprocesses (for example, synthesis of solvents[3]) rely on non-model organisms as they offer exceptional substrate and metabolite diversity, as well as tolerance to metabolic end-products and contaminants. Clostridia in particular were used for industrial acetone–butanol–ethanol (ABE) fermentations in the early-to-mid 20th century because of their unique solventogenic metabolism, but were eventually phased out of use due to the success of petroleum[4]. Acetogenic Clostridia, able to robustly ferment on a variety of abundant C1 gases[5], have recently proven industrially relevant for full commercial-scale ethanol production using emissions from the steelmaking process[6]. However, these strains tend to lack natural machinery to produce such solvents or other more complex products and the tools to engineer them are underdeveloped. While developing tools for engineering Clostridia is ongoing and promising progress has been made[7], discovering methods to speed up metabolic engineering DBT cycles for these and other non-model organisms would accelerate the re-industrialization of such organisms[8].

Cell-free systems provide many advantages for accelerating DBT cycles[9–11]. For example, the open reaction environment allows direct monitoring and easy manipulation of the system. As a result, many groups have used purified systems to study enzyme kinetics and inform cellular expression: testing enzymatic pathway performance in vitro, downselecting promising pathway combinations and implementing those in cells[12,13]. Crude lysates are becoming an increasingly popular alternative to purified systems to build biosynthetic pathways because they provide native-like metabolic networks as well as negate the need for protein purification[14–17]. For instance, dihydroxyacetone phosphate can be made in crude lysates and real-time monitoring can optimize production[16]. In addition, our group has shown that 2,3-butanediol[18], mevalonate[15], *n*-butanol[14,19], limonene[20] and styrene[21] can be made in crude lysates with high productivities. However, to our knowledge, no attempts have been made using cell-free prototyping to improve engineering of industrially relevant, non-model organisms.

To address this opportunity, we report a new iPROBE approach to inform cellular metabolic engineering. The foundational principle is that we can construct discrete enzymatic pathways through modular assembly of cell lysates, containing pathway enzymes produced by cell-free protein synthesis, making the DBT unit cellular lysates rather than genetic constructs or a re-engineered organism (Fig. 1). This reduces the overall time to build pathways from weeks (or months) to a few days, providing an increased capability to test numerous pathways with large numbers of enzyme combinations.

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [2]Chemistry of Life Processes Institute, Northwestern University, Evanston, IL, USA. [3]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA. [4]LanzaTech Inc., Skokie, IL, USA. [5]Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ, USA. [6]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, USA. [7]Simpson Querrey Institute, Northwestern University, Chicago, IL, USA. ✉e-mail: Michael.Koepke@lanzatech.com; m-jewett@northwestern.edu

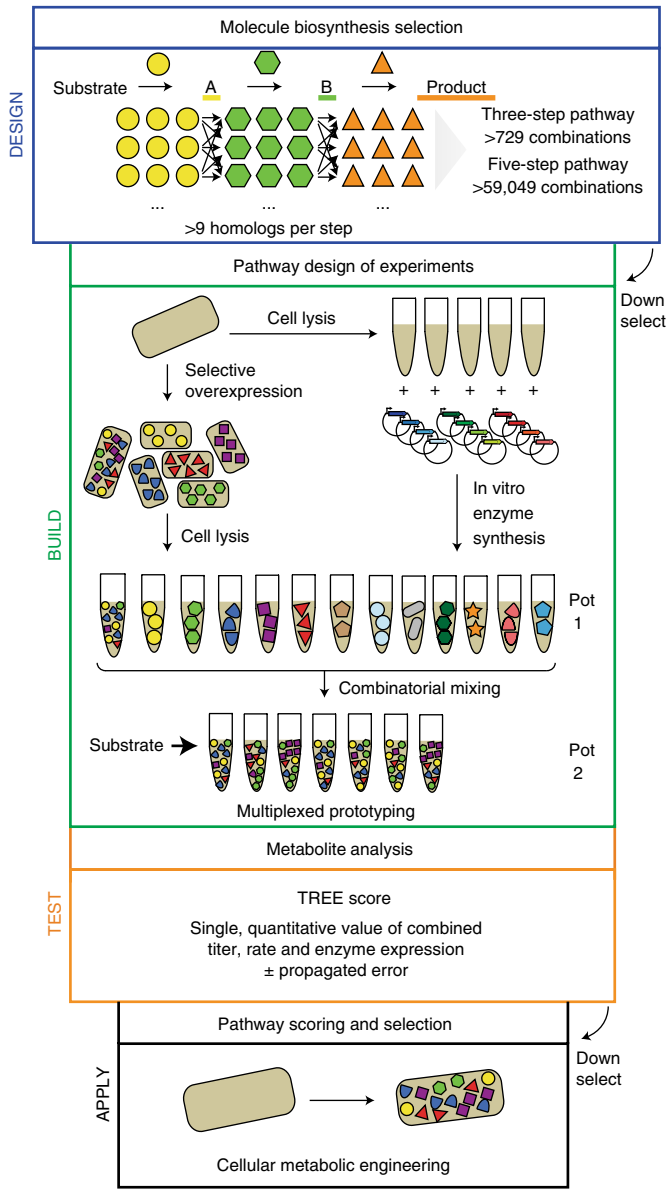

Fig. 1 | A two-pot cell-free framework for iPROBE. A schematic overview of the iPROBE approach following a DBT and apply framework is depicted. Reaction schemes and enzyme homologs are selected (design). Lysates are enriched with pathway enzymes via overexpression before lysis or by CFPS post-lysis and mixed to assemble enzymatic pathway combinations (build). Metabolites are quantified over time and data are reduced into a single quantitative metric for pathway ranking (test). Cell-free selected pathways are implemented in cellular hosts (apply).

We demonstrate iPROBE by optimizing biosynthetic pathways for the production of 3-hydroxybutyrate (3-HB) and *n*-butanol in *Clostridium autoethanogenum*, revealing a strong correlation ($r=0.79$) between in-cell and cell-free pathway performance. Then, we show that we can scale up the best 3-HB-producing *C. autoethanogenum* strain, containing an iPROBE-selected pathway to achieve the highest reported titers of 3-HB at rates of $>1.5\,g\,l^{-1}\,h^{-1}$ in a continuous system using low-cost waste gas as a feedstock.

Results

Establishing the iPROBE framework. Our vision was to demonstrate modular assembly of biosynthetic pathways by mixing multiple *Escherichia coli* (*Eco*) crude cell lysates, each individually enriched with a pathway enzyme, to identify best sets and ratios of enzymes and inform cellular design in an industrially proven[5], non-model host organism, in this case acetogenic *C. autoethanogenum* (Fig. 1). A unique feature of the iPROBE approach, relative to previous work in crude lysate-based cell-free prototyping[14,19,20,22], is that pathways are assembled in two steps (that is, two pots). The first step is enzyme synthesis via cell-free protein synthesis (CFPS) and the second step is enzyme utilization via substrate and cofactor addition to activate small-molecule synthesis (Supplementary Fig. 1). The two-pot iPROBE workflow is important for three reasons. First, it allows for control of enzyme concentration in pathway construction by precise quantification of protein expression yields. Second, the control of enzyme concentrations allows us to assess pathway performance as a function of changing enzyme ratios and ensures enzyme balance. Third, negative physiochemical effects of the CFPS reaction mixtures[19] on small-molecule biosynthesis can be reduced by implementing the controllable two-pot approach.

We selected 3-HB biosynthesis as our first demonstration because it is non-native to *C. autoethanogenum* and because of its importance as a high-value specialty chemical[23]. We first set out to use iPROBE to study the impact of enzyme ratios on pathway performance (Fig. 2a). From acetyl-CoA, a key intermediate in both *E. coli* and *Clostridium*, three enzymes are required to make 3-HB: a thiolase (Thl), a hydroxybutyryl-CoA dehydrogenase (Hbd) and a thioesterase. *E. coli* and *C. autoethanogenum* have native thioesterases that convert 3-hydroxybutyryl-CoA to 3-HB[24], which are often required to be overexpressed for optimal production[25]. However, for screening purposes these native enzymes were sufficient; only two non-native enzymes (Thl and Hbd) were required to be overexpressed in *C. autoethanogenum* cells and *E. coli* lysates. We initially selected a Thl gene from *Clostridium acetobutylicum* (*Cac*) and a Hbd gene from *Clostridium kluyveri* (*Ckl*) (Supplementary Table 1) and expressed them using the *E. coli*-based PANOx-SP CFPS system[26], with soluble concentrations of $5.85\pm0.82\,\mu M$ and $19.31\pm3.65\,\mu M$, respectively. Then, we designed five unique pathway combinations titrating different concentrations of Thl while maintaining a constant concentration of Hbd by mixing different ratios of CFPS reactions (keeping total CFPS reaction added as constant using 'blank' reactions containing no protein produced in vitro). Upon incubation with essential substrates, salts and cofactors (for example, glucose, NAD and CoA), we assessed 3-HB synthesis over time (Fig. 2b). The cell lysate contains endogenous enzymes for glycolysis that regenerate NADH[27] and convert glucose to acetyl-CoA, providing the starting intermediate for 3-HB biosynthesis. As expected, no 3-HB was produced in the absence of Thl. The highest 3-HB titers were observed for $0.5\,\mu M$ *Cac*Thl and $0.5\,\mu M$ *Ckl*Hbd1. We performed a similar titration of *Ckl*Hbd1 while maintaining a constant concentration of *Cac*Thl (Supplementary Fig. 2).

