

Assessing the impact of substrate-level enzyme regulations limiting ethanol titer in *Clostridium thermocellum* using a core kinetic model



Charles Foster ^{a,d}, Veda Sheersh Boorla ^{a,d}, Satyakam Dash ^{a,d}, Saratram Gopalakrishnan ^{a,d}, Tyler B. Jacobson ^{b,d}, Daniel G. Olson ^{c,d}, Daniel Amador-Noguez ^{b,d}, Lee R. Lynd ^{c,d}, Costas D. Maranas ^{a,d,*}

^a Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA

^b Department of Bacteriology, University of Wisconsin—Madison, Madison, WI, USA

^c Thayer School of Engineering at Dartmouth College, Hanover, NH, USA

^d Center for Bioenergy Innovation, Oak Ridge, TN, USA

ARTICLE INFO

Keywords:

Clostridium thermocellum
Hungataclostridium thermocellum
Ruminiclostridium thermocellum
Acetovibrio cellulyticus
Kinetic modeling
Ensemble docking

ABSTRACT

Clostridium thermocellum is a promising candidate for consolidated bioprocessing because it can directly ferment cellulose to ethanol. Despite significant efforts, achieved yields and titers fall below industrially relevant targets. This implies that there still exist unknown enzymatic, regulatory, and/or possibly thermodynamic bottlenecks that can throttle back metabolic flow. By (i) elucidating internal metabolic fluxes in wild-type *C. thermocellum* grown on cellobiose via ¹³C-metabolic flux analysis (¹³C-MFA), (ii) parameterizing a core kinetic model, and (iii) subsequently deploying an ensemble-docking workflow for discovering substrate-level regulations, this paper aims to reveal some of these factors and expand our knowledgebase governing *C. thermocellum* metabolism. Generated ¹³C labeling data were used with ¹³C-MFA to generate a wild-type flux distribution for the metabolic network. Notably, flux elucidation through MFA alluded to serine generation via the mercaptopyruvate pathway. Using the elucidated flux distributions in conjunction with batch fermentation process yield data for various mutant strains, we constructed a kinetic model of *C. thermocellum* core metabolism (i.e. k-therm138). Subsequently, we used the parameterized kinetic model to explore the effect of removing substrate-level regulations on ethanol yield and titer. Upon exploring all possible simultaneous (up to four) regulation removals we identified combinations that lead to many-fold model predicted improvement in ethanol titer. In addition, by coupling a systematic method for identifying putative competitive inhibitory mechanisms using K-FIT kinetic parameterization with the ensemble-docking workflow, we flagged 67 putative substrate-level inhibition mechanisms across central carbon metabolism supported by both kinetic formalism and docking analysis.

1. Introduction

Biological production of biofuels from cellulosic plant feedstocks requires the development of microbial systems for both depolymerization of (hemi)cellulose and efficient fermentation of the released sugars towards reduced biofuel (Lynd et al., 1999). Consolidated bioprocessing (CBP) involves depolymerization and fermentation within the same processing unit by deploying an organism or community of organisms that can carry out both biological functions simultaneously (Lynd et al., 2005). By optimizing growth conditions (Hussain et al., 2017; Saddler and Chan, 1982; Summers et al., 1979) and through genetic engineering (Brown et al., 2011; Higashide et al., 2011; Kannuchamy et al., 2016; Lin

et al., 2015; Lo et al., 2017; Sangkharak et al., 2012; Tian et al., 2017; Tripathi et al., 2010; Xiong et al., 2018b; Ye et al., 2012) of organisms capable of solubilizing cellulose, significant progress has been made in CBP development (Davison et al., 2019; Lynd et al., 2016; Nagarajan et al., 2019; Nargotra et al., 2019; Olson et al., 2012; Wen et al., 2019).

Clostridium thermocellum (*C. therm*) is a gram-positive, cellulolytic thermophile with many attributes that make it an attractive candidate for CBP. In particular, it secretes a number of useful fermentation products naturally (i.e. isobutanol, ethanol, 2,3-butanediol, acetate, L-valine, etc.) (Holwerda et al., 2014; Tian et al., 2016) and has high cellulolytic activity compared to other cellulase-producing organisms such as fungi (Leis et al., 2017). Significant experimental effort has been

* Corresponding author. Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA.

E-mail address: costas@psu.edu (C.D. Maranas).

focused towards not only elucidating the novel features of *C. therm* metabolism, such as atypical glycolytic reactions (Zhou et al., 2013) and unique cofactor specificities (Thompson et al., 2015), but also developing *C. therm* into an industrial production platform for ethanol (Biswas et al., 2015; Deng et al., 2013; Lamed and Zeikus, 1980; Lamed et al., 1988; Lynd et al., 1989; Rydzak et al., 2011; Tian et al., 2017) and C4 alcohols (Holwerda et al., 2014; Lin et al., 2015; Tian et al., 2019). A limited number of experimental efforts have been aimed at understanding the substrate-level regulatory mechanisms at play in individual enzyme catalyzed reactions (Cui et al., 2019; Lamed and Zeikus, 1980; Placzek et al., 2017), but the known regulatory network is limited compared to that of model organisms such as *E. coli* (Placzek et al., 2017). For networks similar in size and pathway coverage, 55 *E. coli* substrate-level regulations were documented in literature and included in a model of core metabolism (Foster et al., 2019), whereas only 22 regulations were included in a *C. therm* core network using data for the entire *Clostridium* genus (Dash et al., 2017).

Kinetic models offer a means for targeting experimental effort to enhance our understanding of substrate-level mechanisms controlling flux. So far, only a limited number of studies have used top-down or bottom-up approaches to better understand *C. therm* metabolism. Several bottom-up models have been aimed at elucidating mechanisms for and describing rates of metabolic events related to substrate utilization (Olsen et al., 2016; Zhang et al., 2015; Zhang and Lynd, 2004). Specifically, Zhang and Lynd (2004) developed a kinetic model to capture the rate of phosphorolytic and hydrolytic β -glucosidic bond cleavage with different substrates, and both Zhang et al. (2015) and Olsen et al. (2016) identified product inhibition of cellulase by glucose with kinetic models. Dash et al. (2017) constructed a top-down kinetic model of *C. therm* core metabolism (k-ctherm118) containing 118 reactions and 53 metabolites from a set of fermentation product yield data for wild-type and a series of genetic mutants grown on cellobiose. The unavailability of complete fluxomic datasets at the time meant that the model was parameterized using only fermentation product yields (Dash et al., 2017), and the lack of internal flux information limited the fidelity of the kinetic model, as internal flux predictions could not be verified with experimentally resolved ones. k-ctherm118 was used to elucidate mechanisms controlling flux under nitrogen limitation (Dash et al., 2017), but an incomplete knowledge of *C. therm*'s substrate-level regulatory network limits non-intuitive prediction and applicability to designing of overproduction strains. More recent efforts have applied ^{13}C -labeling to elucidate the intracellular flux distribution in *C. therm* (Jacobson et al., 2020; Xiong et al., 2018a), and use of these data for parameterization can ensure that kinetic models capture the wild-type phenotype. However, an incomplete substrate-level regulatory network remained a major limitation towards achieving true predictive capability. For example, Dash et al. (2017) through ensemble modeling for robustness analysis (Lee et al., 2014; Theisen et al., 2016) identified ketol-acid reductoisomerase as a point in *C. therm* metabolism where (potentially) missing substrate-level regulation caused non-robust kinetic parameters and thus limited predictive capability.

Recently, data-intensive methods such as SIMMER (Hackett et al., 2016) have emerged for elucidating system-wide substrate-level regulations using kinetic models. SIMMER utilizes *in-vivo* metabolomics and proteomics data across a range of conditions to identify substrate-level regulations which improve the descriptive ability of kinetic rate expressions. Ensemble methods have been developed and applied with problem-specific datasets to elucidate substrate-level regulations in *E. coli* (Christodoulou et al., 2018) and plants (Faraji et al., 2015, 2018; Lee et al., 2011), but the unavailability of such detailed multi-omics datasets for *C. therm* limit the application of such methods. Structural analysis of enzyme-ligand binding by docking studies (Tripathi and Bankaitis, 2017) and use of machine learning models trained on docked enzyme-ligand conformations (Batra et al., 2020; Chandak et al., 2020; Pu et al., 2019) can offer alternative routes for mechanistically investigating enzymatic regulatory interactions. To date, such methods have

been geared primarily towards drug discovery (Amaro et al., 2018; Falcon et al., 2018; Konc et al., 2015), and have not been applied for elucidating metabolic regulations. For example, there are a few case studies investigating drug target capabilities using docking studies of handpicked metabolic enzymes (Gautam et al., 2012). However, high-throughput application of such methods to study enzymatic regulatory events are not known. This work aims to incorporate wild-type ^{13}C -labeling data, kinetic parameterization, and structural analysis to fill in knowledge gaps in *C. therm* metabolism and active substrate-level regulatory network from minimal data and evaluate how substrate-level regulation controls ethanol yield and titer.

As part of this effort, we first generated ^{13}C labeling data for *C. therm* grown on cellobiose (Olson et al., 2017) and then elucidated core metabolic fluxes. We hypothesize that our knowledge of the active *C. therm* metabolic network may be incomplete. ^{13}C -MFA flux elucidation was used not only in kinetic model parameterization, but also to probe for possibly active metabolic pathways that are not currently annotated. Through the deployment of the kinetic model and ensemble docking frameworks, we explore two questions. First, are there inhibitory mechanisms active in *C. therm* metabolism, but currently undocumented, that upon inclusion substantially improve kinetic model parameterization? Second, out of all known *C. therm* substrate-level regulations, which ones control ethanol production?

2. Methods

2.1. Materials

Tracers ($[\text{U}-^{13}\text{C}]$ cellobiose (99 atom% ^{13}C), $[^{13}\text{C}]$ bicarbonate (99 atom% ^{13}C)) were purchased from Cambridge Isotope Laboratories (Andover, MD). All other chemicals were purchased from Sigma-Aldrich (St. Louis, MO).

2.2. Strains and growth conditions

Wild-type *C. thermocellum* (DSM1313) growth and fermentation were carried out in MTC (11.56 g L^{-1} MOPS sodium salt, 3 g L^{-1} trisodium citrate, 1.5 g L^{-1} KH_2PO_4 , 1.5 g L^{-1} NH_4SO_4 , 2.6 g L^{-1} $\text{MgCl}_2 \bullet 6\text{H}_2\text{O}$, 0.13 g L^{-1} $\text{CaCl}_2 \bullet 2\text{H}_2\text{O}$, 0.001 g L^{-1} $\text{FeCl}_2 \bullet 4\text{H}_2\text{O}$, 0.5 g L^{-1} L-cysteine $\text{HCl} \bullet \text{H}_2\text{O}$, 0.004 g L^{-1} p-aminobenzoic acid, 0.002 g L^{-1} biotin, 0.002 g L^{-1} cobalamin, 0.004 thiamin, and 5g L^{-1} cellobiose, pH = 7).

Anaerobic *C. thermocellum* cultures were grown in pressure sealed bottles with butyl rubber stoppers. Cultures were prepared by first filling bottles with a media base containing MOPS solution. Then, oxygen was scrubbed using a vacuum manifold with N_2 gas overlaid, and finally, cell cultures were autoclaved. Other media components were autoclaved separately and added to the cultures using anoxic syringes (syringes depleted of oxygen using 2.5% cysteine HCl solution). After inoculation, culture bottles were placed in a 55 °C water bath for thermophilic growth. A Genesis 20 UV-Vis spectrophotometer equipped with a hunge tube adapter was used to quantify growth via optical density measurement at 600 nm (OD_{600}). Inoculation and sampling of cultures was performed using anoxic syringes. Any culture manipulation requiring a bottle to be opened was carried out in an anaerobic chamber with 5% CO_2 , 5% H_2 , and 90% N_2 atmosphere (<100 ppm O_2).

2.3. Liquid chromatography-mass spectrometry of extracted metabolites

Solvent containing extracted metabolites was dried under N_2 gas, then metabolites were resuspended in Solvent A (97:3 H_2O :methanol containing 10 mM tributylamine adjusted to pH = 8.2 by the addition of acetic acid to ~10 mM). Chromatographic separation occurred on a Acquity UPLC BEH C₁₈ column (2.1x100mm, 1.7 μm particle size; Waters) operating at 25 °C. Solvent A was as described above and solvent B was 100% methanol. The following gradient was used for separation: 0–2.5min, 5% B; 2.5–17min, linear gradient from 5% to 95% B;

17–19.5min, 95% B; 19.5–20min, linear gradient from 95% to 5% B; 20–25min, 5% B. Samples were analyzed on a Vanquish UPLC coupled to a Q Exactive Orbitrap high-resolution mass spectrometer (Thermo-Scientific) by an electrospray ionization source operating in negative mode. Mass spectrometry parameters were Full MS-SIM (single ion monitoring) scanning between 70 and 1000 m/z with a resolution of 70,000 full width at half maximum (FWHM), maximum injection time (IT) of 40 ms, and automatic control gain (ACG) target $1e^6$. LC-MS data were analyzed using the MAVEN software suite (Clasquin et al., 2012; Melamud et al., 2010) and compounds were identified by retention time matching to pure standards and monoisotopic mass.

2.4. Metabolite collection

Anaerobic inoculation was performed on working cultures with a 1:20 or greater dilution from overnight anaerobic cultures grown to an initial OD₆₀₀ of ~0.05 on the same carbon source. 5 mL samples of intracellular metabolites were collected inside the anaerobic chamber via vacuum filtration once the cells reached mid-exponential phase (OD₆₀₀ 0.4–0.5) through 0.45 μ m hydrophilic nylon filters to ensure separation of cells from the media. Metabolism was quenched and metabolites collected by placing used filters cell-side down in 1.5 mL extraction solvent (40% acetonitrile, 40% methanol, 20% water) on dry ice. Cells were washed from filters using extraction solvent. The solvent was then collected, vortexed, and centrifuged for 5 min at 4 °C to remove cellular debris. Supernatant was collected for LC-MS analysis. Labeling experiments with uniformly labeled ¹³C tracers was performed in biological duplicate.

