NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF CHEMICAL ENGINEERING
DEPARTMENT OF PROCESS ANALYSIS AND PLANT DESIGN

A systems approach on the integration of metabolic engineering and processes engineering: the case of kerosene-producing *Saccharomyces cerevisiae*

Thesis submitted for the degree of

Masters in Chemical Engineering

by

Ioannis P. Ntekas

Supervisors: Professor Antonis Kokossis (NTUA)
Professor Vassily Hatzimanikatis (EPFL)

Athens, September 2019

Ευχαριστίες

Μέρος της παρούσας διπλωματικής εργασίας εκπονήθηκε στο εργαστήριο του καθηγητή Βασίλη Χατζημανικάτη στο EPFL, τον οποίον ευχαριστώ θερμά για τη στήριξη και την ευκαιρία που μου έδωσε να αναμετρηθώ με το αντικείμενο της μεταβολικής μηχανικής.

Οφείλω ένα μεγάλο ευχαριστώ στον επιβλέποντα καθηγητή μου Αντώνη Κοκόση για την καθοδήγηση και την ερευνητική ελευθερία που μου παρείχε. Η παρούσα εργασία αποτέλεσε μια σημαντική πρόκληση για εμένα που επηρέασε σαφέστατα την ερευνητική μου ματιά.

Θα ήθελα επίσης να ευχαριστήσω την υποψήφια διδάκτορα του EPFL Τσούκα Σοφία για τη σημαντική βοήθεια στο κομμάτι της μεταβολικής μοντελοποίησης και την παροχή του αρχικού μεταβολικού μοντέλου του *S.cerevisiae*, καθώς και τον υποψήφιο διδάκτορα του ΕΜΠ Κώστα Πυργάκη για τη συνεχή υποστήριξή του στην κατασκευή του μοντέλου υπερδομών.

Το παρόν αποτελεί το τέλος της προπτυχιακής μου πορείας στη σχολή Χημικών Μηχανικών του ΕΜΠ. Οφείλω λοιπόν ένα μεγάλο ευχαριστώ στους γονείς μου και τους φίλους μου για την αμέριστη στήριξη και κατανόηση οπουδήποτε και αν με έβγαλαν τα ταξίδια μου όλα αυτά τα χρόνια. Οτιδήποτε μικρό ή μεγάλο πετυχαίνω θα ήταν αδύνατο χωρίς εσάς.

Γ.Ν

Abstract

The microbial production of fuels and industrial chemicals has been identified as a promising alternative to address the depletion of fossil resources and the climate change, which is tightly correlated to anthropogenic activities. The development of efficient cell-factories requires systematic metabolic engineering of microbial strains to rewire the metabolic network towards the desired behavior. Although minimum separation costs are key determinants of a novel bioprocess viability, downstream process considerations are seldom accounted during the microbial strain design procedure. In this work, an efficient computational strain design workflow is proposed to identify metabolic interventions that succeed high product revenues while demanding minimum separation expenses. The systematic workflow comprises of five modules: In the first module, the Genome-scale Metabolic reconstruction (GEM) of a selected host organism is edited to include metabolic pathways towards a selected product portfolio and economic variables related to the upstream process and the potential product revenue. In the second module, a Mixed-Integer Linear Program (MILP) formulation is addressed to identify alternative sets of reaction eliminations that result in maximum revenue. In the third module, we sample the GEM allowed solution space that correspond to the alternative metabolic strategies and estimate the product stream composition. In the fourth module, based on the exit stream compositions we identify the optimal separation flowsheet and minimum cost for product recovery by solving the corresponding superstructure optimization problem. Finally, the average separation cost and product revenue are used to identify the most promising metabolic strategies.

As a case study, we applied our workflow to rationally design a kerosene producing *S.cerevisiae* strain for minimum downstream separation cost. To this direction, *S.cerevisiae* iMM904 GEM was adapted to include hydrocarbons' producing heterologous pathways. The developed strain design framework was applied to create a pool of alternative metabolic strategies that yield in maximum revenue. Assuming aerobic cell culture conditions in a chemostat array with glucose as the sole carbon source, the models that correspond to the distinct strategies were sampled to estimate the exit stream composition and a distillation supertask problem was solved to identify the minimum separation cost. The applied methodology identified metabolic strategies up to 7-fold more efficient with respect to the initial strain.

The present formulation is the first to our knowledge that aims to bridge the strain design procedure with the downstream process synthesis, paving the way towards microbial strains tailor-made for sustainable biorefinery applications.

Keywords: *S.cerevisiae*; microbial biorefinery; kerosene; genome-scale model; superstructure; distillation supertask; strain design algorithm; metabolic engineering; mixed-integer linear programming

Εκτεταμένη περίληψη

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια της ακαδημαϊκής συνεργασίας μεταξύ της σχολής Χημικών Μηχανικών Ε.Μ.Π και του πανεπιστημίου EPFL, υπό την επίβλεψη των καθηγητών Αντώνη Κοκόση και Βασίλη Χατζημανικάτη. Η εργασία πραγματεύεται το σχεδιασμό ενός υπολογιστικού πλαισίου για το σχεδιασμό μικροβιακών στελεχών με την ικανότητα παραγωγής χημικών και βιοκαυσίμων με έμφαση στο κόστος διαχωρισμού. Το αναπτυχθέν πλαίσιο εφαρμόστηκε επιπλέον για την μελέτη και *in silico βελτιστοποίηση* στελεχών του μύκητα *S.cerevisiae* με την ικανότητα να παράγουν βιοκαύσιμα ανάλογα της κηροζίνης.

Η σύγχρονη κοινωνία είναι άρρηκτα συνδεδεμένη με τη χρήση των ορυκτών καυσίμων τόσο σαν πηγή ενέργειας αλλά και σαν πρώτη ύλη για την παραγωγή προϊόντων. Οι ορατές συνέπειες της συντελούμενης κλιματικής αλλαγής, συνδεδεμένης με την χρήση των ορυκτών καυσίμων, σε συνδυασμό με την εξάντληση των αποθεμάτων πετρελαίου, έχουν οδηγήσει τη διεθνή κοινότητα να αναζητήσει εναλλακτικές.

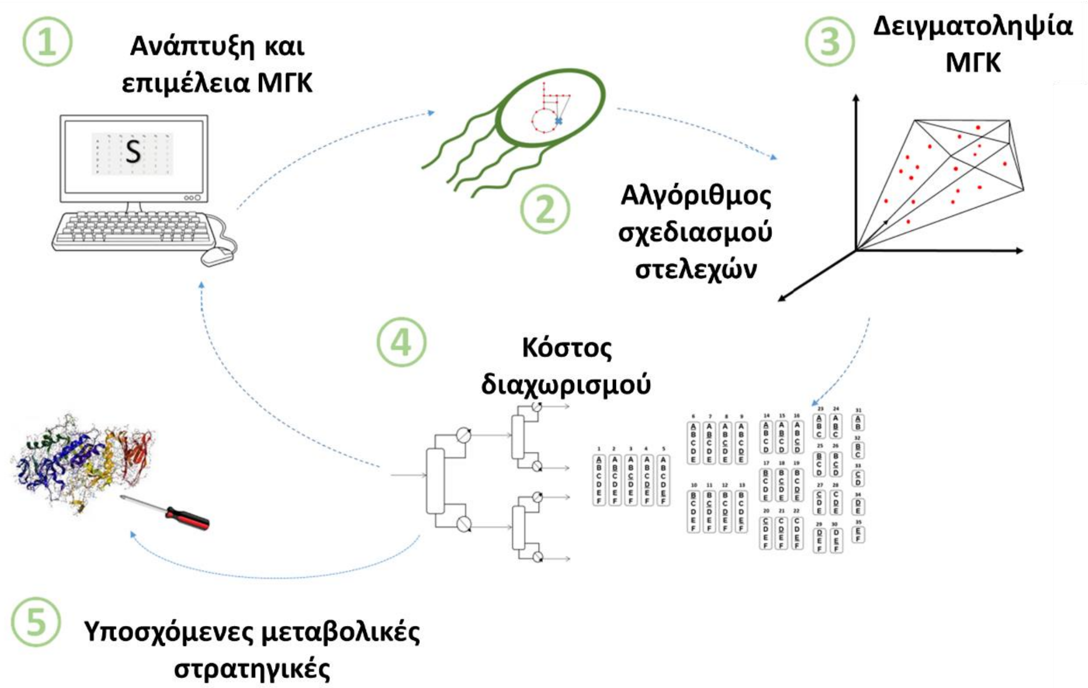
Σε αυτή την κατεύθυνση, η χρήση γενετικά τροποποιημένων μικροοργανισμών για την παραγωγή βιοκαυσίμων και χημικών αποτελεί μια υποσχόμενη βιώσιμη εναλλακτική. Η δημιουργία κατάλληλων στελεχών ικανών να παράγουν χρήσιμα χημικά συχνά προϋποθέτει την επιβολή αλλαγών στον κυτταρικό μεταβολισμό. Επειδή οι αλλαγές αυτές είναι μη προφανείς η σχεδιαστική διαδικασία συχνά υποβοηθάται από τη χρήση Μεταβολικών μοντέλων Γονιδιακής Κλίμακας (ΜΓΚ) που περιέχουν όλη τη διαθέσιμη πληροφορία σχετικά με τις μεταβολικές δυνατότητες ενός οργανισμού. Τα ΜΓΚ αποτελούν στοιχειομετρικές αναπαραστάσεις του συνόλου του μεταβολικού δικτύου υπό τη μορφή συστήματος γραμμικών εξισώσεων και περιορισμών. Η εφαρμογή υπολογιστικών πλαισίων βελτιστοποίησης στα ΜΓΚ μπορεί να αναδείξει τις απαραίτητες γενετικές τροποποιήσεις που θα προσδώσουν στον οργανισμό ένα επιθυμητό χαρακτηριστικό όπως η υπερπαραγωγή ενός χρήσιμου χημικού ή η σύζευξη της παραγωγής με την κυτταρική ανάπτυξη.

Ο διαχωρισμός των προϊόντων της ζύμωσης από το ρεύμα εξόδου του βιοαντιδραστήρα αποτελεί μια από τις βασικές πηγές κόστους στις βιοδιεργασίες διαδραματίζοντας έτσι καθοριστικό ρόλο για τη βιωσιμότητα μιας πιθανής εφαρμογής. Η σύσταση του ρεύματος εξόδου έχει άμεση σχέση με το κόστος διαχωρισμού. Οι διαφορετικές στρατηγικές μεταβολικής μηχανικής που αναδεικνύονται από τη χρήση των αλγορίθμων σχεδιασμού καθορίζουν το σύνολο των ιδιοτήτων του μεταλλαγμένου μικροοργανισμού συνεπώς και τη σύσταση του ρεύματος εξόδου.

Στην παρούσα εργασία, αναπτύσσεται μια καινοτόμος ροή εργασίας με σκοπό την ορθολογική σχεδίαση μικροβιακών στελεχών που αποφέρουν τα μέγιστα έσοδα διατηρώντας παράλληλα χαμηλό κόστος διαχωρισμού. Η προτεινόμενη ροή εργασίας απαρτίζεται από πέντε στάδια. Το πρώτο στάδιο απαρτίζεται από την ανάπτυξη και

επεξεργασία του ΜΓΚ προσθέτοντας μεταβολικά μονοπάτια για την παραγωγή των χρήσιμων χημικών αλλά και μεταβλητές σχετιζόμενες με τη χρηματική αξία των προϊόντων και τα οικονομικά χαρακτηριστικά της ζύμωσης. Στο δεύτερο στάδιο εφαρμόζοντας έναν αλγόριθμο για το σχεδιασμό μικροβιακών στελεχών δημιουργείται μια δεξαμενή μεταβολικών στρατηγικών που οδηγούν σε μεγιστοποίηση των εσόδων. Στο τρίτο στάδιο, εφαρμόζοντας τις μεταβολικές στρατηγικές του τρίτου σταδίου με τη μορφή περιορισμών στο ΜΓΚ, πραγματοποιείται δειγματοληψία στον επιτρεπτό χώρο λύσεων του γραμμικού συστήματος και προσδιορίζεται έτσι η σύσταση του ρεύματος εξόδου. Τα δεδομένα σύστασης του ρεύματος εξόδου του βιοαντιδραστήρα που προκύπτουν για κάθε διαφορετική μεταβολική στρατηγική λειτουργούν σαν είσοδος για το τέταρτο βήμα οπού προσδιορίζεται το ελάχιστο κόστος διαχωρισμού μέσω μιας υπερδομής. Τέλος στο τελευταίο βήμα, οι διαφορετικές στρατηγικές αξιολογούνται βάσει του υπολογιζόμενου μέσου κόστους διαχωρισμού και των μέσων εσόδων.

Συγκεκριμένα για την περίπτωση του *S.cerevisiae* που παράγει κηροζίνη, επεξεργαστήκαμε το ΜΓΚ iMM904 προσθέτοντας μεταβολικά μονοπάτια παραγωγής υδρογονανθράκων και οικονομικούς παράγοντες που συσχετίζονται με τις τιμές των παραγόμενων προϊόντων αλλά και με τη ζύμωση. Η ζύμωση θεωρήσαμε ότι λαμβάνει χώρα σε συστοιχία χημειοστατών σε αερόβιες συνθήκες με σταθερή παροχή γλυκόζης. Χρησιμοποιώντας τον αλγόριθμο σχεδιασμού στελεχών δημιουργήσαμε μια σειρά εναλλακτικών μεταβολικών στρατηγικών που καταλήγουν σε μέγιστο κέρδος. Πραγματοποιώντας δειγματοληψία στον επιτρεπτό χώρο λύσεων των διαφορετικών μοντέλων, προσδιορίσαμε την αντίστοιχη σύσταση του ρεύματος εξόδου της συστοιχίας. Τέλος, υπολογίσαμε το ελάχιστο κόστος διαχωρισμού για το εκάστοτε ρεύμα λύνοντας ένα πρόβλημα υπερδομής αποστακτικών στηλών. Η εφαρμογή της ροής εργασιών κατάφερε να αναδείξει μεταβολικές στρατηγικές που καταλήγουν σε επταπλασιασμό του κέρδους σε σχέση με το αρχικό στέλεχος του μύκητα. Οι μεταβολικές στρατηγικές μπορούν να μεταφραστούν σε εργαστηριακές πρακτικές γενετικής τροποποίησης όπως προσθήκη, διαγραφή, υπερέκφραση ή υποέκφραση γονιδίων και να θέσουν στόχους για την τροποποίηση ενζύμων.


1
Ανάπτυξη και επιμέλεια ΜΓΚ
S
2
Αλγόριθμος σχεδιασμού στελεχών
3
Δειγματοληψία ΜΓΚ
4
Κόστος διαχωρισμού
5
Υποσχόμενες μεταβολικές στρατηγικές

Contents

List of Figures

List of Tables

Chapter 1. Introduction

Biotechnology has served humankind for more than 8.000 years. The first biotechnological applications include the use of microorganisms for the production of fermented foods and beverages such as cheese, yoghurt, beer and wine, which remain at the core of our societies' culinary culture, working as a pillar for the modern food industry. In modern history, the two World Wars have played a crucial role in the acceleration of the industrialization of biology. The acetone-butanol-ethanol fermentation process, which is still in use, was developed at the time of WWI while WWII signalled the industrial scale production of penicillin[1].

In the early 1990s, the advancing field of genetic engineering enabled several more biotechnological success stories with the majority of them being pharmaceuticals-related (recombinant proteins and antibodies). Later the introduction of mathematical modeling and bioinformatics to the study of cellular behavior helped to kick start the science of metabolic engineering –suggesting ways to rationally redirect the cellular metabolism in order to produce a plethora of different chemicals that far exceed the spectrum of food industry. The emerging field of industrial biotechnology has worth over 300 billion USD in revenue and is expected to duplicate by 2025[2].

Apart from the apparent economic potential, the shift to a bio-based economy and the subsequent intensification of industrial biotechnology applications is necessary to tackle the unprecedented challenges that humankind is facing in the modern era. The climate crisis, the depletion of natural resources and the increasing food demands are driving up the inequality gap and bring the planet to its limits. Biotechnology rises as an indispensable tool to address these issues and safeguard a sustainable future for all. Biotechnology is the key element for transforming renewable feedstock to desired chemicals thus dethatching economy from fossil fuels and alleviating the impact of climate crisis. In agriculture, biotechnology advances can yield in more productive, resistant crops that can support the human population. In healthcare, biotech drugs, vaccines and diagnostics are improving health and the quality of life[3,4].

Bio-processes and especially the concept of bio-refinery, where biomass feedstock is transformed to useful biofuels, platform and specialty chemicals and novel products, are described as the most promising routes to establish a sustainable bio-economy[5]. Microbial bio-refineries, which utilize microorganisms and microbial consortia fermentations as the biomass transformation technology present great interest. Bioprocesses demand milder conditions compared to traditional catalytic transformations while the cellular enzymatic toolbox can conduct very specific transformations towards desired chemicals, which are hard or impossible to obtain conventionally. Nevertheless, since these processes are seldom competitive to their petrochemical counterparts, current efforts focus on strain efficacy enhancement with respect to the titer, productivity and yield of the desired product[6]. The development of industrial strains with selected characteristics that can support the commercialization of a bio refinery application is conducted with iterative trial cycles where metabolic interventions are systematically identified and applied to the host organism[6–8]. Because the necessary interventions are not obvious, the process is computer assisted and typically, Mixed Integer Linear Programming (MILP) algorithms are utilized to build the interventions' strategies[9–11].

A parallel endeavor towards commercialization of bio-refinery applications lies on the process synthesis. The upstream and downstream processes, with respect to the fermentation(s), contribute to the total expenditure via the capital and operating cost. Especially the downstream process aiming to products' purification is identified as a major total cost contributor, accounting for up to 80% of the total annual expenditure[12]. The synthesis problem for the downstream process aims to identify the optimum, in terms of cost, selection and sequence of available technologies to separate the vaporizable products from a fermentation broth of known composition. Although, the metabolic network interventions alter the fermentation broths content the microbial development scarcely accounts for downstream insights and the two problems are addressed independently without communicating with each other.

The present thesis aims to address the systems challenge to develop a systematic workflow that will assist to connect metabolic modeling and design strategies with the downstream process synthesis problem. The thesis comprises of six chapters. In the second chapter, we provide the definitions and the necessary background information concerning the key points of the thesis and briefly discuss the state of the art in strain design and downstream process synthesis. In the third and fourth chapter, we present the suggested workflow used to connect the strain design/phenotype prediction problem with the downstream process synthesis problem and discuss the methods used in the individual workflow modules. Furthermore, we specify the methodology for a selected case study concerning the heterologous production of hydrocarbons from *S.cerevisiae* for minimum separation cost. In the fifth chapter, we present and discuss the results for the specific case study and finally in chapter six we discuss our findings, evaluate whether the proposed workflow could assist nowadays strain design procedure and set some future prospects on metabolic and process integration.

Chapter 2. Background and State-of-the-art

2.1 Cellular metabolism

Living cells consist of a large number of compounds and metabolites. While water is the most predominant compound accounting for approximately 70% of the cellular mass, the rest is distributed among complex building blocks such as nucleic acids (DNA and RNA), lipids, proteins and carbohydrates. Synthesis and organization of these macromolecules to form a functioning cell is succeeded by numerous independent reactions. Cellular metabolism is the network of enzyme-catalyzed reactions where nutrients are broken down to form the biomass building blocks and provide the necessary energy to sustain life. Metabolism may be categorized in catabolism and anabolism. Catabolic reactions break down the energy sources to simpler intermediate molecules while anabolic reactions build up the cellular components. Catabolic reactions are typically exergonic while anabolic reactions are endergonic[13,14].

The cellular metabolic network is organized in a structure similar to a 'bow-tie', in the sense that a broad range of substrates are broken down to form a much smaller number of intermediates, which are then channeled towards a large number of distinct biomolecules. Metabolic networks can be divided to smaller subunits called metabolic pathways. Metabolic pathways are series of connected reactions that convert one metabolite to another via anabolic or catabolic ways. The pathways are interconnected by the flow of material and energy; metabolites may participate in several reactions by branching, connecting several reaction sequences while the energy-related universal co-factors (ATP, NADH, NADPH), indispensable for many reactions, add another level of integration between pathways. The pathways are further organized in subnetworks, responsible for a specific cellular function. While there are important structural differences between the metabolic networks of different species, several 'core' subnetworks remain conserved across all organisms.

2.2 Metabolic Engineering

Metabolic engineering is the ad hoc manipulation of the cellular metabolic, regulatory and transport processes, using primarily genetic approaches, in order to enhance the production of desired chemicals[15]. Unlike the early-day approaches applied for chemical overproduction that focused on single-reactions or random mutations, metabolic engineering examines the full network properties to identify potential bottlenecks and develop directed alterations to shift the cellular behavior towards the engineering objective. The imposed genetic manipulations include recombinant DNA techniques and genome-editing techniques such as CRISPR-Cas9 that typically result in gene insertions, deletions upregulations and downregulations[9].

The emerging high-throughput techniques available to decipher omic data (genomes, transcriptomes, proteomes, metabolomes, fluxomes) and the advancements in computational biology have paved the way for systems metabolic engineering. In this systems-level metabolic engineering approach, the omic data and computational techniques broadly used in systems biology together with the synthetic biology design approaches are integrated to the traditional metabolic engineering action framework, allowing better understanding of the cellular functions and engineering capabilities[8].

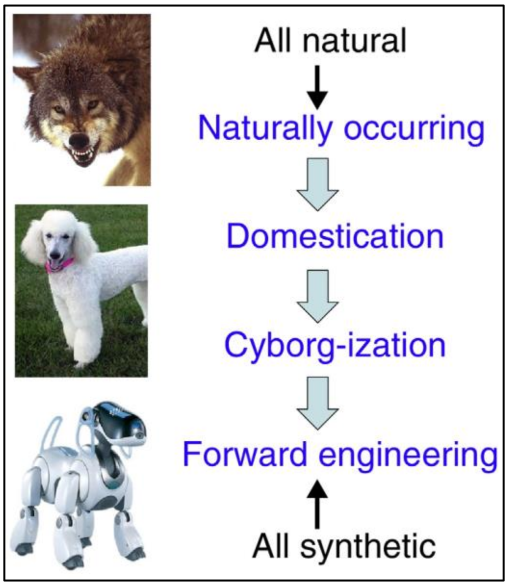


Figure 1: Synthetic biology paves the way towards full predictability of bioengineered systems. The rational strain design resembles the historical wild animals' domestication, to increase productivity and enable the desired response to human instructions[16].

2.2.1 Towards a bio-based economy

The modern society is built upon an on growing dependency towards fossil fuels use. From transport and commodity chemicals to value-added chemicals, fossil fuels serve as a main energy and feedstock provider; in the year 2016, 33% of the global energy consumption and approximately 80% of the liquid transportation fuels were petroleum-derived[17]. The increasing concerns regarding depletion of petroleum resources and the climate change-directly linked to human activity-related Green House Gases (GHGs) emissions have driven the research community to seek for alternative energy options, such as the microbial biosynthesis of advanced biofuels[18]. To this direction, nature's diverse toolbox has been exploited to design novel processes to produce chemicals' building blocks and final products, starting from renewable non-food biomass or even CO_2, leading to neutral or ''negative'' carbon emissions respectively[17,19,20]. The target product can be a natural biological chemical (ethanol, amino acids, etc.) or a molecule that does not normally occur in nature (polylactic acid, 5-methyl-1-heptanol, etc.)[9,8]. Available biochemical reaction databases such as KEGG and Metacyc or retrobiosynthetic tools such as the BNICE.ch framework can be utilized to identify production pathways that meet the specification needs for the selected host organism[21].

2.2.2 Renewable biofuels

Biofuels are arguably the most likely near term renewable alternative to petroleum fuels, with some forms of transportation that cannot be easily electrified (such as long distance trucking, shipping and aviation) having this approach as the only alternative[22].Bioethanol is by far the most produced biofuel, in 2017, the worldwide bioethanol production was more than 95 billion liters with the USA market alone occupying the 2/3 of the total production[23]. Despite the high production titers, bioethanol does not correspond to an ideal candidate for conventional fuel substitution. Direct use of ethanol as a fuel requires engine modifications while its energy content is lower than that of fossil derived gasoline. To this end, an ideal substitute biofuel would enclose precise chemical replacements of the fossil-fuel counterparts, thus being able to be utilized as drop-in fuel in the existing infrastructures without prior engine modifications. This special category of biofuels is referred to as advanced biofuels[24]. Bio-derived hydrocarbons, such as alkanes and alkenes, due to their chemical relativity to conventional fuels and high energy density, consist an ideal renewable biofuel target. Based on the chain length distribution different bio-derived hydrocarbon blends can be used as a replacement to the according fuel type.

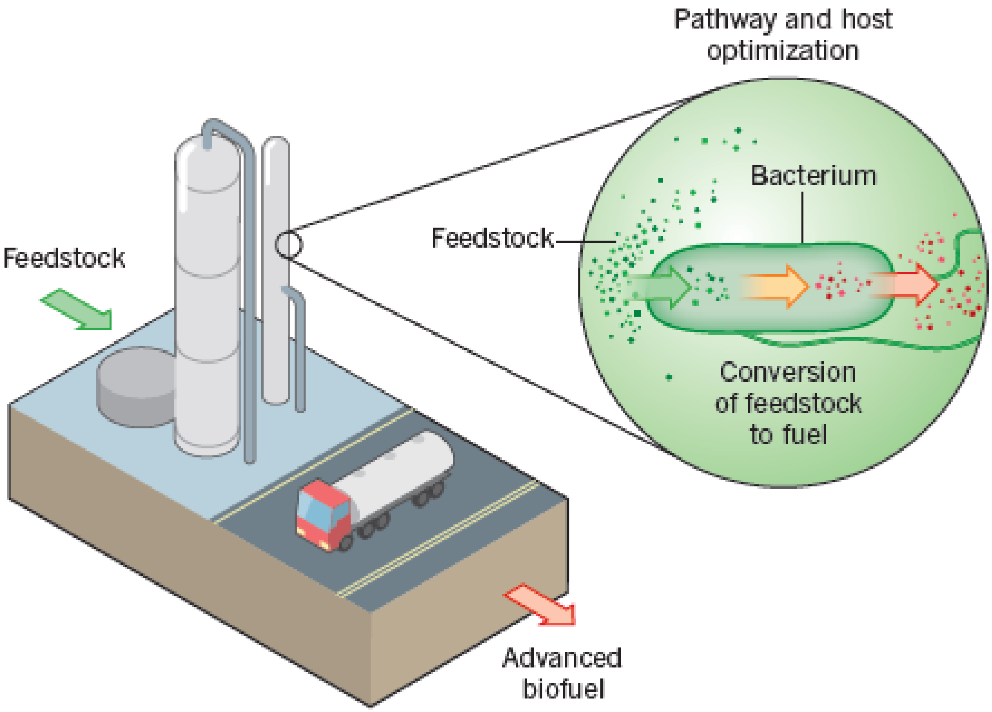

Pathway and host optimization
Bacterium
Feedstock
Feedstock
Conversion of feedstock to fuel
Advanced biofuel

Figure 2: Engineered microorganisms can work as fuel cell-factories to convert Feedstock to advanced biofuels and chemicals[25].

2.2.3 *S.cerevisiae* as a cell-factory

Since *S.cerevsiae* has been broadly used for beer and wine fermentations, its selection as industrial ethanol producer is not surprising. The ethanol fermentation, nowadays, consists a robust, well-studied industrial application, making yeast one of the most preferred host organisms for the production of diverse fuels and chemicals. Moreover, holding a GRAS status by FDA, makes yeast suitable for the production of food-grade products[23]. Over the years, yeast metabolic capabilities have been exploited for the production of various products such as pharmaceuticals (artemisinic acid, human albumin etc.), fuels (alcohols, alkanes, etc.) and platform chemicals (succinic acid, coumaric acid, etc.) and specialties (santalene, valencene, etc.).

2.2.3.1 Yeast hydrocarbon production

Although many microorganisms are able to naturally synthesize alkanes and alkenes as a response against environmental threats, the production levels and the properties of the compounds are not suitable for direct use as drop-in biofuels. For that reason, heterologous hosts are exploited to express several pathways involved in alkanes and alkenes metabolism[26].

Yeast constitute an ideal host candidate for alkanes and alkenes production of the range of kerosene. The accumulated existing knowledge over yeast fermentations facilitates the scale up process. S.cerevisiae exhibits pH-tolerance and has been proven robust in prior applications. Furthermore, since yeast does not naturally produce this class of products the addition of heterologous hydrocarbons producing pathways will evoke minimum cross-talking with the native subsystems[27].

Long-chain alkanes and alkenes production via existing heterologous pathways typically present free fatty acids or free fatty acid-related molecules as a starting point, intersecting in that way with the host's native lipids metabolism[28]. Existing efforts include the heterologous expression of two enzymes from *S. elongatus* the fatty acyl-acid reductase (FAR) that converts fatty acyl-CoA to fatty aldehydes, and the fatty aldehyde deformylating oxygenase (FADO)[29]. The enzymes addition complemented with the *hfd1* gene deletion resulted in a final titer of 22 μg/gDCW. Alternatively to the FAR enzyme, a *Mycobacterium marinum* Carboxylic acid reductase (CAR) has been showcased to 2.7-fold improve the final titers. The main by-product to alkanes and alkenes are the free fatty alcohols, as a result of the native aldehyde reductases activity[30]. The low attainable yields in comparison to the theoretical estimates are partly attributed to the tightly regulated nature of lipids metabolism in yeast. Further metabolic modifications applied towards yield enhancement include the elimination of competing pathways, the increase of co-factor supply by upstream gene additions and the production pathway compartmentalisation[31].

Recent advances include the engineering of Yeast capable of secreting 1-alkenes, minimizing potential stress due to toxicity[32]. Since these long chain hydrocarbons are immiscible, the establishment of an industrial process seems promising; the formation of an organic phase containing the product blend is expected to further facilitate the downstream harvesting process.

The current efforts in order to develop viable microbial biorefineries focus on:

1) lowering the costs related with the suitable microbial host development

2) the establishment of efficient biomass-to-sugars hydrolysis pathways[33].

It is worth mentioning that in many biorefinery applications, the downstream processes account for over than 50% of the total production cost[34]. A sustainable biorefinery shall depend on upstream and downstream processes optimised to meet the product's characteristics and the host's physiological specifications.

2.3 Constraint-based modelling and analysis of metabolism

The emerging high throughput technologies for studying various biological processes and functions at the gene, protein, and metabolite levels, yield in the generation of large amounts of information[35]. Mathematical modeling of metabolism is an invaluable tool to assess and predict the cellular behavior under different environments and genetic backgrounds. The existing approaches to model metabolism can be roughly devided in two groups of approaches: the kinetic modeling and the stoichiometric modeling. In kinetic modeling, different types of mechanistic expressions, such as Michaelis Menten kinetics are utilized to describe reaction rates. The total of the rate expressions constitute a system of ordinary differential equations representing the conservation of mass for each metabolite. Solution of the system results in a time-dependent metabolite concentrations and reaction flux profile. In contrast to kinetic modeling, stoichiometric approaches rely primarily on stoichiometric equations to form a system of linear equations that describe metabolites' mass conservation under steady-state[10].