Developing a metric to quantify pathway performance. We next defined a pathway ranking system to assess pathway activity and inform cellular design. The basis of this ranking system is a single, quantitative metric for our cell-free experiments. We call this metric the TREE score (titer, rate and enzyme expression). The TREE score combines, through multiplication, titer at reaction completion, rate during the most productive phase of biosynthesis and enzyme expression as measured by soluble protein fraction and total enzyme amount. Using our initial set of data (Fig. 2) as a guide, the TREE score is obtained by multiplying 3-HB titer at 24 h, the linear 3-HB production rate between 3 and 6 h and the sum of the average soluble fraction of the pathway enzymes, Thl and Hbd and the inverse of the total enzyme concentration for each of the five pathway combinations (Fig. 2c). While the TREE score rankings are not largely different from the titers ($r=0.89$ for all 3-HB data in this study) or rates ($r=0.91$ for all 3-HB data in this study) alone

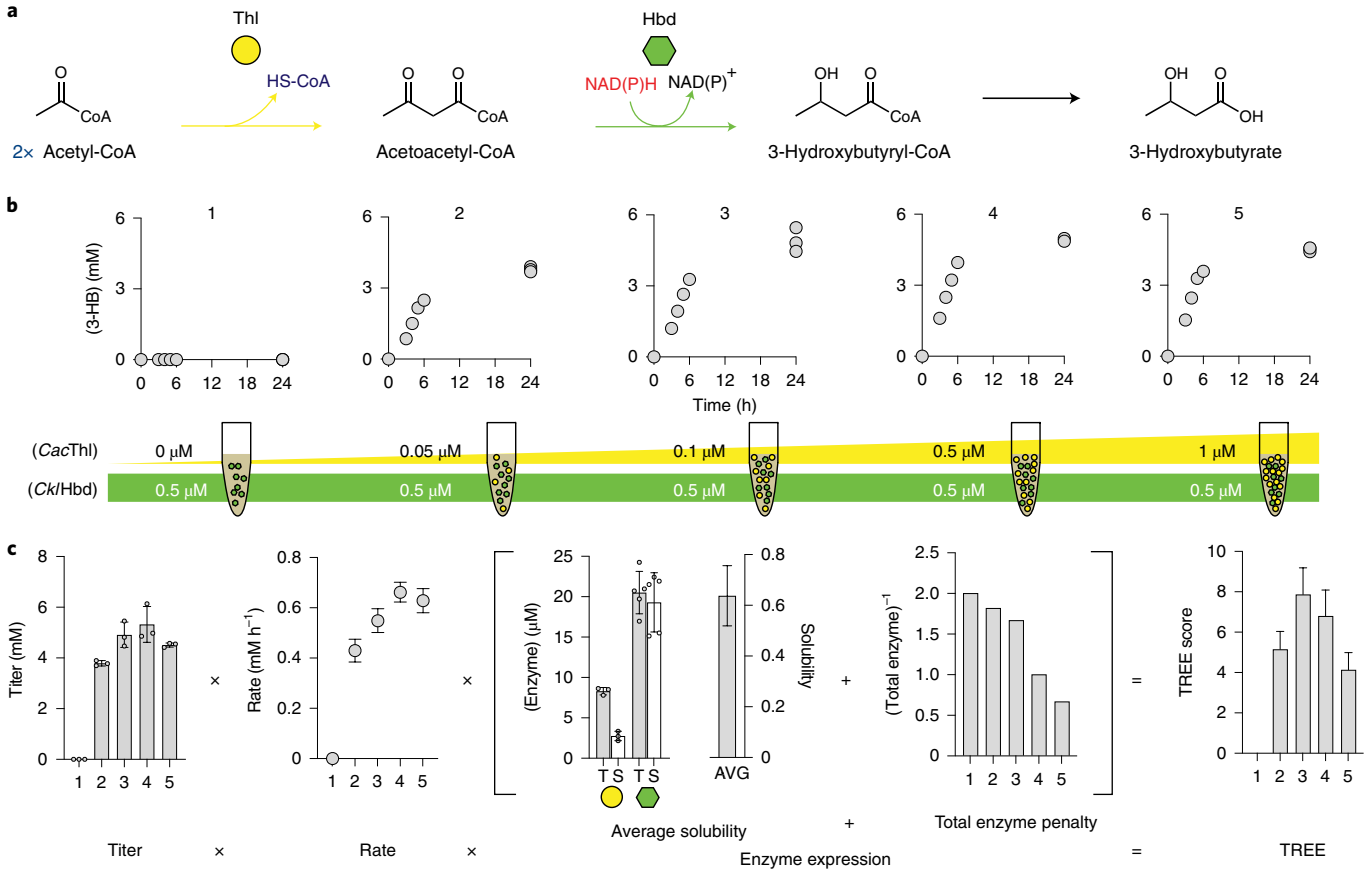

Fig. 2 | Individual pathway enzymes can be tuned in a pathway context and ranked using TREE scores with iPROBE. a, The pathway to produce 3-HB from native metabolism (acetyl-CoA) is presented. **b**, Five pathway designs, titrating *Cac*Thl concentrations, are built with *E. coli* lysates enriched with *Cac*Thl and *Ckl*Hbd1 by CFPS and 3-HB is measured at 0, 3, 4, 5, 6 and 24 h after the addition of glucose. All data are plotted for $n = 3$ independent experiments. **c**, For each of the pathways (1, 2, 3, 4 and 5) the following are shown: 3-HB titer at 24 h ($n = 3$ with bars representing the mean and error bars indicating 1 s.d.), calculated rate of production from linear regression (slope) through 6 h (bars represent standard error of the regression slope), quantified soluble (S) and total (T) enzyme expression ($n = 3$ for *Cac*Thl, $n = 6$ for *Ckl*Hbd1, with bars representing the mean and error bars indicating 1 s.d.) and corresponding average (AVG) soluble fraction of both enzymes with propagated error, the inverse of the total concentration of exogenous soluble enzyme added and the calculated TREE for each pathway combination ($n = 3$ independent experiments) with error bars representing the propagated error as described in Methods.

(all 3-HB pathway TREE scores shown in Supplementary Fig. 3), they exaggerate differences that might arise from each component of the score. For example, combining titer and rate enables use of both in ranking cell-free pathway performance, which is helpful as it is unknown whether one is more or less important for informing cellular design. Additionally, we included the enzyme expression component to penalize a given pathway if in vitro expression is poor, decreasing its overall pathway rank. Typically, enzymes that are either lowly expressed or insoluble in vitro are challenging to express in vivo. Thus, the average solubility of all pathway enzymes overexpressed in the lysate was used to acquire a sense of how difficult the pathway might be to express. The inverse enzyme amount was used to penalize in vitro combinations that might improve a pathway's performance but could be hard to express in cells. While there are multiple ways one could imagine ranking pathways or weighting the TREE score factors, reducing the complexity of available cell-free data was important as it enabled a rapid approach to rank pathways for iPROBE.

iPROBE informs plasmid design in *Clostridium*. We next aimed to validate that cell-free experiments could generate design parameters for DNA construction of biosynthetic pathways in cells, a difficult challenge because gene expression tools are often not as developed in non-model organisms. Selection of promoter regulatory strengths (for example, high, medium and low) for the expression of a coding sequence, in particular, is an essential factor in pathway tuning. Thus, we set out to develop a correlation between specific enzyme concentrations in iPROBE and specific strength regulatory architectures, relative promoter strengths and plasmid copy number for a single operon comprising the 3-HB pathway, for expression in *C. autoethanogenum*. To achieve this goal, we built cell-free pathway combinations for 3-HB by co-titrating (equimolar additions) seven different enzyme concentrations of *Cac*Thl and *Ckl*Hbd1 in our reactions (Supplementary Fig. 4). We ran each cell-free reaction for 24 h and measured the titer of 3-HB produced (Supplementary Fig. 4c). We observed that as the amount of added enzyme increases, the amount of 3-HB increases up to a threshold of 1 μM of each enzyme added. In parallel, we constructed plasmids expressing *Cac*Thl and *Ckl*Hbd1 under eight regulatory architectures of increasing strength and transformed them into separate strains of *C. autoethanogenum*. We ran small-scale bottle fermentations of each strain under anaerobic conditions on carbon monoxide (CO), hydrogen (H_2) and carbon dioxide (CO_2) gas and measured stationary phase titers of 3-HB (Supplementary Fig. 4d). In vivo, we found that increases in expression strength led to higher 3-HB titers, but did not saturate 3-HB expression, as seen in vitro (Supplementary Fig. 4).