2.5. Experimental mutant training datasets

Yield data for ten genetic mutation strains used for training the kinetic model was procured from Dash et al. (2017). Table 1 lists the mutant strains used in the training data for kinetic parameterization as well as their corresponding gene and enzyme level perturbations. Cell

Table 1

Enzyme level perturbations corresponding to genetic perturbations in genetically engineered strains used for training data. The delta symbol (Δ) represents a complete reaction knockout. Enzyme abbreviations are as follows: LDH_L: L-lactate dehydrogenase, PTAr: Phosphotransacetylase, ALCD2x: Alcohol dehydrogenase, ACALD: acetaldehyde dehydrogenase, POR2_i: Pyruvate ferredoxin oxidoreductase, NADOX: Bidirectional hydrogenase, PFL: Pyruvate-formate lyase.

Strains	Genes Perturbed	Enzymes Perturbed	Fold-change in Enzyme Level
LL1036, LL372	Ldh	LDH_L	Δ
LL1041, LL373	Pta	PTAr	Δ
LL1011, LL1042, LL1044, LL374, LL375, LL1043	Ldh	LDH_L	Δ
LL1066	adhE	ALCD2x	0.5
		ACALD	0.5
LL1038, LL1067	adhE	ALCD2x	0.2
		ACALD	0.2
	Ldh	LDH_L	Δ
LL1224	Pfor	POR2_i	Δ
LL1148, LL1149	hydG	NADOX	0.5
	Ldh	LDH_L	Δ
	Pfl	PFL	Δ
	Pta	PTAr	Δ
	Ack	ACKr	Δ
LL350	hydG	NADOX	0.1
LL1147	hydG	NADOX	Δ
	ech		
LL1111	adhE	ALCD2x	0.5
		ACALD	0.5
	Ldh	LDH_L	5

cultures used for generating yield data were also grown under batch anaerobic conditions at 55 °C using AN7 medium with initial cellobiose loading of 2.5 g L⁻¹ (Olson et al., 2017). Experimental data used for training purposes is contained in Supplementary File S2 and is also available at our group's GitHub repository (<https://github.com/maranasgroup/k-ctherm138>).

2.6. ¹³C metabolic flux analysis

We performed flux elucidation and 95% confidence interval estimation for wild-type *C. therm* using the aforementioned set of isotopic labeling data. Metabolic yields for ethanol, acetate, formate, lactate, alanine, malate, aspartate, glutamine, glutamate, proline, pyruvate, valine, leucine, isoleucine, and methionine were measured and used to elucidate the flux distribution. The *C. therm* central carbon metabolism stoichiometric model and atom mapping model constructed by Xiong et al. (2018a) with reactions added for cellobiose consumption and mercaptopyruvate metabolism was adopted as a basis for constructing the elementary metabolite unit (EMU) network used in this study. INCA ¹³C-MFA software implemented in MATLAB was used to generate the EMU network from the atom mapping model, elucidate fluxes and 95% confidence intervals, and perform degrees of freedom analysis to confirm the statistical validity of the inferred flux distribution (Young, 2014). Parameter continuation was used for confidence interval calculation for all net and exchange fluxes in the metabolic network (Antoniewicz et al., 2006; Young, 2014). Before confidence interval estimation, 100 randomly initialized multi-starts were performed to ensure convergence to the minimum SSR and solution reproducibility. Best SSR recovery was 59% on average among all simulations. Confidence intervals were estimated once for the optimal flux distribution.

2.7. K-FIT kinetic parameterization with heterogeneous datasets

Kinetic parameterization was performed using the gradient-based K-FIT algorithm (Gopalakrishnan et al., 2020). A single wild-type flux distribution inferred via ¹³C-MFA from the aforementioned ¹³C-labeling data and mutant strain fermentation yield data were used to train the model. The wild-type flux distribution was used within the K-SOLVE module of K-FIT to infer a set of elementary kinetic parameters at each iteration of the algorithm. As in our previous kinetic model of *C. therm* metabolism (Dash et al., 2017), mutant strain fermentation yields were assumed to correspond to a steady-state condition, and yields were calculated using 100 mmol gDW⁻¹ h⁻¹ cellobiose uptake as a basis. Weighted square residual error function and gradient calculations within K-FIT were updated to reflect these details. K-FIT uses an elementary step decomposition approach to describe each metabolic reaction in the modeled network. Reverse elementary reaction flux and enzyme complex fractional abundance are optimization variables in the algorithm and are used along with the wild-type steady-state flux distribution to calculate a set of elementary kinetic parameters that exactly describes the wild-type flux distribution. The K-FIT algorithm accelerates steady-state metabolite concentration and reaction flux estimation by decomposing the system of ordinary differential equations describing the change in each metabolite species and enzyme complex species with respect to time into two sets of linear algebraic equations, and utilizes established algebraic solution techniques to evaluate each system of equations iteratively until a steady-state is reached. K-FIT minimizes the difference between experimental fluxes and those predicted by the kinetic model under the mutant conditions used in the training datasets. The gradient-directed parameterization method offers significant computational speed-up over metaheuristic methods because it bypasses ODE integration and reduces the number of model evaluations required for convergence. For an in-depth description of the K-FIT algorithm refer to Gopalakrishnan et al. (2019).

As in Foster et al. (2019), the metabolic network used for kinetic parameter generation and ¹³C-MFA in the wild-type strain were made

identical with the exception of reduced cofactor pools, which were simplified to a single pool to avoid linear dependencies between ferredoxin, NADPH, and NADH pools in the stoichiometric matrix. A total of 1,200 multi-starts were initialized randomly, with the model yielding the best SSR value selected as the kinetic model and used in subsequent analysis. The best parameterization was recovered once, with 14 additional models yielding SSR values within 10% of the best parameterization. The next closest model yielded an SSR value within 2% of the best SSR recovered. The set of models yielding SSR within 10% of the best model were used to evaluate standard deviation and mean parameter values for all elementary kinetic parameters in k-ctherm138.

2.8. Discovering *C. thermocellum* substrate-level regulations

We first identified a set of core regulations on metabolic reactions known to be present in the *Clostridium* genus. These included 16 substrate-level inhibitions and three allosteric activations listed in Supplementary File S2 and used in the parameterization of k-ctherm138. Afterwards, we compiled a comprehensive list of 135 bacterial substrate-level inhibitory mechanisms documented in the BRENDA (Placzek et al., 2017) and EcoCyc (Keseler et al., 2017) databases. Starting from a set of elementary kinetic parameters corresponding to a confirmed local minimum (i.e., k-ctherm138), we re-initialized the K-FIT algorithm with one non-core regulation imposed on the model as a competitive inhibition at a time. A regulatory event from this non-core set was only assessed as potentially active in *C. therm* if it lowered the overall SSR while maintaining a required non-zero parameter value. Out of 135 putative competitive inhibitions, K-FIT identified 105 that satisfied these criteria (see Supplementary File S2). These regulations were subjected to ensemble-docking simulations as detailed below.

Enzyme-ligand binding interactions were assessed for their binding strength using docking simulations on ensembles of configurations extracted from molecular dynamics (MD) simulation trajectories. Fig. 1 illustrates the ensemble docking workflow discussed herein. First, we constructed homology models for all 33 *C. therm* enzymes with a K-FIT-

identified hypothetical regulation using trRosetta (Yang et al., 2020). A complete list of the enzymes can be found in Supplementary File S3. Each homology model was prepared for MD simulations using the protein preparation wizard (Sastry et al., 2013) of Maestro wherein the hydrogen bonding network was optimized and a heavy atom-restrained energy minimization with the OPLS_2005 force field (Harder et al., 2016) was carried out. The energy minimized model was placed in a box of explicit solvent water molecules modeled by the SPC water model using the Desmond molecular dynamics package (v2.3) (Bowers et al., 2006) within the Schrodinger software suite (v2019.4). Each MD simulation consisted of the default relaxation protocol of Desmond followed by a 10 ns of production simulation at 1 atm pressure and 310 K temperature, using the NPT ensemble under periodic boundary conditions using particle mesh Ewald. The residual charges from the system were neutralized by adding Na⁺ and Cl⁻ ions to maintain 0.15 M salt concentration. A time step of 2.0 fs and nonbonded cut-off threshold of 9 Å were imposed and the SHAKE (Ryckaert et al., 1977) algorithm was used to keep all bonds involving hydrogen atoms rigid. The MD simulations were run at 55 °C and pH of 7 (consistent with optimal growth conditions for *C. therm* (Akinoshio et al., 2014)). We then extracted snapshots of atomic coordinates from the 10 ns trajectory at 0.1ns interval forming an ensemble of 100 configurations for each enzyme. Using Rosetta force-field (Alford et al., 2017), the 100 configurations were scored to select 20 configurations with lowest potential energy as the ensemble for each enzyme. Since the substrate binding sites of *C. therm* enzymes are not experimentally characterized, we used state-of-the-art binding site prediction software, COACH-D (Wu et al., 2018), which incorporates both structural and sequence conservation methods to arrive at consensus predictions. The homology models of the enzymes built as described above were input to the online COACH-D server to predict binding sites. The binding site information obtained for each enzyme was used to infer the initial coordinates and size of the grid for the search algorithm of docking program SMINA (Koes et al., 2013). Specifically, the centroid of the residues predicted to be a part of the binding site was used as the center of the grid and the sizes of the

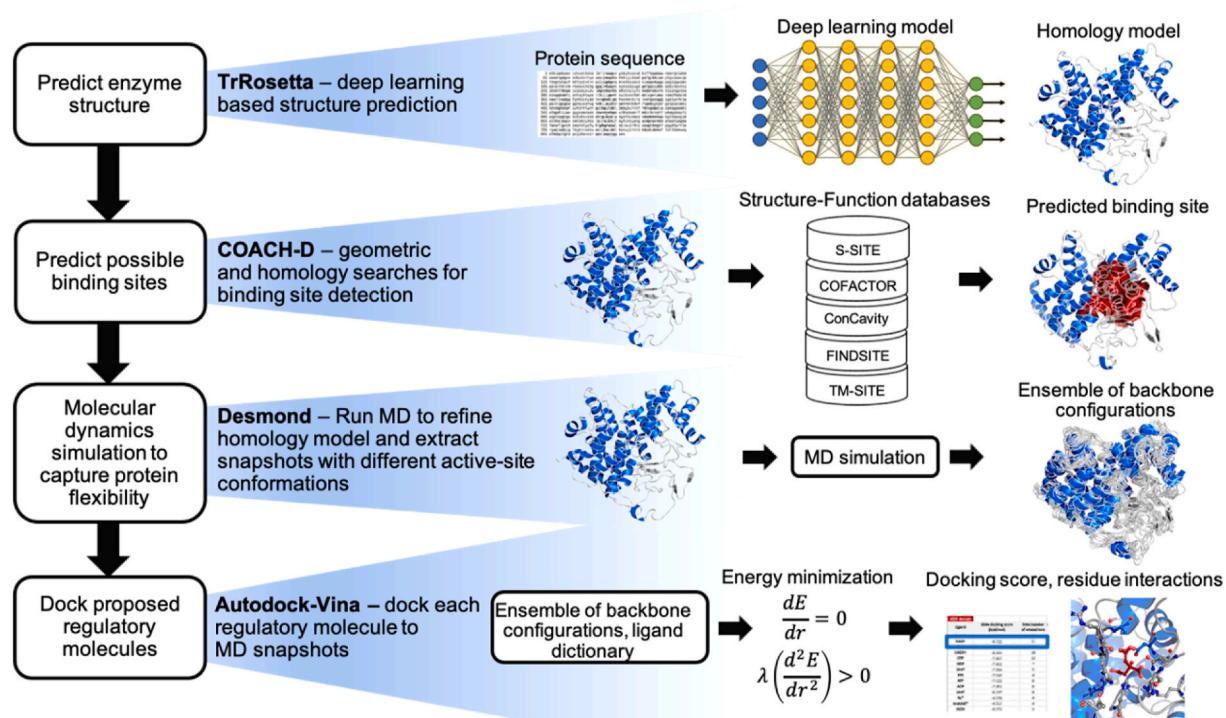


Fig. 1. Ensemble docking workflow for inferring the binding energy of hypothetical substrate-level regulations.

inhibitor to be docked (plus a buffer of 5 Å in the x, y, and z directions) were used as the dimensions of the search grid box. Each of the 105 inhibitions was hence assessed by docking the corresponding inhibitors to each configuration in the extracted ensemble of the corresponding enzymes. For each enzyme-inhibitor pair, the minimum binding energy identified by the docking calculation across the ensemble of 20 configurations was considered as the final predicted binding energy. The Autodock vina (Trott and Olson, 2010) scoring function within SMINA was used for the predictions.