2.3.1 Genome-Scale Metabolic reconstructions (GEMs)

Stoichiometric models have been in use to study the physiology of organisms since 1980s. The knowledge accumulation alongside with the progress in the field of genome annotation led to the creation of Genome Scale Metabolic Reconstructions. GEMs contain links between the occurring reactions and the genes encoding the according enzymes in the form of gene to protein to reaction associations (GPRs). In that way GEMs enclose all known biochemistry taking place inside a specific organism[36]. GEMs are an invaluable tool in systems biotechnology applications such as the construction of metabolic strategies for metabolite overproduction, identification and design of drugs, as well as the study of cellular phenotypes under alternative nutrients or the impact of gene knockouts.

2.3.2 Constraint-based methods

Constraint-based modelling is a broadly used approach to study metabolism. The metabolic network stoichiometric information is encoded in a stoichiometric matrix (S matrix). Each row of the matrix represent a metabolite and each column a reaction. The elements of the matrix are the stoichiometric coefficients of each metabolite in the according reaction. S matrix constitutes a core element of GEMs[37].

2.3.2.1 Flux balance Analysis

Flux Balance Analysis (FBA) is key concept in constraint-based methods and the basis for the construction of numerous other analysis methods. In FBA, constraints are imposed in two ways. The first lies to the FBA assumption that the system is in a pseudo-steady state,meaning that metabolites concentrations inside the cell do not change over time, thus there is no metabolite accumulation. Mathematicaly this is described by the equation:

$$S \cdot v = 0$$

Where S is the stoichiometric matrix and *v* is the vector containing all the reaction fluxes in the metabolic network. In that way S imposes flux balane constraints on the system, ensuring that the total amount of each compound eing produced is equal to the amount consumed. The second type of imposed constraints is the upper and lower bounds given to each reaction flux, typically including laboratory flux measurements (metabolite uptake and secretion rates).

The two types of constraints (stoichiometry related linear equations and bound inequalities) define an allowable solution space. The network may acquire any flux distribution lying inside the solution space.

The aim of flux balance analysis is to find a flux distribution inside the allowable solution space that maximizes or minimizes a specified objective function (Z). A typical task when handling GEMs is the growth prediction under different circumstances. In that case, the objective is the maximization of biomass which is acounted in the stoichiometric matrix as an extra column (biomass reaction) including precursors in stoichiometries simulating biomass production. The stoichiometries are scaled in a manner such that the flux through the biomass reaction is equal to the growth rate (μ).

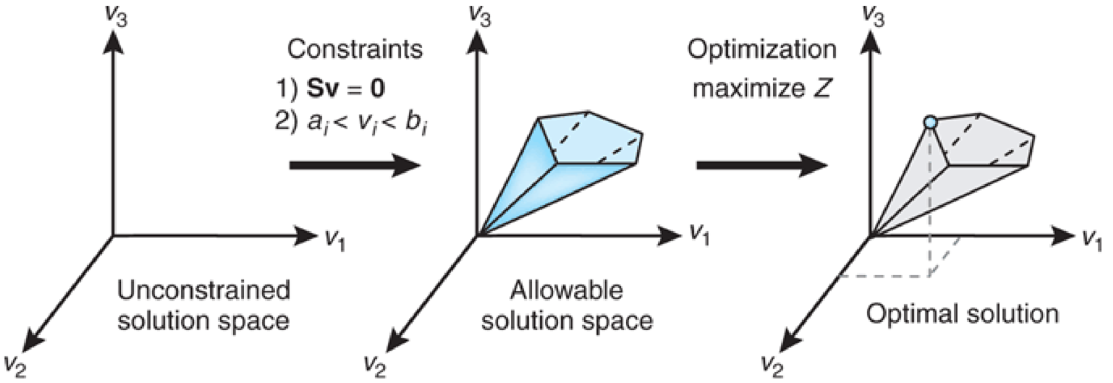


Figure 3: A conceptual basis representation of constraint-based modeling. Without constraints, the metabolic flux distribution may lie at any point in a solution space. When the mass balance constraints (S matrix) and capacity constraints (upper and lower bounds) are imposed, an allowable solution space is defined. The network may acquire any flux distribution within this defined space, while points outside it are denied by the constraints. Through optimization of specific objective functions, FBA identifies a single optimal flux distribution that lies on the edge of the solution space polytope.

2.3.2.2 Thermodynamics-based Flux Analysis

Solutions obtained with FBA are non-unique and sometimes unreliable because they may violate thermodynamic constraints. In order to further constrain the allowed solution space and obtain thermodynamic feasible flux distributions, Henry et. al. proposed the thermodynamics-based flux analysis (TFA) workflow in which extra constraints are added in order to couple reaction directionalities to thermodynamics constraints. In the formulation, metabolite concentrations and Gibbs energy of reactions are added to the model while the Gibbs energy of reaction sign is coupled to the reaction directionality[38].

Table 1: The constraints applied in FBA and TFA formulations[38].

FBA constraints	Mass balance	$S.v = 0$
	Flux capacity	$\underline{v} \leq v \leq \overline{v}$
TFA constraints	Gibbs energy of reaction	$\Delta_r G_i' = \Delta_{r,tpt} G_i' + \sum_{j=1}^{m} n_{i,j}\mu_j$
	Chemical potential	$\mu_j = \Delta_f G_j'^0 + \Delta_{f,err} G_j'^0 + RT \ln x_j$
	Thermodynamic feasibility	$\Delta_r G_i' - K + K * z_i < 0$
	Coupling constraint	$v_i - K * z_i < 0$

Where:

$\Delta_r G_i'$ is the transformed Gibbs free energy of the reaction i

μ_j are the chemical potentials of the reactants j

$\Delta_{r,tpt} G_i'$ is the Gibbs free energy of transport (accouted when the reaction is transport of a compound from one compartment to another

$\Delta_f G_j'^0$ is the standard transformed Gibbs free energy of formation of the compounds

$\Delta_{f,err} G_j'^0$ is the estimated error in the energy of formation

R is the universal gas constant

T is the temperature (here assumed 298 K)

x_j is the molar fraction of the compound j

K is a large (Big-M, $K > \max \Delta_\mathrm{r} G_i'$) value

and z is a binary decision variable

The formulation requires net fluxes to be non-negative. To this direction, each reaction is separated in two: a net forward and a net backward, while the net fluxes are associated such that:

$$v_{net} = v_{forward} - v_{backward}$$

By applying aditional constraints we ensure that at most one of the two reactions are active at a time.

2.4 Computer-aided strain design

The strain design engineering procedure aims to construct microbial strains that overproduce the desired compound. This objective often contradicts to microbial metabolism, which has evolved to favor fast growth. For that reason, the product yield of wild type strains is often much lower than the theoretical maximum yield[10]. The metabolic interventions needed to redirect the metabolic flow towards the desired compound are often multiple and non-intuitive. As an answer to this problem several computer-aided strain design algorithms have emerged, aiming to identify the necessary set of network perturbations to meet the engineering objective. The majority of these approaches is based on a mathematical (or evolutionary) optimization framework that search over a set of potential genetic interventions (gene insertions, eliminations, upregulations and downregulations). One key outcome of the design procedure is to couple cellular growth with the production of the desired compound. In that way, the cellular growth becomes the driving force behind production. If the new functionality is not coupled to cellular growth, it is very likely that under evolutionary pressure it will be lost[10,39,40].

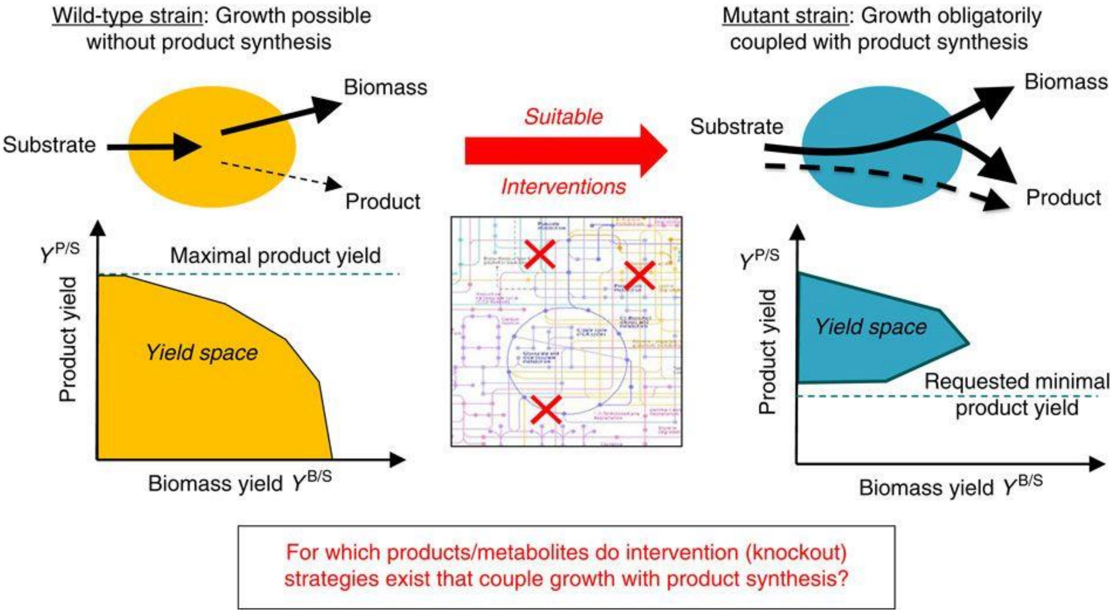

Wild-type strain: Growth possible without product synthesis
Substrate
Biomass
Product
Maximal product yield
$Y^{P/S}$
Product yield
Yield space
Biomass yield $Y^{B/S}$
Suitable
Interventions
Mutant strain: Growth obligatorily coupled with product synthesis
Substrate
Biomass
Product
$Y^{P/S}$
Product yield
Yield space
Requested minimal product yield
Biomass yield $Y^{B/S}$
For which products/metabolites do intervention (knockout) strategies exist that couple growth with product synthesis?

Figure 4: Wild type strains vs. Growth-coupled mutant strains. The desired compound production is not mandatory in the case of the wild type strain. The mutant strain is obliged to produce the desired compound even for zero growth. The chemical in the second case has been transformed to an obligatory biomass formation byproduct[40].

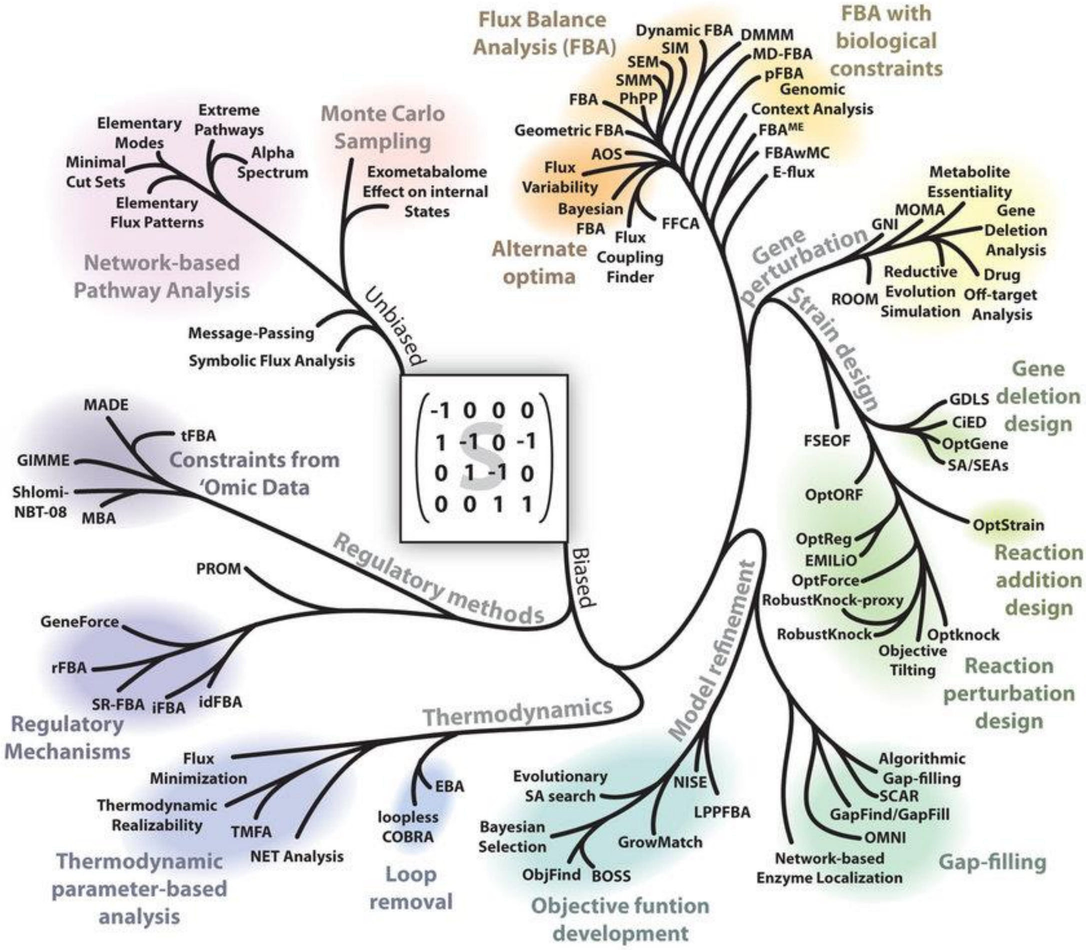


Figure 5: Phylogenetic representation of the alternative constraint-based methods applied to GEMs[41]

2.5 Multicomponent distillation

Distillation is the most common method for the separation of homogeneous mixtures. The process exploits the difference in volatility among the components of a mixture. Repeated vaporization and condensation can lead to virtually complete separation, thus high purity end-products. Besides the high energy costs, distillation is a versatile, robust and well-understood technique[42].

The design of a multicomponent distillation column is based on the decision of the two key components, namely the light (LK) and heavy (HK) component, as well as their recovery in the overhead and bottom product respectively. The separation is aimed to take place between the key components with the light key component kept out of the bottom product and the heavy key component kept out of the top product. The intermediate boiling components will distribute between the products. A number of short-cut methods are available for the design of multicomponent distillation columns. Each addresses different aspects of the column design. Fenske-Underwood-Gilliand

(FUG) is probably the most popular shortcut method applied in distillation column design. The method assumes constant relative volatility alongside the column, which can be approximated as the geometric mean of the relative volatilities at the top and the bottom of the column. The relative volatility of the light key component with respect to the heavy key component will be:

$$(a_{LK/HK})_{avg} = [(a_{LK/HK})_B \cdot (a_{LK/HK})_D]^{1/2}$$

Where B and D denote the distillate and bottom product respectively.

In order to calculate the number of theoretical trays for a distillation column, we first have to calculate the minimum number of trays N_{min} using the Fenske[43] equation and the minimum reflux ratio R_{min} using the Underwood equations[44].

$$N_{min} = \frac{\ln\left[\left({x_{LK}}/{x_{HK}}\right)_D \cdot \left({x_{LK}}/{x_{HK}}\right)_B\right]}{\ln((a_{LK/HK})_{avg})}$$

Where x_{LK} and x_{HK} are the molar fractions of the light and heavy key components respectively.

The minimum Reflux ratio is calculated via the Underwood equations:

i) $\sum_{i=1}^{n} \frac{a_i x_{Fi}}{a_i - \theta} = 1 - \bar{\mathrm{q}}$

ii) $R_{min} + 1 = \sum_{i=1}^{n} \frac{a_i x_{Di}}{a_i - \theta} \quad , 1 < \theta < a_{\mathrm{LK}}$

Where *n* is the number of components, $a_i$is the average geometric volatility for the component I, x_{Fi} and x_{Di} are the molar fractions of the component i at the feed and the distillate respectively, $\bar{\mathrm{q}}$ is the ratio of moles of saturated liquid at the feed stage per feed stream mol. Assuming that the incoming stream is saturated $\bar{\mathrm{q}} = 1$.

The number of theoretical trays can be estimated using Gilliand's correlation for the calculated N_{min} and the selected reflux ratio R.

The separation of a feed mixture in more than two products usually requires multiple distillation columns. The cost-optimal sequence of columns-meaning the one which requires the least annualized investment costs for equipment plus annual utility costs for a given recovery target is not obvious [45]. The different approaches found in the literature addressing the sequencing problem can be classified as approaches that employ heuristics, shortcut methods, mathematical programming methods and methods based on rules and expert systems. While the heuristics are easy to use, they often contradict with each other. The shortcut methods utilize simple models that are seldom reliable[46].

2.6 State of the art in strain design

The typical design process towards a suitable strain for bio-production of fuels and chemicals follows the iterative Design-Build-Test-Learn (DBTL) cycle, commonly involved in engineering practices. The Design involves the choice of the platform-organism and the heterologous pathway to-use, as well as, decisions over strategies able to improve the production characteristics. In the Build module, the DNA parts corresponding to the heterologous pathway are constructed and the platform-organism is transformed. The rest of the metabolic strategies are also imposed to the cell, using primarily gene-editing tools. In the Test module we gather information on the cloning results, strain's efficiency and omics data that help to better comprehend the cellular behavior. The learn module, so far, appears to be the weakest and less described among the rest. It aims to incorporate the attained data, identify the hurdles of the approach and underline alternatives to bypass them.

Nowadays, the process is expensive and laborious; the improvement iterations may last up to eight years while the total cost to develop a commercial strain is on the order of 50 M$. The main focus includes improvement of product titer, yield and production rate (TRY). The design of metabolic strategies to improve the cellular properties can be immensely facilitated by GEMs. Existing algorithms can be used to predict network modifications that lead to the desired properties, such as maximizing yield and coupling the production flux to cellular growth. The latter is important since Adaptive Laboratory Evolution methods (ALE) are commonly used to further improve the strains' characteristics. GEMs can be utilized as well to assist the interpretation of omics data[47]. Constraining a GEM with experimental data obtained in the test module can facilitate the troubleshooting and accelerate the DBTL cycle[33,48].

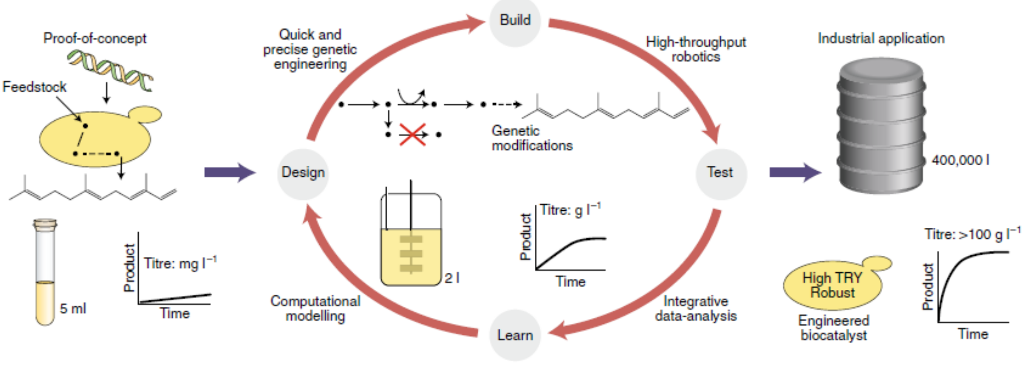


Figure 6: Schematic representation of the DBTL cycle.

The commercial strain can serve as the central biotransformation technology in a microbial biorefinery. Biorefinery applications are highlighted amongst the most promising routes towards the establishment of bio based industry[49]. The usual objective set in biorefinery applications is to optimize the use of resources while minimizing wastes. The biorefinery synthesis and design problem is a relatively complex one, given that multiple feedstock sources can be utilized in different ways and be transformed from different alternative technologies to a multitude of alternative portfolios[50]. Biorefinery products may vary from platform chemicals to specialties.

Bioprocesses' separation typically account for the 60%-80% of the total production cost, directly affecting the project viability. The process synthesis problem is typically handled with one of the following alternative approaches including: enumeration of alternatives, evolutionary modification, and superstructure optimization. The first two approaches are heuristic-based and practically applicable when the number of alternatives is relatively small. Although superstructures are sometimes difficult to develop, they can systematically address the design problem in most cases[12,51].

The commercial strain has been optimized to obtain certain characteristics, namely greater yields and achievable titers. The separation design considerations may begin in the final steps of the strain design, practically when the technology is ready to exit the lab[2]. The separation process synthesis is based on the reactor's exit stream composition. The exit stream, known as the fermentation broth is usually a dilute mixture of metabolites. According to the desired product portfolio, the stream is separated and the according chemicals are recovered. Since the strain's metabolic network has been adjusted to meet the engineering demands, the flux distributions and metabolite concentrations inside the cell are expected to change as well. That means that the fermentation broth content when culturing the commercial strain might differ substantially from the wild-type counterpart. In conclusion, the two problems, the DBTL iterative cycle towards a stable overproducing strain and the separation process synthesis for minimum cost are addressed sequentially.

To tolerate the high separation costs, the Design module should maintain a view over the resulting exit stream and the consequent separation synthesis problem. In that way, the mutant phenotypes- result of the suggested network modifications, can be evaluated with respect to the necessary separation cost to fulfil the requirements of a specified product -portfolio. The common element for the strain design algorithm and the separation synthesis is the fermentation broth content; where the decisions over genetic modifications made in the first defines the variability of the broth contents and subsequently the optimal separation sequence. A workflow, which allows the communication between the two and provides valuable feedback on the design module of the DBTL cycle, should enclose modules that:

1) Prepare and curate the GEM of the host organism(e.g. by adding heterologous pathways necessary for the production of a specified portfolio compounds)
2) Generate a pool of alternative metabolic strategies that yield in maximum potential revenue for the specified portfolio (e.g. using a strain design algorithm)
3) Computationally asses the resulting fermentation broth's composition if the distinct metabolic strategies are applied (e.g. by sampling the allowed solution space of the according GEM)

4) Estimate the minimum separation cost for each case (e.g. address the synthesis problem using superstructure optimization to identify the minimum cost process flowsheet)
5) Compare the different metabolic strategies with respect to the estimated potential profit (considering both the potential revenue and the downstream separation cost) and choose the most promising mutants to be built and tested in the lab. The profit will not be the only decision parameter, since the different strategies will be evaluated with respect to their biological feasibility and wet-lab construction efficacy.

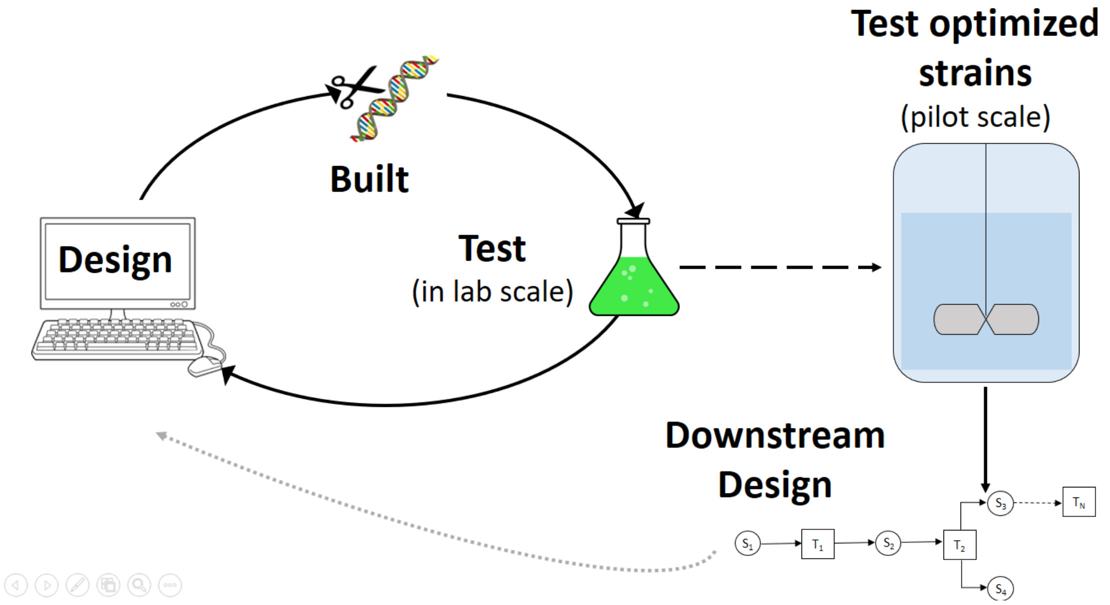


Figure 7: The microbial strain optimization current workflow. The design procedure does not communicate with the downstream process synthesis problem, which happens only after the strain is optimized and cultured in pilot scale. We aim to provide the missing insight from the future process design to the strain design module.

Chapter 3. Problem description and Methodology

3.1 Problem description and workflow outline

3.1.1 Problem description and the main challenges

Typical metabolic engineering approaches applied nowadays to produce efficient cell factories do not take into consideration the separation process synthesis problem that will support the desired bioprocess application. The common practice in computer-aided strain design, usually involves MILP formulations that reveal metabolic strategies (gene deletions and reaction additions, gene upregulations and downregulations) which conclude in maximum yield or growth-coupled production of the target chemical. While these methods are a valuable tool for yield and titer optimization, they do not take into consideration the separation process needs and the subsequent cost that will occur if the proposed metabolic strategies are applied. The different network manipulations may result in different exit stream composition and consequently different separation process design possibilities and yielding costs. It is evident that addressing the strain design problem and the process synthesis problem under a common framework will yield in the rational design of cell-factories that achieve high revenues while preserving low separation costs and favorable downstream process operation conditions. Our main endeavor is thus to develop a systematic workflow that correlates the proposed metabolic interventions of the strain design procedure to potential exit stream compositions and the corresponding optimal downstream process design. The changes in the exit stream content can be approximated by sampling the region of feasible flux distributions of the GEM (solution space).

The primary challenge is to connect the phenotype prediction problem usually addressed via FBA-related methods with the downstream process synthesis procedure. GEMs enclose the total metabolic capacity of organisms. The models are used to predict the microbial behaviour in different regimes and genetic backgrounds. The metabolic capacity is defined by the stoichiometric matrix and the applied capacity and thermodynamic constraints. A prerequisite to address the phenotype prediction problem is to define the cellular objective. Given a specific objective, commonly maximum growth, we can formulate and solve an LP problem that will reveal a flux distribution where the biomass reaction obtains its maximum value. Although the estimated flux distribution satisfies the problem criteria, the solution is non-unique and as a consequence we do not grasp a full idea of the microorganism's secretion capacity. We do not unveil all the potential fermentation broth resulting compositions.

Furthermore, we have to find an effective way to translate the GEM variables to an industrial stream. The variables of the first are described with respect to the bioreactors' biomass content ($mmol \cdot gDCW^{-1} \cdot h^{-1}$) while an industrial stream is characterized by its flow rate, molar composition, etc. Especially in separation processes, we have to identify the critical physicochemical characteristics that will guide the process sequence (on which properties will we base our flowsheeting decisions).

Moreover, we need a systematic way to evaluate the minimum achievable downstream cost for the calculated exit stream compositions. In that way, we will be able to link specific phenotypes with the corresponding downstream cost demands and identify metabolic interventions that lead to maximum profit. The strain design procedures, to our knowledge, have not been connected with the downstream process synthesis. For that reason, we propose a workflow formulation that aims to assist the Design and Learn modules of the DBTL cycle.

3.1.2 Proposed workflow outline

The proposed workflow consists of five steps:

1) In the first step, we reconstruct and curate the Genome-scale model that will work as the basis for the rest of the analysis. To this direction, we identify the heterologous production pathways for a specific portfolio and incorporate them in the Genome-scale model. Next, we proceed with model curation and model reduction using the redGEM[36] and lumpGEM[52] algorithms. Moreover, we incorporate economic factors, related to the upstream process and the products' revenue onto the Genome-scale model.
2) In the second module, using an MILP algorithm we identify alternative metabolic strategies that end up in maximum revenue for a specified set of products and culture conditions.
3) In the third step, we sample the solution space of the GEM after applying the network changes identified in step 2. The changes are expressed in the form of reaction eliminations. The application of the eliminations can be simulated by setting the upper and lower bounds of the reactions to zero. The sampling procedure results in composition and flow-rate estimates regarding the exit stream.
4) In the fourth step, in order to address the separation process synthesis problem we solve a superstructure-based optimization problem. The superstructure encloses all the available equipment alternatives to meet the portfolio specifications. The superstructure yields in a MI(N)LP problem formulation, the solution of which underlines the optimal downstream flowsheet. In that way, we calculate the average separation cost for each metabolic strategy.
5) Finally, we identify the metabolic strategies that yield in maximum revenue at the lowest separation cost. We compare the potential mutant strains with respect to the initial strain performance. The most promising metabolic strategies can be translated in real-life lab decisions.