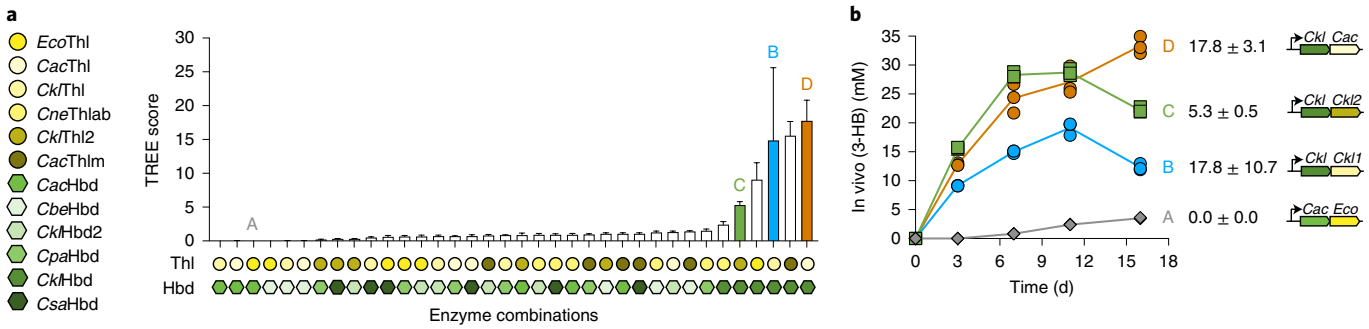

Fig. 3 | Enzymatic pathways can be screened with iPROBE to inform *Clostridium* expression for optimizing 3-HB production. Six homologs were selected for each reaction step. Each possible combination was constructed in cell-free systems using 0.5 μM of each enzyme (36 unique pathway combinations). **a**, 3-HB was measured and TREE scores were calculated and plotted for each iPROBE pathway combination from three independent experiments with propagated error as described in Methods. **b**, We then selected four pathway combinations to test in *C. autoethanogenum* (A, B, C and D). These pathways were built in high-copy plasmids with the highest-strength promoters in single operons. 3-HB was measured for *C. autoethanogenum* fermentations harboring these plasmids. All data are shown from $n = 3$ independent biological samples with lines running through the mean. Labels (A, B, C and D), corresponding TREE scores and genetic designs are used to signify corresponding cell-free combinations.

These data suggest a limitation in expression range with current, well-characterized genetic parts available for use in *C. autoethanogenum*. However, given the trends observed, we used these data to build an initial cell-free-to-cell correlation that connects cell-free enzyme concentrations in iPROBE to cellular plasmid regulatory strength. We found that generally, using <0.1 μM enzyme in vitro corresponds to low regulatory strengths in vivo, using 0.1–0.3 μM enzyme in vitro corresponds to medium strengths in vivo and using >0.3 μM enzyme in vitro corresponds to high strengths in vivo. In principle, this allows us to screen many different pathway combinations in cell-free systems, a key advantage of iPROBE, and provides a rational recommendation for plasmid construction of those pathway combinations in *Clostridium*.

iPROBE down-selects pathways for *Clostridium* expression. To showcase the iPROBE approach, we next screened several possible 3-HB pathway combinations using cell-free experiments, ranked a subset of candidate combinations using the TREE score and showed cellular *C. autoethanogenum* 3-HB biosynthesis from $CO/H_2/CO_2$ gas correlates with cell-free experimental results. To do this, we tested six enzyme homologs of each Thl and Hbd originating from different *Clostridium* species, as these would be the best initial candidates for *Clostridium* expression (Fig. 3a and Supplementary Table 1). We selected all pathway combinations of the 12 enzymes, keeping a fixed total concentration of soluble enzyme added (Supplementary Fig. 5). By measuring 3-HB production over the course of 24 h, along with soluble enzyme expression for each of the enzymes, we are able to calculate TREE scores for each of the 36 pathway combinations in vitro (Fig. 3a). We found that a majority of pathway combinations performed poorly and used iPROBE to identify that the top six pathways contained *Ckl*Hbd1 for the second step. Testing 36 pathway combinations took less than a week to build. In contrast this could have taken more than 6 months in *Clostridium* using standard workflows.

We selected a subset of four pathway combinations from the iPROBE screening to test in *C. autoethanogenum* labeled A (*Ckl*Thl1/*Cac*Hbd), B (*Ckl*Thl2/*Ckl*Hbd1), C (*Ckl*Thl1/*Ckl*Hbd1) and D (*Cac*Thl/*Ckl*Hbd1) (Fig. 3b). These pathways represent our highest-performing pathway, two in the middle (C having a large degree of variability in performance) and one of our low performers. We constructed and transformed DNA with strong regulatory architectures and each of the four pathway enzyme sets into separate strains of *C. autoethanogenum*. We ran small-scale bottle fermentations of each strain under anaerobic conditions on $CO/H_2/CO_2$ gas mixture and measured 3-HB titers at four time points during the fermentation (Fig. 3b). We observed that the best cell-free pathway combination as determined by TREE score (D) also performed the best in *Clostridium* cells, achieving 33.3 ± 1.4 mM 3-HB. The worst pathway combination in cell-free experiments (A) was also the worst performer in *C. autoethanogenum*. The other two pathway combinations (B and C) were not statistically different in the cell-free environment. The exact ranking of pathways B, C and D differ between in vitro and in vivo construction, but all three of these designs were much better than a majority of the 36 combinations tested in the cell-free environment. Notably, we did not observe detectable levels of nonspecific byproducts such as acetoacetate or acetone.

Even though the *E. coli*-based lysate conditions do not necessarily approximate the in vivo Clostridia conditions (for example, pH and aerobic versus anaerobic), our data highlight that the cell-free environment is a powerful prototyping environment for assessing biochemical information and informing cellular design. This is especially true for downselecting pathway combinations that should not be tested in cells (that is, produce little to no product). In fact, the best pathway designs tested in two recently published studies that explored autotrophic 3-HB production in acetogenic Clostridia produced ~4 mM and ~1 mM 3-HB[28,29]. Their pathways correspond to TREE scores of 1.06 ± 0.07 and 0.02 ± 0.00, respectively. Based on our iPROBE screening, we would have suggested not testing these combinations in vivo. For context, our best pathway had a TREE score of 17.76 ± 3.08. In sum, iPROBE offers a modular framework to rapidly assess a large number of pathway combinations, bypassing DNA construction and transformation limitations to facilitate implementation of promising pathway combinations for engineering success in cells.

Cell-free pathway prototyping for *n*-butanol biosynthesis. We next aimed to show that iPROBE could be used to optimize longer pathways. We selected the six-step pathway from acetyl-CoA to *n*-butanol as a model because butanol is an important solvent and drop-in fuel with a US$5 billion per year market (Fig. 4a). The idea was to use iPROBE to optimize cell-free butanol production by constructing several pathway variants. The challenge with this optimization goal is the number of possible permutations. Indeed, testing just six homologs for each of the first four steps of the pathway at three different enzyme concentrations would alone require 314,928 pathway combinations, which exceeds typical analytical pipelines. To manage the landscape of testable hypotheses, we implemented