2.9. Identification of substrate-level regulations limiting ethanol titer and yield

The full combinatorial space of between one and four substrate-level regulations in total was explored to identify the optimal strategy for improving ethanol titer/yield by eliminating regulations. To assess the impact of each individual substrate-level enzyme regulations on ethanol titer/yield, we eliminated the regulation (or combination of regulations) from the model and evaluated the fold-change in steady-state ethanol concentration and yield from the wild-type strain. A total of 5,035 combinations (corresponding to the entire combinatorial space) were tested. The corresponding steady-state ethanol yield and pool size are

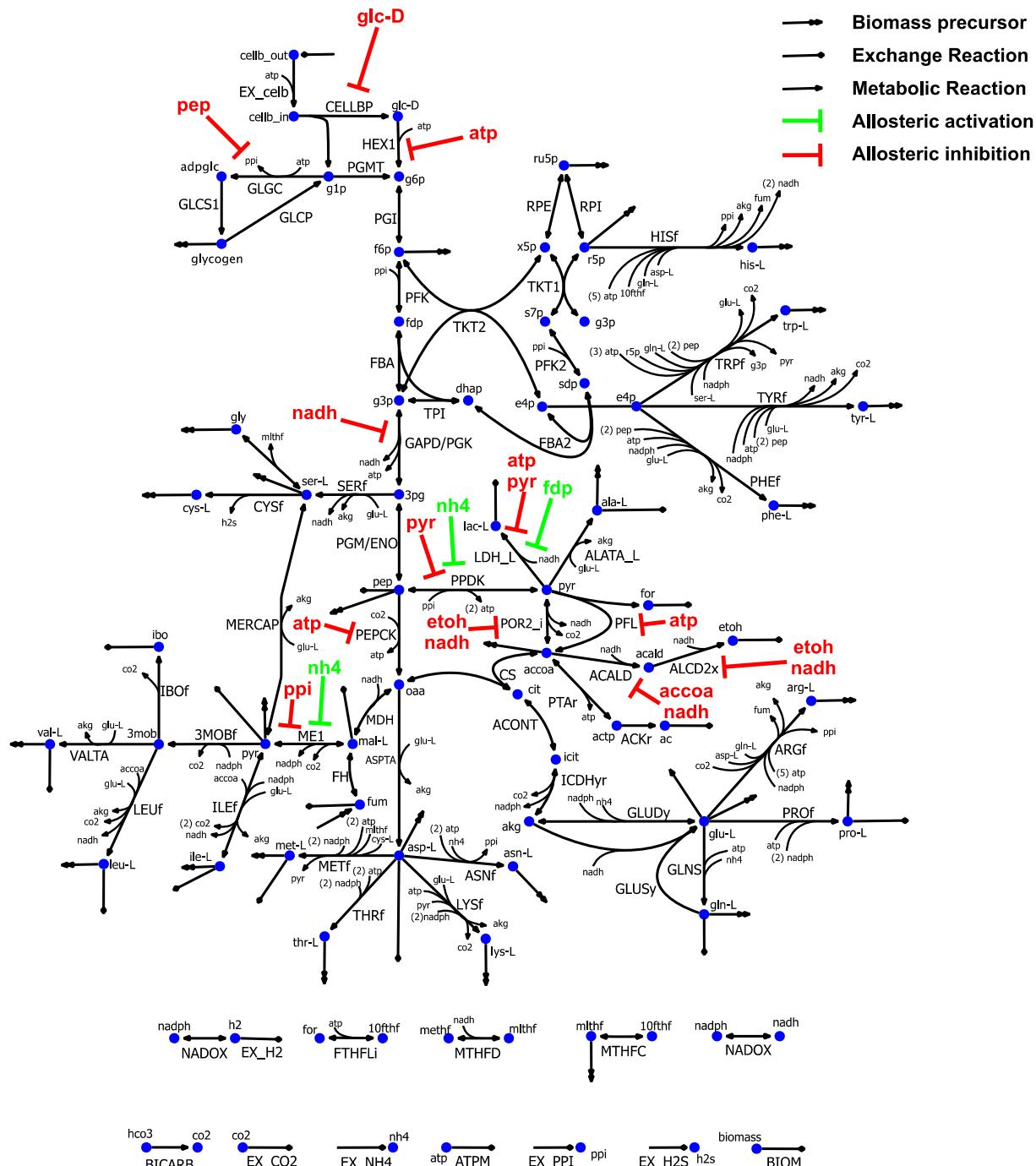


Fig. 2. *C. therm* core metabolic and substrate-level regulatory network used in the construction of k-ctherm138. See Supplementary file S2 for a list of reaction and metabolite abbreviations.

reported in Supplementary File S2.

2.10. Model and code availability

All k-ctherm138 model files and codes required to reproduce the results of this study can be found at our group's GitHub page ([http://github.com/maranasgroup/k-ctherm138](https://github.com/maranasgroup/k-ctherm138)). Codes for the ensemble docking workflow deployed in this study can be found at our group's GitHub page (<https://github.com/maranasgroup/ensembl docking>).

3. Results

3.1. Model coverage

The core metabolic network of *C. therm* central carbon metabolism used in this study (shown in Fig. 2) contains 138 reactions and 81 metabolites from glycolysis, the nonoxidative pentose phosphate (PP) pathway, tricarboxylic acid (TCA) cycle, glycogen storage and recycling pathways, pathways for acetate, lactate, and ethanol secretion, 21 amino acid synthesis pathways, nutrient and cellobiose uptake, and overflow metabolism. The model is similar to the model developed by Dash et al. (2017), with modifications to account for the re-derivation of the biomass drain from the iCth446 genome-scale metabolic model (Dash et al., 2017) to describe the drain of individual amino acids, central carbon metabolites, and cofactors required for macromolecular biosynthesis. Cofactors in the model include NADH, ATP, pyrophosphate, and folate. A single pool for each cofactor was used to eliminate metabolite pool dependencies from the stoichiometric matrix (Gopalakrishnan et al., 2020). The PP pathway is consistent with the most recent *C. therm* GSM model, iCBI655 (Garcia et al., 2020b), and includes sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase (FBA2) and phosphofructokinase (s7p) (PFK2), as well as transketolase 1 (TKT1) and transketolase 2 (TKT2) reactions.

The carbon atom mapping model constructed by Xiong et al. (2018a) with network coverage identical to the kinetic model network used in this study was adopted to describe the carbon transitions in each metabolic reaction in *C. therm* central carbon metabolism. Cofactor and nutrient dependencies identical to those included in the kinetic model were included in the network used for ¹³C-MFA, but NADH and NADPH were separated into distinct pools, with a reversible interconversion representing the ferredoxin NADP reductase reaction. The model was updated to include atom transitions for cellobiose consumption. Atom mapping for the ATP-dependent cellobiose phosphorylase reaction (converting cellobiose to D-glucose and D-glucose 1-phosphate) was gleaned from the KEGG database (Kanehisa and Goto, 2000). A single reaction describing the conversion of pyruvate to serine was also included in the model using glutamate as an amino group donor (representative of the mercaptopyruvate pathway). Carbon mapping for this reaction was inferred from the net carbon transitions of 3-mercaptopuruvate sulfurtransferase, cysteine transaminase, and cystathione β -synthase catalogued in the KEGG database (Kanehisa and Goto, 2000).

The known regulatory network for k-ctherm138 consists of 19 substrate-level inhibitions (16) and activations (3) acting on *C. therm* core metabolic reactions specific to the *Clostridium* genus identified in either the BRENDa database (Placzek et al., 2017) or literature (Cui et al., 2019; Lamed and Zeikus, 1980). Of those regulations, 12 were *C. therm* specific (9 inhibitions, 3 activations), while 7 were not. The k-ctherm138 regulatory network is depicted in Fig. 2, with a description of each regulation included in Supplementary file S2.

3.2. Mercaptopyruvate pathway activity

¹³C-MFA based on *C. therm* central carbon metabolism described by the network developed by Xiong et al. (2018a) with updated biomass synthesis stoichiometry led to a SSR above the threshold for statistical acceptability according to the degrees of freedom (DOF) analysis

performed within INCA (Young, 2014). Simulated mass isotopomer fractional abundances that deviated the most from experimental values included the 3-phosphoglycerate M+2 and pyruvate M+3 mass isotopomers. We therefore explored whether the activity of alternative pathway(s) leading to the re-distribution of labeled and unlabeled carbon incorporated into 3-phosphoglycerate and pyruvate could explain the measured isotopomer fractional abundances. We found that the inclusion of the mercaptopyruvate pathway acting as an alternative production route for L-serine biosynthesis (see Fig. 3) lead to a significant decrease in the SSR from 206 to 102. The reason for the improved capturing of isotopomer abundances is that higher lower glycolytic flux increases fully labeled pyruvate, while lowered gluconeogenic exchange flux decreases the abundance of the M+2 3 pg fragment.

Table 2 compares the overall fitness of the ¹³C-MFA when fluxes were elucidated with and without the mercaptopyruvate pathway in the model, and results indicate that its inclusion is essential to model fitness. An F-test ruled out the null hypothesis for its inclusion with 99.9% confidence implying that the reduction in SSR is highly unlikely to be caused only by the reduction in degrees of freedom associated with adding a pathway. When the mercaptopyruvate pathway was excluded from the metabolic network, lower glycolytic net flux from 3-phosphoglycerate to pyruvate decreased by 4.0 mmol/100 mmol cellobiose uptake in the optimal flux distribution to deliver the required carbon from 3-phosphoglycerate to L-serine, and gluconeogenic (exchange) flux in lower glycolytic reactions delivering unlabeled CO₂ to 3-phosphoglycerate increased by 99.9%. Table 3 compares the experimentally measured 3-phosphoglycerate and pyruvate MDVs with simulated ones with and without the mercaptopyruvate pathway. These seemingly minor readjustments in the 3-phosphoglycerate and pyruvate MDVs caused a 98.5% decrease in their SSR. Admittedly, pyruvate measurements could be error prone without proper derivatization. Note that derivatization was not performed when we measured pyruvate mass isotopomer abundance. However, because the acetyl group in acetyl-CoA is formed from the second and third carbons (C2 and C3) of pyruvate and the first carbon (C1) is lost to decarboxylation, any pair of compounds differing by only an acetyl group can be used to estimate the C2 and C3 isotopomer abundance of pyruvate. Upon comparing glucosamine phosphate and N-acetylglucosaminephosphate mass distribution vectors, we find that the acetyl group contains 41% M+2 abundance (42% when corrected for natural ¹³C abundance). In agreement with N-acetylglucosaminephosphate labeling, pyruvate in our labeling data contains 42% M+2 or M+3 mass isotopomer abundance. We are, therefore, satisfied with the degree of accuracy of our pyruvate measurements without derivatization and that the labeling patterns reflect *C. therm* physiology. A comparison of serine labeling patterns also shed light on the source(s) of serine. Our experiments with ¹³C-bicarbonate/unlabeled cellobiose indicate pyruvate had the highest M+1 abundance (57% M+1 due to exchange flux through pyruvate ferredoxin oxidoreductase), serine had the second highest M+1 abundance (23%), while 3 pg had only 11% M+1. This suggests serine is likely derived from multiple central carbon intermediates, and potentially through the glycine cleavage system, as reverse flux through this pathway would lead to a single ¹³C incorporated at the C1 position of serine. Thus, while the data obtained and subsequent analysis give a good indication that serine is likely derived from multiple central carbon intermediates in *C. therm*, further investigations are needed to ascertain the robustness of findings to changed conditions.

The added pathway contains three enzyme-catalyzed steps as shown in Fig. 3. In the first step (catalyzed by a 3-mercaptopuruvate sulfurtransferase enzyme), pyruvate is converted to mercaptopyruvate. This activity has not been previously annotated for *C. therm*. However, a BLAST search of *C. therm* genome and all known genes encoding 3-mercaptopuruvate sulfurtransferase enzymes identified gene CLO1313_RS12845 (O'Leary et al., 2016) encoding a bifunctional O-acetylhomoserine/O-acetylserine sulfhydrylase as a 50% homologous with 42% sequence coverage of the *Ochrobactrum intermedium*

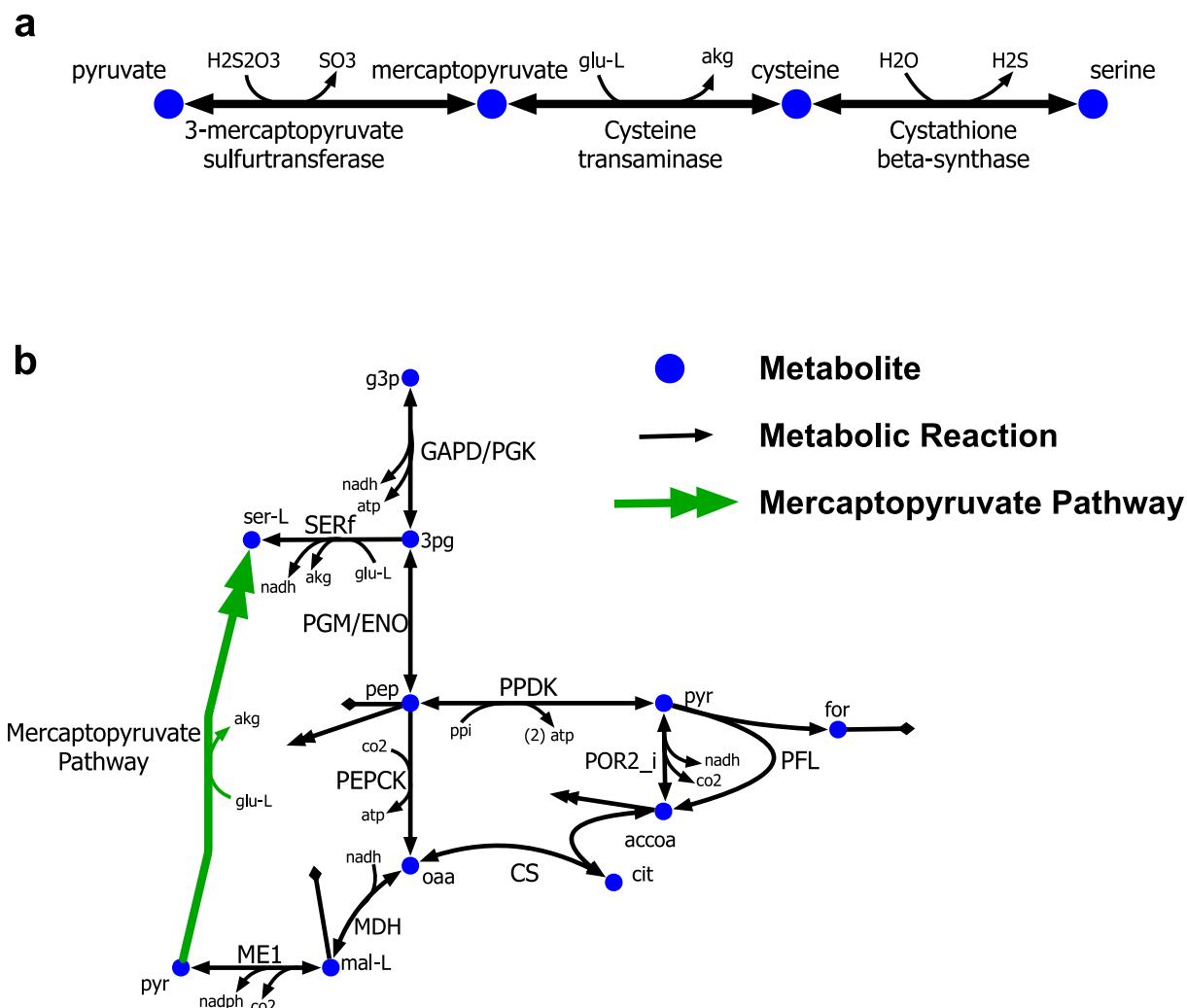


Fig. 3. Mercaptopyruvate pathway for biosynthesis of L-serine from pyruvate (a) Reactions and intermediate metabolites in mercaptopyruvate pathway reactions (b) Mercaptopyruvate pathway role in *C. therm* core metabolism. See Supplementary file S1 for a list of reaction and metabolite abbreviations.