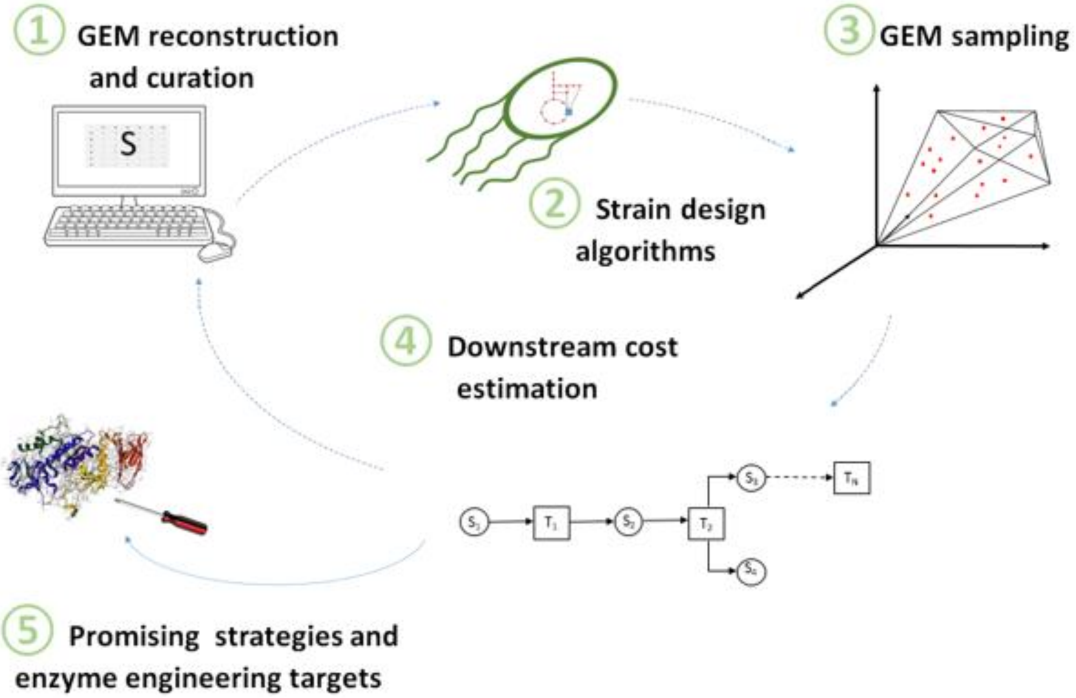

1 GEM reconstruction and curation
S
2 Strain design algorithms
3 GEM sampling
4 Downstream cost estimation
5 Promising strategies and enzyme engineering targets
S1
T1
S2
T2
S3
TN
S4

Figure 8: The strain design workflow

3.1 Genome-scale model reconstruction and curation

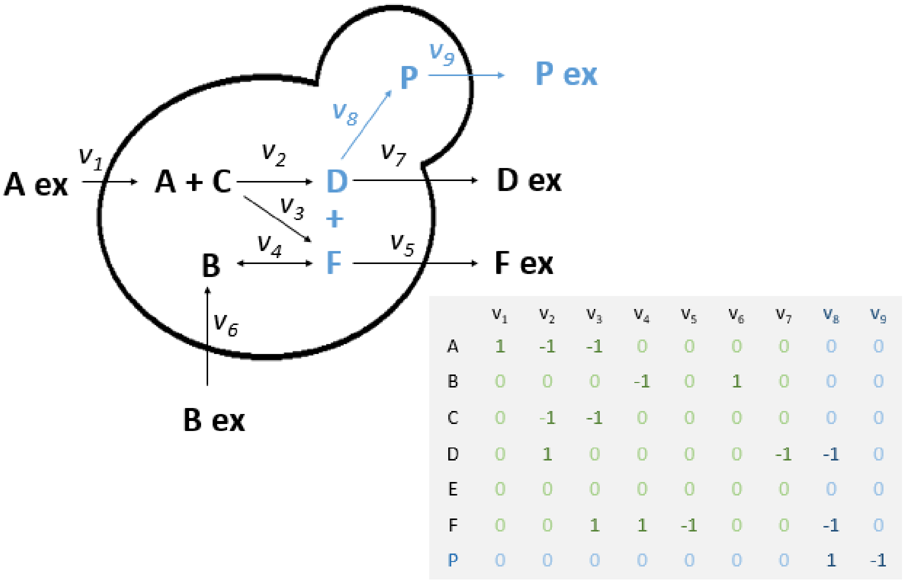
The purpose of the first module is to prepare a functional and practical GEM to conduct the following steps of the analysis. We assume that the designer has already chosen a host organism and the target product-portfolio. In the vast majority of industrial fermentation applications, we utilize one of the so-called platform organisms: *E.coli, S.cerevisiae, C.glutaminicum and A.Niger*. These organisms are well-studied and their genome annotations and metabolic reconstructions are available. For that reason, we do not get into details on the reconstruction procedure, which is extensively described in literature[53]. The GEM preparation is divided in two steps. The first step includes actions concerning the metabolic network while the second step entails economic considerations of the upstream process.

In the first step, the designer identifies the heterologous pathways that enable the production of the target products. The pathways are translated to sequential biochemical reactions where the reactants and the products are cellular metabolites. We augment the stoichiometric matrix according to the novel reactions by adding a number of new columns, equal to the number of the reactions and new lines corresponding to the number of the new metabolites. The pathway selection may lie on literature findings, online biochemical databases or retrobiosynthetic algorithms.

In order to enhance the model accuracy, thermodynamic constraints can be applied. If the portfolio targets include intracellular metabolites, TFA-based methods are mandatory since intracellular concentrations are introduced as variables. In the present study, we focus on metabolites that are secreted to the extracellular matrix.

In this case, the secretion fluxes can be directly used to extract information considering productivity and compounds' concentration in the bioreactor.

The incorporation of experimental data[1] (physiology, fluxomics, metabolomics, etc.) is a common practice that enhances GEM predictability and credibility of the analysis outcome. For practical reasons, GEMs are reduced over specific subsystems of interest. In our methodology we systematically reduce GEMs by applying the redGEM[36] and lumpGEM[52] algorithms. In that way, we to form a core metabolic network including the subsystems of interest while maintaining the paternal GEM main characteristics.

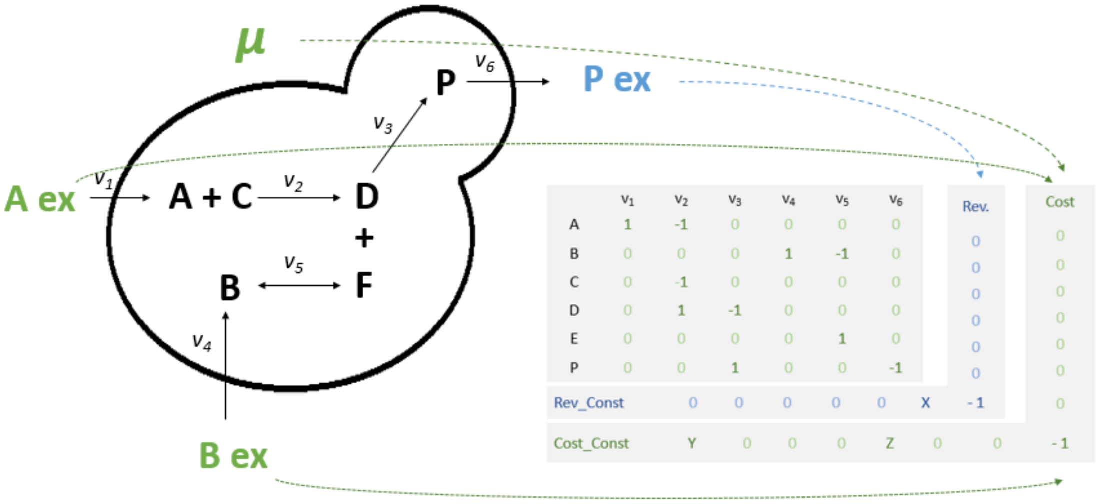


	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
A	1	-1	-1	0	0	0	0	0	0
B	0	0	0	-1	0	1	0	0	0
C	0	-1	-1	0	0	0	0	0	0
D	0	1	0	0	0	0	-1	-1	0
E	0	0	0	0	0	0	0	0	0
F	0	0	1	1	-1	0	0	-1	0
P	0	0	0	0	0	0	0	1	-1

Figure 9: Addition of novel reactions and metabolites to the stoichiometric matrix. With blue, we indicate the heterologous reaction that leads to the production of the desired compound P. The heterologous pathway consists of the reactions V_8 and V_9, which are added as new columns, while the new metabolite P is added as line.

In the second step, we incorporate upstream cost-related and potential revenue variables and constraints onto the GEM context. The upstream cost-related terms correspond to values such as the annualized capital cost of the fermenter, the feedstock price, other utilities etc., while the revenue terms correspond to the resulting revenue if the specified metabolite quantities were sold at the market price. In other words, the potential revenue term is an indicator of the content value. To extract the economics-related constraints and variables we first have to conclude in key upstream process parameters such as the feed stream flow rate and concentration.

[1] The incorporation of experimental expression data (mRNA, protein concentrations, and metabolites concentrations) results in larger in size metabolic and expression models (ME-models)[67]



	v_1	v_2	v_3	v_4	v_5	v_6	Rev.	Cost
A	1	-1	0	0	0	0	0	0
B	0	0	0	1	-1	0	0	0
C	0	-1	0	0	0	0	0	0
D	0	1	-1	0	0	0	0	0
E	0	0	0	0	1	0	0	0
P	0	0	1	0	0	-1	0	0
Rev_Const	0	0	0	0	0	X	- 1	0
Cost_Const	Y	0	0	0	Z	0	0	- 1

Figure 10: A simple metabolic network with the corresponding stoichiometric matrix. The stoichiometric Matrix is augmented by terms accounting for the upstream cost and the products' revenue. The growth rate and the substrate fluxes are used to form the upstream cost while the product secretion flux works as a base for the revenue calculations. The additional constrains' RHS is zero.

3.2 Strain design algorithm

The second module of the workflow comprises of a strain design algorithm suited to propose a pool of alternative metabolic strategies that fulfill the specified criteria set by the designer. In that sense, this part is versatile and alternative available Computer aided strain design algorithms can be utilized. If the designer wishes to computationally asses the downstream separation cost of an already existing pool of alternatives then the algorithmic generation part can be omitted.

3.2.1 OptKnock formulation

The strain design algorithm in use is based on the OptKnock[10,11] MILP formulation proposed by Burgard et.al. Optknock simultaneously handles two competing objective functions: the biological objective of the organism, usually the maximization of cellular growth and the engineering objective set by the designer (usually the overproduction of a desired compound). The formulation aims to identify reaction eliminations that reshape the cellular network in a way that the target chemical production is maximized while the attainable growth rate is at the highest possible levels.

Optknock problem formulation consists of two parts: the outer problem including the engineering objective (maximization of a target flux) and the inner problem of the cellular objective. The outer problem identifies reaction candidates for elimination that maximize the target flux, while the inner problem redistribute metabolic fluxes aiming to maximize the biomass formation in the perturbed network subject to the outer problem-imposed changes. The initial optimization problem is:

$Maximize\ v_{target}$

$subject\ to$

$$\begin{bmatrix} Maximize\ v_{biomass} & \\ subject\ to & \\ \sum_{j \in J} S_{ij} \cdot v_j = 0\ , & \forall\, i \in I \\ LB_j \cdot y_j \leq v_j \leq UB_j \cdot y_j\ , & \forall\, j \in J \\ (inner) & \end{bmatrix}$$

$\sum_{j \in J} (1 - y_j) \leq K$

$y_j \in \{0,1\}, \qquad v_j \in \mathbb{R}, \qquad \forall\, j \in J$

Where,

v_{target}: The target flux to be maximized

$v_{biomass}$: The biomass reaction

y_j: Binary variable that decides whether a network reaction is eliminated. Reactions are eliminated for the value of 1.

K: Maximum allowed reaction eliminations

The initial formulation does not correspond to a linear problem since the inner optimization problem is a constraint to the outer problem. By converting the inner problem to its dual counterpart, congregating the constraints and using the duality property, the authors transformed the problem to a single-level optimization MILP problem:

$Maximize\ v_{target}$

$subject\ to$

$$v_{biomass} = \sum_{j\in J} UB_j \cdot y_j \cdot \mu_j{}^{\mathrm{UB}} - \sum_{j\in J} LB_j \cdot y_j \cdot \mu_j{}^{\mathrm{LB}}$$

$$\sum_{j\in J} S_{ij} \cdot v_j = 0\ , \qquad \forall\, i \in I$$

$$\sum_{i\in I} S_{ij}\lambda_{\mathrm{i}} + \mu_{\mathrm{j}}^{UB} - \mu_{\mathrm{j}}^{LB} = 0\ , \qquad \forall\, j \in j - \{Biomass\}$$

$$\sum_{i\in \mathrm{I}} S_{i,biomass}\lambda_{\mathrm{i}} + \mu_{\mathrm{biomass}}^{UB} - \mu_{\mathrm{biomass}}^{LB} = 0$$

$$LB_j \cdot y_j \le v_j \le UB_j \cdot y_j\ , \qquad \forall\, j \in J$$

$$0 \le \mu_{\mathrm{j}}^{UB} \le \mu_{\mathrm{j}}^{UB,max}\ , \qquad \forall\, j \in J$$

$$0 \le \mu_{\mathrm{j}}^{UB} \le \mu_{\mathrm{j}}^{UB,max}\ , \qquad \forall\, j \in J$$

$$\sum_{j\in J} \left(1 - y_j\right) \le K$$

$$y_j \ \in \{0,1\}, \quad v_j \ \in \mathbb{R}, \ \forall\, j \in J$$

The elimination search is conducted among a predefined reaction list from which we exclude intra and extracellular transport reactions. The solution of the MILP problem identifies alternative metabolic strategies comprising of up to K reaction eliminations. The application of the identified strategies aims to identify strategies that couple revenue with growth. For alternative values of *K* we generate pools of metabolic strategies that will be used to direct the constraints applied in the next step. Each metabolic strategy corresponds to an altered metabolic network, where the identified reactions are knocked-out (the bounds are set to zero).

3.3 GEM sampling

The sampling module aims to provide the connection between the proposed mutant metabolic networks and the superstructure optimization problem. The stoichiometric, capacity and thermodynamic constraints imposed on network reactions, form a solution space, which contains all candidate steady-state solutions. The FBA-based approaches identify only one steady-state solution in which the objective function obtains optimal value. The FBA solution is non-unique and does not provide any information on the range of the metabolic network fluxes that correspond to steady-state solution sets. For that reason, the sampling process is necessary to determine the range of possible steady-state fluxes allowed in the metabolic network under the imposed constraints.

The allowed solution space is uniformly sampled using the artificial centering hit and run (ACHR) algorithm broadly used to estimate flux distributions in metabolic studies[54–56]. The extracellular fluxes are treated according to the same assumptions used to incorporate upstream process data to the GEM to translate each sample to the corresponding fermentation broth composition.

3.4 Separation synthesis using superstructure optimization

Superstructure-based methods are identified as the most systematic way to address the synthesis problem. In general terms, solving a superstructure-based problem firstly requires the representation and generation of a superstructure which encloses all the potentially useful unit operations and the according interconections[12]. Based on the superstructure, we formulate a mixed-integer optimization problem. The problem includes discrete decision variables **y** for the selection of the unit operations and the interconnections and continuous variables **x** that represent flowrates, temperatures, pressures, compositions, etc. The solution of the resulting problem yields in the optimal separation process flowsheet alongside with the optimal operating conditions.

The superstructures yield in MINLP representations of the general form: $Z = \min[f(\boldsymbol{x}, \boldsymbol{y})]\ objective\ function$

$subject\ to$

$$\boldsymbol{h_i}(\boldsymbol{x}, \boldsymbol{y}) \leq \boldsymbol{0}$$

$$\boldsymbol{g_i}(\boldsymbol{x}, \boldsymbol{y}) \leq \boldsymbol{0}$$

$$\boldsymbol{x} \in \boldsymbol{X} = \{\boldsymbol{x} | \boldsymbol{x} \epsilon \mathbb{R}^{\boldsymbol{n}}, \boldsymbol{x}^{\boldsymbol{LB}} \leq \boldsymbol{x} \leq \boldsymbol{x}^{\boldsymbol{UB}}\}$$

$$\boldsymbol{y} \in \boldsymbol{Y} = \{\boldsymbol{y} | \boldsymbol{y} \epsilon \{\boldsymbol{0}, \boldsymbol{1}\}^{\boldsymbol{n}}\}$$

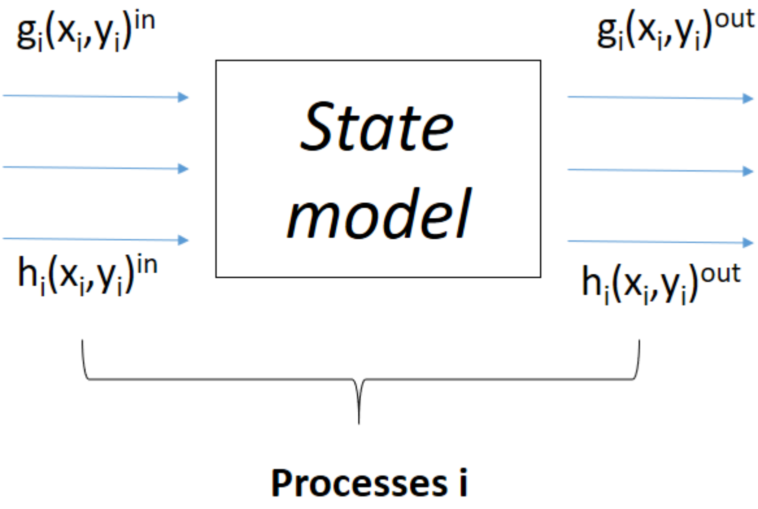


Figure 11: Available processes set the constraints of the mathematical formulation

For example, the State-Task-Network (STN) representation can be used to depict the system. States are defined as the set of physicochemical properties of a stream such as composition, temperature, pressure, particle size, etc. The tasks correspond to the physicochemical transformations that occur between adjacent states[57]. The sampling procedure, alongside with according manipulations (state equation calculations, property databases, etc.) are used to determine the initial feed state based in which the superstructure will be constructed. In the STN representation states and tasks are generally known while the equipment assignment is considered unknown. Each task encloses equipment alternatives that yield in the defined separation. The objective function of the resulting formulation is the minimization of the total separation cost to satisfy the portfolio specifications. That means that the problem includes cost variables corresponding to each potential unit operation.

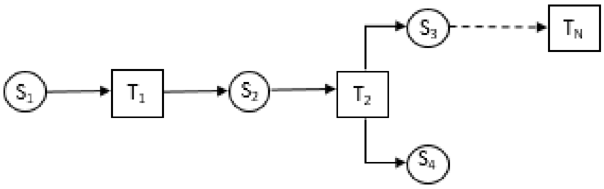


Figure 12: Simple STN superstructure representation

The considered unit operations include distillation, evaporation, adsorption, chromatography, crystallization, filtration, reverse osmosis, etc. Based on the selected alternatives we have to retract the according physicochemical properties in order to calculate the potential intermediate states. We can conduct the calculations with shortcut methods (e.g. Fenske-Underwood-Gilliand for distillation columns) while the cost estimation is based on conceptual cost models since more elaborate calculations are not justified at this stage of the analysis[2,12,34,58].

Chapter 4. Case study: Hydrocarbons producing *S.cerevisiae*

4.1 Application description

Flight-travel-related fossil fuel use and the subsequent CO_2 emissions are expected to increase in the years to come. Unlike the other transportation sectors, there are not sustainable and suitable alternatives to the petrochemically derived kerosene[33]. *S.cerevisiae* has been characterised as one of the most promising platform-organisms to express heterologous genes responsible for long-chain hydrocarbon production (13C-17C), ideal to replace kerosene as drop-in biofuels. During the strain design process, we typically identify genetic interventions that rewire cellular metabolism towards overproduction of the desired compounds. These changes, may as well affect the composition of the fermentation broth and consequently, the downstream separation cost. Since the downstream cost has been proved to be decisive on whether a microbial biorefinery application will be viable, we aim to guide the design procedure towards metabolic strategies that succeed high product revenues while demanding minimum separation expenses.

To this direction, we will apply the proposed workflow to identify metabolic strategies that result in maximum revenue while maintaining low separation costs. In the following paragraphs, we will elaborate on the different workflow modules as they are redefined for the specific case study. *S.cerevisiae* will serve as the host organism which we will genetically modify to produce hydrocarbons (alkanes and alkenes) and fatty alcohols as by-products. We assume that the strain is capable of secreting the produced compounds to the extracellular matrix, thus we do not have to disrupt the cells to harvest the products. We further assume that the cell culture takes place in a chemostat array in aerobic conditions and steady glucose feed. Glucose serves as the only carbon source. Since the target-products are insoluble in water, we assume that the exit stream is split in two. The first split contains the insoluble products, namely the non-native compounds and the rest of the native yeast insoluble secreted metabolites while the second split contains the soluble secreted metabolites. The separation synthesis focuses on the insoluble stream. The components of the stream are identified and matched to specific portfolio products. Because of the nature of the compounds we assume that the products' separation is achieved with sequential distillations. Having only one technology, the superstructure problem is simplified to a supertask.

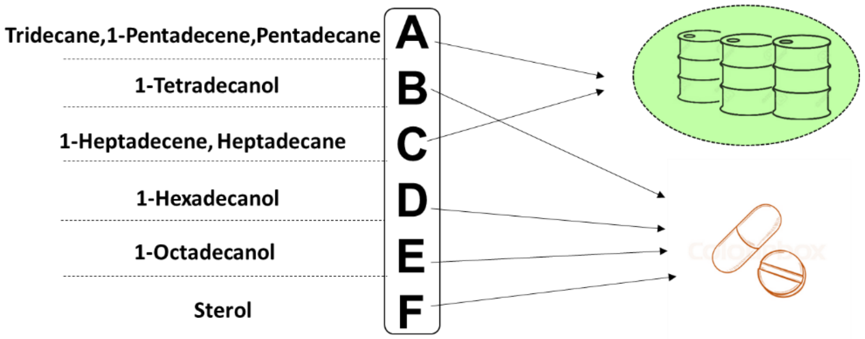

Tridecane,1-Pentadecene,Pentadecane
A
1-Tetradecanol
B
1-Heptadecene, Heptadecane
C
1-Hexadecanol
D
1-Octadecanol
E
Sterol
F

Figure 13: The insoluble metabolites are grouped in product categories

4.2 Genome-scale model curation, reduction and analysis

4.2.1 *S.cerevisiae* for long –chain hydrocarbons production model curation

In the process to evaluate *S.serevisiae* hydrocarbons' production potential and the impact of different metabolic strategies, we incorporated reactions corresponding to 2 heterologous enzymes namely Carboxylic acid reductase (CAR) originated from *Mycobacterium marinum* and a *Synechoccocus elongatus* Aldehyde-deformylating oxygenase (FADO). Alternatively, to the CAR enzyme one could consider using *S.*

elongatus native mechanism for fatty aldehydes production, the Fatty acid reductase (FAR).

As a starting point, we used an extended version of the consensus Yeast GEM iMM904[59]. The model contains nine cellular compartments with 2180 reactions taking place and 1550 participating metabolites. The reactions are divided in 78 subsystems. The Hydrocarbon Producing Strain GEM (HPS) contains 57 extra reactions grouped in a new subsystem, as well as 24 metabolites that do not exist in the initial model. The 20 reactions correspond to transports between the different compartments while the number of unique metabolites is 16, meaning that in the stoichiometric matrix we account for species multiple times if they participate in reactions located in different compartments.

4.2.2 Thermodynamic-based Flux Analysis

Further constraining the HPS GEM, we applied the TFA framework materialized in the form of matTFA, the matlab toolbox implementation of TFA[60]. In that way, we reduce the feasible solution space by adding thermodynamic constraints regarding metabolites' and reactions' Gibbs Free Energy (ΔG). In order to translate the model to the TFA equivalent we enriched the database available with the toolbox, with information considering the fatty-acid metabolism and the newly added reactions and metabolites.

4.2.3 Model reduction

After the thermodynamic curation, the model was reduced using the redGEM[36] and lumpGEM[52] algorithms. The reduction is conducted for aerobic conditions and culture medium containing glucose as the sole carbon source.

First, we selected 9 subsystems of the initial HPS GEM (Acyl Biomass, Biomass, Carnitine Shuttle, Fatty Acid Biosynthesis Mitochondrial, Fatty Acid Biosynthesis, Fatty Acid Degradation, Fatty Acid Elongation, Glycolysis, Heterologous Alkanes Production) around of which the core metabolic network will be constructed. In the next step, after excluding small metabolites, co-factors and inorganics, the algorithm identifies metabolites and reactions pairs between the selected subsystems by directed graph search with respect to the degree of connection D. The selected degree of connection in our case is D=1. The core network is finalized by a second graph search to find connections of the D_1 network with the extracellular space.

Finally, utilizing the lumpGEM[52] algorithm we connect the core network, generated by redGEM, with the Biomass Building Blocks (BBBs) of HPS GEM biomass reaction. The algorithm identifies the smallest alternative subnetworks (S_{min}) that are capable of producing the distinct BBBs with the metabolites found in the redGEM-generated network as a starting point. Finally, lumpGEM calculates unique lumped reactions for all the BBBs. We tested two alternative reduced models with respect to the lumped reactions availability; in the first constructed reduced model we accounted for one lumped reaction per BBB appointed randomly by the algorithm while in the second reduced model, we accounted for all the different lumped reactions that correspond to stoichiometrically balanced subnetworks of size equal to *Smin*.

The reduced models passed a series of tests to evaluate their consistency with the HPS GEM. The performed test consist of 1) calculations of the maximum biomass and products' yields, 2) Single gene-essentiality check for the genes shared with the parent GEM and 3) Thermodynamic flux variability analysis to estimate the feasible range of different subsystems' fluxes.

The gene-essentiality is simulated by setting the lower and upper bounds of a reaction set, which corresponds to the gene under investigation, to zero. Then the usual FBA LP biomass maximization problem is solved; if the system cannot produce biomass the gene is characterized as essential and the gene knock-out as lethal. The procedure is repeated for the whole set of shared genes between the reduced and the parent model.

4.2.4 Flux variability comparison under Biomass and Product yield flux constraints

Here, we attempt to assess the different flux profiles attained in two extreme cases: 1) A biomass proliferating strain hydrocarbon producer and 2) A hydrocarbon overproducing strain.

1) The biomass proliferating strain hydrocarbons producer concept, herein is a strain that while constrained to exhibit growth rates higher or equal to the 95% of the maximum biomass yield, is in addition constrained to produce hydrocarbons in rates higher or equal to the 95% of the maximum yield attained in proliferation conditions. This concept encloses useful information on the characteristics of a strain that while proliferating is a de-facto hydrocarbons producer (the production rates are higher than zero).

2) The second concept corresponds to the opposite scenario, where the strain is constrained to produce hydrocarbons in rates higher or equal to the 95% of the maximum yield while exhibiting growth rates higher or equal to the 95% of the maximum attainable growth rate.

The two scenarios are built for the three alkanes present in the model (n-tridecane, n-pentadecane, n-heptadecane) and each pair was compared according to the fluxes variability for different reaction subsystems.

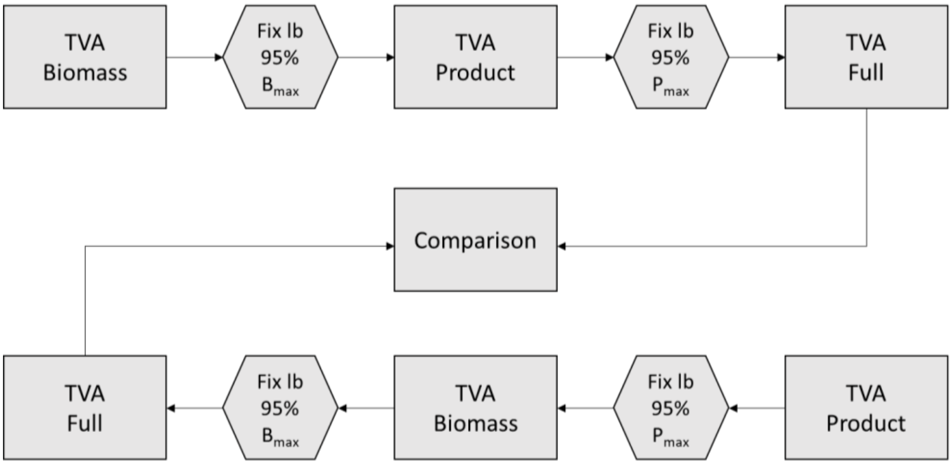


Figure 14: The process we follow to compare the flux variability between the two extreme scenarios.

4.3 Incorporating economic factors onto the Genome-scale model

In order to better assess the potential of the present strain to serve as a kerosene producer, we incorporated upstream cost-related variables and constraints onto the GEM context. In that way, we can directly link the metabolic network properties with the upstream (production) cost and identify network changes that lead to maximum profit scenarios.

Our goal is to enrich the stoichiometric matrix by two sets of elements, where the first set corresponds to variables related with the upstream cost (capital cost, utilities, etc.) and the second to the potential revenue corresponding to specific metabolites produced by the cell, sold at the market price, while the downstream cost is not taken into account.

4.3.1 Upstream process parameters identification and incorporation to the GEM

a) Upstream process basic characteristics

First, we have to define the basic characteristics of the upstream process that will be used to formulate the upstream cost equations. We assume a 994 T/day glucose supply (240 g/L), which corresponds to the amount of glucose generated by the NREL process when hydrolyzing 2000 TDW/day biomass, previously used by *Maravelias et al.* to evaluate the suitability of different microbes and metabolic engineering strategies for the production of fuels and chemicals[34]. The economic parameters presented in this section are also adopted from the aforementioned study unless otherwise stated.

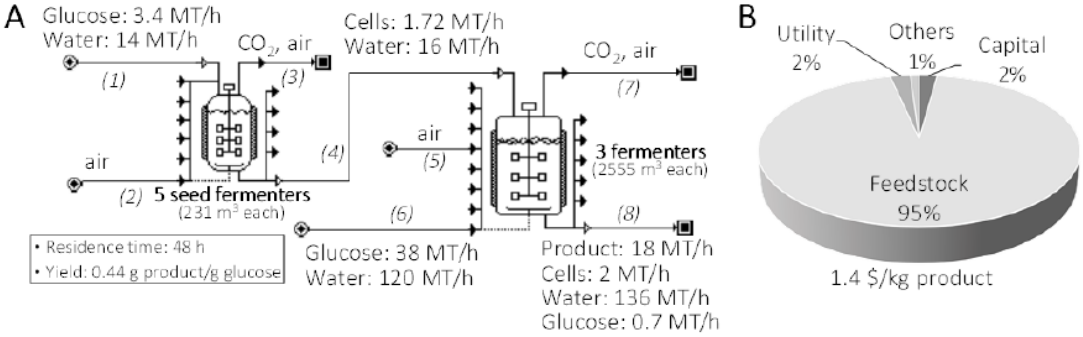


Figure 15: (A) The base case-scenario proposed by Maravelias et al. for the upstream cost calculations (B) The cost distribution for the base-case scenario[34].

The system in study consists of a chemostat array with a steady Feedstock flow rate of 994 T/day. The exit stream is split in two parts: The first part contains the total of the insoluble metabolites secreted by the yeast, present in the reduced model and the second the soluble products. The separation cost for the split is considered negligible. The cells are removed without product loss in any phase. The secreted insoluble metabolites found in the model include alkanes, fatty alcohols, fatty acids and sterol. The soluble metabolites considered in the study include several carboxylic acids and amino acids. While they are far from the total of the compounds that can be found in the stream they will work as an indicator of the potential revenue if they could be retrieved from the fermentation broth without any extra action needed.

In this part, we have to mention that although the GEM contains reactions for the production and secretion of monosaturated fatty alcohols, these molecules are accounted together with their unsaturated counterparts under the properties of the later. Furthermore, due to the vapour pressure similarities between myristate and octadecanol mixtures, the two compounds are accounted as one mix product, uniformly behaving as 1-octadecanol.

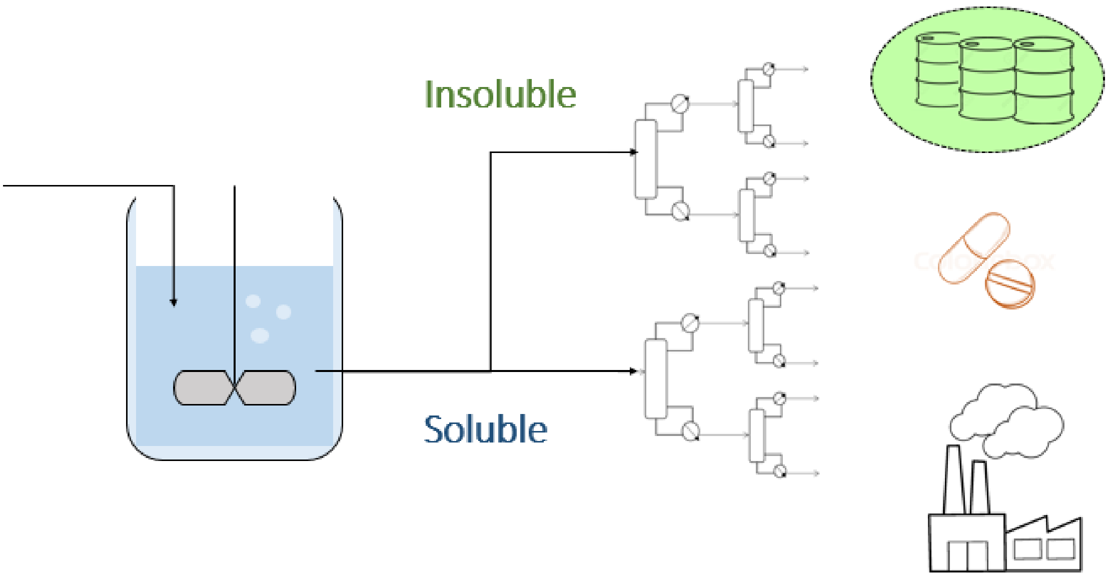


Figure 16: Schematic of the exit-stream split to a stream containing the insoluble products and another that contains the solubles. Although our study focuses on the insoluble products valorization, we proceed with estimating the maximum achievable revenue in the case where all the products were included in the portfolio. The produced compounds may be used as fuels, platform-chemicals and precursors for the production of different molecules or pharmaceuticals.

Table 2: Market prices for the metabolites present in the insoluble products' stream.

		Insoluble		
Name	**MW**	**Product**	**Price ($/T)**	**Use**
Tridecane	184	Kerosine	520	Fuel
Pentadecene	210			
Pentadecane	212			
Heptadecene	238			
Heptadecane	240			
Tetradecanol	214	1-Tetradecanol	1500	Cosmetics, Industrial, Emulsifier
Hexadecanol	242	1-Hexadecanol	1500	
Myristate	228	Octadecanol/ Myristate mix		
Octadecanol	270			
Sterol	385	Sterol	1000	Food, Industrial

Table 3: Market prices for the metabolites present in the soluble products' stream.

		Soluble		
Name	**MW**	**Product**	**Price ($/T)**	**Use**
Formic Acid	46	Formic Acid	600	Food, Insecticide, Industry
Acetic Acid	60	Acetic Acid	300	Food, Pharmaceuticals, Industry

Lysine	146	Lysine	1050	Food
Pyruvic Acid	87	Pyruvic Acid	1000	Food, Industry
Lactic Acid	90	Lactic Acid	1000	Food, Industry
Glutamic Acid	147	Glutamic Acid	800	Food
Fumaric Acid	116	Fumaric Acid	1000	Food, Pharmaceuticals, Industry
Succinic Acid	118	Succinic Acid	1100	Food, Industry
Citric Acid	192	Citric Acid	700	Food, Industry

b) Upstream cost terms

The upstream cost consists of 5 terms:

$$Upstream\ Cost = Substrate\ Cost + Annualized\ CAPEX + Reactor\ OPEX + Labour + Other$$

Excluding the substrate cost, the terms are expressed as a function of the reactor active volume:

Table 4: Cost terms for the upstream process

Cost Term	Expression
Substrate Cost	$Glucose\ Cost + Water\ Cost$
Annualized CAPEX	$CCF \cdot Reactor\ PC\ \cdot Volume$
Reactor OPEX	$Specific\ Power \cdot Electricity\ Cost\ \cdot Volume \cdot Wh$
Labour	$Specific\ Labour\ Needs \cdot Volume \cdot Salary \cdot Wh$
Other	$(IMF + IF) \cdot Reactor\ PC \cdot Volume$

Where *,*

Glucose Cost: Annual glucose purchase cost [M\$/Y]

Water Cost: Annual water purchase cost [M\$/Y]

CCF: Capital Charge Factor

Reactor PC: Reactor Purchase cost per bioreactor volume [M\$/m^{3}*]*

Volume: Bioreactor volume [m^{3}*]*

Specific Power: Electricity needs per bioreactor volume [KW/m^{3}*]*

Electricity Cost: Electricity Cost [\$/KW-h]

Wh: Unit Working hours [h]

Specific Labour Needs: Needs in labour per reactor volume [No workers/m^3]

Salary: Hourly salary per worker [\$/(worker·h)]

IMF: Insurance and maintenance cost factor

IF: Installation Factor

A detailed chart with the economic factors' values used for this study can be found in the appendix.

The upstream cost can be expressed as a two-termed equation where the first term is a linear function of the volume and the second one, the substrate cost, can be assumed constant for our case.

$$Upstream\ Cost = f(V) + Substrate\ Cost$$

$$f(V) = a \cdot V$$

$$a = (Specific\ Power \cdot Electricity\ Cost\ + Specific\ Labour\ Needs \cdot Salary) \cdot Wh \\ + (IMF + IF + CCF) \cdot Reactor\ PC = Const$$

In order to account for the Upstream cost in the GEM we add three variables and the accompanying constraints:

i) The volume variable: The volume in a chemostat system is given by the equation :

$$V = \frac{F}{D} = \frac{F}{\mu}$$

Which can be approximated by a linear relationship for a tight region of growth rates such that:

$$V = \beta \cdot \mu + \gamma$$

The corresponding constraint is:

$$\text{Volume\_const: } V - \beta \cdot \mu = \gamma$$

Where *β* is the slope and *γ* is the intersect of the linear approximation.

ii) The upstream cost variable is as described above, while the corresponding constraint is:

$$\text{Upstream\_const: } Upstream\ Cost - a\ \cdot V = Substrate\ Cost$$

c) Revenue terms

We assume that the bioreactor system in use consists of an array of chemostats in steady state. In this case, the dilution rate (*D*) is equal to the growth rate:

$$D = \mu \ \ (1)$$

Thus, the substrate *[S]* in the exit stream will be:

$$\frac{dS}{dt} = u_S \cdot [B] + D \cdot ([S]_0 - [S]) = u_S \cdot [B] + D \cdot \Delta S = 0 \ \ (2)$$

From (1) and (2):

$$[B] = \frac{\Delta S}{u_s} \cdot D = \frac{\Delta S}{u_s} \cdot \mu \ \ (3)$$

The product *[Pj]* balance will similarly be:

$$\frac{dP_j}{dt} = u_{Pj} \cdot [B] + D \cdot \left(\left[P_j\right]_0 - \left[P_j\right]\right) = 0 \xrightarrow{\left[P_j\right]_0=0} \left[P_j\right] = \frac{\Delta S}{u_s} \cdot u_{Pj} \ (4)$$

Where u_s and u_p are the glucose uptake rate and the product j secretion rate respectively.

The potential annual revenue from the product j will be:

$$Revenue_j \left[\frac{M\$}{Y}\right] = \frac{\Delta S}{u_s} \cdot u_{Pj} \cdot \frac{Mr(P)}{Mr(S)} \cdot Price(P) \cdot 330 \left[\frac{\frac{Tons\ Glc}{day}}{\frac{mmol\ Glc}{gDCW\ h}} \cdot \frac{mmol\ P_j}{gDCW\ h} \cdot \frac{\frac{gr\ P}{mol\ P}}{\frac{gr\ Glc}{mol\ Glc}} \cdot \frac{M\$}{Ton\ P} \cdot \frac{day}{Y}\right] \ (5)$$

While the total potential revenue will be:

$$Revenue = \sum_{j=1}^{n} Revenue_{\ j} \ \ (6)$$

The glucose in the exit stream is considered to be *[S] = 0,02 [S]$_o$* and the glucose uptake rate -5 $\frac{mmol\ Glc}{gDCW\ h}$:

Revenue$_j$_Const: $[357{,}18 \cdot Mr(Pj) \cdot Price(Pj)] \cdot u_{Pj} - \ Revenue_j = 0 \ \ (7)$

For each product j we want to account for, we add one extra variable $Revenue_j$ and an extra constraint Revenue$_j$_Const in the form presented in eq.7. The RHS of the constraint is zero.

d) Gross profit for the upstream process

The gross profit for the upstream part of the process can be directly calculated by subtracting the cost of goods sold (COGS- here in the form of upstream cost) from the potential revenue.

Even though the calculations presented here are a gross approximation of the upstream cost, the profit value can work as an indicator of the available margin for the separation costs to follow the upstream process. The presence of COGS and revenue variables in the model, allows the direct investigation of alternative metabolic network conformations and exit stream compositions that increase profit.

The estimation of the maximum attainable profit for different values of the Biomass reaction can be calculated with the CobraToolbox robustnessAnalysis function.

4.4 Strain design algorithm

For the case study, the engineering objective is the potential revenue from the insoluble metabolites in the exit stream. In other words, we form an objective function were the identified market price of each component works as a weight for the corresponding production flux. Since the cellular growth rate is imposed by the Dillution rate in the case of the chemostat, we narrow down our search for growth rates close to the target *D* value.

$$Maximize\ Revenue_{insolubles}$$

$$subject\ to$$

$$v_{biomass} = \sum_{j\in J} UB_j \cdot y_j \cdot \mu_j{}^{\mathrm{UB}} - \sum_{j\in J} LB_j \cdot y_j \cdot \mu_j{}^{\mathrm{LB}}$$

$$\sum_{j\in J} S_{ij} \cdot v_j = 0\ , \qquad \forall\, i \in I$$

$$\sum_{i\in I} S_{ij}\lambda_{\mathrm{i}} + \mu_{\mathrm{j}}^{UB} - \mu_{\mathrm{j}}^{LB} = 0\ , \qquad \forall\, j \in j - \{Biomass\}$$

$$\sum_{i\in \mathrm{I}} S_{i,biomass}\lambda_{\mathrm{i}} + \mu_{\mathrm{biomass}}^{UB} - \mu_{\mathrm{biomass}}^{LB} = 0$$

$$LB_j \cdot y_j \le v_j \le UB_j \cdot y_j\ , \qquad \forall\, j \in J$$

$$0 \le \mu_{\mathrm{j}}^{UB} \le \mu_{\mathrm{j}}^{UB,max}\ , \qquad \forall\, j \in J$$

$$0 \le \mu_{\mathrm{j}}^{LB} \le \mu_{\mathrm{j}}^{LB,max}\ , \qquad \forall\, j \in J$$

$$\sum_{j\in J}(1 - y_j) \le K$$

$$y_j\ \in \{0,1\}, \quad v_j\ \in \mathbb{R}, \ \forall\, j \in J$$

The OptKnock search for growth-coupling alternatives is conducted to the upstream cost factors-enriched GEM that we discussed in chapter 4. We set the glucose uptake rate $\boldsymbol{v_{glc}}$=-5 mmol · gDCW^{-1} · h^{-1} and specify the list of reactions to consider deleting, as well as the maximum number of allowable reaction eliminations K. The reaction list does not include intra- and extra- cellular transport reactions or lumped reactions. In order to consider alternative solutions of the same length K, we impose integer cuts. The problem construction is conducted with a self-developed matlab function compatible with the matTFA toolbox. The problem is solved using IBM ILOG CPLEX 12.7.1.

4.5 GEM sampling

The FBA solution consists of a single flux distribution corresponding to maximum growth under the given environmental conditions, if the objective is maximization of the biomass reaction. The FBA problem is underdetermined, there are multiple flux distributions that satisfy the imposed constrains while achieving optimal value for the objective function. Since the scope of our application is to estimate the composition of the chemostat array exit stream, we cannot solely rely on the FBA solution. Alternatively, we sample the solution space to attain different flux distributions that are allowed by the mass balance and capacity constraints in order to extract an estimate of the exit stream composition.

In order to proceed with model sampling, we impose the network modifications identified in the previous step and further constrain the glucose uptake rate to $\boldsymbol{v_{glc}}$=-5 mmol · gDCW^{-1} · h^{-1} and biomass reaction to take values $\mu \in \{0.9, 0.11\}$. The sampling is conducted with the COBRA toolbox matlab sampling function[61] available with matTFA[60].The approach used is based on the artificial centered hit and run algorithm(ACHR)[62]. For each case, we generate 5000 samples that are further analyzed to estimate the exit stream content, the produced kerosene characteristics and the potential revenue.

Due to computational time limitations, we cannot proceed with the downstream cost calculations for each sample. To address this issue, we use each sample to construct a vector that comprises of the extracellular fluxes of interest. The vectors are clustered using the kmeans[63] matlab embedded function to form k clusters of vectors of different size. The algorithm partitions the observations into k clusters in which each observation belongs to the cluster with the nearest mean, where the mean works as a prototype of the cluster. Thus, the mean of each cluster will function as the Input for the downstream problem while its size will function as a weight to calculate the average downstream cost $\overline{TC}$ and the average revenue $\overline{TR}$ that will be used to compare the alternative strategies.

$$\overline{TC} = \sum_{k=1}^{C} \frac{W(k)}{n} \cdot TC(k)$$

$$\overline{TR} = \sum_{k=1}^{C} \frac{W(k)}{n} \cdot TR(k)$$

Where,

C is the number of clusters

W(k) is the size of the k cluster

n is the number of samples

TC(k) is the downstream cost calculated for the kth cluster

TR(k) is the revenue calculated for the kth cluster

$\overline{TC}$ is the average downstream cost

$\overline{TR}$ is the average revenue

4.6 Downstream cost calculations

4.6.1 The synthesis problem

The approach used in this study developed by Shah and Kokossis (1997) combines optimization technology in the form of mathematical programming, engineering insights and shortcut design models. These three components are integrated into conceptual models, which include only basic information of the process and set up the background for the application of Conceptual Programming[46].

A superstructure based on the number of distinct separation tasks T is generated by using the list processing technique proposed by Hendry and Hughes (1972). The separation tasks receive a single feed and yield in two products. They are divided into different subsets M according to the product subgroup associated with the feed. The representation enables estimates of compositions and nominal flowrates of feed and products without prior knowledge of the performance of the upstream and downstream tasks. The shortcut methods[2] are employed to calculate the important process parameters (e.g. Fenske-Gilliland method for the calculation of the number of trays, Underwood's method for estimating the minimum reflux ratio, etc.). The process parameters and column basic characteristics are utilized to create the cost-related variables that correspond to each task. Formulation of this problem leads in a mixed-integer linear program, in which the total annualized venture cost is minimized with respect to the tasks sequence.

a) Task cost variables

The total cost of each task is the sum of the capital cost correlated with the column in use and the operating cost linked with the needed utilities.

[2] In order to proceed with the shortcut calculations, it is assumed that the relative volatility remains constant across each distillation column.

The capital cost consists of the column cost, $Cost^{col}$ and a fixed charge cost $Cost^{fix}$. The former is expressed as a function of the vapour to feed ratio V/F and the number of trays NT. The former represents the relative cost of the column in the available design options, while the fixed cost includes the cost of needed supplementary items such as piping.

$$Cost^{col} = a_1 \cdot \left(1 + b_1 \cdot \frac{V}{F}\right) \cdot NT$$

$$Cost^{fix} = a_2 \cdot \left(1 + b_2 \cdot \frac{F}{F_{tot}}\right)$$

The constants α_1, α_2, b_1, b_2 are generic and independent of the separation system, the components or the composition of the feed, merely representing the economic environment.

Table 5: Constants of cost model

No	Generic constant	Value
1	a_1	15.00
2	a_2	765.00
3	a_3	0.05
4	a_4	0.10

The utility cost $Cost^{util}$ for each task is:

$$Cost^{util} = V \cdot (1 + C^{hot} + C^{cold})$$

Where the utility cost indexes C^{hot} and C^{cold} are functions of reboiler temperature (T^{reb}) and condenser temperature(T^{cond}).

Table 6: Discrete cost indices for hot utilities. The discrete cost indices are used to construct continuous cost indicest-dependent functions. For target temperatures outside the ementioned ranges, the cost indices are extrapolated.

No.	Type	Temperature Range (°C)	Cost Index C^{hot}

1	Very low-pressure steam (VLP)	up to 80 °C	0.100
2	Low-pressure steam (LP)	80 °C to 118 °C	0.130
3	Medium pressure steam (MPS)	118 °C to 164 °C	0.169
4	High-pressure steam (HPS)	164 °C to 186 °C	0.187
5	Very high-pressure steam (VHP)	186 °C to 230 °C	0.261
6	Furnace or Hot oil (HO)	230 °C to 260 °C	0.280

Table 7: Discrete cost indices for cold utilities. The discrete cost indices are used to construct continuous cost indicest-dependent functions. For target temperatures outside the ementioned ranges, the cost indices are extrapolated.

No.	Type	Temperature Range (°C)	Cost Index C^{cold}
1	Cooling Water (CW)	45 °C onwards	0.017
2	Chilling water (R1)	0 °C to 45 °C	0.200

b) Constraints of the synthesis problem

The constraints include simple material balances and cost-related expressions. The logical constraints associate integer and continuous variables.

Mass balance expressions:

$$F^{tot} - \sum_{k \in T^{fr}} f_k = 0$$

$$\sum_{k \in T^{fr}} \left(\zeta_{\mathrm{m,t}} \cdot f_t\right) - \sum_{k \in T_m^{inp}} f_k = 0 \;\; \forall \, m \in M$$

Annualized cost expressions for each column:

$$Cost^{tot} = \left[a^{ann} \cdot \left(Cost^{col} + Cost^{fix}\right) + Cost^{util}\right] \cdot \frac{f_t}{F_t} \; \forall \, t \in T$$

Logical constraints:

$$\mathrm{f}_t - F_t \cdot Y_t \leq 0 \, \forall \, t \in T$$

c) The objective function

$$f^{obj} = \sum_{t \in T} Cost_t^{tot}$$

4.6.2 The 6-product stream separation problem

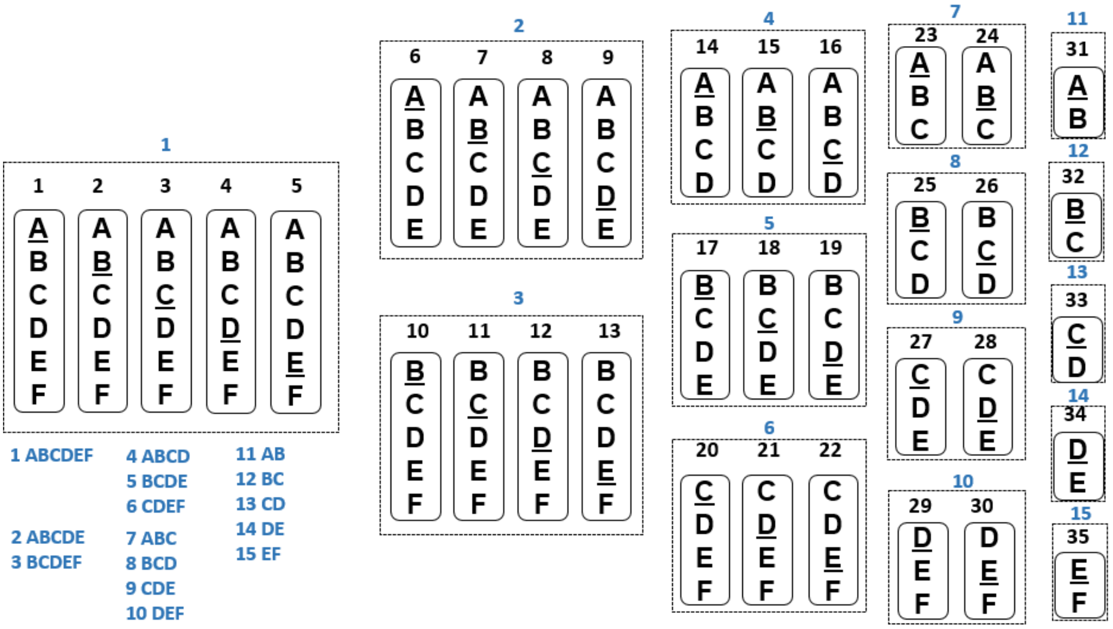


Figure 17: Discrete representation of simple column sequences for a 6-product stream. The 35 tasks are denoted with black enumeration while the 15 subgroups with blue.

Let T={t} be the set of tasks and M={m} be the set of product subgroups. For a single source problem leading to six products there are 35 tasks: t_1,t_2,…,t_{35} and 15 subgroups: m_1,$m_{2,…,}$$m_{15}$. Given the recovery matrix and the molar fraction of components at the feed stream we can calculate the distillate and bottom molar fraction for each task. The distillation columns operate in atmospheric pressure while the pressure drop along each column is considered negligible. We use the Antoine equation to estimate the temperatures at the top and the bottom of each column. Then, using FUG short-cut methods, we calculate the vapour load (*V)*, vapour to feed ratio (*V/F)* and the number of theoretical trays (*NT*) for each task. The operating cost indices C^{hot} and C^{cold} are calculated as well for each task based on the top and bottom temperatures.

Knowing the parameters of the cost variables for each task, we can construct the mixed-integer linear program as described in the previous section. By solving the problem we identify the sequence that succeeds in product separation (at the pre-described purity levels) at minimum cost. The problem matrices construction is conducted by a developed set of matlab functions. The necessary input is the recovery matrix, the stream composition and the components' Antoine coefficients to calculate the condenser and reboiler temperatures. The parameters needed for the problem construction that yield from non-linear equations are approximated using matlabs' built-in vpasolve function. The problem is solved using IBM ILOG CPLEX 12.7.1.

Finally, since the conceptual cost value identified by the minimization problem cannot be used to compare cases of different flowrate, we estimate the CAPEX corresponding to the selected distillation sequence.

The components physicochemical and economic parameters and constants can be found in Appendix A. The detailed calculations followed for the problem construction are presented in Appendix B. The detailed sizing and cost models followed to calculate the sequence CAPEX can be found in Appendix C.

Chapter 5. Case study results

5.1 Results outline

The results presentation follows the exact same pattern with the proposed workflow. First, we discuss about the GEM characteristics and reduce the parent model following two alternative approaches. The two reduced models are compared with the parent model to test their consistency. Furthermore, we utilize the economics-enriched GEM to estimate the maximum potential revenue for different product portfolios. We evaluate whether the hydrocarbons annual revenue is growth-coupled and identify a promising metabolic strategy for revenue maximization.

Moreover, we apply the strain design algorithm to the economics-enriched GEM to create a pool of alternative metabolic strategies yielding in maximum revenue. Each strategy is implemented in the GEM by applying reaction eliminations. The result mutant model is then sampled to obtain alternative allowed flux distributions and asses the hydrocarbon production potentials of the strain. The samples are clustered to groups based on the similarity of the insoluble products' fluxes.

The alternative calculated clusters consequently represent each metabolic strategy. We estimate the minimum downstream separation cost and the potential revenue for the mean of each cluster and proceed with calculating an average mean for each case. Finally, we present the comparative data between the potential mutant strains with respect to the initial strain performance. The performance indicator is the difference between the potential revenue and the downstream cost.

5.2 Genome-scale model curation, reduction and analysis

5.2.1 Model general characteristics

The model contains heterologous reactions producing long-chain alkanes and alkenes. Fatty aldehydes work as intermediates for the alkanes production while fatty alcohols are possible by-products. Schematically the alkanes production process involves the fatty aldehyde production from the corresponding fatty acids and then the aldehyde deformylation towards alkanes and alkenes. Fatty-aldehydes can be transformed to fatty alcohols by the Alcohol dehydrogenase enzymes (ADH)[3] present in yeast's cytosol. We assume that the produced alkanes and alkenes as well as the fatty alcohols are secreted in the extracellular space. The secretion reactions considered here are energy and co-factor independent.

[3] The ADH fatty-aldehydes to fatty-alcohols reactions were not included in the model for each aldehyde separately but are included in the form of pool reactions. For that reason, we included specified reactions for this transformation.

Assuming that the fatty aldehydes and fatty acids can diffuse through the peroxisomal membrane, the compartmentalisation of the production mechanism (map and contain the enzymes inside the peroxisome) does not result in any difference in the attainable flux profiles and final product yields. In all the cases addressed here, we assume that the production reactions can occur both inside the cytosol and inside the peroxisome[4].

Table 8: The heterologous reactions added in the model

Enzyme	Reactions
mmCAR	nadph_x + ttdca_x <=> nadp_x + tdcal_x
	hdca_x + nadph_x <=> hxdcal_x + nadp_x
	hdcea_x + nadph_x <=> nadp_x + hxdceal_x
	nadph_x + ocdca_x <=> nadp_x + ocdcal_x
	nadph_x + ocdcea_x <=> nadp_x + ocdceal_x
seFADO	h_x + 2 nadph_x + o2_x + tdcal_x <=> h2o_x + 2 nadp_x + trdcn_x + for_x
	h_x + 2 nadph_x + o2_x + hxdcal_x <=> h2o_x + 2 nadp_x + for_x + pntdcn_x
	h_x + 2 nadph_x + o2_x + hxdceal_x <=> h2o_x + 2 nadp_x + for_x + pntdcen_x
	h_x + 2 nadph_x + o2_x + ocdcal_x <=> h2o_x + 2 nadp_x + for_x + hptdcn_x
	h_x + 2 nadph_x + o2_x + ocdceal_x <=> h2o_x + 2 nadp_x + for_x + hptdcen_x

[4] Our GEM does not account for toxicity considerations. The compartmentalization of the production mechanism in reality might be proven to raise the production yields because toxic by-products like peroxidase are catabolised[68].

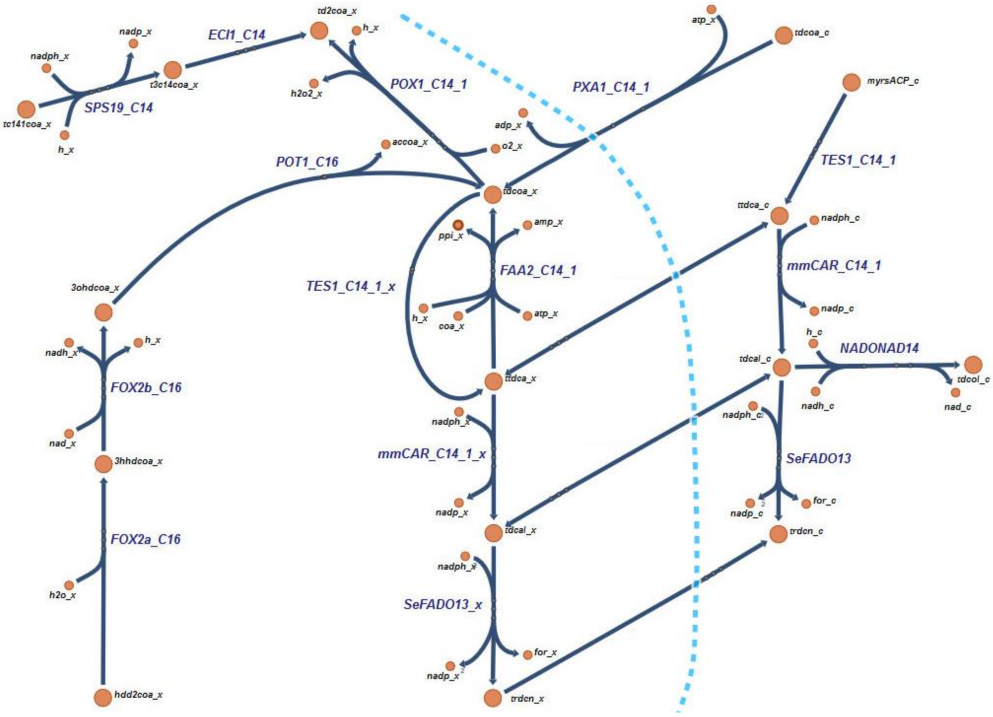


Figure 18: Visualization of the reactions associated with tridecane production. The alkane synthesis happens both in the cytosol (right) and the peroxisome (left).The other hydrocarbons are produced identically from the same enzymes and complexes.

5.2.2 Thermodynamic coverage and reduced models

We were able to assign the Gibbs free energy of formation to 80% of the participating metabolites and the Gibbs free Energy of Reaction to 66% of the reactions. The HPS GEM containing the heterologous reactions and the corresponding transport reactions needed to balance them, was reduced to produce two models: The first containing one lumped reaction per BBB (1 per BBB model) and the second all the possible different lumps corresponding to a *Smin* sized subsystem (Smin model).

The maximum growth rate under thermodynamic constraints and glucose uptake rate $\boldsymbol{v}_{\boldsymbol{glc}}$=-5 mmol · gDCW^{-1} · h^{-1} is 0.4582 h^{-1} for the 1 per BBB model and 0.4850 h^{-1} for the Smin model, while for the original GEM is 0.4872 h^{-1}. The 1 per BBB model contains 395 enzymatic reactions out of a total of 685 with 500 participating metabolites. The Smin model contains 579/827 enzymatic reactions with 496 participating metabolites. The resulting core network after applying the redGEM algorithm is identical in both cases (S matrix size 424 x 606). The difference in size lies in the post processing of the model after the lumped reaction calculations.

Table 9: Reduced models' characteristics

	1 per Biomass Building Block	Smin
Enzymatic Reactions	395	579
Transports	64	61
Lumps	41	187
Total	685	827
Metabolites	500	496
Growth rate (GEM: 0.4872 h^{-1})	0.4582 h^{-1}	0.4850 h^{-1}

The product molecules' flux range appears to be more constrained in the case of the original GEM resulting in lower maximum production yields compared to both the 1 per BBB and the Smin models (Figure 19). This points out that reactions belonging to subsystems excluded from the reduced core network, pose a significant extra constraint in the product yields. Even though, the reduced models appear to reach higher possible levels of hydrocarbons production, the Smin model is more consistent with the original model that the 1 per BBB. The Smin model maximum yields for the different species seem to be sensitive to both the saturation level and the chain length, exhibiting a similar behaviour to the original GEM, while the 1 per BBB model appears to be saturation-insensitiveTable 10. The Smin reduced model appears to be more consistent with the original GEM in terms of maximum biomass and product yields. Henceforth, the model used for the analyses and manipulations will be the Smin reduced model.

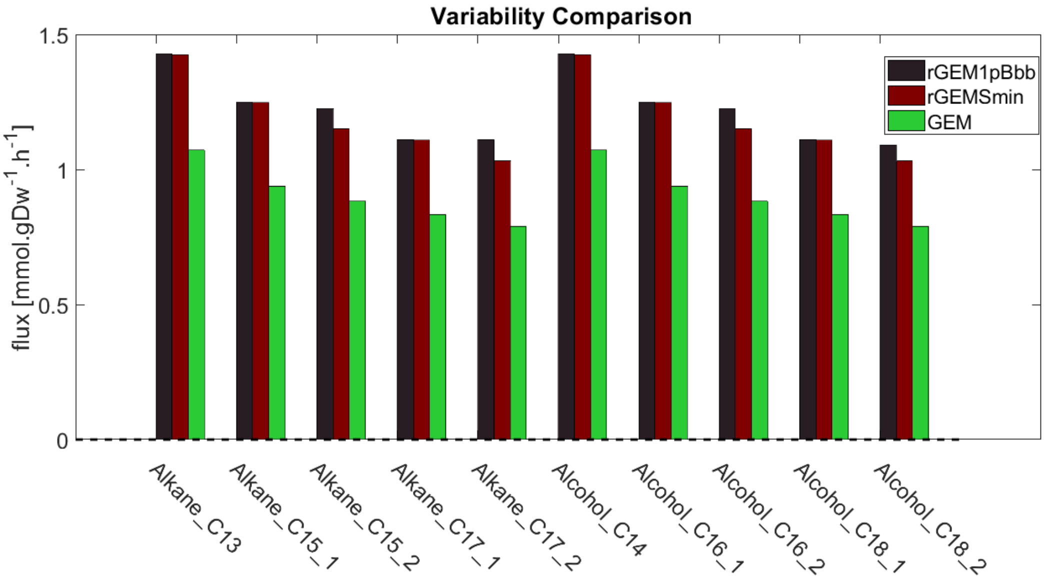


Figure 19 Flux variability comparison between the original model and the two reduced.

Table 10: The maximum attainable fluxes and Yields for all the fatty-acid derived products of interest. The calculations are made for glucose as sole carbon source and glucose uptake rate: 5 mmol·gDCW^{-1}·h^{-1}. The calculations are made for the Smin reduced GEM.

Product	Maximum production flux(mmol·gDCW^{-1}·h^{-1})	$Y_{P/S}$ (mol P/mol S)	$Y_{P/S}$ (g P/g S)
n-Tridecane	1,425	0,285	0,292
n-Pentadecane	1,247	0,249	0,294
1-Pentadecene	1,151	0,230	0,269
n-Heptadecane	1,108	0,222	0,296
1-Heptadecene	1,032	0,206	0,273
1-Tetradecanol	1,425	0,285	0,339
1-Hexadecanol	1,247	0,249	0,335
Hexadecenol	1,151	0,230	0,307
1-Octadecanol	1,108	0,222	0,332
Octadecenol	1,032	0,206	0,307

The Smin model contains 41 sets of lumped reactions resulting in the production of 41 BBBs. The number of alternative minimum sized reaction subsystems that result in a building bloc, as well as, the number of unique lumped reactions (Table 11) varies between the BBBs. For biomass components like mannan (Smin=6) and glucogen (Smin=4) there is only one unique lumped reaction identified for their production while for the amino acid isoleucine (ile-L) there are 24 unique alternatives (Smin=17) .

Table 11: Number of alternative lumped reactions added in the reduced model

Biomass Building Block	**Number of alternative lumped reactions**
13BDglcn, glucogen, mannan, pa_SC, pc_SC, pe_SC, ps_SC, ptd1ino_S, ribflv, tre	**1**
ala-L, asp-L, cys-L, ergst, glu-L, gly, phe-L, ser-L, val-L	**2**
gmp	**3**
amp, asn-L, gln-L, leu-L, met-L, pro-L, thr-L, triglyc_SC, trp-L, tyr-L	**4**
dtmp	**5**
dcmp	**7**
cmp, his-L, lys-L, ump	**8**
damp, dgmp	**9**
arg-L	**12**
zymst	**18**
ile-L	**24**

5.2.3 Smin reduced model consistency checks

In order to investigate whether the reduced model is consistent with the original, we compared the flux variability of reactions belonging to different subsystems after the thermodynamic constraints were imposed. In Figure 20 we observe that the allowed variability of some key enzymatic reactions of Glycolysis appear to be almost identical and slightly more constrained for the reduced model (TDH), while other reactions such as FBA exhibit significant differences. We have to denote that while most of the reactions participating in the fatty acid synthesis and degradation subsystems follow the same pattern between the two models, the maximum achievable flux in the case of the reduced model is considerably lower. A reduced model that contains further subsystems and a higher degree of connectivity is expected to yield in a more consistent version that represents better the predictive capacity of the original GEM.

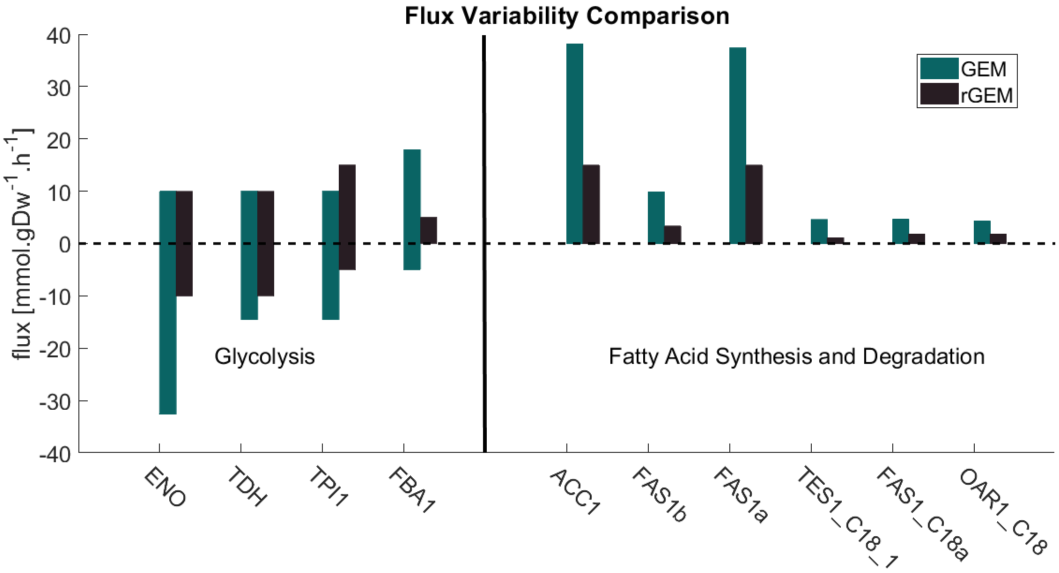


Figure 20: Flux variability comparison for key reactions of the Glycolysis (right) and Fatty acid synthesis and degradation (left) between the original and the reduced model.

We identified 176 (Table 12) common genes between the original and the reduced model, with each gene corresponding to one or more reactions. We performed single-gene deletions to both models and compared the results. As it is expected the lethal single-gene knock-outs appear to be more in the case of the reduced model since there are less rewiring alternatives. Glycolysis is the only refueling alternative for the reduced model, explaining why eight single-gene deletions located in glycolysis subsystem, appear to be lethal for the reduced model and not for the original GEM. We did not detect any false-positive single-gene deletion between the rest of the subsystems initially chosen for the reduction.

Table 12: Gene Essentiality check for the original GEM and the reduced model

Subsystem	**Number of lethal gene K.Os**	
	GEM	**rGEM**
Fatty Acid Biosynthesis	3	3
Fatty Acid Degradation	4	4
Fatty Acid Elongation	4	4
Glycolysis	0	8
Tyr, Trp and Phe Metabolism	0	1
Ox.Phosphorylation	0	1
Phospholipid Biosynthesis	1	0
Transport/Other	3	7
Total	**15**	**28**

5.2.4 Flux variability comparison under Biomass and Product yield flux constraints

The two extreme cases exhibit, as it was expected, very different behaviors. Glycolysis is the only refueling alternative for the metabolic network in the reduced model case. Having the same glucose consumption in both cases explains why most glycolytic reactions appear to be so constrained. Enzymatic complexes catalyze many of the reactions participating in lipids metabolism. The same complexes catalyze the same reaction for lipids of different chain length. As it was expected, the synthesis, elongation and degradation reactions variability depends on the product we are maximizing each time. In that sense, when n-tridecane was set to be the main product of the overproducing strain, the reactions including lipid-related molecules of carbon chain size less or equal to 14 appear to reach higher ranges while for chain size greater than 14 are close to zero. The degradation for the overproducing strain follows the exactly opposite behavior. The biomass proliferating strain in all cases exhibits a similar behavior and the maximum attainable flux remains low.

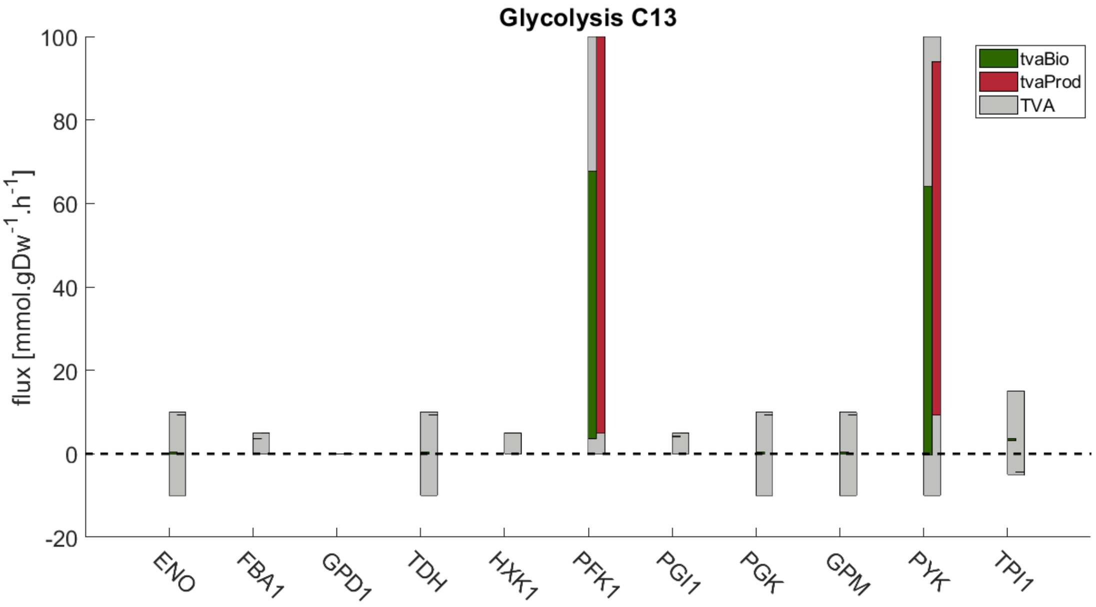


Figure 21: Glycolysis fluxes variability comparison for a n-tridecane overproducing strain(red) and a proliferating strain (green) the gray bar denotes the flux variability for the unconstrained model.

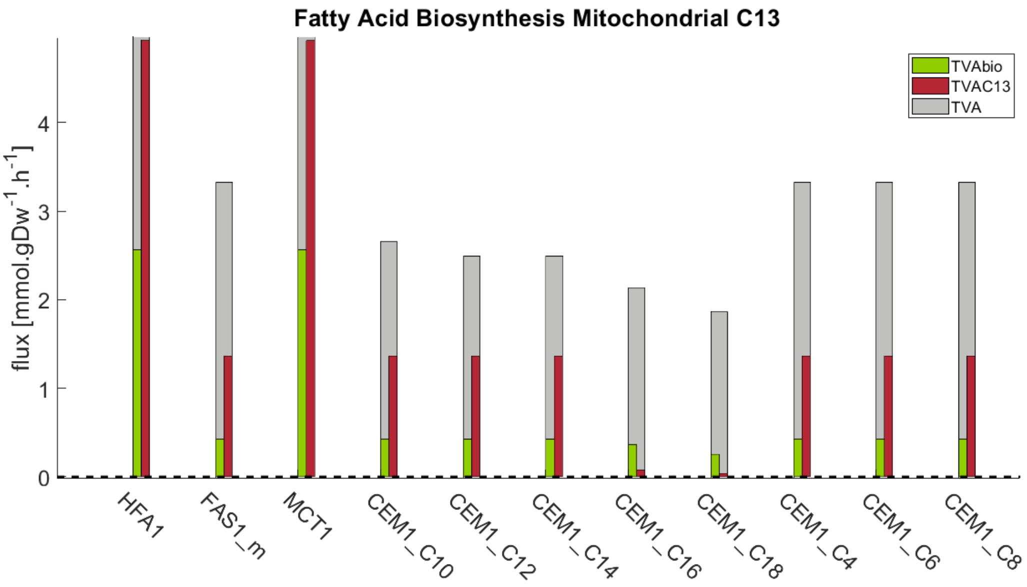


Figure 22: Mitochondrial Fatty acid biosynthesis fluxes variability comparison for a n-tridecane overproducing strain(red) and a proliferating strain (green) the gray bar denotes the flux variability for the unconstrained model.

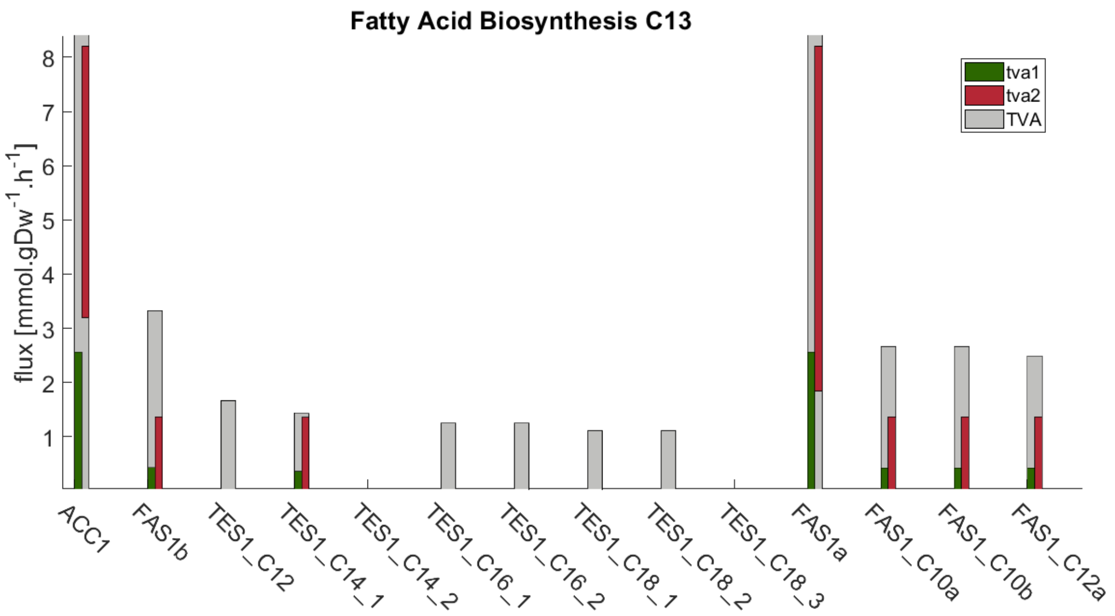


Figure 23: Fatty acid biosynthesis fluxes variability comparison for a n-tridecane overproducing strain(red) and a proliferating strain (green) the gray bar denotes the flux variability for the unconstrained model.

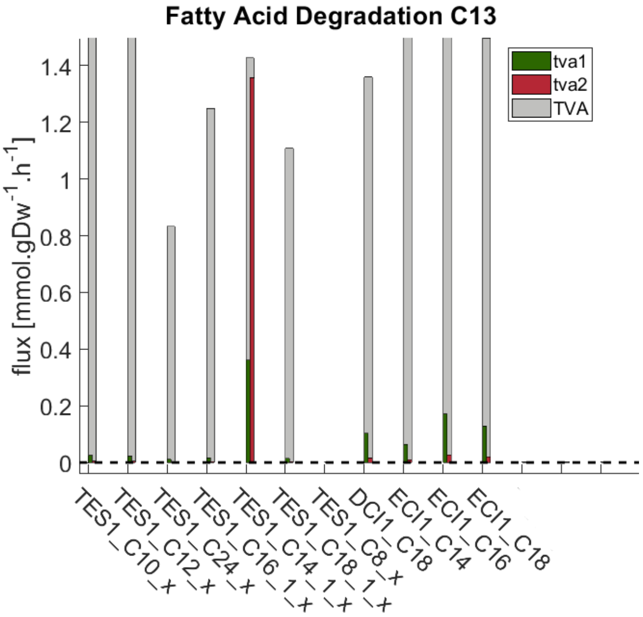


Figure 24: Degradation fluxes variability comparison for a n-tridecane overproducing strain(red) and a proliferating strain (green) the gray bar denotes the flux variability for the unconstrained model.

5.3 Incorporating economic factors onto the Genome-scale model

The production envelopes and pareto fronts for Growth rate versus Revenue and Growth rate versus profit, underline that there is a weak growth-coupling between the cellular growth and the hydrocarbons, the Total insoluble product and Total soluble product fluxes. That means that in the maximum attainable value of the biomass reaction there is de facto production of one or more chemicals belonging to the aforementioned categories. The big difference in potential revenues between the three cases, is explained by the different prices assigned to the products. The prices work as weights on the engineering objective function. The selected soluble products, as well as, the alcohol mixtures correspond to specialty chemicals and consequently the prices assigned to them are far greater than the fuel price. While the average kerosene market price is 520 $/T the market price for high purity fatty alcohol mixtures can be almost triple and double for carboxylic acids purposed for food and industrial applications. The results suggest that the profit margin is greater when the cell factory is reprogrammed to produce specialty chemicals.

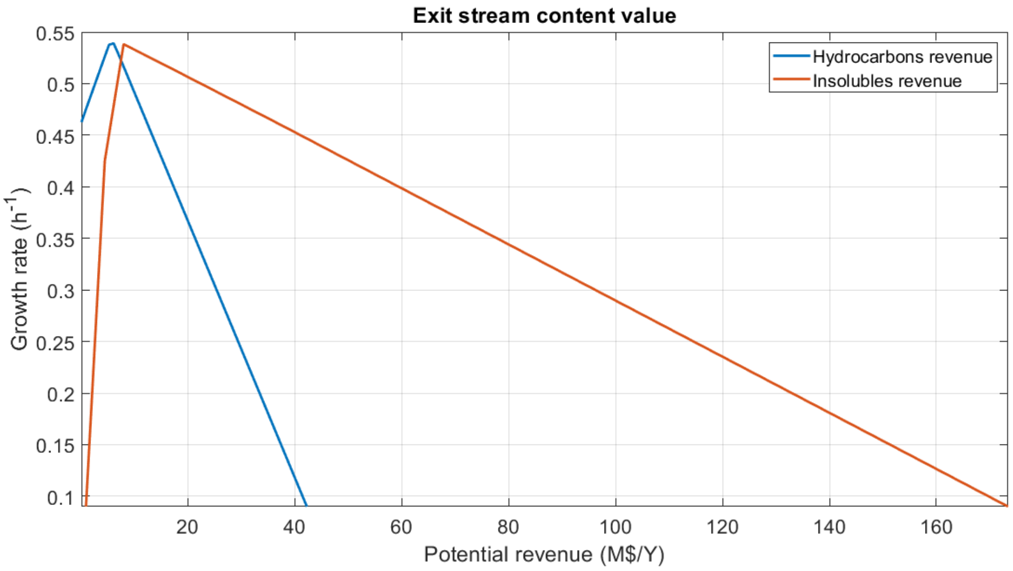


Figure 25: Achieved biomass for different allowed values of potential revenue. The maximum revenue achieved by hydrocarbons production is approximated to 45M$/Y and for the total insoluble stream almost 180M$/Y for almost zero corresponding growth rate. The data shown correspond to the reduced model prior to the thermodynamic constrain- this explains the different shape of the coupling-pitch observed at maximum growth.

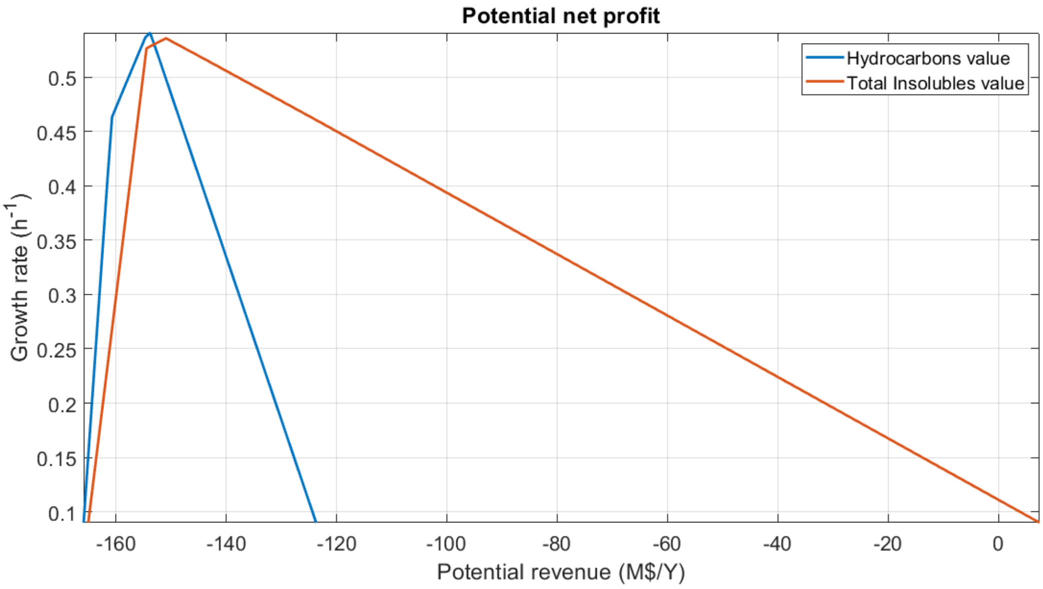


Figure 26: The Upstream process potential net profit for the hydrocarbons stream and the total insoluble products. The suggested bio-kerosene producing process appears unable to break-even while the optimistic scenario for the total insoluble producing process breaks even for dilution rates D=0.1 h^{-1} and leaves a margin of 10 M$ for the downstream process for very high retention times.

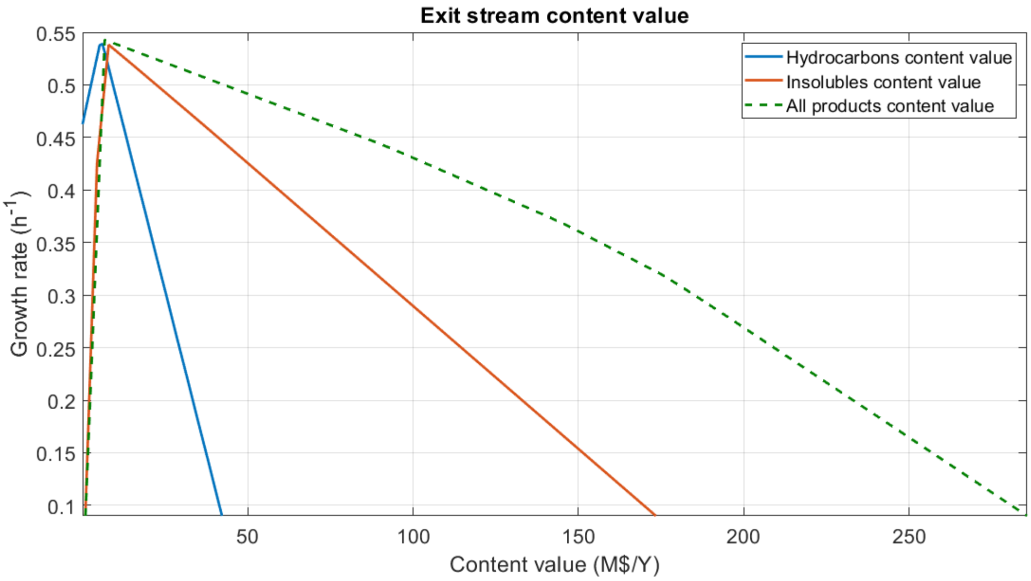


Figure 27:The potential revenue for the hydrocarbons, insolubles and all the products. If all the product portofolio is exploited, the maximum potential revenue can be up to 280M$/Y

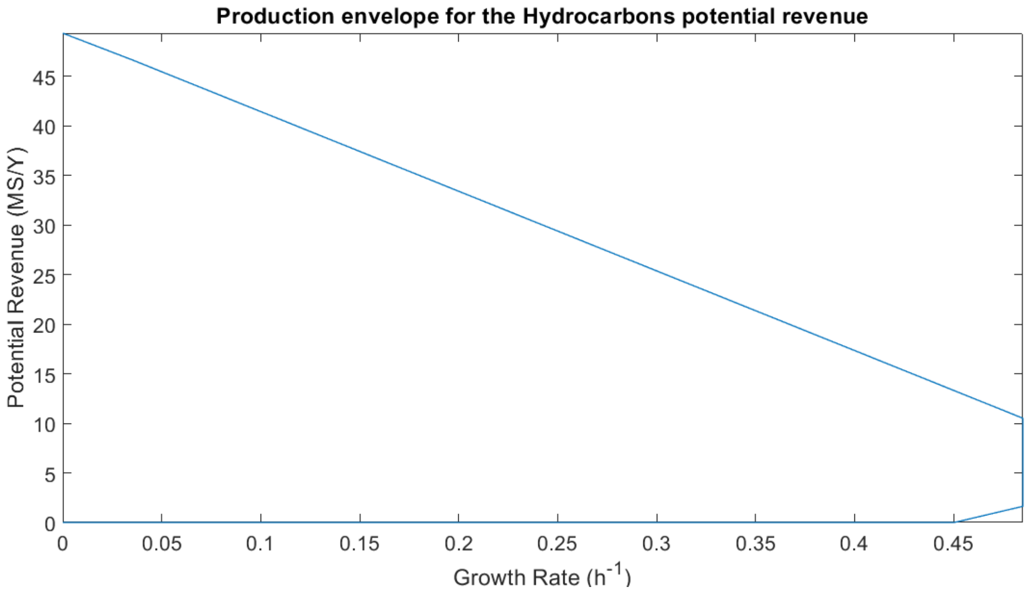


Figure 28: Production envelope for the hydrocarbons potential revenue. The total hydrocarbon flux exhibits weak growth coupling. Meaning that for zero biomass production, the organic flux does not necessarily attains non-zero values, while for maximum biomass production the total organic flux obtains non-zero values. The production envelop corresponds to the reduced model after the thermodynamic constraints are imposed.

The hydrocarbons' potential revenue exhibits a weak growth-coupling behaviour meaning that for zero biomass production the flux comprising of all the hydrocarbons in the model does not have to be non-zero while for maximum biomass the fluxes' value is definitely greater than zero. In order to identify whether a single alkane or alcohol is behind this behaviour we plotted the production envelopes for the individual alkanes and alcohols as well as the total alkanes flux and the total alcohols flux (Figure 29). It appears that while the individual alkane production reactions are not growth coupled the total flux is. A possible explanation to that lies to the fact that alkanes comprise the main drain for AcCoA-derived molecules. In higher biomass production rates the alkane reactions are needed to balance the higher fatty acid production. The fatty-alcohol production fluxes do not exhibit the same behaviour, neither individually nor in total.

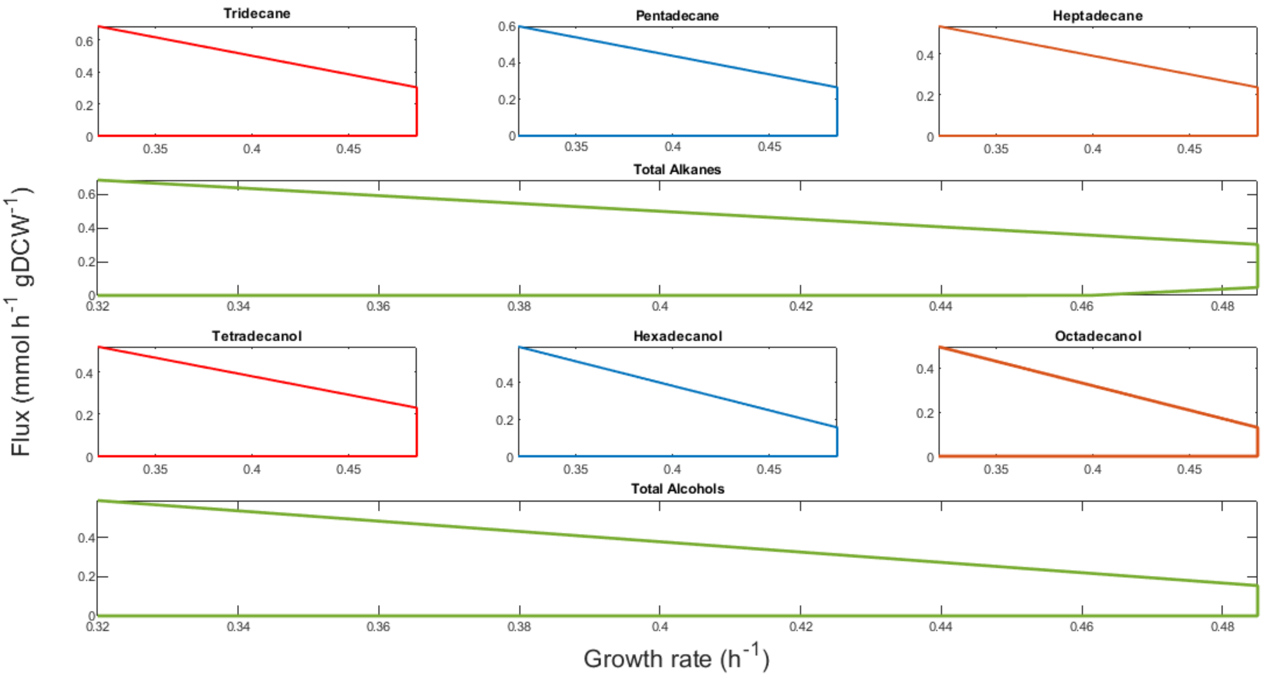


Figure 29: Comparative production envelopes for individual Alkanes, individual alcohols, Total Alkanes and total alcohol flux. While the individual reactions do not appear to be growth-coupled, the total alkane flux exhibits growth-coupled behaviour.

As showcased in the production envelope for the hydrocarbon stream potential revenue, for D=0.1 h^{-1} the revenue lies between the two extreme values of 42M$/Y and 0M$/Y. Attempting to identify key fluxes that differentiate between the two extreme states we conducted a flux variability analysis, where the fluxes variability of the upper extreme state (growth rate set to 0.1 and revenue to 42M$/Y) were compared against the lower extreme state (growth rate set to 0.1 and revenue to 0M$/Y) (Figure 30). The analysis identified that the FDH reaction catalysed by formate dehydrogenase appears to obtain its maximum values $v \in \{0.59, 1.13\}$ in the case of maximum revenue and lower-ranged values for zero revenue $v \in \{0.59, 0.43\}$.

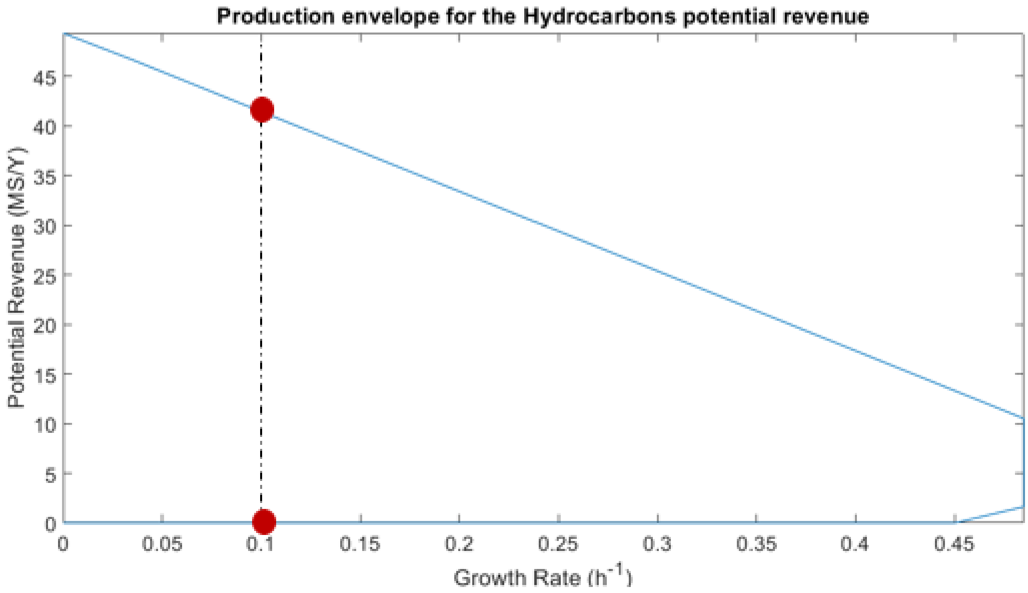


Figure 30:The two boundary revenue conditions for D=0.1h^{-1}.The upper limit is 42M$ while the lower is 0M$.

As an attempt to simulate an FDH upregulated strain (FDH↑) we will constrain the reaction to take values $v \in \{0.8,1.13\}$. While this strain is able to achieve lower maximum growth rates, the potential revenue from hydrocarbons, thus the hydrocarbon production, is strongly coupled with the cellular growth (Figure 31).

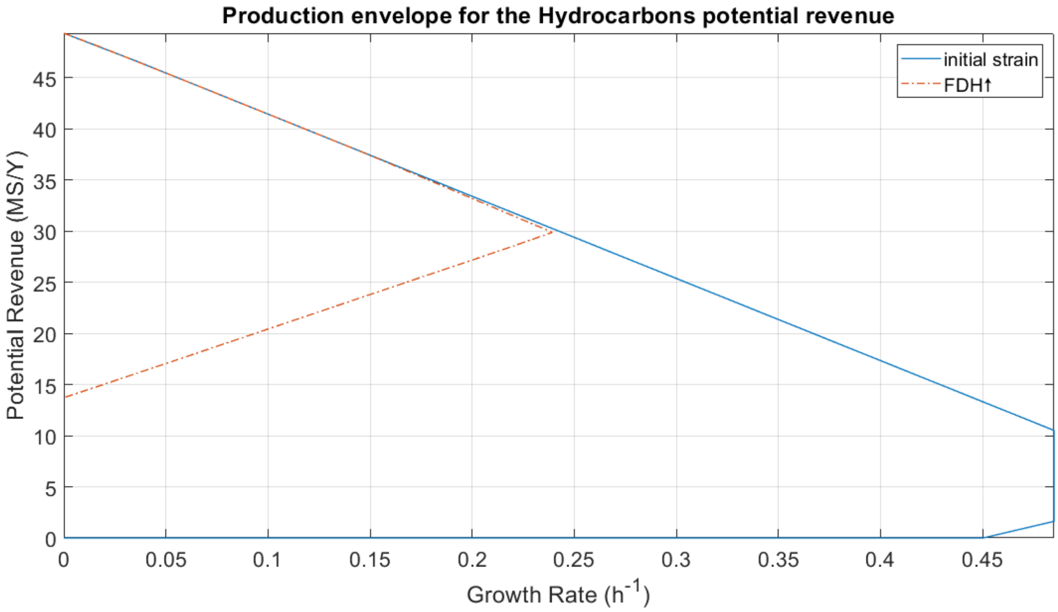


Figure 31: Production envelope for the hydrocarbons potential revenue for the initial strain and the FDH↑ strain.

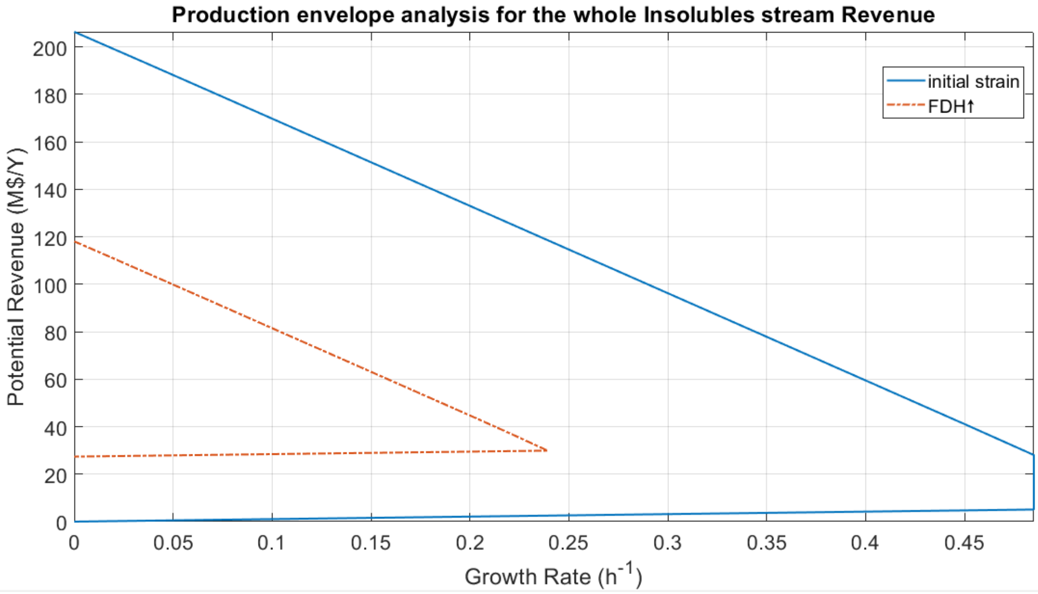


Figure 32: Production flux envelope for the total insoluble products' stream for the initial strain and the FDH↑ strain.

5.4 Strain design algorithm

The OptKnock-based strain design algorithm was used to identify reaction eliminations that would result in coupling of the growth rate with the insoluble stream potential revenue inside the Dilution rate working-region. Although the production of the total stream is already growth-coupled as it was showcased in the previous section, we shall target to strong growth coupling that would ensure higher product yields. The search was performed for the initial strain and the FDH↑ strain, for the chemostat D working region and for the complete allowed region of growth rates. The allowed number of eliminations varied from 1-20 and the alternative solutions obtained for each length was limited to three.

We were unable to obtain alternative metabolic strategies that yield in stronger growth coupling for both cases and for the total allowed region of growth rates. That might be explained from the reduced alternatives that the GEM under study entails. Another possible explanation could be the existence of pool reactions in the fatty acids subsystems that restrain eliminations that would yield in strong growth-coupling. The set of pool reactions is considered transports thus, it is excluded from the search. Furthermore, better results might be obtained for more thorough search comprising of larger size of allowed deletions and more iterations, while bearing in mind that a big number of deletions might be practically infeasible.

The inclusion of the design algorithm in the workflow procedure is incremental. The strains to be tested and optimized with ALE in the lab have to exhibit growth-coupled behaviour for the production of the desired chemical. The suggested reaction eliminations suggested by the algorithm, even though do not yield in stronger coupling or improved yield will serve as a pool of potential metabolic strategies.

We will test whether this pool of suggested eliminations will have any impact on the downstream cost. The reaction elimination radically changes the solution space leading to possible alterations in the exit stream's estimated composition.

Strategy No	Number of deletions	FDH ↑	Deletions
0	0	No	Initial strain
1	1	No	SeFADO17_2_x
2	1	No	POT1_C10
3	1	No	POT1_C12
4	2	No	POT1_C10, POT1_C12
5	2	No	SeFADO17_1_x, SeFADO17_2_x
6	2	No	POT1_C10, POT1_C14
7	3	No	POT1_C10, POT1_C12, POT1_C14
8	3	No	POT1_C10, POT1_C12, POT1_C16
9	3	No	POT1_C10, POT1_C12, POT1_C18
10	4	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16
11	4	No	POT1_C10, POT1_C12, POT1_C14, POT1_C18
12	4	No	POT1_C10, POT1_C12,POT1_C14, HFA1
13	5	No	POT1_C12, HFD1, FOX2_a, FOX_2b, POX_C12
14	5	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18
15	5	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16, HFA1
16	7	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16,POT1_C18, HFA1, POX1_C10
17	7	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, HFA1, POX1_C12
18	7	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, POX1_C10, POX1_C12
19	8	No	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, HFA1, POX1_C10, POX1_C12
20	1	FDH ↑	SeFADO17_2_x
21	1	FDH ↑	SeFADO17_1_x
22	1	FDH ↑	POT1_C10
23	3	FDH ↑	SeFADO15_2_x, SeFADO17_1_x, SeFADO17_2_x
24	3	FDH ↑	POT1_C12, NADH2-u6cm, SeFADO17_2_x
25	3	FDH ↑	POT1_C10, SUCD2_u6m, FAS1_C8a
26	5	FDH ↑	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18
27	5	FDH ↑	POT1_C10, POT1_C12, POT1_C14, POT1_C16, HFA1
28	5	FDH ↑	POT1_C10, POT1_C12,POT1_C14, POT1_C16, POX1_C10
29	7	FDH ↑	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, HFA1, POX1_C10
30	7	FDH ↑	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, HFA1, POX1_C12
31	7	FDH ↑	POT1_C10, POT1_C12, POT1_C14, POT1_C16, POT1_C18, HFA1, POX1_C14_1
32	0	FDH ↑	

The POT1 is an Acetyl-CoA acyl transferase enzyme, located in yeast peroxisome and participates in beta- oxidation process (fatty acids degradation). The enzyme catalyses the degradation of a 3-oxoacyl- CoA molecule to acetyl-CoA and the according acyl-CoA molecule. Given that beta-oxidation is a competing pathway to the heterologous hydrocarbons producing pathways, since both have fatty acids as a starting point, disruption of beta-oxidation process might be proven useful for yield and titer maximization. Since the enzyme's substrates vary in chain-length, while the proposed strategies indicate elimination of chain-length-specified reactions, a possible realization of such strategies would include enzyme engineering of POT1 to exhibit low affinity towards the chain-lengths we shall avoid.

The POX1 enzyme participates, as well, in beta-oxidation further oxidizing acyl-CoA compounds. HFA1 enzyme, located in mitochondria, is responsible for the mitochondrial conversion of Acetyl-CoA to Malonyl-CoA.

5.5 GEM sampling

The sampling procedure was followed for the total pool of metabolic strategies applied in the initial strain and the FDH↑ strain. Here, for brevity, we will present and compare the sampling results for the initial strain and the FDH↑ strain. First, we will show the potential revenue of the insoluble products' stream and the average kerosene composition in the case of the initial strain. Secondly, we will follow the same procedure for the FDH↑ strain and present a brief comparison between the individual product fluxes in the two cases.

For a dilution rate D=0.09-0.11 h^{-1} the approximated potential revenue for the initial strain is 15 M$/Y(Figure 33). The average kerosene based on the fluxes distribution for the various samples

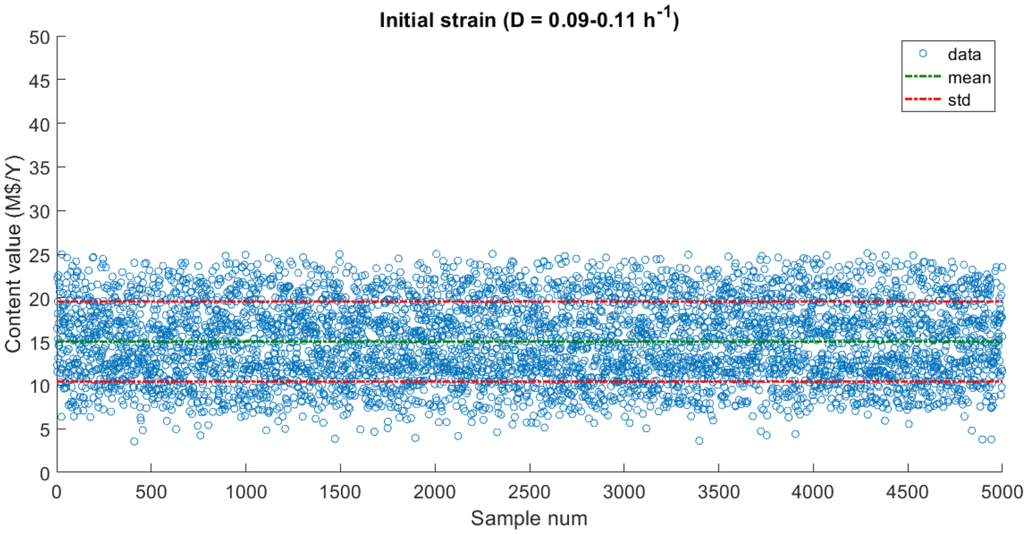


Figure 33:Potential insoluble products' revenue for different samples of the initial strain. The mean revenue is approximately 15 M$/Y while the std is 4.9.

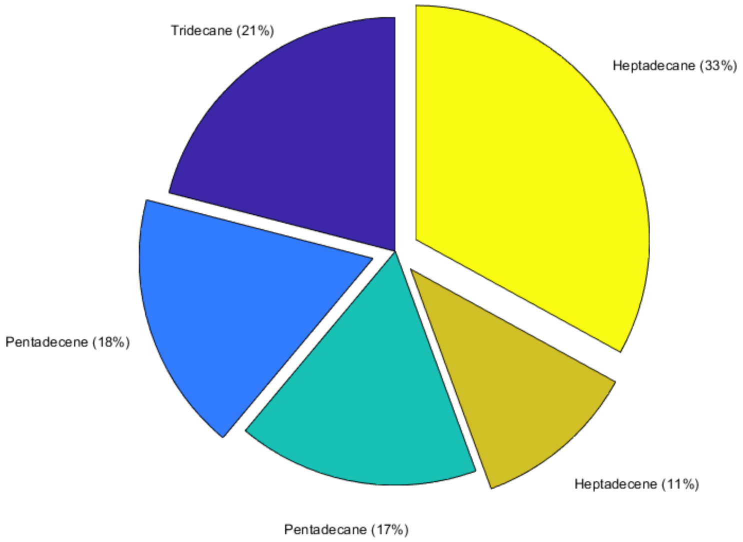


Figure 34:Kerosene composition (initial strain). Heptadecane appears to be the most abundant component of the kerosene produced from the initial strain, accounting for the 33% of the total stream. The rest of the components: Tridecane(21%), Pentadecene(18%), Pentadecane(17%), Heptadecene (11%).

For a dilution rate D=0.09-0.11 h^{-1} the approximated potential revenue for the FDH↑ strain is 36 M$/Y (Figure 35). The average kerosene based on the fluxes distribution for the various samples mostly consists of tridecane (82%), n-pentadecane comes second accounting for 16% of the total kerosene product while n-heptadecane and the saturated hydrocarbons account for the rest 2%.

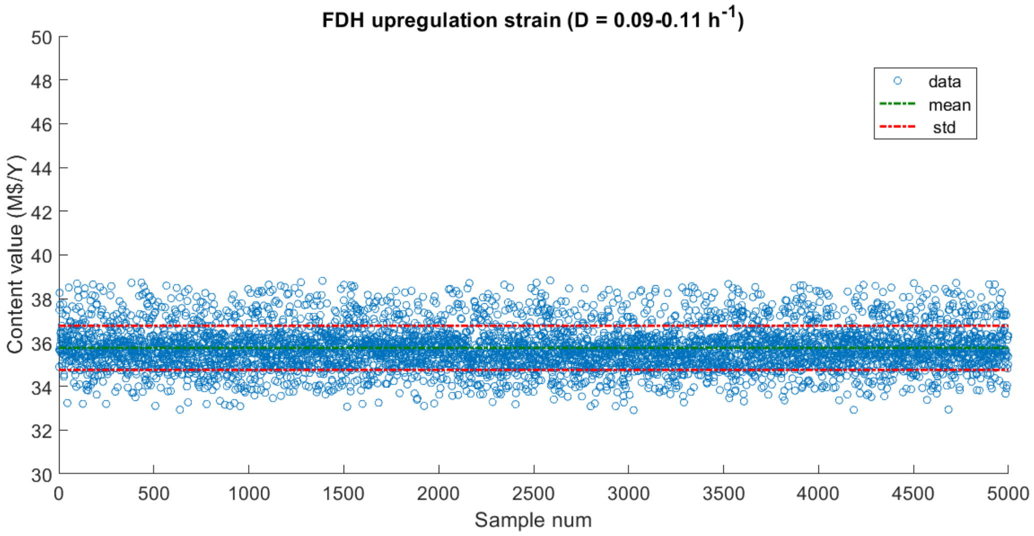


Figure 35:Potential insoluble products' revenue for different samples of the FDH upregulated strain. The mean revenue is approximately 36 M$/Y while the std is 1.1.

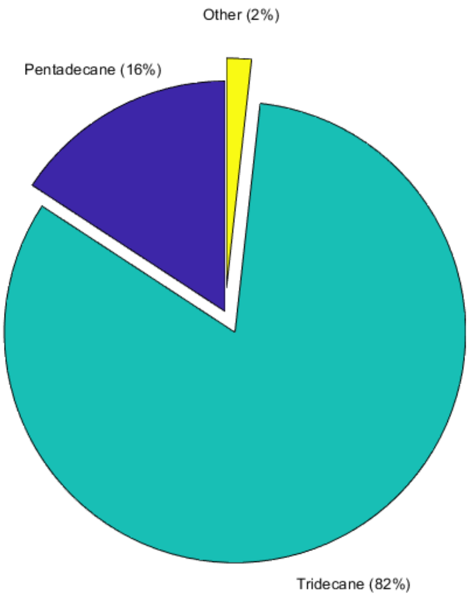


Figure 36: Kerosene composition (FDH↑). Tridecane covers 82% of the total product. Pentadecane comes second in abundancy with 16%.

As estimated by the sampling procedure, FDH↑ appears to double the potential revenue while radically altering the profile of the produced kerosene. Tridecane is favoured over heptadecane while pentadecane, in both cases, remains in comparable levels. As it appears from the probability-flux plot (Figure 37), apart kerosene related products (1 and 3) the rest appear to drop.

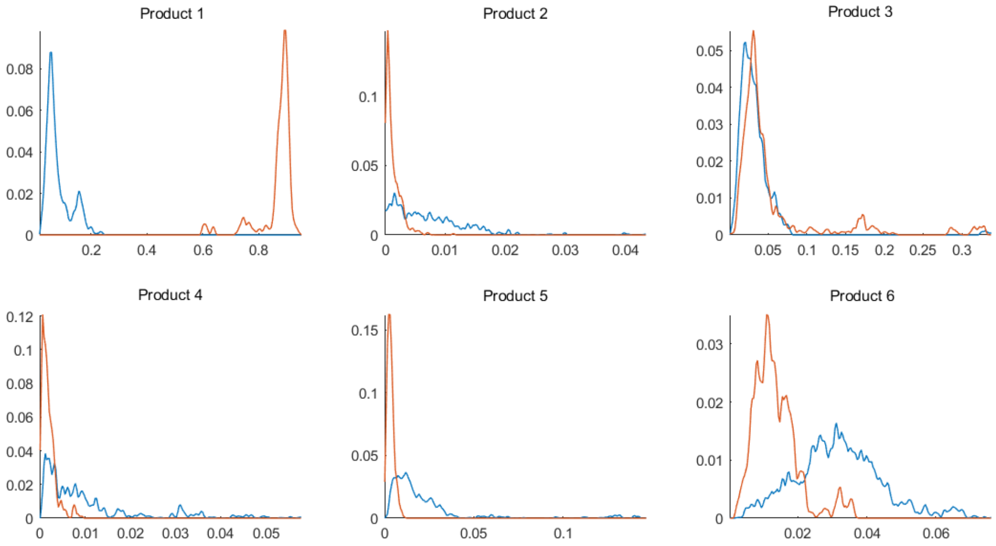


Figure 37 : Probability-Flux diagrams for the 6 distillation products

5.6 Minimum downstream cost estimation

In this section, we will present the preparatory calculations for the sequencing problem, including the task columns basic characteristics. Based on the column characteristics, we will proceed with the annual CAPEX estimations. Furthermore, we will showcase that the sequence identified as optimal using the conceptual cost is consistent with the CAPEX estimations for all the alternative separation routes. The calculations presented here refer to the initial strain without any further modifications.

5.6.1 Downstream cost estimation for the initial strain

Parameters estimation

The samples were clustered in 15 groups using the k-means algorithm as described in the sampling procedure. The calculations for the Molar flow are based in the upstream process presented, for a 994 T/day glucose supply (240 g/L). For each alternative metabolic strategy cluster set, we calculate the molar fraction of the stream components and the total molar flow.

Table 13: Exit stream composition for the various clusters

Product	**Cluster No.(molar fraction)**														
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***
A	0,133	0,475	0,052	0,065	0,216	0,067	0,196	0,219	0,049	0,306	0,050	0,159	0,253	0,106	0,029
	0,501	0,063	0,129	0,029	0,080	0,074	0,071	0,087	0,045	0,085	0,015	0,089	0,045	0,033	0,030
	0,053	0,054	0,101	0,045	0,062	0,116	0,076	0,057	0,144	0,114	0,127	0,152	0,035	0,083	0,065
B	0,027	0,036	0,104	0,009	0,021	0,027	0,125	0,040	0,053	0,042	0,031	0,033	0,324	0,047	0,019
C	0,029	0,033	0,056	0,036	0,275	0,104	0,064	0,037	0,025	0,055	0,041	0,062	0,063	0,030	0,012
	0,042	0,065	0,122	0,715	0,054	0,259	0,076	0,243	0,414	0,094	0,537	0,104	0,080	0,390	0,666
D	0,037	0,049	0,072	0,011	0,031	0,070	0,074	0,034	0,024	0,054	0,023	0,059	0,056	0,042	0,020
E	0,066	0,105	0,126	0,044	0,076	0,104	0,109	0,136	0,099	0,079	0,073	0,147	0,030	0,101	0,072
F	0,112	0,121	0,239	0,045	0,184	0,180	0,210	0,148	0,148	0,172	0,104	0,195	0,114	0,168	0,086
Molar Flow (kmol/h)	**12,223**	**10,024**	**7,128**	**18,777**	**8,607**	**9,344**	**6,662**	**8,474**	**12,473**	**6,794**	**14,565**	**6,375**	**11,303**	**11,574**	**15,742**
Cluster Size	**165**	**265**	**686**	**150**	**321**	**198**	**483**	**180**	**249**	**632**	**202**	**831**	**68**	**282**	**288**

For the first cluster, after we estimate the top and bottom composition, based on the feed and the recovery matrix, we can first apply the Antoine equation to calculate the Top and Bottom temperatures and the components' volatilites and then calculate the columns' basic characteristics using the FUG shortcut methods. We assume that the components are 98% recovered to the product fraction they belong and the rest 2% is distributed to the neighbouring products.

Table 14: The product fraction calculations based on the recovery matrix. The results correspond to the task 1 (including the products ABSDEF) and the relative volatility is calculated for an A/B separation (1-tetradecanol is the Light-Key component).

i	Component	Xif	Recovery fractions in Product						Relative Volatility
			A	B	C	D	E	F	
1	n-tridecane	0,1334	0,98	0,02	0	0	0	0	2,0986
2	1-pentadecene	0,5012	0,98	0,02	0	0	0	0	1,0875
3	n-pentadecane	0,0529	0,98	0,02	0	0	0	0	1,0387
4	1-tetradecanol	0,0266	0,01	0,98	0,01	0	0	0	1,0000
5	1-heptadecene	0,0291	0	0,01	0,98	0,01	0	0	0,5576
6	n-heptadecane	0,0418	0	0,01	0,98	0,01	0	0	0,5345
7	1-hexadecanol	0,0371	0	0	0,01	0,98	0,01	0	0,3107
8	1-octadecanol	0,0664	0	0	0	0,01	0,98	0,01	0,0972
9	zymosterol	0,1117	0	0	0	0	0,02	0,98	0,0001
Product fractions (Xip):			0,6739	0,0405	0,0701	0,0377	0,0676	0,1101	
Feed flowrate			**12,2 kmol/h**						

After calculating the distillate and bottom compositions, we can calculate the Top and Bottom temperatures corresponding to the condenser and reboiler temperatures respectively (Figure 38).

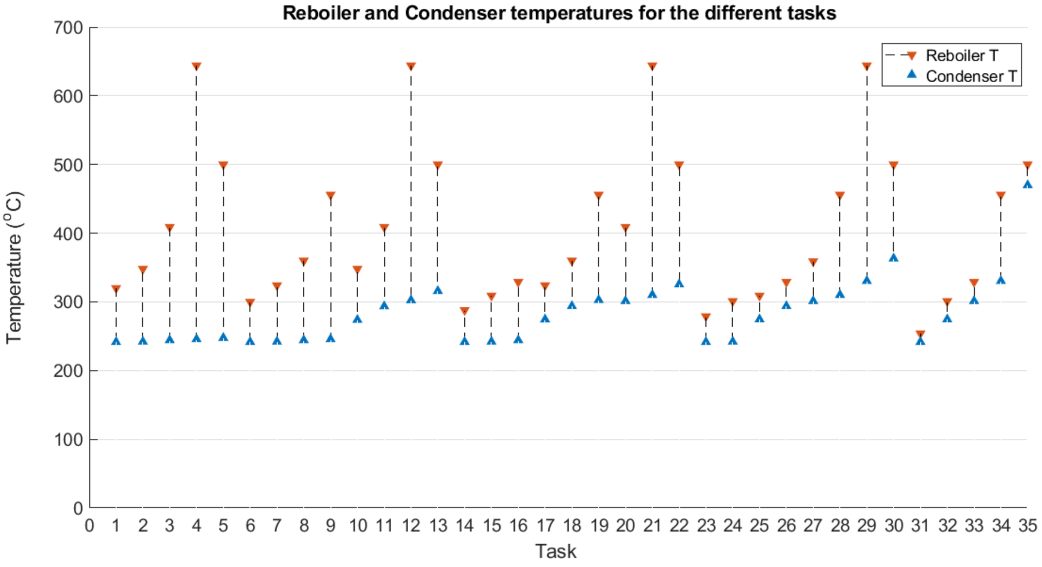


Figure 38: The reboiler and condenser temperatures as calculated for the 35 different tasks.

The shortcut calculations yield in the necessary parameters to 1) construct the sequencing problem and 2) estimate the annual venture cost.

Table 15: Estimated parameters for the sequencing problem construction. The reflux ratio corresponds to 1.1R_{min} as calculated by the second part of the Underwood equation.

Task	Theoretical Trays	Reflux ratio (R)	Vapour to feed ratio (V/F)	Cold utility cost index	Hot utility cost index
1	497	13,102	9,503	0,017	0,279
2	40	2,184	2,274	0,017	0,280
3	44	1,604	2,043	0,017	0,281
4	23	1,226	1,831	0,017	0,282
5	3	1,075	1,847	0,017	0,284
6	497	13,102	10,679	0,017	0,279
7	40	2,184	2,556	0,017	0,280
8	44	1,604	2,296	0,017	0,281
9	23	1,226	2,057	0,017	0,282
10	38	5,373	0,523	0,017	0,288
11	42	2,773	1,121	0,017	0,299
12	22	1,504	1,034	0,017	0,304
13	3	1,065	1,280	0,017	0,314
14	497	13,101	11,557	0,017	0,279
15	40	2,183	2,765	0,017	0,280
16	44	1,603	2,483	0,017	0,281
17	38	5,996	0,894	0,017	0,288
18	42	2,802	1,772	0,017	0,299
19	22	1,510	1,626	0,017	0,304
20	42	3,388	1,063	0,017	0,305
21	22	1,618	0,980	0,017	0,310
22	3	1,064	1,261	0,017	0,323
23	497	13,100	12,111	0,017	0,279
24	40	2,178	2,894	0,017	0,280
25	38	5,936	1,320	0,017	0,288
26	42	2,751	2,613	0,017	0,299
27	42	3,390	1,741	0,017	0,305
28	22	1,616	1,604	0,017	0,310
29	21	2,425	0,583	0,017	0,324
30	3	1,055	0,994	0,017	0,351
31	496	13,087	13,288	0,017	0,279
32	38	5,468	1,701	0,017	0,288
33	42	3,286	2,756	0,017	0,305
34	21	2,415	1,205	0,017	0,324
35	3	1,041	0,765	0,017	0,411

5.6.2 Distillation task sequence for the minimum annual venture cost

The optimization problem solved for the Cluster No.1, identified the sequence 1-10-20-29-35 as the one resulting in minimum total cost for a value TCOST= 3.996 M$/Y

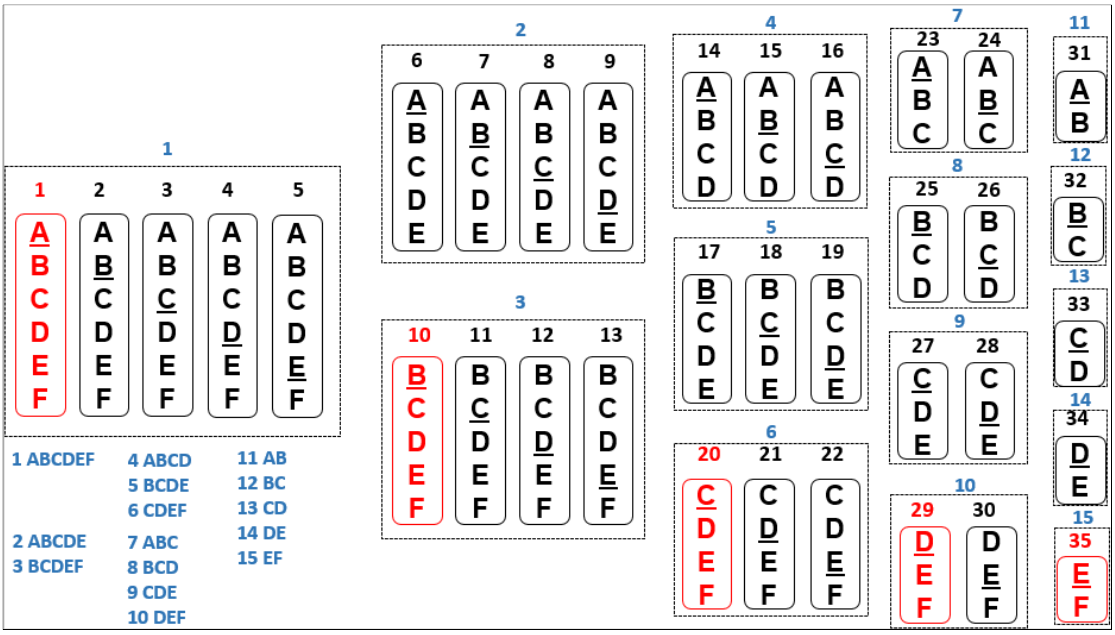


Figure 39: The identified task sequence that minimizes the total cost. With red the identified tasks 1-10-20-29-35 that result in minimum conceptual total cost TCOST= 3.996 M$/Y.

Table 16: Estimated cost variables from the conceptual cost model for the selected tasks.

Estimated variables	Task 1	Task 10	Task 20	Task 29	Task 35
$Flowrate\ (kmol/h)$	12,223	3,986	3,491	2,634	2,173
$Cost^{col}\ (K\$)$	10997,312	584,919	663,476	324,175	46,721
$Cost^{fix}(K\$)$	841,500	789,947	786,847	781,483	778,598
$Cost^{util}(K\$/Y)$	150,567	2,723	4,905	2,057	2,374
$Cost^{tot}(K\$/Y)$	2885,333	320,318	339,930	257,464	193,023

Table 17: Distillate composition for the identified as minimum cost tasks

Component	xd				
	Task 1	Task 10	Task 20	Task 29	Task 35
1	0,194	0,002	0,000	0,000	0,000
2	0,729	0,007	0,000	0,000	0,000
3	0,077	0,001	0,000	0,000	0,000
4	0,000	0,963	0,000	0,000	0,000
5	0,000	0,011	0,409	0,000	0,000
6	0,000	0,016	0,586	0,000	0,000
7	0,000	0,000	0,005	0,982	0,000
8	0,000	0,000	0,000	0,018	0,966
9	0,000	0,000	0,000	0,000	0,034

Table 18: Bottom composition for the identified as minimum cost tasks.

Component	xb				
	Task 1	Task 10	Task 20	Task 29	Task 35
1	0,008	0,000	0,000	0,000	0,000
2	0,031	0,000	0,000	0,000	0,000
3	0,003	0,000	0,000	0,000	0,000
4	0,081	0,001	0,000	0,000	0,000
5	0,089	0,101	0,001	0,000	0,000
6	0,128	0,145	0,002	0,000	0,000
7	0,114	0,130	0,170	0,002	0,000
8	0,204	0,232	0,308	0,370	0,006
9	0,342	0,391	0,518	0,628	0,994

Table 19: Basic design and operating characteristics for the selected columns.

Characteristic	Task 1	Task 10	Task 20	Task 29	Task 35
Number of Trays	497	38	42	21	3
Reflux ratio	13,102	5,373	3,388	2,425	1,041
Vapour to feed ratio	9,503	0,523	1,063	0,583	0,765
Treb (°C)	331	356	430	647	500
Tcond (°C)	260	274	301	331	470

In a similar manner, we estimate the optimal separation sequences and the annualized total cost for the rest of the clusters (Table 20). We observe that the prominent favored sequence for the total of the clusters is 1-10-20-29-35, which is the same as the one identified as optimal for the Cluster No. 1. Moreover, the clusters 6 and 15 differentiate from the prominent behavior since the identified as optimal sequence in these cases is 1-11-20-32-35. In Figure 40 we present the relative annualized cost results as for the cluster size.

Table 20: Complete cost optimization results for the initial strain clusters.

Cluster No.	Optimal sequence	Total Cost (k$/Y)
1	1-10-20-29-35	3996,066
2	1-10-20-29-35	2670,949
3	1-10-20-29-35	3537,728
4	1-10-20-29-35	2518,198
5	1-10-20-29-35	2669,436
6	1-11-29-32-35	3143,838
7	1-10-20-29-35	2737,717
8	1-10-20-29-35	2699,478
9	1-10-20-29-35	3087,186
10	1-10-20-29-35	2878,596
11	1-10-20-29-35	2859,987
12	1-10-20-29-35	3525,718
13	1-10-20-29-35	2365,802
14	1-10-20-29-35	2726,517
15	1-11-29-32-35	2747,717

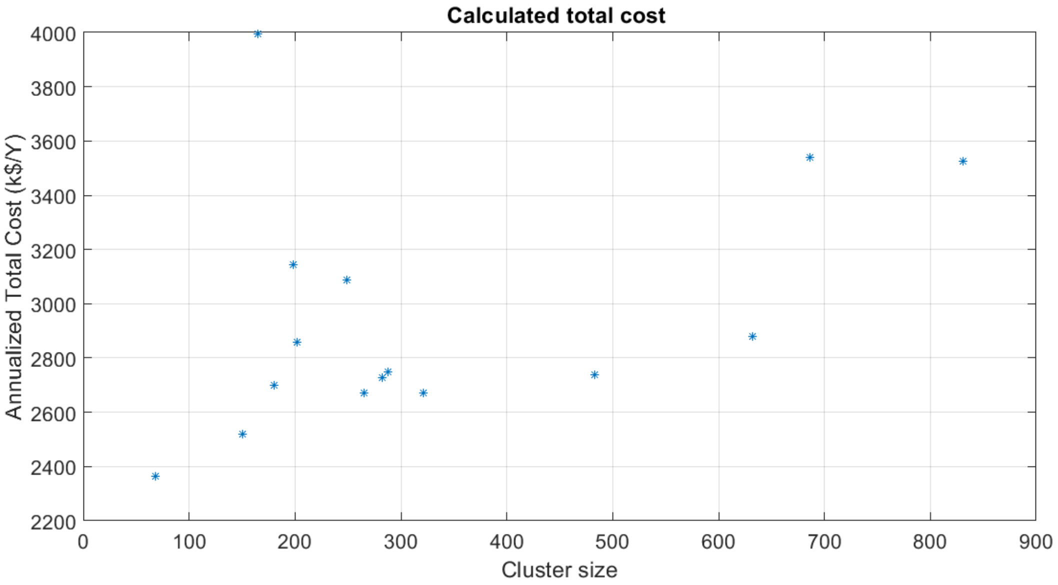


Figure 40: Calculated annualized total (conceptual) cost vs. cluster size.

5.6.3 Annualized capital cost estimation

In order to extract the average cost for the alternative engineered strains and compare the strategies we proceed with detailed capital cost estimations based on the cost model and assumptions found in Appendix. We assume that the utilities cost remain the same as calculated on the previous section.

Table 21: Capital cost estimation for the optimal sequence of the cluster No.1

Cost variable	Task 1	Task 10	Task 20	Task 29	Task 35
Distillation column annualized capital cost (M$/Y)	12,517	0,399	0,610	0,273	0,107
Distillation column trays and tower internals annualized capital cost (M$/Y)	26,186	0,262	0,479	0,164	0,051
Heat exchangers and furnace capital cost (M$/Y)	0,042	0,001	0,002	0,001	0,111
Distillation column total annualized CAPEX (M$/Y)	38,745	0,662	1,091	0,438	0,268
Utilities annual cost (k$/Y)	139,890	6,993	13,001	6,534	6,184
Total annualized CAPEX (M$/Y)	**77,630**	**1,330**	**2,195**	**0,883**	**0,542**

Table 22: Cost Estimation for the initial strain clusters

Cluster No.	Distillation column annualized capital cost (M$/Y)	Distillation column trays and tower internals annualized capital cost (M$/Y)	Heat exchangers and furnace capital cost (M$/Y)	Utilities annual cost (k$/Y)	Total annualized CAPEX (M$/Y)
1	13,906	27,142	0,156	0,173	41,376
2	5,642	6,893	0,170	0,068	12,773
3	9,299	14,970	0,110	0,084	24,463
4	6,342	7,560	0,173	0,137	14,213
5	5,439	6,471	0,127	0,060	12,097
6	8,311	12,165	0,125	0,095	20,697
7	5,700	7,232	0,120	0,058	13,110
8	5,596	6,778	0,144	0,064	12,582
9	11,334	16,642	0,324	0,203	28,503
10	6,003	7,786	0,126	0,059	13,974
11	8,501	12,399	0,148	0,135	21,182
12	8,362	12,813	0,124	0,070	21,369
13	6,176	7,927	0,134	0,116	14,354
14	6,843	9,042	0,140	0,092	16,117
15	7,383	9,692	0,152	0,129	17,356

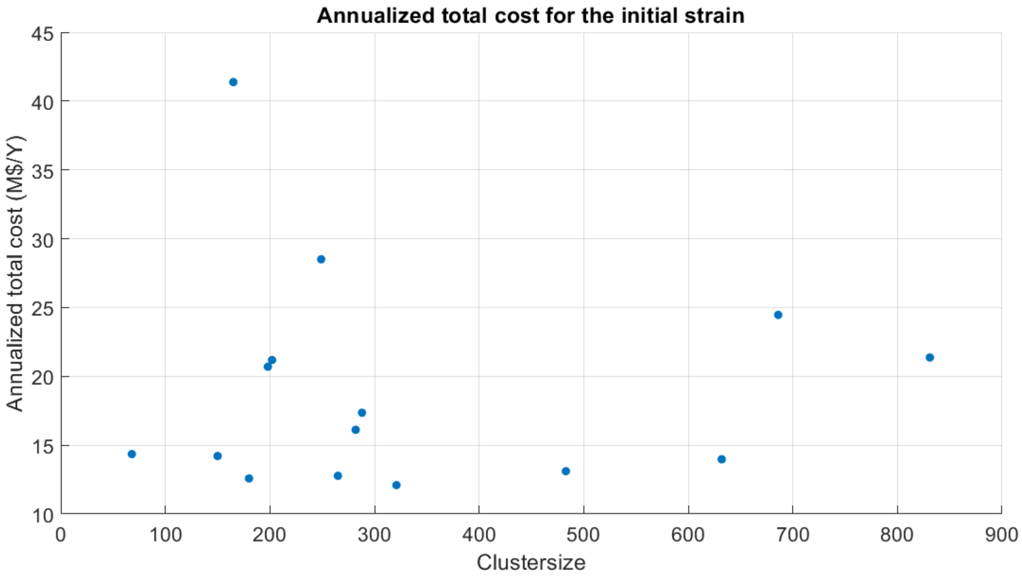


Figure 41: Estimated annualized total (real) cost vs. cluster size.

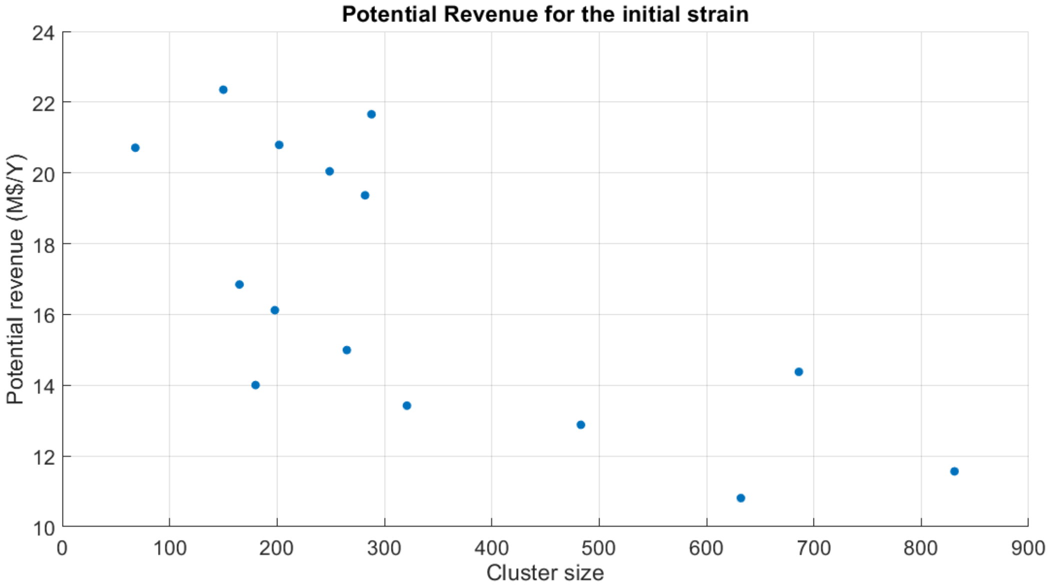


Figure 42: Estimated annual potential revenue vs. cluster size.

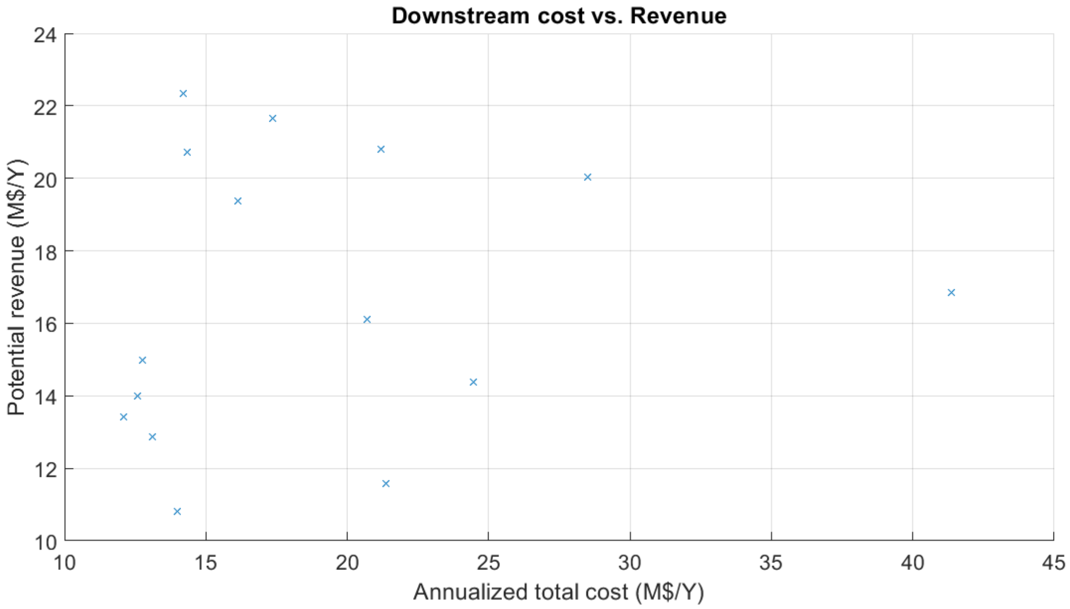


Figure 43: Comparative presentation of the annual cost demands versus the attainable revenue for the 15 clusters of the initial strain.

We proceed with calculating the annualized total downstream cost and total annual revenue for the initial strain:

$$\overline{TC}_{initial\ strain} = \sum_{k=1}^{15} \frac{W(k)}{n} \cdot TC(k) = 18{,}84 \pm 6{,}41\ M\$/Y$$

$$\overline{TR}_{initial\ strain} = \sum_{k=1}^{15} \frac{W(k)}{n} \cdot TR(k) = 14{,}99 \pm 3{,}65\ M\$/Y$$

5.6.4 Consistency check for the conceptual cost model

We will evaluate whether the distillation task sequence identified by the algorithm as optimal based on the conceptual cost value, is consistent with the actual estimated annual capital cost for the different alternative routes. For the 6-products distillation problem, there are 42 alternative identified routes (Table 23) that lead to full separation.

The consistency check was performed for the 15 feed streams corresponding to the initial strain (Table 13: Exit stream composition for the various clusters). We calculated the annual downstream cost for the 42 alternative routes and examined whether the identified as optimum from the sequencing optimization program holds the minimum value.

The total downstream cost calculations are in accordance with the optimization results for all the clusters except cluster 6 and 15 ; the sequence identified as the most cost-efficient by the program in both cases is the sequence 6 while the cost estimation underlines the sequence 5 as the one with the minimum cost. The program-identified sequence is 0.4 M$/Y and 0.57M$/Y greater than the identified optimal for the clusters 6 and 15 respectively.

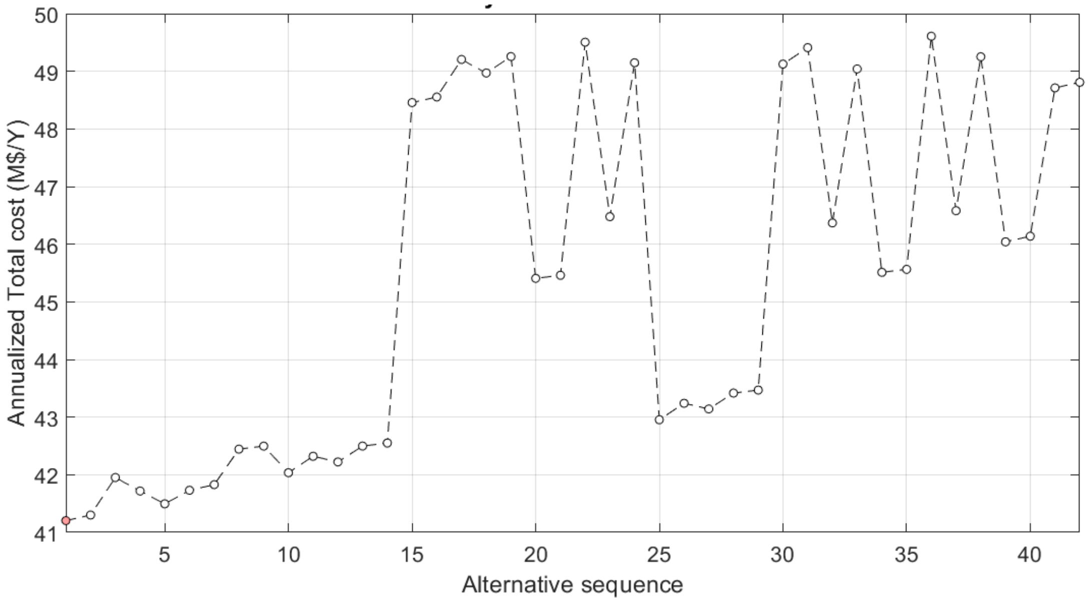


Figure 44:Consistency check for the cluster No 1. The identified as optimal sequence agrees with the estimated one

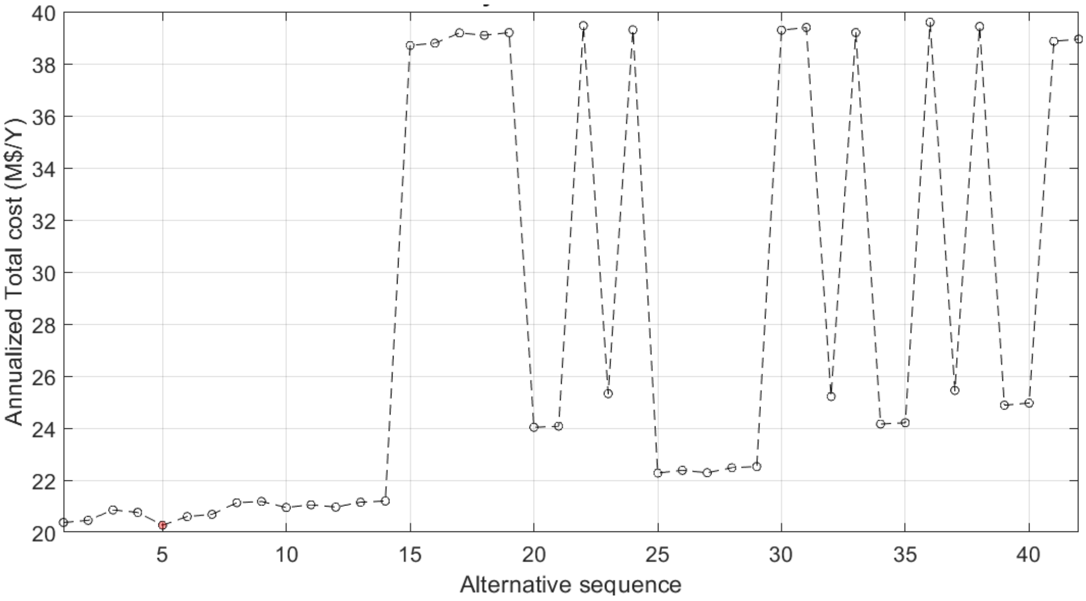


Figure 45: Consistency check for the cluster No 6. The identified as optimal sequence is not in accordance with the estimated one. The program identified sequence 6 as the optimal sequence while the cost estimations underline 5 is the one with the minimum cost. Anyhow, the difference between the two sequences can be assumed negligible.

Table 23: The alternative distillation task routes for a 6-product stream.

No	**Route**
1	**1 - 10 - 20 - 29 - 35**
2	**1 - 10 - 20 - 30 - 34**
3	**1 - 10 - 21 - 33 - 35**
4	**1 - 10 - 22 - 27 - 34**
5	**1 - 10 - 22 - 28 - 32**
6	**1 - 11 - 32 - 29 - 35**
7	**1 - 11 - 32 - 30 - 34**
8	**1 - 12 - 35 - 25 - 33**
9	**1 - 12 - 35 - 26 - 32**
10	**1 - 13 - 17 - 27 - 34**
11	**1 - 13 - 17 - 28 - 33**
12	**1 - 13 - 18 - 32 - 34**
13	**1 - 13 - 19 - 25 - 33**
14	**1 - 13 - 19 - 26 - 32**
15	**2 - 31 - 20 - 29 - 35**
16	**2 - 31 - 20 - 30 - 34**
17	**2 - 31 - 21 - 33 - 35**
18	**2 - 31 - 22 - 27 - 34**
19	**2 - 31 - 22 - 28 - 33**
20	**3 - 23 - 32 - 29 - 35**
21	**3 - 23 - 32 - 30 - 34**
22	**3 - 24 - 31 - 29 - 35**
23	**3 - 24 - 31 - 30 - 34**
24	**4 - 35 - 14 - 25 - 33**
25	**4 - 35 - 14 - 26 - 32**
26	**4 - 35 - 15 - 31 - 33**
27	**4 - 35 - 16 - 23 - 32**
28	**4 - 35 - 16 - 24 - 31**
29	**5 - 6 - 17 - 27 - 34**
30	**5 - 6 - 17 - 28 - 33**
31	**5 - 6 - 18 - 32 - 34**
32	**5 - 6 - 19 - 25 - 33**
33	**5 - 6 - 19 - 26 - 32**
34	**5 - 7 - 31 - 27 - 34**
35	**5 - 7 - 31 - 28 - 33**
36	**5 - 8 - 34 - 23 - 32**
37	**5 - 8 - 34 - 24 - 31**
38	**5 - 9 - 14 - 25 - 33**
39	**5 - 9 - 14 - 26 - 32**
40	**5 - 9 - 15 - 31 - 33**
41	**5 - 9 - 16 - 23 - 32**
42	**5 - 9 - 16 - 24 - 31**

5.6.5 Comparative results for the alternative metabolic strategies

Here we present the according results for the alternative tested metabolic strategies. We observe that the strains can be clustered in two distinct groups. The FDH↑ engineered strains exhibit double profit with respect to the initial strain constraining the upregulated strains at the far top of the diagram. Furthermore, we observe that the average cost to obtain the desired portfolio appears to be lower for the upregulated strains. A possible explanation could be the rewiring of the fluxes towards the lighter products (especially towards tridecane). After the first column that the light kerosene fraction is removed, the remaining stream of minor flowrates can be easily separated.

The different metabolic strategies applied to the initial strain without upregulation (bottom-right) do not appear to improve the separation cost, although in most cases the attainable potential revenue appears to reach higher levels than the initial strain.

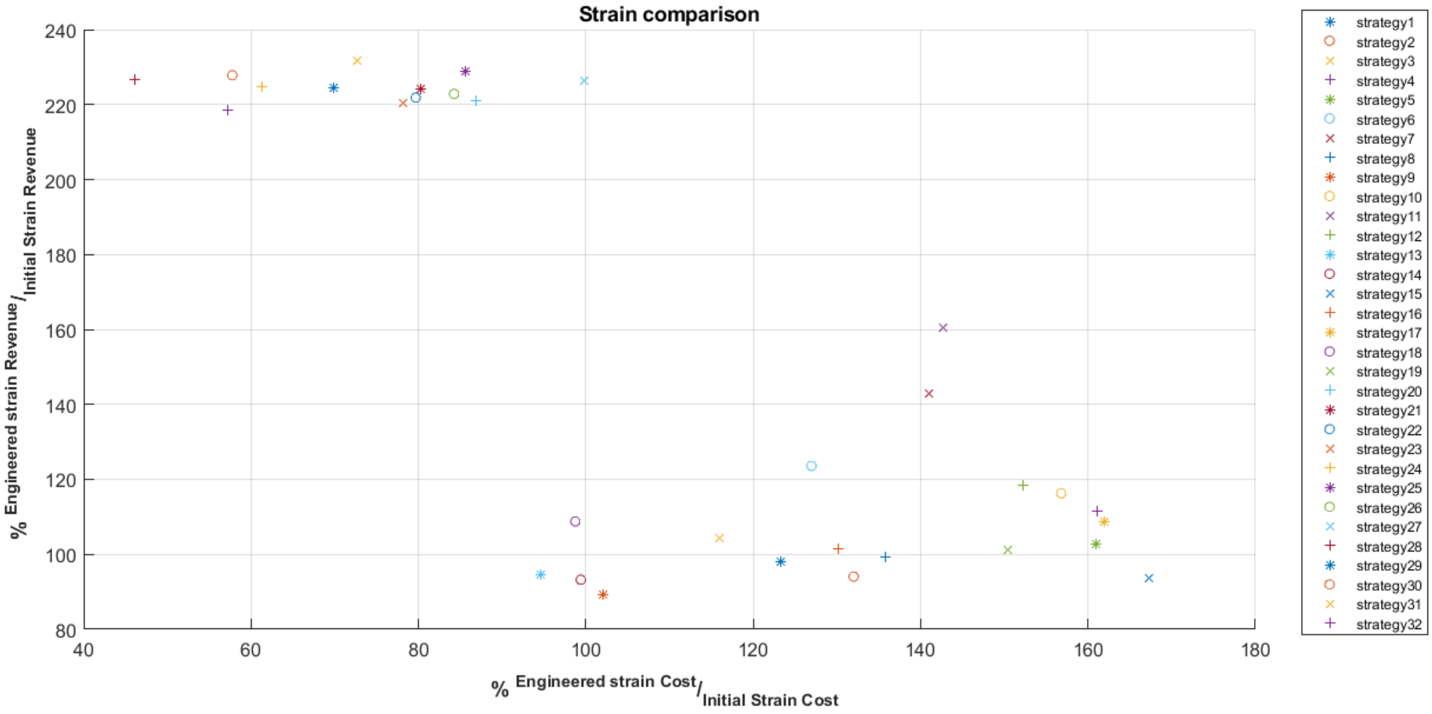


Figure 46: Comparing the Engineered strains estimated relative revenue and relative downstream cost with respect to the initial strain.

Table 24: Comparison of the mean downstream cost and mean revenue for the alternative engineered strains.

Strategy	Mean downstream cost	Mean revenue	Revenue - Cost	Efficiency
0	18,84 ± 6,41	14,99 ± 3,65	-3,85	0%
1	23,22 ± 11,35	14,71 ± 3,09	-8,51	-121%
2	24,87 ± 13,57	14,10 ± 3,65	-10,77	-180%
3	21,86 ± 11,15	15,65 ± 8,57	-6,22	-61%
4	30,34 ± 21,25	16,73 ± 4,13	-13,61	-253%
5	30,33 ± 17,30	15,42 ± 5,04	-14,91	-287%
6	23,92 ± 12,35	18,53 ± 2,78	-5,39	-40%
7	26,56 ± 13,82	21,44 ± 6,41	-5,11	-33%
8	25,59 ± 11,70	14,88 ± 3,01	-10,71	-178%
9	19,24 ± 9,13	13,39 ± 5,97	-5,85	-52%
10	29,54 ± 13,44	17,43 ± 5,73	-12,11	-214%
11	26,88 ± 32,27	24,04 ± 14,14	-2,84	26%
12	28,67 ± 14,35	17,77 ± 3,59	-10,90	-183%
13	17,81 ± 6,88	14,19 ± 4,07	-3,62	6%
14	18,73 ± 5,62	13,98 ± 3,16	-4,75	-23%
15	31,52 ± 20,95	14,04 ± 5,08	-17,49	-354%
16	24,52 ± 12,69	15,20 ± 5,56	-9,32	-142%
17	30,50 ± 20,23	16,29 ± 3,53	-14,21	-269%
18	18,60 ± 6,14	16,30 ± 4,87	-2,30	40%
19	28,35 ± 16,94	15,16 ± 4,29	-13,18	-242%
20	16,36 ± 4,84	33,15 ± 1,06	16,79	536%
21	15,14 ± 5,28	33,60 ± 1,77	18,46	579%
22	15,01 ± 5,05	33,26 ± 0,95	18,25	574%
23	14,72 ± 6,49	33,03 ± 1,34	18,31	576%
24	11,55 ± 5,46	33,72 ± 2,60	22,16	676%
25	16,13 ± 5,27	34,30 ± 1,98	18,17	572%

26	15,87 ± 5,46	33,40 ± 0,64	17,53	555%
27	18,81 ± 5,63	33,92 ± 1,03	15,11	492%
28	8,69 ± 2,79	33,98 ± 1,23	25,29	757%
29	13,16 ± 6,24	33,67 ± 0,63	20,51	633%
30	10,88 ± 5,68	34,15 ± 1,71	23,27	704%
31	13,70 ± 6,19	34,74 ± 1,29	21,04	646%
32	10,77 ± 5,06	32,74 ± 0,88	21,96	670%

The efficiency comparison among the alternative metabolic strategies is based on the difference between the mean revenue and the mean downstream cost. Although the feedstock annual price was proven to be the highest among the annual expenses (apx. 150 M$/Y), it is assumed equal for the different cases. A novel upstream future process may provide glucose or alternative carbon sources from non-food biomass to much lower prices, enabling the development of a microbial cell factory like the one described. Indicatively, if the glucose comes from the NREL upstream process the feedstock cost may almost to half[34]. The present study aimed to identify strategies that reduce the separation cost while preserving high product revenues as a result the pseudo-profit value (Profit = Mean Revenue – Mean Downstream Cost) is an accurate measure to compare the alternative strategies.

We observe that the strategies 11, 13, 18 and 20-32 appear to be improved with respect to the initial strain. The first three strategies correspond accordingly to 4, 5 and 7 reaction eliminations without FDH upregulation, while the rest correspond to the total of the FDH↑ strains for all the different lengths of reaction eliminations. The calculated standard deviation in the strains without upregulation appears to obtain greater values. This may be explained by the fact that the FDH flux in the case of the upregulated strains was constrained to obtain values in a specific smaller range, subsequently constraining the solution space.

The FDH↑ strains exhibit a 5-7 fold potential increase in profit, thus the upregulation of the fdh gene appears as a promising strategy towards sustainable kerosene biorefineries. The translation of the reaction elimination procedure to real-life laboratory techniques encloses some difficulties in the case of lipids metabolism since a single enzyme is responsible for the catalysis of different reactions. As a result, a simple gene knock-out approach is not applicable. Nevertheless, the identified strategies may work as enzyme engineering objectives. For example, the strategy No. 28 that appears to be the most promising (757% profit increase with respect to the initial strain) involves FDH↑ and POT1 reaction eliminations for substrates of chain length: C10, C12, C14, C16 and POX1 for C10. The two enzymes may be engineered to exhibit less affinity towards these substrates using directed evolution techniques. Alternatively, we could search for corresponding enzymes in different microorganisms, which exhibit properties closer to the desired and replace the native yeast enzymes. The same thinking can be followed for the rest of the strategies. We have to bear in mind though that certain strategies may not lead in viable strains. For instance, the disruption of the respiratory chain as indicated in the strategies 24 and 25 is potentially not applicable for *S.cerevisiae* growing in aerobic conditions. The complete pool of metabolic strategies shall be tested in the parent GEM context before proceeding with experimental verification.

Chapter 6. Conclusion and Future research

6.1 Bridging strain design with downstream process synthesis

In the present study, we achieved to propose a “bridge” between the strain design procedures typically followed during the Design module of the DBTL cycle with the downstream process synthesis and more specifically for the examined case study the distillation-sequencing problem. We displayed that alternative metabolic strategies may induce important changes in the fermentation broth attainable compositions that directly affect the downstream process expenditures necessary to obtain a specific product portfolio. In the following, we will present an overview of the main finding on each module of the proposed workflow. Furthermore, we will provide some insights.

6.1.1 Genome-scale model curation, reduction and analysis

We developed a GEM that corresponds to a mutant *S.cerevisiae* strain capable of producing kerosene-range hydrocarbons. The model was thermodynamically curated and reduced. The reduced model exhibits a maximum growth rate $\mu=0.4850\ h^{-1}$ under aerobic conditions and glucose as the only carbon source. The estimated maximum theoretical yields (mol Product/mol Glucose) for n-tridecane, n-pentadecane, 1-pentadecene, n-heptadecane, 1-heptadecene were 0.285, 0.249, 0.230, 0.222, 0.206 respectively.

The network changes are directly linked with the extracellular matrix. The extracellular matrix is closely defined by the GEM since we have assumed that its composition depends on the product secretion fluxes. In our case, the extracellular matrix and the following portfolio definition took place after the network systematic reduction. The resulting extracellular matrix was subsequently reduced as well, simplifying the separation but also constraining the portfolio alternative targets. Given the lack of experimental data and the nature of our task which was to showcase the connection between the strain design and the synthesis this simplification is acceptable. Anyhow, in future implementations the reduction step shall maintain a vivid communication with the downstream considerations in the sense that the potential alternative compositions occurring in the initial GEM shall be maintained in the reduced model. The consistency checks in the future reduced models used in the workflow shall thoroughly investigate the plurality and the range of the extracellular fluxes.

6.1.2 Incorporating economic factors onto the Genome-scale model

The upstream process assumptions is the key element to translate cellular fluxes to stream composition enabling in that way, the rest of the analysis. By assigning market prices to the target portfolio chemicals we were able to assess the maximum attainable potential revenue. When we valorise the kerosene compounds the potential maximum revenue is 45M$/Y and raises up to 180M$/Y when the target portfolio contains the total of the insoluble stream products. The revenue values exhibit a growth-coupled behaviour since the revenue value that corresponds to maximum growth rate is non-zero.

For aerobic fermentation in a chemostat array with dilution rate D= 0.1 h^{-1} , steady glucose supply 994 T/day (240 g/L) and glucose uptake rate -5 mmol · $gDCW^{-1} \cdot h^{-1}$ we observed that the potential revenue from hydrocarbons may fluctuate from 0M$/Y to 42M$/Y. Comparing the cellular fluxes variability for the two cases we identified that FDH↑ upregulation may lead to strong growth-coupling of the hydrocarbons revenue reassuring non-zero revenue values.

The incorporation of economic factors onto the GEM context, enables the extraction of useful data concerning the process viability. In the present case study, we verified that the feedstock cost remains the most important bottleneck in the upstream process. Given that glucose market price is comparable to those of fuels, it is easy to understand that specialty chemicals represent a more profitable target market. The utilization of alternative non-food competing feedstock stands as the only promising direction.

6.1.3 Strain design algorithm

The strain design algorithm is the core of the Design module of the DBTL cycle. The incorporation of the revenue terms inside the GEM untaps a new dynamic. Using the total potential revenue as the engineering objective in an OptKnock-like formulation, we can unravel different portfolios that result to the desired properties (e.g. growth-coupling). We estimate that this approach can be used to simultaneously identify promising product combinations to be produced microbially. In the present study, we were not able to identify a metabolic strategy that results in stronger growth-coupling of the insoluble stream potential revenue. This is partially due to the existence of pool reactions in the lipid metabolism subsystems and the less number of rewiring alternatives that the reduced model entails.

6.1.4 GEM sampling

In the case study, the alternative metabolic strategies were tested *in silico* by applying the according capacity constraints to the GEM. The sampling procedure was applied to assess alternative phenotypic microbial states that correspond to a chemostat array with dilution rate D= 0.1 ± 0.01 h^{-1} and steady glucose supply 994 T/day (240 g/L). The alternative strategies result in different product secretion ranges thus different stream composition capacities. For example, the two representative cases of the initial strain and the FDH↑ strain exhibit utterly different hydrocarbon production potential.

In the second case the alcohols and sterol production drops significantly while the estimated revenue doubles. The kerosene produced by the initial strain is rich in n-hexadecane (33%) with the rest of the components: 21% n-tridecane, 18% 1-pentadecene, 17% n-pentadecane, 11% n-heptadecene. The FDH↑ strain yields in a kerosene mixture where n-tridecane is the most abundant compound (82%) and pentadecane is in similar levels to the initial strain produced kerosene (14%).

6.1.5 Minimum downstream cost estimation

The downstream process synthesis and cost estimation focused on the separation of the stream that contains the insoluble products. For this specific case the superstructure formulation can be represented as a distillation supertask problem. Using an MILP formulation we identified the minimum separation cost for the alternative stream compositions that correspond to the different metabolic strategies. We developed an automated matlab set of functions that generates and solves the MILP problem using the recovery matrix and the stream initial composition as inputs.

The Antoine equation is used to estimate the distillation column top and bottom temperatures while the FUG shortcut method is applied to calculate the columns' basic operational characteristics. Solution of the problems yield sin a conceptual minimum cost estimation for the corresponding stream separation. Conducting a consistency check for the initial strain we concluded that the conceptual cost-based algorithm is generally in accordance with more detailed cost estimation methods.

The tool can be expanded to address more complex sequencing problems that take into consideration alternative unit operations such as filtration, reverse osmosis, chromatography and crystallisation. In that way, we will be able to expand our workflow for the production of alternative chemicals and result in more reliable downstream cost estimates.

6.1.6 Promising metabolic strategies

Applying our developed workflow, we tested *in silico* 33 alternative metabolic strategies and calculated the corresponding mean downstream separation cost and mean product revenue. Although the feedstock price in our case study is the highest among the expenses our study focuses on separation-product revenue trade-offs. We identified metabolic strategies applied to the initial strain that enhance strain's efficiency up to 26%. The identified metabolic strategies applied at the FDH↑ strain yield efficiencies 500-700% for the production of the specific portfolio.

The most promising identified metabolic strategy corresponds to an FDH↑ strain with five reaction deletions: FDH **↑** POT1_C10, POT1_C12, POT1_C14, POT1_C16, POX1_C10 **Δ** that practically disrupt beta-oxidation for fatty acids with chain lengths 10-16 carbon atoms. The translation of such strategies to laboratory practice entails an extra level of difficulty. Since the same enzyme catalyzes the reaction for substrates of different chain lengths, we cannot delete the corresponding gene because that would lead to collateral deletion of desired reactions. A possible interpretation of the strategy would involve enzyme engineering towards catalysts that exhibit low affinity towards the undesired chain-length substrates. Alternatively, we can search for corresponding enzymes from different organisms that exhibit behavior closer to the desired.

6.2 The advanced DBTL cycle

The proposed workflow can directly assist the typical DBTL cycle. It is evident that the application of the workflow requires a valid GEM. The workflow estimations provide a range for the fermentation broth composition and the subsequent downstream cost. It makes sense that provision of experimental data will further constrain the model and improve the accuracy of our calculations. By that, we propose that our workflow can work at the interface of Design and Learn modules making sure that the proposed metabolic strategies don't deviate from the minimum cost target.

6.3 Towards metabolic – technoeconomic models (MTEs)

Our initial goal was to formulate a problem where stoichiometric matrix is merged with all the technoeconomic parameters and variables that frame the bioprocess, including the upstream part and the downstream-separation part. In that way we would be able to directly identify network interventions that result to maximum total profit and estimate the optimal separation technologies' sequence and operation conditions.

Although the upstream process terms and revenue terms can be easily incorporated to the GEM, the superstructure representation poses some difficulties. Usually the development of the reconstruction requires a priori knowledge of the starting stream composition to define the properties of the distinct tasks. A future challenge that we currently try to tackle is to develop a framework that will enable the unification of the two types of problems leading to a new generation type of MTE models.

MTEs can be used to address the total microbial biorefinery synthesis problem in parallel to the strain design module. The total biorefinery synthesis problem aims to identify optimal routes from feedstock to processes to portfolios that maximize profit. In a strain design framework where we examine the total unified MTE we can propose tailor-made microbial strains that consume the ideal feedstock to produce the target portfolios- reassuring in that way maximum profit and establishing a systematic umbrella approach to lead future biorefinery synthesis.

References

1. d'Espaux, L. *et al.* Engineering high-level production of fatty alcohols by Saccharomyces cerevisiae from lignocellulosic feedstocks. *Metab. Eng.* **42,** 115–125 (2017).

2. Kokossis, A. C., Tsakalova, M. & Pyrgakis, K. Design of integrated biorefineries. *Comput. Chem. Eng.* **81,** 40–56 (2015).

3. Home - 2019 - United Nations Sustainable Development. Available at: https://www.un.org/sustainabledevelopment/. (Accessed: 21st September 2019)

4. Lokko, Y. *et al.* Biotechnology and the bioeconomy—Towards inclusive and sustainable industrial development. *N. Biotechnol.* **40,** 5–10 (2018).

5. Amoah, J., Kahar, P., Ogino, C. & Kondo, A. Bioenergy and Biorefinery: Feedstock, Biotechnological Conversion, and Products. *Biotechnol. J.* **14,** 1800494 (2019).

6. Gustavsson, M. & Lee, S. Y. Prospects of microbial cell factories developed through systems metabolic engineering. *Microb. Biotechnol.* **9,** 610–617 (2016).

7. Caspeta, L. & Nielsen, J. Economic and environmental impacts of microbial biodiesel. *Nat. Biotechnol. 2013 319* (2013).

8. Lee, J. W. *et al.* Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* **8,** 536–546 (2012).

9. Long, M. R., Ong, W. K. & Reed, J. L. Computational methods in metabolic engineering for strain design. *Curr. Opin. Biotechnol.* **34,** 135–141 (2015).

10. Maranas, C. D. & Zomorrodi, A. R. *Optimization methods in metabolic networks*.

11. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84,** 647–657 (2003).

12. Wu, W., Henao, C. A. & Maravelias, C. T. A superstructure representation, generation, and modeling framework for chemical process synthesis. *AIChE J.* **62,** 3199–3214 (2016).

13. Stephanopoulos, G., Aristidou, A. A. & Nielsen, J. H. *Metabolic engineering : principles and methodologies*.

14. Ingalls, B. P. *Mathematical modeling in systems biology : an introduction*. (The MIT Press, 2013).

15. Stephanopoulos, G. Synthetic Biology and Metabolic Engineering. *ACS Synth. Biol.* **1,** 514–525 (2012).

16. Martínez-García, E. & de Lorenzo, V. Molecular tools and emerging strategies for deep genetic/genomic refactoring of Pseudomonas. *Curr. Opin. Biotechnol.* **47,** 120–132 (2017).

17. Zhang, Y., Nielsen, J. & Liu, Z. Metabolic engineering of *Saccharomyces cerevisiae* for production of fatty acid-derived hydrocarbons. *Biotechnol. Bioeng.* **115,** 2139–2147 (2018).

18. Foo, J. L., Susanto, A. V., Keasling, J. D., Leong, S. S. J. & Chang, M. W. Whole-cell biocatalytic and de novo production of alkanes from free fatty acids in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **114,** 232–237 (2017).

19. Lee, S. Y. *et al.* A comprehensive metabolic map for production of bio-based chemicals. *Nat. Catal.* **2,** 18–33 (2019).

20. Becker, J. & Wittmann, C. Advanced Biotechnology: Metabolically Engineered Cells for the Bio-Based Production of Chemicals and Fuels, Materials, and Health-Care Products. *Angew. Chemie Int. Ed.* **54,** 3328–3350 (2015).

21. Soh, K. C. & Hatzimanikatis, V. DREAMS of metabolism. *Trends Biotechnol.* **28,** 501–508 (2010).

22. Karatzos, S., McMillan, J., Task, J. S.-R. for I. B. & 2014, undefined. The potential and challenges of drop-in biofuels. *task39.org*

23. Nielsen, J. Yeast Systems Biology: Model Organism and Cell Factory. *Biotechnol. J.* **14,** 1800421 (2019).

24. Fairley, P. Introduction: Next generation biofuels. *Nature* **474,** S2–S5 (2011).

25. Kung, Y., Runguphan, W. & Keasling, J. D. From Fields to Fuels: Recent Advances in the Microbial Production of Biofuels. *ACS Synth. Biol.* **1,** 498–513 (2012).

26. Kang, M.-K. & Nielsen, J. Biobased production of alkanes and alkenes through metabolic engineering of microorganisms. *J. Ind. Microbiol. Biotechnol.* **44,** 613–622 (2017).

27. Kang, M.-K. & Nielsen, J. Biobased production of alkanes and alkenes through metabolic engineering of microorganisms. *J. Ind. Microbiol. Biotechnol.* **44,** 613–622 (2017).

28. Shi, S., Octavio Valle-Rodriguez, J., Khoomrung, S., Siewers, V. & Nielsen, J. Functional expression and characterization of five wax ester synthases in Saccharomyces cerevisiae and their utility for biodiesel production. *Biotechnol. Biofuels* **5,** 7 (2012).

29. Buijs, N. A., Zhou, Y. J., Siewers, V. & Nielsen, J. Long-chain alkane production by the yeast *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **112,** 1275–1279 (2015).

30. Fernandez-Moya, R. & Da Silva, N. A. Engineering Saccharomyces cerevisiae for high-level synthesis of fatty acids and derived products. *FEMS Yeast Res.*

17, (2017).

31. Zhang, Y., Nielsen, J. & Liu, Z. Metabolic engineering of *Saccharomyces cerevisiae* for production of fatty acid-derived hydrocarbons. *Biotechnol. Bioeng.* **115,** 2139–2147 (2018).

32. Zhou, Y. J., Hu, Y., Zhu, Z., Siewers, V. & Nielsen, J. Engineering 1-Alkene Biosynthesis and Secretion by Dynamic Regulation in Yeast. *ACS Synth. Biol.* **7,** 584–590 (2018).

33. Zhou, Y. J., Kerkhoven, E. J. & Nielsen, J. Barriers and opportunities in bio-based production of hydrocarbons. *Nat. Energy* 1 (2018). doi:10.1038/s41560-018-0197-x

34. Wu, W., Long, M. R., Zhang, X., Reed, J. L. & Maravelias, C. T. A framework for the identification of promising bio-based chemicals. *Biotechnol. Bioeng.* (2018). doi:10.1002/bit.26779

35. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5,** 264–276 (2003).

36. Ataman, M., Hernandez Gardiol, D. F., Fengos, G. & Hatzimanikatis, V. redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLOS Comput. Biol.* **13,** e1005444 (2017).

37. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–248 (2010).

38. Soh, K. C. & Hatzimanikatis, V. Constraining the Flux Space Using Thermodynamics and Integration of Metabolomics Data. in 49–63 (Humana Press, New York, NY, 2014). doi:10.1007/978-1-4939-1170-7_3

39. Jensen, K., Broeken, V., Hansen, A. S. L., Sonnenschein, N. & Herrgård, M. J. OptCouple: Joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs. *Metab. Eng. Commun.* **8,** e00087 (2019).

40. von Kamp, A. & Klamt, S. Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms. *Nat. Commun.* **8,** 15956 (2017).

41. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10,** 291–305 (2012).

42. Smith, R. *Chemical Process Design and Integration*. *John Wiley & Sons, Ltd* (Wiley, 2005). doi:10.1529/biophysj.107.124164

43. Fenske, M. R. Fractionation of Straight-Run Pennsylvania Gasoline. *Ind. Eng. Chem.* **24,** 482–485 (1932).

44. Underwood, A. J. V. Fractional Distillation of Multicomponent Mixtures. *Ind. Eng. Chem.* **41,** 2844–2847 (1949).

45. Westerberg, A. W. The synthesis of distillation-based separation systems. *Comput. Chem. Eng.* **9,** 421–429 (1985).

46. Shah, P. B. & Kokossis, A. Design targets of separator and reactor-separator systems using conceptual programming. *Comput. Chem. Eng.* **21,** S1013–S1018 (1997).

47. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164,** 1185–1197 (2016).

48. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987 (2015).

49. Tsagkari, M., Couturier, J.-L., Kokossis, A. & Dubois, J.-L. Early-Stage Capital Cost Estimation of Biorefinery Processes: A Comparative Study of Heuristic Techniques. *ChemSusChem* **9,** 2284–97 (2016).

50. Kokossis, A. C. & Yang, A. On the use of systems technologies and a systematic approach for the synthesis and the design of future biorefineries. *Comput. Chem. Eng.* **34,** 1397–1405 (2010).

51. Pyrgakis, K. A. & Kokossis, A. C. A Total Site Synthesis approach for the selection, integration and planning of multiple-feedstock biorefineries. *Comput. Chem. Eng.* **122,** 326–355 (2019).

52. Ataman, M. & Hatzimanikatis, V. lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLOS Comput. Biol.* **13,** e1005513 (2017).

53. Wang, H. *et al.* RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Comput. Biol.* **14,** e1006541 (2018).

54. Herrmann, H. A., Dyson, B. C., Vass, L., Johnson, G. N. & Schwartz, J.-M. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Syst. Biol. Appl.* **5,** 32 (2019).

55. Saa, P. A. & Nielsen, L. K. ll-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinformatics* **32,** 2330–2337 (2016).

56. Schellenberger, J. & Palsson, B. Ø. Use of Randomized Sampling for Analysis of Metabolic Networks. *J. Biol. Chem.* **284,** 5457–5461 (2009).

57. Yeomans, H. & Grossmann, I. E. A systematic modeling framework of superstructure optimization in process synthesis. *Comput. Chem. Eng.* **23,** 709–731 (1999).

58. Pyrgakis, K. A. *et al.* A process integration approach for the production of biological iso-propanol, butanol and ethanol using gas stripping and

adsorption as recovery methods. *Biochem. Eng. J.* **116,** 176–194 (2016).

59. Herrgård, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* **26,** 1155–1160 (2008).

60. Salvy, P. *et al.* pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. *Bioinformatics* **35,** 167–169 (2019).

61. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6,** 1290–1307 (2011).

62. Kaufman, D. E. & Smith, R. L. Direction Choice for Accelerated Convergence in Hit-and-Run Sampling. *Oper. Res.* **46,** 84–95 (1998).

63. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28,** 100 (1979).

64. Douglas, J. M. (James M. *Conceptual design of chemical processes*. (McGraw-Hill, 1988).

65. Yaws, C. L. *The Yaws Handbook of Vapor Pressure : Antoine coefficients*.

66. NIST WebBook. Available at: https://webbook.nist.gov/. (Accessed: 26th September 2019)

67. Salvy, P. & Hatzimanikatis, V. ETFL: A formulation for flux balance models accounting for expression, thermodynamics, and resource allocation constraints. *bioRxiv* 590992 (2019). doi:10.1101/590992

68. Zhu, Z. *et al.* Enabling the synthesis of medium chain alkanes and 1-alkenes in yeast. *Metab. Eng.* **44,** 81–88 (2017).

Appendix A: Upstream process parameters

Table: Upstream process parameters

Fermenter	
Temperature (ºC)	30
Pressure (atm)	1
Specific power (kW/m3)	0,3
PC (purhcase cost) base size (m3)	3500
PC Base cost ($)	700000
Max size per equipment (m3)	2750

Feedstock	
Glucose ($/kg)	0,5
Water ($/m3)	0,85

Utility	
Chilled water ($/T)	0,4
Electricity ($/KW-h)	0,1

Other	
Costing year	2016
Operating days per year	330
Operating capacity (%)	100
Capital charge factor	0,1

Appendix B: Preparatory calculations for the sequencing problem

Here we present the calculation steps followed to construct the sequencing problem parameters. We constructed an algorithm to conduct the calculations in matlab environment. The matlab functions used here are available upon request.

Let us assume we have a 3-product separation by distillation sequencing problem. A stream containing 60% n-tridecane, 10% n-pentadecane and 30% n-heptadecane with a feed flowrate 1000kmol/h needs to be separated in 3 products A,B,C as indicated in the recovery matrix :

Table 25: The recovery fraction matrix

component/Product	A	B	C
n-tridecane	0.98	0.02	0.00
n-pentadecane	0.01	0.98	0.01
n-heptadecane	0.00	0.02	0.98

The supertask problem involves 3 product subgroups and 4 tasks

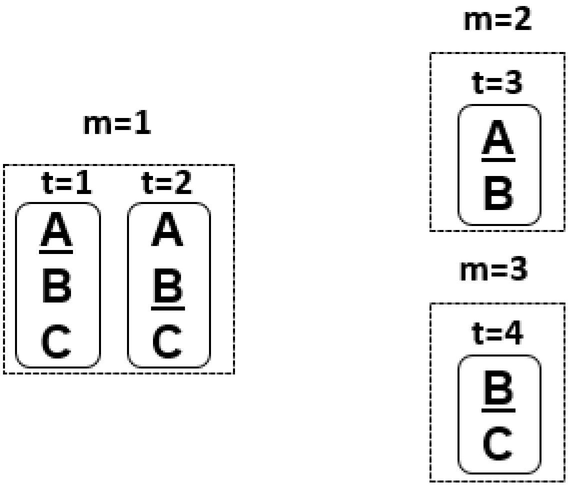


Figure 47: The subgroups and tasks for the 3-product distillation problem

First, we have to estimate the distillation tasks resulting compositions using the recovery matrix and the given stream. Then we will apply the Antoine equation to estimate the Top and Bottom temperatures and the relative volatilities of the components and finally we will apply the FUG shortcut methods to estimate each task's basic characteristics-needed to construct the MILP problem. The problem is solve as indicated in section 3.6.1 thus, we will not get into further detail on problem formulation and solution.

The products' fraction for each product subgroup is calculated as indicated by the recovery matrix. For the first subgroup m=1 (A,B,C) it will be:

Table 26: Table for the first subgroup

ABC			RECOVERY FRACTION MATRIX		
i	cp	xif	A	B	C
1	n-C13	0.6	0.98	0.02	0.00
2	n-C15	0.1	0.01	0.98	0.01
3	n-C17	0.3	0.00	0.02	0.98
Product Fraction		xip	0.589	0.116	0.295

Based on the results of the first subgroup we can now construct the Feed and Product vertices for the rest:

Table 27:Table for the second subgroup

AB			RECOVERY FRACTION MATRIX		
i	cp	xif	A	B	C
1	n-C13	0,851064	0,98	0,02	0
2	n-C15	0,140426	0,01	0,98	0,01
3	n-C17	0,008511	0	0,02	0,98
Product Fraction		xip	0,835	0,155	0,010

Table 28: Table for the third subgroup

BC			RECOVERY FRACTION MATRIX		
i	cp	xif	A	B	C
1	n-C13	0,029197	0,98	0,02	0
2	n-C15	0,240876	0,01	0,98	0,01
3	n-C17	0,729927	0	0,02	0,98
Product Fraction		xip	0,031	0,251	0,718

Table 29: Distillate and Bottom molar fractions

		Distillate molar fraction xd				Bottom molar fraction xd			
i	cp	**t=1**	**t=2**	**t=3**	**t=4**	**t=1**	**t=2**	**t=3**	**t=4**
1	n-C13	0,998	0,851	0,998	0,002	0,029	0	0,110	0
2	n-C15	0,002	0,140	0,002	0,94	0,241	0,003	0,889	0,003
3	n-C17	0	0,009	0	0,058	0,730	0,997	0,001	0,997

Table 30: Fraction of feed ζ that yields stream mεM needed to construct the MILP

ζ(m,t) table				
m/t	**1**	**2**	**3**	**4**
1	0	0	0	0
2	0	0,705	0	0
3	0,411	0	0	0

Continuing, we can apply Antoine equation to estimate Top and Bottom temperatures for columns working on atmospheric pressure:

Table 31: Top and Bottom temperatures

	t=1	**t=2**	**t=3**	**t=4**
Ttop (ºC)	236	239	236	272
Tbot (ºC)	288	299	265	299

Table 32: Relative volatilities for each task

		Relative volatility			
		t=1	**t=2**	**t=3**	**t=4**
i	**place of cut cp**	**1**	**2**	**1**	**2**
1	n-C13	2,2	4,4	2,2	4,4
2	n-C15	1,0	2,0	1,0	2,0
3	n-C17	0,5	1,0	0,5	1,0

Finally applying the FUG shortcut methods, we get:

Table 33: Basic column characteristics as calculated by FUG.

	t=1	**t=2**	**t=3**	**t=4**
Nmin	10,88844	12,28388	10,87541	12,26919
Ntheor	28	32	28	30
R	2,205666	1,709604	2,160593	4,996359
V/F	1,888137	1,910271	2,640507	1,506531

The problem is formulated as described in Chapter 3.6.1 and solved to identify that the sequence **1-4** yields in minimum separation cost **2.87M$/Y**

Appendix C: Cost models

Heat exchangers cost[64]

$$\text{InstalledCost(\$)} = \left(\frac{M\&S}{280}\right) \cdot 101.3 \cdot A^{0.65} \cdot (2.29 + (F_d + F_p) \cdot F_m)$$

$$\text{A(area)} = \frac{HEATDUTY}{U \cdot \Delta TLM}\ [ft^2]$$

M&S (projected for 2019): 1638.9

F_d = 0.85 (U-tube heat exchanger)

F_p = 0.0 (Pressure up to 10 bar)

F_m=2.81 (Stainless steel)

Overall heat transfer coefficient U= 200 BTU·h^{-1}·ft^{-2}·F^{-1}

Furnace cost[64]

$$\text{InstalledCost(\$)} = \left(\frac{M\&S}{280}\right) \cdot 5070 \cdot Q^{0.85} \cdot (1.23 + F_d + F_m + F_p)$$

F_d = 1.0 (cylindrical)

F_p = 0.0 (Pressure up to 500psi)

F_m=0.0 (Carbon steel)

Distillation columns cost[64]

$$\text{InstalledCost(\$)} = \left(\frac{M\&S}{280}\right) \cdot 101.9 \cdot D^{1.066} \cdot H^{0.802} \cdot (2.18 + F_m \cdot F_p)$$

$$\text{D(diameter), H(Height)} = \text{[ft]}$$

F_p = 1.0 (Pressure up to 3.4 bar)

F_m= 2.25 (Stainless steel)

Distillation column trays and tower internals[64]

$$\text{InstalledCost(\$)} = \left(\frac{M\&S}{280}\right) \cdot 4.7 \cdot D^{1.55} \cdot H \cdot (F_s + F_t \cdot F_m)$$

$$\text{D(diameter), H(Height)} = \text{[ft]}$$

F_s = 2.2 (Tray spacing (12in))

F_m= 1.7 (Stainless steel)

F_t= 1.8 (Tray type (Bubble cap))

Height and Diameter calculations

$$\mathrm{H = Tray\ spacing \cdot Ntheor + vapour\ disengaging\ space}$$

Ntheor: The theoretical number of trays

Vapour disengaging space = 13.12 ft

$$\mathrm{D} = \sqrt{\frac{Area}{\pi}}$$

$$\mathrm{Area} = \frac{vapour\ traffic}{v^{flood} \cdot (1-\varphi)}$$

$$\mathrm{v}_{max}^{flood} = CS \cdot \sqrt{\frac{\rho^{\mathrm{liq}} - \rho^{vap}}{\rho^{vap}}}$$

$\mathrm{v}^{flood} = 0.7 \cdot \mathrm{v}_{max}^{flood}$ and ϕ=0.12

ρ^{liq} is the liquid density estimated based on the molar fraction and ρ^{vap} the vapour density calculated with the ideal gases equation of state for T= T_{cond} .

Annualization of Capital cost $\mathrm{ACC = CAPEX} \cdot \frac{i \cdot (1+i)^n}{(1+i)^n - 1}$

Interest rate i = 0.05

Number of years n = 25

Negligible Insurance and Maintenance cost

Appendix D: Component Properties

$$Log_{10}(P) = A - \frac{B}{(T+C)}$$

P: component vapour pressure in mmHg

T: temperature ^{o}C

A,B,C Antoine equation compound constants

	A	B	C
Tridecane	7,00756	1690,67	174,22
Pentadecene	7,01555	1781,974	162,582
Pentadecane	7,02359	1789,95	161,38
Tetradecanol	7,41181	2003,29	168,13
Heptadecene	7,03925	1877,91	151,53
Heptadecane	7,0143	1865,1	149,2
Hexadecanol	6,1586	1380	91
Octadecanol	4,32298	685,976	10,85
Sterol	2	3500	0

The alkanes, alkenes and fatty-alcohol Antoine constants are presented as found at The Yawns handbook of vapour pressure: Antoine coefficients.[65]

The condenser and reboiler duty are estimated based on the vaporization enthalpy ΔH_{vap} [KJ/mol] . Using the available data found in the NIST database [66] we constructed the following approximations to calculate the enthalpies at different temperatures and p=1 atm.

ΔH_{vap} (tridecane) = -0.0885*T+92.783;

ΔH_{vap} (pentadecene) = -0.0655*T+ 90.171;

ΔH_{vap} (pentadecane) = -0.095*T+ 103.5;

ΔH_{vap} (tetradecanol) = -0.2556*T+ 186.13;

ΔH_{vap} (heptadecene) = -0.0757*T + 101.89;

ΔH_{vap} (heptadecane) = -0.1381*T+ 133.4;

ΔH_{vap} (hexadecanol)= -0.2292*T+ 183.25;

ΔH_{vap} (octadecanol) = -0.2226*T+ 189.25;