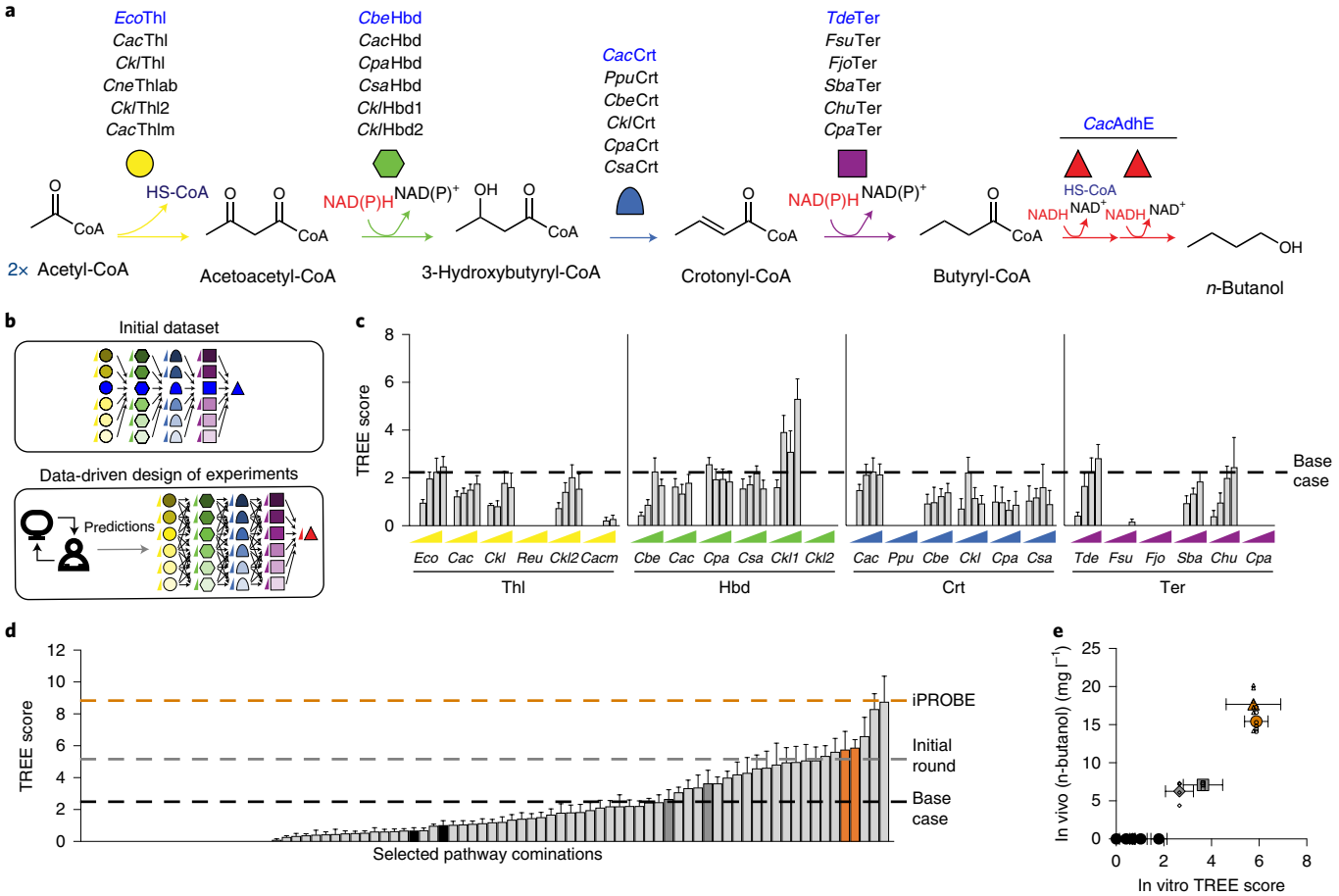

Fig. 4 | Cell-free pathway testing combined with data-driven design of experiments quickly screens 205 unique pathway combinations and selects pathways for cellular butanol production. a, A reaction scheme and corresponding enzyme homologs for the production of butanol are presented. **b**, An initial dataset is collected from reactions testing each homolog at five concentrations individually with the base-case set of enzymes (blue). Then, neural network-based design of experiments is implemented using the data presented in **c** to predict enzyme sets that are constructed with iPROBE. **c**, TREE scores were calculated from 24-h butanol time courses of the initial 120 pathway combinations. The TREE score for the base-case set is represented by a dashed line. **d**, Eighty-three predicted and hand-selected pathway combinations were built and TREE scores were calculated for each. The black dashed line represents the base-case set of enzymes, the gray dashed line corresponds to the best case from **c** and the orange dashed line represents the highest TREE score achieved through the data-driven iPROBE approach. **e**, Nine pathway combinations were constructed and transformed in *C. autoethanogenum*. End-point in vivo titers of butanol production ($n = 6$ for circles and triangles; $n = 4$ for squares and diamonds; y axis error bars represent 1 s.d.) are plotted against in vitro TREE scores for the corresponding pathway combination. All TREE score error bars are derived from $n = 3$ independent experiments and show propagated error as described in Methods.

a neural network-based, machine-learning algorithm to predict beneficial pathway combinations, following the construction of an initial dataset.

In creating the initial dataset, we chose six homologs of each Thl, Hbd, Crt and Ter (Fig. 4a and Supplementary Table 1). We tested five concentrations for each enzyme homolog in a pathway context, consisting of our base set of enzymes (Fig. 4b), totaling 120 pathway combinations. We built these combinations in cell-free reactions, measured butanol production over time and calculated TREE scores for each (Fig. 4c). In total, we collected these additional data in five experimental sets of 20 pathway combinations with each set taking 5 d (3 d of HPLC time included). A majority of the enzyme homologs did not out-perform the original enzyme set (*Eco*Thl/*Cbe*Hbd/*Cac*Crt/*Tde*Ter), which has been extensively characterized[30–32]. However, we found that substituting *Ckl*Hbd1 can double the TREE score at high concentrations, in agreement with an independent study that found a 1.6-fold improvement in ABE fermentation with *C. acetobutylicum* by replacing native Hbd with *Ckl*Hbd1[33].

With this initial dataset collected, we identified ten neural network architectures based on a combination of heuristic search for model design and tenfold cross-validation (training and testing) for model scoring. We then used a gradient-free optimization strategy to maximize butanol production. We utilized the ten best architectures (most accurate predictions and highest model entropy) to make pathway combination predictions (homolog set and enzyme ratios), which we could then build with the cell-free framework (Fig. 4b). Design predictions suggesting enzyme concentrations of <0.01 µM (a majority <2 nM) were ruled out due to experimental constraints and we built the remaining 43 predictions in cell-free reactions (Supplementary Fig. 6a). We compared the results with two additional sets of experiments: (1) a set of varying enzyme ratios using only the base set enzymes (21 pathway combinations; Supplementary Fig. 6b) and (2) a hand-selected set of 18 pathway combinations based on our understanding of the biosynthetic pathway (Supplementary Fig. 6c). In total we tested 205 unique pathway combinations (base-set combinations, initial round combinations and data-driven designs) (Fig. 4d; all data shown in Supplementary Fig. 7).

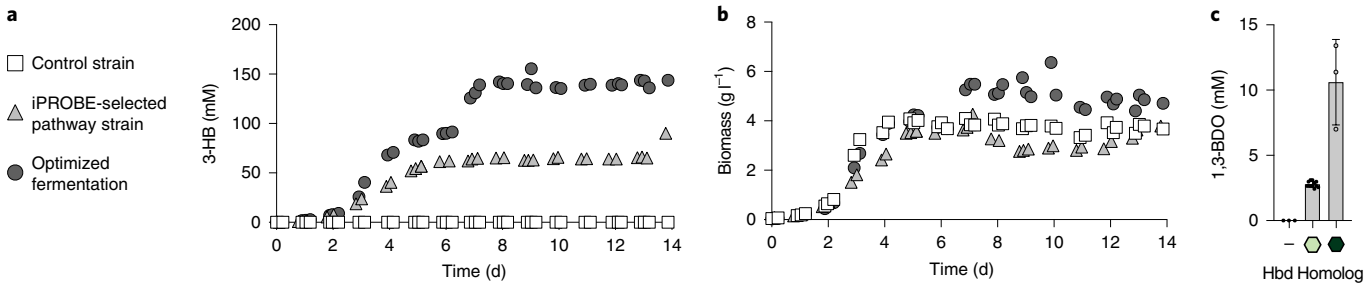

Fig. 5 | *Clostridium* fermentations show improved production of 3-HB and identification of a new route to 1,3-BDO. a,**b**, The iPROBE-selected optimal pathway, *Cac*Thl and *Ckl*Hbd1, for 3-HB production is built in a *C. autoethanogenum* strain and run in a 14-d continuous fermentation on $CO/H_2/CO_2$ gas as sole energy and carbon source. Comparing fermentation of the control strain of *C. autoethanogenum* (white squares) to the initial (light gray triangles) and optimized (dark gray circles) fermentations of the iPROBE-selected pathway expression strain, 3-HB (**a**) and biomass (**b**) were monitored. Each circle represents a single data point and is representative of two independent experiments. **c**, 1,3-BDO was measured during steady-state fermentation and all measurements are shown for the control strain (–; $n=3$), the iPROBE-selected pathway expression strain (light green hexagon; $n=12$) and strain containing the *Cne*PhaB homolog in place of Hbd (dark green hexagon; $n=3$). Error bars represent 1 s.d.

Nearly 20% of the total pathway combinations screened had higher TREE scores than our base case and we achieved over fourfold higher TREE scores (~2.5 times higher titer and 58% increase in rate) over the base-case pathway combination. The consensus enzyme set for top-performing pathways included *Eco*Thl, *Ckl*Hbd1, *Cac*Crt and *Tde*Ter with variations in enzyme concentrations. Five of the top six TREE scores each arose from pathways predicted from the neural network-based approach and were better than our hand-selected set, highlighting that deep-learning approaches can be used with a lack of a priori knowledge of the pathway.