Table 2

A comparison of SSR and degrees of freedom upon flux elucidation with either traditional serine synthesis pathway as the sole serine source or the mercapto-pyruvate pathway as the sole serine source.

	Without mercaptopropane pathway	With mercaptopropane pathway
Degrees of freedom	144	142
Upper threshold for statistically acceptable SSR	179.1	176.9
SSR	206.3	102.3

3-mercaptopyruvate sulfurtransferase enzyme (Kanehisa and Goto, 2000; O’Leary et al., 2016), suggesting it may exhibit promiscuity towards pyruvate. In the second step, 3-mercaptopyruvate is converted to L-cysteine by cysteine transaminase, (present in *C. therm*, encoded by gene CLO1313_RS07420 (Thompson et al., 2016)). Cystathionine β -lyase is an alternative means for converting pyruvate to L-cysteine, and is also annotated in *C. therm* (CLO1313_RS01965). The final step is the conversion of L-cysteine to L-serine. A bidirectional BLAST search revealed that gene CLO1313_RS11805 (cysteine synthase) (O’Leary et al., 2016) is a 55% homologue with 99% sequence coverage to *Helicobacter cinaedi* cystathionine β -synthase (reversible (Jhee and Kruger,

Table 3

Table 3
 Comparison of 3-phosphoglycerate (3 pg) and pyruvate (pyr) experimentally measured mass isotopomer fractional abundance, that estimated without the mercaptopyruvate pathway, and that estimated with the mercaptopyruvate pathway included in model (and the associated weighted squared errors).

Mass isotopomer	Experimental Fractional Abundance	With mercaptopyruvate pathway		Without mercaptopyruvate pathway	
		Fractional abundance	Squared error	Fractional abundance	Squared error
3 pg M+2	0.0945	0.0946	0.0007	0.1166	19.5683
3 pg M+3	0.3452	0.3542	0.0001	0.3374	2.4241
pyr M+2	0.2703	0.2726	0.0558	0.2888	3.4398
pyr M+3	0.1466	0.1391	0.5567	0.0885	33.6912

2005), known to catalyze conversion of L-cysteine to L-serine (Kanehisa and Goto, 2000)). qPCR analysis (Dumitache et al., 2017; Sander et al., 2015) indicates CLO1313_RS12845, CLO1313_RS07420, CLO1313_RS01965, and CLO1313_RS11805 are all expressed in both planktonic and sessile *C. therm* cell cultures.

Note that inclusion of atom transitions for either L-serine deaminase or the glycine cleavage system offered similar resolution of 3-phosphoglycerate and pyruvate mass isotopomer distributions as the mercaptopyruvate pathway. These reactions are not annotated for *C. therm.* A

bidirectional BLAST indicates the closest sequence alignment to L-seine deaminase is homoserine dehydrogenase (CLO1313_RS09815, 39% homologue, 53% sequence coverage of Steinernema glaseri L-seine deaminase). The best evidence for the glycine cleavage system in *C. therm* is for an annotated RNA/DNA binding cold shock protein performing decarboxylation (P-protein) (CLO1313_RS12925, 49% homologue, 93% sequence coverage Roseobacter sp. P-protein), tryptophan synthase subunit exhibiting promiscuous dihydrolipoamide dehydrogenase (L-protein) activity (CLO1313_RS04300, 57% homologue, 45% sequence coverage Stylophora pistillata L-protein), and glycine hydroxymethyltransferase exhibiting promiscuous amino-methyltransferase (T-protein) activity (CLO1313_RS05865, 47% homologue, 41% sequence coverage Anaerolineae bacterium T-protein).

The thermodynamic plausibility of each pathway was also assessed. Table 4 lists the standard Gibbs free energy of transformation (ΔG°) for reactions in the mercaptopyruvate pathway, L-serine deaminase, and glycine cleavage system as predicted by dGpredictor (Wang et al., 2021) and eQuilibrator (Noor et al., 2012). Results indicate that at standard conditions, flux from pyruvate to L-serine and forward flux through the glycine cleavage pathway involve positive though relatively small values for the standard Gibbs energy of the reaction alluding to possible reversibility. The ^{13}C -MFA flux distribution that best agrees with the labeling data involves a reversed flux through either the mercaptopyruvate pathway, glycine cleavage system, or L-serine deaminase (i.e., from pyruvate to L-serine in the case of the mercaptopyruvate pathway and L-serine deaminase, or from CO_2 and 5,10-methylenetetrahydrofolate to glycine in the case of the glycine cleavage system). Removal of the pyruvate mass distribution vectors from the dataset yields best fit flux distributions that flip the flux direction towards the forward direction of either the glycine cleavage system or the mercaptopyruvate pathway. Inclusion of either forward pathway in the model provides the same improvement of fit (F-test-confirmed, 99.9% confidence). Note that the glycine cleavage system is known to operate in the reverse direction under anaerobic conditions (Kikuchi et al., 2008). Therefore based on the preponderance of evidence of (i) literature supporting reverse glycine cleavage system flux and (ii) the ΔG° of all reactions in the mercaptopyruvate pathway are positive but as small as other glycolytic reactions that operate in the forward direction (e.g., FBA $\Delta G^\circ = 5.3 \pm 0.2 \text{ kcal mol}^{-1}$, GLYCL $\Delta G^\circ = 5.8 \pm 1.2 \text{ kcal mol}^{-1}$), we chose to retain the pyruvate mass distribution vectors in the subsequent analyses. In addition, we chose to retain the mercaptopyruvate pathway in k-ctherm138 because (i) it substantially improves model fit, (ii) its

inclusion passed an F-test ($p\text{-value} = 2.7\text{e}^{-5}$), and (iii) two out of the three carbon transitions are annotated in *C. therm* while the third shares reasonable sequence identity (i.e., 59%, 99% coverage) with *C. therm* protein loci. In contrast, we excluded the L-serine deaminase reaction, sharing an identical carbon atom transition with the mercaptopyruvate pathway, as the sequence alignment evidence was borderline. Note that if the enzymes identified as mercaptopyruvate pathway homologues encoded unique activities, a simple knockout experiment would have sufficed to determine whether the pathway was active or not. However, each identified homologue catalyzes multiple required steps in metabolism (Garcia et al., 2020a), which would confound interpretation of such results.

3.3. Resolution of lower glycolysis and malate shunt flux

The best fit metabolic flux distribution and 95% confidence interval upper and lower bounds are illustrated in Fig. 4 alongside flux variability analysis (FVA) ranges from an identical network (constrained by substrate uptake and product excretion). The incorporation of experimental labeling data into flux elucidation caused a contraction in flux ranges for 50 of 70 reactions across *C. therm* central carbon metabolism. The best fit flux distribution yielded a 76/24 pyruvate phosphate dikinase (PPDK)/malate shunt flux split ratio. The 95% confidence interval could be retained for split ratios between 83/16 and 59/41. The experimentally derived confidence interval for the PPDK/malate shunt split estimated by Olson et al. (2017) using dynamic ^{13}C -labeling experiments is completely subsumed within the ^{13}C -MFA estimated 95% confidence interval. The obtained results confirm that both PPDK and the malate shunt are significant sources of pyruvate in *C. therm*. The split ratio was constrained by the introduction of a unique carbon labeling pattern at the oxaloacetate node in the malate shunt. Reverse flux from the (Re)-citrate synthase reaction delivered 9 mmol/100 mmol cellobiose uptake from citrate towards oxaloacetate and acetyl-CoA. This reverse flux allowed for the incorporation of unlabeled carbon into oxaloacetate (and subsequently malate) at the third and fourth carbon positions, helping to resolve the M+2 malate mass isotopomer abundance. (Re)-citrate synthase was allowed to carry exchange flux based on the observations made by Mall et al. (2018). INCA ^{13}C -MFA model and flux maps are included in Supplementary Files S4 and S5 at our group's github repository (<https://github.com/maranasgroup/k-ctherm138>).

3.4. k-ctherm138 parameterization

A total of 1,200 multi-starts were performed to ensure convergence to the best model possible and solution reproducibility. The best model parameterized yielded an SSR value of 167.4, and 14 additional multi-starts converged to models with SSR values within 10% of the optimal value, with the next best model yielding an SSR of 170.3. Parameterization of k-ctherm138 took a total of 42 h of clock time. Of 122 predicted metabolic yield values to which data was fitted, 78% fell within a single standard deviation of the corresponding experimental value. Overall, K-FIT offered a 9.5-fold improvement in SSR over a genetic algorithm (GA) based implementation (Dash et al., 2017), as the total SSR for the GA parameterization was 1,593. Of the 967 elementary kinetic parameters contained in k-ctherm138, 819 (85%) had a coefficient of variation (CV) less than one across the top 15 models, indicating partial resolution. However, only 93 (10%) parameters had a CV less than 0.25, indicating most parameters could assume a wide range of values. Because the only mutant data available for model training corresponds to a small subset of reactions (fermentation product exchange fluxes only), and because the number of parameters far outweigh the number of training datapoints, a high degree of parameter uncertainty across the metabolic network is expected. The majority of well-resolved (i.e., CV < 0.25) parameters (81%) belonged to glycolysis, pyruvate metabolism, and exchange reactions because of their proximity to fitted yield datapoints, while 16%

Table 4

Standard Gibbs free energy of transformation and associated uncertainty for reactions in the mercaptopyruvate pathway, L-serine deaminase, and glycine cleavage system estimated by dGPredictor and eQuilibrator (pH = 7.0, pMG = 3.0 (eQuilibrator only), Ionic Strength = 0.1 M). Abbreviations: ΔG° : standard Gibbs free energy of transformation, SD: standard deviation, mpyr: 3-mercaptopropionate, trdrd: reduced thioredoxin, trdox: oxidized thioredoxin, glu-L: L-glutamate, akg: 2-oxoglutarate, cys-L: L-cysteine, ser-L: L-serine, pyr: pyruvate, gly: glycine, thf: tetrahydrofolate, mlthf: 5,10-methylenetetrahydrofolate.

Reaction	Stoichiometry	dGPredictor (kcal mol ⁻¹)		eQuilibrator (kcal mol ⁻¹)	
		ΔG°	SD	ΔG°	SD
3-mercaptopropionate sulfurtransferase	mpyr + trdrd \rightarrow pyr + H_2S + trdox	-2.9	8.1	-3.0	2.3
Cysteine transaminase	mpyr + glu-L \rightarrow cys-L + akg	3.6	8.0	4.2	2.2
cystathione β -synthase	cys-L + H_2O \rightarrow ser-L + H_2S	6.4	1.4	6.2	2.0
cystathione β -lyase	cys-L + H_2O \rightarrow pyr + NH_4 + H_2S	-0.70	1.4	-0.93	2.0
L-serine deaminase	ser-L \rightarrow pyr + NH_4	-7.1	0.82	-7.1	0.43
Glycine cleavage system	gly + NAD + thf \rightarrow CO_2 + mlthf + NADH + NH_4	5.8	1.2	5.5	1.7

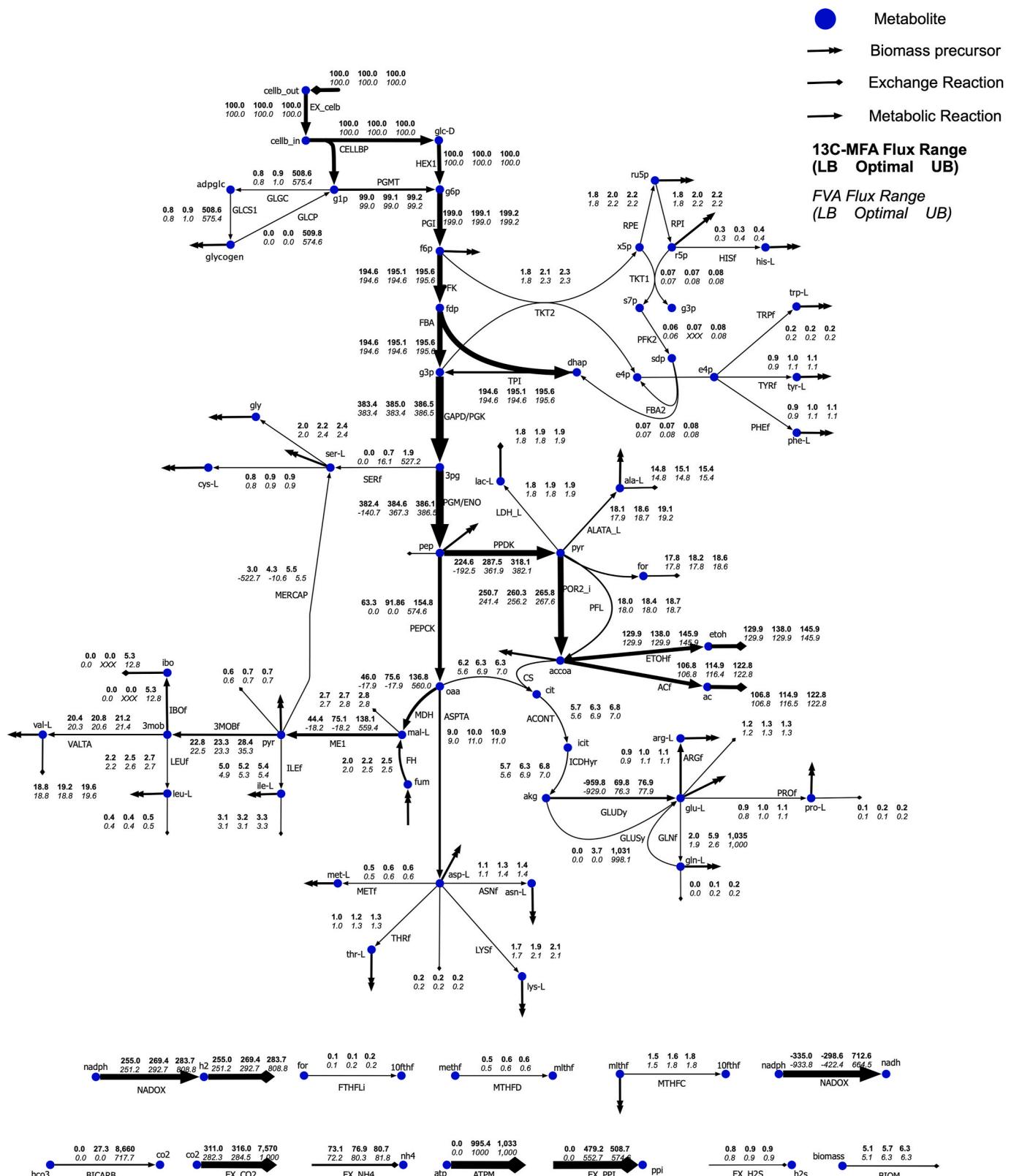


Fig. 4. Optimal ^{13}C -MFA and 95% confidence interval upper and lower bounds of wild-type *C. therm* grown on cellobiose. 100 mmol cellobiose uptake was used as a basis. Italicized values are FVA flux ranges for identical network and conditions. Line thickness proportional to best-fit flux value. See Supplementary file S1 for a list of reaction and metabolite abbreviations.

belonged to the irreversible elementary steps in biomass precursor sink reactions. The proximity of well-resolved parameters to measured yield datapoints supports the prevailing opinion that the parameter resolution in top-down kinetic parameterization is highly dependent upon data availability. Elementary kinetic parameters for k-ctherm138 and all models yielding SSR within 10% of k-ctherm138, mean parameter values, standard deviations/coefficients of variations, and k-ctherm138-predicted yield and metabolite concentrations are reported in Supplementary file S2.