Analysis of the iPROBE pathway combinations showed several key design parameters. First, we observed that there are specific enzyme homologs and concentrations that improve the TREE score. For example, the chosen Thl does not seem to matter in the 20 top-performing combinations, whereas the selection of Hbd does (Supplementary Fig. 8a); *Ckl*Hbd1 was superior to the rest. Second, iPROBE enabled identification of enzymes not to test in Clostridia. Specifically, *Cne*Thl, *Ckl*Hbd2, *Ppu*Crt, *Fsu*Ter, *Fjo*Ter and *Cpa*Ter all underperformed in the cell-free context. Being limited by throughput in non-model organisms, it is important to identify both promising and poor enzyme candidates. Third, we noticed that in the 20 top-performing combinations, Hbd is present at significantly ($P<0.001$) higher concentrations and the median Crt concentration is lower, though not significantly, than the initial 0.3 µM (Supplementary Fig. 8b). This suggests that higher concentrations of Hbd and Ter relative to Crt are optimal for effective pathway operation, which can be further investigated in vitro by measuring metabolite fluxes and achieved in vivo by constructing plasmids with proper genetic architectures.

We next assessed iPROBE's ability to inform cellular design by constructing representative pathway combinations from the iPROBE screening in *C. autoethanogenum* strains to produce butanol. While we tried to cover a wide range of TREE scores, challenges with transformation limited us to two pathway combinations scoring among the top five combinations (*Cac*Thl/*Ckl*Hbd1/*Cac*Crt/*Tde*Ter and *Eco*Thl/*Ckl*Hbd1/*Cac*Crt/*Tde*Ter), two pathway combinations in the middle range of the dataset and five pathway combinations near the tail end of all combinations tested (Supplementary Fig. 9a). To avoid diverting flux toward 3-HB, we identified and knocked out a native thioesterase that was able to hydrolyze 3-HB-CoA from our screening strain. After monitoring butanol production over the course of 6 d (Supplementary Fig. 9a), we observed a promising correlation between in vivo expression in *C. autoethanogenum* and TREE scores from iPROBE (Fig. 4e). This emphasizes that selecting top-performing pathways from iPROBE can improve production in *Clostridium* and decreases the number of strains that need to be tested. Of note, the iPROBE data suggested that balancing Hbd and Ter expression to keep crotonyl-CoA at minimal concentrations improves pathway performance. This hypothesis is corroborated by the in vivo results. We see lower butanol production when Hbd is expressed highly and Ter is expressed lowly but higher production when both are expressed highly.

While overall butanol production was low in *C. autoethanogenum*, we were able to increase production using the iPROBE-selected *Ckl*Hbd1 from 0 mM to 0.2 ± 0 mM. In addition, when comparing two butanol synthesis pathways in vivo (one with the standard *Cac*Hbd and one with the iPROBE-selected *Ckl*Hbd1) we increased butanol production sixfold from 0.2 ± 0 mM to 1.1 ± 0 mM (Supplementary Fig. 9b) by replacing the *trans*-2-enoyl-CoA reductase (Ter) enzyme with the ferredoxin-dependent electron bifurcating enzyme complex (Bcd–EtfA:EtfB) naturally used for these activities in Clostridia[34]. This is not surprising in light of a recent study that showed that Ter is detrimental to ABE fermentation when introduced in *C. acetobutylicum*[35]. Using the Bcd–EtfA:EtfB complex, we were also able to increase production to 22.0 ± 0.1 mM by manipulating the plasmid architecture (Supplementary Fig. 9c). For comparison, the best previously reported butanol production in engineered acetogenic Clostridia was ~2 mM[36]. Moreover, the Bcd–EtfA:EtfB complex is extremely oxygen-sensitive[37] and has so far been inactive in *E. coli* lysates[30], highlighting an area for potential improvement of iPROBE (that is, compatibility of *E. coli* lysates with non-model organisms). Taken together, we observed that iPROBE strongly correlated with cellular performance (Supplementary Fig. 10, $r=0.79$) for 20 pathways tested for both 3-HB production and butanol synthesis. Overall, this work demonstrates the power of coupling data-driven design of experiments with a cell-free prototyping framework to select feasible subsets of pathways worth testing in vivo for non-model organisms.

Scaled-up fermentations of iPROBE-selected pathway. Next, the best-performing iPROBE-selected strain for 3-HB production was chosen for process scale-up from 0.1-l bottle fermentations to 1.5-l continuous fermentations using $CO/H_2/CO_2$ gas as the sole carbon and energy source. Over a 2-week fermentation, we monitored 3-HB and biomass in a control strain and our iPROBE-selected strain with and without optimized fermentation conditions (Fig. 5a,b). In optimized fermentations, we observed high titers of 3-HB, ~15 $g\,l^{-1}$ (140 mM) at rates of >1.5 $g\,l^{-1}\,h^{-1}$ in a continuous system. This is not only higher than the previously reported concentration in acetogenic *Clostridium*[28,29], but to our knowledge also exceeds the previously highest reported concentration for traditional model organisms like *E. coli* (titer of ~12 $g\,l^{-1}$ and rate

of ~0.25 g l^{-1} h^{-1} in a fed batch system)[24,38] and yeast (titer of ~12 g l^{-1} and rate of ~0.05 g l^{-1} h^{-1} in a fed batch system)[39] without any additional genomic modifications to optimize flux into the pathway. As expected with an acetogenic host, we observed the production of acetate as a byproduct during fermentation (Supplementary Fig. 11). We anticipate that genome modifications to increase 3-HB flux could further improve fermentation titers, as was seen in a recent study reporting a 2.6-fold improvement in 3-HB production in a related *Clostridium* by downregulation of two native genes[28].

Surprisingly, we also observed production of a new metabolite, 1,3-butanediol (1,3-BDO), at 3–5% of the 3-HB levels and up to 0.5 g l^{-1} (Fig. 5c). This is attributed to nonspecific activity of a native aldehyde:ferredoxin oxidoreductase and alcohol dehydrogenase able to reduce 3-HB to 3-hydroxybutyraldehyde and further to 1,3-BDO. Indeed, no 1,3-BDO was observed when transforming the pathway into a previously generated aldehyde:ferredoxin oxidoreductase knockout strain[40]. These enzymes have been shown to reduce a range of carboxylic acids to their corresponding aldehydes and alcohols through reduced ferredoxin[36,41]. While the *(R)*-(−)-form of 1,3-BDO has been produced via other routes[42,43], when using the *Ckl*-derived Hbd we also detected the *(S)*-(+)-form of 1,3-BDO as determined by chiral analysis, which to our knowledge has never before been produced in a biological system. This chiral specificity is determined by the chosen 3-hydroxybutyryl-CoA dehydrogenase, either (*S*)-specific *Ckl*-derived Hbd or (*R*)-specific *Cne*-derived PhaB. Given that 1,3-BDO is used in cosmetics and can also be converted to 1,3-butadiene used in nylon and rubber production with a US$20 billion per year market[23,44], the discovery of this pathway is important. In summary, iPROBE provides a quick and powerful framework to optimize and discover biosynthetic pathways for cellular metabolic engineering efforts, including those in non-model hosts.

Discussion

We demonstrate a modular cell-free platform, called iPROBE, for constructing biosynthetic pathways with a quantitative metric for pathway performance selection (the TREE score). We establish that iPROBE can be used to engineer and improve small-molecule biosynthesis in non-model organisms that can be arduous to manipulate. In one example, iPROBE enabled the construction of a strain of *C. autoethanogenum* that produces high titers and yields of 3-HB in continuous fermentations (~20× higher than the previous highest report). The scale-up work also led to the identification of a new route to 1,3-BDO, for which we could produce the (*R*)- and (*S*)-isomer depending on enzyme selection. In another example, we used iPROBE with data-driven design of experiments to test 205 pathway combinations in vitro for the production of butanol and showed increased butanol production in acetogenic Clostridia by testing a further subset of designs in vivo. Notably, iPROBE demonstrates a strong correlation with in vivo pathway performance.

Despite the inherent contextual differences in *E. coli* lysates and Clostridia cells (for example, oxygen sensitivity), we have successfully demonstrated that iPROBE facilitates cellular design in three ways: (1) identifying sets of enzymes that work well together to produce a desired biological chemical; (2) downselecting poorly performing pathway/enzyme candidates; and (3) evaluating optimal ratios of enzymes and potential synergy between enzymes before embarking on laborious experiments in these organisms. iPROBE complements existing enzymatic assay approaches and provides the advantage of its combinatorial capability rather than an ability to select any single enzyme alone. This can accelerate DBT cycles (weeks with iPROBE instead of many months in *Clostridium*). While not all issues with engineering non-model organism expression are mitigated by iPROBE, it complements and enhances in vivo strategies.

Future developments of iPROBE could seek to improve the ability to design and optimize biosynthetic pathways in non-model organisms. For example, efforts to mimic physiochemical conditions of the organism of interest (for example, cofactors) and various conditions that mimic the phase of fermentation used during biochemical production (for example, batch versus semicontinuous and aerobic versus anaerobic) could be explored. In addition to enhancing the prototyping environment, refining the TREE score (for example, by weighting each factor) with more in vitro to in vivo correlation data will help to identify the minimal amount of cell-free data needed to accurately inform in vivo pathway performance. Indeed, we believe that the TREE score metric serves as a starting point and anticipate it evolving and improving in subsequent works. Finally, we note that while we focused on *Clostridium* as a cellular factory, many of our findings and tools could be applied to conventional hosts.

Looking forward, we anticipate that iPROBE will facilitate DBT cycles to decrease the number of strains that need to be engineered in vivo and the time required to achieve desired process objectives. This will increase the flexibility of biological processes to adapt to new markets, expand the range of fossil-derived products that can be displaced with bioderived alternatives and enhance the economic benefits for co-produced fuels.

Online content

Any Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-020-0559-0.

Received: 25 October 2019; Accepted: 6 May 2020;
Published online: 15 June 2020

References

1. Nielsen, J. et al. Engineering synergy in biotechnology. *Nat. Chem. Biol.* **10**, 319–322 (2014).
2. Nielsen, J. & Keasling, J. D. Engineering cellular metabolism. *Cell* **164**, 1185–1197 (2016).
3. Green, E. M. Fermentative production of butanol–the industrial perspective. *Curr. Opin. Biotechnol.* **22**, 337–343 (2011).
4. Jones, D. T. & Woods, D. R. Acetone-butanol fermentation revisited. *Microbiol. Rev.* **50**, 484–524 (1986).
5. Tracy, B. P., Jones, S. W., Fast, A. G., Indurthi, D. C. & Papoutsakis, E. T. Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. *Curr. Opin. Biotechnol.* **23**, 364–381 (2012).
6. Takors, R. et al. Using gas mixtures of CO, CO_2 and H_2 as microbial substrates: the do's and don'ts of successful technology transfer from laboratory to production scale. *Microb. Biotechnol.* **11**, 606–625 (2018).
7. Charubin, K., Bennett, R. K., Fast, A. G. & Papoutsakis, E. T. Engineering *Clostridium* organisms as microbial cell-factories: challenges & opportunities. *Metab. Eng.* **50**, 173–191 (2018).
8. Connelly Jr., T. M. et al. *Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals* (The National Academies Press, 2015).
9. Dudley, Q. M., Karim, A. S. & Jewett, M. C. Cell-free metabolic engineering: biomanufacturing beyond the cell. *Biotechnol. J.* **10**, 69–82 (2015).
10. Morgado, G., Gerngross, D., Roberts, T. M. & Panke, S. Synthetic biology for cell-free biosynthesis: fundamentals of designing novel in vitro multi-enzyme reaction networks. *Adv. Biochem Eng. Biotechnol.* **162**, 117–146 (2018).
11. Silverman, A. D., Karim, A. S. & Jewett, M. C. Cell-free gene expression: an expanded repertoire of applications. *Nat. Rev. Genet.* **21**, 151–170 (2020).
12. Bogorad, I. W., Lin, T. S. & Liao, J. C. Synthetic non-oxidative glycolysis enables complete carbon conservation. *Nature* **502**, 693–697 (2013).
13. Zhu, F. et al. In vitro reconstitution of mevalonate pathway and targeted engineering of farnesene overproduction in *Escherichia coli*. *Biotechnol. Bioeng.* **111**, 1396–1405 (2014).
14. Karim, A. S. & Jewett, M. C. A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab. Eng.* **36**, 116–126 (2016).
15. Dudley, Q. M., Anderson, K. C. & Jewett, M. C. Cell-free mixing of *Escherichia coli* crude extracts to prototype and rationally engineer high-titer mevalonate synthesis. *ACS Synth. Biol.* **5**, 1578–1588 (2016).

16. Hold, C., Billerbeck, S. & Panke, S. Forward design of a complex enzyme cascade reaction. *Nat. Commun.* **7**, 12971 (2016).
17. Kelwick, R. et al. Cell-free prototyping strategies for enhancing the sustainable production of polyhydroxyalkanoates bioplastics. *Synth. Biol.* **3**, ysy016 (2018).
18. Kay, J. E. & Jewett, M. C. Lysate of engineered *Escherichia coli* supports high-level conversion of glucose to 2,3-butanediol. *Metab. Eng.* **32**, 133–142 (2015).
19. Karim, A. S., Heggestad, J. T., Crowe, S. A. & Jewett, M. C. Controlling cell-free metabolism through physiochemical perturbations. *Metab. Eng.* **45**, 86–94 (2018).
20. 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).
21. Grubbe, W. S., Rasor, B. J., Krüger, A., Jewett, M. C. & Karim, A. S. Cell-free styrene biosynthesis at high titers. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.05.979302 (2020).
22. Karim, A. S. & Jewett, M. C. Cell-free synthetic biology for pathway prototyping. *Methods Enzymol.* **608**, 31–57 (2018).
23. Clomburg, J. M., Crumbley, A. M. & Gonzalez, R. Industrial biomanufacturing: the future of chemical production. *Science* **355**, 6320 (2017).
24. Tseng, H. C., Martin, C. H., Nielsen, D. R. & Prather, K. L. Metabolic engineering of *Escherichia coli* for enhanced production of (*R*)- and (*S*)-3-hydroxybutyrate. *Appl. Environ. Microbiol.* **75**, 3137–3145 (2009).
25. McMahon, M. D. & Prather, K. L. Functional screening and in vitro analysis reveal thioesterases with enhanced substrate specificity profiles that improve short-chain fatty acid production in *Escherichia coli*. *Appl. Environ. Microbiol.* **80**, 1042–1050 (2014).
26. Jewett, M. C. & Swartz, J. R. Mimicking the *Escherichia coli* cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol. Bioeng.* **86**, 19–26 (2004).
27. Jewett, M. C., Calhoun, K. A., Voloshin, A., Wuu, J. J. & Swartz, J. R. An integrated cell-free metabolic platform for protein production and synthetic biology. *Mol. Syst. Biol.* **4**, 220 (2008).
28. Woolston, B. M., Emerson, D. F., Currie, D. H. & Stephanopoulos, G. Rediverting carbon flux in *Clostridium ljungdahlii* using CRISPR interference (CRISPRi). *Metab. Eng.* **48**, 243–253 (2018).
29. Fluchter, S. et al. Anaerobic production of poly(3-hydroxybutyrate) and its precursor 3-hydroxybutyrate from synthesis gas by autotrophic Clostridia. *Biomacromolecules* **20**, 3271–3282 (2019).
30. Atsumi, S. et al. Metabolic engineering of *Escherichia coli* for 1-butanol production. *Metab. Eng.* **10**, 305–311 (2008).
31. Inui, M. et al. Expression of *Clostridium acetobutylicum* butanol synthetic genes in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **77**, 1305–1316 (2008).
32. Shen, C. R. et al. Driving forces enable high-titer anaerobic 1-butanol synthesis in *Escherichia coli*. *Appl. Environ. Microbiol.* **77**, 2905–2915 (2011).
33. Nguyen, N. P., Raynaud, C., Meynial-Salles, I. & Soucaille, P. Reviving the Weizmann process for commercial *n*-butanol production. *Nat. Commun.* **9**, 3682 (2018).
34. Li, F. et al. Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. *J. Bacteriol.* **190**, 843–850 (2008).
35. Qi, F. et al. Improvement of butanol production in *Clostridium acetobutylicum* through enhancement of NAD(P)H availability. *J. Ind. Microbiol. Biotechnol.* **45**, 993–1002 (2018).
36. Köpke, M. et al. *Clostridium ljungdahlii* represents a microbial production platform based on syngas. *Proc. Natl Acad. Sci. USA* **107**, 13087–13092 (2010).
37. Chowdhury, N. P., Kahnt, J. & Buckel, W. Reduction of ferredoxin or oxygen by flavin-based electron bifurcation in *Megasphaera elsdenii*. *FEBS J.* **282**, 3149–3160 (2015).
38. Gao, H. J., Wu, Q. & Chen, G. Q. Enhanced production of D-(−)-3-hydroxybutyric acid by recombinant *Escherichia coli*. *FEMS Microbiol. Lett.* **213**, 59–65 (2002).
39. Yun, E. J. et al. Production of (S)-3-hydroxybutyrate by metabolically engineered *Saccharomyces cerevisiae*. *J. Biotechnol.* **209**, 23–30 (2015).
40. Liew, F. et al. Metabolic engineering of *Clostridium autoethanogenum* for selective alcohol production. *Metab. Eng.* **40**, 104–114 (2017).
41. Perez, J. M., Richter, H., Loftus, S. E. & Angenent, L. T. Biocatalytic reduction of short-chain carboxylic acids into their corresponding alcohols with syngas fermentation. *Biotechnol. Bioeng.* **110**, 1066–1077 (2013).
42. Kataoka, N. et al. Enhancement of (R)-1,3-butanediol production by engineered *Escherichia coli* using a bioreactor system with strict regulation of overall oxygen transfer coefficient and pH. *Biosci. Biotechnol., Biochem.* **78**, 695–700 (2014).
43. Nemr, K. et al. Engineering a short, aldolase-based pathway for (R)-1,3-butanediol production in *Escherichia coli*. *Metab. Eng.* **48**, 13–24 (2018).
44. Jing, F. et al. Direct dehydration of 1,3-butanediol into butadiene over aluminosilicate catalysts. *Catal. Sci. Technol.* **6**, 5830–5840 (2016).