3.5. Probing the documented bacterial substrate-level regulatory space for unidentified regulatory mechanisms in *C. therm* central carbon metabolism

The k-ctherm138 regulatory network is illustrated in Fig. 2 and contains all documented substrate-level regulations acting on *C. therm* enzymes from across the *Clostridium* genus. However, there are likely

regulations active in *C. therm* that have not been experimentally validated, as significantly less effort has been put into understanding substrate-level interactions of *C. therm* enzymes than other bacterium species such as *E. coli*. For k-ecoli74 (Foster et al., 2019) (an *E. coli* network similar in size) 55 inhibitions were considered and ported from the BRENDa (Placzek et al., 2017) and Ecocyc (Keseler et al., 2017) databases. We, therefore, developed a method to expand the scope of the *C. therm* regulatory network by identifying a set of putative regulations that have not been documented in *C. therm*, but whose inclusion would significantly improve yield prediction. We tested one-at-a-time the potential of each bacterial regulation from the comprehensive list outlined in the methods to improve model fitness. Note that parameterization with each added regulation was re-initialized from a confirmed local minimum. Those which improved model fitness to the training data while assuming non-zero parameter values were retained as potential regulations, whereas the ones which did not improve fitness were

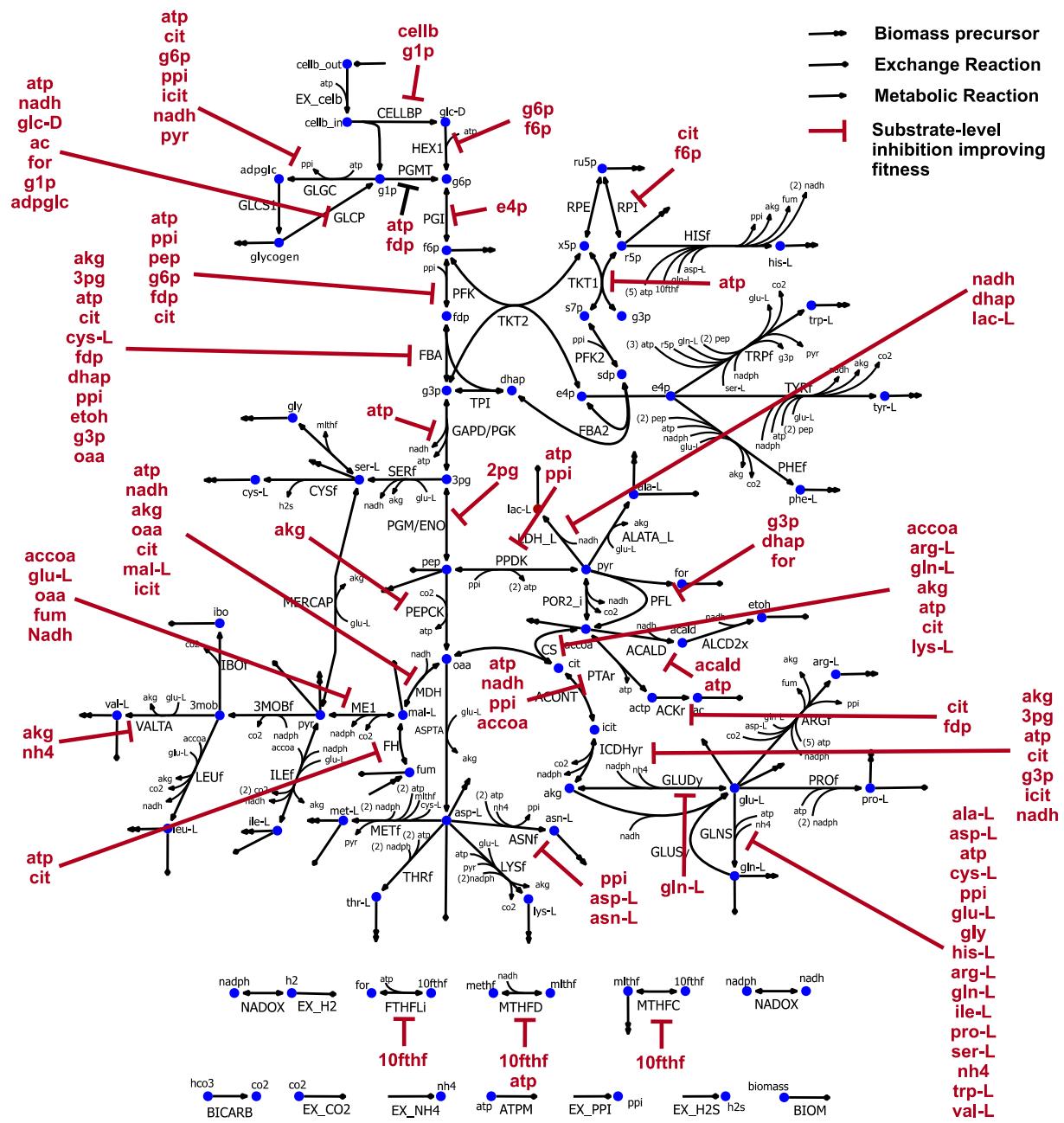


Fig. 5. k-ctherm138 metabolic network, with regulations shown to improve fitness upon addition to k-ctherm138 and re-initialization from local minimum with K-FIT algorithm.

discarded. This procedure identified 105 putative substrate level inhibitions across core metabolism improving model fitness to training data with non-zero parameter values (illustrated in Fig. 5). The improvement in model fitness is reported in Supplementary File S2. Regulations improving fitness did so by at least 3% with NH₄⁺ inhibition of aldehyde dehydrogenase improving fitness the most, decreasing the SSR by 9%. NH₄⁺ and a number of other positive monovalent alkali metals, positive bivalent alkali earth metals, and transition metal ions have been shown to inhibit *E. coli* (Jo et al., 2008) and *Klebsiella pneumoniae* (Raj et al., 2010) aldehyde dehydrogenase activities, but these inhibitions have yet to be explored in *C. therm*.

Even though as many as 105 inhibitions across central carbon metabolism have the capacity to improve kinetic model fitness, it is unclear whether they truly capture substrate-level interactions or are simply artifacts of the fitting process. To generate an orthogonal line of evidence, we perform ensemble docking simulations on the identified regulations and assign a confidence score based on the results. For each inhibition, binding energy of the putative competitive inhibitor in the

active site of the respective enzyme was calculated and compared with positive control binding energies obtained from crystal structures of protein complexes with the same inhibitor molecule bound. Positive controls in essence quantify the range of binding energies observed when the ligand in question interacts with a protein. They were established by first identifying all protein complexes bound to the inhibitor molecules in question from the protein data bank (PDB) (Berman et al., 2000). Positive control binding scores were then calculated by minimizing binding energy between the inhibitor molecules in the binding pocket of corresponding protein crystal structures. For 39 of 46 inhibitors, at least 20 crystal complexes were found and used in the positive control simulations (see Supplementary File S3 for number of complexes used for each inhibitor). For each inhibitor, the mean binding energy and standard deviation of all identified complexes was calculated and used for subsequent analysis (positive control binding energies for each inhibitor are detailed in Supplementary File S3). Any putative inhibitor-enzyme complex that scored a binding energy less than (lower binding energy is more favorable) or equal to its respective mean

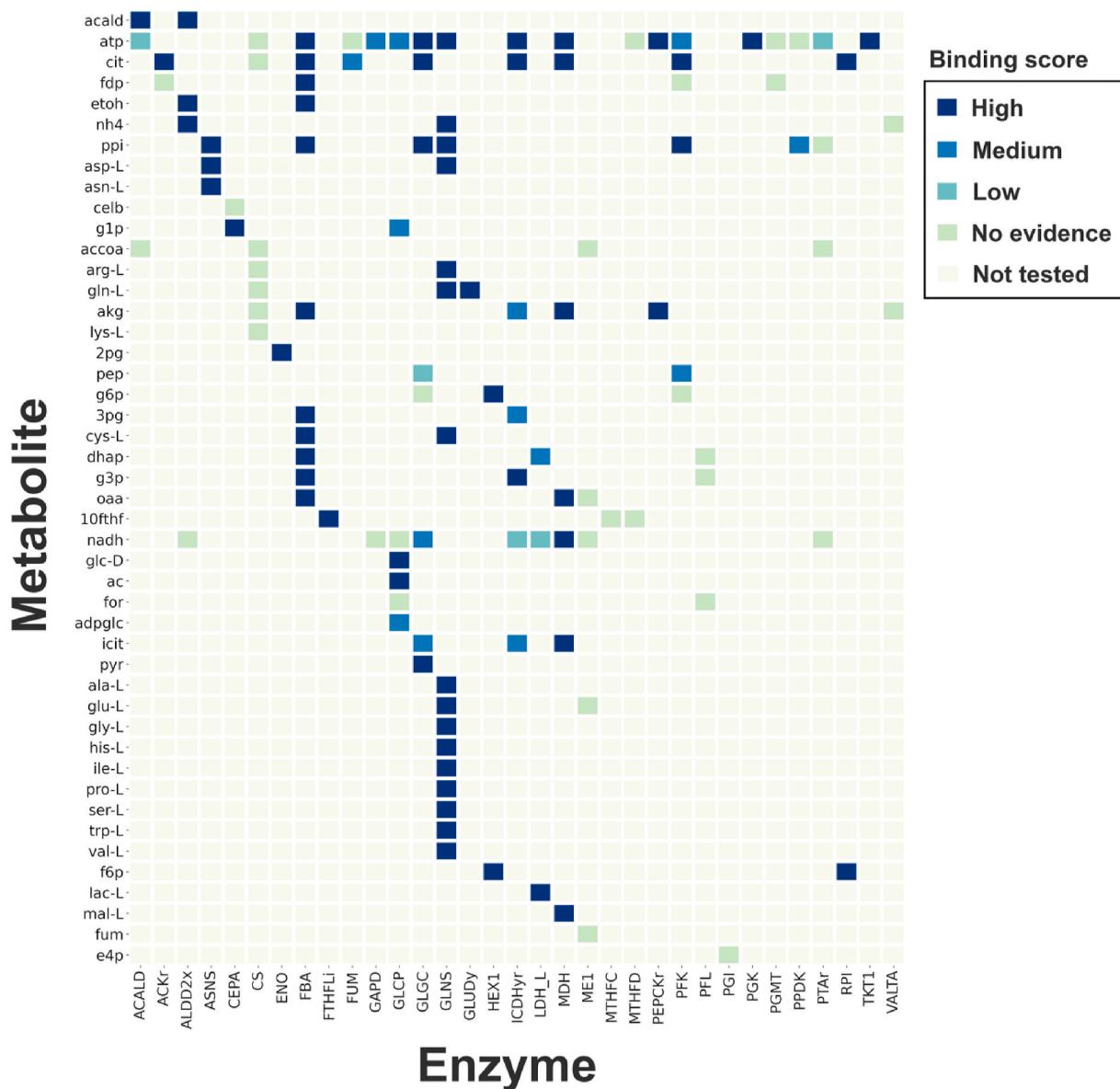


Fig. 6. All possible inhibitor - enzyme pairs are shown as a grid with those improving kinetic model fitness to training data (and thus tested via ensemble docking) colored. Inhibitions assigned high (greater than or equal to mean control binding score), medium (within +0.5 SD of control mean), low (within +1.0 SD), and negative (outside + 1.0 SD) binding scores are shown with colors dark blue, blue, cyan and light green respectively.

positive control was assigned a ‘high’ binding score. Those that obtained binding energy scores smaller in magnitude than, but within ± 0.5 or ± 1.0 standard deviations of their respective mean positive control values were assigned ‘medium’ or ‘low’ binding scores, respectively. Similarly, putative inhibitor-enzyme complexes with binding energies scores $+1.0$ standard deviation below their respective mean positive control values were assigned ‘no evidence’ regulation assessments. Fig. 6 shows the binding scores assigned to all enzyme-inhibitor pairs tested. Note that only the pairs predicted to improve k-therm138 fitness to training data were tested using structural studies and assigned binding scores (all other enzyme-inhibitor pairs are labeled as ‘not tested’ in Fig. 6).

As shown in Fig. 6, a total of 67 out of 105 inhibitions assessed by ensemble docking simulations were assigned a high binding score. Of the remaining inhibitions, 15 and 5 were assigned medium and low binding scores respectively, while the rest were assigned a negative binding score. Notably, our workflow identifies ATP as the most prominent inhibitor across *C. therm* central carbon metabolism with 13 of 18 tested inhibitions receiving a high or medium binding score. These

regulations span glycolysis, glycogen synthesis and degradation, the pentose phosphate pathway, and the malate shunt. Citrate and pyrophosphate were also identified as prevalent inhibitors. While citrate inhibitions receiving high or medium scores were primarily on glycolysis and the TCA cycle, pyrophosphate inhibitions receiving high or medium scores were on amino acid metabolism and glycolytic reactions. Docking citrate, isocitrate, and 2-oxoglutarate to malate dehydrogenase yielded high scores. This result implicates malate dehydrogenase as a control point for the malate shunt/left branch of TCA cycle, and suggests that the increased activity in the right branch of the TCA cycle could throttle flux through the left branch. Thus, the activity of the left branch is intrinsically dependent upon the activity of the right, despite their disconnection in *C. therm*. The results also implicate fructose bisphosphate aldolase and glutamine synthetase as highly regulated enzymes. Along with malate dehydrogenase (7), all inhibitions tested for fructose bisphosphate aldolase (11) and glutamine synthetase (16) received high scores; central carbon intermediates primarily inhibit fructose bisphosphate aldolase, while amino acids bind strongly with glutamine synthetase. Each inhibition tested has been identified experimentally in at

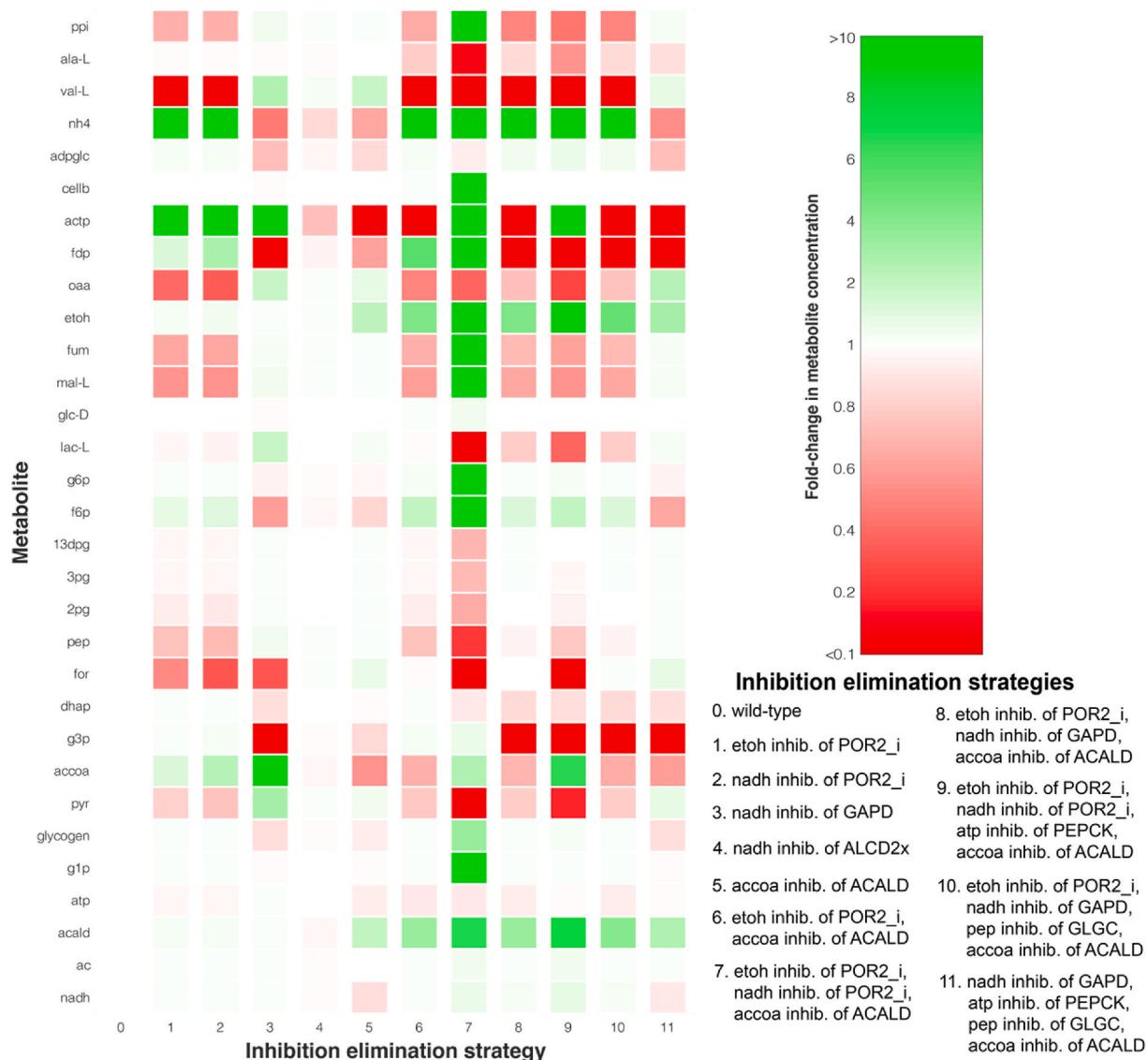


Fig. 7. Fold-change in glycolytic, pyruvate metabolism, and ethanol/acetate metabolism metabolite concentration for top elimination strategies with between one and four substrate-level regulation deletions. Reaction/enzyme abbreviations: POR2_i: pyruvate ferredoxin oxidoreductase, GAPD: glyceraldehyde 3-phosphate dehydrogenase, ALCD2x: alcohol dehydrogenase, ACALD: acetaldehyde dehydrogenase, GLGC: glucose 1-phosphate adenyltransferase, PEPCK: phosphoenolpyruvate carboxykinase. Metabolite abbreviations are defined in Supplementary file S2.

least one other bacteria, but none have been confirmed in a *Clostridium* organism (Placzek et al., 2017). Thus, these results can serve as starting point for expanding our knowledgebase of the substrate-level regulatory network in *C. therm*, and the genus *Clostridium* so that we may better identify intervention targets enabling their development as production platforms.

3.6. Probing *C. therm* substrate-level regulations for limitations to ethanol production

We used k-ctherm138 to pinpoint which substrate-level regulations limited ethanol titer/yield by successively removing a single substrate-level regulation (out of 19) at a time. Predictions are for batch culture conditions (consistent with experimental training data, Sections 2.2, Supplementary File S1). Because batch fermentation yields were assumed to correspond to a metabolic steady-state during parameterization and because the SSF-Estimator module of the K-FIT algorithm was used to evaluate fold-change in metabolite concentration and fermentation yields, all concentration and yield predictions are at metabolic steady-state. Because k-ctherm138 was trained to metabolic yield data, it was used to predict product yield per unit of substrate utilization and fold change in steady-state metabolite concentration upon network perturbation, and not absolute metabolic rates. Fig. 7 depicts the central carbon intermediate metabolite concentration fold-changes from the wild-type strain predicted by k-ctherm138 for several of the top regulatory relief strategies for improving *C. therm* ethanol titer. Because K-FIT predicts fold-change in metabolite concentration rather than absolute metabolite concentration, a 1.4 g L⁻¹ batch culture wild-type ethanol titer (Tian et al., 2016) is used as a basis for comparison and adds context in subsequent analysis. Note that because k-ctherm138 was not trained with intracellular metabolite concentration data, predicted metabolite concentrations in perturbation strategies that deviate significantly from the wild-type strain can exceed physiologically relevant ranges. Nevertheless, the qualitative trends in the predicted pool sizes remain consistent with cellular phenotype. Strategies involving the elimination of alcohol dehydrogenase product inhibition by ethanol lead to physiologically irrelevant predicted ethanol titers and were not considered in subsequent analysis. Acetyl-CoA competitive inhibition of acetaldehyde dehydrogenase (ACALD) (Fig. 7, strategy 5) was the single regulation limiting ethanol titer most. Its removal resulted in a 3.0-fold increase in steady-state ethanol titer from 1.4 to 4.2 g L⁻¹, but the metabolome was not significantly re-distributed. Perturbed metabolite concentrations were contained primarily to ethanol and acetate metabolism. A -1.6-fold decrease in steady-state acetyl-CoA concentration was observed, as the competitive inhibition of ACALD no longer needed to be overcome to increase the pool sizes of acetaldehyde and ethanol, both of which increased in steady-state concentration (2.8 and 3.0-fold, respectively). In other single regulation relief strategies for which ethanol concentrations increased (Fig. 7, strategies 1–4), acetyl-CoA concentration also increased (between 2.1 and 11.2-fold increase) to overcome the inhibition on ACALD. In pyruvate ferredoxin oxidoreductase (POR2_i) inhibition relief strategies, pyruvate concentration decreased by -1.18-fold and -1.25-fold because elevated pyruvate concentration was not needed maintain POR2_i activity. Malate shunt intermediate concentrations also decreased by between -1.25-fold and -2.1-fold. These strategies were not as effective, yielding at most a 1.4-fold increase in steady-state ethanol concentration to 2.0 g L⁻¹ (only 21% of the increase observed upon removal of acetyl-CoA inhibition of ACALD) and no increase in ethanol yield. Thus, acetyl-CoA inhibition of ACALD is the single most significant substrate-level regulation limiting ethanol titer. This target is consistent with the experimental results pointing to mutations in NADH binding domains of the *adhE* enzyme yielding significant improvement in ethanol titer (Brown et al., 2011; Zheng et al., 2015).

Regulatory relief strategies that included the combined elimination of acetyl-CoA inhibition of ACALD and ethanol and NADH inhibitions of

POR2_i lead to the greatest increase in ethanol concentration (9.4-fold increase to at least 13.2 g L⁻¹). When these three regulations were eliminated simultaneously (Fig. 7, strategy 7), large perturbations in metabolite concentration from the wild-type strain were observed across central carbon metabolism. In upper glycolysis, glucose 6-phosphate (>20-fold), fructose 6-phosphate (>20.0-fold), and fructose 1,6-bisphosphate (>20.0-fold) concentrations increased, while lower glycolytic metabolite concentrations decreased. Similar to the POR2_i single regulation relief strategies, acetyl-CoA concentration increased by 3.4-fold despite the decrease in pyruvate concentration. However, since the substrate-level inhibition on ACALD was also removed, ethanol concentration increased by 9.8-fold. This regulation elimination strategy also resulted in the greatest increase in ethanol yield, from 137 to 146 mmol per 100 mmol cellobiose consumed. Combining this strategy with the elimination of ATP inhibition of phosphoenolpyruvate carboxykinase (PEPCK) (Fig. 7, strategy 9) yielded the maximum increase in ethanol titer (10.4-fold increase to 14.6 g L⁻¹). The metabolome redistribution in this strategy was similar to relief strategy 7, but changes in upper and lower glycolytic metabolite concentrations from the wild-type strain were less pronounced. Malate shunt utilization also increased by 26 mmol gDW⁻¹ h⁻¹ per 100 mmol gDW⁻¹ h⁻¹ cellobiose consumed in this strategy. Overall, the relief strategies identified *in silico* provide several metabolic engineering targets for improving ethanol yields *in vivo*, and are consistent with homologue substitution targets identified experimentally (Hon et al., 2017; Tian et al., 2017). Homologue substitution experiments targeting POR2_i and *adhE* simultaneously lead to an increase in batch fermentation ethanol titer to 15 g L⁻¹ (Hon et al., 2017). Upon testing the identified ethanol-limiting inhibitions using ensemble docking, ATP inhibition of PEPCK was found to have a high confidence and phosphoenolpyruvate inhibition of glucose 1-phosphate adenyltransferase was found to have a medium confidence, while the rest had either a low or exhibited no evidence. (See Supplementary File S3 for details).

4. Discussion

The analysis introduced here puts forth an integrated workflow for leveraging process and isotope labeling data to parameterize a kinetic model of metabolism and assess the dispensability of different pathways under investigation. In addition, putative substrate-level regulations are ranked-ordered based on how detrimental their removals are to model fitting. An orthogonal avenue of assessing substrate-level regulation is introduced that relies on MD simulation followed by ligand docking to assess the structural compatibility of interaction between the putative competitive inhibitors and the enzymatic active site. We believe that the proposed framework has the potential to reduce the amount of experimental data needed to generate kinetic models, which is particularly important for non-model organisms. This includes the majority of non-model organisms for which the lack of reliable regulatory information is a primary hurdle (Foster et al., 2020; Placzek et al., 2017; Strutz et al., 2019).

While using process data to train kinetic models can ensure the kinetics of terminal points in metabolism are captured, isotopic labeling data are necessary to reliably constrain flux ranges in networks with branched pathways. For example, in the k-ctherm118 kinetic model of *C. therm* central carbon metabolism (Dash et al., 2017), the reported wild-type flux distribution estimated using only flux balance analysis cycled carbon between the malate shunt and PPDK, and the flux split ratio was not consistent with isotopic labeling experiments (Olson et al., 2017). This inaccuracy was propagated to kinetic parameters. Using isotopic tracer data, we were able to resolve the flux split ratio between the malate shunt and PPDK in our wild-type flux distribution. Thus, our kinetic parameters reliably capture pyruvate metabolism physiology. However, uncertainty remains in this split ratio in mutant strains, and reliably predicting the usage of these competing pathways under perturbed conditions falls outside of the scope of model k-ctherm138.

Isotopic labeling data can also be used to distinguish between dispensable and required pathways to achieve good parameterization fit. Through analysis of the goodness of fit in response to pathway addition/removal on ^{13}C -MFA results, careful interpretation of protein sequence alignments and transcriptomics or proteomics data, pathways necessary to describe observed isotopic labeling distributions can be identified. In the absence of isotopic labeling data, a common strategy is FBA-based flux projection form a core to a genome-scale model. However, as shown for *E. coli* (Foster et al., 2019), projecting fluxes elucidated using a simplified network to a larger network can lead to inaccuracies/wide ranges in flux values which propagate to kinetic parameters. In addition, assigning parameters based on arbitrary flux assignments due to their wide ranges can lead to unrealistic parameter values in pathways which could otherwise be fully resolved via ^{13}C -MFA (Gopalakrishnan and Maranas, 2015; Khodayari and Maranas, 2016).

The need to capture substrate-level regulation to accurately describe metabolic phenotypes was clearly demonstrated with yeast glycolysis (Smallbone et al., 2013). In non-model organisms with limited substrate-level regulatory information, the only option is to simply port all regulatory information that can be gleaned for related organisms, as oftentimes very limited regulations are documented specific to the non-model organism of interest (Greene et al., 2019). The obvious problem with this strategy is the model will likely include many incorrect substrate-level regulations. Therefore, systematic criteria need to be applied before a regulation enters the kinetic model. Approximately 25% (25 of 105) of regulations identified as improving k-ctherm138 fitness to training data were by either ATP or NADH. Of these, 17 were on reactions with at least one native cofactor or substrate containing an adenine moiety. The presence of the same moiety (e.g., adenine) in both inhibitor and native substrate appears to be a good indicator that the enzyme in question forms favorable interactions with the inhibitor in question. This is confirmed by our ensemble docking results for inhibitions involving ATP, for which six of nine inhibitions on catalytic domains for metabolites containing adenine moiety scored high or medium. Moving beyond improvements in model fitness, structural analysis can investigate into the mechanistic basis for adopted substrate-level regulations. In this study, we have deployed an ensemble docking workflow and used a positive control set to assess the plausibility of competitive inhibitions. However, incorporation of a negative control set may also be necessary, as this could help us pinpoint the extent of binding interaction threshold needed to declare a binding event as competitive inhibition. Assessment of competitive inhibition using protein structural analysis requires only measurement of binding strength as opposed to that of allosteric regulations (i.e., uncompetitive, non-competitive inhibition) which may require monitoring long range inter-domain interactions and/or conformational changes of active sites. An ensemble-docking approach scored using multiple snapshots from MD trajectory might be necessary for this task as static docking are prone to predicting incorrect binding poses when only a single receptor conformation is considered (Totrov and Abagyan, 2008). Longer MD simulations could also be useful to monitor changes in residue interaction networks connecting active and allosteric sites that are prevalent in allosteric regulations (Bowerman and Wereszczynski, 2016). Machine learning models based on protein structure and sequence may provide an alternative means to predict substrate-level interactions assuming that sufficiently large datasets on such interactions become available. Both in the case of MD simulations and machine learning based inference, negative interaction data (Cheng et al., 2017; Najm et al., 2021) would be important to establish thresholds on the extent of binding interactions and the unbiased training of models, respectively.

Beyond capturing regulation, capturing network-wide proteome shifts in response to genetic and environmental perturbations remains a challenge for the successful deployment of large-scale kinetic modeling frameworks. Resource allocation models (Yang et al., 2018) provide a systematic way of capturing condition-dependent changes in enzyme level, and are thus well-suited for integration with kinetic models of

metabolism. By enforcing resource allocation model predictions as perturbations to enzyme level within kinetic modeling frameworks, the true in-vivo kinetics of the system can be captured. As we continue to develop interoperable workflows between metabolic and structural modeling it is also important to establish ways to link protein-ligand binding energetics and kinetic parameters. This will ultimately enable us to better predict with a kinetic model how point mutations, gene up/down regulations, and environmental perturbations can lead to desired metabolic outcomes.

Ultimately, for useful kinetic models of non-model organisms to be readily developed and deployed, adaptable workflows for regulatory network elucidation and kinetic parameterization from limited data are necessary. To maximize the potential of these tools, it is of critical importance for modelers to adhere to reproducibility standards and make concerted efforts to share models on platforms accessible to the broader scientific community. This would help accelerate the integration of large-scale kinetic models of metabolism in design-build-test-learn cycles to inform metabolic network redesign.

Author statement

Charles Foster: Conceptualization, Methodology, Software, Formal Analysis, Writing- Original draft preparation, Investigation **Veda Sheersh Boorla:** Conceptualization, Methodology, Software, Writing – Original Draft **Satyakam Dash:** Methodology **Saratram Gopalakrishnan:** Methodology **Tyler Jacobson:** Data Curation, Writing – Review and Editing **Daniel G. Olson:** Writing – Review and Editing **Daniel Amador-Noguez:** Supervision, Funding Acquisition, Writing – Review and Editing **Lee R. Lynd:** Supervision, Funding Acquisition, Project Administration, Conceptualization, Writing – Review and editing **Costas D. Maranas:** Conceptualization, Writing – Review and Editing, Supervision, Project Administration, Funding Acquisition.

Acknowledgments

Funding provided by the Center for Bioenergy Innovation, a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy. Funding also provided by NSF Award Number MCB-1615646.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ymben.2021.12.012>.

References

- Akinoshio, H., Yee, K., Close, D., Ragauskas, A., 2014. The emergence of *Clostridium thermocellum* as a high utility candidate for consolidated bioprocessing applications. *Front. Chem.* 2, 66.
- Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., Labonte, J.W., Pacella, M.S., Bonneau, R., Bradley, P., Dunbrack Jr., R.L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., Gray, J.J., 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theor. Comput.* 13, 3031–3048.
- Amaro, R.E., Baudry, J., Chodera, J., Demir, O., McCammon, J.A., Miao, Y., Smith, J.C., 2018. Ensemble docking in drug discovery. *Biophys. J.* 114, 2271–2278.
- Antoniewicz, M.R., Kelleher, J.K., Stephanopoulos, G., 2006. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metab. Eng.* 8, 324–337.
- Batra, R., Chan, H., Kamath, G., Ramprasad, R., Cherukara, M.J., Sankaranarayanan, S., 2020. Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking studies. *J. Phys. Chem. Lett.* 11, 7058–7065.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. *Nucleic Acids Res.* 28, 235–242.

- Biswas, R., Zheng, T., Olson, D.G., Lynd, L.R., Guss, A.M., 2015. Elimination of hydrogenase active site assembly blocks H2 production and increases ethanol yield in Clostridium thermocellum. *Biotechnol. Biofuels* 8, 20.
- Bowerman, S., Wereszczynski, J., 2016. Detecting allosteric networks using molecular dynamics simulation. *Methods Enzymol.* 578, 429–447.
- Bowers, K.J., Sacerdoti, F.D., Salmon, J.K., Shan, Y., Shaw, D.E., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M. A., 2006. Molecular dynamics—Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC, 6.
- Brown, S.D., Guss, A.M., Karpinet, T.V., Parks, J.M., Smolin, N., Yang, S., Land, M.L., Klingeman, D.M., Bhandiwad, A., Rodriguez Jr., M., Raman, B., Shao, X., Mielenz, J. R., Smith, J.C., Keller, M., Lynd, L.R., 2011. Mutant alcohol dehydrogenase leads to improved ethanol tolerance in Clostridium thermocellum. *Proc. Natl. Acad. Sci. U. S. A.* 108, 13752–13757.
- Chandak, T., Mayginnes, J.P., Mayes, H., Wong, C.F., 2020. Using machine learning to improve ensemble docking for drug discovery. *Proteins* 88, 1263–1270.
- Cheng, Z., Huang, K., Wang, Y., Liu, H., Guan, J., Zhou, S., 2017. Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.* 11, 9.
- Christodoulou, D., Link, H., Fuhrer, T., Kochanowski, K., Gerosa, L., Sauer, U., 2018. Reserve flux capacity in the pentose phosphate pathway enables Escherichia coli's rapid response to oxidative stress. *Cell Syst.* 6.
- Clasquin, M.F., Melamud, E., Rabinowitz, J.D., 2012. LC-MS data processing with MAVEN: a metabolic analysis and visualization engine. *Curr. Protoc. Bioinf.* (Chapter 14), Unit14 11.
- Cui, J., Olson, D.G., Lynd, L.R., 2019. Characterization of the Clostridium thermocellum AdhE, NfnAB, ferredoxin and Pfor proteins for their ability to support high titer ethanol production in Thermoanaerobacterium saccharolyticum. *Metab. Eng.* 51, 32–42.
- Dash, S., Khodayari, A., Zhou, J., Holwerda, E.K., Olson, D.G., Lynd, L.R., Maranas, C.D., 2017. Development of a core Clostridium thermocellum kinetic metabolic model consistent with multiple genetic perturbations. *Biotechnol. Biofuels* 10, 108.
- Davison, S.A., Keller, N.T., van Zyl, W.H., den Haan, R., 2019. Improved cellulase expression in diploid yeast strains enhanced consolidated bioprocessing of pretreated corn residues. *Enzym. Microb. Technol.* 131.
- Deng, Y., Olson, D.G., Zhou, J., Herring, C.D., Joe Shaw, A., Lynd, L.R., 2013. Redirecting carbon flux through exogenous pyruvate kinase to achieve high ethanol yields in Clostridium thermocellum. *Metab. Eng.* 15, 151–158.
- Dumitracche, A., Klingeman, D.M., Natzke, J., Rodriguez Jr., M., Giannone, R.J., Hettich, R.L., Davison, B.H., Brown, S.D., 2017. Specialized activities and expression differences for Clostridium thermocellum biofilm and planktonic cells. *Sci. Rep.* 7, 43583.
- Falcon, W.E., Ellingson, S.R., Smith, J.C., Baudry, J., 2018. Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *J. Phys. Chem.* 123, 5189–5195.
- Faraji, M., Fonseca, L.L., Escamilla-Trevino, L., Barros-Rios, J., Engle, N., Yang, Z.K., Tschaplinski, T.J., Dixon, R.A., Voit, E.O., 2018. Mathematical models of lignin biosynthesis. *Biotechnol. Biofuels* 11, 34.
- Faraji, M., Fonseca, L.L., Escamilla-Trevino, L., Dixon, R.A., Voit, E.O., 2015. Computational inference of the structure and regulation of the lignin pathway in *Panicum virgatum*. *Biotechnol. Biofuels* 8, 151.
- Foster, C.J., Gopalakrishnan, S., Antoniewicz, M.R., Maranas, C.D., 2019. From Escherichia coli mutant 13C labeling data to a core kinetic model: a kinetic model parameterization pipeline. *PLoS Comput. Biol.* 15, e1007319.
- Foster, C.J., Wang, L., Dinh, H.V., Suthers, P.F., Maranas, C.D., 2020. Building kinetic models for metabolic engineering. *Curr. Opin. Biotechnol.* 67, 35–41.
- Garcia, S., Thompson, A.R., Giannone, R.J., Dash, S., Maranas, C., Trinh, C.T., 2020a. Development of a genome-scale metabolic model of Clostridium thermocellum and its applications for integration of multi-omics datasets and strain design. *bioRxiv* 1–35.
- Garcia, S., Thompson, R.A., Giannone, R.J., Dash, S., Maranas, C.D., Trinh, C.T., 2020b. Development of a genome-scale metabolic model of Clostridium thermocellum and its applications for integration of multi-omics datasets and computational strain design. *Front. Bioeng. Biotechnol.* 8, 772.
- Gautam, B., Singh, G., Gulshan, W., Farmer, R., Singh, S., Singh, A.K., Jain, P.A., Yadav, P.K., 2012. Metabolic pathway analysis and molecular docking analysis for identification of putative drug targets in *Toxoplasma gondii*: novel approach. *Bioinformation* 8, 134–141.
- Gopalakrishnan, S., Dash, S., Maranas, C., 2020. K-FIT: an accelerated kinetic parameterization algorithm using steady-state fluxomic data. *Metab. Eng.*
- Gopalakrishnan, S., Dash, S., Maranas, C.D., 2019. K-FIT: an Accelerated Kinetic Parameterization Algorithm Using Stead-State Fluxomic Data. *bioRxiv*.
- Gopalakrishnan, S., Maranas, C.D., 2015. 13C metabolic flux analysis at a genome-scale. *Metab. Eng.* 32, 12–22.
- Greene, J., Daniell, J., Köpke, M., Broadbelt, L., Tyo, K.E.J., 2019. Kinetic ensemble model of gas fermenting Clostridium autoethanogenum for improved ethanol production. *Biochem. Eng. J.* 148, 46–56.
- Hackett, S.R., Zanotelli, V.R., Xu, W., Goya, J., Park, J.O., Perlman, D.H., Gibney, P.A., Botstein, D., Storey, J.D., Rabinowitz, J.D., 2016. Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* 354.
- Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., Wang, L., Lupyan, D., Dahlgren, M.K., Knight, J.L., Kaus, J.W., Cerutti, D.S., Krilov, G., Jorgensen, W.L., Abel, R., Friesner, R.A., 2016. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theor. Comput.* 12, 281–296.
- Higashide, W., Li, Y., Yang, Y., Liao, J.C., 2011. Metabolic engineering of Clostridium cellulolyticum for production of isobutanol from cellulose. *Appl. Environ. Microbiol.* 77, 2727–2733.
- Holwerda, E.K., Thorne, P.G., Olson, D.G., Amador-Noguez, D., Engle, N.L., Tschaplinski, T.J., van Dijken, P., Lynd, L.R., 2014. The exometabolome of Clostridium thermocellum reveals overflow metabolism at high cellulose loading. *Biotechnol. Biofuels* 7, 1–11.
- Hon, S., Olson, D.G., Holwerda, E.K., Lanahan, A.A., Murphy, S.J.L., Maloney, M.I., Zheng, T., Papanek, B., Guss, A.M., Lynd, L.R., 2017. The ethanol pathway from Thermoanaerobacterium saccharolyticum improves ethanol production in Clostridium thermocellum. *Metab. Eng.* 42, 175–184.
- Hussain, A.A., Abdel-Salam, M.S., Abo-Ghalia, H.H., Hegazy, W.K., Hafez, S.S., 2017. Optimization and molecular identification of novel cellulose degrading bacteria isolated from Egyptian environment. *J. Genet. Eng. Biotechnol.* 15, 77–85.
- Jacobson, T.B., Korosh, T.K., Stevenson, D.M., Foster, C., Maranas, C., Olson, D.G., Lynd, L.R., Amador-Noguez, D., 2020. In vivo thermodynamic analysis of glycolysis in Clostridium thermocellum and Thermoanaerobacterium saccharolyticum using (13)C and (2)H tracers. *mSystems* 5.
- Jhee, K., Kruger, W.D., 2005. The role of cystathionine beta-synthase in homocysteine metabolism. *Antioxidants Redox Signal.* 7, 813–822.
- Jo, J.E., Mohan Raj, S., Rathnasingh, C., Selvakumar, E., Jung, W.C., Park, S., 2008. Cloning, expression, and characterization of an aldehyde dehydrogenase from Escherichia coli K-12 that utilizes 3-Hydroxypropionaldehyde as a substrate. *Appl. Microbiol. Biotechnol.* 81, 51–60.
- Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.
- Kannuchamy, S., Mukund, N., Saleena, L.M., 2016. Genetic engineering of Clostridium thermocellum DSM1313 for enhanced ethanol production. *BMC Biotechnol.* 16 (Suppl. 1), 34.
- Keseler, I.M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D.A., Weaver, D., Collado-Vides, J., Paulsen, I., Karp, P.D., 2017. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.* 45, D543–D550.
- Khodayari, A., Maranas, C.D., 2016. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat. Commun.* 7, 13806.
- Kikuchi, G., Motokawa, Y., Yoshida, T., Hiraga, K., 2008. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proc. Jpn. Acad. Sci.* 84, 246–263.
- Koes, D.R., Baumgartner, M.P., Camacho, C.J., 2013. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* 53, 1893–1904.
- Konc, J., Lesnik, S., Janezic, D., 2015. Modeling enzyme-ligand binding in drug discovery. *J. Cheminf.* 7, 48.
- Lamed, R., Zeikus, J.G., 1980. Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in Clostridium thermocellum and thermoanaerobium brockii. *J. Bacteriol.* 114, 569–578.
- Lamed, R.J., Lobos, J.H., Su, T.M., 1988. Effects of stirring and hydrogen on fermentation products of Clostridium thermocellum. *Appl. Environ. Microbiol.* 54, 1216–1221.
- Lee, Y., Chen, F., Gallego-Giraldo, L., Dixon, R.A., Voit, E.O., 2011. Integrative analysis of transgenic alfalfa (*Medicago sativa* L.) suggests new metabolic control mechanisms for monolignol biosynthesis. *PLoS Comput. Biol.* 7, e1002047.
- Lee, Y., Lafontaine Rivera, J.G., Liao, J.C., 2014. Ensemble Modeling for Robustness Analysis in engineering non-native metabolic pathways. *Metab. Eng.* 25, 63–71.
- Leis, B., Held, C., Bergkemper, F., Dennermark, K., Steinbauer, R., Reiter, A., Mechelke, M., Moerch, M., Graubner, S., Liebl, W., Schwarz, W.H., Zverlov, V.V., 2017. Comparative characterization of all cellosomal cellulases from Clostridium thermocellum reveals high diversity in endoglucanase product formation essential for complex activity. *Biotechnol. Biofuels* 10, 240.
- Lin, P.P., Mi, L., Morioka, A.H., Yoshino, K.M., Konishi, S., Xu, S.C., Papanek, B.A., Riley, L.A., Guss, A.M., Liao, J.C., 2015. Consolidated bioprocessing of cellulose to isobutanol using Clostridium thermocellum. *Metab. Eng.* 31, 44–52.
- Lo, J., Olson, D.G., Murphy, S.J., Tian, L., Hon, S., Lanahan, A., Guss, A.M., Lynd, L.R., 2017. Engineering electron metabolism to increase ethanol production in Clostridium thermocellum. *Metab. Eng.* 39, 71–79.
- Lynd, L.R., Grethelein, H.E., Wolkin, R.H., 1989. Fermentation of cellulosic substrates in batch and continuous culture by Clostridium thermocellum. *Appl. Environ. Microbiol.* 55, 3131–3139.
- Lynd, L.R., Guss, A.M., Himmel, M.E., Beri, D., Herring, C., Holwerda, E.K., Murphy, S.J., Olson, D.G., Payne, J., Rydzak, T., Shao, X., Tian, L., Worthen, R., 2016. Advances in consolidated bioprocessing using Clostridium thermocellum and Thermoanaerobacter saccharolyticum. *Ind. Biotechnol.* 365–394.
- Lynd, L.R., van Zyl, W.H., McBride, J.E., Laser, M., 2005. Consolidated bioprocessing of cellulosic biomass: an update. *Curr. Opin. Biotechnol.* 16, 577–583.
- Lynd, L.R., Wyman, C.E., Gerngross, T.U., 1999. Biocommodity engineering. *Biotechnol. Prog.* 15, 777–793.
- Mall, A., Sobotta, J., Huber, C., Tschirner, C., Kowarschik, S., Bacnik, K., Mergelsberg, M., Boll, M., Hugler, M., Eisenreich, W., Berg, I.A., 2018. Reversibility of citrate synthase allows autotrophic growth of thermophilic bacterium. *Science* 359, 563–567.
- Melamud, E., Vastag, L., Rabinowitz, J.D., 2010. Metabolomic analysis and visualization engine for LC-MS data. *Anal. Chem.* 82, 9818–9826.

- Nagarajan, D., Lee, D.-J., Chang, J.-S., 2019. Recent insights into consolidated bioprocessing for lignocellulosic biohydrogen production. *Int. J. Hydrogen Energy* 44, 14362–14379.
- Najm, M., Azencott, C.A., Playe, B., Stoven, V., 2021. Drug target identification with machine learning: how to choose negative examples. *Int. J. Mol. Sci.* 22.
- Nargotra, P., Sharma, V., Bajaj, B.K., 2019. Consolidated bioprocessing of surfactant-assisted ionic liquid-pretreated *Parthenium hysterophorus* L. biomass for bioethanol production. *Bioresour. Technol.* 289, 121611.
- Noor, E., Bar-Even, A., Flamholz, A., Lubling, Y., Davidi, D., Milo, R., 2012. An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics* 28, 2037–2044.
- O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745.
- Olsen, J.P., Alasepp, K., Kari, J., Crays-Bagger, N., Borch, K., Westh, P., 2016. Mechanism of product inhibition for cellobiohydrolase Cel7A during hydrolysis of insoluble cellulose. *Biotechnol. Bioeng.* 113, 1178–1186.
- Olson, D.G., Horl, M., Fuhrer, T., Cui, J., Zhou, J., Maloney, M.I., Amador-Noguez, D., Tian, L., Sauier, U., Lynd, L.R., 2017. Glycolysis without pyruvate kinase in *Clostridium thermocellum*. *Metab. Eng.* 39, 169–180.
- Olson, D.G., McBride, J.E., Shaw, A.J., Lynd, L.R., 2012. Recent progress in consolidated bioprocessing. *Curr. Opin. Biotechnol.* 23, 396–405.
- Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., Schomburg, D., 2017. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* 45, D380–D388.
- Pu, L., Govindaraj, R.G., Lemoine, J.M., Wu, H.C., Brylinski, M., 2019. DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.* 15, e1006718.
- Raj, S.M., Rathnasingh, C., Jung, W.-C., Selvakumar, E., Park, S., 2010. A Novel NAD+-dependent aldehyde dehydrogenase encoded by the *puuC* gene of *Klebsiella pneumoniae* DSM 2026 that utilizes 3-hydroxypropionaldehyde as a substrate. *Biotechnol. Bioproc. Eng.* 15, 131–138.
- Ryckaert, J.P., Cicotti, G., Berendsen, H.J.C., 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 321–341.
- Rydzak, T., Levin, D.B., Cicek, N., Sparling, R., 2011. End-product induced metabolic shifts in *Clostridium thermocellum* ATCC 27405. *Appl. Microbiol. Biotechnol.* 92, 199–209.
- Saddler, J.N., Chan, M.K.-H., 1982. Optimization of *Clostridium thermocellum* growth on cellulose and pretreated wood substrates. *Eur. J. Appl. Microbiol. Biotechnol.* 16, 99–104.
- Sander, K., Wilson, C.M., Rodriguez Jr., M., Klingeman, D.M., Rydzak, T., Davison, B.H., Brown, S.D., 2015. *Clostridium thermocellum* DSM 1313 transcriptional responses to redox perturbation. *Biotechnol. Biofuels* 8, 211.
- Sangkharak, K., Vangsirikul, P., Janthachat, S., 2012. Strain improvement and optimization for enhanced production of cellulase in *Cellulomonas* sp. TSU-03. *Afr. J. Microbiol. Res.* 6, 1079–1084.
- Sastray, G.M., Adzhigirey, M., Day, T., Annabhimmoju, R., Sherman, W., 2013. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234.
- Smallbone, K., Messiaha, H.L., Carroll, K.M., Winder, C.L., Malys, N., Dunn, W.B., Murabito, E., Swainston, N., Dada, J.O., Khan, F., Pir, P., Simeonis, E., Spasic, I., Wishart, J., Weichart, D., Hayes, N.W., Jameson, D., Broomhead, D.S., Oliver, S.G., Gaskell, S.J., McCarthy, J.E., Paton, N.W., Westerhoff, H.V., Kell, D.B., Mendes, P., 2013. A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS Lett.* 587, 2832–2841.
- Strutz, J., Martin, J., Greene, J., Broadbelt, L., Tyo, K., 2019. Metabolic kinetic modeling provides insight into complex biological questions, but hurdles remain. *Curr. Opin. Biotechnol.* 59, 24–30.
- Summers, R.J., Boudreault, D.P., Srinivasan, V.R., 1979. Continuous cultivation for apparent optimization of defined media for *cellulomonas* sp. and *Bacillus cereus*. *Appl. Environ. Microbiol.* 38, 66–71.
- Theisen, M.K., Lafontaine Rivera, J.G., Liao, J.C., 2016. Stability of ensemble models predicts productivity of enzymatic systems. *PLoS Comput. Biol.* 12, e1004800.
- Thompson, R.A., Dahal, S., Garcia, S., Nookae, I., Trinh, C.T., 2016. Exploring complex cellular phenotypes and model-guided strain design with a novel genome-scale metabolic model of *Clostridium thermocellum* DSM 1313 implementing an adjustable cellulosome. *Biotechnol. Biofuels* 9, 194.
- Thompson, R.A., Layton, D.S., Guss, A.M., Olson, D.G., Lynd, L.R., Trinh, C.T., 2015. Elucidating central metabolic redox obstacles hindering ethanol production in *Clostridium thermocellum*. *Metab. Eng.* 32, 207–219.
- Tian, L., Conway, P.M., Cervenka, N.D., Cui, J., Maloney, M., Olson, D.G., Lynd, L.R., 2019. Metabolic engineering of *Clostridium thermocellum* for n-butanol production from cellulose. *Biotechnol. Biofuels* 12, 186.
- Tian, L., Papanek, B., Olson, D.G., Rydzak, T., Holwerda, E.K., Zheng, T., Zhou, J., Maloney, M., Jiang, N., Giannone, R.J., Hettich, R.L., Guss, A.M., Lynd, L.R., 2016. Simultaneous achievement of high ethanol yield and titer in *Clostridium thermocellum*. *Biotechnol. Biofuels* 9, 116.
- Tian, L., Perot, S.J., Hon, S., Zhou, J., Liang, X., Bouvier, J.T., Guss, A.M., Olson, D.G., Lynd, L.R., 2017. Enhanced ethanol formation by *Clostridium thermocellum* via pyruvate decarboxylase. *Microb. Cell Factories* 16, 171.
- Trotov, M., Abagyan, R., 2008. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* 18, 178–184.
- Tripathi, A., Bankaitis, V.A., 2017. Molecular docking: from lock and key to combination lock. *J. Mol. Med. Clin. Appl.* 2, 1–19.
- Tripathi, S.A., Olson, D.G., Argyros, D.A., Miller, B.B., Barrett, T.F., Murphy, D.M., McCool, J.D., Warner, A.K., Rajgarhia, V.B., Lynd, L.R., Hogsett, D.A., Caiazza, N.C., 2010. Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a pta mutant. *Appl. Environ. Microbiol.* 76, 6591–6599.
- Trott, O., Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461.
- Wang, L., Upadhyay, V., Maranas, C.D., 2021. dGPredictor: Automated Fragmentation Method for Metabolic Reaction Free Energy. *bioRxiv*.
- Wen, Z., Li, Q., Liu, J., Jin, M., Yang, S., 2019. Consolidated bioprocessing for butanol production of cellulolytic Clostridia: development and optimization. *Microb. Biotechnol.*
- Wu, Q., Peng, Z., Zhang, Y., Yang, J., 2018. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.* 46, W438–W442.
- Xiong, W., Lo, J., Chou, K.J., Chao, W., Magnusson, L., Dong, T., Maness, P., 2018a. Isotope-Assisted metabolite analysis sheds light of central carbon metabolism of a model cellulolytic bacterium *Clostridium thermocellum*. *Front. Microbiol.* 9, 1–11.
- Xiong, W., Reyes, L.H., Michener, W.E., Maness, P.-C., Chou, K.J., 2018b. Engineering cellulolytic bacterium *Clostridium thermocellum* to co-ferment cellulose- and hemicellulose-derived sugars simultaneously. *Biotechnol. Bioeng.* 115, 1755–1763.
- Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D., 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. Unit. States Am.* 117, 1496–1503.
- Yang, L., Yurkovich, J.T., King, Z.A., Palsson, B.O., 2018. Modeling the multi-scale mechanisms of macromolecular resource allocation. *Curr. Opin. Microbiol.* 45, 8–15.
- Ye, X., Zhang, C., Zhang, Y.H., 2012. Engineering a large protein by combined rational and random approaches: stabilizing the *Clostridium thermocellum* cellobiose phosphorylase. *Mol. Biosyst.* 8, 1815–1823.
- Young, J.D., 2014. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* 30, 1333–1335.
- Zhang, P., Wang, B., Xiao, Q., Wu, S., 2015. A kinetics modeling study on the inhibition of glucose on cellulose of *Clostridium thermocellum*. *Bioresour. Technol.* 190, 36–43.
- Zhang, Y.H.P., Lynd, L.R., 2004. Kinetics and relative importance of phosphorolytic and hydrolytic cleavage of celldextrins and cellobiose in cell extracts of *Clostridium thermocellum*. *Appl. Environ. Microbiol.* 70, 1563–1569.
- Zheng, T., Olson, D.G., Tian, L., Bomble, Y.J., Himmel, M.E., Lo, J., Hon, S., Shaw, A.J., van Dijken, J.P., Lynd, L.R., 2015. Cofactor specificity of the bifunctional alcohol and aldehyde dehydrogenase (Adhe) in wild-type and mutant *Clostridium thermocellum* and thermoanaerobacterium saccharolyticum. *J. Bacteriol.* 197, 2610–2619.
- Zhou, J., Olson, D.G., Argyros, D.A., Deng, Y., van Gulik, W.M., van Dijken, J.P., Lynd, L.R., 2013. Atypical glycolysis in *Clostridium thermocellum*. *Appl. Environ. Microbiol.* 79, 3000–3008.