Methods

Bacterial strains and plasmids. *E. coli* BL21(DE3) (NEB) was used for preparation of cell extracts, which were used to express all exogenous proteins in vitro[22]. A derivate of *C. autoethanogenum* DSM10061 obtained from the German Collection of Microorganisms and Cell Cultures GmbH (DSMZ) was used for in vivo characterization and fermentations[45]. For butanol production, this strain was used with a native thioesterase (CAETHG_1524) knockout made using Triple Cross recombination as described previously[46].

Twenty-three enzymes were examined in this study (Supplementary Table 1). DNA for all enzyme homologs tested were codon-adapted for *E. coli* using IDT codon optimizer. Non-clostridial sequences were codon-adapted for *C. autoethanogenum* using a LanzaTech in-house codon optimizer, and all native clostridial genes were used as is. *E. coli*- and *C. autoethanogenum*-adapted sequences are listed in Supplementary Notes 1 and 2, respectively. For the cell-free work, the pJL1 plasmid (Addgene, 69496) was used. The modular pMTL80000 plasmid system[47] along with *acsA*[40], *fdx*[40], *pta*[48] and *pfor*[49] promoters were used for *C. autoethanogenum* plasmid expression.

Cell extract preparation. *E. coli* BL21(DE3) cells were grown, collected, lysed and prepared using previously described methods[14,50].

iPROBE reactions. CFPS reactions were performed to express each enzyme individually using a modified PANOx-SP system described in previous pubications[26,51]. Fifty-microliter CFPS reactions were carried out for each individual enzyme in 2-ml microcentrifuge tubes. Enzyme concentrations in CFPS reactions were quantified by ^{14}C-leucine incorporation during in vitro translation. Then reactions performed for identical enzymes were pooled together when multiple reaction-tube volumes were needed to keep a consistent 50-µl reaction volume and geometry for every CFPS reaction. Based on molar quantities of exogenous enzymes in each CFPS reaction determined by radioactive measurement, CFPS reactions were mixed to assemble complete biosynthetic pathways in 1.5-ml microcentrifuge tubes. CFPS reactions constitute 15 µl of a 30-µl-total second reaction. When the total CFPS reaction mixture constituted less than 15 µl, 'blank' CFPS reaction was added to make the total amount of CFPS reaction up to 15 µl. The 'blank' reactions consist of a typical CFPS reaction with no DNA added. The 15-µl CFPS mixture was then added to fresh extract (8 mg ml^{-1}), kanamycin (50 µg ml^{-1}), glucose (120 mM), magnesium glutamate (8 mM), ammonium glutamate (10 mM), potassium glutamate (134 mM), glucose (200 mM), Bis Tris (pH 7.8) (100 mM), NAD (3 mM) and CoA (3 mM); final reaction concentrations are listed. Reactions proceeded over 24 h at 30 °C. Measurements from samples were taken at 0, 3, 4, 5, 6 and 24 h.

Quantification of protein produced in vitro. CFPS reactions were performed with radioactive ^{14}C-leucine (10 µM) supplemented in addition to all 20 standard amino acids. We used trichloroacetic acid to precipitate radioactive protein samples. Radioactive counts from trichloroacetic acid-precipitated samples was measured by liquid scintillation to quantify soluble and total yields of each protein produced as previously reported (MicroBeta2; PerkinElmer)[26,27]. All enzyme expression data are listed in Supplementary Table 2.

Metabolite quantification. HPLC was used to analyze 3-HB and *n*-butanol. We used an Agilent 1260 series HPLC system via a refractive index detector. 3-HB and *n*-butanol were separated with 5 mM sulfuric acid as the mobile phase and one of two column conditions: (1) an Aminex HPX-87H or Fast Acids anion exchange column (Bio-Rad Laboratories) at 35 or 55 °C and a flow rate of 0.6 ml min^{-1} or (2) an Alltech IOA-2000 column (Hichrom) at 35 or 65 °C and flow rate of 0.7 ml min^{-1} as described earlier[52]. 1,3-BDO was measured using gas chromatography analysis, employing an Agilent 6890N gas chromatograph equipped with an Agilent CP-SIL 5CB-MS (50 m × 0.25 µm × 0.25 µm) column, autosampler and a flame ionization detector (FID) as described elsewhere[52]. For chiral analysis of *(S)*-(+)-1,3-BDO and (R)-(−)-1,3-BDO an Agilent 6890N gas chromatograph equipped with a Restek Rt-bDEXse 30 m × 0.25 mm ID × 0.25 µm df column and an FID was used. Samples were prepared by heating for 5 min at 80 °C, followed by 3-min centrifugation at 14,000 r.p.m. Exactly 400 µl of supernatant was then transferred to a 2-ml glass autosampler vial and 100 µl of an internal standard solution (5-methyl-1-hexanol and tetrahydrofuran in ethanol) was added. The capped vial was then briefly vortexed. Sample vials were transferred to an autosampler for analysis using a 1-µl injection, a split ratio of 60 to 1 and an inlet temperature of 230 °C. Chromatography was performed with an oven program of 50 °C with a 0.5-min hold to a ramp of 3 °C min^{-1} to 70 °C to a ramp of 2 °C min^{-1} to 100 °C with a final ramp at 15 °C min^{-1} to 220 °C with a final 2-min hold. The column flow rate was 30 cm s^{-1} using helium as the carrier gas. The FID was kept at 230 °C. Quantitation was performed using a linear internal standard calibration.

TREE score calculations. The TREE score was calculated by multiplying the titer by the rate by enzyme expression metric.

$$\text{TREE score} = \text{titer} \times \text{rate} \times \left(\text{average solubility} + [\text{total enzyme}]^{-1}\right)$$

The titer is the metabolite concentration (mM) in the cell-free reaction at 24 h, when the reaction is complete. The error associated with the titer is one s.d. of $n = 3$ independent experiments. The rate is the slope of the linear regression of metabolite concentrations (mM h^{-1}) taken at 3, 4, 5 and 6 h time points ($n = 4$). The rate-associated error is the standard error of the slope calculated by the linear regression. The average soluble fraction term is calculated by first determining the soluble fraction (soluble protein/total protein, $n = 3$ independent experiments) for each individual enzyme via ^{14}C-leucine incorporation. The average soluble fraction is then the average value of soluble enzyme fractions (mM soluble/mM total protein) (in this case, five enzymes) and the error associated with the soluble fraction term is propagated error. The concentration of total enzyme is calculated by the addition of the final concentrations of each enzyme (µM soluble protein). The final error on the TREE score is the propagated error of each individual component. Data used to generate TREE scores were not overlaid plots because this value and the propagated error do not represent a distribution of data. All 3-HB data including TREE scores are provided in Supplementary Dataset 1. All butanol data including TREE scores are provided in Supplementary Dataset 2.

In vivo gas fermentations. In vivo cultivation and small-scale bottle fermentation studies were carried out as described earlier using a synthetic gas blend, representative of waste gases from steel manufacturing, consisting of 50% CO, 10% H_2, 40% CO_2 (Airgas)[49]. Continuous fermentations were carried out in 1.5-l continuous stirred tank reactors with constant gas flow as described elsewhere[52,53].

Design of experiments using neural networks. A neural-network-based approach was used to explore the vast landscape of possible experimental designs. We first processed the cell-free butanol dataset and then developed and optimized the neural network to provide cell-free pathway recommendations. Modeling enzymatic pathways requires a mix of continuous and categorical variables. Because many machine-learning algorithms require numeric input and output variables, we used one-hot encoding, which is a process that converts categorical variables into a numerical format that machine-learning algorithms can use. This method treats categorical variables as multidimensional binary inputs that must sum to one. The concentration values were used as is, resulting in a 30-variable input matrix: 25 variables representing the categorical variation (that is, different homologs) and 5 representing the concentration. We used these features to build our deep neural network regressions.

We generated and evaluated neural network architectures based on a combination of heuristic search for model design and tenfold cross-validation for model scoring[54]. We limited our model architecture search to fully connected layers but varied the number of hidden layers (between 5 and 15 layers) and the number of nodes in each layer (between 5 and 15 nodes). We first randomly generated hundreds of model architectures based on these criteria. Using a genetic algorithm we performed crossovers and mutations on current model architectures, which were then trained using the back-projection method and scored using tenfold cross-validation[55]. Although no direct regularization methods were used the cross-validation step reduces the chance of over fitting. We proceeded using the genetic algorithm hundreds of times with thousands of iterations. Of the final 100 model architectures created, the top 10 models were chosen such that these models had the highest scores and highest design entropy. This ensures model diversity, which highlights data ambiguity (that is, model conclusions drawn from the same dataset).

We then optimized each of the ten models using the Nelder Mead Simplex, which provides a gradient-free optimization strategy to find the local minimum;[56] in our case maximum butanol production. This method generated 10 recommendations per model yielding a total of 100 recommendations. From this we selected the top ten recommendations that maximized both the average predicted butanol production (TREE score) and maximized the input entropy. This method ensured that we were not over-sampling in an area and set the basis for our hybrid exploration and exploitation-based sampling strategy. Each of these represents an exploitation-based recommendation, but through enforcing diversity in models and the recommendation vector we also were able to explore the sample space. From these optimized models, ten predictions were selected from each of the top ten architectures to be constructed in the cell-free environment. We removed predictions that were impossible experimentally (that is, concentrations too low to pipette accurate volumes).

Statistics and reproducibility. All statistical information provided in this manuscript is derived from $n = 3$ or greater independent experiments unless otherwise noted in figure legends. Error bars on metabolite and protein quantification in vitro and in vivo represent one s.d. derived from these experiments. All error bars on TREE score values are propagated errors as described in the TREE score calculations in Methods. These data do not represent a distribution of measured data but rather a calculation with propagated errors. In comparing the significance of enzyme concentration on TREE scores for butanol production in Supplementary Fig. 8b we used the Mann–Whitney *U*-test to determine whether enzyme concentrations of the enzyme combinations that produced the top 20 TREE scores were greater than the enzyme concentrations of the entire dataset.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All cell-free data generated and shown in this manuscript are provided in Supplementary Table 2 and Supplementary Datasets 1 and 2 (.xlsx). Any additional data or unique materials presented in the manuscript may be available from the authors upon reasonable request and through a materials transfer agreement.

References

45. Heijstra, B. D., Kern, E., Koepke, M., Segovia, S. & Liew, F. M. Novel bacteria and methods of use thereof. US patent 20130217096A1 (2013).
46. Liew, F. et al. Gas fermentation: a flexible platform for commercial scale production of low-carbon fuels and chemicals from waste and renewable feedstocks. *Front. Microbiol.* **7**, 694 (2016).
47. Heap, J. T., Pennington, O. J., Cartman, S. T. & Minton, N. P. A modular system for *Clostridium* shuttle plasmids. *J. Microbiol. Methods* **78**, 79–85 (2009).
48. Nagaraju, S., Davies, N. K., Walker, D. J., Kopke, M. & Simpson, S. D. Genome editing of *Clostridium autoethanogenum* using CRISPR/Cas9. *Biotechnol. Biofuels* **9**, 219 (2016).
49. Köpke, M. et al. 2,3-Butanediol production by acetogenic bacteria, an alternative route to chemical synthesis, using industrial waste gas. *Appl. Environ. Microbiol.* **77**, 5467–5475 (2011).
50. Kwon, Y. C. & Jewett, M. C. High-throughput preparation methods of crude extract for robust cell-free protein synthesis. *Sci. Rep.* **5**, 8663 (2015).
51. Jewett, M. C. & Swartz, J. R. Substrate replenishment extends protein synthesis with an in vitro translation system designed to mimic the cytoplasm. *Biotechnol. Bioeng.* **87**, 465–472 (2004).
52. Koepke, M., Jensen, R. O., Behrendorff, J. B. Y. H. & Hill, R. E. Genetically engineered bacterium comprising energy-generating fermentation pathway. US patent 9,738,875 (2017).
53. Valgepea, K. et al. Maintenance of ATP homeostasis triggers metabolic shifts in gas-fermenting acetogens. *Cell Syst.* **4**, 505–515 (2017).
54. Haykin, S. *Neural Networks: A Comprehensive Foundation* (Prentice Hall PTR, 1994).
55. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **4**, 65–85 (1994).
56. Nelder, J. A. & Mead, R. A simplex method for function minimization. *Computer J.* **7**, 308–313 (1965).

Acknowledgements

We thank A. M. Mueller, R. T. Tappel, W. Allen, L. Tran and S. D. Brown (LanzaTech) for conversations regarding this work. In addition, we thank C. Reynolds (Lockheed Martin) for conversations on the design of experiments using neural networks. This work is supported by the US Department of Energy, Office of Biological and Environmental Research in the Department of Environment Office of Science under award number DE-SC0018249. M.C.J. gratefully acknowledges the David and Lucile Packard Foundation and the Camille Dreyfus Teacher–Scholar Program. We also thank the following investors in LanzaTech's technology: BASF, CICC Growth Capital Fund I, CITIC Capital, Indian Oil Company, K1W1, Khosla Ventures, the Malaysian Life Sciences, Capital Fund, L. P., Mitsui, the New Zealand Superannuation Fund, Petronas Technology Ventures, Primetals, Qiming Venture Partners, Softbank China and Suncor.

Author contributions

A.S.K., S.D.S., M.K. and M.C.J. designed the study. A.S.K., Q.M.D. and M.C.J. developed the cell-free framework. A.S.K., S.A.C., J.T.H., W.S.G. and B.J.R. performed all cell-free experiments. A.S.K. and Q.M.D. analyzed cell-free data. A.J. performed *Clostridium* strain engineering for 3-HB and 1,3-BDO. T.A. performed *C. autoethanogenum* gas fermentation for 3-HB and 1,3-BDO. Y.Y., F.E.L., R.O.J., S.G. and M.K. performed *C. autoethanogenum* strain engineering and gas fermentation for butanol. A.J., Y.Y. and M.K.. analyzed *C. autoethanogenum* data. A.Q. developed analytical methods for 3-HB, 1,3-BDO and butanol. D.C., M.T., M.Kr. and J.S. performed all design of experiments using neural networks. A.S.K., M.K. and M.C.J. wrote the manuscript.

Competing interests

A.J., T.A., S.G., A.Q., Y.Y., F.E.L., R.O.J., S.D.S. and M.K. are employees of LanzaTech, which has commercial interest in gas fermentation with *C. autoethanogenum*. Production of 3-HB, 1,3-BDO and 1-butanol from C1 gases has been patented (US patents 9,738,875 and 9,359,611). A.S.K. and M.C.J. are co-inventors on the US provisional patent application 62/173,818 that incorporates discoveries described in this manuscript. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-020-0559-0.

Correspondence and requests for materials should be addressed to M.K. or M.C.J.

nature research

Corresponding author(s): Michael C. Jewett

Last updated by author(s): Apr 8, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Agilent ChemStation (Agilent Technologies, Inc.) was used to collect metabolite data. No custom software was used.
Data analysis	Microsoft Excel was used for data analysis. GraphPad Prism 8 was used for some of the graphs in this manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All cell-free data generated and shown in this manuscript are provided in Supplementary Dataset 1 and 2 (.xlsx). Any additional data presented in the manuscript may be available from the authors upon reasonable request and through a materials transfer agreement.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All sample sizes used are listed in the manuscript. Quantification of protein production from n>=3 CFPS reactions were used to enable calculation of mean and standard deviation as is regular practice in cell-free literature. Second pot reactions for biosynthesis of desired metabolites were run with n=1 at time point 3, 4, 5, and 6 h to get single measurements to determine rate of reaction and were run with n=3 at time point 24 h to enable calculation of mean and standard deviation. For in vivo experiments n>=3 for all cases except for Supplementary Figure 9B with Bcd (n=2 for this case only). Mean and standard deviation was calculated here as well. TREE scores were calculated as described in the literature with propagated error. In those cases, one time course was used to calculate TREE score based on the sample size for the time course listed above.
Data exclusions	No data was excluded.
Replication	All attempts at replication were successful.
Randomization	Samples were organized by experimental variables then characterized fully and reported in completion. Therefore, samples were not randomized.
Blinding	Blinding was not relevant. Animal or human participants were not used in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging