Fortschritt-Berichte VDI

VDI

Reihe 3

Verfahrenstechnik

Dipl.-Ing. Dipl.-Wirtsch.-Ing.
Juan José Wilhelm Victoria Villeda,
Berlin

Nr. 950

Reaction Network Generation and Evaluation for the Design of Biofuel Value Chains

Berichte aus der
Aachener Verfahrenstechnik - Prozesstechnik

RWTH Aachen University

Reaction Network Generation and Evaluation for the Design of Biofuel Value Chains

Generierung und Evaluierung von Reaktionsnetzwerken für die Auslegung von Biokraftstoff-Wertschöpfungsketten

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Juan José Wilhelm Victoria Villeda

Berichter: Univ.-Prof. Dr.-Ing. Wolfgang Marquardt
Univ.-Prof. Dr.-Ing. André Bardow
Tag der mündlichen Prüfung: 28.10.2016

Fortschritt-Berichte VDI

Reihe 3

Verfahrenstechnik

Dipl.-Ing. Dipl.-Wirtsch.-Ing.
Juan José Wilhelm Victoria Villeda,
Berlin

Nr. 950

Reaction Network Generation and Evaluation for the Design of Biofuel Value Chains

Berichte aus der
Aachener Verfahrenstechnik - Prozesstechnik

RWTH Aachen University

Victoria Villeda, Juan José Wilhelm

Reaction Network Generation and Evaluation for the Design of Biofuel Value Chains

Fortschr.-Ber. VDI Reihe 3 Nr. 950. Düsseldorf: VDI Verlag 2017.
202 Seiten, 29 Bilder, 60 Tabellen.
ISBN 978-3-18-395003-4, ISSN 0178-9503,
€ 71,00/VDI-Mitgliederpreis € 63,90.

Für die Dokumentation: Biofuels – Value Chain Design – Formal Reaction Network Generation – Biofuel blend – Network evaluation – Optimization-based analysis – Elementary mode analysis – Network topology analysis

The present contribution addresses engineers and researchers in process systems engineering and computational chemistry. It provides a holistic approach to the design of biofuel synthesis value chains, consisting of several analytic stages. The model-based centers around the generation of reaction networks that head from substrates to desired products. The generation advances in in a formal manner, relying of fundamentals of chemistry, restrictable by empirical knowledge. By investigating the molecular structure of the network intermediates, quantitative statements on the selectivity of network reactions are derived. The networks are decomposed into elementary nodes, non-further decomposable sequences of reactions. Optimization-based techniques are employed to determine optimal synthesis pathways and biofuel blends. The employed evaluation criteria allow for deriving statements on techno-economic performance and network topology.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter http://dnb.ddb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at http://dnb.ddb.de.

D82 (Diss. RWTH Aachen University, 2016)

Als Manuskript gedruckt. Printed in Germany.
ISSN 0178-9503
ISBN 978-3-18-395003-4

Vorwort

Die hier vorliegende Arbeit entstand während meiner Zeit als Stipendiat der NRW Forschungsschule "Brennstoffgewinnung aus nachwachsenden Rohstoffen" (BrenaRo) am Lehrstuhl für Prozesstechnik der Aachener Verfahrenstechnik an der RWTH Aachen.

Mein besonderer Dank gilt meinem Doktorvater, Herrn Professor Dr.-Ing. Wolfgang Marquardt für die Förderung und Unterstützung während dieser Zeit. Ohne seine Begeisterungsfähigkeit und Offenheit gegenüber neuen Ansätzen wäre diese Arbeit nicht möglich gewesen. Weiterhin danke ich dem Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen für die Bereitstellung der Stipendiatenstelle. Schließlich danke ich Professor Dr.-Ing. André Bardow für die Übernahme des Koreferats und Univ.-Prof. Dr.-Ing. Hubertus Murrenhoff für die Übernahme des Prüfungsvorsitzes.

Allen Mitarbeitern des Lehrstuhls danke ich für die freundschaftliche und kollegiale Atmosphäre, die das Leben und Arbeiten in Aachen sehr angenehm gestaltet hat. Die enge Zusammenarbeit mit Manuel Dahmen und Manuel Hechinger sowie die Kooperation mit Kirsten Ulonska, Jörn Viell und Anna Voll innerhalb der TMFB-Gruppe sowie Christian Redepenning, Sebastian Recker und Mirko Skiborowski aus der Synthesegruppe war stets sehr angenehm und produktiv. Weiterhin danke ich Manuel Dahmen für das Korrekturlesen dieser Arbeit.

Mein größter Dank gilt meiner Familie. Was ihr zeit meines Lebens für mich geleistet habt und für mich bedeutet, kann man nicht in Worte fassen. Alles was ich kann, bin und erreicht habe, verdanke ich euch. Ohne eure bedingungslose Unterstützung wäre ich nicht dort, wo ich jetzt bin. Dafür bin ich euch unendlich dankbar.

Berlin, im Oktober 2016 *Juan José Wilhelm Victoria Villeda*

Contents

Notation

Latin symbols

A	Atom Set	[–]
A	Stoichiometric Matrix	[–]
AE	Atom Efficiency	[%]
ARE	Average Relative Error	[%]
ALR	Annual Loan Repayment	[$/a]
AM	Adjacency Matrix	[–]
AS	Adjacency Scheme	[–]
B	Substrate Matrix	[–]
b	Vector of Fluxes Leaving a Network	[–]
BM	Bond Electron Matrix	[–]
C	Conversion	[%]
CE	Carbon Efficiency	[%]
CN	Cetane Number	[–]
E	Product Matrix	[–]
E	Edge Set	[–]
EC	Energy Consumption	[–]
EI	Environmental Impact	[–]
Em	Emissions Indicator	[–]
F	Flux Matrix	[–]
f	Flux Vector	[–]
H	Enthalpy	[kJ/kmole]
IC	Investment Cost	[$]
LD_{50}	Median Lethal Dose	[mg/kg]
MB	Molecular Body	[–]
MRE	Maximum Relative Error	[%]
MW	Molecular Weight	[g/mole]
N	Quantity	[–]

N	Absolute Molar Flux Matrix	[mole]
n	Normalized Molar Flux Matrix	[mole/mole]
p	Reaction Patterns Vector	[–]
p	Price	[$/kg] or [$/l]
Q	Formal Electric Charge (Molecule)	[–]
q	Formal Electric Charge (Atom)	[–]
r	Reaction	[–]
RC	Resource Consumption	[–]
rp	Resulting Patterns Vector	[–]
S	Selectivity	[%]
s	Substance	[–]
T	Temperature	[K]
t	Plant Lifetime	[y]
tp	Triggering Patterns Vector	[–]
TAC	Total Annualized Costs	[$/a]
TAR	Total Annualized Revenues	[$/a]
TP	Toxicity Potential	[–]
TS	Total Selectivity	[–]
TT	Transition Table	[–]
v	Stoichiometric Coefficient	[–]
V	Vertex Set	[–]
VE	Valence Electrons	[–]
VS	Valence Scheme	[–]
VSC	Valence Scheme Combinations	[–]
x	Fraction	[–]
Y	Yield	[%]
y	Molar Fraction	[–]
z	Interest Rate	[%]

Greek symbols

α	Weighting Factor of an Elementary Mode
γ	Scaling Factor of Normalized Fluxes
η	Efficiency
σ	Permutation of Valence Scheme Bonds
χ	Atomic Share

Abbreviations

2-BF	2-Butylfuran
2-BTHF	2-Butyltetrahydrofuran
3-MTHF	3-Methyltetrahydrofuran
BrenaRo	Brennstoffgewinnung aus nachwachsenden Rohstoffen
BtL	Biomass to Liquid
C	Carbon
CI	Compression Ignition
CO_2	Carbon Dioxide
DB	Double Bond
DHF	Dihydrofuran
EM	Elementary Mode
FAME	Fatty Acid Methyl Esters
FIME	Family of Isomeric Molecular Ensembles
FTS	Fischer-Tropsch Synthesis
GAMS	General Algebraic Modeling System
H	Elementary Hydrogen
H_2	Molecular Hydrogen
H_2O	Water
IA	Itaconic Acid
LCA	Life Cycle Assessment
LP	Linear Programming
ME	Molecular Ensemble
MIP	Mixed-Integer Programming
MILP	Mixed-Integer Linear Programming
NBP	Normal Boiling Point
NLP	Non-Linear Programming
O	Oxygen
QSPR	Quantitative Structure Property Relationship
RNFA	Reaction Network Flux Analysis
SB	Single Bond
SMILES	Simplified Molecular Input Line Entry Specification
TB	Triple Bond
THF	Tetrahydrofuran
TMFB	Tailor-Made Fuels from Biomass
WTW	Well-to-Wheel (Analysis)

Subscripts and superscripts

$boil$	At Boiling Point
C	Carbon Atom
com	Combustion
$form$	Formation
FS	Feedstock
H	Hydrogen Atom
in	Flux Entering a Network
k	Iteration Variable
L	In Liquid State
$melt$	At Melting Point
T	Target
O	Oxygen Atom
out	Flux Leaving a Network
P	Product
$Path$	Pathways
R	Reaction
Rf	Refunctionalization
S	Substrate
T	Transposed Matrix
tot	Total
U	Upper Triangular Matrix
VS	Valence Scheme
VSC	Valence Scheme Combination
VT	Valence Scheme Transition
$-$	Out(-degree of a Vertex)
$+$	In(-degree of a Vertex)

Kurzfassung

Die stetige Verknappung fossiler Ressourcen sowie der steigende Energiebedarf erfordern eine Neuausrichtung der chemischen Industrie bezüglich der verwendeten Rohstoffe. Da die Menge des fossilen Kohlenstoffs begrenzt ist, ist es erforderlich, alternative Quellen zu erschließen, deren Verfügbarkeit auf lange Zeit gesichert ist. Dieser Wandel wird eine zentrale Rolle in der Entwicklung der chemischen Wertschöpfungsketten im Laufe der nächsten Jahre und Jahrzehnte einnehmen.

In den letzten Jahren hat sich Biomasse als wahrscheinlichster alternativer Kohlenstofflieferant herauskristallisiert. Die bisherigen Wertschöpfungsketten der chemischen Industrie sind aufgrund der Beschaffenheit der Biomasse jedoch nicht oder nur teilweise übertragbar. Daher geht mit der Änderung der Rohstoffquelle sowohl die Identifikation neuer Chemikalien mit gewünschten Eigenschaften als auch die Entwicklung neuer Prozesse einher.

Da die Verschiebung zu erneuerbaren Rohstoffen auch den größten Abnehmer fossiler Energieträger, den Verkehrssektor, betrifft, steht die Herstellung von Biokraftstoffen als essentielle Herausforderung der nächsten Jahre im Fokus wissenschaftlicher Aufmerksamkeit. Zum Zweck der systematischen Identifikation von Biokraftstoffen wurde an der RWTH Aachen der Exzellenzcluster "Tailor-Made Fuels From Biomass" (zu Deutsch "Maßgeschneiderte Kraftstoffe aus Biomasse") ins Leben gerufen. Dieser Forschungsverbund hat sich zum Ziel gesetzt, Kraftstoffe der nächsten Generation vorzuschlagen, die sowohl auf der Anwendungs- als auch auf der Herstellungsseite optimale Eigenschaften aufweisen. Diese optimalen Eigenschaften umfassen sowohl die wirtschaftliche und nachhaltige Synthese von Kraftstoffen aus Biomasse als auch eine emissionsarme und effiziente Verbrennung im Kolbenmotor.

In diesem Kontext fallen der computer-basierten Prozesstechnik zwei Aufgaben zu. Sie umfassen zum einen die Identifikation von Kraftstoffen, die definierte Eigenschaften erfüllen; zum anderen müssen dazugehörige Herstellungsprozesse vorgeschlagen und im Rahmen eines konzeptionellen Prozessentwurfs systematisch ausgearbeitet werden. Die Identifikation geeigneter Reaktionspfade ist der erste Schritt des konzeptionellen Prozessentwurfs. Die Auswahl von Reaktionspfaden erfolgte bisher meist auf der Basis experimenteller Untersuchungen und Heuristiken. Systematische Evaluierungkonzepte fußten auf manuell zusammengetragenen Reaktionsnetzwerken auf Basis von Literaturrecherchen.

Diese Vorgehensweise ist jedoch nicht nur fehleranfällig, sondern auch zeitaufwändig und begrenzt dadurch die Anzahl der untersuchbaren Fälle. Ganzheitliche, modellbasierte Ansätze zur Generierung, Identifikation und Evaluierung optimaler Synthesepfade im Rahmen der Biokraftstoffsynthese sind bisher kaum verfügbar.

Der Schwerpunkt dieser Arbeit liegt daher auf der computer-basierten Generierung und Auswertung von Reaktionsnetzwerken. Die Basis bildet eine graphentheoretische Formulierung von Molekülen und Reaktionen, wodurch die Entwicklung von kompakten und effizienten Algorithmen zur Modifikation der betrachteten Substanzen ermöglicht wird. Dadurch können, ausgehend von benutzerdefinierten Substraten, Reaktionspfade zu gewünschten Zielsubstanzen generiert werden. Die Formulierung erlaubt es auch, solche Reaktionen zu erzeugen und als Teil des Syntheseprozesses vorzuschlagen, die bisher noch nicht in der Literatur bekannt sind.

Zur Identifikation der einzelnen Reaktionspfade werden kombinatorische Methoden zur Analyse biologischer Netzwerke adaptiert. Ein mehrstufiger Ansatz aus Kombinatorik und Optimierung wird vorgeschlagen, der nicht nur eine ökonomische und ökologische Bewertung der Reaktionspfade ermöglicht, sondern auch die Topologie des Netzwerks erschließt und Aussagen über die Robustheit einer Syntheseentscheidung erlaubt. Aus einer Datenbank organischer Reaktionen werden experimentelle Daten abgerufen und in die Evaluierung integriert. Eingebettet in ein modellbasiertes Produkt-Prozess-Design können so Synthesepfade zu maßgeschneiderten Kraftstoffkandidaten systematisch identifiziert und für weiterführende Untersuchungen vorgeschlagen werden. In einem nachfolgenden Schritt wird geprüft, bis zu welchem Grad die Ausnutzung des Ausgangsmaterials erhöht werden kann, wenn auftretende Abfallströme in den Produktstrom integriert werden. Vorraussetzung ist hierbei, dass die resultierende Mischung ebenfalls die gewünschten Eigenschaften besitzt.

Generierung und Evaluierung von Reaktionsnetzwerken werden abschließend für zwei Fälle exemplarisch durchgeführt. Betrachtet wird dabei die Synthese von alternativen biobasierten Dieselkraftstoffen. Der erste Fall analysiert die Synthese von 3-MTHF ausgehend von Itakonsäure. Dieser Prozess wurde im Exzellenzcluster TMFB bereits detailliert untersucht. Es zeigt sich, dass das automatisch generierte Netzwerk eine Vielzahl bisher nicht betrachteter Reaktionen beinhaltet. Die angewendete Lösungsstrategie offenbart eine Vielzahl an Synthesepfaden und weist die wichtigsten Reaktionen des Netzwerks aus. Die Synthese von 3-MTHF leidet jedoch an den hohen Kosten des Substrats Itakonsäure, die auch durch die Integration der Abfallströme nicht wettgemacht werden können. Daher werden in einer zweiten Studie die strukturell ähnlichen Moleküle 2-BF und 2-BTHF vorgeschlagen und untersucht, die ausgehend vom aktuell günstigeren Furfural produziert werden können. Die durchgeführte Analyse zeigt, dass diese Substanzen effizient hergestellt werden können. Insbesondere dann, wenn Abfallströme in den Produktstrom integriert

werden, erscheint eine Synthese von 2-BTHF zu gegenwärtigen Marktpreisen von Furfural wirtschaftlich möglich.

Die präsentierten Methoden sind in einem Softwarepaket vereint. Dieser rein computerbasierte Ansatz beschleunigt und unterstützt den wissenschaftlichen Prozess der Identifikation neuer Kraftstoffe, indem detaillierte Netzwerke mit hohem Informationsgehalt in kurzer Zeit bereitgestellt werden. Darauf aufbauende Experimente können zielgerichtet geplant und ausgeführt werden, da vielversprechende Reaktionspfade schon vor Beginn der Versuche bekannt sind.

Abstract

The combination of continuously depleting fossil resources and the steadily increasing demand for energy pose upcoming challenges to the utilization of feedstock in chemical industry. Since the availability of fossil resources is limited by quantity, exploitation of alternative, long-term available sources is necessary. This transition will be a dominant center piece in the design of value chains in chemical industry within in the next years and decades.

More and more, biomass takes the stage as most promising alternative carbon source. However, current value chains are not or only to a limited extend transferrable due to the chemical and structural composition of biomass. Thus, a change in the carbon source will come hand in hand with the identification of novel chemical compounds with desired properties as well as with the development of corresponding production processes.

This shift towards biomass feedstock will affect all consumers of fossil energy carries, also the transportation sector as single largest consumer. This development positions biofuel production in the focus of academic research in the next years, both as challenge and also as opportunity. At RWTH Aachen University, the Cluster of Excellence "Tailor-Made Fuels from Biomass" was established to systematically identify and propose biofuels. Its overall objective is to propose next generation biofuels that exhibit optimal performance from an overall perspective, considering the production process as well as the thermo-physical properties. Thus, optimal performance takes into account the economic and sustainable synthesis of fuels from biomass and their efficient combustion in internal combustion engines at low emissions.

In this context, the task of computer-aided process systems engineering is twofold. It comprises on the one hand the identification of fuels that exhibit defined properties and on the other hand the identification and design of the corresponding production processes in a systematic conceptual process design approach. The first task to address here is to propose and evaluate suited reaction pathways from feedstock to desired product. Most often, the choice of reaction pathways was made based on experimental investigations and empirical knowledge. Systematic concepts for evaluating reaction pathway alternatives founded on manually assembled reaction networks, derived from exhaustive literature research. However, this approach is susceptible for incompleteness and time intense, thus posing methodological limits to the number investigatable scenarios. Holistic, model-based

approaches for generating, identifying and evaluating optimal synthesis pathways in the context of biofuel value chain design are only rarely available.

The contribution of this work is the computer-based generation and evaluation of reaction networks. It founds on formalisms of graph theory to abstract molecules and reactions, leading to compact and efficient algorithms for altering chemical substances. In this manner, reaction pathways are established from a user-defined feedstock to defined target substances, summing up to reaction networks. This way of abstracting the principles of chemical synthesis also allows for generating and proposing such reactions as part of the synthesis process that are not reported in literature so far.

Combinatorial methods from systems biology are employed to identify individual reaction pathway alternatives in the generated networks. A multi-stage approach of combinatorial and optimization-based methods is proposed to assess not only economic and ecological, but also topological aspects of the reaction pathways to allow for statements on the robustness of a design task. Experimental data, if available, is retrieved from a data base of organic reactions and included into the evaluation process. Embedded into a model-based product-process design, synthesis pathways towards tailored biofuel candidates can be systematically identified and proposed for further experimental investigations. In a subsequent step, the potential of blending unconverted intermediates and desired product while maintaining imposed property constraints is elucidated for the sake of increasing feedstock utilization.

Concluding this contribution, generation and evaluation of reaction networks is demonstrated by the example of two case studies, targeting the synthesis of bio-based diesel fuel surrogates. The synthesis of 3-MTHF starting from itaconic acid is topic of the first case study. This process was already investigated in detail in the Cluster of Excellence. It is shown that the computer-generated reaction network comprises a plethora of reactions that are so far not reported in literature. Likewise, the number of pathways identified by applying the evaluation routine distinctly increases in comparison to the results derived from manually assembled networks. Economic evaluation leads to the statement that this combination of feedstock and product suffers applicability due to the currently high prices of itaconic acid, which also cannot be compensated by integrating unconverted intermediates. Thus, the investigation of structurally similar compounds 2-BF and 2-BTHF is proposed and performed. Despite structural similarity, these substances can be derived from furfural, which is currently traded at lower market prices than itaconic acid. The analysis reveals that these substances can be produced sustainable from the provided feedstock, especially by integrating unconverted intermediates and final product. Under the presumed assumptions, the synthesis of 2-BTHF from furfural is economically viable under today's market conditions.

The presented methods are provided in a single software package. This computer-based

approach accelerates and supports the scientific process of biofuel identification by providing detailed reaction networks with a high information density in a short time span. With the information of promising reaction pathways at hand at already very early stage of the assessment, experimental investigation campaigns can be supported or even guided to achieve highest information gain with least effort.

1 Introduction

The availability of carbon feedstock is crucial for mobile propulsion, power generation and chemicals manufacturing and consequently the maintenance of nowadays standards of living. Although fossil carbon resources are currently abundantly available, the increasing demand will steadily deplete the residual amounts. Anticipating the future growth and development of developing markets in India, China, South East Asia and South America, it is obvious that the demand for fossil resources will rather increase than decrease (Sari and Soytas, 2007), affecting not only the costs of chemicals manufacturing, power generation and domestic heating, but also mobile propulsion and global cargo distribution. In addition, emissions stemming from the combustion of fossil carbon resources increase the atmospheric CO_2 concentration, acting as a key driver of global warming (Metz, 2007). Although the need for alternative energy carriers is evident, establishing them in the market is difficult since the requirements are high: long-term replacement of fossil fuels, replenishment on a short time-scale and abundant availability are mandatory prerequisites. In addition, emissions from production and combustion processes should be avoided or, if inevitable, should not contribute to the accumulation of harmful substances in either soil, water or air.

Fuels from biomass are one promising alternative to fossil fuels; biomass is available in sufficient amounts and replenishes on a reasonable time scale. Liquid energy carriers from biomass have attracted great interest in research and industry within the last decades (Naik et al., 2010). Their high energy densities and similarity to fossil fuels in terms of thermophysical properties, short and long term storage and distribution allow to maintain the concept of internal combustion engines with only minor modifications. In contrast to the combustion of fossil energy carriers, combustion of biobased energy carriers releases only CO_2 that is already part of the global carbon cycle (Post et al., 1990).

There is a manifold of challenges to face in biorenewable fuel production.The biomass feedstock has an oxygen to carbon ratio up to 1:1, leading to thermophysical feedstock properties that vary significantly from those required to serve as gasoline or diesel fuel. At the same time, the oxygen is present in various functional arrangements such as hydroxyl, ether, carbonyl and carboxylic acid groups, where each of these groups requires different processing steps for their refunctionalization or cleavage. Therefore, processing biomass is significantly different to the processing of crude oil (Daoutidis et al., 2013). However,

this feedstock shift must not be considered a threat, but rather an opportunity to critically review and eventually redesign existing processing paradigms. The challenge in the context of biorenewable fuel production is to establish efficient processing steps and truly sustainable value chains (Marquardt et al., 2010, Victoria Villeda et al., 2012).

1.1 The multi-dimensional context of sustainable biofuels

Sustainable production that considers economic, ecological and social aspects (Fiksel, 2002) has to account for the complete biofuel life cycle consisting of biomass cultivation, biomass processing to biofuels, biofuel combustion and the interactions of these processes with the ecosystem (cf. Figure 1.1). CO_2 serves as a major substrate for biomass growth through photosynthesis (Post et al., 1990). Large-scale biomass cultivation may have a negative impact on the ecosystem, e.g. based on land-use change (Kim et al., 2009) and additional secondary effects due to water consumption and the use of fertilizer. Emissions from biofuel combustion (e.g. CO_2, soot or NO_x) and by-products of the manufacturing process (e.g. CH_4) constitute hazards to the ecosystem (Fearnside, 2000) while the ecosystem itself influences biomass cultivation in terms of climate conditions, e.g., temperature and humidity. Hence, a complex system has to be accounted for when discussing the sustainability of biofuel value chains.

First generation biofuels such as fermentative bioethanol production from corn or sugarcane and catalytic production of fatty acid methyl esters (FAME) from plant oil already exceed a global annual production of 120 billion liters (OECD/FAO, 2011). However, these biofuels only make use of the sugar-, starch- and oil-bearing parts of the plants and compete with the food chain - either indirectly by using arable land for cultivation, or directly by increasing the market price of crops (FOA, 2010). In contrast, lignocellulosic biomass

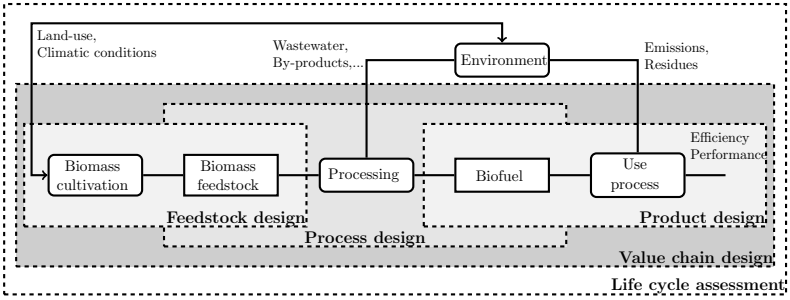


Figure 1.1 – Abstraction of environmental system influencing biofuel value chains (adapted from Victoria Villeda et al. (2012))

such as straw and bagasse is a material which is not competing with human nutrition and can also be converted into bioethanol (Bjerre et al., 1996). Novel approaches also propose to use algae as lignocellulosic feedstock for both, bioethanol (Li et al., 2014) and biodiesel production (Viêgas et al., 2015), since algae constitutes a feedstock whose cultivation is independent of arable land.

Research in bioethanol production has focussed on improving the individual steps of bioethanol from lignocellulose, leading to energy efficient pretreatment and highly selective fermentation. However, the yield of lignocellulosic bioethanol production is lower than the one starting from corn or sugar cane and it requires more efforts in the pretreatment of the feedstock (Sarkar et al., 2012).

Biomass-to-Liquid (BtL) processes that rely on the Fischer-Tropsch-Synthesis can, theoretically, process biomass independently of its heritage and composition and require less pretreatment effort. They furthermore allow for producing diesel fuel from lignocellulosic feedstock (Anderson et al., 1984). To this end, the biomass is broken down to C_1 building blocks (syngas) with subsequent catalytic recombination to form long chain hydrocarbons. However, a targeted production of a certain compound is aggravated through the nature of the used catalyst, which always yields a product spectrum rather than a defined substance (Dry, 2002).

Besides their individual shortcomings, the mentioned processes have in common that the molecular structure of the desired biofuel is predetermined. However, when considering the molecular structure as a degree of freedom, the tailoring of novel, defined fuel structures obtained from targeted refunctionalization of biomass monomers offers the opportunity of tailoring molecules to specific applications (Janssen et al., 2011, Hoppe et al., 2016). Combining a highly selective production process and a product that is in optimal accordance with its use process presumably yields a fuel that outperforms existing biofuels. In order to achieve a biofuel as sustainable and efficient as possible, not only the production process has to be accounted for, but rather the complex interactions of influences from the specific aspects of the value chain.

Consequently, 4 major tasks arise in the context of sustainable production of novel fuels (Victoria Villeda et al., 2012):

(i) Biofuels with desired engine-relevant properties need to be identified,

(ii) novel optimized production processes have to be elaborated for viable and sustainable market positioning,

(iii) product- and process-related impacts on the ecosystem have to be accounted for, and

(iv) optimal designs of biomass cultivation and distribution have to be worked out.

An optimal biofuel hence is a compromise between engine performance, production, environmental impact and sustainable feedstock cultivation, where optimality has to be defined in terms of measurable quantities covering all relevant aspects of the life cycle.

1.2 Towards the integration of product and process design

A model-based description of the value chain offers the opportunity to design each step from biomass cultivation to engine combustion. However, the multitude of interrelated dependencies, the variety of possible feedstock-product-process combinations and the complexity of the underlying phenomena render an entirely model-based description of the value chain a desired, but at least distant goal. Nevertheless, elaborated methodologies already exist for 2 distinct sub-problems, i.e., model-based product and process design. To achieve overall optimal solutions, product and process design need to be combined to form an integrated approach to the design of value chains. Isolated consideration of product and process performance might prevent the identification of an overall optimal biofuel because trade-offs between product performance and production effort cannot be established in an objective manner. While steps are already taken to extend the scope towards considering the entire value chain, the modeling detail of each step varies significantly.

Life cycle assessment (LCA) is a tool that is frequently employed to assess the economic, environmental and also social aspects of biofuels and their supply and production chains. It origins from approaches in the late 1960s to analyse the efficiency and environmental issues of resource use for the production of materials (see Hunt et al. (1996)). In biofuel production, LCA is often referred to as Well-to-Wheel (WTW) analysis. WTW analysis is used to evaluate biofuel value chains from the biomass feedstock to the energy at the wheel with respect to selected product and process performance criteria (Wang, 1999, Edwards et al., 2004, Yan et al., 2010).

Due to the high level perspective from which LCAs are carried out, they can be used to consider aspects of raw material extraction, manufacturing, transport, use and disposal of residues for a variety of biofuel value chains in a single approach. Almost every biofuel that was or is currently discussed as fossil fuel surrogate was analysed by LCA tools considering the influence of different aspects such as type of feedstock or pretreatment processes. Life cycle assessments for first generation biofuels (fuels from crops) were, amongst others, performed by Kim and Dale (2005) on bioethanol from corn and biodiesel from soybean and by Halleux et al. (2008) on bioethanol from sugar beet and rapeseed methyl ester. The topics of their investigations reached from evaluating environmental impact and economic performance of the value chain to assessing the influence of cropping systems on energetic efficiency. Considering second generation biofuels (fuels from cellulosic biomass), Xie

et al. (2011) evaluated energy use and greenhouse gas emissions for the Fischer-Tropsch synthesis and compared it to fuels from coal. Kumar and Murthy (2012) compared greenhouse gas emissions and energy requirement of different pretreatment processes for ethanol production from grass straws. Increases in production yields led to increasing interest in microalgae as feedstock in biodiesel production (Lardon et al., 2009, Campbell et al., 2011, Collet et al., 2013).

The variety of fuels considered and criteria assessed shows that LCA is a versatile tool to evaluate key performance indicators (KPIs) of biofuel value chains. In order to achieve the holistic assessment of the fuel's overall life cycle, the assessed LCA tools rely on rather simplified models to manage the complexity of the model. In addition, the identification of novel biofuels and the design of corresponding value chains is not the goal of the analysis. Thus, LCA is an analysis tool to assess the overall performance of identified and established biofuel value chains with a high level of abstraction.

More detailed evaluations are required to better understand the relations that exist between feedstock, product and process. Experimental campaigns were carried out to identify the impact of varying feedstock composition on the properties of biodiesel and biodiesel/diesel blends (Kinast, 2003, Canakci and Sanli, 2008, Gui et al., 2008). The data from such experimental campaigns was then employed by Chang and Liu (2009) to develop an integrated biodiesel processing model. It allows for simulating the process performance of entire biodiesel manufacturing chains for different feedstock compositions. Since biodiesel is a blend of multiple components, the relation between feedstock and product is of high interest and product design principles can be employed. Similar investigations were carried out for the production of ethanol from different feedstock (Huang et al., 2009) and different pretreatment processes (Aden and Foust, 2009). Contrasting biodiesel, ethanol is a defined compound. Hence product design is not applicable and the sole focus of experimental campaigns and process models is on improving the production process performance and sustainability (Cardona and Sánchez, 2007).

Garcia and You (2015) presented an approach towards product and process design of biofuels that is based on a process technology superstructure. They constructed a network that consists of 129 substances and 193 technologies, from which optimal processing pathways towards defined products are selected. Each processing step is described by a set of key characteristics such as energy consumption, losses and yields. This network is formulated as a mixed integer linear programming problem (MILP) and solved subject to a two-dimensional objective function comprising total annualized cost and environmental impact. Biomass feedstock is supplied from various sources, reaching from crops, sugar cane and soybeans over soft- and hardwood to microalgae. The comprised set of processing steps and feedstocks allows for constructing processes for first, second and third generation biofuels. As such, this approach amalgamates the previous superstructure approaches on

biomass to fuel (Kim et al., 2013), microalgae to fuel (Gong and You, 2014) and latest findings that were published in bioprocessing literature (for the complete list of 64 considered contributions we refer to Garcia and You (2015)).

While the aforementioned approach starts from a broad view of the biofuel processing chain and then increases the model detail, a methodology incorporating very detailed knowledge on product and process design methods was proposed by Gani and Pistikopoulos (2002). In a two-staged framework, an inverse problem is solved to determine the optimal set of properties of a not yet specified substance, leading to optimal process performance. In a subsequent step, a molecular design problem is solved to determine suitable molecular structures exhibiting the identified desired properties.

Although this framework integrates product and process design methods, it is still focussing on the processing part of the overall value chain. The product is only considered from the perspective of its production process rather than of its use process.

The research in the Cluster of Excellence "Tailor-Made Fuels from Biomass" (TMFB) at RWTH Aachen University (Marquardt et al., 2010, Janssen et al., 2011) follows the implications provided in recent contributions towards sustainable utilization of biomass (Sanders et al., 2007, Marquardt et al., 2010) by maintaining the rich molecular structure of biomass. Instead of breaking biomass to C_1 building blocks and synthesize complex molecules out of it, the synthesis power of nature shall be exploited to the maximum extent possible. Previous product-process design approaches are extended by considering the fuel's molecular structure as a degree of freedom. Such a strategy requires the identification of biofuel candidates as well as the proposition of novel reaction pathways towards their synthesis. A major challenge is the unavailability of experimental data, concerning both, the thermophysical properties of the biofuel candidates and the performance of their synthesis pathways and combustion.

In this context, Hechinger et al. (2010) presented an integrated product and process design approach for identifying gasoline surrogate candidates. Quantitative structure property relationships (Katritzky and Fara, 2005) were employed to predict the relevant thermophysical properties (enthalpy of combustion, normal boiling point and liquid density) of a set of proposed structures. Systematic process design identifies and evaluates pathways for their synthesis, based on a network of reactions constructed from extensive literature research. Optimal pathways are identified by solving an inverse problem (Voll and Marquardt, 2012b,a) optimizing an economic or ecological objective function. The product design part of this approach was later improved by Hechinger et al. (2012), who proposed a stepwise identification campaign for fuels suitable for spark ignition engines. Mathematically rigorous structure generators such as those presented by Gugisch et al. (2012) or Dahmen et al. (2013) are employed to generate an extensive pool of substances. For the substances in this pool, thermophysical properties are calculated with predictive property

models. Biofuel candidates are identified by comparing the predicted values against desired engine-relevant property data. Dahmen et al. (2012) extended this approach to diesel fuels by deriving and applying a predictive model for cetane numbers.

Concerning improvements in process design, a manually constructed holistic network, comprising the knowledge of synthetic chemistry, was compiled by Kowalik et al. (2012) and serves as basis for identifying synthetic pathways towards desired substances. However, such approaches do not contribute to the identification of novel reactions outside the known scope. Marvin et al. (2013) identified manually constructed reaction networks as major drawback in the design of biofuel synthesis pathways and computationally derived production routes to substances that perform optimal in biofuel-gasoline blends. An automated reaction network generator (Rangarajan et al., 2010) was employed to produce reaction pathways towards these substances. Like Besler et al. (2009) and Voll and Marquardt (2012a,b), they use a number of criteria to evaluate the networks and to find promising production routes. In addition, Marvin et al. (2013) include kinetic data for reaction schemes, thus enabling a detailed evaluation of the network. However the accuracy of the provided data is questionable since reaction kinetics are not only influenced by the occurring scheme, but also by molecular constitution of the reacting substances, catalysts, solvent and reaction conditions, which are not considered in their approach. In a similar manner, Yim et al. (2011) presented the use of the reaction network generator of Hatzimanikatis et al. (2004) to derive biochemical pathways towards 1,4-butanediol from glucose using *E. coli* and investigate further potentials assuming metabolic improvements. Both, Marvin et al. (2013) and Yim et al. (2011) employ reaction network generators that are set up on an empirical basis, meaning that reaction networks are constructed from generalized information that is already available in literature. Novel reaction mechanisms can thus not be proposed with their approach.

1.3 Contribution of this thesis

By defining the fuel's molecular structure as degree of freedom and employing molecular structure generators to generate a pool of candidate substances, a major challenge arises: the synthesis pathways towards promising compounds are likely to be unreported in literature. Manual assembly of reaction networks is not the best-suited means since only substances and reactions are collocated that were known at the time of assembly. As such, the final networks will either be small, incomplete or even void. This contribution provides an extension of the process design approach as carried out by Hechinger et al. (2010) to also assess synthesis pathways towards unreported substances. As presented by Marvin et al. (2013), reaction network generators can be used to derive synthesis networks towards

biofuel candidates. The advantage is twofold: on the one hand, reaction generators provides an initial set of reactions and intermediates for experimental campaigns; on the other hand, an adequate formulation of the reaction network generation is a means to elucidate unknown reaction schemes and provide alternatives to existing ones.

The approach of Marvin et al. (2013) relies to a certain degree on data from literature since networks can only be constructed based on user-provided reaction schemes. In the context of TMFB, where not only the molecular structure of a fuel is a degree of freedom, but also the synthesis pathways for their production, this can turn out to be a major drawback since novel reaction mechanisms cannot be identified in this manner. Therefore, a reaction network generator is required that constructs reaction networks only based on fundamentals of chemistry. Similar to molecular structure generators as for instance presented by Gugisch et al. (2012), such a generator derives all feasible derivatives of a provided set of molecules, but in addition elucidates the underlying reaction schemes, iteratively processes the generated substances and links them through reactions to provide a network reaching from substrate to target compounds. Such a formal reaction network generator was presented by Fontain and Reitsam (1991), however they stated that the complexity increases exponentially with the size of the investigated substances.

In this thesis, a reaction network generator is proposed that breaks down the combinatorial complexity of formal reaction network generation into smaller sub problems, without losing the formal character. By considering only the non-hydrogen atoms of the provided substrate(s), larger molecules can be processed. However, the formulation still offers the opportunity to include various kinds of empirical knowledge on e.g. reaction patterns and molecular constitution to target the network generation into a desired direction; however it is not reliant on such input. Furthermore, an approach for rapidly estimating the selectivity of individual reactions in the generated networks is proposed that is based on the molecular constitution of the network substances.

The generated networks serve as a basis for identifying and evaluating synthesis pathways for the production of promising biofuel candidates. The process design as presented by Hechinger et al. (2010), Voll and Marquardt (2012a,b) and Marvin et al. (2013) only assesses the performance of the fuel synthesis pathways in the network. However, the structure of the network bears valuable information. In systems biology and metabolic engineering, methodologies are already employed to derive statements on the number of available production routes and the importance of individual reaction steps in a metabolic network (Gagneur and Klamt, 2004, Papin et al., 2004). These combinatorial assessments are known as elementary mode analysis and yield non-decomposable pathways that link feedstock and target in a network (Stelling et al., 2002). Any valid steady-state flux distribution is representable by a non-negative linear combination of elementary modes (Papin et al., 2004).

The present contribution proposes an approach that joins combinatorial assessment of elementary modes with optimization-based pathway analysis to evaluate biofuel synthesis networks. It employs the concept of elementary mode analysis to first determine all non-further decomposable reaction sequences from substrate to target compound and then derives the optimal flux distribution as a linear combination of elementary modes in an optimization problem. This approach extends the determination of the best performing synthesis pathways by statements on the number of synthesis alternatives to derive the desired target and on the frequency of occurrence of individual network reactions in the elementary modes. This information is especially valuable in the lab-based verification of the identified pathways; computer-generated reactions may not be performable, thus the higher the number of alternatives, the more opportunities are available to synthesize a desired target. The number of occurrence of reactions represents their importance in the network; the more often they are employed in elementary modes, the more synthesis pathways rely on their real-life applicability. The proposed approach identifies such reactions and thus can serve as guideline for experimental investigation campaigns.

Furthermore, the process design as presented by Hechinger et al. (2010) and Voll and Marquardt (2012a,b) targets the production of pure substances. Marvin et al. (2013) presented a product-process design approach to the model-based design of mixtures by estimating the properties of biofuel-gasoline blends where gasoline is externally provided. However, neither one approach considers unconverted intermediates as constituents of the final mixture, although they occur in significant amounts. As combustion engines do not require a single compound fuel and the fuel's structure and composition is a degree of freedom, a mixture of the desired target and the unconverted intermediates is likely to increase the performance of the production process while maintaining the required properties. This thesis is the first scientific contribution to propose an approach to determine if and to which amount a target compound can be mixed with unconverted intermediates to form a biofuel blend. An inverse problem is formulated to evaluate the contribution of each unconverted intermediate. Linear mixing rules are employed to determine the properties of the resulting fuel blend.

This thesis is structured as follows: Chapter 2 gives an overview on the underlying graph-theoretical formalisms that are implemented to describe molecules, reactions and networks. These formalisms are required to abstract the chemistry and allow for automated reaction network generation. Furthermore, it presents the principles of reaction network generation. Chapter 3 presents how empirical knowledge allows for narrowing the network generation to a desired molecular space. Chapter 4 presents an estimation of the selectivity of reactions by identifying symmetric configurations of functional motifs in the network substances. Furthermore it is shown how several networks can be combined and how the network is reduced to only those reactions that participate in the synthesis of a desired substance.

Chapter 5 presents the novel, multi-stage evaluation strategy for reaction networks. Based on graph theory, the networks are decomposed into elementary, non-decomposable linear sequences of reactions, called elementary modes, which represent the feasible network pathways. An optimization-based formulation identifies the optimal combination of elementary modes subject to an economically and/or ecologically motivated objective function. The evaluation routine is complemented by an analysis to integrate unconverted intermediates and the desired product. It provides a first glimpse towards the possibilities of producing mixtures instead of pure substances.

Chapter 6 presents 2 case studies. The first focuses on comparing the presented methodology against manually assembled reaction networks to show the benefits of computational a reaction network generation. The production of 3-MTHF from itaconic acid was chosen as a reference. Automatic reaction network generation provides a network one order of magnitude larger in terms of substances and reactions contained, compared to the manually assembled one. Economic analysis of the reaction pathways showed, that the production of 3-MTHF lacks economic feasibility due to the currently high market prices of the feedstock itaconic acid.

The second case study evaluates the production of 2-BF and 2-BTHF, which were identified in TMFB as highly relevant substances by qualifying as biofuel based on their thermophysical properties. Although structurally similar to 3-MTHF, they can be derived from furfural, which is currently available at lower market prices than itaconic acid. Pathways for the synthesis of these substances are then identified and evaluated, first targeting the production of a pure substance, and secondly including unconverted intermediates into the product stream.

The focus of this contribution is to provide means for identifying and ranking different production alternatives and provide guidelines for further investigations. Consequently, the evaluation results should not be expected to yield results comparable to the level of detail and the accuracy of those derived from detailed flowsheets.

2 Reaction network generation: Introduction to the basic algorithm

Reaction network generators are computational tools that serve as a means to elucidate reaction mechanisms and generate reaction networks. They have been subject of academic research for the past 4 decades (Corey and Wipke, 1969). Their importance was underlined by awarding the Nobel Prize in chemistry in 1990 to James E. Corey (James, 1993) for his work in the field of theory and methodology of organic synthesis (Corey and Cheng, 1989). With increasing computational power, the ability to solve more and more sophisticated and detailed tasks emerged, such that reaction generators are now an essential tool in various fields of chemistry and engineering. Reaction network generators are employed in the design of drugs and chemicals (Fontain and Reitsam, 1991), combustion (Song, 2004) and pyrolysis modeling (Broadbelt et al., 1994), petrochemical processing (Quann and Jaffe, 1992) and metabolic engineering (Hatzimanikatis et al., 2005).

The task of a reaction network generator is to compute networks where one or multiple substrates are modified in chemical reactions under the presence of reactants (chemical substances that are consumed during the progress of a reaction) and to link them with target compounds. A network is generated in stages, where the substances generated in one stage serve as substrates of the subsequent one. The generation process can either be constrained by including empirical restrictions such as reaction rules or constraints on the molecular constitution or even be reliant on such knowledge to generate the reaction networks. Common to all reaction network generators is the recursive character; every alteration is applied to each substance in each stage of the network. Thus, reaction network generators require a stopping criterion, which is chosen according to the formulation of the generator. They either proceed until a certain substance is found or no more modifications can be applied to the generated substances (Fontain and Reitsam, 1991). Reaction network generators that include reaction kinetics further offer the opportunity to compare the production rate of substances to a user-defined reference rate (Song, 2004). The generation ends when the production rates of all substances fall below this threshold.

Reaction network generation can be applied in 2 directions: either by following the progress of reactions and generating the derivatives of a substance (*forward or total synthesis*), or by inverting the reactions to find predecessors or building blocks of the target

compound (*retro-synthesis*).

According to Tomlin et al. (1997), reaction network generators need to satisfy five essential requirements, that are, independent of the specific area of application, common to all network generators:

(i) Unambiguous representation of molecules and reactions,

(ii) internal representation of molecules,

(iii) internal representation of reaction rules,

(iv) iterative application of alterations to the molecular structure of all provided and generated molecules for systematic network generation, and

(v) employment of a systematic procedure to limit combinatorial explosion.

An unambiguous description of reactions and molecules is inevitable for unique labeling of the generated substances. Thus, mathematical abstractions of molecules and reactions have to be defined for internal representation. Several possibilities have been developed to describe molecules and reactions such that they can be assessed and processed by computational algorithms. This thesis restricts its scope to the two-dimensional (2D) structure of molecules, since considering the three-dimensional (3D) structure does not increase the level of detail in the presented approach, but only the amount of data to be processed.

The fifth requirement of Tomlin et al. (1997) is most important for the computational applicability of reaction network generators. With increasing size of the substrates, the number of products increases exponentially. Chemical fundamentals and empirical information on reactions and molecules need to be provided by the algorithm and or the user to target the generation process, leading to different degrees of rigor of the reaction network generation.

Ugi et al. (1979) postulate that reaction network generation can be categorized into 3 types of rigor, which are

(i) empirical methods, where the networks are generated upon reaction-specific data derived from reaction libraries,

(ii) semiformal methods, in which the generated reactions are derived from a set of generic reaction rules, and

(iii) formal methods based on graph theory, performing reactions only based on the valence rule of the atoms.

Out of these, only formal methods allow for elucidating unresolved chemical reaction mechanisms, but come at the cost of exponential growth of the number of generated substances and reactions. This restricts formal methods to the investigation of networks with rather low molecular weight substances.

Synthesis applications often employ substances with a comparatively high number of atoms. In reaction network generation, this asks for a guided generation process, which is realized by introducing empirical knowledge into the network generation process. Such restrictions need to be chosen carefully, since the benefit of targeting the network generation comes at the cost of a reduced product spectrum and the danger of excluding interesting reactions from further consideration.

The subsequent section gives an overview on different reaction network generators presented in literature. This overview is then used to differentiate the reaction network generator presented in the remainder of this chapter from the ones available in literature.

2.1 Reaction network generators - a literature review

The very first reaction network generators were developed to support the chemical synthesis of organic compounds, formulated as empirical retro-synthetic problems. The first reaction network generator was OCSS (**O**rganical **C**hemical **S**imulation of **S**ynthesis), proposed by Corey and Wipke (1969). OCSS was the predecessor of the more commonly known software package LHASA (**L**ogic and **H**euristics **A**pplied to **S**ynthetic **A**nalysis) (Corey et al., 1972), which is under continuous development concerning the incorporated database of retro-synthetic reactions. Since then, various retro-synthetic software tools were presented, such as SECS (Wipke et al., 1977), SYNCHEM (Gelernter et al., 1977), EROS (Gasteiger and Jochum, 1978), SYNGEN (Hendrickson, 1990) and HOLOWIN (Barberis et al., 1996). They differ in the internal representation of reactions and molecules and in the construction and pruning of the reaction network. While retro-synthetic tools support the design of organic syntheses, forward reaction network generators are commonly employed to a wider set of applications. Very early formulations were used to describe conversion processes in petrochemical refineries (Liguras and Allen, 1989a,b, McDermott et al., 1990, Quann and Jaffe, 1992). Due to the complex character of crude oil, the description was not targeted at tracing one certain substance, but rather describing the composition, the occurring reactions and the properties of the resulting mixture. Quann and Jaffe (1992) described individual hydrocarbons as a vector of structural increments; several of these vectors are combined with individual weights to form the vector of a hydrocarbon mixture. Reactions are introduced as modifications of the composition of the mixture vector.

The graph-theoretical *Bond Electron Matrix*-notation of molecules (Ugi et al., 1979) led

to formal reaction network generation. Fontain and Reitsam (1991) introduced the first formal reaction network generator RAIN (**R**eaction **A**nd **I**ntermediate **N**etworks) with the aim of deducing reaction pathways between substrates and target compounds, while the generation is only guided by a small set of constraints.

Besides the graph-based representation of reactions and molecules, linguistic description is an alternative frequently used in reaction network generation. Prickett and Mavrovouniotis (1997a,b,c) introduced the **R**eaction **D**escription **L**anguage (RDL) that incorporates language-like formalisms. The syntax consists of a sequence of individual commands that sum up to a stepwise description of the performed reaction. Molecules are described as simple text strings and reactions take place according to defined, sequential modifications. This formalism is then provided to a network generation routine that modifies the specified substrates based on the user-defined rules. RDL was extended by Hsu et al. (2008) to the representation of catalytic reactions.

KING (**Ki**netic **N**etwork **G**enerator) by Di Maio and Lignola (1992) is the first application of network generation to the field of combustion chemistry. It was introduced for the representation of complex networks occuring in combustion processes. It uses graph-theoretical description of molecules and reactions. Reaction kinetics were included by using a library of elementary reactions that contains the reactions of the combustion mechanisms. Broadbelt et al. (1994) applied this formalism to the description of pyrolysis degradation systems, Wong et al. (2004) extended it to the production of nano-particles, Hatzimanikatis et al. (2004) to biochemical transformations, Kruse et al. (2002) to polymerization and depolymerization and Khan et al. (2009) to the formation of tropospheric ozone. Other graph-based descriptions of combustion reactions were introduced by Warth et al. (2000) through EXGAS and by Ratkiewicz and Truong (2003) in the form of COMGEN. EXGAS is a tool that generates detailed kinetic models of the gas phase oxidation of alkanes and ethers. The reactions are generated relying on a database which contains information on particular species and also on generic elementary reactions from the chemistry of hydrocarbon oxidation. EXGAS is linked to additional software packages providing the generated reaction networks with information on thermophysical properties and kinetic data. The derived models are directly usable for the simulation of the reaction mechanisms. COMGEN identifies uni- and bimolecular sub-patterns in the substances and applies generic reaction patterns for their modification. The most common reaction generator in combustion engineering is the RMG (**R**eaction **M**echanism **G**enerator) package (Song, 2004) developed at Massachusetts Institute of Technology (MIT), which provides a larger set of elementary reactions and also incorporates pressure- and temperature-dependence of the developed mechanisms.

Increasing interest emerged in the computational derivation of metabolic networks. The complexity of metabolic networks demands for a computational approach to systemati-

cally identify the metabolisms that can be performed by the organism under consideration. Hatzimanikatis et al. (2005) extended the network generator of Broadbelt et al. (1994) to explore the diversity of metabolic networks. The results were astonishing as a multitude of novel biochemical routes to the synthesis of the investigated substrate-product combinations were unraveled. The network generation was coupled with a thermodynamic assessment of the production pathways to identify the thermodynamically most favorable ones. Faeder et al. (2005) and Blinov et al. (2006) introduced a package for dynamic rule-based generation of reaction networks for protein-protein interactions, which are particularly prominent in signal transduction. Mayeno et al. (2005) presented the reaction network generator BioTRANS that predicts the metabolites resulting from the exposure of organisms to 4 common drinking water pollutants and generates the emerging substances of enzymatic processing within the organism. SynBioSS (**Syn**thetic **Bio**logy **S**oftware **S**uite) developed by Hill et al. (2008) is one of the first implementations made available online. It allows for dynamic modeling and simulation of metabolisms of synthetic biological systems; kinetic parameters are derived from a compiled database of metabolic reactions.

Within the last years, the generated networks became a basis for early stage evaluation of large scale production processes. Moity et al. (2014) recently described the in-silico design of pathways towards biobased solvents starting from itaconic acid as substrate. Yim et al. (2011) generated pathways using the reaction network generator of Hatzimanikatis et al. (2005) to derive pathways for the direct enzymatic synthesis of 1,4-butanediol from common metabolic intermediates. The generated model was used for optimizing the anaerobic operation of *E. coli*. Recently, Marvin et al. (2013) proposed to use the language-based software RING (**R**ule **I**nput **N**etwork **G**enerator) (Rangarajan et al., 2012a,b) for simultaneous generation and evaluation of biomass upgrading routes for the synthesis of biobased components to be used in biofuel-gasoline blends.

An overview of intensively discussed reaction network generators in literature is presented in Table 2.1. Along with their naming comes a short description of the field of application, the underlying formalism and its degree of rigor, corresponding to the classification of Ugi et al. (1979). The ordering is chronologically by their first appearance in literature.

Table 2.1 – Reaction network generators, their applications and formulations

Year/Name/Publications	Application and formal basis	Type[1]
1991 RAIN (Fontain and Reitsam, 1991)	General formal system for the generation of reaction networks (i) Molecules are provided in matrix notation and converted into combinations of valence schemes (ii) Alteration of molecules through modification of valence schemes, reactions are not user-defined but internally generated with respect to mathematical and chemical constraints	f
1992 KING (Di Maio and Lignola, 1992)	Networks of combustion chemistry of hydrocarbons and related compounds (i) Matrix notation of molecules (ii) Alteration of molecules through reaction matrices, reactions schemes are user-defined	sf
1992 LHASA (Johnson et al., 1992, Johnson and Marshall, 1992a,b)	Generation of synthesis networks by retro-synthesis in organic chemistry (i) Identification of strategic bonds that can be split (ii) Database of retro-synthetic reactions, reactions are applied to strategic bonds	e
1992 SOL (Quann and Jaffe, 1992)	Description of composition, reactions and properties of complex hydrocarbon mixtures (i) Molecules are described as a vector of structural increments, mixtures are linear combinations thereof (ii) Reactions modify the composition of structural increment vectors	sf

Continued on next page

[1] e = empirical, sf = semiformal, f = formal

Table 2.1 – continued from previous page

Year/Name/Publications	Application and formal basis	Type
	Modeling of gas phase pyrolysis, biochemical reactions and nano-particle synthesis	
1994 NETGEN (Broadbelt et al., 1994)	(i) Matrix notation of molecules (ii) Alteration of molecules through reaction matrices (iii) Elementary reactions are lumped into reaction families, parameters for Arrhenius equation are estimated by means of linear free energy relationships	sf
1997 RDL (Prickett and Mavrovouniotis, 1997a,b,c)	Modeling of complex reaction systems by applying sequences of elementary steps for general applications (i) Linguistical representation of molecules (ii) Linguistical representation of reaction steps, introduction of reaction description language	sf
2000 EXGAS (Warth et al., 2000)	Gas phase oxidation networks of gasoline-related compounds (alkanes and ethers) (i) Tree-like description of molecules and radicals (ii) Incorporation of generic elementary hydrocarbon oxidation reactions and database reactions for particular species	sf
2003 COMGEN (Ratkiewicz and Truong, 2003, 2006)	Network of elementary reactions of hydrocarbon gas phase chemistry, kinetic simulations (i) Molecular species are internally stored as connectivity tables and externally as linguistic constructs (ii) Reactions are externally provided as 1D-string and internally as connectivity tables, identifying patterns and modifying the connectivity of the molecules (iii) Generic rate estimates for reaction classes are used for kinetic simulations	sf

Continued on next page

Table 2.1 – continued from previous page

Year/Name/Publications	Application and formal basis	Type
2004 BNICE (Hatzimanikatis et al., 2004, Hatzimanikatis et al., 2005)	Construction of biological reaction network, subsequent thermodynamic evaluation (i) Matrix notation of molecules (ii) Alteration of molecules through reaction matrices	sf
2004 RMG (Song, 2004)	Rate-based combustion network generation of alkanes and similar compounds (i) Molecules are encoded as chemgraphs containing 3 attributes: chemical elements, free electrons and chemical bonds (ii) Reactions are clustered in reaction templates which consist of a set of generic, elementary reaction steps that modify the attributes of the chemgraphs (iii) Reaction family kinetics are retrieved from a rate library	sf
2005 Bio-TRANS (Mayeno et al., 2005)	Reaction network generation for organic chemical mixtures, assessment of kinetic properties (i) Linguistic description of molecules or molecular mixtures (ii) Linguistic set of reactions rules	sf
2006 Bio-NETGEN (Blinov et al., 2006)	Biological reaction network generation, dynamic network simulation (i) Graph-theoretical description of molecular entities, vertices are functional units of the molecule, edges represent inter- or intramolecular bonds, molecules can have a set of additional variables (e.g. a status) and fixed attributes (e.g. molecular weight) (ii) Reaction rules alter the variable attributes, remove/add edges and replace molecular entities with one or multiple molecular entities	sf

Continued on next page

Table 2.1 – continued from previous page

Year/Name/Publications	Application and formal basis	Type
2010 RING (Rangarajan et al., 2010, 2012b, Marvin et al., 2013)	Generation and analysis of complex reaction networks of various chemistries (i) SMILES notation of molecules (ii) Reaction schemes are provided a priori and compiled into RDL	sf

Since TMFB is considering the fuel's molecular structure as a degree of freedom (Marquardt et al., 2010), the reaction pathways and mechanisms leading to identified biofuel candidates will likely be unreported in literature. The ability to propose reactions based only on fundamentals of chemistry without any empirical knowledge included is therefore a prerequisite for the employed generator; otherwise, the generation process would be limited to the known space of chemistry. Only through a formal formulation of reaction network generation, novel reaction pathways towards identified, high potential biofuel candidates can be proposed without being reliant on prior experimental investigations or experimental data in the first place. In fact, the reaction network generator serves as a means to target experimental investigation by providing detailed information on intermediates and available synthesis pathways.

So far, RAIN proposed by Fontain and Reitsam (1991) is the only formal generator presented in literature. However, this generator lacks the ability to process larger molecules due to computational limitations stemming from the exponentially increasing complexity of formal network generation. Therefore a novel reaction network generator (named ReNeGen, **Re**action **Ne**twork **Gen**erator) is introduced that is capable of performing formal reaction network generation, sharing RAIN's perception of atoms as an assembly of so-called valence schemes, but processes them in a way that breaks down the generation process into subproblems of smaller complexity. While RAIN considers all atoms of a molecule (including hydrogen) and thereof generates the derivatives, ReNeGen only considers non-hydrogen atoms in the combinatorial part of the calculation to drive down complexity. Therefore the procedure is split into 2 sub-routines: One to compute the covalent bonds each atom establishes, and one to determine the adjacency of the atoms. The results from these 2 sub-problems are then superposed to form the products of given substrates. This approach is expected to be computationally beneficial since the sizes of the individual problems are considerably reduced compared to the algorithm proposed by Fontain and Reitsam (1991). As such, larger molecules, as frequently encountered in the processing of biorenewables (Werpy et al., 2004), can be processed. In addition, the proposed implementation offers the ability to include empirical knowledge to a user-defined

degree to target the network generation process.

The following sections will introduce the underlying formalisms of ReNeGen. Basic concepts of cheminformatics, such as representation of molecules, reactions and networks are presented in detail, since they form the basis of the subsequently presented formalisms of ReNeGen. The reaction network generation for the substrates formic acid and hydrogen is used as an accompanying example to visualize the algorithmic procedures.

2.2 Computational representation of molecules

Molecules are computationally either represented through character strings or mathematical graphs. Both descriptions will be used in the context of this thesis, therefore they are introduced in this section. Formalisms and expressions of graph theory are explained in more detail in the mathematical preliminaries that can be found in Appendix A.

When discussing reaction network generation, the definition of the terms *molecular ensemble* (ME) and *family of isomeric molecular ensembles* (FIME) is necessary. A ME is an arrangement of one or multiple chemical substances. Multiple MEs are isomeric if they share an equal sum formula, no matter of their molecular constitution. The set of all perceivable isomeric MEs to a given sum formula is called the family of isomeric molecular ensembles (FIME).

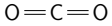
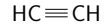
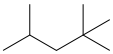
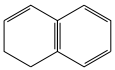
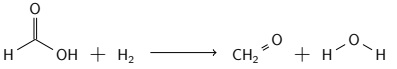
2.2.1 Molecules described by character strings

Character strings store the information of the 2D or 3D molecular constitution in a compact sequence of characters. Several established formalisms exist, of which SMILES (Weininger, 1988), SMARTS (Jeliazkova and Kochev, 2011) and InChi (Heller and McNaught, 2009) are most frequently applied. In this thesis, only SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification) is used.

SMILES is a language with a simple vocabulary for the description of atoms and bonds. The atoms are represented using their atomic symbols (C, H, O, N, S, etc.). Aliphatic atoms are written in uppercase and aromatic atoms in lowercase letters. Usually, hydrogen atoms are not explicitly included in the SMILES notation, as they do not add new information to the representation but only increase the length of the SMILES notation. A small grammar set is included to represent atomic arrangements. Branching is denoted by parenthesis notation, rings are represented by integer values. The number of rings is represented by the highest integer in the notation of a molecule.

SMILES notation is not restricted to the description of single compounds; it can also represent MEs as a single character sequence. Disconnected compounds are separated by a dot ".". SMILES notation can furthermore represent reactions, indicating the direction

Table 2.2 – Examples of SMILES notation of different structural motifs

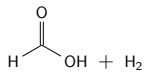
Structure	SMILES notation
 Propane	CCC
 Carbon dioxide	O=C=O
 Ethyne	C#C
 Iso-octane	CC(C)CC(C)C(C)
 2-Methylfuran	O1c(ccc1)C
 Naphthalene	c1ccc2cccccc2c1
 Formic acid hydrogenation	[HH].OC=O >>O=C.O

of the reaction by ">>". Examples of the previously described notations are presented in Table 2.2.

One molecule can be expressed by a multitude of SMILES notations (see Table 2.3). Introducing a canonical SMILES representation avoids ambiguous notations by labeling a molecule with a unique SMILES notation (Weininger et al., 1989). Unfortunately, there are several, but distinct canonizations available. Hence, in order to avoid ambiguous SMILES notation, the same routine/software always has to be employed to convert a SMILES notation into a canonical one. One such routine is provided in the Open Babel software package (OBoyle et al., 2011) that is used in this thesis for the generation of canonical SMILES. Example 1 presents the ambiguity of SMILES notation.

Example 1. *The SMILES codes denoted in Table 2.3 illustrate the ambiguity of SMILES notation of an ME comprising formic acid and hydrogen. The SMILES notation in the last column is the canonical one generated by Open Babel (OBoyle et al., 2011).*

Table 2.3 – Ambiguity of SMILES notation of formic acid and hydrogen

Structure	Possible SMILES representations				Canonical
	O=CO.[HH]	OC=O.[HH]	C(O)=O.[HH]	C(=O)O.[HH]	O=CO.[HH]
	[HH].O=CO	[HH].OC=O	[HH].C(O)=O	[HH].C(=O)O	

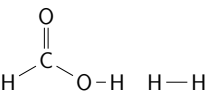
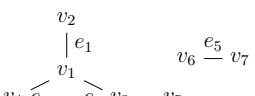
2.2.2 Molecules described by graphs

Graph-based approaches represent the structure of an atom, consisting of a core (which is the nucleus and electrons of all non-valence shells) and the valence electrons (electrons in the valence shell), as a graph. In a molecule, the individual atom cores are bound by sharing valence electrons in covalent bonds. In the mathematical description of a ME, the cores are considered as vertices V and the covalent bonds as edges E of graph $G(V, E)$. G is undirected, which is consistent with the undirected character of covalent bonding in molecules.

The number of atoms is $N_A = |V(G)|$; vertices are labeled from 1 to N_A, thus $V(G) = \{v_1, v_2, ..., v_{N_A}\}$. The number of bonds is $N_B = |E(G)|$; edges are labeled from 1 to N_B, thus $E(G) = \{e_1, e_2, ..., e_{N_B}\}$. An edge e is an ordered pair of 2 vertices v_i and v_j and can also be denoted as $e = \{v_i, v_j\}$. The atom set $A(G) = \{a_1, a_2, ...a_{N_A}\}$ denotes the types of all atoms in $V(G)$ in corresponding order; thus it is its atomic labeling. Example 2 presents the graph-based representation of the ME of formic acid and hydrogen.

Example 2. *Table shows the graph representation of formic acid and hydrogen, depicting the graph G(V,E), the vertex set V(G), the atom set A(G) and the edge set E(G).*

Table 2.4 – Graph representation $G(V, E)$ of formic acid and hydrogen, described by the vertex set $V(G)$, the atomic vector $A(G)$ and the edge set $E(G)$

Structure	G(V,E)	V(G)	A(G)	E(G)
		$\begin{Bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{Bmatrix}$	$\begin{Bmatrix} C \\ O \\ O \\ H \\ H \\ H \\ H \end{Bmatrix}$	$\begin{Bmatrix} \{v_1, v_2\} \\ \{v_1, v_3\} \\ \{v_1, v_4\} \\ \{v_3, v_5\} \\ \{v_6, v_7\} \end{Bmatrix}$

The arrangement of atoms in a ME can be represented through matrix notation. 2 formats are commonly employed, namely the adjacency matrix **AM** and the bond electron

Table 2.5 – **AM** and **BM** notation of formic acid and hydrogen

A(G)	**AM**	**BM**
$\begin{Bmatrix} C \\ O \\ O \\ H \\ H \\ H \\ H \end{Bmatrix}$	$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$

matrix **BM**. **AM** denotes the adjacency of vertices in a molecular graph. It is defined as

$$\mathbf{AM} = (am_{i,j}) \in \{0,1\}^{N_A \times N_A} \tag{2.1}$$

$$am_{i,j} = \begin{cases} 1 \text{ if } \{v_i, v_j\} \in E \\ 0 \text{ if } \{v_i, v_j\} \notin E \end{cases} \tag{2.2}$$

Since an atom i cannot establish a bond with itself, the diagonal entries $am_{i,i} = 0$.

The **BM** notation, defined as

$$\mathbf{BM} = (bm_{i,j}) \in \{0,1,2,3\}^{N_A \times N_A} \tag{2.3}$$

by Ugi et al. (1979), maintains the properties of the **AM** notation and further includes the bond order. An entry $bm_{i,j} = 1$ represents a single, $bm_{i,j} = 2$ a double and $bm_{i,j} = 3$ a triple bond between atoms v_i and v_j. Both, **AM** and **BM** notation, are employed in this thesis. The **BM** notation of MEs forms the basis of reaction generation while the **AM** notation is used for determining the uniqueness of generated MEs. Example 3 illustrates the difference of **AM** and **BM** notation.

Example 3. *Table 2.5 presents **AM** and **BM** notation of formic acid and hydrogen. The matrices are arranged in accordance to the ordering of the atom set A(G). Note the main difference between these notations, which is the representation of bond orders in **BM** notation (e.g. in $bm_{1,2}$).*

Constitutional information of a ME is contained in its **BM** notation. The vector **ve**, defined by

$$\mathbf{ve} \in \{0,1,2,3\}^{1 \times N_A}, \tag{2.4}$$

contains the number of valence electrons ve_i that are bound in covalent bonds of every

atom i. This value is calculated by the of row i in the **BM** notation according to

$$ve_i = \sum_{j=1}^{N_A} bm_{i,j}. \tag{2.5}$$

N_{VE} represents the total number of valence electrons of the ME by the sum of the valence electrons per atom via

$$N_{VE} = \sum_{i=1}^{N_A} ve_i. \tag{2.6}$$

The electric charge q of atom i is retrieved by subtracting the number of bound valence electrons from the core charge c_i, such that

$$q_i = c_i - ve_i. \tag{2.7}$$

The core charge of each atom is provided in the periodic table, see e.g. Furniss et al. (1989). The electric charge Q of a ME is calculated by

$$Q = \sum_{i=1}^{N_A} q_i. \tag{2.8}$$

The number of single bonds (SB), double bonds (DB) and triple bonds (TB) of atom i is calculated via

$$N_{SB,i} = \sum_{j=1}^{N_A} bm_{i,j} \quad \forall i : bm_{i,j} = 1 \tag{2.9}$$

$$N_{DB,i} = \sum_{j=1}^{N_A} bm_{i,j}/2 \quad \forall i : bm_{i,j} = 2 \tag{2.10}$$

$$N_{TB,i} = \sum_{j=1}^{N_A} bm_{i,j}/3 \quad \forall i : bm_{i,j} = 3. \tag{2.11}$$

As only C, O, and H are considered here, the number of occurrences of an carbon (N_C), oxygen (N_O) and hydrogen (N_H) in a ME is also contained in VE, since every atom type has its own characteristic number of valence electrons,. N_C is received from accounting for the number of entries ve where $ve_i = 4$. Equivalently, N_O is received from accounting for the number of entries where ($ve_i = 2$) and N_H where ($ve_i = 1$). This framework can be extended to other atoms by accounting for their respective number of valence electrons. In case two atom types share the same amount of valence electrons, the information stored

in the atomic labeling $A(G)$ can be used to distinguish.

2.3 Computational representation of reactions

A chemical reaction is the conversion of a ME of substrates into an isomeric ME of products. The conversion is achieved by redistributing valence electrons by breaking existing and forming new covalent bonds. A comprehensive formalism for the computational description of chemical reactions was introduced by Dugundji and Ugi (1973). The MEs of substrates **B** and products **E** are denoted in **BM** notation. A reaction is denoted as a reaction matrix

$$\mathbf{R} = (r_{i,j}) \in \{-3, ..., 3\}^{N_A \times N_A}. \tag{2.12}$$

The values in $r_{i,j}$ represent the redistribution of valence electrons in $b_{i,j}$ which are caused by the reaction. Triple bonds are the highest chemically feasible bond order; therefore $r_{i,j} \in \{-3, ..., 3\}$. Since **B** and **E** are isomeric ensembles, **R** has to be of the size $N_A \times N_A$.

In accordance to chemical knowledge, the number of valence electrons N_{VE} remains constant during a reaction; no valence electrons are allowed to be added to or removed from the ME. Therefore, it is imposed that

$$\sum_{i=1}^{N_A} \sum_{j=1}^{N_A} r_{i,j} \stackrel{!}{=} 0. \tag{2.13}$$

Likewise to **B** and **E**, **R** is symmetric, since the alteration of the bond between atoms i and j is equivalent to the alteration of the bond between j and i. This requirement is expressed through

$$r_{i,j} = r_{j,i}. \tag{2.14}$$

For entries $r_{i,j} < 0$, the bond order of $b_{i,j}$ is reduced by $|r_{i,j}|$, while for entries $r_{i,j} > 0$ the bond order of $b_{i,j}$ is increased by $r_{i,j}$. Bonds can only be reduced to a minimum value of zero and increased to a maximum value of 3. Every other bond modification is not feasible. These constraints are expressed by

$$|r_{i,j}| \leq b_{i,j} \quad \forall r_{i,j} < 0 \tag{2.15}$$

and

$$r_{i,j} + b_{i,j} \leq 3 \quad \forall r_{i,j} > 0. \tag{2.16}$$

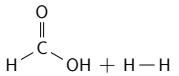
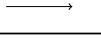
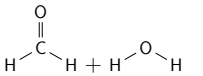
The reaction is performed by the addition of substrate matrix $\mathbf{B}$ and reaction matrix $\mathbf{R}$:

$$\mathbf{B} + \mathbf{R} = \mathbf{E}. \tag{2.17}$$

Example 4 presents matrix based notation of chemical reactions for the hydrogenation of formic acid to formaldehyde and water.

Example 4. *Table 2.6 presents substrate matrix* $\boldsymbol{B}$*, reaction matrix* $\boldsymbol{R}$ *and product matrix* $\boldsymbol{E}$ *for the hydrogenation of formic acid. Matrix* $\boldsymbol{R}$ *fulfills the imposed constraints on electron neutrality, cf. Equation* (2.13)*, on symmetry, cf. Equation* (2.14) *and positive bond orders, cf. Equations* (2.15) *and* (2.16)*.*

Table 2.6 – Computational notation of formic acid hydrogenation

B		**R**		**E**
$\begin{pmatrix} 0&1&2&1&0&0&0\\ 1&0&0&0&1&0&0\\ 2&0&0&0&0&0&0\\ 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1\\ 0&0&0&0&0&1&0 \end{pmatrix}$	+ $\begin{pmatrix} 0&-1&0&0&0&1&0\\ -1&0&0&0&0&0&1\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 1&0&0&0&0&0&-1\\ 0&1&0&0&0&-1&0 \end{pmatrix}$	=	$\begin{pmatrix} 0&0&2&1&0&1&0\\ 0&0&0&0&1&0&1\\ 2&0&0&0&0&0&0\\ 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0 \end{pmatrix}$
				

During the reaction, the covalent bonds between the carbon atom and the hydroxy group $r_{1,2} = r_{2,1} = -1$ *and the hydrogen atoms* $r_{6,7} = r_{7,6} = -1$ *are dissected. New bonds are formed between the dissected hydroxy group and one of the 2 hydrogen atoms* $r_{2,7} = r_{7,2} = 1$ *as well as between the carbon atom and the residual hydrogen atom* $r_{1,6} = r_{6,1} = 1$*.*

2.4 Formalisms of the reaction generator

The task of a formal reaction generator is to generate the family of isomeric molecular ensembles (FIME) to a provided substrate ME. To this end, ReNeGen perceives molecules is the same manner as RAIN (Fontain and Reitsam, 1991), i.e. by abstracting their covalent bonding by means of *valence schemes* (VS) and reactions based on *valence scheme*

transitions (VT) that convert valence schemes into one another. However, the underlying procedures of computing reaction products differ significantly. Fontain and Reitsam (1991) use the information of valence scheme transitions to compute reaction matrices $\mathbf{R}$ by introducing the information on valence scheme transitions in every feasible permutation, considering all atoms present in $\mathbf{B}$. This leads to the construction of matrices of size $N_A \times N_A$. The reaction matrices are then one by one applied to the substrate ME stated in $\mathbf{B}$. The complexity grows exponentially with every additional atom in $\mathbf{B}$, a common characteristic to formal reaction network generators (Fontain, 1995, Tomlin et al., 1997).

ReNeGen aims at reducing this combinatorial complexity. It considers only the non-hydrogen atoms in the combinatorial part of the generation. This modification builds on the fact that hydrogen atoms are always bound by single bonds and thus only add complexity (in terms of problem size), but not variation (in terms of bond orders) to the generation process. This approach leads to an entirely novel formulation of the formal network generation. The complexity is broken down from a single problem that exponentially grows with the number of atoms N_A to 2 smaller problems that grow exponentially only with the number of non-hydrogen atoms $\tilde{N}_A$. The first problem is the generation of valence scheme combinations. The second problem is the computation of adjacency schemes of the non-hydrogen atoms and their superposition with information on covalent bonding from the valence scheme combinations. This leads to the formation of *molecular bodies*, which denote adjacency and covalent bonding of the non-hydrogen atoms. The hydrogen atoms of the ME are afterwards used to equilibrate eventually existing formal electric charges in these molecular bodies. Excess hydrogen forms molecular hydrogen (H_2). For each sub-problem, generic constraints from fundamental understandings of chemistry are derived to confirm the feasibility of the output of each sub-problem. As such, the structure of the products is solely determined from modifications in the covalent bonding of the non-hydrogen atoms. This way, the complexity is kept to a minimum by constantly excluding non-feasible valence scheme combinations, atomic adjacency schemes and molecular bodies from further consideration.

In a postprocessing step, multiple instances of product MEs are identified and removed and the main substance of each ME is determined. The identification of the main substance is necessary since only the main product of a generated ME serves as substrate to subsequent reactions. The generated substances are included into a network of substances and reactions that is constructed based on the generator's output.

The following sections present the aforementioned aspects in more detail.

2.4.1 Valence schemes and valence scheme transitions

The atoms of a ME are bound in valence schemes. These schemes are defined individually for each of the implemented atom type (C, O, H) and describe how atoms distribute their valence electrons to covalent bonding. Each valence scheme is characterized by a specific combination of single bonds (SB), double bonds (DB) and triple bonds (TB). This framework can easily be extended to consider atoms outside the current scope by conveying their ways of establishing bonds to neighboring atoms into valence scheme notation. In events where 2 distinct types of atoms have the same number of valence electrons, they can be distinguished by additionally taking their atomic labeling into account.

The valence schemes of the implemented atom types are presented in Table 2.7. Equations (2.9)-(2.11) are employed to determine order and the number of each bond type for each atom and, on this basis, to assign a valence scheme.

Table 2.7 – Valence schemes (VS) of carbon, oxygen and hydrogen with number of valence electrons (ve), single bonds (N_{SB}), double bonds (N_{DB}) and triple bonds (N_{TB})

i	VS	ve	N_{SB}	N_{DB}	N_{TB}
1		4	4	0	0
2		4	2	1	0
3		4	1	0	1
4		4	0	2	0
5		2	2	0	0
6		2	0	1	0
7		1	1	0	0

The transition of valence schemes is the governing formalism of the reaction generation. A transition table $\mathbf{TT} \in \{0,1\}^{7\times7}$ states in an entry $tt_{i,j}$, whether a transition of valence scheme i to valence scheme j is allowed (1) or not (0). It is never allowed to perform transitions that change the type of an atom. An exemplary transition table is presented in Table 2.8. Enabled transitions are concatenated for every atom i in its set of valence scheme transitions VT_i, which contains the indices of the non-zero entries of row i in $\mathbf{TT}$.

Table 2.8 – Exemplary **TT**, every chemically feasible transition is enabled

	–C<	=C<	–C≡	=C=	–O–	=O	–H
–C<	1	1	1	1	0	0	0
=C<	1	1	1	1	0	0	0
–C≡	1	1	1	1	0	0	0
=C=	1	1	1	1	0	0	0
–O–	0	0	0	0	1	1	0
=O	0	0	0	0	1	1	0
–H	0	0	0	0	0	0	1

Example 5 presents valence schemes (VS) and valence scheme transitions (VT) of formic acid and hydrogen according to the valence schemes denoted in Table 2.7 and the transition table in Table 2.8.

Example 5. *Table 2.9 presents how the atoms of formic acid and hydrogen distribute their valence electrons to single, double and triple bonds. Aligning this information with Table 2.7 gives the valence scheme of each atom. Using the information provided in Table 2.8 leads to the valence scheme transitions. The valence schemes for hydrogen are only listed for the sake of completeness, they are not required in the further conduct of the reaction generator.*

Table 2.9 – Constitutional features, valence schemes (VS) and valence scheme transitions (VT) of the atoms in the exemplary ME

A(G)	**BM**	N_{SB}	N_{DB}	N_{TB}	VS	VT
$\begin{pmatrix} C \\ O \\ O \\ H \\ H \\ H \\ H \end{pmatrix}$	$\begin{pmatrix} 0&1&2&1&0&0&0 \\ 1&0&0&0&1&0&0 \\ 2&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&1 \\ 0&0&0&0&0&1&0 \end{pmatrix}$	2 2 0 1 1 1 1	1 0 1 0 0 0 0	0 0 0 0 0 0 0	2 5 6 7 7 7 7	$\{1,2,3,4\}$ $\{5,6\}$ $\{5,6\}$ $\{7\}$ $\{7\}$ $\{7\}$ $\{7\}$

2.4.2 Combination of valence schemes

The generation of the FIME has to consider every combination of valence schemes for each atom. Such a combination denotes how the atoms distribute their valence electrons

to covalent bonds in an molecular ensemble. These combinations are formed from the Cartesian product of the valence scheme transition sets VT_i. The set of valence scheme combinations VSC results to

$$VSC = VT_1 \times ... \times VT_{\tilde{N}_A}. \tag{2.18}$$

The number of valence scheme combinations, N_{VSC}, is

$$N_{VSC} = \prod_{i=1}^{\tilde{N}_A} |VT_i|. \tag{2.19}$$

A valence scheme combination is addressed by the index k with $k \in \{1, ..., N_{VSC}\}$, while the individual valence schemes of VSC_k are addressed by $VSC_{k,i}$.

It is necessary to ensure that a valence scheme combination VSC_k is feasible. A feasible valence scheme combination requires (i) that single, double and triple bonds $N_{SB}(VSC_k)$, $N_{DB}(VSC_k)$ and $N_{TB}(VSC_k)$, each appear in even numbers since each occurrence of a bond type needs exactly one corresponding counterpart, and (ii) if bond types appear only twice in a VSC_k, they have to stem from 2 distinct valence schemes. Example 6 presents the set of valence scheme combinations of the non-hydrogen atoms of the exemplary ME.

Example 6. *16 unique valence scheme combinations VSC_k can be constructed for the exemplary ME, which are listed in Table 2.10.*

The analysis of the number of bond types reveals that the following combinations are not feasible:

- *Combinations 2 and 3 fail since the number of double bonds is odd (no atom can connect to the double bond at the oxygen atom).*
- *Combination 5 fails since the number of double bonds is odd (no atom can connect to the double bond at the carbon atom).*
- *Combination 8 fails since the number of double bonds is odd (at least one double bond cannot be bound).*
- *Combinations 9-12 fail since the number of triple bonds is odd (no atom can connect to the triple bond at the carbon atom).*
- *Combination 13 fails although the amount of double bonds is even, but they belong to the same valence scheme.*
- *Combinations 14 and 15 fail since the number of double bonds is odd (at least one double bond at the carbon atom cannot be bound).*

Table 2.10 – The set of valence scheme combinations of the non-hydrogen atoms of formic acid and hydrogen

VSC_k	C	O	O	N_{SB}	N_{DB}	N_{TB}
1	−C(−)(−)−	−O−	−O−	8	0	0
2	−C(−)(−)−	−O−	=O	6	1	0
3	−C(−)(−)−	=O	−O−	6	1	0
4	−C(−)(−)−	=O	=O	4	2	0
5	=C<	−O−	−O−	6	1	0
6	=C<	−O−	=O	4	2	0
7	=C<	=O	−O−	4	2	0
8	=C<	=O	=O	2	3	0
9	−C≡	−O−	−O−	5	0	1
10	−C≡	−O−	=O	3	1	1
11	−C≡	=O	−O−	3	1	1
12	−C≡	=O	=O	1	2	1
13	=C=	−O−	−O−	4	2	0
14	=C=	−O−	=O	2	3	0
15	=C=	=O	−O−	2	3	0
16	=C=	=O	=O	0	4	0

This assessment already reduces the amount of feasible valence scheme combinations to only 5 out of 16.

2.4.3 Computing the adjacency of the non-hydrogen atoms

The previously derived valence scheme combinations VSC_k define which valence schemes of non-hydrogen atoms can occur simultaneously in a molecular ensemble. However, they only provide statements on the covalent bonds orders, not on the atomic adjacency in the ME. A single valence scheme combination can be feasible for different atomic adjacency arrangements. Vice versa, different valence scheme combinations may be applicable to one atomic adjacency. These adjacency arrangements are addressed by *adjacency schemes*, denoted as **AS**.

An adjacency scheme **AS** is defined similar to an adjacency matrix **AM**, but represents only the non-hydrogen atoms. An entry $as_{i,j}$ denotes whether atom i is connected to atom j ($as_{i,j} = 1$). Hence, an adjacency scheme is defined as

$$\mathbf{AS} = (as_{i,j}) \in \{0,1\}^{\tilde{N}_A \times \tilde{N}_A}. \tag{2.20}$$

Since **AS** are symmetric matrices and bonds between atoms are undirected, it is sufficient to compute the upper triangular matrices $\mathbf{AS}^U$ to describe the adjacency of the non-hydrogen atoms. Atom i can establish bonds up to ve_i connections to every atom but itself ($as_{i,i} \stackrel{!}{=} 0$). Since only the upper triangular matrix is considered, the number of atoms that can be adjacent to atom i is $\tilde{N}_A - i$. The maximum number of bonds N_b of atom i is

$$N_{b,i}^{max} = min(ve_i, \tilde{N}_A - i). \tag{2.21}$$

The minimum number of bonds is always

$$N_{b,i}^{min} = 0. \tag{2.22}$$

The actual value of established bonds towards other non-hydrogen atoms ($N_{b,i}$) can be less than $N_{b,i}^{max}$ since atom i does not have to establish bonds to non-hydrogen atoms only. It can also establish bonds to hydrogen atoms. In case that $N_{b,i}^{min} = 0$, atom i does not establish any bond towards a non-hydrogen atom, but only towards hydrogen atoms. Also, an established bond can involve more than one valence electron, meaning that an adjacency $as_{i,j} = 1$ can contain up to 3 valence electrons in the final molecule.

The adjacency of an atom i can take various arrangements, which are determined by generating the unique permutations of a set $a \in \{0,1\}^{1\times\tilde{N}_A-i}$, with N_b elements equaling 1, the residual $\tilde{N}_A - i - N_{b,i}$ elements equaling 0. The permutations of the set a are determined for every $N_b \in \{N_{b,i}^{min}, ..., N_{b,i}^{max}\}$ of atom i and are concatenated in a set A_i. The Cartesian product

$$AC = A_1 \times ... \times A_{\tilde{N}_A} \tag{2.23}$$

generates combined sets of atom-wise adjacency arrangements for all non-hydrogen atoms. The number of possible combinations of adjacency arrangements, N_{AC}, is the product of the cardinalities of the sets A_i, stated as

$$N_{AC} = \prod_{i=1}^{\tilde{N}_A} |A_i|. \tag{2.24}$$

A combination of adjacency arrangements is addressed by AC_l, where $l \in \{1, ..., N_{AC}\}$. The individual adjacency arrangements in AC_l are addressed by $AC_{l,i}$. Each AC_l is introduced into a zero matrix $\mathbf{AS} = (as_{i,j}) \in 0^{\tilde{N}_A\times\tilde{N}_A}$, starting at $as_{i,i+1}$, resulting in the adjacency schemes **AS**.

Example 7 presents the computation of the adjacency schemes for the exemplary ME.

Example 7. *The example of formic acid and hydrogen contains $N_A = 7$ atoms with $\tilde{N}_A = 3$ non-hydrogen atoms. Table 2.11 denotes the individual row configurations A_i for each atom i. Atom $i = 1$ can be encountered in 4 and atom $i = 2$ in 2 different adjacency*

Table 2.11 – Determination of the atom-wise row configurations

i	ve_i	$\tilde{N}_A - i$	$N_{b,i}^{min}$	$N_{b,i}^{max}$	$N_{b,i}$	A_i
1	4	2	0	2	0 1 2	{0,0} {1,0},{0,1} {1,1}
2	2	1	0	1	0 1	{0} {1}
3	2	0	0	0	0	$\emptyset$

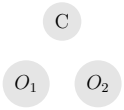
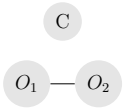
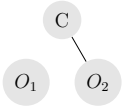
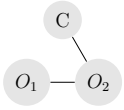
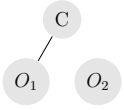
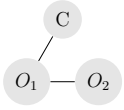
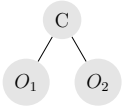
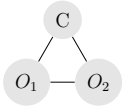
arrangements. The adjacency of atom $i = 3$ can only be an empty set since the set A_3 is of zero size. 8 combinations of adjacency arrangements emerge that are denoted in Table 2.12.

Table 2.12 – The set of combinations of atom-wise adjacency arrangements for the exemplary ME

i	AC_l							
	1	2	3	4	5	6	7	8
1	{0,0}	{0,0}	{0,1}	{0,1}	{1,0}	{1,0}	{1,1}	{1,1}
2	{0}	{1}	{0}	{1}	{0}	{1}	{0}	{1}
3	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

From these combinations of adjacency arrangements result upper triangular matrices of the adjacency schemes $\boldsymbol{AS}^U$, which are denoted in Table 2.13. They are presented together with a graphical depiction of the adjacency scheme that visualizes the connectivity between the non-hydrogen atoms.

Table 2.13 – Upper triangular matrices $\mathbf{AS}^U$ and corresponding depictions of atomic adjacency

l	$\mathbf{AS}_l^U$	Graph	l	$\mathbf{AS}_l^U$	Graph
1	$\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\end{pmatrix}$		2	$\begin{pmatrix}0&0&0\\0&0&1\\0&0&0\end{pmatrix}$	
3	$\begin{pmatrix}0&0&1\\0&0&0\\0&0&0\end{pmatrix}$		4	$\begin{pmatrix}0&0&1\\0&0&1\\0&0&0\end{pmatrix}$	
5	$\begin{pmatrix}0&1&0\\0&0&0\\0&0&0\end{pmatrix}$		6	$\begin{pmatrix}0&1&0\\0&0&1\\0&0&0\end{pmatrix}$	
7	$\begin{pmatrix}0&1&1\\0&0&0\\0&0&0\end{pmatrix}$		8	$\begin{pmatrix}0&1&1\\0&0&1\\0&0&0\end{pmatrix}$	

2.4.4 Combining valence schemes and adjacency schemes

Subsequently, valence scheme combinations and adjacency schemes are brought together to introduce bond orders into the atomic adjacency. The resulting matrices are called *molecular bodies* (**MB**). They are defined similar to **BM**-matrices, but represent only the non-hydrogen atoms $\tilde{N}_A$, such that they are of size $\tilde{N}_A \times \tilde{N}_A$.

The bonds of the valence scheme combination specified in $VSC_{k,i}$ are introduced into the non-zero elements of row i of an adjacency scheme $\mathbf{AS}_l$. The bonds of $VSC_{k,i}$ can be arranged in m different ways. $\sigma_{k,i,m}$ denotes the m-th arrangement of valence scheme i in valence scheme combination k. The arrangements of all valence schemes are presented in Table 2.14.

Table 2.14 – Arrangements σ of the bonds in the valence schemes

m	VS_1	VS_2	VS_3	VS_4	VS_5	VS_6
1	{1,1,1,1}	{1,1,2}	{1,3}	{2,2}	{1,1}	{2}
2		{1,2,1}	{3,1}			
3		{2,1,1}				

For a combination of valence schemes VSC_k, the arrangements $\sigma_{k,i,m}$ are determined.

Several arrangements per valence scheme can occur, such that all combinations of arrangements $C\sigma_k$ for VSC_k have to be generated by forming the Cartesian product

$$C\sigma_k = \sigma_{k,1,m} \times ... \times \sigma_{k,\tilde{N}_A,m}. \tag{2.25}$$

The number of combinations of arrangements in $C\sigma_k$, $N_{C\sigma,k}$, is determined by

$$N_{C\sigma,k} = \prod_{i=1}^{\tilde{N}_A} |\sigma_{k,i}|. \tag{2.26}$$

An individual combination of arrangements in $C\sigma_k$ is addressed by $C\sigma_{k,o}$, where $o \in \{1, ..., N_{C\sigma,k}\}$. A combination of arrangements $C\sigma_{k,o}$ is introduced into the upper triangular matrix of an adjacency scheme, $\mathbf{AS}^U$, by assigning the elements of $\sigma_{k,i,m,n=1,...,N_b}$ of $C\sigma_{k,o}$ to the non-zero elements of $\mathbf{AS}^U$, which gives the upper triangular matrix of a molecular body, $\mathbf{MB}^U$. The final molecular body $\mathbf{MB}$ is calculated from the addition of $\mathbf{MB}^U$ and its transposed matrix by

$$\mathbf{MB} = \mathbf{MB}^U + (\mathbf{MB}^U)^T. \tag{2.27}$$

A generated $\mathbf{MB}$ has to be checked for feasibility, which is determined considering bond orders and electrical charge. A row i of $\mathbf{MB}$ has to contain all bonds $\sigma_{k,i,m} > 1$. Since the number of bonds $N_{b,i}$ can be less than the number of bonds in $\sigma_{k,i,m}$, not every bond has to be transferred to the adjacency scheme. It has to be kept in mind that the adjacency schemes only represent the connection of the non-hydrogen atoms. Therefore, the bonds that are not assigned to an adjacency scheme are those that are established towards the hydrogen atoms. However, hydrogen atoms can only form single bonds. Hence, every bond $\sigma_{k,i,m} > 1$ has to be established between non-hydrogen atoms and thus needs to be included in $\mathbf{MB}$. The charge of row i in a generated $\mathbf{MB}$ has to be

$$q_i \geq 0 \tag{2.28}$$

to be chemically viable. Positive charges ($q_i > 0$) can be equilibrated by establishing bonds towards hydrogen atoms. If the electrical charge Q of the molecular body (determined by Equation (2.8)) is

$$Q \leq N_H, \tag{2.29}$$

sufficient hydrogen is available to equilibrate the molecular body. Only then, a generated $\mathbf{MB}$ is feasible. Example 8 presents step by step the construction of molecular bodies of

Table 2.15 – Arrangements of the valence schemes of VSC_7 (left) and combinations of arrangements $C\sigma_7$ (right)

m	$\sigma_{7,1,m}$	$\sigma_{7,2,m}$	$\sigma_{7,3,m}$
1	{1,1,2}	{1,1}	{2}
2	{1,2,1}		
3	{2,1,1}		

$\longrightarrow$

i	1	2	3
$C\sigma_{7,1}$	{1,1,2}	{1,1}	{2}
$C\sigma_{7,2}$	{1,2,1}	{1,1}	{2}
$C\sigma_{7,3}$	{2,1,1}	{1,1}	{2}

Table 2.16 – Introduction $C\sigma_7$ into $\mathbf{AS}_{U,4}$ and applying feasibility criteria on bond order and electric charge of the generated molecular bodies **MB**

Nr.	$\mathbf{AS}^U$	$C\sigma_7$	$\mathbf{MB}^U$	**MB**	Bonds feasible	q_i	$Q \leq N_H$
1	$\begin{pmatrix}0&1&1\\0&0&0\\0&0&0\end{pmatrix}$	$\{1,1,2\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix}0&1&1\\0&0&0\\0&0&0\end{pmatrix}$	$\begin{pmatrix}0&1&1\\1&0&0\\1&0&0\end{pmatrix}$	no	2 1 1	yes
2	$\begin{pmatrix}0&1&1\\0&0&0\\0&0&0\end{pmatrix}$	$\{1,2,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix}0&1&2\\0&0&0\\0&0&0\end{pmatrix}$	$\begin{pmatrix}0&1&2\\1&0&0\\2&0&0\end{pmatrix}$	no	1 1 0	yes
3	$\begin{pmatrix}0&1&1\\0&0&0\\0&0&0\end{pmatrix}$	$\{2,1,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix}0&2&1\\0&0&0\\0&0&0\end{pmatrix}$	$\begin{pmatrix}0&2&1\\2&0&0\\1&0&0\end{pmatrix}$	yes	1 0 1	yes

the exemplary ME.

Example 8. *Valence scheme combination 7 and adjacency scheme 4 are used to explain the combination of adjacency schemes and valence scheme combinations. VSC_7 contains the valence schemes VS2, VS5 and VS6. Their possible arrangements (in agreement to Table 2.14) are denoted left in Table 2.15. Thereof result 3 combinations of arrangements, denoted right in Table 2.15. The combinations of arrangements are then introduced into the adjacency scheme, presented in Table 2.16. The bonds in the molecular bodies 1 and 2 turn out to be infeasible. A bond of order 2 was not conveyed to the first row of molecular body 1. Molecular body 2 misses a double bond in row 2 and has a double bond in row 3, where only single bonds are allowed to be established. Only molecular body 3 contains all required bonds of order > 1 in the corresponding rows. The formal electric charge Q of molecular body 3 equals 2, hence it is also feasible from the point of charge equilibration in terms of Equation* (2.29)*, since $N_H = 4$.*

Introducing the $C\sigma$ of all valence scheme combinations into the generated adjacency schemes yields 10 feasible molecular bodies, which are presented in Table 2.17.

Table 2.17 – Molecular bodies of exemplary ME

Nr.	**MB**	Structure	Nr. of underlying **AS**	Q
1	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 2 & 0 \end{pmatrix}$		2	4
2	$\begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$		3	4
3	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$		4	4
4	$\begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$		5	4
5	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$		6	4
6	$\begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$		7	0
7	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$		7	4
8	$\begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$		7	2
9	$\begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$		7	2
10	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$		7	2

2.4.5 Equilibrating the formal electric charge of the molecular bodies

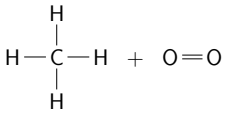
The molecular bodies can contain a formal electric charge ($Q > 0$) which is equilibrated by establishing covalent bonds between charged non-hydrogen atoms and hydrogen atoms. To this end, a **MB** is transferred to a matrix $\mathbf{E} \in \{0,1,2,3\}^{N_A \times N_A}$ (a product matrix in the sense of section 2.3), with

$$e_{i,j} = mb_{i,j}. \tag{2.30}$$

The atoms in **E** are denoted in the order of the atomic vector A(G). Bonds are formed between a charged non-hydrogen atom i ($q_i \geq 1$) and a charged hydrogen atom j ($q_j = 1$) by setting the corresponding entries $e_{i,j} = e_{j,i} = 1$. If the number of hydrogen atoms is higher than the number of free valence electrons in a molecular body ($N_H > Q$), molecular hydrogen (H_2) is formed. Since the substances in every substrate MEs invariably consist of equilibrated, non-charged molecules, excess hydrogen always is present in even amounts, if any occurs. The set of equilibrated molecular bodies forms the family of isomeric molecular ensembles. Example 9 presents the resulting FIME for the exemplary ME of formic acid and hydrogen.

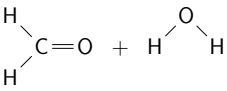
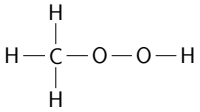
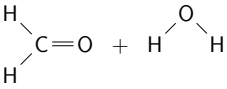
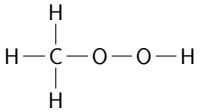
Example 9. *Table 2.18 lists the complete set of equilibrated molecular bodies, i.e. the products, which can be obtained from the substrates formic acid and hydrogen.*

Table 2.18 – The set of generated product MEs

ME	**E**	Structure
1	$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	

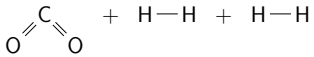
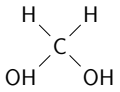
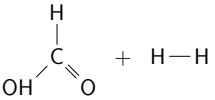
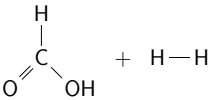
Continued on next page

Table 2.18 – continued from previous page

ME	**E**	Structure
2	$\begin{pmatrix} 0 & 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	
3	$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$	
4	$\begin{pmatrix} 0 & 2 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$	
5	$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$	

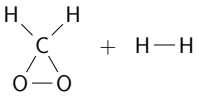
Continued on next page

Table 2.18 – continued from previous page

ME	**E**	Structure
6	$\begin{pmatrix} 0 & 2 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	
7	$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$	
8	$\begin{pmatrix} 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	
9	$\begin{pmatrix} 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	

Continued on next page

Table 2.18 – continued from previous page

ME	**E**	Structure
10	$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	

2.5 Postprocessing of the generated MEs

Visual investigation of Table 2.18 shows that some MEs are generated twice. This redundancy is not desired and only one instance of each unique ME shall be maintained. Furthermore, if the formulation of the reaction network targets at production of a pure substance, a molecule within a ME has to be chosen as the main product to serve as substrate to the subsequent reactions in the network.

2.5.1 Check for uniqueness

In order to identify redundant structures, a unique encoding for each ME is required. In Cheminformatics, Morgan's algorithm (Morgan, 1965, Figueras, 1993) is commonly employed to identify unique molecules in a set comprising redundancy.

Consider a ME composed of N_A atoms and represented by its adjacency matrix **AM**. Morgan's algorithm iteratively computes so-called connectivity values v_i^k for each atom i. A value v_i^k represents the number of unique paths of a certain length k in a molecular graph that start at molecule i. The connectivity values v_i^1 equal the number of neighbors of atom i, which is equivalent to the number of paths of length 1 (refer to Appendix A for the definition of neighborhood in a graph and the length of paths). The connectivity values for k>1 are computed from

$$v^{k+1} = \mathbf{AM} \cdot v^k. \tag{2.31}$$

This procedure is carried out repeatedly until $k = N_A$.

In this form, Morgan's algorithm is not sufficient to identify redundancy. 2 identical

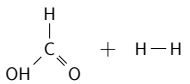
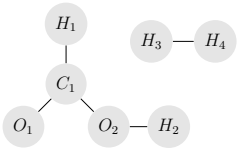
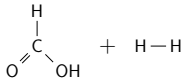
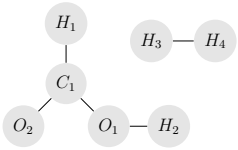
MEs can still lead to different vectors v^k since atoms can still be arranged differently. A canonization is achieved by sorting the atoms by decreasing number of valence electrons and connectivity values. This canonization leads to a unified representation of the Morgan vectors and avoids the problems arising from different computational encodings of identical MEs. In order to identify the FIME, the Morgan vectors of all MEs are computed, canonized and compared against each other.

Example 10 shows the application of Morgan's algorithm to 2 structurally identical, but differently encoded MEs.

Example 10. *Visual investigation of Table 2.18 reveals that the MEs 2 & 4, 3 & 5 and 8 & 9 are structurally equivalent. Since only one instance of each ME needs to be maintained, Morgan's algorithm is employed to detect redundancy.*

Table 2.19 presents the computed Morgan vectors for the MEs 8 & 9. Since the individual atoms are arranged differently, the comparison of the Morgan vectors implies that 2 different MEs are present.

Table 2.19 – Morgan's Algorithm applied to different representations of ME 7

ME	Structure	Adjacency	A(G)	v^1	v^2	v^3
8			$\begin{pmatrix} C_1 \\ O_1 \\ O_2 \\ H_1 \\ H_2 \\ H_3 \\ H_4 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 4 \\ 3 \\ 4 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 4 \\ 6 \\ 4 \\ 4 \\ 1 \\ 1 \end{pmatrix}$
9			$\begin{pmatrix} C_1 \\ O_1 \\ O_2 \\ H_1 \\ H_2 \\ H_3 \\ H_4 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 4 \\ 4 \\ 3 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 6 \\ 4 \\ 4 \\ 4 \\ 1 \\ 1 \end{pmatrix}$

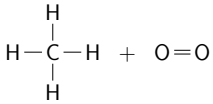
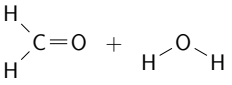
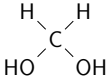
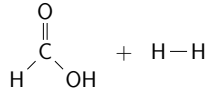
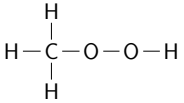
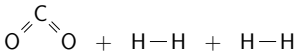
Resorting the entries v_i^k by decreasing valence electrons and connectivity values leads to the canonized representation of the Morgan vectors presented in Table 2.20.

Applying this canonized version of Morgan's algorithm to all MEs in Table 2.18 leads to the FIME, presented in Table 2.21.

Table 2.20 – Sorting the connectivity values v_i^k in decreasing order reveals that MEs 8 & 9 are structurally identical since the entries of v_i^k are identical in every case

A(G)	ME 8			ME 9		
	v^1	v^2	v^3	v^1	v^2	v^3
$\begin{pmatrix} C \\ O \\ O \\ H \\ H \\ H \\ H \end{pmatrix}$	$\begin{pmatrix} 3 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 4 \\ 4 \\ 3 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 6 \\ 4 \\ 4 \\ 4 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 4 \\ 4 \\ 3 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 6 \\ 4 \\ 4 \\ 4 \\ 1 \\ 1 \end{pmatrix}$

Table 2.21 – Graphical representation of FIME of the exemplary ME

ME	Structure	ME	Structure
1		2	
3		4	
5		6	
7			

2.5.2 Identification of the main reaction product

The generated MEs often comprise more than one substance. However, in some applications, only the main product of a generated ME shall be represented in the network and provided as substrate to the subsequent network reactions. The main product needs to be determined by a chosen, adequate criterion. To address this task, it is inevitable to identify type and quantity of the substances in a ME.

Consider the graph representation $G(V, E)$ of an arbitrary ME. The individual substances in the ME are subgraphs $G_i(V_i, E_i)$ of G. They are determined by investigating the connectivity of the vertices $V(G)$ in terms of their neighborhood. Neighboring vertices are members of the same subgraph G_i and represent a distinct substance in the ME.

An iterative routine (with iteration count k) is employed to identity the subgraphs of G. The set V^* contains the vertices of G that are so far not assigned to any set V_i. Consequently, at the beginning of the first iteration, V^* equals V. The subgraphs G_i of G are determined following the steps listed subsequently:

1. Take the element $v_1^* \in V^*$ and form the vertex set $V_{i=1}^{k=1}$.
2. Determine the neighborhood vertices $N_G(V_i^k)$.
3. Set $V_i^{k+1} = V_i^k \cup N_G(V_i^k)$.
4. If $|V_i^{k+1}| = |V_i^k|$, the connected component is complete and $V_i = V_i^k$. Else, set $k = k + 1$ and return to step 2.
5. Update V^* to $V^* = V^* \setminus V_i$ which excludes vertices that are member of V_i from V^*.
6. If $|V^*| > 0$, set $i = i + 1$ and $k = 1$ and return to step 1. Else, exit.

Example 11 presented the algorithmic procedure of identifying the constituents of ME 7 from Table 2.21.

Example 11. *Table 2.22 presents the algorithm, starting at vertex v_1. 3 subgraphs are present in ME 7 of Table 2.21, each requiring 3 iterations to be determined.*

Table 2.22 – Determining the molecular composition of ME 7

i	k	Graph	V_i^k	$\lvert V_i^k \rvert$	**BM**
		$V^* = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$			
1	1	$\mathsf{v_1}$	$\{v_1\}$	1	$\begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$
	2	$\mathsf{v_2 - v_1 - v_3}$	$\{v_1, v_2, v_3\}$	3	
	3	$\mathsf{v_2 - v_1 - v_3}$	$\{v_1, v_2, v_3\}$	3	
		$V^* = \{v_4, v_5, v_6, v_7\}$			
2	1	$\mathsf{v_4}$	$\{v_4\}$	1	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
	2	$\mathsf{v_4 - v_5}$	$\{v_4, v_5\}$	2	
	3	$\mathsf{v_4 - v_5}$	$\{v_4, v_5\}$	2	
		$V^* = \{v_6, v_7\}$			
3	1	$\mathsf{v_6}$	$\{v_6\}$	1	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
	2	$\mathsf{v_6 - v_7}$	$\{v_6, v_7\}$	2	
	3	$\mathsf{v_6 - v_7}$	$\{v_6, v_7\}$	2	
		$V^* = \emptyset$			

The main product can be determined using different criteria such as thermophysical properties like molecular weight (MW) or enthalpy of combustion (ΔH_{com}), which are estimated using QSPR models (see Appendix D). Information based on graph theory can also be used as criterion to account for the redistribution of the substrates atoms amongst the reaction products, deriving the *atomic share* χ_i of a formed molecule i. χ_i sets in relation the number of atoms that are conveyed from the substrates to a molecule i and the total number of atoms of the substrates. This is expressed as the ratio of the size of the subset $V(G_i)$ to the superset $V(G)$ as

$$\chi_i = \frac{|V(G_i)|}{|V(G)|}. \tag{2.32}$$

The applied criterion has to be chosen in correspondence to the goals of the network generation. In terms of biofuel synthesis, it is most viable to pursue those substances in the network that contain the highest enthalpy of combustion. Otherwise, if the targeted application requires a maximum in material use, criteria focussing on the amount of conserved material (highest molecular weight, atomic share) are better suited. Depending on the chosen criterion, different reaction products be will identified as main product. The generated instances of the main product equal its stoichiometric coefficient in the reaction. The impact of using different criteria is illustrated in Example 12.

Example 12. *Consider the ME of formic acid and hydrogen from previous examples as*

substrate (denoted in substrate matrix $\boldsymbol{B}$ *in Table 2.23) and ME 7 in Table 2.21 as product (denoted as product matrix* $\boldsymbol{E}$ *in Table 2.23). The individual molecules in* $\boldsymbol{B}$ *and* $\boldsymbol{E}$ *are separated by dashed lines and labeled by* $s_i^{\mathbf{B}}$ *and* $s_i^{\mathbf{E}}$*, respectively.* $\boldsymbol{B}$ *contains 2 substances* $s_i^{\mathbf{B}}$ *(i=1,2) where* $s_1^{\mathbf{E}}$ *refers to the main substrate (MS) in* $\boldsymbol{B}$ *(formic acid).* $\boldsymbol{E}$ *contains 3 substances* $s_i^{\mathbf{E}}$ *(i=1,2,3) with no main product labeled so far.*

The substances in $\boldsymbol{E}$ *are evaluated regarding aforementioned criteria.* $s_1^{\mathbf{E}}$ *contains 3 atoms, thus the atomic share of* $s_1^{\mathbf{E}}$ *is* χ_1*=3/7. The atomic shares of* $s_2^{\mathbf{E}}$ *and* $s_3^{\mathbf{E}}$ *are* 2/7 *each. In terms of atomic share,* $s_1^{\mathbf{E}}$ *is the main product. The same statement holds true when considering the molecular weight (MW) of the products. However, it changes when assessing the enthalpy of combustion. In this case, substances* $s_2^{\mathbf{E}}$ *and* $s_3^{\mathbf{E}}$ *are determined as main reaction products. This illustrates how strongly the decision criterion to identify the main reaction product influences the synthesis design task.*

Table 2.23 – Application of atomic share (χ), molecular weight (MW) and enthalpy of combustion (ΔH_{com}) to determine the main reaction product

B	$s_i^{\mathbf{B}}$	**E**	$s_i^{\mathbf{E}}$	χ_i	ΔH_{com}	MW
-	-	-	-	-	MJ/kmol	kg/kmol
$\left(\begin{array}{cccccc\|cc} 0 & 1 & 2 & 1 & 0 & & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & & 1 & 0 \end{array}\right)$	$\begin{array}{c} \\ \\ 1 \\ \\ \\ \\ 2 \end{array}$	$\left(\begin{array}{cccc\|cccc} 0 & 2 & 2 & & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & & 0 & 0 & 1 & 0 \end{array}\right)$	$\begin{array}{c} \\ 1 \\ \\ \\ 2 \\ \\ 3 \end{array}$	3/7 2/7 2/7	0 241.86 241.86	44.01 2.016 2.016

2.6 Reaction network formulation

Graph-based representation of reaction networks is an established formalism in metabolic engineering (Stephanopoulos, 1999). It is also suited for representing reaction networks in chemical synthesis. A reaction network is formulated as a graph $G(S, R)$ where the vertices in S represent substances and the edges in R refer to the reactions. Substances are addressed by s_i, where $i \in \{1, ..., N_S\}$ with N_S being the number of substances in the network. Reactions are addressed by r_j, where $j \in \{1, ..., N_R\}$ with N_R being the number of reactions in the network. A reaction r always connects 2 substances, thus it can also be denoted as $r = \{s_i, s_j\}$. The reaction heads from s_i to s_j and represents the progress of the denoted reaction. It has to be stressed that r_j is not related to the entries of the reaction matrix $\mathbf{R}$ (cf. to Section 2.3).

Accumulation of substances in the network is not allowed; source and sink reactions transport pure substrate into and pure product out of the network, respectively. Unlike

every other reaction in the network, they do not represent chemical reactions, hence they will be addressed as *pseudo-reactions*. The vertex from which the substrate is supplied is called the *source*, the vertex where the desired product is removed is called the *sink*. Every other vertex in the network is called an *intermediate*.

Reaction r_j may require reactants to be performed; therefore a supply reaction r_j^{in} is added to provide those substances. Unconverted substances are removed via reactions r_j^{out}. Example 13 depicts such a reaction network and explains the meaning of the individual reactions.

Example 13. *An exemplary network is visualized in Figure 2.1.* $r_1, ..., r_8$ *are the network reactions.* r_1 *is the source reaction of the network, while* r_8 *is the sink reaction.* s_1 *is the source and* s_7 *is the sink vertex of the network. Substances that are required to perform* r_j *are supplied via* r_j^{in}*. Unconverted substances are removed via* r_j^{out}*. Since reactions* r_1 *and* r_8 *do not represent chemical transformations, no reactants are added and they are always performed at full conversion. Hence, the values for* r_1^{in}*,* r_1^{out}*,* r_8^{in} *and* r_8^{out} *in this example are 0.*

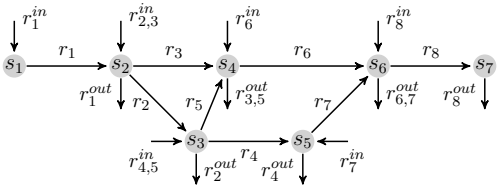


Figure 2.1 – Representation of an exemplary reaction network

A reaction network is constructed iteratively by processing a set of substrates to generate a set of products. The processing of all substrates (cf. Section 2.4) including the subsequent postprocessing of the product matrices $\mathbf{E}$ (cf. Section 2.5) constitutes a *network stage*. The first network stage contains only the main substrate s_{MS}, such that $S = \{s_{MS}\}$. The main products of one stage serve as substrates to the subsequent one.

A single substrate can lead to several products; vice versa, several substrates can lead to the same main product. This leads to a thorough interconnection of the substances in the network. However, each substance shall only be represented once in the reaction network. The set of vertices S is a steadily growing set which contains all unique network substances. It is initialized with the substrates that are provided via the source reactions. The main reaction products that are generated in each stage are added to S as long as

they are not already a member. A product s_P is added to S according to

$$S = \begin{cases} S \cup s_P & \text{if } s_P \notin S \\ S & \text{else.} \end{cases} \tag{2.33}$$

Likewise to the introduction of new substances, further reactions are also included in the network at each stage to connect substrates and their products. A new reaction is then included into the set of reactions R according to

$$R = \begin{cases} R \cup r_j & \text{if } r_j \notin R \\ R & \text{if } else. \end{cases} \tag{2.34}$$

Concerning the addition of reactions, it is not sufficient to only compare main product and reaction substrate. If this were the case, only one reaction could be established despite different reaction mechanisms emerging from the presence of different reactants. As such, a new reaction $r_j = \{s_S, s_P\}$ is added to the set of reactions R, if there is no reaction present that links the substrate ME (where s_S is main substrate) to the product s_P. Multiple reactions can thus emerge from s_S towards s_P reflecting the influence of differing reactants.

Example 14 schematically presents the construction 2 consecutive stages of a reaction network.

Example 14. *Consider the construction of a reaction network that starts at substance s_1. The set of vertices S comprises only this very substance such that $S = \{s_1\}$, while the set of reactions R consequently is empty (see Figure 2.2). It is assumed that the substances s_2,...,s_5 are the main products of s_1. It is further assumed that s_3 and s_4 are identical.*

The generated substances s_2 - s_5 are one by one screened whether they are already members of the network by employing Equation (2.33)*, adding new elements to S and R accordingly. Since the substances s_3 and s_4 are identical, s_4 is not included in the network. At network stage 2, the set of substances is $S = \{s_1, s_2, s_3, s_5\}$ and the set of reactions is $R = \{\{s_1, s_2\}, \{s_1, s_3\}, \{s_1, s_5\}\}$. The substances s_2, s_3 and s_5 serve as substrates in the subsequent network stage.*

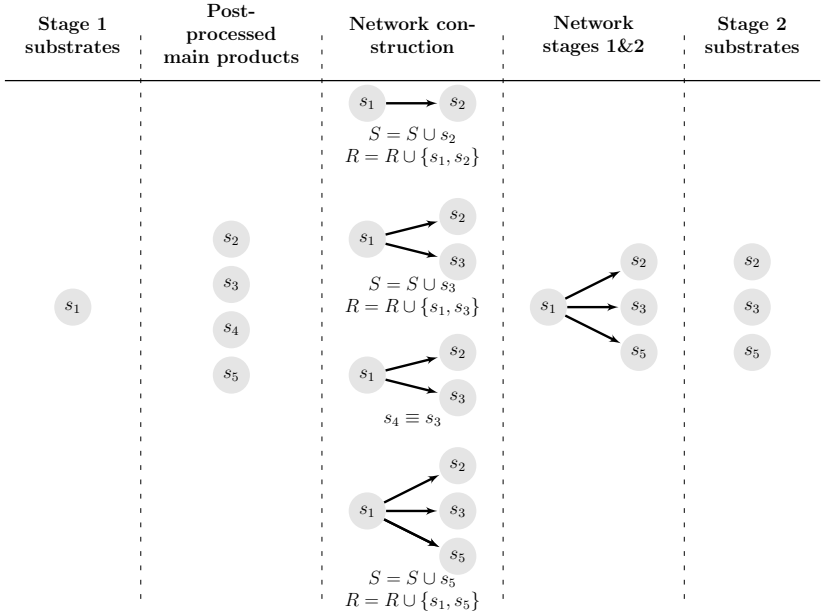


Figure 2.2 – Schematic illustration of the construction of a reaction network

2.7 Workflow of reaction network generation

ReNeGen is based on 2 routines; one generates the reactions to a given product, the other one constructs the network based on these results.

Figure 2.3 shows the programming flowchart of the reaction generator. Starting from a provided molecular ensemble (ME), the first step is to identify valence schemes (VS) and thereof the valence scheme transitions (VT), cf. Section 2.4.1. The VT sets are then used to compute the sets of valence scheme combinations (VSC) that are feasible for the considered ME, cf. Section 2.4.2. These VSC determine which covalent bonds occur simultaneously in the products. Subsequently the adjacency schemes of the non-hydrogen atoms are determined, cf. Section 2.4.3.

For each adjacency scheme **AS**, it is checked whether a VSC can be introduced in accordance to the formalisms stated in Section 2.4.4. Since a single VSC can fit into an **AS** in various arrangements, all arrangements of the bonds in VSC have to be determined (Cσ) and applied to the considered **AS**. Resulting molecular bodies may carry a formal electric charge and have to be equilibrated with hydrogen that is available in the ME, cf. Section 2.4.5.

This procedure is performed for $l = \{1, ..., N_{AS}\}$ (adjacency schemes), $k = \{1, ..., N_{VSC}\}$ (valence scheme combinations) and $o = \{1, ..., N_{C\sigma,k}\}$ (combinations of arrangements). The procedure ends when no more **AS** are available.

The network generation is summarized in the programming flowchart shown in Figure 2.4. It starts with providing the parameters

- substrate,
- reactants (substances that are added to perform certain reactions, e.g. hydrogen and acetone, chosen from an implemented list of substances),
- stages (the maximum number of network stages that are generated),
- main product identification criterion, and the
- target compound.

The substrate is externally provided and forms the root of the reaction network; every substance in the network is a derivative of the substrate. Reactants are required to perform certain reactions. For instance, hydrogenations can only take place in the presence of hydrogen. The substances are usually consumed during the process of a reaction, either by covalent bonding with the main substance or by becoming part of a reaction by-product, which is then removed. Multiple reactants can be provided to the network generation, the total number of reactants is denoted as N_{Reac}. It is also possible to investigate and perform reactions without the presence of reactants. The maximum number of network stages has to allow for sufficient rearrangements of the substrate and its derivatives to form the desired product. The maximum number of stages is not necessarily the real number of stages that are required; it is rather an upper bound that serves as stopping criterion for the network generation. The true number of stages might be less, but is always dependent on the defined generation task. The criterion for main product identification can be chosen from the criteria presented in Section 2.5.2. The main product is the substance which will be linked to the substrate by the network reactions.

Reaction network generation is an iterative process. For the computational implementation, the introduction of a list and a stack is required, cf. Figure 2.4. The list stores the substances that serve as substrates to a network stage. These substances are pushed to the stack at the beginning of a stage (push) and the list is cleared. The substances are one by one retrieved (popped) from the top of the stack to be processed in the reaction network generator. At the very beginning of the network generation, the substrate is the only substance that is written to the list of unprocessed substances.

The generation process starts by verifying that the maximum number of stages is not exceeded. If so, the substances denoted on the list of unprocessed substances are pushed

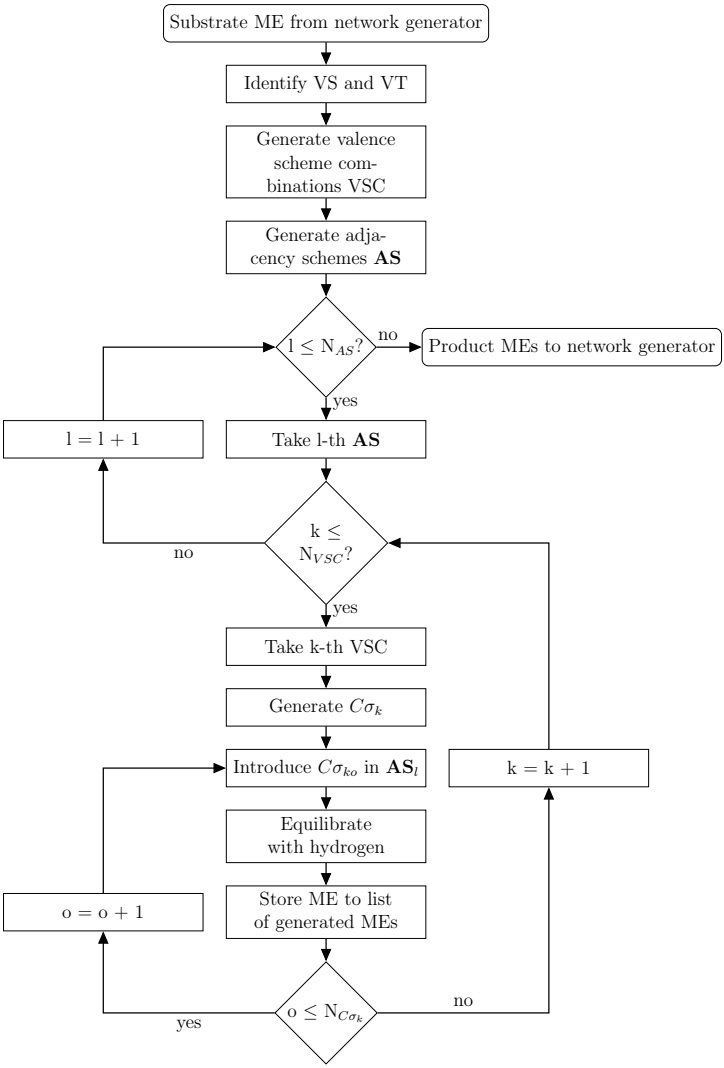


Figure 2.3 – Programming flowchart of reaction generation

to the stack. Since the stack is continuously emptied, it is checked whether unprocessed substances are still in the stack. If so, a substance is popped from the stack and combined with the i-th reactant (where $i \in \{1, ..., N_{Reac}\}$) that was specified by the user to form a molecular ensemble (ME). This ME is then passed to the reaction generator, described in Section 2.4 and depicted in Figure 2.3. The generated MEs are first screened for uniqueness (cf. Section 2.5.1) to avoid redundant results. Afterwards, the main product of each generated ME is determined (cf. Section 2.5.2). The main product of each ME is added to the list of unprocessed substances.

In a single network stage, the algorithm processes every combination of unprocessed molecules and reactants. When no more unprocessed substances are available in one stage, the algorithm proceeds to the next one, as long as the maximum number of network stages is not exceeded. Otherwise, the algorithm exits and the network generation is completed.

The runtime for the presented example of the hydrogenation of formic acid is 2.3 seconds on a PC equipped with an Intel Core i5 Quadcore CPU @ 3.2 GHz and 8 GB RAM. The routine contains approximately 1500 lines of MATLAB code.

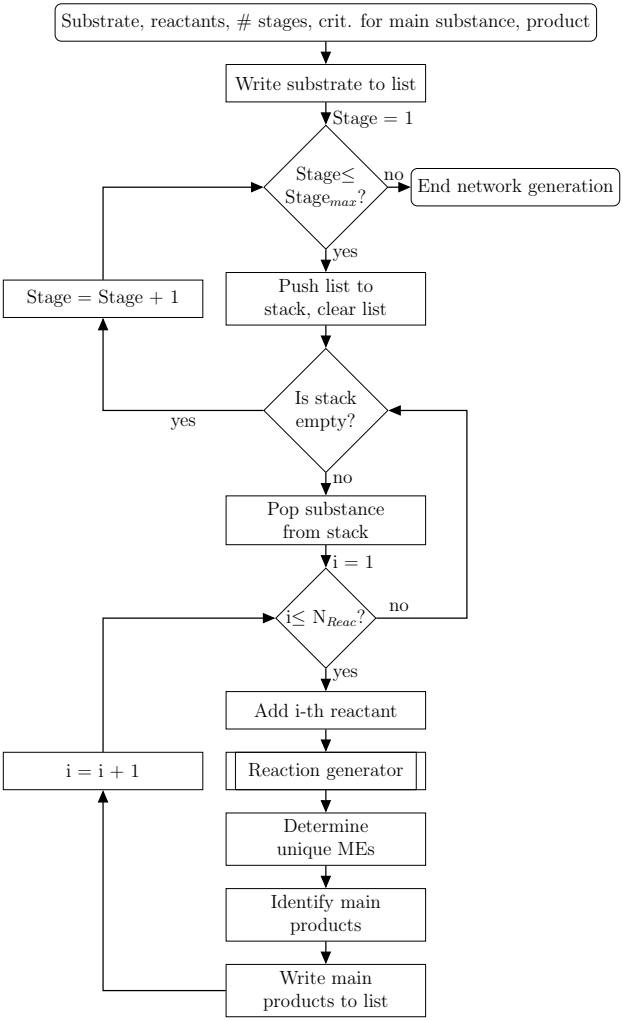


Figure 2.4 – Programming flowchart of reaction network generation

2.8 Conclusions

In this chapter, the reaction network generator ReNeGen was introduced to generate the Family of Isomeric Molecular Ensembles (FIME) of a specified Molecular Ensemble (ME). The reaction generator is based on the formalisms introduced by Fontain and Reitsam (1991) in their generator RAIN, which abstracts a ME into valence schemes and perform modifications through valence scheme transitions. The formulation of ReNeGen builds on fundamental laws of chemistry and does not require any empirical knowledge to operate. It is therefore suited for elucidating the entire scope of isomeric products to a molecular ensemble under investigation and simultaneously yields corresponding reaction mechanisms. This contrasts to empirical and semiformal implementations of reaction network generators such as RING which has most recently been reported in the context of biofuel synthesis networks or those generators which are commonly employed in biotechnology. These generators can only generate substances and reactions in a pre-defined scope based on externally provided reaction mechanisms.

In contrast to RAIN, where every atom in a molecular ensemble is considered, the presented methodology reduces the combinatorial complexity of reaction network generation. Instead of directly computing **R**-matrices from the information on valence scheme transitions and generate products thereof (as done by RAIN), the procedure is decomposed into sub-problems of smaller complexity. Valence scheme combinations are only computed for non-hydrogen atoms, which are then superposed with their atomic adjacency. Formal electric charges of these molecular bodies are equilibrated by establishing bonds towards previously excluded hydrogen atoms. A computational comparison of the algorithms in terms of runtime and maximum size of processed molecules was not possible since RAIN was not available. However, the presented approach is expected to determine the products with less computational effort, especially in terms of processing substances with a high H/C and H/O ratio, where $N_A >> \tilde{N}_A$.

3 Reaction network generation: Including empirical knowledge

The complete generation of all derivatives of a given substrate is the value proposition of formal methods for reaction network generation, especially in case there is no knowledge available on the synthesis task under investigation. However, formal network generation also have to offer the functionality to account for empirical knowledge and user-defined constraints within the generation process. Contrasting other formulations, empirical knowledge is incorporated in a top-down framework into formal reaction network generation. This means that a holistic generation task, which relies on fundamentals of chemistry, is focussed to a certain molecular space, rather than constructing the network solely based on empirical knowledge (bottom-up framework of semi-formal and empirical generators).

The following sections present means that are available in ReNeGen to focus the output by constraining either bond dissection and formation or the constitution and thermophysical properties of the substances. The principles of using restrictions on the valence scheme transitions, the molecular constitution and reaction rules was first introduced by Fontain and Reitsam (1991) in their contribution on formal reaction network generation. The QSPR models presented in Dahmen et al. (2012) are employed to calculate thermophysical properties of network substances and exclude or maintain them based on their predicted values.

3.1 Restrictions on the transition of valence schemes

The transition table is the backbone of the reaction generation, as it determines the allowed transitions of the valence schemes in the molecular ensembles (see Section 2.4.1). Due to their importance to the entire generation process, valence scheme transitions should only be disabled if they

(i) do not contribute or even hinder reaching the target of the synthesis,

(ii) are unrealistic and unlikely to be achieved (based on empirical knowledge), or

(iii) lead to the inversion of an already generated reaction.

Example 15 presents the impact of a restricted transition table on the valence scheme combinations.

Example 15. *The transition table in Table 3.1 allows only transitions where bond orders are maintained or reduced. Such a setup guides network generation from substrates with high degree of unsaturation to target compounds with lower degrees of unsaturation. As shown in the case studies in Chapter 6, such a design task is commonly encountered in the synthesis of biofuels. From this configuration results the reduced set of valence scheme combinations presented in Table 3.2. According to the constraints on valence scheme combinations stated in Section 2.4.2, only the combinations 1 and 4 are feasible.*

Table 3.1 – Transition table restricted to transitions that maintain or reduce the bond order, values for holistic generation denoted in parenthesis

	–C(–)–	=C<	–C≡	=C=	–O–	=O	–H
–C(–)–	1	0 (1)	0 (1)	0 (1)	0	0	0
=C<	1	1	0 (1)	0 (1)	0	0	0
–C≡	1	1	1	0 (1)	0	0	0
=C=	1	1	0 (1)	1	0	0	0
–O–	0	0	0	0	1	0 (1)	0
=O	0	0	0	0	1	1	0
–H	0	0	0	0	0	0	1

Table 3.2 – Set of valence scheme combinations based the restricted transition table, non-hydrogen atoms only

VSC_k	C	O	O	N_{SB}	N_{DB}	N_{TB}
1	–C(–)–	–O–	–O–	8	0	0
2	–C(–)–	=O	–O–	6	1	0
3	=C<	–O–	–O–	6	1	0
4	=C<	=O	–O–	4	2	0

3.2 Reaction rules

Reaction rules restrict the generation process of ReNeGen even tighter than modifications of the transition table. They demand to only perform alterations of those covalent bonds in a molecular ensemble that are declared modifiable by the rule specifications. Rules are iteratively applied to all (generated and provided) substances in the network. Thus it is of most importance that these rules are generally applicable to a manifold of different substances. In this contribution, reaction rules are retrieved from a chemical textbook (Furniss et al., 1989) and a literature report (Alonso et al., 2010). The collocated set includes only such reactions that are commonly applied in the processing of biorenewable feedstock.

A comprehensive set of generic reaction rules was compiled in collaboration by the author and his colleague Manuel Dahmen to serve as a generation guideline in a computational molecular structure generator (Dahmen et al., 2013). This very set of reactions rules is incorporated in the present contribution (presented in Table B.1). The collected reaction rules target at

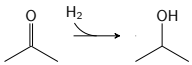
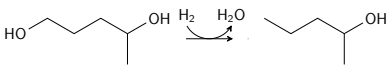
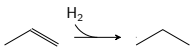
(i) altering and cleaving the (oxygenated) functionality,

(ii) decreasing the number of unsaturated bonds, and

(iii) increasing the molecular weight of the molecules.

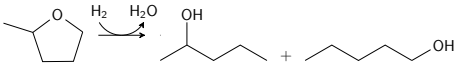
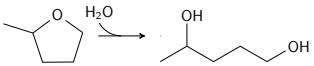
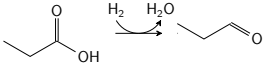
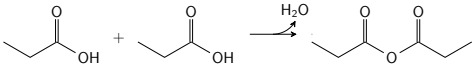
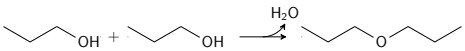
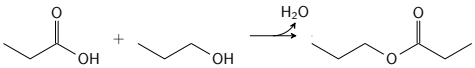
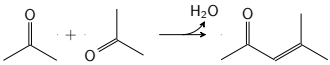
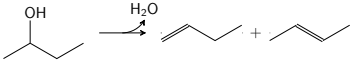
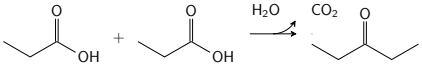
The molecular weight increase can either be achieved through carbon-carbon couplings (i.e. ketonization) or carbon-oxygen-carbon couplings (i.e. etherification). An expert check was performed by chemists within the TMFB Cluster of Excellence to ensure the applicability of this set of reactions to diverse substances.

Table 3.3 – The set of reaction rules in ReNeGen

Nr.	Name	Example reaction
1	Ketone/Aldehyde hydrogenation	
2	Alcohol hydrogenation	
3	Alkene hydrogenation	

Continued on next page

Table 3.3 – continued from previous page

Nr.	Name	Example reaction
4	Heterocycle hydrogenation	
5	Heterocycle hydrolysis	
6	Carboxylic acid hydrogenation	
7	Formation of acid anhydrides	
8	Etherification	
9	Esterification	
10	Aldol condensation	
11	Alcohol dehydration	
12	Ketonization	

The reactions allow only certain bonds to be split and maintain major parts of the covalent bonding stated in **B**. A reaction rule is only applicable to a molecular ensemble **B**, if it contains certain structural arrangements (called *patterns*) in sufficient amounts. A pattern is defined by a specific arrangement of covalent bonds, their bond order and

the type of atoms they connect. The atoms and their covalent bonding that form these patterns are the only motifs that are allowed to be modified; every other motif of the ME is maintained. Thus, the number of possible valence scheme combinations and adjacency schemes drastically decreases. In total, 10 different molecular patterns are distinguished (Table 3.4).

The patterns and their frequency of occurrence are identified using algorithms from group contribution methods (Joback and Reid, 1987, Constantinou and Gani, 1994) that investigate the type, the covalent bonding and the neighborhood of an atom to determine the pattern it is bound in. The patterns that occur in $\mathbf{B}$ are concatenated in the vector

$$\mathbf{p^B} \in \mathbb{N}^{1\times 10}, \tag{3.1}$$

where each entry $p_i^{\mathbf{B}}$ accounts for the number of occurrences of pattern i in $\mathbf{B}$.

The applicability of a reaction rule demands for the occurrence of specific patterns in certain minimal amounts. For each reaction, a vector

$$\mathbf{p^R} \in \mathbb{N}^{1\times 10}, \tag{3.2}$$

is defined that denotes the quantity of required patterns in the entries $p_i^{\mathbf{R}}$. A reaction rule is applicable to a molecular ensemble $\mathbf{B}$, if

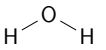
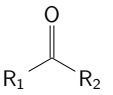
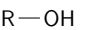
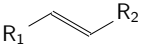
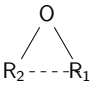
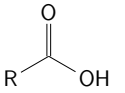
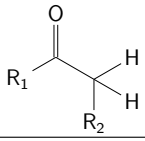
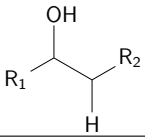
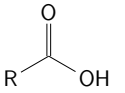
$$p_i^{\mathbf{B}} \geq p_i^{\mathbf{R}} \; \forall \; i. \tag{3.3}$$

The vectors $\mathbf{p^R}$ of all reaction rules are presented in Appendix B.

The indices of the modifiable bonds $b_{i,j}$ have to be determined and the sets of valence scheme transitions, valence scheme combinations and adjacency schemes are derived accordingly. In terms of valence scheme transitions, it has to be considered which covalent bonds at which atom are allowed to be modified. Concerning the adjacency schemes, the sets of atom-wise adjacency, A_i (see Section 2.4.3), have to account for atomic adjacency which is not allowed to be modified. The restrictions of the reaction rules render only small sets of valence scheme combinations and adjacency schemes feasible. The procedure of reaction network generation (Section 2.4) is then executed on these sets.

Example 16 presents the influence of a carboxylic acid hydrogenation reaction rule on the alterations of the ME of formic acid and hydrogen.

Table 3.4 – Distinguished patterns to determine applicability of reaction rules

p_i	Name	Structure	Annotation
1	Hydrogen pattern		
2	Water pattern		
3	Carbonyl pattern		R_1 is a carbon atom, R_2 is either a carbon atoms (forms a ketone) or a hydrogen atom (forms an aldehyde)
4	Hydroxy pattern		R is a carbon atom
5	Olefinic pattern		R_1 and R_2 can be any kind of atom
6	Heterocyclic pattern		R_1 and R_2 are members of the same carbon cycle or hetero cycle
7	Carboxylic acid pattern		R is a carbon or hydrogen atom
8	Aldol condensation pattern		R_1 and R_2 are carbon or hydrogen atoms
9	Hydroxy condensation pattern		R_1 and R_2 can be any kind of atom
10	Ketonization pattern		R is a carbon atom

Example 16. *The hydrogenation of formic acid to formaldehyde is schematically presented in Figure 3.1. Hydrogen is used as reactant to alter carboxylic group to a carbonyl group under the formation of water.*

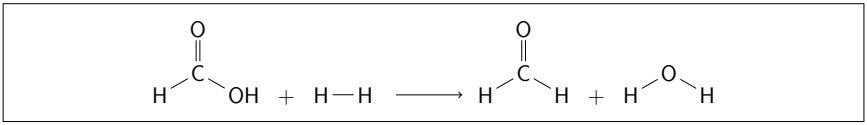


Figure 3.1 – Reaction scheme of carboxylic acid hydrogenation of formic acid

The carboxylic acid hydrogenation demands for the presence of one instance of a carboxylic acid pattern ($p_7^{\boldsymbol{R}} = 1$) and one hydrogen pattern ($p_1^{\boldsymbol{R}} = 1$) in $\boldsymbol{B}$. Thus, $\boldsymbol{p}^{\boldsymbol{R}}$ results in

$$\mathbf{p^R} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0). \tag{3.4}$$

Analysis of the patterns present in molecular ensemble $\boldsymbol{B}$ gives

$$\mathbf{p^B} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0). \tag{3.5}$$

Since $p_i^{\boldsymbol{B}} \geq p_i^{\boldsymbol{R}}$ for all patterns i, the reaction rule is applicable.

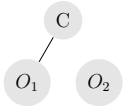
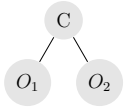
The $\boldsymbol{BM}$-notation of formic acid and hydrogen is stated in Table 2.9. The patterns are encountered at the atoms 1, 3, 6 and 7 in the bonds $b_{1,3} = b_{1,3} = 1$ (-OH group of carboxylic acid) and $b_{6,7} = b_{7,6} = 1$ (molecular hydrogen). Out of these atoms, only atoms 1 and 3 are considered in constructing valence scheme combinations and adjacency schemes since they are non-hydrogen atoms. The single bond of the carbon atom (i=1) towards the oxygen atom (i=3) is allowed to be altered, while one single and one double bond have to be maintained. Thus, the bonds of the carbon atom can only be arranged as in valence scheme 2. At the oxygen atom (i=3), one single bond is allowed to be altered while one single bond has to be maintained. Thus, the bonds of the oxygen atom can only be arranged as in valence scheme 5. The valence scheme of every other atom in the ME is maintained. This results in only one feasible valence scheme combination, which is depicted in Table 3.5.

Table 3.5 – Restricted valence scheme combinations of the non-hydrogen atoms of formic acid and hydrogen

VSC_k	C	O	O	N_{SB}	N_{DB}	N_{TB}
1	=C⟨	–O–	=O	4	2	0

The bond and hence the adjacency between atoms 1 and 2 is not allowed to be altered. The upper triangular matrices of the feasible adjacency schemes are depicted in Table 3.6.

Table 3.6 – Upper triangular matrices $\mathbf{AS}^U$ and corresponding depictions of atomic adjacency

$\mathbf{AS}_l$	$\mathbf{AS}_l^U$	Graph	$\mathbf{AS}_l$	$\mathbf{AS}_l^U$	Graph
1	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$		2	$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	

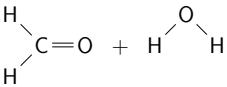
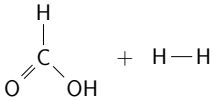
The valence scheme combination is then assigned in every permutation to the non-zero elements of the adjacency schemes, as shown in Table 3.7. Only the molecular bodies 3 and 6 fulfill the requirements on bond orders and formal electric charge.

Table 3.7 – Introduction the valence scheme combination into the available adjacency schemes and applying feasibility criteria on bond order and electric charge on the generated molecular bodies

Nr.	$\mathbf{AS}^U$	$C\sigma_1$	$\mathbf{MB}^U$	$\mathbf{MB}$	Bonds feasible	q_i	$Q \leq N_H$
1	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{1,1,2\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	no	3 1 2	no
2	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{1,2,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	no	3 1 2	no
3	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{2,1,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	yes	2 0 2	yes
4	$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{1,1,2\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	no	2 1 1	yes
5	$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{1,2,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$	no	1 1 0	yes
6	$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\{2,1,1\}$ $\{2\}$ $\{1,1\}$	$\begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	yes	1 0 1	yes

After equilibrating the electric charge of the molecular bodies 3 and 6, the molecular ensembles presented in Table 3.8 result.

Table 3.8 – Output of reaction network generation

ME	**E**	Structure
1	$\begin{pmatrix} 0 & 2 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$	
2	$\begin{pmatrix} 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$	

ME 1 in Table 3.8 is the provided substrate ME and will be discarded by the subsequent network generation step. ME 2 is the desired product ME.

3.3 Constraints on the molecular constitution and thermophysical properties

The previous sections presented ways to restrict the product scope during the generation process. Additional means can be taken at the end of generating a network stage to further target the network. This includes constraining the molecular constitution and the thermophysical properties of all identified main substrates in the network. These constraints are used to avoid that substances with undesirable motifs or properties serve as substrates to the subsequent network stage. Computed substances are screened for their molecular motifs and thermophysical properties directly after their identification as main reaction product to minimize the computational effort spent on their processing in case they will be discarded from further consideration.

3.3.1 Constraining the molecular constitution

Molecular motifs are identified again using algorithms from group contribution methods (Joback and Reid, 1987, Constantinou and Gani, 1994). Constitutional features that are constrainable in ReNeGen with a lower and upper bound are

(i) atoms per type,

(ii) ring size,

(iii) number of rings, and

(iv) number of certain functional groups.

Minimum and maximum number of each atom type are the most important constitutional aspects of the substances in **E**. Setting a reasonable upper bound for the number of carbon, oxygen and hydrogen atoms avoids polymerization of the substrates, while setting reasonable lower bounds avoids their decomposition to small size molecules. These bounds always have to be defined in accordance to the considered case. The ring size plays an important role in terms of stability of a cyclic compound, as rings with less than 4 or more than 7 atoms are usually expected to be unstable (Furniss et al., 1989). Also the occurrence of multiple rings within one molecule can be undesirable. Certain functional groups (e.g. peroxides) are also undesired in many synthesis applications as they pose major concerns to safety issues.

3.3.2 Restrictions on thermophysical properties

In the same manner, predicted thermophysical properties of generated substances serve as disqualifying criterion for further consideration. Following Hechinger et al. (2012) and Dahmen et al. (2013), predictive property models (see Appendix D) are employed to estimate a set of thermophysical property data for each substance. The computed values are compared against user-defined upper and lower limits on the specific property. Substances that exceed the specified limits are not considered as substrates to the subsequent network stage. It needs to be stressed that the employed property models incorporate a prediction error, which are reported in Appendix D. The applied property constraints should be set such that the inaccuracies in model prediction are accounted for.

3.4 Conclusions

ReNeGen provides the opportunity to include empirical knowledge to target the formal generation process into a specific direction. Empirical knowledge is included in a top-down approach, focussing the formal generation task into a desired chemical space. The

restrictions only reduce the number of valence scheme combinations, adjacency schemes and substrates per network stage without affecting the formal generation routine or ReNeGen. Hence, intermediates and products are still generated based on formalized understandings of chemistry on atomic adjacency and covalent bonding. This contrasts to bottom-up frameworks of semiformal and empirical reaction networks generators, where the provided empirical knowledge is the single basis to network generation. Thus, the effect of empirical knowledge on formal and empirical network generation are opposed: In the case of formal generation, less empirical knowledge will result in more comprehensive networks, while the size of networks derived by empirical generators will shrink. Especially in design tasks, where only little is known about the reaction mechanisms and the generated intermediates, semiformal network generators have to rely on additional assumptions to have a sufficiently large set of information to generate a network of meaningful size. Formal reaction network generators are not relying on such assumptions; they generate reaction networks based on fundamentals of chemistry and allow for including only that amount of empirical knowledge that is known about the design task. Furthermore, the effect of introducing an empirical assumption on the resulting network can be assessed one by one for plausibility.

4 Reaction network generation: Network manipulation

Up to this point, there are several aspects of the generated reactions and networks that narrow the scope of the assessment, contradict behavior observed in practice or lead to substances that do not have any connection to the target compounds.

The assessment scope is narrowed down since a generated reaction network only links one substrate and one target. However, it is sometimes favorable to combine several distinct networks into a larger one that comprises multiple substrates and targets, allowing for comparing synthesis pathways of multiple substrate-target combinations.

The reactions in the networks are, so far, assumed to exhibit full selectivity towards their products, which is rarely achieved in practice. Molecular functionality in symmetric arrangements avoids selective modifications of the molecules, thus leading to significant yield reductions. This aspect has to be considered to allow for more realistic statements on the performance of investigated pathways.

Since the resulting networks can be quite complex in terms of the number of substances and reactions, the networks should be reduced to only those substances that participate in the formation of a target compound. Every other substance is not of interest and should hence be deleted from the network to reduce computational complexity.

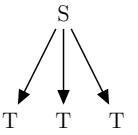
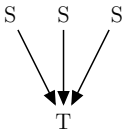
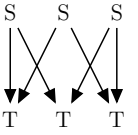
In detail, this chapter focusses on how to

1. merge several distinct networks into a single one,
2. estimate the selectivity of reactions in a novel approach that is based on molecular motifs in the reacting substances, and
3. perform a reduction of the reaction network to only those substances and reactions that participate in the formation of a target compound.

4.1 Merging multiple networks

A reaction network generated by ReNeGen represents the synthesis alternatives of a single target compound from a single substrate, where substrate and target are linked through

Table 4.1 – Four types of substrate-target configurations and their implications to the synthesis design. The term "best" in the column "Implications" refers to the calculated minimum/maximum value of a not further specified objective function.

Type	Configuration	Implications
Single substrate, single target		Gives the best pathway of the specified substrate-target combination
Single substrate, multiple targets		Allows for comparing the performance towards different targets that can be derived from the provided substrate
Multiple substrates, single target		Allows for identifying the best substrate for a certain target
Multiple substrates, multiple targets		Identifies best substrate-target combination

one or multiple pathways, representing the actual sequence of reactions for target synthesis. If several networks of distinct substrate-target combinations share common intermediates, they can be merged into a single one. This way, a more profound investigation of the design task is possible by analyzing and comparing different substrate-target combinations. Four types of configurations of substrate-target combinations can be encountered (see Table 4.1).

Networks are merged by forming the union of their graph representations. Consider an arbitrary, but finite number of n graphs G_1,...,G_n, each containing a vertex sets V_1,...,V_n and an edge set E_1,...,E_n. The union of these graphs, $G(V, E)$, is represented as

$$G = \bigcup_{i=1}^{n} G_i = G_1 \cup ... \cup G_n = (V_1 \cup ... \cup V_n, E_1 \cup ... \cup E_n). \tag{4.1}$$

Example 17 schematically presents the merger of two networks.

Example 17. *In the left part of Figure 4.1, two networks G_1 and G_2 are presented. Both*

networks start from the provided substrates S_1 and S_2, respectively, and are directed towards the targets T_1 and T_2, respectively. The grey box highlights the reactions and substances that are common to G_1 and G_2. The union of both networks, $G = G_1 \cup G_2$, is presented in the right part of Figure 4.1. The sets of vertices and edges of the merged network G(V,E)

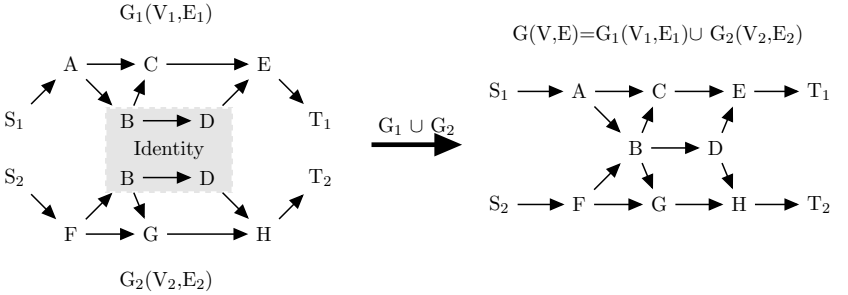


Figure 4.1 – Example of merging two networks

are

$$V = \{S_1, S_2, A, B, C, D, E, F, G, H, T_1, T_2\}$$
$$E = \{\{S_1, A\}, \{S_2, F\}, \{A, B\}, \{A, C\}, \{B, C\}, \{B, D\}, \{B, G\},$$
$$\{C, E\}, \{D, E\}, \{D, H\}, \{F, B\}, \{F, G\}, \{G, H\}, \{E, T_1\}, \{H, T_2\}\}.$$

The resulting network contains two substrates and two targets. The initial formulation comprised two separate combinations of substrates and targets, namely S_1&T_1 and S_2&T_2. In the merged network, novel combinations of substrates and targets occur (namely S_1&T_2 and S_2&T_1), that were not represented in the separate networks.

4.2 Estimation of the selectivity of reactions

Multi-step synthesis pathways should include reactions that are highly selective towards the desired reaction products to minimize the losses at each stage. An exact determination of the selectivity requires, besides the knowledge of the reacting substances, at least information on reaction conditions as well as solvents and molecular configuration of the catalyst. So far, the selectivity of only very specific reactions can be estimated (e.g. the hydrodesulfurization of dibenzothiophene and 4,6-dimethyldibenzothiophene (Farag, 2010)). These estimations require detailed kinetic models of the underlying reaction mechanism. However, it has to be assumed that most of the generated reactions may not have been

investigated experimentally or analyzed theoretically.

The selectivity of reactions is, amongst other causes, highly dependant on the molecular structure of the involved molecules, especially considering the functional groups contained and their arrangement. In a facilitating approach, symmetric arrangement of molecular functionality, i.e. the occurrence of two or more instances of the same functionality at different positions in a molecule, is considered here as single cause for non-selectiveness. Non-selective reaction will yield multiple substances in equal amounts as one or multiple instances of the symmetric functionality are cleaved simultaneously. This assumption coincides with experimental experience, for instance the selective hydrogenation of Glycerol (Oh et al., 2011). Highly selective cleavages of the secondary hydroxy group, which occur only once in the molecule, were achieved. However, selective hydrogenation of one of the two primary hydroxy groups could not be performed and always led to a mixture of 1,2-propanediol and 2-propanol.

A novel computational approach for rapidly estimating the selectivity of reactions is presented in this contribution. It aims at identifying reactions where symmetric arrangements affect the selectivity and assigning an approximated selectivity value based on molecular functionality. It has to be stressed that this approach does not intend to provide an accurate calculation of the actual selective behavior of a reaction; the intention is to have a means at hand that supports the identification of synthesis bottlenecks solely based on the molecular structure of the intermediates. At the same time, a reasonably estimated selectivity value shall be incorporated into the analysis that is (i) on the same level of accuracy and (ii) on a comparable time horizon as the overall assessment presented in this thesis.

The selectivity of reactions is triggered by a set of 150 manually identified arrangements (called *triggering patterns*), distinguishing between different types of molecular functionality and their first degree neighborhood (for the definition of neighborhood, cf. Appendix A). A triggering pattern thus is not only classified by its incorporated functional group type, but also whether it is bound to a primary, secondary or tertiary carbon atom, a chain, a ring, an aromatic or a heterocyclic arrangement. If such triggering patterns occur in symmetric arrangements, they induce non-selective reactions. To assess the selectivity of reactions, a substrate i is represented by a vector

$$\mathbf{tp}_i \in \mathbb{N}^{150\times 1} \tag{4.2}$$

where each $tp_{y,i}$ accounts for the quantitative occurrence of a certain triggering pattern y.

Besides the substance that carries symmetric arrangements of triggering patterns, the corresponding products of the non-selective reactions have to be identified. Hence, *resulting patterns* are introduced, each representing a triggering pattern after refunctionalization.

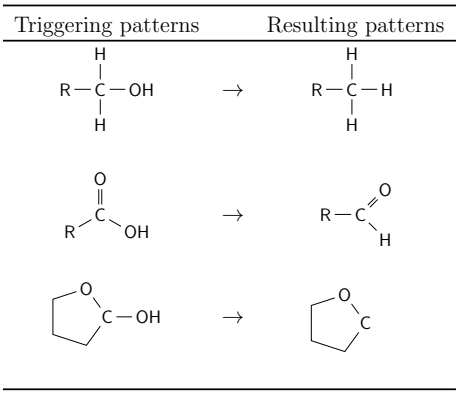


Figure 4.2 – A representative set of triggering patterns and their corresponding resulting patterns

Thus, each substance i is additionally represented by a vector

$$\mathbf{rp}_i \in \mathbb{N}^{150\times 1} \tag{4.3}$$

where each $rp_{y,i}$ accounts for the quantitative occurrence of a certain resulting pattern y in substance i. The patterns and their frequency of occurrence are identified using algorithms from group contribution methods (Joback and Reid, 1987, Constantinou and Gani, 1994). Figure 4.2 presents several combinations of triggering and resulting patterns as an illustrating example. The complete set of triggering and resulting patterns is provided in Appendix C.

In case,

$$tp_{y,i} \geq 2, \tag{4.4}$$

a symmetric arrangement of the triggering pattern tp_y is present in substance i and renders a selective refunctionalization of one instance of tp_y impossible. Such a triggering pattern will be referred to by using the index y^*.

Substances j that are produced by refunctionalizing $tp_{y^*,i}$ are identified by the following criteria:

1. It has to be ensured that only $tp_{y^*,i}$ is refunctionalized. Every other triggering pattern has to be unprocessed, such that

$$tp_{y,j} = tp_{y,i} \quad \forall \quad y \setminus y^*. \tag{4.5}$$

2. A refunctionalization of one instance of $tp_{y^*,i}$ always gives one instance of $rp_{y^*,i}$. Therefore, $rp_{y^*,j}$ has to increase proportionally to the number of refunctionalizations of $tp_{y^*,i}$. Consequently, the sum of triggering and resulting patterns in j has to equal the sum of triggering and resulting patterns in i, stated as

$$rp_{y^*,j} + tp_{y^*,j} = rp_{y^*,i} + tp_{y^*,i}. \tag{4.6}$$

3. The number of refunctionalizations, N_{Rf}, is calculated from

$$N_{Rf} = tp_{y^*,i} - tp_{y^*,j}. \tag{4.7}$$

Since every refunctionalization requires one reaction, the network distance $d(i,j)$ (see Appendix A for its definition) between substances i and j has to be

$$d(i,j) = N_{Rf}. \tag{4.8}$$

Those substances j that satisfy these criteria are concatenated in a subset $V^* \in V$. If so far there is no reaction between substance i and any substance in V^*, novel reactions are added to the set of network reactions.

It is assumed that each of the substances in V^* is produced in equal amounts. This is a pragmatic assumption aiming to have a means at hand to rapidly assign a value to the selectiveness of reactions and deepen the understanding of the design task under investigation. However, this approach is not deterministic since it does not consider all aspects that determine selectivity. The calculated selectivity values need to be verified by experimental investigations. The selectivity S_j of reaction r_j results in this approach from

$$S_j = \frac{1}{|V^*|}. \tag{4.9}$$

This implies that the higher the number of $tp_{y,i}$, the lower is the selectivity towards a substance $j \in V^*$. Example 18 presents the proposed selectivity estimation for the hydrogenation of 3-(hydroxymethyl)-1,6-hexanediol. It illustrates how the substances in the network are assessed and the network is modified to account for symmetric functionality and its impact on the selectivity of the network reactions.

Example 18. *Figure 4.3 presents the network for 3-(hydroxymethyl)-1,6-hexanediol hydrogenation as it is generated by ReNeGen. 3-(hydroxymethyl)-1,6-hexanediol, labeled as substance 1, contains three instances of a primary hydroxy group bound to a carbon chain*

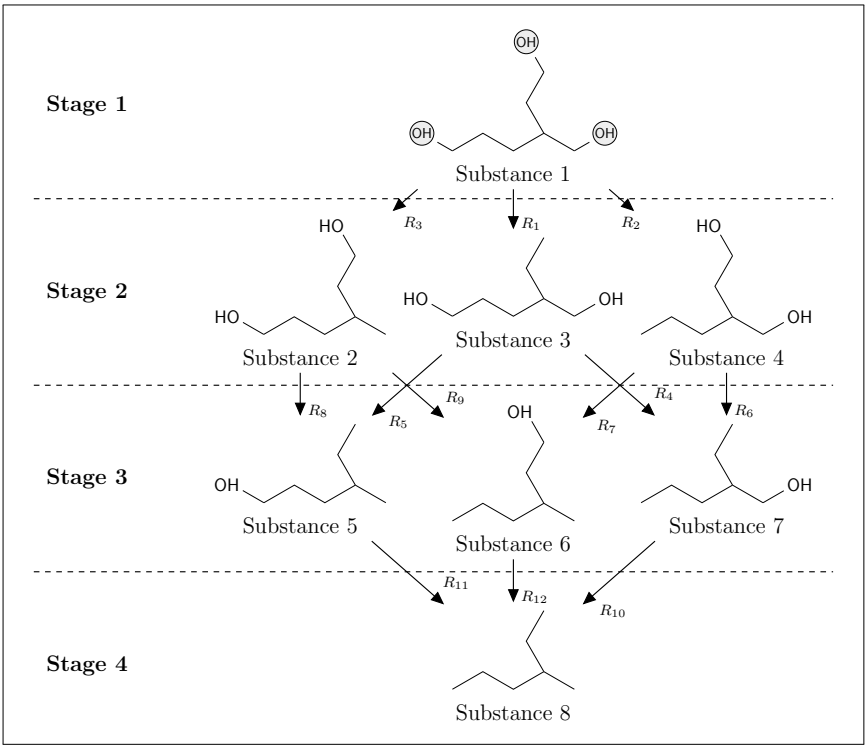


Figure 4.3 – Reaction network for 3-(Hydroxymethyl)-1,6-hexanediol deoxygenation under the presence of hydrogen and performing only hydroxy condensation reactions. The symmetric arrangement of hydroxy groups in substance 1 are highlighted in the gray circles

(tp_3), highlighted by the gray circles. The corresponding entry in vector $\boldsymbol{tp}$ *is*

$$tp_{3,1} = 3. \tag{4.10}$$

Table 4.2 denotes for each substance i the number of occurrences of triggering pattern tp_3 and the number of resulting patterns rp_3. Since $tp_{3,1} \geq 2$, the condition stated in Equation (4.4) is fulfilled and a selective refunctionalization of $tp_{3,1}$ cannot be performed. $\boldsymbol{tp}_i$ *and* $\boldsymbol{rp}_i$ *cannot be presented here due to their size. However, $tp_{3,1}$ is the only triggering pattern present in substance 1. Therefore, the condition stated in Equation (4.5) is fulfilled for every network substance. The sum $tp_{3,1} + rp_{3,1}$ results to 3 for each network substance, which states that also the condition in (Equation (4.6)) is fulfilled. Furthermore,*

Table 4.2 – Occurrence of triggering and resulting pattern 3 in the example

s_i	$tp_{3,i}$	$rp_{3,i}$	$tp_{3,i} + rp_{3,i}$	N_{Rf}	$d_{1,j}$
2	2	1	3	1	1
3	2	1	3	1	1
4	2	1	3	1	1
5	1	2	3	2	2
6	1	2	3	2	2
7	1	2	3	2	2
8	0	3	3	3	3

the distance $d(1, j)$ of every substance j equals the number of refunctionalization N_{Rf}, fulfilling the third condition (Equation (4.8)). Hence, substances 2-8 result from non-selective refunctionalization of $tp_{3,1}$. Additional reactions are introduced between substance 1 and substances 5,6,7 and 8. Dashed arrows show these reactions in Figure 4.4.

According to Equation (4.9), the selectivity of reactions r_1-r_3, r_{13}-r_{15} and r_{19} is 1/7.

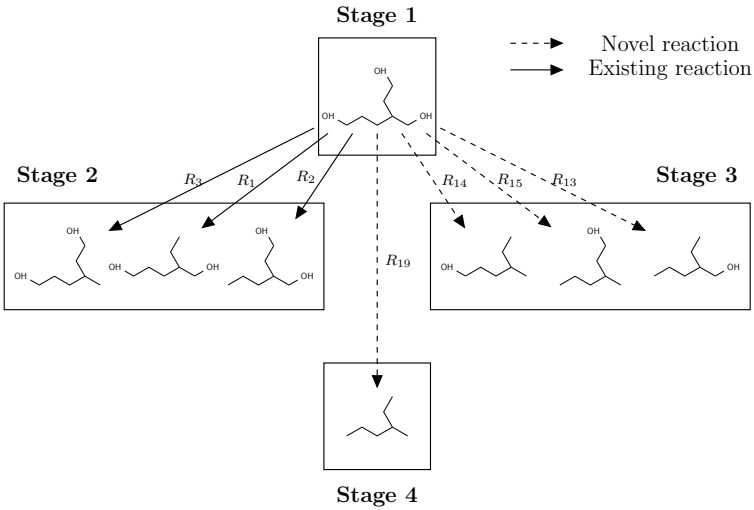


Figure 4.4 – Introduction of novel reaction into the network to account for cleaving multiple instances of a functionality in symmetric arrangements

4.3 Network reduction

The generated networks often contain reactions and substances that are not involved in the production of the desired target (called *dead ends*). They negatively influence the computational efficiency of subsequent evaluation algorithms. If only those substances and reactions that contribute to the synthesis of a product are maintained in the network, information is provided more obvious and processing of the network requires less computational effort.

This contribution is first to propose an approach for the reduction of synthesis reaction networks that is based on graph theory and the analysis of the in- and outdegrees of the network vertices (see Appendix A for the definition of in- and outdegree).

Only vertices that are head vertices of sink reactions (pseudo-reactions where product is removed from the network, see Section 2.6) are allowed to have an outdegree $d^-(s_i) = 0$. Every other vertex in the network with an outdegree $d^-(s_i) = 0$ is an illegitimate sink of the network. These vertices and their adjacent edges can be removed without impairing the synthesis of the target compounds. Since preceding vertices may also have no connection to the target compounds, the routine is applied iteratively until all illegitimate sinks are removed.

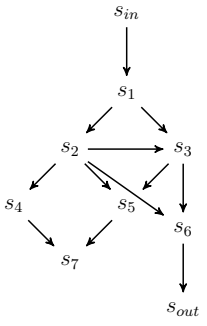
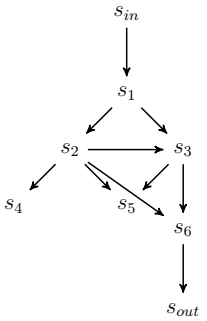
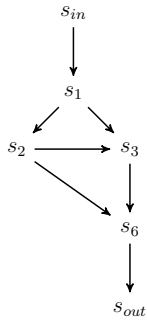
The algorithm consists of four distinct steps:

1. Determine the outdegree $d^-(s_i)$ of each substance in the network.
2. Identify illegitimate sinks of the network with $d^-(s_i) = 0 \; \forall \; s_i \notin$ target compounds. If there are none, exit. Otherwise, continue with step 3.
3. Remove dead ends from V.
4. Remove those edges from E that led to the dead ends and return to step 1.

Example 19 presents the application of this algorithm to an exemplary network.

Example 19. *The substance represented by vertex s_{out} is the legitimate sink of the network that is presented in Figure 4.3. It is the only vertex allowed to have $d^-(s_i) = 0$. Screening the vertex degrees (step 1) reveals that $d^-(s_7) = 0$ although s_7 is not declared to be a sink of the network (step 2). Thus s_7 is a dead end and is removed (step 3), as well as the reactions $\{s_4, s_7\}$ and $\{s_5, s_7\}$ (step 4). The second iteration reveals that vertices s_4 and s_5 are dead ends. Consequently, they are removed, along with the reactions $\{s_2, s_4\}$, $\{s_3, s_4\}$ and $\{s_3, s_5\}$. In the next iteration step, no more dead ends are detected, such that the algorithm exits.*

Table 4.3 – Example of the network reduction algorithm

Iteration	Network graph	$E(G)$	$V(G)$	$d^-(s_i)$
0		$\begin{Bmatrix} \{s_{in}, s_1\} \\ \{s_1, s_2\} \\ \{s_1, s_3\} \\ \{s_2, s_3\} \\ \{s_2, s_4\} \\ \{s_2, s_5\} \\ \{s_2, s_6\} \\ \{s_3, s_5\} \\ \{s_3, s_6\} \\ \{s_4, s_7\} \\ \{s_5, s_7\} \\ \{s_6, s_{out}\} \end{Bmatrix}$	$\begin{Bmatrix} s_{in} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_{out} \end{Bmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 4 \\ 2 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$
1		$\begin{Bmatrix} \{s_{in}, s_1\} \\ \{s_1, s_2\} \\ \{s_1, s_3\} \\ \{s_2, s_3\} \\ \{s_2, s_4\} \\ \{s_2, s_5\} \\ \{s_2, s_6\} \\ \{s_3, s_5\} \\ \{s_3, s_6\} \\ \{s_6, s_{out}\} \end{Bmatrix}$	$\begin{Bmatrix} s_{in} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_{out} \end{Bmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 4 \\ 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$
2		$\begin{Bmatrix} \{s_{in}, s_1\} \\ \{s_1, s_2\} \\ \{s_1, s_3\} \\ \{s_2, s_3\} \\ \{s_2, s_6\} \\ \{s_3, s_6\} \\ \{s_6, s_{out}\} \end{Bmatrix}$	$\begin{Bmatrix} s_{in} \\ s_1 \\ s_2 \\ s_3 \\ s_6 \\ s_{out} \end{Bmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 0 \end{pmatrix}$

4.4 Conclusions

Combining individual reaction networks allows for a comparison of multiple substrate-target combinations. The individual pathways can be compared such that not only superior pathways, but also superior substrate-target combinations can be detected.

The newly introduced estimation of selectivity allows for a more realistic representation of the computed reactions. The selectiveness of a reaction strongly influences the production performance of a desired product. Only by assessing the selectivity, the true potential of target substances can be detected and misleading statements can be avoided. This contribution is first to propose such an approach for a rapid estimation of reaction selectivity. The quality of the network representation benefits from selectivity estimation in two ways: (i) it unravels reaction sequences that incorporate the same refunctionalization and could possibly be performed in one reaction step and (ii) it shows which substances in the network can only be produced at low selectivity. The presented methodology enables insights into the influence of the substrates's molecular structure on the selectivity of a reaction.

A new approach to the reduction of chemical synthesis networks based on the analysis of the network topology was presented. It reduces the network to only those substances and reactions that participate in the synthesis of a desired product. The network reduction is required to improve the performance of subsequent evaluation steps by reducing its size.

5 Network evaluation strategy

It was mentioned in Section 1.2 that an optimization-based evaluation strategy for biofuel synthesis, named Reaction Network Flux Analysis (RNFA) (Voll and Marquardt, 2012b), is currently employed in TMFB to identify attractive pathways towards biofuels candidates with respect to a provided objective function. However, this technique does not take the structure of the network into account, although the assessment of its topology allows for statements about the so-called *robustness* of a synthesis network in terms of number of available pathways and the importance of certain reactions. A robust network provides multiple alternative pathways to produce the target compound, which is desirable, since the practical feasibility of computationally generated reactions and hence of certain pathways is not guaranteed. Information on the robustness of a reaction network can be retrieved from its decomposition into unique reaction sequences, called *elementary modes* (Papin et al., 2004). The final flux distribution in the network is either a single elementary mode or a linear combination of several.

This chapter presents a multi-stage evaluation scheme that not only aims at determining an optimal flux distribution, but also identifies important reactions in the network and allows for statements about the robustness of the design task. Besides evaluating the design task of single-component biofuels, an additional evaluation step is proposed that incorporates streams of unconverted intermediates into the final product. In previous evaluation strategies (Voll and Marquardt, 2012a, Marvin et al., 2013), the potential of these streams was not further elucidated, only their loss was accounted for. However, when combined with the final product, these streams inherently bear the potential to increase the feedstock utilization when blended with the target compound, of course under consideration of the initially imposed property constraints, which are not allowed to be violated. As such, this approach constitutes a more integrated use of the processed biomass feedstock.

The following sections will first describe the previously employed optimization-based evaluation technique and subsequently the novel approach that also assesses the network topology and the integration of waste streams.

5.1 Optimization-based network evaluation

The mathematical representation of the steady-state flux distribution in reaction networks is

$$\mathbf{A} \cdot \mathbf{f} = 0. \tag{5.1}$$

$\mathbf{A}$ represents the stoichiometric matrix, defined as $\mathbf{A} \in \mathbb{Q}^{s \times r}$. The elements $a_{i,j}$ are the stoichiometric coefficients of the network reactions, which are negative if a substance i is consumed in reaction j, and positive if a substance is formed. The rows of $\mathbf{A}$ refer to the substances in the reaction network $i \in \{1, ..., s\}$ and the columns refer to the reactions $j \in \{1, ..., r\}$. $\mathbf{A}$ also includes the substrate supply to the network and the removal of the final product, in order to close the overall material balance. Vector $\mathbf{f}$ in Equation (5.1) is of size $r \times 1$ and summarizes the fluxes of the reactions (Varma and Palsson, 1994). The molar flux of a substance i in reaction j in the network can thus be retrieved by multiplying f_j with the corresponding stoichiometric coefficient $a_{i,j}$. In systems biology, these models are used to describe the metabolism of an organism. They are employed to support the design of experiments (Wittmann and Heinzle, 2001) and to elucidate the potential improvements by metabolic re-engineering of an organism of interest (Stephanopoulos, 1999).

RNFA was developed on the basis of related concepts that were originally introduced in the metabolic engineering community, to evaluate reaction networks for biofuel synthesis applications (Voll and Marquardt, 2012b,a). The steady-state flux model presented in Equation (5.1) was modified by introducing a vector $\mathbf{b}$ to account for non-ideal network behavior due to incomplete conversion and non-selective reactions. The novel formulation is as follows:

$$\mathbf{A} \cdot \mathbf{f} = \mathbf{b} \tag{5.2}$$

$\mathbf{b} \in \mathbb{R}^{s \times 1}$ in Equation (5.2) summarizes all fluxes that leave the network.

The system of linear equations stated in Equations (5.1) and (5.2) is in most cases under-determined. It contains more reactions than substances and one substance can be derived by multiple sequences of reactions. RNFA provides an optimal solution by solving an inverse problem which incorporates an objective function motivated by sustainability arguments. Criteria that can serve as objective functions are summarized in Appendix E.

The formulation of the optimization problem is

$$\begin{aligned} \min_{f,b} & \left\{ \begin{matrix} \phi_1 \\ \vdots \\ \phi_n \end{matrix} \right\} \\ s.t. \quad & \mathbf{A} \cdot \mathbf{f} = \mathbf{b} \\ & f, b \geq 0 \\ & \mathbf{f} \in \mathbb{R}^r, \mathbf{b} \in \mathbb{R}^s. \end{aligned} \tag{5.3}$$

ϕ_n in Equation (5.3) represents different objective functions. The formulation allows for multi-objective evaluation of the optimization problem. RNFA is implemented in the General Algebraic Modeling System GAMS (Brooke et al., 1998).

5.2 Multi-stage network evaluation

Optimization-based evaluation of reaction networks, as presented by Voll and Marquardt (2012b), Yim et al. (2011) and Marvin et al. (2013), gives the optimal solution to a certain objective function. Such techniques usually do not assess the structural properties of the network (topology). The topology contains information on the number of alternative pathways available to derive the target compound. Substrate-target combinations with a high number of synthesis pathways are favorable since multiple alternatives are available to achieve the desired target. Furthermore, crucial reactions to a chosen synthesis can be identified by accounting for their number of occurrences in the pathways.

Combinatorial approaches, which are commonly referred to as *elementary mode analyses*, are employed in metabolic engineering to determine all synthesis pathways in a network (Stelling et al., 2002). The results from elementary mode analysis not only give the available pathways but are also useful to determine key aspects of the network topology such as the network structure (pathway length, reaction participation) (Papin et al., 2004), network robustness (functionality of pathways under perturbation) (Stelling et al., 2002) and network fragility (Gagneur and Klamt, 2004). Evaluation criteria as a function of the network fluxes are employed to assess key performance indicators of each pathway, with the final flux distribution towards a desired target compound consisting of only one or of linear combinations of these pathways. To this end, an LP-problem is formulated to determine the contribution of each synthesis pathway to the final flux distribution to optimize an imposed objective function. In this thesis, elementary mode analysis is adapted for the evaluation of biofuel synthesis networks. The performance criteria collocated by Hechinger et al. (2010) and Voll and Marquardt (2012a,b) are employed to serve as objective functions

to set up the LP problem. They rely on molar- or mass-related thermophysical properties of the network fluxes and compounds and allow for assessing also more elaborated criteria such as investment cost and environmental impact. The collocated evaluation criteria are presented in Appendix E.

In previous publications on biofuel synthesis design (Voll and Marquardt, 2012a,b, Marvin et al., 2013), no attention was paid to unconverted intermediates; they were removed from the network without further consideration. However, the composition of the desired product is defined as a degree of freedom; it is only limited in terms of exhibiting required thermophysical properties. Hence, the unconverted intermediates can be used in conjunction with the initially defined target compound to form a fuel blend that fulfills the imposed property constraints. An optimization-based methodology will be presented as additional step to the multi-stage network evaluation to assess the potential of this approach towards increasing feedstock utilization.

The following sections will discuss each stage of the solution strategy in more detail.

5.2.1 Decomposition of the network into individual pathways

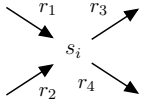
Combinatorial decomposition of reaction networks into individual pathways is an established methodology in metabolic engineering and systems biology, known as *elementary mode analysis* (Stelling et al., 2002, Gagneur and Klamt, 2004, Terzer and Stelling, 2008). An elementary mode (EM) denotes a unique sequence of reactions to produce a desired target compound from a specified substrate. The terminus *elementary* relates to the fact that an EM cannot be decomposed into two or more other modes.

The EMs of a network can be deduced without knowing any flux rate, providing an external parameter or relying on an objective function. Each EM represents a solution vector $\mathbf{f}$ to the initial problem formulation presented in Equation (5.1). The EMs define the boundaries of the flux distribution in the network; an optimal flux distribution is then described by a single EM or linear combination of several.

Elementary mode analysis is commonly carried out by determining the null space of the stoichiometric matrix, which leads to repeated calculation of individual EMs (Stelling et al., 2002). A novel approach is presented here that gives inherently computes the EMs without redundancy, which otherwise is the case in null space calculations, such that only distinct pathways are constructed. Its computational efficiency was so far not compared against established methodologies.

The identification of EMs suggested in this work relies on forming prolonging reaction paths from ordered pairs of reactions. Ordered pairs of reactions are 2-tuples of network reactions, where the first member of the tuple is a reaction entering a vertex s_i, and the second member is a reaction leaving the same vertex. To identify those tuples, the entering

Table 5.1 – Entering and leaving reactions at a substance vertex s_i and the formation of ordered pairs of reactions

$G(S,R)$	$link_i^+$	$link_i^-$	op
	$\{r_1, r_2\}$	$\{r_3, r_4\}$	$\{r_1, r_3\}$ $\{r_1, r_4\}$ $\{r_2, r_3\}$ $\{r_2, r_4\}$

and leaving reactions at each vertex i of the network are identified and denoted in the sets $link_i^+$ and $link_i^-$, respectively. This concept is adapted from Fenves (1967) and is based on the work of Ford and Fulkerson (1962). The ordered pairs at a vertex s_i are constructed by forming the Cartesian product (see Table 5.1 for an illustrating example)

$$op_i = link_i^+ \times link_i^-. \tag{5.4}$$

The set of ordered pairs OP of the entire network is the union of the individual sets of ordered pairs op_i, expressed by

$$OP = \bigcup_{i=1}^{N_s} op_i. \tag{5.5}$$

An element in OP is addressed by op_q with $q \in \{1, ..., N_{OP}\}$. N_{OP} is the number of ordered pairs. $op_{q,1}$ is the first (entering) and $op_{q,2}$ the second (leaving) member of an ordered pair.

EMs are described as sequences of reactions, forming reaction network paths. A path j of length k is addressed by p_j^k. The individual reactions in a path are addressed by $p_{j,i}^k$ with $i \in \{1, ..., k\}$. All m paths of length k are concatenated in the set

$$P^k = \{p_1^k, ..., p_m^k\}. \tag{5.6}$$

Three requirements are imposed on a path to qualify as an EM:

1. A path begins at source reaction r_{source}.

2. A path ends at sink reaction r_{sink}.

3. A path does not contain a reaction twice, otherwise a loop would be formed.

The EMs are constructed by stepwise increasing the pathway length k by adding ordered pairs that are incident to $\mathrm{p}_{j,k}^k$. The initial set of paths P^2 is the set of ordered pairs, such

that

$$P^2 = OP \quad \forall \quad op_q : op_{q,1} = r_{source}. \tag{5.7}$$

Equation 5.7 requires that the elementary mode analysis starts at the network's source reaction r_{source}; hence, only such ordered pairs are considered in P^2 that contain r_{source} in op_1. New path members are always added to that end of the path that is averted to the source. New path members are identified by

$$p^k_{j,k} = op_{q,1}, \tag{5.8}$$

stating that these ordered pairs are incident to $p^k_{j,k}$. Several ordered pairs can be incident to $p^k_{j,k}$, resulting in multiple paths p^{k+1} from p^k. Only such ordered pairs are allowed to be considered where

$$op_{q,2} \notin p^k_j. \tag{5.9}$$

This ensures that reaction $op_{q,2}$ is so far not contained in p^k_j, such that no loops are formed.

Assume that IOP denotes the ordered pairs that fulfill Equations (5.8) and (5.9). NOP is the number of elements in IOP. The new paths of length $k+1$ are then formed by

$$\begin{aligned} p^{k+1}_1 &= p^k_j \cup IOP_{1,2} \\ p^{k+1}_2 &= p^k_j \cup IOP_{2,2} \\ &. \\ &. \\ p^{k+1}_{NOP} &= p^k_j \cup IOP_{NOP,2} \end{aligned}$$

The paths j with $p^k_{j,1} = r_{source}$ and $p^k_{j,k} = r_{sink}$ denote completed EMs and are stored as a vector $\mathbf{f}^{\,r\times 1}$, where the entries f_r denote whether reaction r is active in this particular path ($f_r = 1$) or not ($f_r = 0$). By starting from the source of the network and following the progress of the network reactions while avoiding circular flows, only unique pathways are generated and concatenated in the **f**-vectors; extraction of distinct pathways is not required.

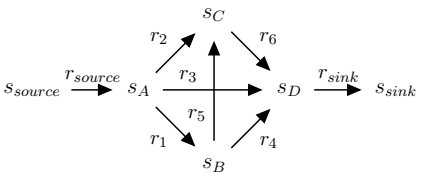
All vectors $\mathbf{f}_1, ..., \mathbf{f}_{em}$ are concatenated in the flux matrix $\mathbf{F} \in (0,1)^{r\times em}$. The rows of $\mathbf{F}$ refer to the network reactions $j \in \{1, ..., r\}$, the columns represent the EMs $m \in \{1, ..., em\}$. The number of elementary modes em is an important information to derive from the network's topology. The more EMs leading to the target compound, the more alternatives are still available in case that a chosen pathway fails. A high number of alternatives

is desirable, since the feasibility of the computationally generated reactions cannot be guaranteed. The column sum of **F** gives the number of reactions per EM. The row sum of **F** gives the number of occurrences of a reaction j in all EMs. Since EMs are not further decomposable, they fail if any one reaction of the mode cannot be performed. Frequently encountered reactions determine the feasibility of many production pathways and hence should be in the center of attention when transferring the results from the computational model to the experimental research.

Example 20 illustrates the application of the proposed elementary mode analysis to a small representative network.

Example 20. *Consider the network depicted left in Figure 5.1, which is composed of 6 vertices and 8 reactions. The elementary mode analysis aims at identifying all existing pathways between the vertices s_{source} and s_{sink}.*

The first step is to derive the sets $link_i^+$ and $link_i^-$ by denoting the entering and leaving reactions at each vertex. These sets for the exemplary network are presented right in Figure 5.1.



s_i	$link_i^+$	$link_i^-$
s_{source}	$\emptyset$	$\{r_{source}\}$
s_A	$\{r_{source}\}$	$\{r_1, r_2, r_3\}$
s_B	$\{r_1\}$	$\{r_4, r_5\}$
s_C	$\{r_2, r_5\}$	$\{r_6\}$
s_D	$\{r_3, r_4, r_6\}$	$\{r_{sink}\}$
s_{sink}	$\{r_{sink}\}$	$\emptyset$

Figure 5.1 – Exemplary network (left) and corresponding sets $link_i^+$ and $link_i^-$ (right)

The sets $link_i^+$ and $link_i^-$ provide the basis to form the ordered pairs of reactions. Every ordered pair of entering and leaving reactions at each compound is formed to result in the set

$$OP = \{\{r_{source}, r_1\}, \{r_{source}, r_2\}, \{r_{source}, r_3\}, \{r_1, r_4\}, \{r_1, r_5\},$$
$$\{r_2, r_6\}, \{r_5, r_6\}, \{r_3, r_{sink}\}, \{r_4, r_{sink}\}, \{r_6, r_{sink}\}\}$$

In total, 10 ordered pairs of reactions exist in the network. They are used to construct the pathways that connect source and sink. The ordered pairs that contain r_{source} as entering reaction serve as the initial pathways P^2. Stepwise, incident ordered pairs are identified to prolong each path until it reaches r_{sink}. Table 5.2 presents the stepwise increase of the length of the pathways by denoting the length of the pathways k and the comprised reactions. The last column denotes whether the current path denotes a sequence from source to sink or not.

Table 5.2 – stepwise construction of the network paths by increasing the length k and combining paths and ordered pairs

k	P^k	IOP	r_{source},$r_{sink} \in P^k$
2	$\{r_{source}, r_1\}$	$\{r_1, r_4\}$,$\{r_1, r_5\}$	no
2	$\{r_{source}, r_2\}$	$\{r_2, r_6\}$	no
2	$\{r_{source}, r_3\}$	$\{r_3, r_{sink}\}$	no
3	$\{r_{source}, r_1, r_4\}$	$\{r_4, r_{sink}\}$	no
3	$\{r_{source}, r_1, r_5\}$	$\{r_5, r_6\}$	no
3	$\{r_{source}, r_2, r_6\}$	$\{r_6, r_{sink}\}$	no
3	$\{r_{source}, r_3, r_{sink}\}$	$\emptyset$	yes
4	$\{r_{source}, r_1, r_4, r_{sink}\}$	$\emptyset$	yes
4	$\{r_{source}, r_1, r_5, r_6\}$	$\{r_6, r_{sink}\}$	no
4	$\{r_{source}, r_2, r_6, r_{sink}\}$	$\emptyset$	yes
5	$\{r_{source}, r_1, r_5, r_6, r_{sink}\}$	$\emptyset$	yes

The indices of the reactions in the individual paths are transferred to the flux matrix $\boldsymbol{F}$*. 4 EMs exist in this example. Besides the pseudo-reactions (*r_{source} *and* r_{sink}*), that appear in every reaction,* r_1 *and* r_6 *participate in 2 EMs each and are the most frequently encountered reactions in* $\boldsymbol{F}$*. From this theoretical example results the implication that validating the feasibility of* r_5 *should be of highest interest.*

$$\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4\} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

The network reactions r_j span a r-dimensional Cartesian coordinate system where each reaction represents a unit vector $\mathbf{b}_j$ with $_{j\in\{1,\ldots,r\}}$. Each column of $\mathbf{F}$ is a specific combination of the network reactions. It represents an edge of a convex polyhedral cone, the so-called *flux cone* (Wagner and Urbanczik, 2005). Each point within the flux cone is a valid state of the network. Several EMs can be linearly combined and scaled with a non-negative factor $\boldsymbol{\alpha} = \alpha_m \in [0, 1]^{1\times em}$ to result in a flux distribution

$$\mathbf{v} = \sum_{m=1}^{em} \alpha_m \cdot \mathbf{f}_m, \quad \alpha_m \geq 0. \tag{5.10}$$

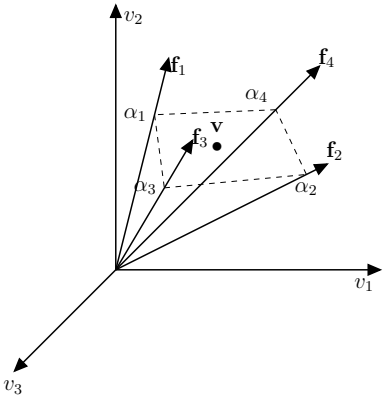


Figure 5.2 – Flux cone for a network with 3 reactions and 4 resulting EMs

Each α_m represents the share of elementary mode m in the final flux distribution. The individual contributions α_m sum up to

$$\sum_{m=1}^{em} \alpha_m = 1. \tag{5.11}$$

A qualitative representation of a flux cone is presented in Figure 5.2. The EMs are scaled with α_m to give the final flux distribution $\mathbf{v}$. This is an idealizing scenario since $\mathbf{v}$ does not consider incomplete conversion or non-selective reaction pathways. The next section presents how such non-ideal network reactions are accounted for.

5.2.2 Flux balancing in elementary modes

The fluxes in each EM are calculated by balancing the incoming and leaving streams of the vertices. Yield constraints on reactions from incomplete conversion and non-selective behavior of the network reactions are considered. A cutout of a reaction sequence is presented in Figure 5.3 to explain the calculation of the molar fluxes. Assume that the reactions r_{source}, r_j and r_{j+1} are active in an elementary mode m. r_{source} provides the main substrate to the network. Reaction r_j is of the form

$$A \; + \; B \; \longrightarrow \; C \; + \; D. \tag{5.12}$$

The main substrate is denoted by A, the provided reactant by B, the desired product by C and the formed by-product by D.

A molar flux $n_{i,j,m}$ denotes the flux of substance i in reaction j of elementary mode m.

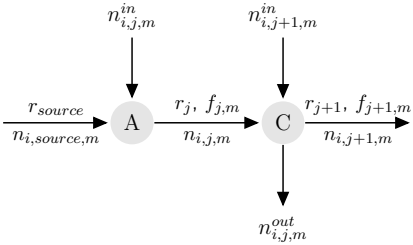


Figure 5.3 – Analogy between network and process representation of reaction pathways

It is normalized with respect to 1 mole of substrate. r_{source} provides the stream of pure main substrate $n_{A,source,m}$ to the network. The required reactant B is provided via stream $n^{in}_{B,j,m}$. Reactants are always provided in stoichiometric amounts such that they are not limiting the reaction. The stoichiometric coefficient $v_{i,j}$ of reaction r_j defines the amount of each produced compound to be

$$n_{i,j,m} = (n_{A,source,m} + n^{in}_{B,j,m}) \cdot v_{i,j}. \tag{5.13}$$

This formulation implies a reaction yield of 100%. However, reaction yields are usually less than 100% due to non-ideal conversion and/or selectivity. The yield Y_j of a reaction r_j is the product of conversion C_j and selectivity S_j, such that

$$Y_j = C_j \cdot S_j. \tag{5.14}$$

Considering incomplete conversion, the composition of $n_{i,j,m}$ is then

$$n_{i,j,m} = \underbrace{Y_j \cdot (n_{A,source,m} + n^{in}_{B,j,m}) \cdot v_{i,j}}_{\textit{produced substances } C \textit{ and } D} + \underbrace{(1 - Y_j) \cdot (n_{A,source,m} + n^{in}_{B,j,m})}_{\textit{unconverted substances } A \textit{ and } B}. \tag{5.15}$$

The right hand side of Equation (5.15) consists of two terms. The first term refers to the produced substances, the second to unconverted substrate and reactant. From this mixture, the desired product C is separated, assuming sharps splits. The molar flux of the desired product, $n_{C,j,m}$, results in

$$n_{C,j,m} = Y_j \cdot n_{A,source,m} \cdot v_{A \to C,j}. \tag{5.16}$$

Unconverted amounts of A, unconverted reactant B and the generated by-product D are

concatenated in the flux $n^{out}_{i,j,m}$. The composition of $n^{out}_{i,j,m}$ is calculated to be

$$n^{out}_{i,j,m} = Y_j \cdot (n_{A,j,m} + n^{in}_{B,j,m}) \cdot v_{i \neq C,j} + (1 - Y_j) \cdot (n_{A,j,m} + n^{in}_{B,j,m}). \tag{5.17}$$

$n^{out}_{i,j,m}$ thus includes every substance except the desired product C, which is expressed by considering only those stoichiometric coefficients where $i \neq C$.

The normalized fluxes rates are summarized in matrix notation, such that

$$\mathbf{n}_m = (n_{i,j,m}) \in \mathbb{R}^{s \times r} \tag{5.18}$$

$$\mathbf{n}^{in}_m = (n^{in}_{i,j,m}) \in \mathbb{R}^{s \times r} \tag{5.19}$$

$$\mathbf{n}^{out}_m = (n^{in}_{i,j,m}) \in \mathbb{R}^{s \times r}. \tag{5.20}$$

Each matrix in Equations (5.18) - (5.20) concatenates the fluxes of the substances i (rows) in the reactions j (columns) for each elementary mode m. Multiplication of the normalized fluxes with a scaling factor γ, which represents the molar amount of substrate provided to the network, gives the absolute molar fluxes, as presented in Equations (5.21) - (5.23).

$$\mathbf{N}_m = \gamma \cdot \mathbf{n}_m \tag{5.21}$$

$$\mathbf{N}^{in}_m = \gamma \cdot \mathbf{n}^{in}_m \tag{5.22}$$

$$\mathbf{N}^{out}_m = \gamma \cdot \mathbf{n}^{out}_m \tag{5.23}$$

5.2.3 Determination of the optimal flux distribution

An optimal network flux distribution, which is optimal subject to a specified objective function, is a linear combination of the elementary modes in the network. The flux distribution is determined by setting up and solving an optimization problem, in which the contributions $\boldsymbol{\alpha} \in [0, 1]^{1 \times em}$ of each elementary mode m to the flux distribution are determined. In order to maintain the scaling of the network, the sum of all individual contributions α_m is 1. Multiplication of the contributions α_m with the molar fluxes $\mathbf{N}_m$, $\mathbf{N}^{in}_m$ and $\mathbf{N}^{out}_m$ yields

the absolute fluxes in the network. The formulation of the optimization problem is

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \phi \\
\text{subject to} \quad & \phi = \sum_{m=1}^{em} \alpha_m \cdot \phi_m, \\
& \phi_m = f(\mathbf{N}_m, \mathbf{N}_m^{in}, \mathbf{N}_m^{out}), \\
& \mathbf{N} = \sum_{m=1}^{em} \alpha_m \cdot \mathbf{N}_m, \\
& \mathbf{N}^{in} = \sum_{m=1}^{em} \alpha_m \cdot \mathbf{N}_m^{in}, \\
& \mathbf{N}^{out} = \sum_{m=1}^{em} \alpha_m \cdot \mathbf{N}_m^{out}, \\
& \sum_{n=1}^{em} \alpha_m = 1, \\
& \alpha_m \in [0,1]^{1 \times em}.
\end{aligned}
\tag{5.24}
$$

$\boldsymbol{\phi}$ is the objective function, which is linearly composed of the individual contributions ϕ_m from each EM. ϕ_m is a function of either the entering ($\mathbf{N}_m^{in}$), leaving ($\mathbf{N}_m^{out}$) or reaction fluxes ($\mathbf{N}_m$), or ratios/combinations thereof. Objective functions that are meaningful in the context of biofuel synthesis planning are presented in Appendix E.

5.2.4 Integration of intermediate waste streams

The workflow presented in the previous chapter is designed to evaluate the synthesis pathways towards a pure substance, in the sense of the formulation of Hechinger et al. (2010). However, incomplete reaction yields lead to streams of unconverted substances that are concatenated in $\mathbf{N}_m^{out}$. The evaluation strategies introduced so far (Yim et al., 2011, Voll and Marquardt, 2012a,b, Marvin et al., 2013) do not consider the potential inherently contained in these substances, but rather consider them as waste or do not consider them at all. However, if the real-life use case is not restricted to using a pure substance, the target compound can be blended with unconverted intermediates, as long as imposed product specifications on thermophysical properties are not violated.

The use case of a fuel, as considered in TMFB, imposes constraints on the thermophysical properties, but not on the chemical composition of the product. These property constraints can also be satisfied by a mixture, which is produced from integrating the streams of unconverted intermediates and the target compound. This integration increases the quantitative

fuel output while maintaining the feedstock consumption, which in return increases the degree of feedstock utilization and therefore the economic efficiency.

The properties of the mixture are calculated based on linear mixing of the properties calculated by the QSPR models denoted in Appendix D. Linear mixing rules impose negligible intermolecular interactions of the mixture partners, such that it gets less accurate with the presence of polar constituents in the mixture (Kontogeorgis and Folas, 2009). Since the network intermediates often contain oxygen in functional arrangements (which is one cause to polarity (Furniss et al., 1989)), linear mixing rules can be inadequate for a accurate description of the considered mixtures. Therefore, the framework for calculating mixture properties is flexible in a way that linear mixing rules can be easily exchanged by more sophisticated methods, e.g. a combination of equation of state and linear and quadratic mixing rules (Kontogeorgis and Folas, 2009). In the case of using only linear mixing rules, a property of the mixture, P_{mix}, results from the contributions P_i of each substance i multiplied with its molar fraction y_i, such that

$$P_{mix} = \sum_{i=1}^{s} y_i \cdot P_i. \tag{5.25}$$

The mixture has to fulfill the property constraints that are imposed on the target, which is expressed by

$$P_{min} \leq P_{mix} \leq P_{max}. \tag{5.26}$$

Since this mixture formation task is under-determined and leads to an infinite number of solutions, it is formulated as an optimization problem. The target function is to maximize the feedstock utilization that is defined as the molar ratio of output to feedstock.

The available amounts of unconverted intermediates and target compound are concatenated in $\mathbf{N}^{out}$. The ratio of integration of each stream in $\mathbf{N}^{out}$ is denoted by $\mathbf{x} \in [0,1]^s$. The values x_i represent the fraction of the stream of substance i that is included in the final mixture. x_i takes values between 0 (stream completely neglected) and 1 (stream completely integrated). y_i represents the molar fraction of component i in the final mixture. y_i can take values between 0 (substance not included) and 1 (output consists only of compound i). The sum of the individual contributions y_i has to sum up to 1. The formulation

of the optimization problem results in

$$\begin{aligned} \max_x & \frac{\sum_{i=1}^{s} x_i \cdot (\sum_{j=1}^{r} N_{i,j}^{out})}{N_{FS,source}^{in}} \\ s.t. \quad & y_i = \frac{x_i \cdot N_{i,j}^{out}}{\sum_{i=1}^{n} x_i \cdot N_{i,j}^{out}} \\ & P_{mix} = \sum_{i=1}^{s} y_i \cdot P_i \\ & P_{min} \leq P_{mix} \leq P_{max} \\ & \sum_{i=1}^{s} y_i = 1 \\ & x_i \in [0,1] \\ & y_i \in [0,1] \end{aligned} \tag{5.27}$$

The formulation is a non-linear, constrained optimization problem, which is then solved using the optimization environment TOMLAB (Holmström, 1999), available in MATLAB (2010).

5.3 Discussion of the evaluation strategies

Optimization-based (Varma and Palsson, 1994, Schilling et al., 2000) and combinatorial evaluation (Papin et al., 2004) strategies are both frequently applied in the study of metabolic networks. Despite their original heritage, they can both be employed in other fields such as the flux analysis of chemical reaction networks. In the following sections, optimization-based network evaluation strategies, as employed by Voll and Marquardt (2012b), Yim et al. (2011), Marvin et al. (2013), are compared against the newly introduced multi-stage evaluation procedure.

5.3.1 Optimization based evaluation strategy

Optimization-based techniques are used for determining the stationary flux distribution in metabolic networks (Varma and Palsson, 1994, Schilling et al., 2000). In their basic application, they yield the flux distribution of the pathway which is optimal with respect to a meaningful objective function. Lee et al. (2000) presents a workflow to derive all available pathways in a network. It relies on binary variables u_j which represent the activity of one reaction in a pathway. This statement is incorporated into the optimization

problem via

$$f_j \leq U_j u_j \quad \forall j \in \{1, ..., N_r\}, \tag{5.28}$$

$$\mathbf{u} \in \{0, 1\}^{N_r}. \tag{5.29}$$

U_j is a sufficiently large parameter that has to be chosen according to the expected fluxes. A variable u_j is set to one if a flux f_j occurs. Essentially, every u_j is identical to a row in $\mathbf{F}$. These flux variables are then integrated into the optimization problem to yield a mixed-integer problem (MIP) formulation. Additional constraints are included that change the composition of u for each run, until every feasible reaction sequence in the network is elucidated. The optimization is performed for each identified reaction sequence. Thus each U_j has to be chosen such that in every run, it suffices to the requirements of the current run. If the parameter is not adequately chosen, then the calculated results will either be incomplete or truncated, since false or incomplete reaction sequences will be calculated.

Voll and Marquardt (2012b) presented the only contribution in the context of biofuel synthesis to include an MIP strategy as introduced by Lee et al. (2000) to enumerate all available pathways in their synthesis. The evaluation is carried out under the objective of maximizing product yield. A change of the objective function requires performing the complete optimization process again. In order to get all flux distributions optimal to all criteria, the procedure has to be carried for every evaluation criterion (cf. Figure 5.4). Furthermore, neither Voll and Marquardt (2012b) nor Yim et al. (2011) and Marvin et al. (2013) assess the topology of the synthesis network to derive statements on network robustness and importance of individual reactions.

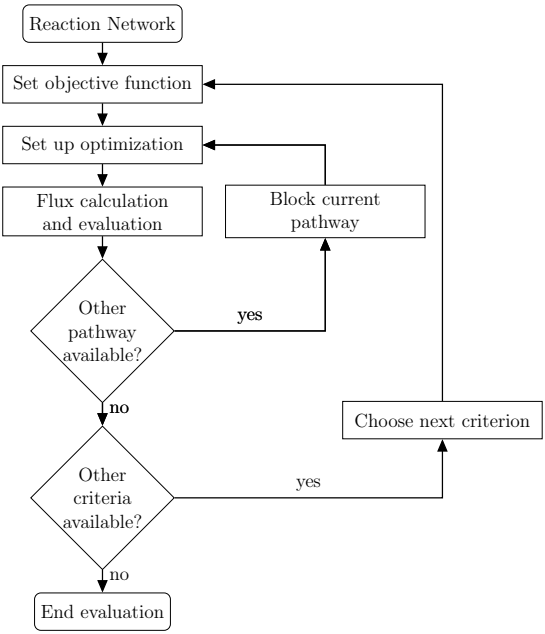


Figure 5.4 – Depiction of an optimization-based evaluation strategy to identify and evaluate all pathways in a reaction network

5.3.2 Multi-stage evaluation strategy

A conceptual representation of the proposed multi-stage evaluation strategy is presented in Figure 5.5. Starting from the network representation, the network is first decomposed into its EMs. These EMs determine the boundaries of the network flux distribution. The entering, leaving and internal molar fluxes $\mathbf{N}_m^{in}$, $\mathbf{N}_m^{out}$ and $\mathbf{N}_m$ of each EM are then calculated. The optimal flux distribution is a linear combination of the individual linear elementary modes. A LP problem is formulated in Equation (5.24) to determine the individual contributions α_m of each elementary mode m to the final flux distribution. The potential of waste stream integration can then be determined by forming a mixture of desired product and non-converted intermediates.

Elementary mode analysis neither requires an objective function nor any externally provided parameter. The determined EMs allow for statements on the robustness of a design task and the importance of each network reaction. In contrast to optimization-based approaches, a variation of the objective function does not require the recalculation of the network fluxes, since they are already known and stated in $\mathbf{N}_m$, $\mathbf{N}_m^{in}$ and $\mathbf{N}_m^{out}$. Only a new

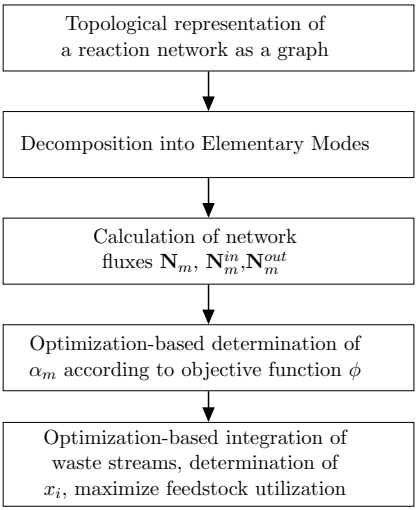


Figure 5.5 – Multi-stage evaluation of reaction networks

linear combination of EMs needs to be derived for a differing objective function according to Equation (5.24).

It has to be stressed that if a reaction network contains more than say 10^5 EMs, elementary mode analysis gets time-consuming and requires large computational power (Papin et al., 2004). The networks should then not be decomposed into EMs to avoid computational difficulties. However, the threshold of 10^5 will certainly raise with increasing computational power in the future.

5.4 Conclusions

This chapter introduces a novel strategy for the evaluation of biofuel reaction pathways, that is based on a combination of elementary mode analysis (Gagneur and Klamt, 2004, Terzer and Stelling, 2008) and optimization-based determination of optimal flux distributions. The integration of streams of unconverted intermediates into the final product allows for increasing the degree of utilization of the processed feedstock and widens the scope of the assessment towards biofuel blends and their potentials.

Elementary mode analysis decomposes the network into unique pathways that connect sink and source of a network. These modes allow for statements on the importance of distinct reactions, the robustness of a solution to the design task and the boundaries for the flux distribution in the network (Stelling et al., 2002). The determination of the

EMs does not require the provision of external parameters or an objective function. The additional information was so far not retrieved in any previous evaluation strategy for biofuel reaction networks.

In contrast to purely optimization-based evaluation strategies as performed previously (Yim et al., 2011, Voll and Marquardt, 2012a,b, Marvin et al., 2013), the multi-stage approach relies on a completely evaluated network in terms of occurring network flux distributions and reaction fluxes. Optimal flux distributions are determined from the evaluated EMs, objective functions are functions of the calculated fluxes. Calculations with a novel objective function do not require a repeated evaluation of the network fluxes, but only a new combination of the EMs, rendering the evaluation strategy more time-efficient than previous approaches.

The integration of waste streams has never been considered in literature and is enabled by the fuel design approach of TMFB, which defines the biofuel's molecular structure as additional degree of freedom. The biofuel is only constrained by a set of required properties, not by its molecular composition. Thus, mixtures can be created that give the desired properties without restrictions on the molecular constitution. The effectiveness of this approach can most certainly be increased if the composition of the mixture is designed from the very beginning. Also, the use of linear mixing rules often strongly simplifies the real behavior of the mixture. However, by exchanging Equation (5.25) with a more rigorous model, the mixture behavior can be described with more accuracy.

The values of the evaluation criteria should be taken with caution, as many assumptions and simplifications have been made in order to allow for an early estimation of process performance. However, with a reflective interpretation of the performance indicators, the evaluation gives helpful information that supports determining the most promising combinations of substrates and targets and also the best performing associated synthesis pathways.

6 Case studies

This chapter presents the application of the previously described methodologies to two synthesis tasks. The first task describes the synthesis of 3-methyltetrahydrofuran (3-MTHF) from itaconic acid (IA). This combination of substrate and target was identified in the Cluster of Excellence "Tailor-Made Fuels from Biomass" (TMFB) as a high performing biofuel synthesis process (Voll and Marquardt, 2012b). 3-MTHF qualifies as a biofuel for compression ignition (CI) engines by fulfilling property constraints that were defined within TMFB.

The second case study evaluates the synthesis processes of two novel biofuels that were identified by TMFB researchers as both, biofuel candidates and intermediates to the synthesis of 1-octanol (Julis and Leitner, 2012). The targeted substances, 2-butylfuran (2-BF) and 2-butyltetrahydrofuran (2-BTHF), are predicted to exhibit the required thermophysical properties which qualify them as highly promising biofuels for CI engines. The economic and ecological performance of their production processes is assessed for the first time in this study.

6.1 Defining a reference process

The calculated performance indicators of the proposed synthesis routes need to be set into relation to existing biofuel synthesis processes to evaluate their performance compared to established processes. Only such routes that are comparable to or even outperform such references can be considered as competitive alternatives. To ensure comparability of the proposed and the reference synthesis process, the performance indicators of both processes are determined using the same metrics.

3-MTHF is a fuel that is supposed to be used in CI engines. Currently, the production of fatty acid methyl esters (FAME) is the most frequently employed production process for CI-suited biofuels. FAME fuel is produced via trans-esterification of a triglyceride under the presence of methanol, giving three instances of FAME and the side product glycerol by a base-catalyzed reaction (Knothe et al., 2005). The conversion of this processes is approximately 100% with a selectivity of 98% (Ma and Hanna, 1999). FAME production leads to a multi-component mixture, containing fatty acid methyl esters of different length

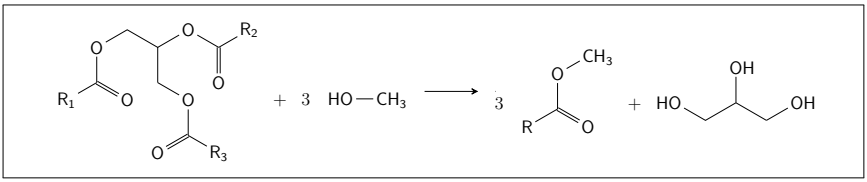


Figure 6.1 – Schematic description of FAME production

and degree of unsaturation along with the side product glycerol.

The synthesis of FAME fuels is schematically presented in Figure 6.1. The vegetable oil is a triglyceride with three alkyl residues, R_1 - R_3, which each usually contain between 14-18 partially unsaturated carbon atoms. Each feedstock has a certain ratio between the different combinations of chain length and degree of unsaturation. Even vegetable oils stemming from one plant carry alkyl residues of varying length and saturation. One ratio that is often encountered in nature is 18:2 carbon atoms to double bonds (Ma and Hanna, 1999). To simplify the definition of a reference process, it is imposed that one representative from the (18:2)-fraction, linoleic acid, serves as general alkyl residue in the triglyceride. In this way, the synthesis does not yield a mixture of different fatty acid methyl esters but a defined substance, such that the reference process and the novel process can be evaluated on equal terms.

The annual reference amount of production is 190 million liters, which corresponds to the size of bioethanol production facilities currently under construction (Ethanol Producer Magazine, 2014). FAME is assumed to be produced from rapeseed oil, which is currently traded at 1,000 \$/t (Index mundi, 2014). Methanol, which serves as a reactant, is currently traded at approximately 600 \$/t (Methanex, 2014). A plant lifetime of 10 years and an interest rate of 8% are attributed to FAME production.

6.2 Lignin gasification for hydrogen production

Biomass is composed of the three main constituents cellulose, hemicellulose and lignin. In the envisioned concept of fuel production in TMFB, the biofuel will be produced in a biorefinery (Kamm et al., 2008), an integrated facility for the simultaneous production of various products from biomass. The biomass is transported to the site and disintegrated afterwards. The disintegration can for instance be performed by means of Organosolv (McDonough, 1992) or the newly developed OrganoCat process (vom Stein et al., 2011). The disintegration separates the biomass constituents from each other. Platform chemicals such as those presented by Werpy et al. (2004) are then synthesized from cellulose and

hemicellulose. Several approaches propose to make material use of lignin by producing lignin-based carbon fibers (Kadla et al., 2002) or fuels and bulk chemicals (Zakzeski et al., 2010). However, these approaches are not yet commercially viable.

In the presented case studies, lignin is considered as an optional hydrogen source. In a gasification process, that is installed parallel to the biofuel production process (cf. Figure 6.2), hydrogen is produced from lignin. This process bears energetic losses ΔE and emits CO_2, which has to be accounted for in the evaluation.

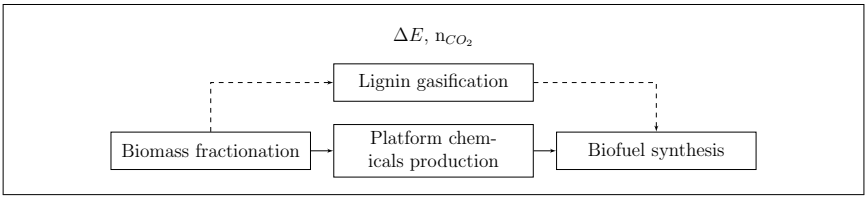


Figure 6.2 – Schematic process for lignin gasification

The conversion of lignin into hydrogen and carbon dioxide is quantified using the findings of Baumlin et al. (2006), who estimated the maximum hydrogen yield from Kraft-lignin gasification to be

$$C_6H_{6.3}O_{2.2} + 9.8\ H_2O \rightarrow 12.95\ H_2 + 6\ CO_2. \tag{6.1}$$

Hydrogen is produced from lignin via a fast pyrolysis of the biomass with subsequent cracking of residual vapors, steam reforming, water-gas-shift reaction and steam gasification of the residual pyrolysis char. These steps are concatenated in one reaction step and evaluated with the same metrics as presented in Appendix E.

6.3 Synthesis of 3-MTHF from itaconic acid

IA was identified as platform chemical by Werpy et al. (2004) based upon the rich amount of contained functionality. It is derived from glucose by aerobic fungal fermentation under the presence of *Aspergillus terreus* (Okabe et al., 2009). TMFB research is focussing on the production of IA through fermentation from green biomass using the corn smut fungus *Ustilago maydis* (Klement et al., 2012).

3-MTHF was identified in TMFB as a potential biofuel candidate (Wimmer et al., 2010) for use in CI engines. It combines a high energy density (H_{com}) with a boiling point (T_{boil}) sufficiently low to avoid oil dilution (which occurs if the boiling temperature of the unburnt fuel is higher than that of the engine oil). The melting point (T_{melt}) is sufficiently

Table 6.1 – Fuel-relevant properties of 3-MTHF and fuel property constraints relevant to CI engines, predicted by using the QSPR models presented in Appendix E

Structure		T_{boil}	ρ_L	ΔH_{com}	T_{melt}	CN
–	–	K	kg/l	MJ/kg	K	–
	Prediction	373.1	0.87	-33.2	162.0	40
	Constraint	≤623.15	≥0.7	≤-32	≤ 253.15	≥30

low to enable appropriate cold flow behavior. Besides these properties, combustion relevant properties such as auto-ignition behavior under compression are crucial for property functionality of CI engines. Just recently Dahmen et al. (2012) proposed a QSPR model for the prediction of cetane numbers (CN) and thermophysical properties of diesel fuel candidates. These models are used in this thesis to calculate the CN and the thermophysical properties of the network substances. The estimated properties of 3-MTHF are presented in Table 6.1, along with the imposed fuel property constraints.

6.3.1 Manually constructed reaction network

Previous approaches to the evaluation of 3-MTHF production from itaconic were based on a manually assembled network. Voll and Marquardt (2012a) used the reaction network presented in Figure 6.3 to identify promising synthesis pathways. It is based on data retrieved from exhaustive literature research using the web-based chemicals and reactions library SciFinder (American Chemical Society, 2014). The reactions were chosen such that the gravimetric heating value of IA is successively increased.

It turned out that only a rather small amount of reactions are known that lead from IA to 3-MTHF. Some of them are quantified experimentally regarding conversion and selectivity or have just been proposed and still need to be demonstrated in practice. In total, the manually assembled network comprises 8 substances, 8 reactions, and 2 elementary modes. It is stated in the literature that reactions R_3 and R_4 have to occur simultaneously (Geilen et al., 2010). This implies that the two arising elementary modes will occur together in the resulting flux distribution.

Voll and Marquardt (2012a) report total annualized costs (TAC) ranging from 50 to approximately 80 M$/a for a production capacity equivalent to the energetic content of 100,000 tons of ethanol per year. Different scenarios were investigated that represent 3-MTHF production with and without gasification of lignin to replace external hydrogen. The prices for feedstock and hydrogen were set to 50 $/t and 2,700 $/t, respectively.

Voll and Marquardt (2012a) also evaluated the environmental impact of the production

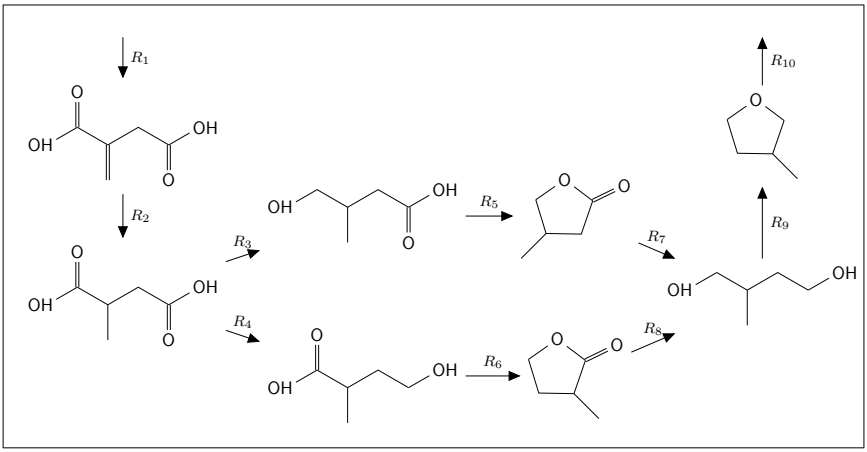


Figure 6.3 – Manually assembled reaction network for the synthesis of 3-MTHF from itaconic acid (Voll and Marquardt, 2012b)

pathways. It was stated that the environmental impact of the processes increases when including lignin gasification for hydrogen production. This effect is caused by the additional CO_2 emissions from the lignin gasification. However, the calculated values of the individual contributions were not related to a reference process such as bioethanol or FAME, but to the ecologically worst performing pathway in the network. To elucidate how the proposed synthesis pathways compare against an established biofuel production process and not amongst each other, FAME production was chosen as a reference entity in this thesis.

6.3.2 Automated reaction network generation for 3-MTHF synthesis

The computational generation of the reaction network for 3-MTHF synthesis from IA was performed according to the settings listed in Table 6.2. These settings were chosen to reflect the design paradigms of the network constructed by Voll and Marquardt (2012a) and restrict ReNeGen to a comparable molecular space.

The number of stages is set high to ensure the generation of every relevant reaction. The only reactant that is added is hydrogen (as only oxygenated functionality shall be removed). In addition, the system is allowed to isomerize (perform a reaction without the presence of any reactant). Several reaction rules are further included in order to target the generation. They focus on alteration and cleavage of the oxygenated and unsaturated functionality. The main product of a generated molecular ensemble (ME) is determined by the highest value of ΔH_{com} to ensure that the substances with the highest gravimet-

Table 6.2 – Scenario definition for 3-MTHF synthesis of IA

Setting	Value
Substrate	Itaconic acid
Target	3-Methyltetrahydrofuran
Number of stages	15
Reactant	H_2
Reaction rules	1-7, 11
Main product criterion	ΔH_{com}
Biomass composition	0.6:0.2:0.2

ric energy content are forwarded to the next network stage. The biomass composition (cellulose:hemi-cellulose:lignin) is required to quantify hydrogen production from lignin. It is chosen to represent an average, but not further determined biomass. For simplicity, the individual values describe the molar fractions of the elementary compositions of cellulose, hemicellulose and lignin, i.e. $C_6H_{10}O_5$, $C_5H_{10}O_5$ and $C_6H_{6.3}O_{2.2}$.

Table 6.3 – Structural constraints on reaction network intermediates for 3-MTHF synthesis from IA

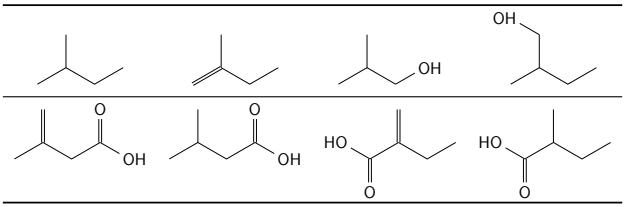
Structural feature	Setting
Elementary composition	$C_5H_{0-10}O_{0-5}$
Number of rings	0-1
Ring size	4-6

The generated substances have to satisfy the requirements on their molecular constitution that are stated in Table 6.3. The elementary composition defines the molecular space in which the molecules can be refunctionalized, but inhibits polymerization and atomization. Each molecule is allowed to carry up to one ring; multiple, also fused rings and ring systems are not allowed. The ring size is restricted to only such arrangements that are stable according to expert knowledge.

Topological operations on the generated network

The generated network contains 52 intermediates and 131 reactions. Network reduction is carried out requiring 5 iterations, excluding 16 substances and 31 reactions from the network that do not participate in any synthesis pathway of 3-MTHF. A representative set of excluded substances is provided in Table 6.4. The substances are excluded since they carry less oxygen than 3-MTHF or lack the molecular motifs to be refunctionalized to reach 3-MTHF.

Table 6.4 – Representative set of substances excluded through network reduction



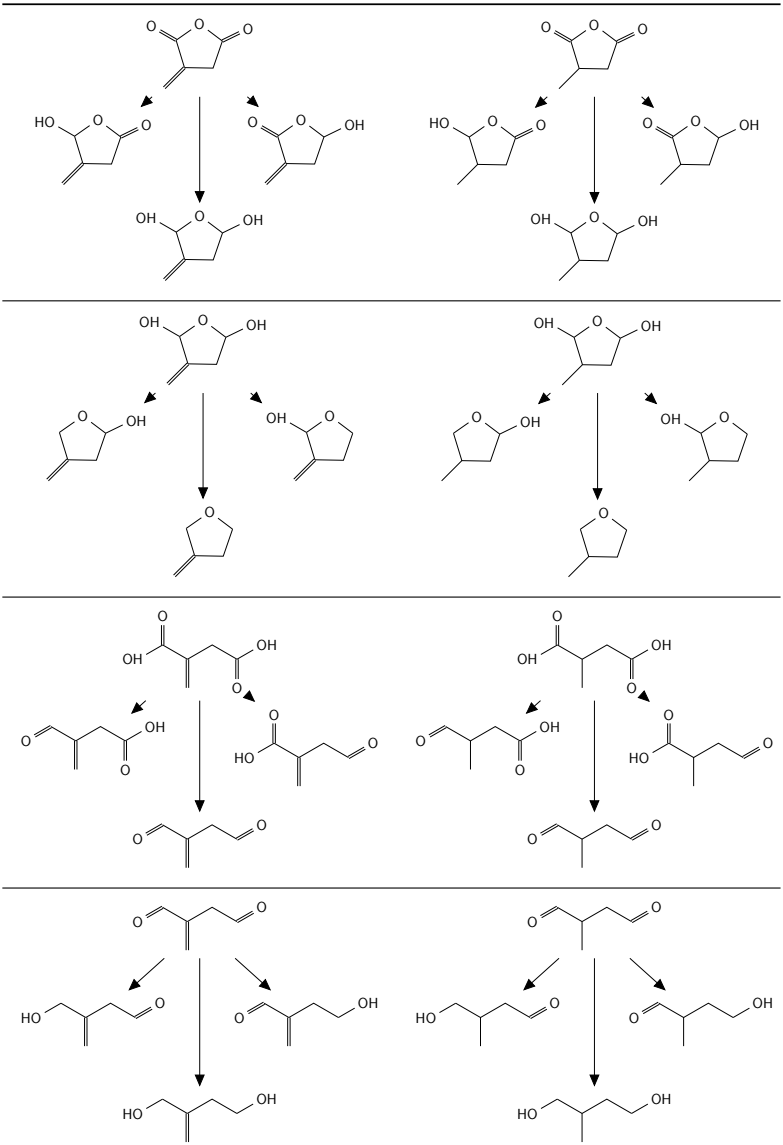
8 molecules carry a symmetric occurrence of molecular functionality. Table 6.5 presents these substances and the corresponding set of products.

8 additional reactions are included into the network, that represent reactions where multiple instances of a symmetric functionality are processed. The number of symmetric functionalities is two in every case, leading to non-selective formation of 3 products. Therefore the selectivity of each of these reactions is $\frac{1}{3}$. Every other reaction in the network is estimated to achieve full selectivity.

The resulting network is presented in Figure 6.4, in which substrate (IA) and product (3-MTHF) are highlighted. The graphical depiction of the molecules was generated by using the Open Source software Package Indigo Depict (Services, 2013), the graphical network representation was generated by using the Open Source software package Graphviz (Ellson et al., 2002).

The comparison of the manually assembled network (Figure 6.3) to the computationally generated one (Figure 6.4 and also attached in larger format to the back cover of this document) shows that a multitude of novel reactions towards a manifold of additional intermediates outside the literature scope are unraveled, bound into a complex interconnection of the network intermediates. Both, the number of substances and reactions is about one order of magnitude higher compared to the manually assembled network, leading to a variety of pathways to synthesize 3-MTHF from IA.

Table 6.5 – Visualization of non-selective reactions in the 3-MTHF synthesis network



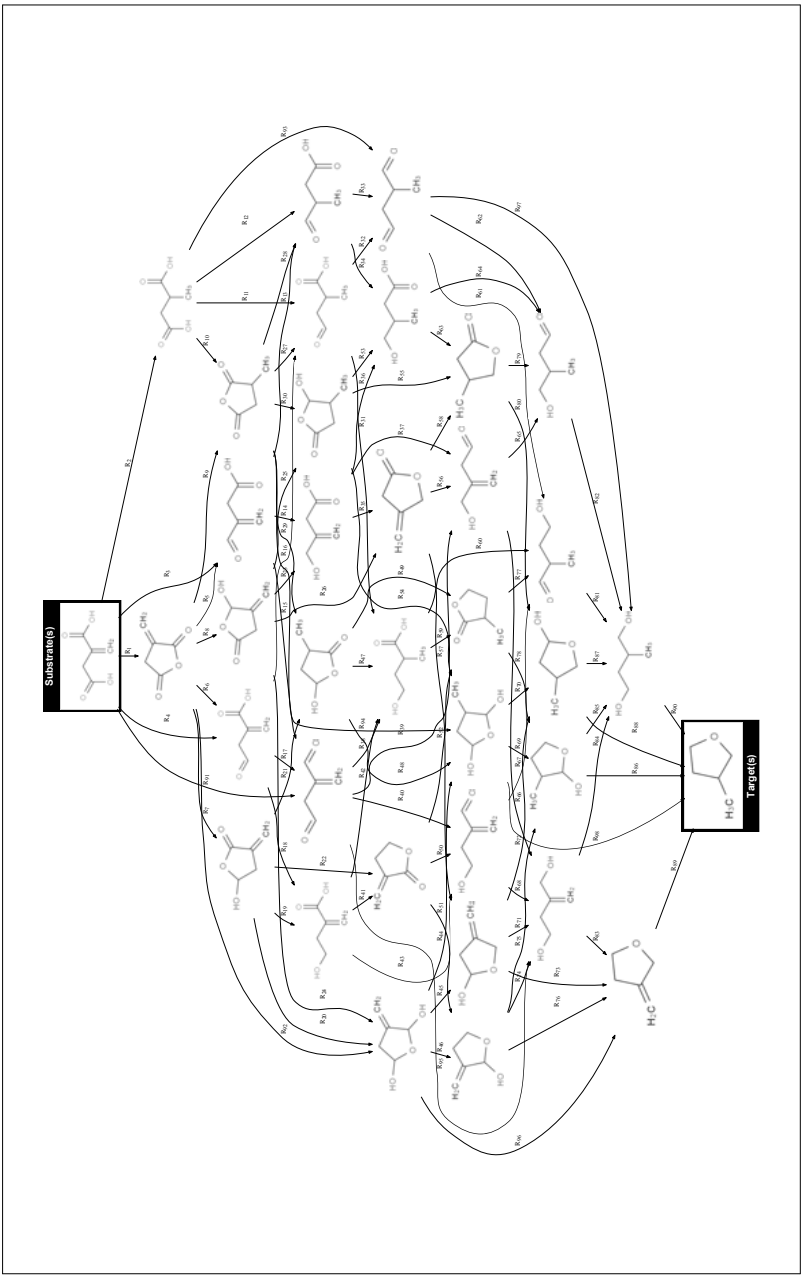


Figure 6.4 – Computationally generated reaction network for the synthesis of 3-MTHF from itaconic acid

6.3.3 Evaluation of the reaction network

Subsequently, the generated network is evaluated using the proposed multi-stage evaluation strategy. The computed results are compared against FAME production as reference process. Unlike to previous assessments, it is further elucidated whether the produced amount of 3-MTHF can be blended with streams of unconverted intermediates and whether efficiency gains arise thereof. Throughout the evaluation, process configurations with and without lignin gasification are considered.

The elementary mode analysis of the generated network leads to 388 available pathways to synthesize 3-MTHF from IA. The lengths of the pathways vary between 4 and 8 sequential reactions as presented left in Figure 6.5. Most elementary modes contain 7 sequential reactions.

Conferring to diagram on the right side of Figure 6.5, 60 elementary modes are sequences of fully selective reactions. More frequently encountered are elementary modes that contain 1 or 2 reactions which cannot be performed at full selectivity. Those elementary modes as stand-alone pathways only exhibit low feedstock conversion. An example of such a low performing reaction sequence is presented in Figure 6.6. Two reactions, each equipped with a selectivity $S_j = \frac{1}{3}$, occur within the sequence. To this end, the estimated selectivity of this elementary mode is only $\frac{1}{9}$.

The reactions that occur most frequently in the elementary modes are shown in Table 6.6, where also their absolute and relative number of occurrences is presented. Frequently encountered reactions are positioned at the very beginning (r_1 and r_2) and at the very end of the network (r_{89} and r_{90}). If either r_1 or r_{90} turn out to be infeasible, more than 60% of the elementary modes would not be performable. However, the importance of these reactions was already identified in literature, which is underlined by the fact, that experimental data on the reactions r_1, r_2, and r_{90} is already available.

Conversion data was gathered through an extensive literature research, using the web-based reaction library SciFinder (American Chemical Society, 2014). Experimentally determined conversion data was only available for 6 out of 98 reactions (cf. to Appendix F). Reactions without reported data were assumed to exhibit a conversion of 97%. Based on these values, the normalized flow rates of each elementary mode are determined.

Molecular weight (MW), normal boiling point (T_{boil}), enthalpy of combustion (ΔH_{com}), liquid density (ρ_L), melting point (T_{melt}), cetane number (CN) and median lethal dose (LD_{50}) are calculated for each substance, using the QSPR models reported by Dahmen et al. (2012). The estimated property data of the network substances is provided in Appendix F.

The normalized flow rates of each elementary mode are scaled by a user-defined value γ (representing the annual amount of produced fuel in mole/a), to receive absolute molar

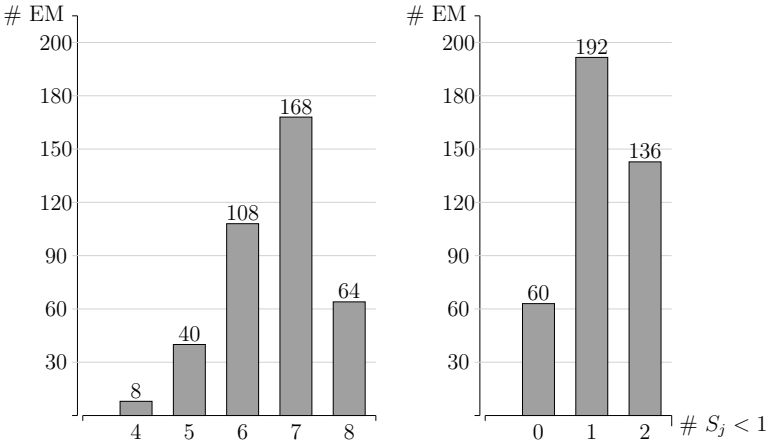


Figure 6.5 – Histograms of (left) length distribution of elementary modes and (right) number of non-selective reactions per elementary mode for 3-MTHF synthesis

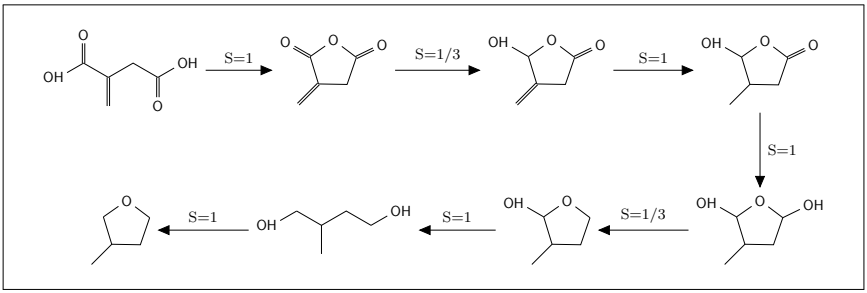
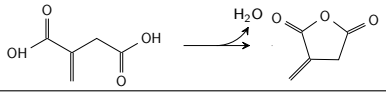
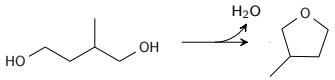
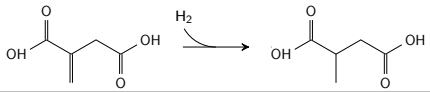
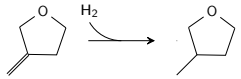


Figure 6.6 – Example of low performing elementary mode due to low selectivity in two reactions for 3-MTHF synthesis

Table 6.6 – Most frequent reactions in 3-MTHF synthesis, measured in absolute and relative amounts

Nr.	Reaction	Abs. occ.	Rel. occ.
1		245	63.1%
90		245	63.1%
2		60	15.5%
89		60	15.5%

flows. However, production capacities are usually not provided in mole/a but rather in l/a. Also in this case study, the annual production capacity is stated in l/a, denoted as γ_{vol}. This scaling has to be converted into the molar scaling γ. With knowledge of the liquid density $\rho_{L,T}$ and the molecular weight MW_T of the target, the molar amount of feedstock provided to the network γ is calculated from

$$\gamma = \gamma_{vol} \cdot \frac{\rho_{L,T}}{MW_T}. \tag{6.2}$$

Assessment of economic evaluation criteria requires information on the costs of feedstock and reactants. These costs are listed in Table 6.7, together with the assumed biofuel sales price.

Table 6.7 – Prices for feedstock, biofuel, reactant and auxiliary requirements

IA	Fuel	Hydrogen$_{ext}$	Hydrogen$_{int}$
\$/kg	\$/l	\$/kg	\$/kg
0.5 assumption	1.2 assumption	2.7 (Ruth, 2011)	0 assumption

The prices for IA and hydrogen are projected to an established fuel market in the future. Assuming a reaction that produces IA from glucose at full yield, current glucose market prices of about 350 - 400 \$/t (Alibaba.com, 2014a) will lead to IA prices of about 500 \$/t (without considering further processing costs as energy demand or auxiliary materials). However, it has to be stressed that current IA prices are about 2,000-3,000 \$/kg (Al-

ibaba.com, 2014b). Their inclusion into the evaluation would render the entire production pathways economically unfeasible.

TAC are set as objective function, such that the optimization problem results in

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \{TAC\} \\ s.t.\ \ &\mathbf{N} = \boldsymbol{\alpha} \cdot \mathbf{N}_m \\ &\mathbf{N}_m = \gamma_{mole} \cdot \mathbf{n}_m \\ &\sum_{m=1}^{388} \alpha_m = 1 \\ &\alpha_m \in [0,1]. \end{aligned} \tag{6.3}$$

A plant lifetime of 10 years and an interest rate of 8% are assumed. The TAC are then calculated as presented in Appendix E. Solving this optimization problem results in a linear combination of 9 equally contributing elementary modes that make up the optimal flux distribution.

6.3.4 Discussion of the evaluation results

The optimal flux distribution contains several non-selective reactions, depicted in Figure 6.7. The non-selective behavior is due to the symmetric occurrence of two carbonyl groups bound to a THF ring. These groups are first refunctionalized into hydroxy groups and subsequently completely removed. Each of the 9 elementary modes that are contained in the network achieves a selectivity of 1/9 regarding the conversion of IA into 3-MHTF, while the linear combination of these EM achieves full selectivity.

This particular flux distribution performs optimal for both process scenarios with and without lignin gasification. It includes 3 reactions with experimentally determined conversions that are higher than the default value of 97%. The flux distribution achieves a yield of 88% from IA to 3-MTHF. The energetic losses ΔE sum up to 6.44 MJ per kg 3-MTHF which equals an energetic efficiency (in terms of heat of combustion) of 83.7%. From these values, investment cost (IC) and total annualized cost (TAC) can be calculated according to Equations (E.8) and (E.9). The individual contributions to the financial structure are denoted in Table 6.8, distinguishing the process configurations with and without lignin gasification. The major contribution to the TAC is the cost of feedstock in both process configurations.

The identified flux distribution for the production of 3-MTHF from IA is hydrogen-intensive with a consumption of 4.56 mole H_2 per mole 3-MTHF, leading to high cost burdens for external hydrogen supply. In the configuration without lignin gasification,

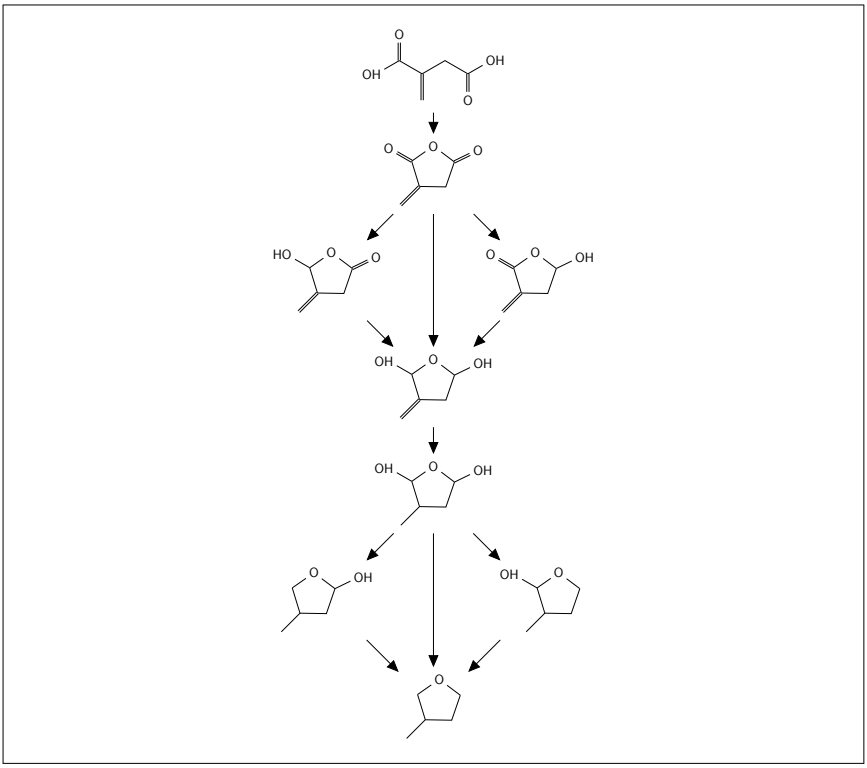


Figure 6.7 – Highest performing combination of pathways with TAC as objective function

the cost of hydrogen is the second largest contribution. However, in case of a process with lignin gasification, the annual loan repayment (ALR) is the second largest contribution and the cost of hydrogen is significantly decreased. The gasification of lignin produces enough hydrogen to reduce the requirement for external hydrogen to 0.99 mole H_2 per mole 3-MTHF. The required process for lignin gasification increases the overall investment cost and thereby also the annual loan repayment. However, the tradeoff between savings from internal hydrogen production and increasing ALR is in favor of the process configuration with lignin gasification, stated by the higher annual revenues. (cf. Table 6.8).

The economic viability of production processes strongly depends upon the market price of the feedstock. Table 6.9 presents the maximum allowable feedstock price p_{FS}^{max} for IA, calculated from Equation (E.11) presented in Appendix E. The process alternative with lignin gasification can cope with higher feedstock costs and is economically viable for IA

Table 6.8 – Cost structure and revenues for 3-MTHF production, with and without lignin gasification for hydrogen production

	Synthesis of 3-MTHF	
Lignin gasification	no	yes
	M$	M$
Investment cost	63.7	150.6
Total annual costs	197.7	171.0
Loan repayment	8.3	19.5
Feedstock	142.0	142.0
Reactants	47.5	9.5
Ext. hydrogen	47.5	9.5
Revenues	228.0	228.0
Total annual revenues	30.3	57.0

market prices up to 700 $/t. Although this is a significant higher value than the assumed value of 500 $/t, it is still far below current market prices of IA.

Table 6.9 – Maximum allowable feedstock (FS) price for 3-MTHF production process, with and without lignin gasification for hydrogen production

Target	Lignin gasif.	ALR	Reactants	Revenue	Cash flow for FS	Required FS	p_{FS}^{max}
–	–	M$/a	M$/a	M$/a	M$/a	kt/a	$/t
3-MTHF	no	8.3	47.5	228.0	172.2	283.9	606.3
	yes	19.5	9.5	228.0	199.1	283.9	701.2

The findings on the pathway performance without lignin gasification compare well to the values reported by Voll and Marquardt (2012a) on the synthesis of 3-MTHF from IA. They state a yield of 91.27% for their best performing pathway at a hydrogen demand of 4.79 mole/mole IA. Their energetic efficiency sums up to 86.5%, which is slightly higher than the value stated here, which results from higher reported yields of the reactions stated in their network. The optimal pathway requires three steps to reach 3-MTHF from IA without any branching of synthesis pathway. In contrast, the optimal synthesis proposed here comprises 6 steps and branching occurs twice due to non-selective reactions. It is obvious that some of the reactions stated by Voll and Marquardt (2012a) lump several reactions into a single one, which in contrast are resolved by ReNeGen as individual reaction steps. Hence it can be assumed that several individual reaction steps generated by ReNeGen can be merged into a single reaction step. Currently there is no general modeling framework available for model-based identification of such reaction compartments. Also, ReNeGen is

not capable of determining whether several reactions can be lumped; this information has to provided through experimental research and can then be included network generation and evaluation process (e.g. by evenly distributing the non-ideality of the lumped reaction on the individual reactions provided by ReNeGen). Investment cost are not directly comparable since Voll and Marquardt (2012a) calculated IC for the year 1993, while the present contribution used updated values for the year 2013. However, since IC is a function of energetic loss, similar values will result from both contributions.

6.3.5 Comparison against the reference process

Figure 6.8 graphically depicts the comparison of the financial structures of four scenarios of 3-MTHF production against the reference process. The scenarios represent process configurations with and without lignin gasification and incorporate real and ideal conversions ($C_j = 1$ for any reaction). The scenarios with ideal conversion are depicted to show the emerging potential from improving the performance of the reactions. The constituents of the TAC of a production scenario are

(i) the annual loan repayments (ALR),

(ii) the costs of feedstock supply (Feed) and

(iii) the costs of reactants (Rea.).

They are compared against the total annualized revenues (TAR) that result from selling the annually produced amount of fuel (190 Ml/a) at the specified sales price of 1.2 \$/l. It needs to be stressed that results are calculated using the metrics in Appendix E, which are by their empirical character inaccurate in their prediction.

All scenarios for 3-MTHF production from IA are economically viable, since the TAC are lower than the TAR. The scenarios that assume ideal conversion show that, based on the employed evaluation metrics, the TAC can still be decreased by 15% by a more efficient conversion of the feedstock (cf. Figure 6.8). Similar to the real cases, the ideal process configuration with lignin gasification is economically more promising than the process configuration without.

The scenarios for 3-MTHF production economically outperform the reference process through lower expenses on feedstock supply. However, the assumed price for rapeseed oil is twice as high as for IA. Taking the current market price of IA, the prices for feedstock would be 4-6 times higher in the 3-MTHF production scenarios, ending in cost structures that are far off from being competitive.

Costs for reactants are higher for processes without lignin gasification compared to FAME

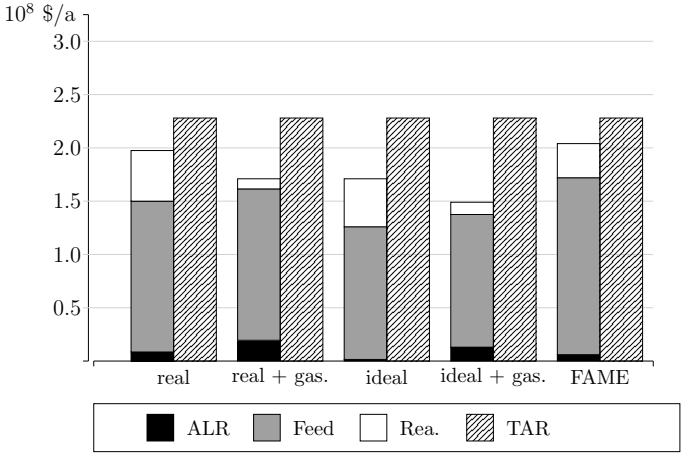


Figure 6.8 – Financial structure of FAME production and of real and ideal process configurations for the synthesis of 3-MTHF from IA, considering processes with and without lignin gasification

Table 6.10 – Environmental aspects of 3-MTHF production

Main product	Gas.	RC	EC	Em	TP	EI
3-MTHF	no	0.77	2.26	1.00	1.00	5.03
	yes	0.77	2.26	–	1.00	–

production. They can be reduced by providing hydrogen from lignin gasification, in which case they are less than for FAME production.

Annual loan repayments for the investment cost (IC) are rather low for both, the production scenarios for 3-MTHF and FAME. The IC are calculated based on the energetic losses of the processes (cf. Appendix E). Based on the employed evaluation criteria, IC are calculated to be smaller for FAME production than for 3-MTHF production. However, ALR only plays a minor role in the economic performance of any process configuration.

Table 6.10 denotes the values for resource consumption (RC), energy consumption (EC), emissions indicator (Em) and toxicity potential (TP) of 3-MTHF production related to the values of the reference process. They sum up to the environmental impact (EI) of the process, as described in (Appendix E).

Both production scenarios exhibit worse environmental impacts than the reference process due to higher energy consumption (EC) resulting from lower yields. The resource consumption (RC) of the 3-MTHF processes are lower, since the FAME process leads to

the formation of glycerol as a side product. The emissions potential (Em) of the process configuration without lignin gasification is equal to the reference case, since both processes do not lead to the formation of substances containing a global warming potential. However, for the production scenario with lignin gasification, The Em value cannot be calculated as the corresponding value of FAME-production is 0 and goes as denominator into the equation. The amount of CO_2 that is produced per unit of 3-MTHF results to 2.45 kg CO_2 per kg 3-MTHF. The toxicity potentials (TP) of both process configurations are equal to the TP of FAME production, since both fuels, 3-MTHF and FAME, exhibit similar LD_{50} values. Although the emission potential of 3-MTHF production with lignin gasification cannot be put in relation to FAME production, it can be stated that all proposed scenarios exhibit a higher environmental impact than the reference process, leading to the conclusion, that the production of 3-MTHF is less environmentally benign than that of FAME.

6.3.6 Integration of intermediate waste streams

Due to the high sensitivity of the process economics on the feedstock, the supplied feedstock should be exploited to the highest extent possible. To this end, the integration of waste streams, as presented in Section 5.2.4, is employed to elucidate the potential of blending 3-MTHF with unconverted intermediates. The final mixture still has to meet the property constraints that are stated in Table 6.1.

The mixture resulting from waste stream integration increases the feedstock utilization to 97% (compared to 88% in the non-integrated scenario). It is composed of 4 substances, whose properties are listed in Table 6.11 along with the properties of the mixture. The molar fraction of each substance is denoted by $\boldsymbol{y}_i$. The major part of the mixture (approx. 90%) still consists of 3-MTHF. ΔH_{com} sets the tightest constraint to the formation of the mixture; it is the property that reaches its upper bound. The other property constraints still would allow for further integration, although CN is already very close to its lower bound.

It has to be stressed that the predicted values, especially those for CN, have to be assessed critically. The CN data set for model development is by far the smallest and contains substances that are commonly perceived as fuels. The integrated substances have never been measured concerning their CN, thus they differ significantly from the molecules in the training data. However, the calculated negative values do not contradict with the meaning of cetane numbers. The CN is a relative value related to the auto-ignition behavior of two reference substances, such that negative CN values represent an auto-ignition behavior below the lower reference value.

The overall volumetric output (cf. Table 6.12) of the integrated process results to almost

Table 6.11 – Properties of individual waste streams and of fuel mixture

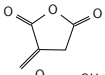
Structure	y_i	MW	ΔH_{com}	T_{boil}	ρ_l	CN	T_{melt}
–	–	kg/kmol	MJ/kg	K	kg/l	–	K
	0.08	116.13	-20.71	503.12	1.26	-63.35	377.91
	0.01	102.15	-25.67	436.72	1.03	-11.96	263.50
	0.01	102.15	-25.49	445.24	1.01	-17.10	258.92
	0.90	86.15	-33.16	373.09	0.87	40.09	161.97
mixture	1	88.89	-32.00	384.97	0.90	30.62	181.45

Table 6.12 – Comparison of the revenues of 3-MTHF production process with and without integration of waste streams

Pure product		Mixture		Gain
$\dot{V}_{fuel}$	Revenues	$\dot{V}_{fuel}$	Revenues	
Ml/a	M$/a	Ml/a	M$/a	%
190.0	228.0	209.7	251.6	10.4

210 Ml/a, which equals a relative gain of more than 10%. The revenues from fuel sales increase proportionally to the increase in volumetric output. Therefore, integrated process configurations significantly increase the annualized revenues compared to the production of pure 3-MTHF.

The higher revenues from fuel sale allow for economic operation even if the prices for feedstock are higher. The maximum allowable feedstock costs for the integrated process configurations with and without lignin gasification are presented in Table 6.13. In the case of an integrated production process with lignin gasification, the highest allowable feedstock price is 784 $/t. Although this leads to a further increase of more than 10% compared to the non-integrated scenario, the value is still far below current market prices of IA.

The presented case study was executed in 101.3 CPU sec. Reaction network generation including the selective estimation required 72.1 CPU sec. and the multi-stage evaluation from elementary mode analysis to integration of intermediate waste streams 29.2 CPU sec. on a PC equipped with an Intel Core i5 Quadcore CPU @ 3.2 GHz and 8 GB RAM.

Table 6.13 – Maximum allowable feedstock (FS) prices for integrated 2-MTHF production processes, with and without lignin gasification for hydrogen production

Target	Lignin gasif.	ALR	Reactants	Revenue	Cash flow for FS	Required FS	p_{FS}^{max}
–	–	M$/a	M$/a	M$/a	M$/a	kt/a	$/t
3-MTHF	no	8.3	47.5	251.6	195.8	283.9	690
	yes	19.5	9.5	251.6	222.6	283.9	784

6.4 Synthesis of 2-BF and 2-BTHF from furfural

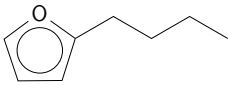
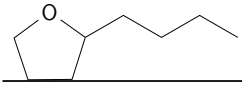
Although 3-MTHF is a suitable fuel for use in CI engines, economic drawbacks result from the high price of its feedstock IA. Therefore, a second case study is presented to synthesize substances that are structurally similar to 3-MTHF, but can be produced from a feedstock that is available at lower market price.

Julis and Leitner (2012) presented several substances that can be derived from furfural, which either could serve as intermediates in biofuel synthesis pathways (e.g. towards 1-octanol) or qualify as biofuels themselves. Furfural is available at a market price of about 1,000-1,200 $/t (Alibaba.com, 2014c), which is significantly below that of IA. It has been highlighted as a key building block of a bioeconomy, expecting tremendous potential in its use as platform chemical (Lange et al., 2012, Cai et al., 2014). De Jong and Marcotullio (2010) performed a conservative estimation of furfural production in a biorefinery and estimated a market price of approximately 1,000 $/t. They assume an acidic biomass hydrolysis with subsequent dehydration of the derived C_5 sugars. The annual production of furfural is approximately 10,000 t. The process economics strongly rely upon the biomass market price. However, is is expected that significant improvements can be achieved through further investigation of the reaction paths and by integrating reaction and product separation.

2-BF and 2-BTHF, which are two of the substances presented by Julis and Leitner (2012), can be derived from furfural, contain a heterocyclic motif and fulfill the imposed property constraints. They are presented in Table 6.14, along with their property values estimated by using the QSPR models presented in Appendix D.

Visual investigation reveals that the two substances are structurally very similar to 3-MTHF, they only differ in length and positioning of the alkyl side chain and the degree of unsaturation of the heterocycle in the case of 2-BF. 2-BF and 2-BTHF themselves are structurally akin; while 2-BF contains a furan ring, 2-BTHF contains the fully hydrogenated derivative of this heterocycle. In both cases, the butyl side chains are located at the same position. The predicted property values well exceed the imposed requirements which allows for integrating larger amounts of unconverted intermediates into the final

Table 6.14 – Property data of 2-BF and 2-BTHF, predicted by using the QSPR models presented in Appendix D

Structure	T_{boil}	ρ_l	ΔH_{com}	T_{melt}	CN
-	K	kg/l	MJ/kg	K	–
	421.2	0.9	-35.8	176.2	35.8
	436.2	0.85	-37.2	170.2	86.8

product without violating the property constraints.

6.4.1 Reaction network generation for 2-BF and 2-BTHF synthesis

The computational generation of the reaction network was performed according to the settings in Table 6.15.

Table 6.15 – Scenario definition for 2-BF and 2-BTHF synthesis from furfural

Setting	Value
Substrate	Furfural
Targets	2-BF, 2-BTHF
Number of stages	15
Reactants	H_2, Acetone
Reaction rules	1-7, 11-12
Main product criterion	ΔH_{com}
Biomass composition	0.6:0.2:0.2

Furfural contains less carbon atoms than 2-BF and 2-BTHF. To increase the carbon content, acetone is provided as a reactant which can serve as means to increase the length of the alkyl side chain. Acetone was manually chosen; it is the reactant required for a aldol condensation, a reaction capable of establishing novel carbon-carbon bonds to prolong for instance carbon side chains, which is required in this case. Constraints on structure and composition of the network substances are presented in Table 6.16. The major difference to the settings of 3-MTHF synthesis is the extended carbon range in the elementary composition of the intermediates. It has to be extended to represent the synthesis of a C_8 fuel from a C_5 feedstock.

Table 6.16 – Structural constraints on reaction network intermediates in 2-BF and 2-BTHF synthesis from furfural

Structural feature	Setting
Elementary composition	$C_{5-8}H_{0-16}O_{0-3}$
Number of rings	0-1
Ring size	4-6

The network generation is performed individually for the combinations of furfural as feedstock and 2-BF and 2-BTHF as targets respectively, resulting in two distinct networks. These networks are merged according to the operations presented in Section 4.1, to give a combined network of both substrate-target combinations. This combined network is presented in Figure 6.9 (and also attached in larger format to the back cover of this document), highlighting the feedstock furfural and the products 2-BF and 2-BTHF. It can be seen that the synthesis pathways towards 2-BF and 2-BTHF share several intermediates, due to the structural similarity of both substances. This combined network contains 60 reactions and 28 intermediates.

In total, 93 elementary modes link furfural with the targets (cf. Figure 6.10, left). 90 elementary modes lead to the production of 2-BTHF, while only 3 lead towards 2-BF. The pathways towards 2-BTHF all comprise 6 sequential steps, while those towards 2-BF comprise 4 sequential steps, rendering 2-BF synthesis slightly favorable due to less processing steps. None of the intermediates contains any of the distinguished functional groups in a symmetric arrangement. Therefore, all 93 elementary modes are estimated to contain only reactions that perform at full selectivity (cf. Figure 6.10, right).

Elementary mode analysis indicates that the production of 2-BTHF is more robust since more pathways are available. The most frequent reactions in these elementary modes for both products are presented in Table 6.17. Reaction r_3 is a key reaction in both synthesis tasks. In case that r_3 cannot be performed, two thirds of the synthesis pathways for 2-BTHF production will not be performable. Even more severe is the effect on the synthesis of 2-BF. Every pathway for the production of 2-BF incorporates r_3, meaning that 2-BF cannot be produced if r_3 cannot be performed. Fortunately, experimental data are available in literature that demonstrate the feasibility of r_3 (Alvarez-Ibarra et al., 1992) at a conversion of 95%. However, subsequent experimental investigations should be carried out to improve the performance and confirm the applicability in processes beyond lab scale.

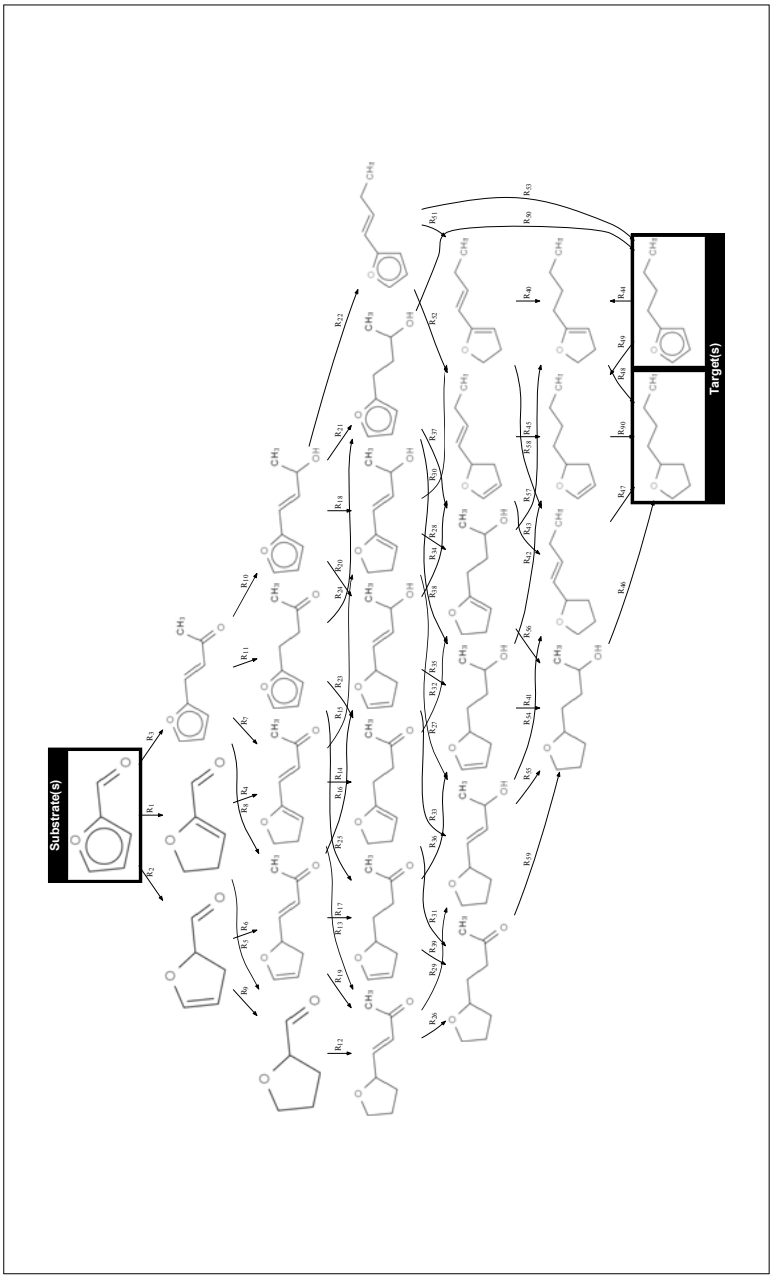


Figure 6.9 – Merged reaction network for the synthesis of 2-BF and 2-BTHF from furfural

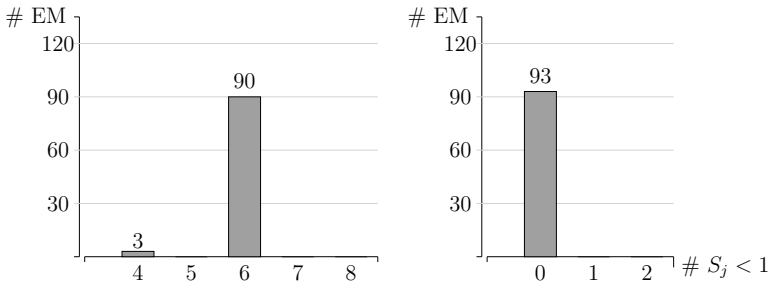
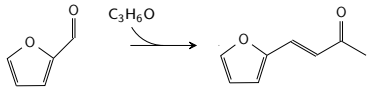
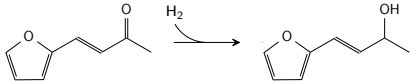
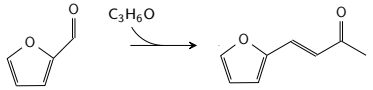
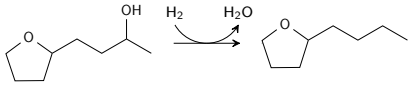


Figure 6.10 – Histograms of length distribution of elementary modes (left) and number of non-selective reactions per elementary mode (right) for the synthesis pathways from furfural to 2-BF and 2-BTHF

Table 6.17 – Most frequent reactions in 2-BF (r_3 and r_{10}) and 2-BTHF synthesis (r_3 and r_{57}), measured in absolute and relative amounts

r_j	Scheme	Abs. occ.	Rel. occ.
r_3		3	100%
r_{10}		2	66.67%
r_3		60	66.67%
r_{57}		40	44.44%

C_3H_6O: Acetone

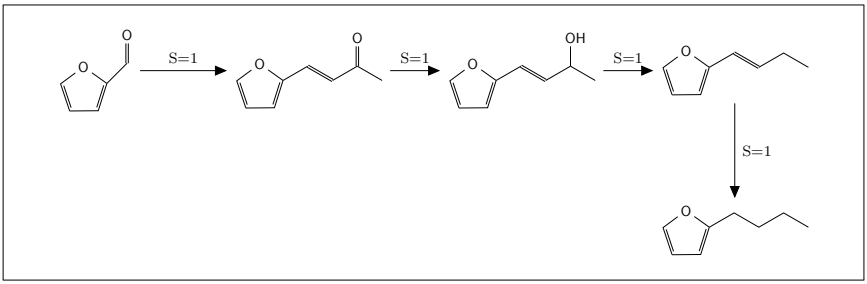


Figure 6.11 – Best performing pathway for the synthesis of 2-BF from furfural

6.4.2 Evaluation of the reaction network

Experimentally determined conversion data were available for 6 reactions. Every other reaction was assumed to achieve a conversion of 97%. As in the previous case study, the optimization problem was set up to minimize the TAC. Identical costs for feedstock, reactant and biofuel were assumed to evaluate the two case studies on common basis and allow for a direct comparison. The price of acetone is set to 1,000 \$/t, in accordance to the value reported by Beale et al. (2008). The production of each substance is considered with and without lignin gasification and each combination of substrate and target is considered individually.

6.4.3 Discussion of the evaluation results

The optimal flux distribution for the production of 2-BF is a single elementary mode, depicted in Figure 6.11. 2-BF is produced by a series of 4 sequential reactions, where each reaction achieves full selectivity.

The optimal pathway for the production of 2-BTHF also consists of a single elementary mode, comprising 6 sequential reactions that perform at full selectivity. It is depicted in Figure 6.12.

The cost structures of the identified pathways towards 2-BF and 2-BTHF are shown in Table 6.18, both considering scenarios with and without lignin gasification. The values indicate that the process configurations without lignin gasification suffer from high costs for reactants, while the costs for feedstock are reduced compared to 3-MTHF production. The inclusion of lignin gasification does not significantly reduce the costs of the reactants, although the demand for external hydrogen vanishes for the production of 2-BF. The high costs for reactants result from the additional demand for acetone. However, the difference between process configurations with and without lignin gasification is lower, especially for

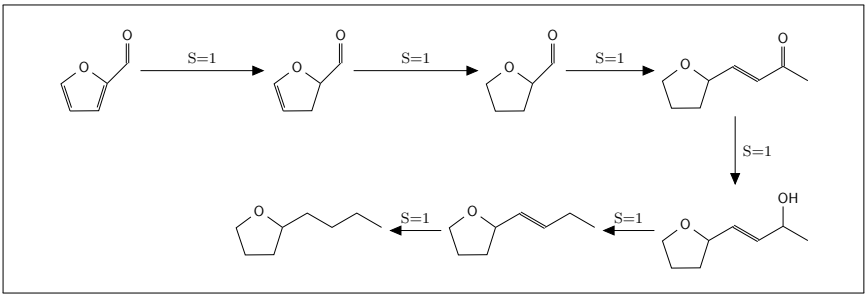


Figure 6.12 – Best performing pathway for the synthesis of 2-BTHF from furfural

the production of 2-BTHF, where the additional ALR equals the generated savings from internal hydrogen provision.

Table 6.18 – Financial structure of the optimal pathway for 2-BF and 2-BTHF production from furfural

	Synthesis of 2-BF		Synthesis of 2-BTHF	
Lignin gasification	no	yes	no	yes
	M$	M$	M$	M$
Investment cost	74.7	136.4	92.3	148.9
Total annual costs	181.0	174.8	180.1	180.1
Loan repayment	9.7	17.7	12.0	19.3
Feedstock	74.4	74.4	70.0	70.0
Reactants	94.2	82.8	98.2	90.8
Ext. hydrogen	11.4	0.0	17.8	10.4
Acetone	82.8	82.8	80.4	80.4
Revenues	228.0	228.0	228.0	228.0
Total annual revenues	47.0	53.1	47.9	47.9

The maximum allowable costs for feedstock for both, 2-BF and 2-BTHF synthesis, are presented in Table 6.19. Every process configuration allows for maximum feedstock prices that are higher than those of process configuration in case of 3-MTHF production. However, the highest allowable feedstock prices p_{FS}^{max} are still below the market price of furfural.

Table 6.19 – Maximum allowable feedstock (FS) prices for 2-BF and 2-BTHF production processes, with and without lignin gasification for hydrogen production

Target	Lignin gasif.	ALR	Reactants	Revenue	Cash flow for FS	Required FS	p_{FS}^{max}
–	–	M$/a	M$/a	M$/a	M$/a	kt/a	$/t
2-BF	no	9.7	96.9	228.0	121.4	148.8	816
	yes	17.7	85.4	228.0	124.9	148.8	839
2-BTHF	no	12.0	98.2	228.0	117.8	140.0	841
	yes	19.3	90.1	228.0	118.6	140.0	847

6.4.4 Comparison against the reference process

A graphical comparison of the cost structure of four process configurations (with and without lignin gasification, real and ideal conversions) against the reference process is depicted in Figure 6.13 for 2-BF and in Figure 6.14 for 2-BTHF, respectively.

The two production alternatives exhibit very similar economic behavior. The costs of feedstock are significantly lower for 2-BF and 2-BTHF production than for FAME production. High costs for reactants have a major impact on the TAC and are multiple times higher than for FAME production. The ideal scenarios show that further improvement of the reaction conversion can reduce the TAC by an additional 10%. It is interesting to notice, that the production process of 2-BTHF without lignin gasification under the assumption of ideal conversion performs better than the process with lignin gasification. This shows, that it is not always beneficial to replace external hydrogen by hydrogen from lignin gasification.

The environmental impacts (cf. Table 6.20) of 2-BF and 2-BTHF production without lignin gasification are, in total, below those of FAME production. Both production alternatives without lignin gasification exhibit significantly lower values of energy consumption. In contrast to 2-BTHF production, 2-BF production shows a more efficient use of the provided resources (RC). The amount of emissions of the process configurations without lignin gasification are equal to the value of FAME production. Due to the same reason as in the 3-MTHF case study, a quantitative value for the emissions indicator cannot be established for the processes with lignin gasification. However, it can be stated that the emissions of CO_2 per kg of product are 1.41 kg/kg for 2-BF production and 2.01 kg/kg for 2-BTHF production, which is less than for 3-MTHF. All 4 process configurations have the same toxicity potential as FAME. In total, the production of 2-BF in a process without lignin gasification is estimated to be most environmentally favorable. Although not achieving the same low value as 2-BF, 2-BTHF production without lignin gasification is also environmentally more benign than the FAME process.

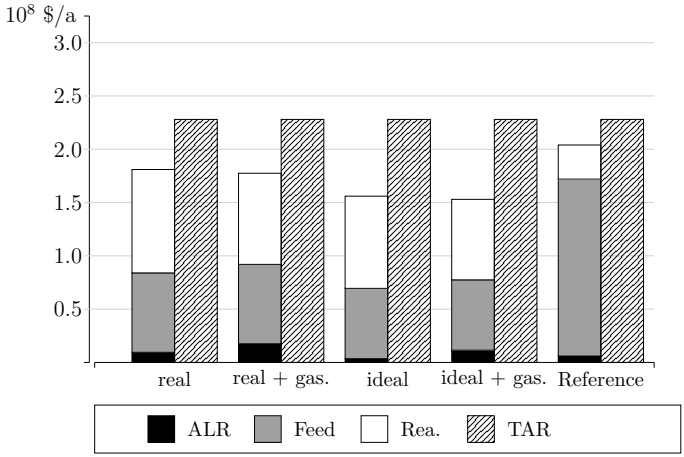


Figure 6.13 – Financial structure of the four scenarios of 2-BF production

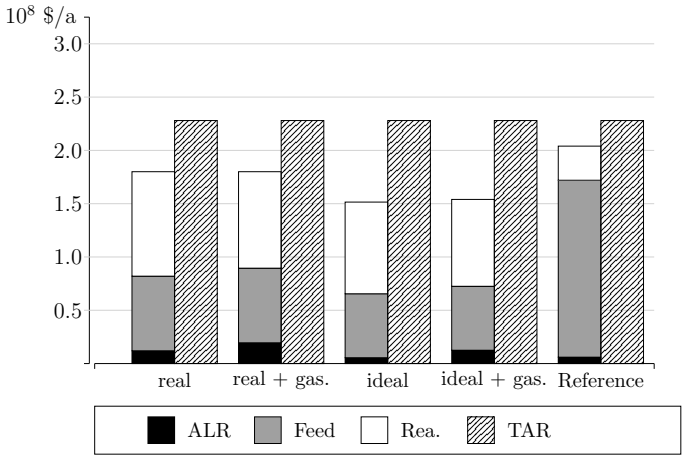


Figure 6.14 – Financial structure of the four scenarios of 2-BTHF production

Table 6.20 – Environmental aspects of 2-BF and 2-BTHF production

Target	Lignin gasif.	RC	EC	Em	TP	EI
2-BF	no	0.91	0.62	1.00	1.00	3.53
	yes	0.91	0.62	–	1.00	–
2-BTHF	no	1.24	0.59	1.00	1.00	3.83
	yes	1.24	0.59	–	1.00	–

6.4.5 Integration of intermediate waste streams

The identified production routes for 2-BF and 2-BTHF can be improved by integrating the streams of unconverted intermediates. Since the property data of 2-BF and 2-BTHF shows a significant gap between estimated and specified property data, it can be expected that the product streams of both substances have a high potential for integrating unconverted intermediates.

Both production alternatives can include all occurring streams of unconverted intermediates and exhibit a feedstock utilization of 100%. Every occurring flux of unconverted intermediates is integrated into the final product. Tables 6.21 and 6.22 present the composition and properties of the resulting mixtures.

Integration of waste streams results in a five and six component mixture in the case of 2-BF and 2-BTHF, respectively. The constrained thermophysical properties are within the allowed range, rendering both mixtures feasible for use in CI engines. Also interesting to look at is the economic performance of the processes. Since the processes themselves are not changed, they exhibit the same TAC as presented earlier. However, the revenues from fuel sales increase, since the integration of waste streams leads to a significant increase in volumetric fuel output. The revenues of the integrated production processes are listed and compared to the non-integrated processes in Table 6.23. The numbers reveal that the increase in volumetric fuel output leads to a significant increase in the revenues generated from fuel sale.

These increases in revenues lead to the maximum feedstock prices p_{FS}^{max} for the integrated scenarios presented in Table 6.24. It turns out, that the scenario of 2-BTHF production with lignin gasification allows for the highest feedstock price, which is 1,246 \$/t. Also, the process configuration without lignin gasification could operate at the current market prices of furfural. The integrated production processes of 2-BF can also cope with furfural prices higher than 1,000 \$/t, but they do not reach the high values of the integrated production of 2-BTHF, due to the lower increase in volumetric fuel output.

The presented case study was executed in 87.6 CPU sec., where 59.1 CPU sec. were

Table 6.21 – Properties and components of the mixture of 2-BF and unconverted intermediates

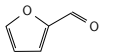
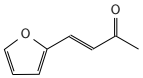
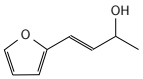
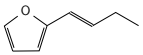
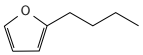
Structure	y_i	MW	ΔH_{com}	T_{boil}	ρ_l	CN	T_{melt}
–	–	kg/kmol	MJ/kg	K	kg/l	–	K
	0.05	96.1	-32.3	413.6	1.13	-12.9	222.5
	0.01	136.2	-28.4	473.1	1.04	-1.7	227.2
	0.03	138.2	-29.4	488.8	1.05	-21.4	219.2
	0.02	122.2	-35.3	421.9	0.91	22.2	168.7
	0.89	124.2	-35.8	421.7	0.90	35.8	175.8
mixture	1	123.2	-34.9	423.6	0.91	31	179.6

Table 6.22 – Properties and components of the mixture of 2-BTHF and unconverted intermediates

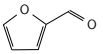
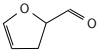
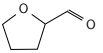
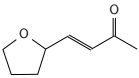
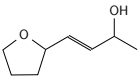
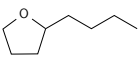
Structure	y_i	MW	ΔH_{com}	T_{boil}	ρ_l	CN	T_{melt}
–	–	kg/kmol	MJ/kg	K	kg/m^3	–	K
	0.03	96.1	-23.3	413.6	1.14	-12.9	222.5
	0.03	98.1	-25.0	429.2	1.09	31.5	238.5
	0.03	100.1	-25.5	433.9	1.06	55.4	198.9
	0.03	140.2	-30.0	488.9	0.98	38.1	205.6
	0.03	142.2	-31.1	500.6	0.99	16.5	195.3
	0.85	128.2	-37.2	435.9	0.84	86.8	170.0
mixture	1	126.3	-35.8	438.2	0.87	78.1	176.0

Table 6.23 – Comparison of the economic performance of 2-BF and 2-BTHF production processes, with and without waste stream integration

Target	Pure product		Mixture		Gain
-	$\dot{V}_{fuel}$	Revenues	$\dot{V}_{fuel}$	Revenues	
-	Ml/a	M$/a	Ml/a	M$/a	%
2-BF	190.0	228.0	217.0	260.4	14.2
2-BTHF	190.0	228.0	224.5	269.4	18.2

Table 6.24 – Maximum allowable feedstock prices for integrated 2-BF and 2-BTHF production processes, with and without lignin gasification for hydrogen production

Target	Lignin gasif.	ALR	Reactants	Revenue	Cash flow for FS	Required FS	p_{FS}^{max}
–	–	M$/a	M$/a	M$/a	M$/a	kt/a	$/t
2-BF	no	9.7	96.9	260.4	153.8	148.8	1,034
	yes	17.7	85.4	260.4	157.3	148.8	1,057
2-BTHF	no	12.0	86.7	269.4	170.7	140.0	1,219
	yes	19.3	75.6	269.4	174.5	140.0	1,246

required for the generation of the reaction network and 28.5 CPU sec. for the multi-stage evaluation from elementary mode analysis to integration of intermediate waste streams.

6.5 Conclusions

Two case studies were presented in this chapter. The first case study targeted the production of 3-MTHF from IA. This process was already investigated in the Cluster of Excellence TMFB and serves as a reference to demonstrate the benefits of automated reaction network generation and the novel network evaluation strategy.

Automatic reaction network generation revealed a multitude of reactions and intermediates that were so far not considered by the previous studies (Voll and Marquardt, 2012a,b). Elementary mode analysis showed that a large number of production alternatives is available, although many of them suffer from low estimated selectivity values. The most important reactions in the network were identified based on elementary mode analysis; these findings have to be taken into consideration in any further investigations on the production of 3-MTHF. The production of 3-MTHF is economically viable if feedstock prices of an established market are assumed. However, this assumption is highly questionable, since current prices for IA are multiple times higher. Furthermore, the identified 3-MTHF pro-

duction pathway is estimated to be environmentally less benign than the reference process.

These two aspects led to the decision to perform a second case study, targeting substances that are structurally similar to 3-MTHF, but start from a different feedstock. Furfural was chosen as feedstock based on the current interest in literature, its suitability as platform chemical and the lower market prices compared to IA. Reaction networks were automatically generated towards both substances, starting from furfural.

The production of 2-BTHF is the more robust alternative, since more pathways are available. Besides, the most crucial reactions for both synthesis tasks are already known and quantified experimentally. The evaluation revealed that economically viable production processes can be established towards the synthesis of both substances. In contrast to 3-MTHF synthesis, the main cost factor is no longer the feedstock, but the required reactants. However, feedstock prices still act as significant cost factor. Lignin gasification has only a smaller effect on the economics of the process than it is the case in 3-MTHF production. The environmental impacts of the identified processes are lower than for the reference process, due to a higher efficiency in the utilization of the provided resources.

It turned out that the feedstock utilization for both, 2-BF and 2-BTHF production, can be increased to 100% by incorporating the occurring flows of unconverted intermediates. Due to the increased volumetric output, the production of 2-BTHF can cope with feedstock prices of more than 1,200 \$/t, which is already above current market prices of furfural. Although the results are promising, three aspects have to be stressed: (i) not all cost aspects are considered in the presented assessment, (ii) the impact of the increased demand for furfural on its market price is not accounted for and (iii) the employed models comprise uncertainties that impede the accuracy of the analysis, thus the results demand for experimental and real life verification. However, candidate production scenarios can be proposed, assessed and evaluated only based on very early knowledge of the reaction pathways. The derived results should be used to guide further research and serve as basis of subsequent decisions.

Based on the presented findings, 2-BF and 2-BTHF are highly interesting biofuel candidates that should be the topic of more detailed investigations. According to the modeling results they both exhibit desired thermophysical properties and can be produced in an economically promising and ecologically benign way. The generated networks and the identification of the most important reactions serve as a sound basis to guide experimental research. The integration of waste streams in particular has to be checked for feasibility, since the economic potentials of the processes considerably increases by producing a mixture instead of a pure substance as a biofuel.

The case studies showed that the required time to construct and evaluate networks is reduced considerably by using automated reaction network generation. While manual assembly of the reaction network takes hours to days, the automatic reaction network

generation only required several seconds.

The second case study specifically demonstrated that ReNeGen and the novel multi-stage evaluation strategy in conjunction with a product design approach is a very effective tool for identifying, evaluating and proposing application-tailored biofuels and their corresponding production processes.

7 Conclusions and outlook

This contribution presents a computational, model-based approach for generating and evaluating synthesis networks towards novel biofuels. A reaction network generator was developed that is capable of generating synthesis networks solely based on fundamentals of chemistry. In comparison to the only previous approach towards formal reaction network generation by Fontain and Reitsam (1991), the complexity is broken down into sub-problems, expected to be computationally less complex than the initial formulation of formal network generation, implemented in RAIN. Hydrogen atoms are neglected in the combinatorial part of the generation process which significantly reduces the size of the reaction network generation task. This approach requires an entire novel formulation of the generation process that builds on introducing information on the covalent bonding of single atoms into atomic adjacency schemes. Formerly excluded hydrogen atoms are then either used to equilibrate formal electric charges or form molecular hydrogen. Post-processing routines ensure the uniqueness of the generated reaction products and identify those substances that will be processed in subsequent network stages. This comprehensive and efficient approach to formal network generation can be accompanied by user-provided empirical knowledge, which, in a top-down manner, targets the formal network generation process into a desired direction, but is neither a prerequisite nor mandatory for the reaction network generation itself.

The formal formulation was chosen to meet the needs of the Cluster of Excellence "Tailor-Made Fuels from Biomass" (TMFB). In TMFB, novel fuel candidates and their corresponding synthesis pathways shall be proposed. Inevitably, this leads to situations where only little or no information is available in literature about an envisioned design task. So far, this posed a major challenge to the previous TMFB approach of constructing reaction networks based on literature data. Marvin et al. (2013) showed how to automatically generate reaction networks towards biofuel candidates. However, their generator has to rely on a manifold of assumptions to generate a network of meaningful size. To be not reliant on assumptions and to use the available information to the maximum extent possible without diluting it amongst assumptions, the formal formulation of ReNeGen with its capability to include empirical knowledge to a user-defined degree constitutes the best alternative to cope with TMFB requirements.

The molecular constitution of the substances in the network is investigated to provide

an estimate of the selectiveness of the network reactions. Up to now, this approach has not been reported in literature, neither in computational chemistry nor in computational reaction network generation.

Elementary mode analysis forms the basis of a novel, multi-stage evaluation methodology for reaction networks, that extends recent approaches to retrieving information on the robustness of synthesis decision and importance of individual reaction from the network's topology. The derived elementary modes are the basis for determining an optimal flux distribution. Furthermore, they allow for statements about the robustness of synthesis networks and the importance of individual reactions. As a concluding step, the integration of waste streams into the final product is proposed as a means to increase the efficiency of the derived process configurations.

The thesis is concluded by the application of the presented methodologies to two distinct case studies. The first case study, the synthesis of 3-MTHF from itaconic acid, serves as a reference to compare manual and automated reaction network generation. It turned out that automated reaction network generation provides reaction networks that are one order of magnitude larger than manual assembled ones. The economic assessment revealed that the process suffers from the high market prices of itaconic acid.

The second case targets the identification of biofuels that are structurally similar to 3-MTHF, but can be derived from a less expensive feedstock. 2-BF and 2-BTHF were proposed as suitable CI engine fuel candidates by TMFB researchers (Julis and Leitner, 2012). These substances can be derived from furfural, which is available at lower market prices than itaconic acid. Assessment of the topology of the network revealed that a manifold of reaction pathways are available, especially for the synthesis of 2-BTHF. The most important reactions in the network were identified based on their frequency of occurrence in the elementary modes. Economic evaluation of the optimal flux distribution allows for the assumption that the production of 2-BTHF in process configurations with waste stream integration is viable at current furfural market prices.

Three main implications arise on the further course of conduct concerning the use and current state of implementation of the presented methodology, which will be presented in the following sections.

7.1 Increasing property model detail and accuracy

Up to now, only sharp splits are assumed that ideally separate the main product from the residual components in the stream. This assumption cannot be transferred to the final application. Non-sharp splits will occur that decrease the amount of generated main product which by implication leads to the processing of mixtures. At this point, linear mixing rules

are no longer sufficient; rather, more detailed thermodynamic models (see Kontogeorgis and Folas (2009)) need to be employed in order to adequately describe the mixture.

Such more detailed modeling of thermodynamic behavior facilitate (i) an assessment of energy requirement of separation tasks and (ii) a more rigorous description of biofuel mixture properties. The first aspect was already addressed by Voll (2013), who proposed a close integration of network evaluation and process design. Separation steps introduced after each network reaction are evaluated by means of simple performance indicators to assess the potential of the separation task (such as proposed in Jaksland et al. (1995) who incorporate quotients of pure compound properties to assess the separation potential).

The second aspect targets a more rigorous model-based foundation for the design and production of biofuel mixtures. A first step towards the production of biofuel blends was taken by considering the waste streams as part of the final product. However, the integration is performed in addition to the optimal production of a pure substance. It can be assumed that by targeting a mixture from the very beginning, even more efficient processes can be identified.

7.2 Design of sustainable value chains

A thorough integration of the presented methodologies into a wider context, considering the multi-dimensionality of sustainability in terms of economical, ecological and social effects (Scott-Cato, 2009) will allow for assessing the performance of a product-process combination as a whole. This requires an integrated consideration of biomass cultivation, production of the biobased product, the effects of product use and production process on the environment and the impact of the environment on biomass cultivation. Thinking beyond a pure metric of measuring effects, an integrated consideration will allow for decision-making on the type of utilized biomass based upon market prices or even the design of an optimal biomass composition.

Most challenging to achieve is a detailed description of the ecosystem. First approaches could incorporate already established models that globally abstract the environment by means of energy and material flows (Bakshi and Fiksel, 2003). However, more detailed models are required to assess additional local factors such as land-use change, biomass harvest and water consumption, but also global factors such as the increase of greenhouse gas accumulation in the atmosphere and average surface temperature of the planet.

Extending this approach to consider an integrated biorefinery (Fernando et al., 2006) allows for flexible shifts of the product spectrum to actively react on shifts in market prices and availability of biomass feedstock.

7.3 Synthesis design outside the biofuel scope

In this thesis, ReNeGen was solely applied to biofuel synthesis tasks. The overview on reaction network generators in Section 2.1 shows that various other fields of application exist. By including a wider set of atom types with their corresponding valence schemes, various other synthesis tasks outside the scope of biofuel production can be addressed. Likewise, a wider set of empirical knowledge can be included in ReNeGen. As such, ReNeGen will be suitable for a variety of synthesis design tasks in organic chemistry.

Appendices

A - Mathematical preliminaries

Appendix A provides a short introduction to the mathematical terms of graph and set theory used in this thesis. The collocation of fundamentals in graph theory was compiled from Diestel (2006), while the work of Deiser (2004) was used for compiling the mathematical preliminaries of set theory.

A.1 Graph theory

A simple graph $G(V, E)$ is an ordered pair of a non-empty vertex set $V(G)$ and a non-empty edge set $E(G)$. Each element of $e \in E(G)$ is said to be an edge joining two vertices $v_i \in V(G)$ and $v_j \in V(G)$, thus $e = \{v_i, v_j\}$. In an undirected graph, the edge $\{v_i, v_j\}$ is identical to the edge $\{v_j, v_i\}$. This is not the case in a directed graph where each edge e has a determined direction. In directed graphs, the position of a vertex on an edge leads to a distinction of its labeling. v_i is called the *tail* of the edge and v_j is the *head*. v_j is the a *successor* of v_i, and v_i is a *predecessor* of v_j.

$|V(G)|$ is the number of vertices and $|E(G)|$ is the number of edges in G. Two vertices v_i and v_j are called *adjacent* if $\{v_i, v_j\} \in E(G)$. An edge e and a vertex v are called incident, if $v \in e$, meaning that v is a vertex on e.

Subgraph

Assume two graphs, G and G' with

$$G \cup G' := (V \cup V', E \cup E') \tag{A.1}$$

and

$$G \cap G' := (V \cap V', E \cap E'). \tag{A.2}$$

If $G \cap G' = \emptyset$, then G and G' are disjunct. If $V' \subseteq V$ and $E' \subseteq E$, then G' is a subgraph to G (and G a supergraph to G'), stated as $G' \subseteq G$.

Distance and neighborhood

The *distance* between two vertices i and j, $d(i,j)$, is the number of edges in the shortest path connecting them. Consider the network in Figure A.1 as an example. The distances between vertex A and the other vertices in the network are denoted by $d(A,j)$. A vertex can have multiple neighboring vertices, called *neighborhood*, which are denoted $N(v)$. For instance, $N(A) = \{B, C, D, E\}$ in Figure A.1. The consideration of neighborhood can be extended to account for vertices that are in a *neighborhood distance* i to vertex v. This neighborhood is then referred to as $N^i(v)$. Again referring to Figure A.1, $N^2(A) = \{F, G, H\}$.

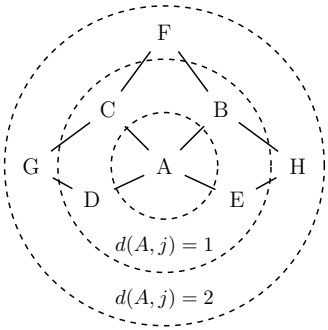


Figure A.1 – The distance in networks

Degree of a vertex

The *degree of a vertex* $d(v)$ is the number of edges that are incident to v. In undirected graphs, the degree of a vertex v is equal to $|N^1(v)|$. In directed graphs, *in-* and *outdegree* of a vertex v are distinguished. The number of head endpoints adjacent to a vertex v is called outdegree $d^-(v)$, representing the number of successors; the number of tail endpoints adjacent to v is called indegree $d^+(v)$, representing the number of predecessors. v is a source of a graph if $d^+(v) = 0$ and a sink if $d^-(v) = 0$.

Paths in a graph

A *path* $P(V,E)$ is a nonempty graph in the form of

$$P = \{v_0, v_1, ..., v_k\}, \tag{A.3}$$

which denotes a vertex path, or

$$P = \{\{v_0, v_1\}, \{v_1, v_2\}, ..., \{v_{k-1}, v_k\}\}, \tag{A.4}$$

which denotes an edge path. v_0 and v_k are the terminal vertices of P. The number of elements in a path is called the *length*. A path of length k, where k denotes the number of elements in P, is referred to as P^k.

Connectivity

A nonempty graph G is considered connected, if any two vertices are connected by paths. If G is not completely connected, G is supergraph to a set of subgraphs G', called *connected components*. A connected component G' is a subgraph, if it contains all edges $\{v_i, v_j\} \in E$, where $v_i, v_j \in V'$.

Union of two graphs

Consider two graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$. The *union* of these two graphs is:

$$G_1 \cup G_2 := (V_1 \cup V_2, E_1 \cup E_2). \tag{A.5}$$

Likewise, the definition can be extended to be applicable to an arbitrary, but finite, number of graphs G_1,...,G_n through:

$$\bigcup_{i=1}^{n} G_i = G_1 \cup ... \cup G_n := (V_1 \cup ... \cup V_n, E_1 \cup ... \cup E_n). \tag{A.6}$$

Both definitions are equally applicable to undirected as well as directed graphs. Consider the following figure as an example to the union of three undirected graphs.

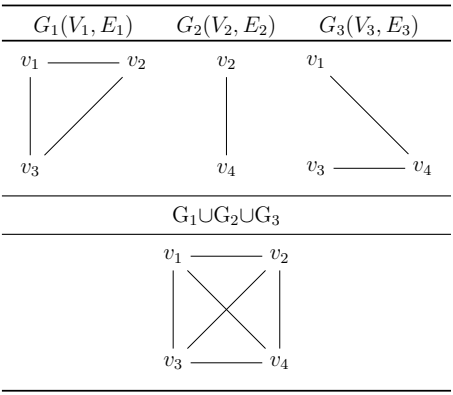


Figure A.2 – Example of merging three individual networks

A.2 Set theory

Cardinality

The *cardinality* of a set is the measure of the number of the elements in the set. The cardinality of a set A is usually denoted $|A|$.

Cartesian product

The *Cartesian product* of two sets A and B is defined as

$$A \times B := \{(a, b) | a \in A, b \in B\}. \tag{A.7}$$

The Cartesian product for a finite number of sets, $A_1, ..., A_n$ is defined by

$$A_1 \times ... \times A_n := \{(a_1, a_n) | a_i \in A_i \; for \; \{i = 1, ..., n\}\}. \tag{A.8}$$

The number of n-tuples from the Cartesian product of n sets is

$$|A_1 \times ... \times A_n| := |A_1| \cdot ... \cdot |A_n|. \tag{A.9}$$

B - Reactions rules

The following table lists the pattern vectors of the implemented reaction rule.

Table B.1 – The set of reaction rules in ReNeGen

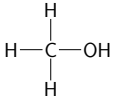
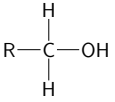
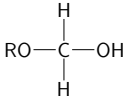
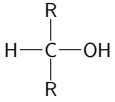
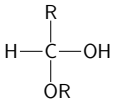
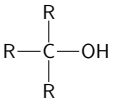
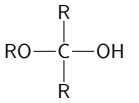
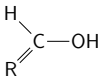
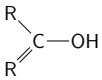
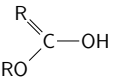
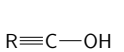
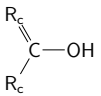
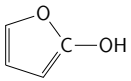
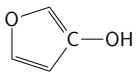
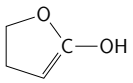
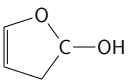
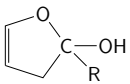
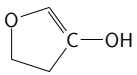
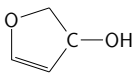
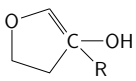
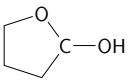
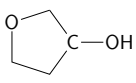
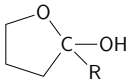
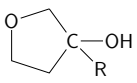
Nr.	Name	Pattern vector
1	Ketone/Aldehyde hydrogenation	$\mathbf{p^B}$ = (1,0,1,0,0,0,0,0,0,0)
2	Alcohol hydrogenation	$\mathbf{p^B}$ = (1,0,0,1,0,0,0,0,0,0)
3	Alkene hydrogenation	$\mathbf{p^B}$ = (1,0,0,0,1,0,0,0,0,0)
4	Heterocycle hydrogenation	$\mathbf{p^B}$ = (1,0,0,0,0,1,0,0,0,0)
5	Heterocycle hydrolysis	$\mathbf{p^B}$ = (0,1,0,0,0,1,0,0,0,0)
6	Carboxylic acid hydrogenation	$\mathbf{p^B}$ = (1,0,0,0,0,0,1,0,0,0)
7	Formation of acid anhydrides	$\mathbf{p^B}$ = (0,0,0,0,0,0,2,0,0,0)
8	Etherification	$\mathbf{p^B}$ = (0,0,0,2,0,0,0,0,0,0)
9	Esterification	$\mathbf{p^B}$ = (0,0,1,0,0,0,0,1,0,0)
10	Aldol condensation	$\mathbf{p^B}$ = (0,0,1,0,0,0,0,1,0,0)
11	Alcohol dehydration	$\mathbf{p^B}$ = (0,0,0,0,0,0,0,0,1,0)
12	Ketonization	$\mathbf{p^B}$ = (0,0,0,0,0,0,0,0,0,2)

C - Triggering and resulting patterns for determining non-selective reactions

Table C.1 lists the molecular motifs that lead to non-selective reactions if occurring at least twice in one molecule (*triggering patterns*). They are dissected into hydroxyl, carbonyl, carboxylic acid and olefin group as well as hydroxy condensation pattern. Distinctions are made between carbon chains, carbon rings and 5- and 6-membered heterocyclic compounds. The heterocyclic compounds are further distinguished into furan and pyran, respectively, as well as partially (dihydrofuran (DHF), dihydropyran (DHP)) and fully hydrogenated derivatives (tetrahydrofuran (THF), tetrahydropyran (THP)). Table C.2 lists the molecular motifs that result from refunctionalizing the corresponding motif in Table C.1 (*resulting patterns*).

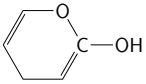
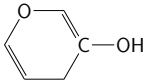
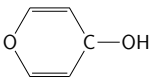
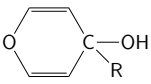
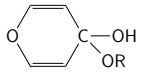
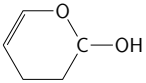
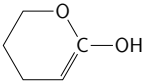
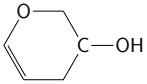
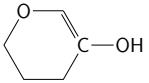
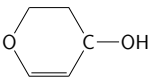
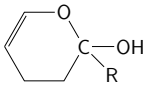
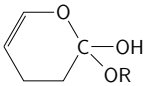
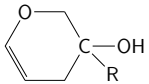
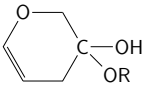
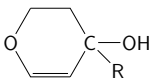
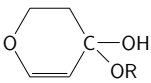
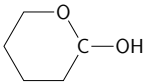
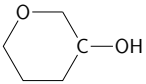
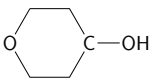
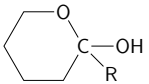
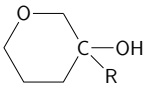
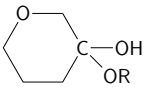
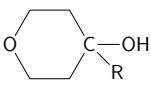
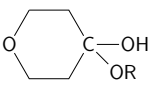
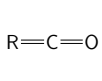
In accordance with the Erlenmeyer rule, only arrangements with no more than two oxygen atoms attached to a single carbon atom are considered. Arrangements with a higher number of oxygen atoms adjacent to a single carbon atom are not stable, but rather will undergo a condensation of one hydroxy group (Furniss et al., 1989). Only one of these oxygen atoms is allowed to be in a hydroxy formation, the other one has to be an ether, an ester or part of a carboxylic acid. The presented patterns cover molecular motifs that are commonly encountered in biomass respectively its derivatives.

Table C.1 – Distinguished *triggering patterns* for selectivity assessment and estimation, ordered by functionality. The residual R represents a carbon atom that is part of an arbitrary molecule. R_c represents an a pure carbon ring of arbitrary size.

	Hydroxy Group			
Carbon chain				
				
				
Carbon cycle				
Furan				
DHF				
				
THF				

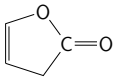
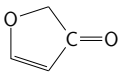
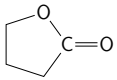
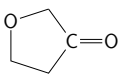
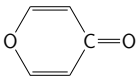
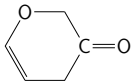
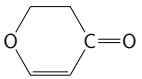
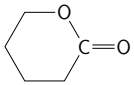
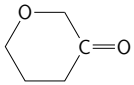
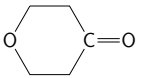
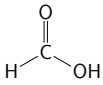
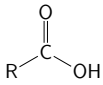
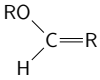
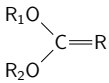
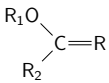
Continued on next page

Table C.1 – continued from previous page

Pyran				
				
DHP				
				
				
THP				
				
		Carbonyl group		
Carbon chain				
Carbon cycle				

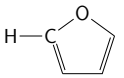
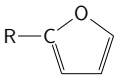
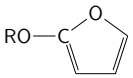
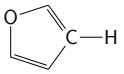
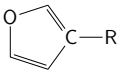
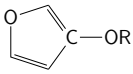
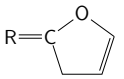
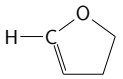
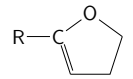
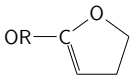
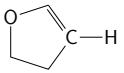
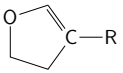
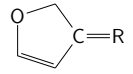
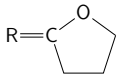
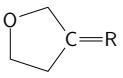
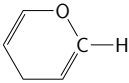
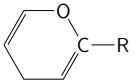
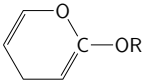
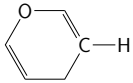
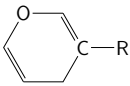
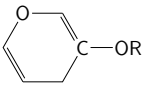
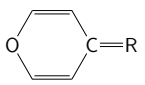
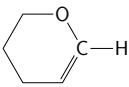
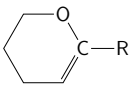
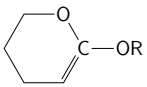
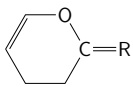
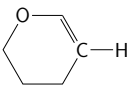
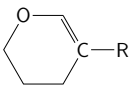
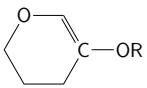
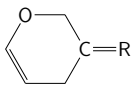
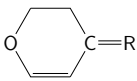
Continued on next page

Table C.1 – continued from previous page

DHF			
THF			
Pyran			
DHP			
THP			
Carboxylic acid group			
Carbon chain			
Olefinic group			
Carbon chain			  
Carbon cycle			

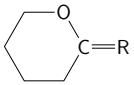
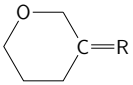
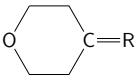
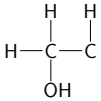
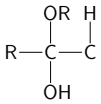
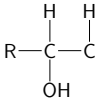
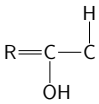
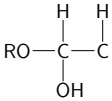
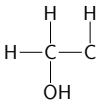
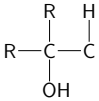
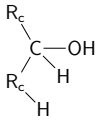
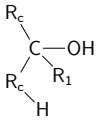
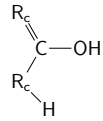
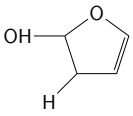
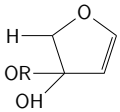
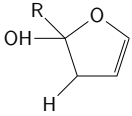
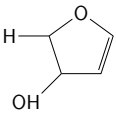
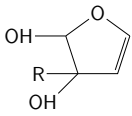
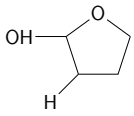
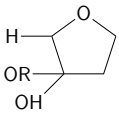
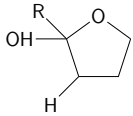
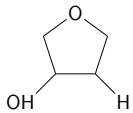
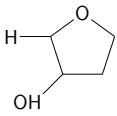
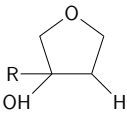
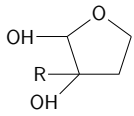
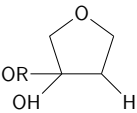
Continued on next page

Table C.1 – continued from previous page

Furan				
				
DHF				
				
THF				
Pyran				
				
DHP				
				
				

Continued on next page

Table C.1 – continued from previous page

THP				
Hydroxy condensation pattern				
Carbon chain	 	 	 	
Carbon cycle				
DHF	 			
THF	 	 	 	 

Continued on next page

Table C.1 – continued from previous page

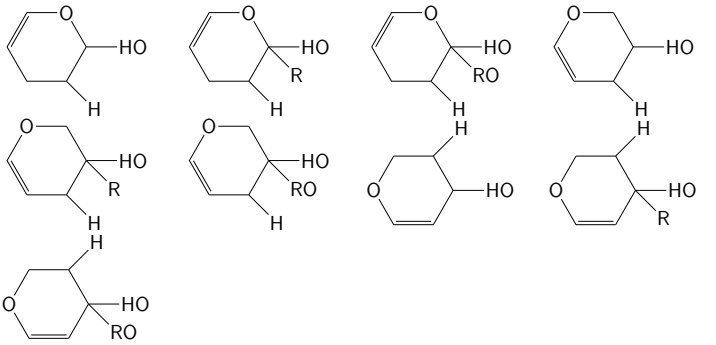
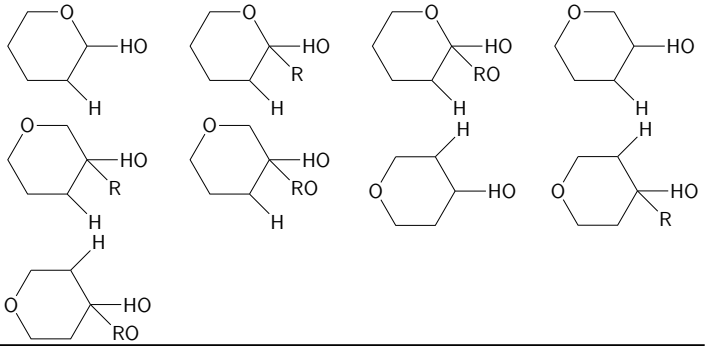
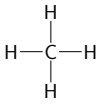
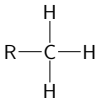
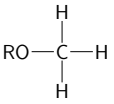
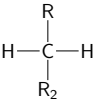
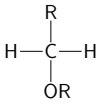
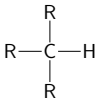
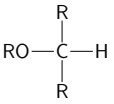
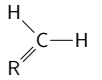
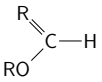
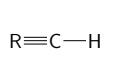
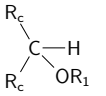
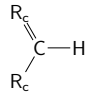
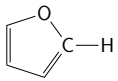
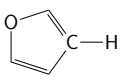
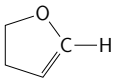
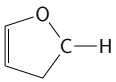
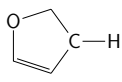
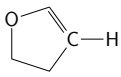
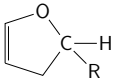
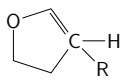
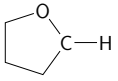
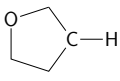
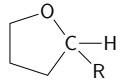
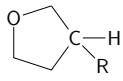
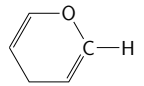
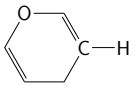
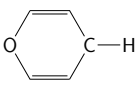
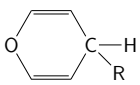
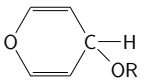
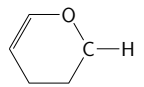
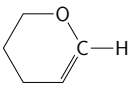
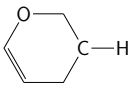
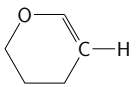
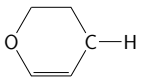
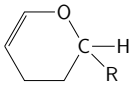
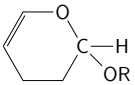
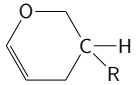
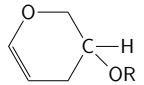
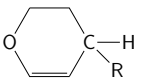
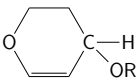
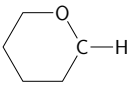
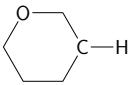
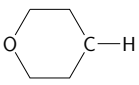
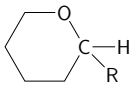
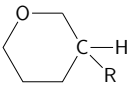
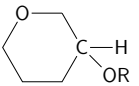
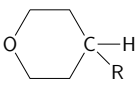
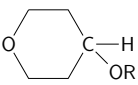
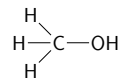
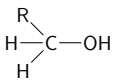
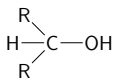
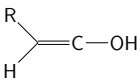
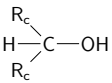
DHP	
THP	

Table C.2 – Distinguished *resulting patterns* in the context of selectivity estimation, ordered by functionality. The residuals R always start with a carbon atom. R_c is a rest that consists of a pure carbon ring (no hetero atoms).

	Hydroxy Group			
Carbon chain				
				
				
Carbon cycle				
Furan				
DHF				
				
THF				

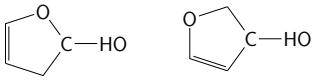
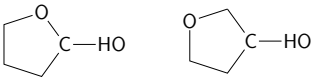
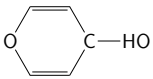
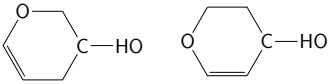
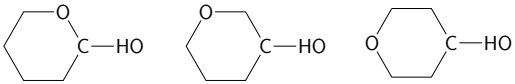
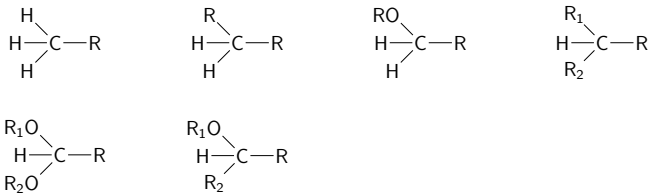
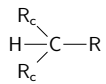
Continued on next page

Table C.2 – continued from previous page

Pyran				
				
DHP				
				
				
THP				
				
Carbonyl group				
Carbon chain				
Carbon cycle				

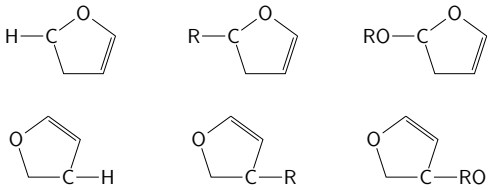
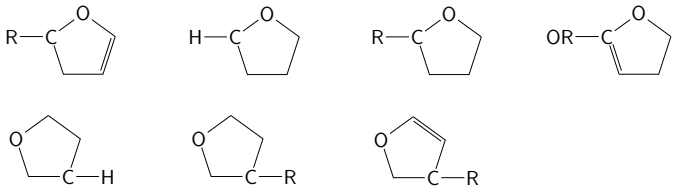
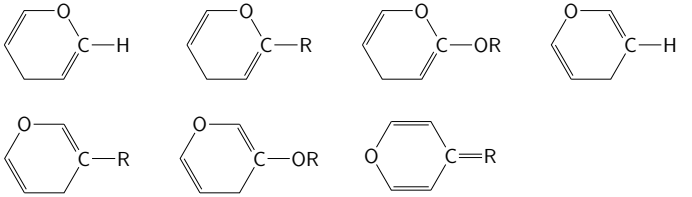
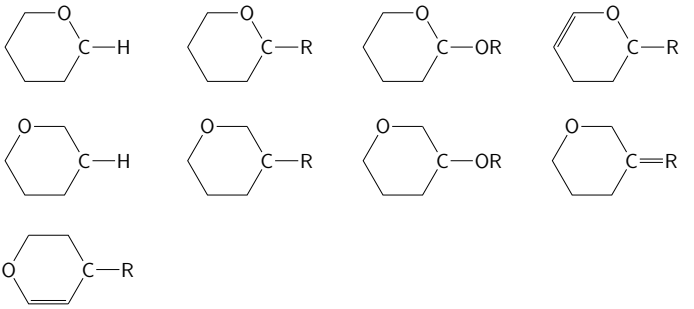
Continued on next page

Table C.2 – continued from previous page

DHF	
THF	
Pyran	
DHP	
THP	
Carboxylic acid group	
Carbon chain	
Olefinic group	
Carbon chain	
Carbon cycle	

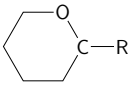
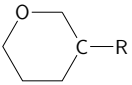
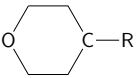
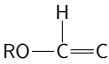
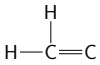
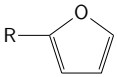
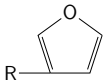
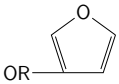
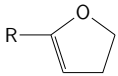
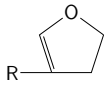
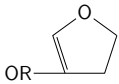
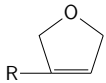
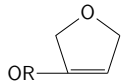
Continued on next page

Table C.2 – continued from previous page

Furan	
DHF	
THF	
Pyran	
DHP	

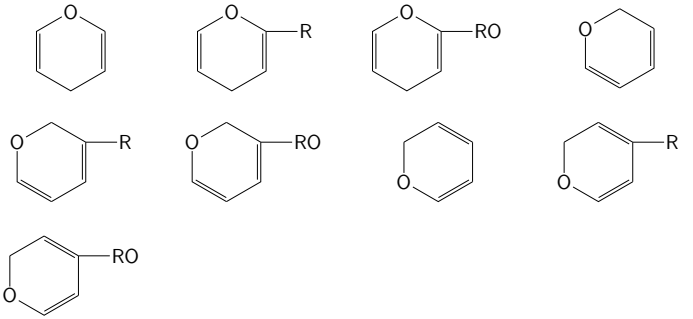
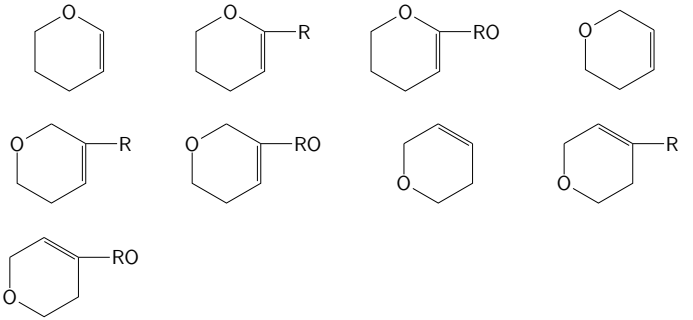
Continued on next page

Table C.2 – continued from previous page

THP				
Hydroxy condensation pattern				
Carbon chain				
				
Carbon cycle				
DHF				
				
THF				
				

Continued on next page

Table C.2 – continued from previous page

DHP	
THP	

D - Property models

The evaluation of de-novo synthesis pathways requires knowledge on the thermophysical properties of unmeasured substances. Quantitative structure property relationships (QSPRs) are employed to calculate a defined set of properties to each substance. Since the number of substances in the network is often large, the property models have to embody a compromise between accuracy and calculation effort.

The QSPR models that were derived by Dahmen et al. (2012) are used in this thesis. QSPR models employ molecular descriptors (Todeschini and Consonni, 2008), calculated from the molecule's two- or three-dimensional structure, to correlate a property of interest with a molecule's structural features. A property P is described as a linear combination of the descriptor variables D_i of the following form:

$$P = a_1 \cdot D_1 + a_2 \cdot D_2 + a_3 \cdot D_3 + ... \tag{D.1}$$

D_i is a subset of descriptors and a_i are weights derived by an adequate regression method. QSPR models are commonly employed for the prediction of thermophysical properties. For detailed information on the model building workflow that is used to derive the QSPR models employed in this thesis, refer to the work of Dahmen et al. (2012). A general overview on the historical development of QSPR models has been provided by Katritzky and Fara (2005).

Table D.1 lists the split between training and external validation data along with the average relative error (ARE) and the maximum relative error (MRE) of each model. The last column provides the data source for the model building molecules and their property values.

Table D.1 – Information upon QSPR models and modeling data

Property	# Training compounds # Ext. Val. compounds	ARE [%] ARE [%]	MRE [%] MRE [%]	Data source
CN	205 26	31.92 25.40	1153[a] 243.75	(Dahmen et al., 2012)
ΔH_{com}	846 94	0.62 0.43	10.13 5.80	(Rowley et al., 2003)
LD_{50}	525 64	1.54 1.72	23.04 10.98	ICAS[b]
ρ_l	624 69	1.44 1.50	15.41 6.93	(Rowley et al., 2003)
T_{boil}	817 98	1.47 1.03	13.61 3.52	(Rowley et al., 2003)
T_{melt}	821 94	8.34 9.90	64.13 47.78	(Rowley et al., 2003)

[a]The observed deviation stems from the database value being 1 and the predicted value 12.53, next largest MRE is 383.12%

[b]Model was build from simulated LD_{50} data of a representative set of substances calculated by software package ICAS, employing group contribution method as presented in (Hukkerikar et al., 2012)

E - Evaluation criteria

An overview on the evaluation criteria is provided in Table E.1. They are distinguished into material, economic, environmental and energetic criteria. They are discussed in detail in the following sections. The collocation of evaluation criteria to evaluate reaction networks was first performed by Hechinger et al. (2010) and Voll and Marquardt (2012a,b).

Table E.1 – Evaluation criteria sorted by category

	Name	Unit	Formula	
Material	Reactant consumption	mole/a	$N_i^{in} = \sum_{j=1}^{r} N_{i,j}^{in}, \quad i \setminus FS$	$(E.1)$
	Feedstock consumption	mole/a	$N_{FS}^{in} = \sum_{j=1}^{r} N_{i,j}^{in}, \quad i = FS$	$(E.2)$
	Product formation	mole/a	$N_T^{out} = \sum_{j=1}^{r} N_{i,j}^{out} \quad i = T$	$(E.3)$
	Substances with GWP	mole/a	$N_i^{in} = \sum_{j=1}^{r} N_{i,j}^{out}, \quad i \in \{CH_4, CO_2, CO\}$	$(E.4)$
	Atom efficiency	–	$AE = \frac{M_T^{out}}{\sum_{i=1}^{s} M_i^{in}}$	$(E.5)$
	Carbon efficiency	–	$CE = \frac{N_T^{out} \cdot nC_T}{\sum_{i=1}^{s} N_i^{in} \cdot nC_i}$	$(E.6)$
	Synthesis yield	–	$Y = \frac{N_T^{out}}{N_{FS}^{in}}$	$(E.7)$
Economic	Investment cost	\$	$IC = 3 \cdot \Delta E^{0.84}$	$(E.8)$
	Total annualized costs	\$/a	$TAC = \frac{IC \cdot z}{1-(1+z)^{-t}} + \sum_{i=1}^{s} C_i$	$(E.9)$
	Total annualized revenues	\$/a	$TAR = \sum_{i=1}^{s} R_i - TAC$	$(E.10)$
	Maximum feedstock price	\$/kg	$p_{FS}^{max} = \frac{R_T - \frac{IC \cdot z}{1-(1+z)^{-t}}}{M_{FS}^{in}}$	$(E.11)$
	Minimum fuel price	\$/kg	$p_T^{min} = \frac{C_{FS} + \frac{IC \cdot z}{1-(1+z)^{-t}}}{M_T^{out}}$	$(E.12)$

Continued on next page

Table E.1 – continued from previous page

	Name	Unit	Formula	
Environment	Energy consumption	MW/kg	$EC = \frac{\Delta E}{M_T^{out}}$	(E.13)
	Resource consumption	kg/kg	$RC = \frac{M_T^{out}}{\sum_{i=1}^{s} M_i^{in}}$	(E.14)
	Emission impact	kg/kg	$Em = \frac{M_{CO}^{out} + M_{CO_2}^{out} + M_{CH_4}^{out}}{M_T^{out}}$	(E.15)
	Toxicity potential	kg	$TP = M_T^{out} \cdot TS_T$	(E.16)
	Environmental impact	–	$EI = \frac{EC}{EC_{ref}} + \frac{RC}{RC_{ref}} + \frac{Em}{Em_{ref}} + \frac{TP}{TP_{ref}}$	(E.17)
Energetic	ΔH_{com} efficiency	–	$\eta_{E,com} = \frac{N_T^{out} \cdot \Delta H_{com,T}}{\sum_{i=1}^{s} N_i^{in} \cdot \Delta H_{com,i}}$	(E.18)
	ΔH_{form} efficiency	–	$\eta_{E,form} = \frac{N_T^{out} \cdot \Delta H_{form,T}}{\sum_{i=1}^{s} N_i^{in} \cdot \Delta H_{form,i}}$	(E.19)

FS: Feedstock
T: Synthesis target
Material prices p_i in \$/kg
The mass stream M_i of a substance is calculated via $M_i = N_i \cdot MW_i$
ΔH_{com} and ΔH_{form} are of unit kJ/mole

E.1 Material balance related criteria

The absolute flux $N_{i,j}$ of each substance i in reaction j is known from the flux evaluation presented in Chapter 5. The material streams of reactants (Equation (E.1)), main substrate (Equation (E.2)) and main target (Equation (E.3)) are most interesting since they strongly influence the efficiency, economic competitiveness and sustainability of the flux distribution. Another special interest is in the formation of methane and carbon dioxide (Equation (E.4)), since they contain a global warming potential (Protocol, 1997).

Atom and carbon efficiency (Equations (E.5) and (E.6)) quantify the material efficiency from the substrate to the target compound. Atom and carbon efficiency are important concepts in green chemistry for evaluating the sustainability of a production process. They are the most widely applied measures of the efficiency of a process or product (Lapkin and Constable, 2008). Low carbon and atom efficiencies are undesired since they imply that significant amounts of main substrate are lost in the synthesis. The synthesis yield (Equation (E.7)) quantifies the conversion of main substrate into the main target.

E.2 Economic criteria

Lange (2001) derived criteria to evaluate the investment costs (IC) of the process, based on the energetic loss ΔE (in MW/a) of enthalpy of combustion ΔH_{com}:

$$\Delta E = \frac{\sum_{i=1}^{s} N_i^{in} \cdot \Delta H_{com,i} - N_T^{out} \cdot \Delta H_{com,T}}{31536000\ s} \tag{E.20}$$

is the annual difference in heat of combustion of the provided substances (substrates and reactants) on the one side and the main target on the other side. The investment costs are calculated via Equation (E.8). Since the equation was derived from investments based on the year 1993, they are updated to February 2014 (Cheresources.com, 2014) using the CEPCI index (Lozowski, 2012), according to

$$IC_{2014} = \frac{CEPCI_{2014}}{CEPCI_{1993}} \cdot IC_{1993}. \tag{E.21}$$

The calculation of the costs C_i of a substance i requires the knowledge of its market price p_i in \$/kg. The cost result to

$$C_i = p_i \cdot \sum_{j=1}^{r} \mathbf{N}_{i,j}^{in} \cdot MW_i. \tag{E.22}$$

Likewise, the revenues R_i are calculated from

$$R_i = p_i \cdot \sum_{j=1}^{r} \mathbf{N}_{i,j}^{out} \cdot MW_i. \tag{E.23}$$

$$\tag{E.24}$$

With further information upon the lifetime of the plant t (in years) and the interest rate z (in percent) one can calculate the total annualized costs as presented in Equation (E.9). The total annualized revenues (Equation (E.10)) result from the difference of the revenues earned from selling the produced substances and the total annualized cost. Maximum feedstock costs or minimum sales price of the product can be calculated from combining Equations (E.9) and (E.10) and setting the total annual revenues to zero.

E.3 Environmental criteria

Voll and Marquardt (2012a) proposed an adoption of the Eco-Efficiency Analysis (EEA) of Saling et al. (2002) to assess the environmental impact of the synthesis pathways. The original formulation contains six individual contributions, namely energy consumption,

resource consumption, emissions, land use, toxicity and risk potential. Land use and risk potential are neglected in this evaluation as the required input data is not available. The EEA is a relative metric, meaning that a reference entity must be chosen a priori. Thus, an established process such as FAME production from vegetable oil serves as a reference. These processes are evaluated in the same manner as the synthesis pathways.

Energy Consumption (EC) is a measure for the energy required to produce a unit of target substance, relating the energy consumption of the process to the amount of substance produced (Equation (E.13)). It considers the amount of energy that is lost in terms of heat of combustion. Neither consumption of electric energy nor heat are considered.

The Resource Consumption (RC) measures the amount of feedstock and reactants used per unit of target substance (Equation (E.14)).

The Emission Indicator (Em) accounts for the amount of emissions released per unit of target substance (Equation (E.15)). Only air emissions are considered, soil and water are neglected. Voll (2013) states that air emissions are expected to have the most severe environmental impact and can hardly be avoided. Thus, the emission indicator accounts for those substances that are expected to promote climate change, which are CO, CO_2 and CH_4. Global Warming Potential (GWP) values are those derived by the United Nations Framework Convention on Climate Change (Protocol, 1997). Each individual pollutant is weighted with its GWP to calculate the total GWP of the synthesis pathway.

The Toxicity Potential (TP) measures the toxicity of a synthesis pathway. The toxicity of a substance is represented by its LD_{50}-value (amount of substance killing the median of a test population). Saling et al. (2002) propose a classification of certain intervals of LD_{50} values in accordance the German Chemicals Act, assigning toxicity scores to substances within defined LD_{50} intervals (cf. Table E.2) The toxicity score of the main product is multiplied by the amount of produced main product to give the toxicity potential of a synthesis pathways (Equation (E.16)).

All four contributions (RC, EC, Em and TP) are then related to the reference entities EC_{ref}, RC_{ref}, Em_{ref}, TP_{ref}. The Environmental Impact (EI) is the sum of the individual increments (Equation (E.17)).

Table E.2 – Classification of LD_{50} values for the calculation of Toxicity Potential (TP) in Eco-Efficiency Analysis (EEA) as introduced by Saling et al. (2002)

Symbol	Concentration interval, LD_{50} value, rat, oral, 2h	Assigned TP value in EEA
T+, very toxic	$LD_{50} \leq 25$ mg/kg	1000
T, toxic	25 mg/kg $< LD_{50} \leq 200$ mg/kg	100
Xn, harmful	200 mg/kg $< LD_{50} \leq 2000$ mg/kg	10
Xi, irritant	LD_{50}>2000 mg/kg	1

E.4 Energetic criteria

The energetic efficiency of a pathway can either be measured in terms of the conservation of the heat of combustion ΔH_{com} or the enthalpy of formation ΔH_{form} (cf. Equations (E.18) and (E.19)). Energetic efficiency relates the amount of energy leaving the network as main product to the amount of energy that are provided to the network. ΔH_{com} and ΔH_{form} of each substance are calculated via the QSPR models presented in Appendix D.

While the conservation of ΔH_{com} is interesting in terms of fuel production, the conservation of ΔH_{form} is a good measure for the quality of a reaction. The enthalpies of formation are calculated at standard state as the operation conditions of the reaction are not known.

F - Case study data

F.1 Data of 3-MTHF synthesis from itaconic acid

Table F.1 lists intermediates of the presented case study with their canonical SMILES representation derived by OpenBabel (OBoyle et al., 2011) and the predicted property values that are important for the evaluation of the reaction pathways.

Table F.1 – Properties of pure substances in 3-MTHF synthesis network

Smiles	MW	CN	H_{com}	LD_{50}	ρ_l	T_{boil}	T_{melt}
–	kg/kmol	–	MJ/kg	mg/kg	kg/l	K	K
OC(=O)CC(=C)C(=O)O	130.1	-67.4	-14.3	1565.7	1.32	602.8	382.7
O=C1OC(=O)C(=C)C1	112.1	-45.7	-17.6	368.1	1.60	512.4	322.4
OC(=O)CC(C(=O)O)C	132.1	-64.2	-15.2	1626.3	1.25	586.7	351.0
O=CC(=C)CC(=O)O	114.1	-16.2	-18.3	4404.8	1.20	533.5	321.5
O=CCC(=C)C(=O)O	114.1	-21.9	-18.5	3344.4	1.21	534.1	335.9
OC1CC(=C)C(=O)O1	114.1	-57.7	-18.9	374.2	1.29	521.5	356.2
O=C1CC(=C)C(O1)O	114.1	-69.4	-19.2	374.5	1.31	509.1	349.8
O=C1CC(C(=O)O1)C	114.1	-42.6	-18.2	383.8	1.22	492.2	294.1
O=CCC(C(=O)O)C	116.1	-19.5	-19.6	3463.7	1.15	519.1	292.6
O=CC(CC(=O)O)C	116.1	-14.3	-19.5	4562.9	1.14	519.2	278.7
OCC(=C)CC(=O)O	116.1	-76.8	-19.8	4506.6	1.18	552.2	314.7
O=CCC(=C)C=O	98.1	39.5	-24.2	8980.8	1.08	457.9	263.5
OCCC(=C)C(=O)O	116.1	-68.6	-19.7	3406.0	1.18	551.4	343.0
OC1OC(C(=C)C1)O	116.1	-63.4	-20.7	379.2	1.26	503.1	377.9
OC1OC(=O)C(C1)C	116.1	-54.6	-19.5	389.4	1.22	497.7	322.2
C=C1CCOC1=O	98.1	-16.9	-23.9	1139.0	1.11	459.0	287.2
O=C1CC(C(O1)O)C	116.1	-57.0	-19.5	390.7	1.23	497.8	324.9
C=C1COC(=O)C1	98.1	-33.6	-23.9	1524.8	1.10	463.0	276.5
OCCC(C(=O)O)C	118.2	-64.6	-20.7	3542.7	1.12	537.3	312.9
O=CCC(C=O)C	100.1	42.3	-25.5	9310.3	1.03	444.5	216.2
OCC(CC(=O)O)C	118.1	-63.4	-20.5	4693.9	1.12	541.8	295.1
OCC(=C)CC=O	100.1	-22.8	-25.8	9174.1	1.07	480.3	249.4
OCCC(=C)C=O	100.1	-9.3	-25.5	9121.5	1.06	478.5	265.2

Continued on next page

Table F.1 – continued from previous page

Smiles	MW	CN	H_{com}	LD_{50}	ρ_l	T_{boil}	T_{melt}
–	kg/kmol	–	MJ/kg	mg/kg	kg/l	K	K
OC1OC(C(C1)C)O	118.1	-51.6	-20.6	396.2	1.20	495.2	338.0
OC1CC(=C)CO1	100.1	-29.8	-25.5	1541.2	1.07	459.5	296.8
C=C1CCOC1O	100.1	-23.9	-25.8	1151.8	1.09	444.9	305.9
O=C1OCCC1C	100.1	-12.8	-24.4	1188.4	1.05	439.2	258.3
CC1CC(=O)OC1	100.1	-20.5	-24.2	1593.2	1.03	446.4	257.5
CC(C=O)CCO	102.2	-5.9	-26.8	9465.4	1.02	465.8	222.0
CC(CO)CC=O	102.2	-9.7	-26.7	9537.6	1.02	471.3	216.3
OCCC(=C)CO	102.2	-67.8	-27.0	9387.2	1.05	502.6	269.8
CC1CCOC1O	102.3	-12.0	-25.7	1206.3	1.03	436.7	263.5
CC1COC(C1)O	102.2	-17.1	-25.5	1610.0	1.01	445.2	258.9
C=C1COCC1	84.1	24.7	-33.4	4111.0	0.92	388.2	199.1
OCCC(CO)C	104.2	-53.8	-27.7	9798.1	1.01	496.1	246.9
CC1COCC1	86.2	40.1	-33.2	4316.7	0.87	373.1	162.0

Table F.2 presents the conversions C_j for each reaction that are incorporated into the reaction pathway evaluation. If an experimentally derived value was available in literature, the corresponding publication is provided. For every other reaction, a default conversion of 0.97 is assumed. Pseudo reactions (not listed in this table) are always assumed to exhibit a conversion rate of 1, as they do not constitute a reaction in its original meaning of converting molecules.

Table F.2 – Reaction conversions in 3-MTHF case study

R_j	C_j	Reference	R_j	C_j	Reference
R1	1	(Robert et al., 2011)	R50	0.97	–
R2	1	(Huang et al., 2010)	R51	0.97	–
R3	0.97	–	R52	0.94	(Bartholomäus et al., 2013)
R4	0.97	–	R53	0.97	–
R5	0.97	–	R54	0.97	–
R6	0.97	–	R55	0.97	–
R7	0.97	–	R56	0.97	–
R8	0.97	–	R57	0.97	–
R9	0.97	–	R58	0.97	–
R10	0.88	(Midgley and Thomas, 1987)	R59	0.97	–

Continued on next page

Table F.2 – continued from previous page

R_j	C_j	Reference	R_j	C_j	Reference
R11	0.97	–	R60	0.97	–
R12	0.97	–	R61	0.97	–
R13	0.97	–	R62	0.97	–
R14	0.97	–	R63	0.75	(Krohn and Riaz, 2004)
R15	0.97	–	R64	0.97	–
R16	0.97	–	R65	0.97	–
R17	0.97	–	R66	0.97	–
R18	0.97	–	R67	0.92	(Mori, 2008)
R19	0.97	–	R68	0.97	–
R20	0.97	–	R69	0.97	–
R21	0.97	–	R70	0.97	–
R22	0.97	–	R71	0.97	–
R23	0.97	–	R72	0.97	–
R24	0.97	–	R73	0.97	–
R25	0.97	–	R74	0.97	–
R26	0.97	–	R75	0.97	–
R27	0.97	–	R76	0.97	–
R28	0.97	–	R77	0.97	–
R29	0.97	–	R78	0.92	(Adrio and Hii, 2011)
R30	0.97	–	R79	0.97	–
R31	0.97	–	R80	0.92	(Mori, 2008)
R32	0.97	–	R81	0.97	–
R33	0.97	–	R82	0.97	–
R34	0.97	–	R83	0.97	–
R35	0.97	–	R84	0.97	–
R36	0.97	–	R85	0.97	–
R37	0.97	–	R86	0.97	–
R38	0.97	–	R87	0.97	–
R39	0.97	–	R88	0.97	
R40	0.97	–	R89	0.97	–
R41	0.97	–	R90	0.93	(Olah et al., 1981)
R42	0.97	–	R91	0.97	–
R43	0.97	–	R92	0.97	
R44	0.97	–	R93	0.93	–
R45	0.97	–	R94	0.97	–
R46	0.97	–	R95	0.97	–
R47	0.97	–	R96	0.97	–
R48	0.97	–	R97	0.97	–
R49	0.97	–	R98	0.97	–

F.2 Data of 2-BTHF and 2-BF synthesis from furfural

Table F.3 presents the molecules of the 2-Butyltetrahydrofuran and 2-Butylfuran from furfural. Included in the table is the corresponding thermophysical property data.

Table F.3 – Properties of pure substances in 2-BF and 2-BTHF synthesis network

Smiles	MW	CN	H_{com}	LD_{50}	ρ_l	T_{boil}	T_{melt}
–	kg/kmol	–	MJ/kg	mg/kg	kg/l	K	K
O=Cc1ccco1	96,1	-12.9	-23.3	32.2	1.14	413.6	222.5
O=CC1=CCCO1	98.1	36.4	-24.7	35.9	1.11	423.9	225.5
O=CC1CC=CO1	98.1	31.5	-25.0	36.4	1.09	429.2	238.5
CC(=O)C=Cc1ccco1	136.2	-1.7	-28.4	1571.5	1.04	473.1	227.2
O=CC1CCCO1	100.1	55.4	-25.5	38.9	1.06	433.9	198.9
CC(=O)C=CC1=CCCO1	138.2	22.7	-29.7	1818.7	1.01	474.6	226.2
CC(=O)C=CC1CC=CO1	138.2	20.8	-29.7	1854.7	0.99	481.3	247.6
CC(C=Cc1ccco1)O	138.2	-21.4	-29.4	1770.7	1.05	488.8	219.2
CC(=O)CCc1ccco1	138.2	5.6	-28.9	1705.4	1.02	468.9	224.2
CC(=O)C=CC1CCCO1	140.2	38.1	-30.0	2031.4	0.98	488.9	205.6
CC(C=CC1=CCCO1)O	140.2	3.5	-30.5	2047.7	1.03	493.5	225.0
CC(=O)CCC1=CCCO1	140.2	30.9	-30.3	1996.6	0.99	463.3	226.9
CC(C=CC1CC=CO1)O	140.2	-1.1	-30.7	2093.3	1.00	495.2	243.9
CC(=O)CCC1CC=CO1	140.2	33.3	-30.3	2036.7	0.97	475.1	246.4
CC(CCc1ccco1)O	140.2	-7.7	-29.9	1941.6	1.03	486.1	221.4
CCC=Cc1ccco1	122.2	22.2	-35.3	248.7	0.91	421.9	168.7
CC(C=CC1CCCO1)O	142.2	16.5	-31.1	2296.0	0.99	500.6	195.3
CC(=O)CCC1CCCO1	142.2	54.1	-30.6	2200.5	0.95	481.1	207.7
CC(CCC1=CCCO1)O	142.2	18.1	-31.3	2279.4	0.99	477.0	222.4
CCC=CC1=CCCO1	124.2	50.0	-36.4	281.6	0.89	424.0	181.8
CC(CCC1CC=CO1)O	142.2	21.9	-31.3	2320.6	0.96	487.9	249.6
CCC=CC1CC=CO1	124.2	44.7	-36.6	287.0	0.87	426.3	192.2
CCCCc1ccco1	124.2	35.8	-35.8	266.1	0.90	421.7	175.8
CC(CCC1CCCO1)O	144.2	42.4	-31.5	2514.1	0.96	497.0	214.9
CCC=CC1CCCO1	126.2	61.5	-36.9	311.4	0.86	435.2	148.3
CCCCC1=CCCO1	126.2	61.4	-37.1	306.1	0.87	414.9	184.5
CCCCC1CC=CO1	126.2	65.2	-37.1	310.1	0.85	425.8	201.9
CCCCC1CCCO1	128.2	86.8	-37.2	332.0	0.84	435.9	170.0

Table F.4 presents the reaction conversions used in the case study. 6 reactions were found in literature, default conversions of 0.97 were assumed for every other reaction.

Table F.4 – Reaction conversions in 2-BF and 2-BTHF case study

R_j	C_j	Reference	R_j	C_j	Reference
R1	0.97	–	R31	0.97	–
R2	0.97	–	R32	0.97	–
R3	0.95	(Alvarez-Ibarra et al., 1992)	R33	0.97	–
R4	0.97	–	R34	0.97	–
R5	0.97	–	R35	0.97	–
R6	0.97	–	R36	0.97	–
R7	0.97	–	R37	0.97	–
R8	0.97	–	R38	0.97	–
R9	0.97	–	R39	0.97	–
R10	0.99	(He et al., 2012)	R40	0.97	–
R11	0.99	(Yamashita et al., 1980)	R41	0.97	–
R12	0.97	–	R42	0.97	–
R13	0.97	–	R43	0.97	–
R14	0.97	–	R44	0.97	–
R15	0.97	–	R45	0.97	–
R16	0.97	–	R46	0.97	–
R17	0.97	–	R47	0.97	–
R18	0.97	–	R48	0.97	–
R19	0.97	–	R49	0.97	–
R20	0.97	–	R50	0.97	–
R21	0.7	(Zhao et al., 2011)	R51	0.97	–
R22	0.97	–	R52	0.97	–
R23	0.97	–	R53	0.97	–
R24	0.97	–	R54	0.97	–
R25	0.73	(Waidmann et al., 2013)	R55	0.97	–
R26	0.97	–	R56	0.97	–
R27	0.97	–	R57	0.97	–
R28	0.97	–	R58	1	(Fischer et al., 1996)
R29	0.97	–	R59	0.97	–
R30	0.97	–	R60	0.97	–

G - Software availability and handling

Appendix G describes where to obtain source code of ReNeGen, which additional software environment is required to operate ReNeGen at full functionality and how to specify a reaction network generation.

G.1 Software availability

The software source code will be made available at the open source website `Openscience.org`. Questions concerning ReNeGen can be directed to `renegen.software@gmail.com`.

G.2 Setting up the software environment

ReNeGen is implemented and tested in MATLAB 2010a. MATLAB licences can be obtained at https://de.mathworks.com/. In the current state of development, ReNeGen comes without a graphical user interface. Hence, changes of the options, constraints and the paths to input, output, temporary, data and external software directories have to be directly denoted in the main file (*main.m*).

The subsequently listed software is required in addition to MATLAB to operate ReNeGen at its presented functionality:

- **TOMLAB Optimization Environment**, a commercial optimization and modeling platform comprising solvers that exceed the functionality of those available in MATLAB for solving optimization problems. Licences and free trial versions can be obtained at http://tomopt.com/tomlab/.

- **Open Babel**, an open source chemical toolbox for searching, converting, analyzing and storing the different ways of computationally representing chemical data. In ReNeGen, it is used to convert SMILES notation into BEM and AM notation and vice versa. The software is available at http://www.openbabel.org/.

- **Dragon**, a commercial software to calculate molecular descriptors for estimating thermophysical properties of the network substances by using the

QSPR models presented in Appendix D. Licenses can be obtained at http://www.talete.mi.it/products/dragon_description.htm.

- **Graphviz**, an open source graph visualization software used in ReNeGen to visualize the generated networks. It is available at http://www.graphviz.org/.

- **Indigo Depict**, which is part of the Indigo toolkit, an open source organic chemistry toolbox. Indigo is used to convert SMILES notations of molecules into their graphical depictions. It is available at http://lifescience.opensource.epam.com/indigo/index.html.

The paths to the additionally required software have to be denoted in the *PathStruct* array of the main file. This array also denotes the directions to the input, output, data and temporary directories. The paths to softwares and folders have to be stated in an absolute manner, not relative to the ReNeGen directory. An exemplary specification of the PathStruct array is presented subsequently, with absolute paths to be included instead of the dots (...).

```
PathStruct = struct   ('Data' ,'...\ReNeGen\data\',...
                      'Dragon','...\ReNeGen\src\dragon\',...
                      'Graphviz','...\graphviz-2.38\release\bin\dot\',...
                      'Indigo','...\ReNeGen\src \Indigo\indigo-depict.exe',...
                      'Input' ,'...\ReNeGen\Input\',...
                      'OBabel','...\OpenBabel-2.3.1\',...
                      'Output' ,'...\ReNeGen\Output\',...
                      'Temp' ,'...\ReNeGen\temp\',...
                      'TOMLAB','...\MATLAB\R2010a\toolbox\tomlab\');
```

G.3 Setting up a reaction generation task

ReNeGen offers ten options to specify a reaction network generation task. Several of these options can contain multiple entries, such as the provided substrates or the desired targets. In accordance to MATLAB syntax, multiple entries are separated by commas.

The main substrates for the reaction network generation are chosen from a set of platform chemicals an denoted in option 1. The user can choose from the substances that are presented in Table G.1 by denoting the corresponding arguments. The list of platform chemicals can be extended by adding further SMILES notations to the MATLAB data file *PlatformChemicals.mat*, which is located in ReNeGen's data directory.

Table G.1 – Main substrates currently available in ReNeGen

Argument	Substance
1	Glycerol
2	Lactic acid
3	3-Hydroxypropionic acid
4	Succinic acid
5	Furfural
6	Levulinic acid
7	Xylitol
8	2,5-Furandicarboxylic acid
9	Sorbitol
10	Itaconic acid
11	5-Hydroxymethylfurfural
12	Acetone

The targeted products are stated in SMILES notation in option 2.

Option 3 contains the reactants that are added to a substrate to form a molecular ensemble that undergoes reaction generation. The user can choose multiple reactants amongst the alternatives in Table G.2 by denoting the corresponding identifier. The available set of reactants can be extended by including further substances in the MATLAB data file *Reactants.mat*, which is located in ReNeGen's data directory.

Table G.2 – Reactants currently available in ReNeGen

Argument	Substance
-2	Main network substrate(s)
-1	Current substrate
0	None/Isomerization
1	Hydrogen
2	Acetone
3	Methanol
4	Ethanol
5	n-Butanol
6	CO_2
7	H_2O

Option 4 denotes the maximum number of reaction stages that are carried out by the generator. It can be any positive integer value.

Option 5 states whether reaction rules are included or not (arguments are 'yes'/'no'). The desired reaction rules are included by stating their activity (arguments are 0/1) in the vector *ReactionRules* in the main file.

Option 6 is used to state which thermophysical properties are calculated. The user can choose multiple amongst the properties presented in Table G.3.

Table G.3 – Thermophysical properties

Argument	Property	Unit
'MW	Molecular weight	[g/mole]
'CETANE'	Cetane number	[–]
'FP'	Flash point	[K]
'HCOM'	Enthalpy of combustion	[J/kmole]
'HVP'	Enthalpy of vaporization at 298K	[J/kmole]
'LDN'	Liquid density at 298K	[kmole/m]
'NBP'	Normal boiling point	[K]
'MP'	Melting point	[K]
'TP'	Toxicity potential	[mg/g]

Option 7 states the identifier criterion of the main reaction product. The available arguments are 'MW' for molecular weight, 'HCOM' for enthalpy of combustion and 'AS' for atomic share.

The default reaction yield is stated in option 8. This value is used in the network evaluation if no value from experimental investigations is available in ReNeGen's reaction database. This database is located in ReNeGen's data directory under the name *Yield_Database.xlsx*. Novel reactions can be added in SMILES notation; the substrates are denoted in the first column, the products in the second column and the reaction yield in the third column. The user has to provide the SMILES notation in the canonical representation of Open Babel. ReNeGen automatically compares the generated reactions to those available in the database and includes available data on reaction yields into the evaluation.

Option 9 states the settings of the reference process. The user specifies the reference scenario by providing a name argument (available arguments are 'Ethanol' and 'FAME') and a numeric production quantity in liters per year. The reference reaction network and process specifications are chosen based on the name argument.

Option 10 states which criteria are plotted after reaction network generation and evaluation. The user can choose amongst the criteria listed in Table G.4. Three diagrams can be created with two criteria plotted against each other in each diagram. The first stated criterion represents the abscissa, the second on the ordinate data.

Table G.4 – Evaluation criteria available for 2D plots

Argument	Property	Unit
'YIELD'	Total yield	[%]
'STEPS'	Number of reaction steps	[–]
'IC'	Investment cost	[$]
'AIC'	Annualized investment costs	[$/a]
'TAC'	Total annualized costs	[$/a]
'Pmax'	Max. feedstock price	[$/kg]
'Rmin'	Min. biofuel sales price	[$/l]
'TAR'	Total annual revenues	[$/a]
'EEC'	Energy efficiency	[%]
'EI'	Environmental impact	[–]
'CE'	Carbon efficiency	[%]
'AE'	Atom efficiency	[%]
'H2'	H_2 consumption	[mole/mole]

To restrict the generated network substances to a desired molecular and/or thermophysical space, constraints can be imposed on

- maximum atom count (in vector *COHMax*),
- minimum atom count(in vector *COHMax*),
- maximum number of rings(in vector *nRingsMax*),
- minimum number of rings (in vector *nRingsMin*),
- maximum ring size (in vector *nRingSizeMax*),
- minimum ring size (in vector *nRingSizeMin*), and
- property constraints on thermophysical properties chosen in option 4 (in vectors *PropMax* and *PropMin*).

The constitution of the provided biomass is required for the economic process evaluation and has to be denoted in the *Biomass_Comp* vector. This vector states the individual fractions of cellulose, hemicellulose and lignin.

The costs for the platform chemical(s) in $/t have to be stated in the vector *Prices*. Prices for reactants are automatically retrieved from the file *Reactants.mat* which is located in ReNeGen's data directory.

A report file will be generated in LATEX syntax that summarizes the settings and results of the network generation task and can be compiled into a pdf document.

The following example shows, in MATLAB syntax, the settings of the 2-BF and 2-BTHF synthesis case study presented in Chapter 6.

Example 21. *The task of the generation is a reaction network that contains the synthesis pathways from furfural to 2-BF and 2-BTHF. The options were set as following (provided in MATLAB syntax).*

Options{1} = [5];
Options{2} = [{'CCCCC1CCCO1'},{'CCCCc1ccco1'}];
Options{3} = [0,1,2];
Options{4} = 15;
Options{5} = 'Yes';
Options{6} = [{'MW'},{'CETANE'},{'HCOM'},{'TP'},{'LDN'},{'NBP'},{'MP'}];
Options{7} = [{'HCOM'}];
Options{8} = 0.97;
Options{9} = [{'FAME'},{190*10^6}];
Options{10} = [{'TAC'},{'IC'};{'YIELD'},{'IC'}];

The vector ReactionRules is set to:

$$ReactionRules = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 1\ 1];$$

The constraints of this reaction task are:

- Maximum atom count ($COHMax = [8\ 3\ 16]$),
- Minimum atom count($COHMax = [5\ 0\ 0]$),
- Maximum number of rings($nRingsMax = 1$),
- Minimum number of rings ($nRingsMin = 0$),
- Maximum ring size ($nRingSizeMax = 6$),
- Minimum ring size ($nRingSizeMin = 4$), and
- Property constraints on thermophysical properties chosen in option 4.

No property constraints are set, so the vectors PropMax *and* PropMin *are:*

$$PropMax = [\ \text{inf inf inf inf inf inf inf}];$$
$$PropMin = [\text{-inf -inf -inf -inf -inf -inf -inf}];$$

The biomass composition is:

$$Biomass\_Comp = [0.6\ 0.2\ 0.2];$$

The price for the platform chemical furfural, defined in option 2, is:

$$Prices = 1000;$$

Bibliography

Aden, A. and Foust, T. (2009). Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. *Cellulose*, 16(4):535–545.

Adrio, L. A. and Hii, K. K. (2011). An expedient synthesis of olfactory lactones by intramolecular hydroacylalkoxylation reactions. *Euro J Org Chem*, 2011(10):1852–1857.

Alibaba.com (2014a). `http://www.alibaba.com/showroom/glucose-price.html`. [Online, accessed 29.01.2014].

Alibaba.com (2014b). `http://www.alibaba.com/showroom/price-itaconic-acid.html`. [Online, accessed 29.01.2014].

Alibaba.com (2014c). `http://www.alibaba.com/showroom/furfural-price.html`. [Online, accessed 28.03.2014].

Alonso, D. M., Bond, J. Q., and Dumesic, J. A. (2010). Catalytic conversion of biomass to biofuels. *Green Chem*, 12(9):1493–1513.

Alvarez-Ibarra, C., Arias Perez, M. S., Fernandez, M. J., Serrano, D., and Sinisterra, V. (1992). Efficient Wittig-Horner and improved Claisen-Schmidt synthesis of acyclic α-enones with a 2-furyl or 3, 4-methylenedioxyphenyl group at the β-position. *J Chem Res-S*, 10(10):326–327.

American Chemical Society (2014). `https://scifinder.cas.org`. [Online, accessed 27.01.2014].

Anderson, R., Kölbel, H., and Rálek, M. (1984). *The Fischer-Tropsch Synthesis*, volume 16. Academic Press New York.

Bakshi, B. R. and Fiksel, J. (2003). The quest for sustainability: Challenges for process systems engineering. *AIChE J*, 49(6):1350–1358.

Barberis, F., Barone, R., and Chanon, M. (1996). HOLOWin: A fast way to search for tandem reactions with computer. Application to the taxane framework. *Tetrahedron*, 52(46):14625–14630.

Bartholomäus, R., Dommershausen, F., Thiele, M., Karanjule, N. S., Harms, K., and Koert, U. (2013). Total synthesis of the postulated structure of fulicineroside. *Chem-Eur J*, 19(23):7423–7436.

Baumlin, S., Broust, F., Bazer-Bachi, F., Bourdeaux, T., Herbinet, O., Toutie Ndiaye, F., Ferrer, M., and Lédé, J. (2006). Production of hydrogen by lignins fast pyrolysis. *Int J Hydrogen Energy*, 31(15):2179–2192.

Beale, A., Laughlin, M., Lidback, A., and Cooke, B. (2008). Acetone market report.

Besler, A., Harwardt, A., and Marquardt, W. (2009). Reaction networks A rapid screening method. In .

Bjerre, A. B., Olesen, A. B., Fernqvist, T., Plöger, A., and Schmidt, A. S. (1996). Pretreatment of wheat straw using combined wet oxidation and alkaline hydrolysis resulting in convertible cellulose and hemicellulose. *Biotechnology and bioengineering*, 49(5):568–577.

Blinov, M. L., Yang, J., Faeder, J. R., and Hlavacek, W. S. (2006). Graph theory for rule-based modeling of biochemical networks. In *Transactions on Computational Systems Biology VII*, pages 89–106. Springer.

Broadbelt, L. J., Stark, S. M., and Klein, M. T. (1994). Computer generated pyrolysis modeling: on-the-fly generation of species, reactions, and rates. *Ind Eng Chem Res*, 33(4):790–799.

Brooke, A., Kendrick, D., Meeraus, A., Raman, R., and America, U. (1998). The General Algebraic Modeling System. Technical report, GAMS Development Corporation.

Cai, C. M., Zhang, T., Kumar, R., and Wyman, C. E. (2014). Integrated furfural production as a renewable fuel and chemical platform from lignocellulosic biomass. *J Chem Technol Biot*, 89(1):2–10.

Campbell, P. K., Beer, T., and Batten, D. (2011). Life cycle assessment of biodiesel production from microalgae in ponds. *Bioresource technology*, 102(1):50–56.

Canakci, M. and Sanli, H. (2008). Biodiesel production from various feedstocks and their effects on the fuel properties. *J Ind Microbiol Biotechnol*, 35(5):431–441.

Cardona, C. A. and Sánchez, Ó. J. (2007). Fuel ethanol production: process design trends and integration opportunities. *Bioresource technology*, 98(12):2415–2457.

Chang, A.-F. and Liu, Y. (2009). Integrated process modeling and product design of biodiesel manufacturing. *Ind Eng Chem Res*, 49(3):1197–1213.

Cheresources.com (2014). `http://www.cheresources.com/invision/topic/19749-recent-cepci/`. [Online, accessed 29.01.2014].

Collet, P., Spinelli, D., Lardon, L., Hélias, A., Steyer, J.-P., and Bernard, O. (2013). Life-cycle assessment of microalgal-based biofuels. *Biofuels from algae*, pages 287–312.

Constantinou, L. and Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, 40(10):1697–1710.

Corey, E. and Wipke, W. T. (1969). Computer-assisted design of complex organic syntheses. *Sci*, 166:178–192.

Corey, E. J. and Cheng, X.-M. (1989). *The Logic of Chemical Synthesis*. John Wiley New York.

Corey, E. J., Wipke, W. T., Cramer III, R. D., and Howe, W. J. (1972). Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *J Am Chem Soc*, 94(2):421–430.

Dahmen, M., Hechinger, M., Victoria Villeda, J. J., and Marquardt, W. (2012). Towards model-based identification of biofuels for compression ignition engines. *SAE Int J Fuels Lubr*, 5(3):990–1003.

Dahmen, M., Victoria Villeda, J. J., and Marquardt, W. (2013). Refunctionalization of bio-based platform chemicals into novel biofuels: A computational approach. In *3rd International Conference on Sustainable Chemical Product and Process Engineering, Dalian, China, 27-30.05.2013*.

Daoutidis, P., Marvin, W. A., Rangarajan, S., and Torres, A. I. (2013). Engineering biomass conversion processes: a systems perspective. *AIChE Journal*, 59(1):3–18.

De Jong, W. and Marcotullio, G. (2010). Overview of biorefineries based on co-production of furfural, existing concepts and novel developments. *Int J Chem React Eng*, 8(1).

Deiser, O. (2004). *Einführung in die Mengenlehre*. Springer.

Di Maio, F. and Lignola, P. (1992). KING, a kinetic network generator. *Chem Eng Sci*, 47(9):2713–2718.

Diestel, R., editor (2006). *Graphentheorie*, volume 4. Springer.

Dry, M. E. (2002). High quality diesel via the fischer–tropsch process–a review. *Journal of Chemical Technology and Biotechnology*, 77(1):43–50.

Dugundji, J. and Ugi, I. (1973). An algebraic model of constitutional chemistry as a basis for chemical computer programs. In *Comput Chem*, pages 19–64. Springer.

Edwards, R., Mahieu, V., Griesemann, J.-C., Larive, J.-F., and Rickeard, D. J. (2004). Well-to-wheels analysis of future automotive fuels and powertrains in the European context. *SAE Trans*, 113(4):1072–1084.

Ellson, J., Gansner, E., Koutsofios, L., North, S. C., and Woodhull, G. (2002). Graphviz open source graph drawing tools. In *Graph Drawing*, pages 483–484. Springer.

Ethanol Producer Magazine (2014). `http://www.ethanolproducer.com/plants/listplants/US/Existing/Sugar-Starch/page:1/`. [Online, accessed 29.01.2014].

Faeder, J. R., Blinov, M. L., Goldstein, B., and Hlavacek, W. S. (2005). Rule-based modeling of biochemical networks. *Complexity*, 10(4):22–41.

Farag, H. (2010). Hydrodesulfurization of dibenzothiophene and 4, 6-dimethyldibenzothiophene over NiMo and CoMo sulfide catalysts: Kinetic modeling approach for estimating selectivity. *J Colloid Interface Sci*, 348(1):219–226.

Fearnside, P. M. (2000). Global warming and tropical land-use change: Greenhouse gas emissions from biomass burning, decomposition and soils in forest conversion, shifting cultivation and secondary vegetation. *Clim Chang*, 46(1-2):115–158.

Fenves, S. J. (1967). *Computer Methods in Civil Engineering*. Prentice-Hall Inc.

Fernando, S., Adhikari, S., Chandrapal, C., and Murali, N. (2006). Biorefineries: Current status, challenges, and future direction. *Energy & Fuels*, 20(4):1727–1737.

Figueras, J. (1993). Morgan revisited. *J Chem Inf Comput Sci*, 33(5):717–718.

Fiksel, J. (2002). Sustainable development through industrial ecology. In Lankey, R. and Anastas, P., editors, *Advancing Sustainability through Green Chemistry and Engineering, American Chemical Society*.

Fischer, R., Frank, J., Henkelmann, J., Merger, F., Ruehl, T., Siegel, H., and Weyer, H.-J. (1996). Preparation of 2-methyl-1, 4-butanediol and 3-methyltetrahydrofuran. US Patent 5,536,854.

FOA (2010). Bioenery and food security. Technical report, United Nations.

Fontain, E. (1995). *Kombinatorik und chemische Metrik formaler Reaktions- und Strukturgenerierung*. PhD thesis, Technische Universität München.

Fontain, E. and Reitsam, K. (1991). The generation of reaction networks with RAIN. 1. The reaction generator. *J Chem Inf Comput Sci*, 31(1):96–101.

Ford, L. and Fulkerson, D. R. (1962). *Flows in Networks*, volume 3. Princeton University Press.

Furniss, B. S., Hannaford, A., Smith, P., and Tatchell, A. (1989). *Vogels textbook of practical organic chemistry*. Prentice Hall.

Gagneur, J. and Klamt, S. (2004). Computation of elementary modes: A unifying framework and the new binary approach. *BMC Bioinformatics*, 5(1):175.

Gani, R. and Pistikopoulos, E. N. (2002). Property modelling and simulation for product and process design. *Fluid Phase Equilib*, 194:43–59.

Garcia, D. J. and You, F. (2015). Multiobjective optimization of product and process networks: General modeling framework, efficient global optimization algorithm, and case studies on bioconversion. *AIChE Journal*, 61(2):530–554.

Gasteiger, J. and Jochum, C. (1978). EROS: A computer program for generating sequences of reactions. In *Organic Compunds*, pages 93–126. Springer.

Geilen, F., Engendahl, B., Harwardt, A., Marquardt, W., Klankermayer, J., and Leitner, W. (2010). Selective and flexible transformation of biomass-derived platform chemicals by a multifunctional catalytic system. *Angew Chem Ger Edit*, 122(32):5642–5646.

Gelernter, H., Sanders, A., Larsen, D., Agarwal, K., Boivie, R., Spritzer, G., and Searleman, J. (1977). Empirical explorations of SYNCHEM. *Sci*, 197(4308):1041–1049.

Gong, J. and You, F. (2014). Global optimization for sustainable design and synthesis of algae processing network for co2 mitigation and biofuel production using life cycle optimization. *AIChE Journal*, 60(9):3195–3210.

Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C., and Wassermann, A. (2012). MOLGEN 5.0, a molecular structure generator. In *Advances in Mathematical Chemistry*.

Gui, M. M., Lee, K., and Bhatia, S. (2008). Feasibility of edible oil vs. non-edible oil vs. waste edible oil as biodiesel feedstock. *Energy*, 33(11):1646–1653.

Halleux, H., Lassaux, S., Renzoni, R., and Germain, A. (2008). Comparative life cycle assessment of two biofuels ethanol from sugar beet and rapeseed methyl ester. *The International Journal of Life Cycle Assessment*, 13(3):184–190.

Hatzimanikatis, V., Li, C., Ionita, J. A., and Broadbelt, L. J. (2004). Metabolic networks: Enzyme function and metabolite structure. *Curr Opin Struc Biol*, 14(3):300–306.

Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D., and Broadbelt, L. J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics*, 21(8):1603–1609.

He, P., Liu, X., Zheng, H., Li, W., Lin, L., and Feng, X. (2012). Asymmetric 1, 2-Reduction of Enones with Potassium Borohydride Catalyzed by Chiral N, N'-Dioxide–Scandium (III) Complexes. *Organic letters*, 14(19):5134–5137.

Hechinger, M., Dahmen, M., Victoria Villeda, J., and Marquardt, W. (2012). Rigorous generation and model-based selection of future biofuel candidates. In *Proceedings of the 11th International Symposium on Process Systems Engineering*.

Hechinger, M., Voll, A., and Marquardt, W. (2010). Towards an integrated design of biofuels and their production pathways. *Comput Chem Eng*, 34(12):1909–1918.

Heller, S. and McNaught, A. (2009). The IUPAC international chemical identifier (InChI). *Chem Inter*, 31(1):7–9.

Hendrickson, J. (1990). The SYNGEN approach to synthesis design. *Anal Chim Acta*, 235:103–113.

Hill, A. D., Tomshine, J. R., Weeding, E. M., Sotiropoulos, V., and Kaznessis, Y. N. (2008). SynBioSS: The synthetic biology modeling suite. *Bioinformatics*, 24(21):2551–2553.

Holmström, K. (1999). The TOMLAB optimization environment in Matlab.

Hoppe, F., Heuser, B., Thewes, M., Kremer, F., Pischinger, S., Dahmen, M., Hechinger, M., and Marquardt, W. (2016). Tailor-made fuels for future engine concepts. *International Journal of Engine Research*, 17(1):16–27.

Hsu, S., Krishnamurthy, B., Rao, P., Zhao, C., Jagannathan, S., and Venkatasubramanian, V. (2008). A domain-specific compiler theory based framework for automated reaction network generation. *Comput Chem Eng*, 32(10):2455–2470.

Huang, H.-J., Ramaswamy, S., Al-Dajani, W., Tschirner, U., and Cairncross, R. A. (2009). Effect of biomass species and plant size on cellulosic ethanol: a comparative process and economic analysis. *Biomass and Bioenergy*, 33(2):234–246.

Huang, K., Zhang, X., Emge, T. J., Hou, G., Cao, B., and Zhang, X. (2010). Design and synthesis of a novel three-hindered quadrant bisphosphine ligand and its application in asymmetric hydrogenation. *Chem Commun*, 46(45):8555–8557.

Hukkerikar, A. S., Kalakul, S., Sarup, B., Young, D. M., Sin, G., and Gani, R. (2012). Estimation of environment-related properties of chemicals for design of sustainable processes: Development of group-contribution+ (GC+) property models and uncertainty analysis. *J Chem Inf Comput Sci*, 52(11):2823–2839.

Hunt, R. G., Franklin, W. E., and Hunt, R. (1996). LCAHow it came about. *The international journal of life cycle assessment*, 1(1):4–7.

Index mundi (2014). `http://www.indexmundi.com/de/rohstoffpreise/?ware=rapsol`. [Online, accessed 29.01.2014].

Jaksland, C. A., Gani, R., and Lien, K. M. (1995). Separation process design and synthesis based on thermodynamic insights. *Chem Eng Sci*, 50(3):511–530.

James, L. (1993). *Nobel Laureates in Chemistry, 1901-1992*, volume 1. Chemical Heritage Foundation.

Janssen, A. J., Kremer, F. W., Baron, J. H., Muether, M., Pischinger, S., and Klankermayer, J. (2011). Tailor-made fuels from biomass for homogeneous low-temperature diesel combustion. *Energy & Fuels*, 25(10):4734–4744.

Jeliazkova, N. and Kochev, N. (2011). AMBIT-SMARTS: Efficient searching of chemical structures and fragments. *Mol Inf*, 30(8):707–720.

Joback, K. G. and Reid, R. C. (1987). Estimation of pure-component properties from group-contributions. *Chem Eng Commun*, 57(1-6):233–243.

Johnson, A. P. and Marshall, C. (1992a). Starting material oriented retrosynthetic analysis in the LHASA program. 2. Mapping the SM and target structures. *J Chem Inf Comput Sci*, 32(5):418–425.

Johnson, A. P. and Marshall, C. (1992b). Starting material oriented retrosynthetic analysis in the LHASA program. 3. Heuristic estimation of synthetic proximity. *J Chem Inf Comput Sci*, 32(5):426–429.

Johnson, A. P., Marshall, C., and Judson, P. N. (1992). Starting material oriented retrosynthetic analysis in the LHASA program. 1. General description. *J Chem Inf Comput Sci*, 32(5):411–417.

Julis, J. and Leitner, W. (2012). Synthesis of 1-Octanol and 1, 1-Dioctyl Ether from Biomass-Derived Platform Chemicals. *Angew Chem Int Edit*, 51(34):8615–8619.

Kadla, J., Kubo, S., Venditti, R., Gilbert, R., Compere, A., and Griffith, W. (2002). Lignin-based carbon fibers for composite fiber applications. *Carbon*, 40(15):2913–2920.

Kamm, B., Kamm, M., and Gruber, P. R. (2008). *Biorefineries - Industrial processes and products: Status quo and future directions*. Wiley-VCH.

Katritzky, A. R. and Fara, D. C. (2005). How chemical structure determines physical, chemical, and technological properties: An overview illustrating the potential of quantitative structure-property relationships for fuels science. *Energy & Fuels*, 19(3):922–935.

Khan, S. S., Zhang, Q., and Broadbelt, L. J. (2009). Automated mechanism generation. Part 1: Mechanism development and rate constant estimation for VOC chemistry in the atmosphere. *J Atmos Chem*, 63(2):125–156.

Kim, H., Kim, S., and Dale, B. E. (2009). Biofuels, land use change, and greenhouse gas emissions: Some unexplored variables. *Environ Sci Technol*, 43(3):961–967.

Kim, J., Sen, S. M., and Maravelias, C. T. (2013). An optimization-based assessment framework for biomass-to-fuel conversion strategies. *Energy & Environmental Science*, 6(4):1093–1104.

Kim, S. and Dale, B. E. (2005). Life cycle assessment of various cropping systems utilized for producing biofuels: Bioethanol and biodiesel. *Biomass and Bioenergy*, 29(6):426–439.

Kinast, J. (2003). Production of biodiesels from multiple feedstocks and properties of biodiesels and biodiesel/diesel blends. Technical report, National Renewable Energy Laboratory.

Klement, T., Milker, S., Jäger, G., Grande, P. M., Domínguez de María, P., and Büchs, J. (2012). Biomass pretreatment affects Ustilago maydis in producing itaconic acid. *Microb Cell Fact*, 11:43.

Knothe, G., Van Gerpen, J. H., Krahl, J., et al. (2005). *The biodiesel handbook*, volume 1. AOCS press Champaign, IL.

Kontogeorgis, G. M. and Folas, G. K. (2009). *Thermodynamic Models for Industrial Applications: From Classical and Advanced Mixing Rules to Association Theories*. John Wiley & Sons.

Kowalik, M., Gothard, C. M., Drews, A. M., Gothard, N. A., Weckiewicz, A., Fuller, P. E., Grzybowski, B. A., and Bishop, K. J. (2012). Parallel optimization of synthetic pathways within the network of organic chemistry. *Angewandte Chemie International Edition*, 51(32):7928–7932.

Krohn, K. and Riaz, M. (2004). Total synthesis of (+)-xyloketal D, a secondary metabolite from the mangrove fungus Xylaria sp. *Tetrahedron Lett*, 45(2):293–294.

Kruse, T. M., Woo, O. S., Wong, H.-W., Khan, S. S., and Broadbelt, L. J. (2002). Mechanistic modeling of polymer degradation: A comprehensive study of polystyrene. *Macromol*, 35(20):7830–7844.

Kumar, D. and Murthy, G. S. (2012). Life cycle assessment of energy and GHG emissions during ethanol production from grass straws using various pretreatment processes. *The International Journal of Life Cycle Assessment*, 17(4):388–401.

Lange, J.-P. (2001). Fuels and chemicals manufacturing; guidelines for understanding and minimizing the production costs. *Cattech*, 5(2):82–95.

Lange, J.-P., van der Heide, E., van Buijtenen, J., and Price, R. (2012). Furfurala promising platform for lignocellulosic biofuels. *ChemSusChem*, 5(1):150–166.

Lapkin, A. and Constable, D. (2008). *Green chemistry metrics*. Wiley-Blackwell.

Lardon, L., Helias, A., Sialve, B., Steyer, J.-P., and Bernard, O. (2009). Life-cycle assessment of biodiesel production from microalgae. *Environmental science & technology*, 43(17):6475–6481.

Lee, S., Phalakornkule, C., Domach, M. M., and Grossmann, I. E. (2000). Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput Chem Eng*, 24(2):711–716.

Li, K., Liu, S., and Liu, X. (2014). An overview of algae bioethanol production. *International Journal of Energy Research*, 38(8):965–977.

Liguras, D. K. and Allen, D. T. (1989a). Structural models for catalytic cracking. 1. Model compound reactions. *Ind Eng Chem Res*, 28(6):665–673.

Liguras, D. K. and Allen, D. T. (1989b). Structural models for catalytic cracking. 2. Reactions of simulated oil mixtures. *Ind Eng Chem Res*, 28(6):674–683.

Lozowski, D. (2012). Chemical engineering plant cost index (CEPCI). *Chem Eng-New York*, 119:84.

Ma, F. and Hanna, M. A. (1999). Biodiesel production: A review. *Bioresource Technol*, 70(1):1–15.

Marquardt, W., Harwardt, A., Hechinger, M., Kraemer, K., Viell, J., and Voll, A. (2010). The biorenewables opportunity-toward next generation process and product systems. *AIChE J*, 56(9):2228–2235.

Marvin, W. A., Rangarajan, S., and Daoutidis, P. (2013). Automated generation and optimal selection of biofuel-gasoline blends and their synthesis routes. *Energy & Fuels*, 27:3585–3594.

MATLAB (2010). *Version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

Mayeno, A. N., Yang, R. S., and Reisfeld, B. (2005). Biochemical reaction network modeling: Predicting metabolism of organic chemical mixtures. *Environ Sci Technol*, 39(14):5363–5371.

McDermott, J. B., Libanati, C., LaMarca, C., and Klein, M. T. (1990). Quantitative use of model compound information: Monte Carlo simulation of the reactions of complex macromolecules. *Ind Eng Chem Res*, 29(1):22–29.

McDonough, T. J. (1992). The chemistry of organosolv delignification. Technical report, Institute of Paper Science and Technology Atlanta, Georgia.

Methanex (2014). `http://www.methanex.com/products/methanolprice.html`. [Online, accessed 29.01.2014].

Metz, B. (2007). *Climate change 2007 - Mitigation of climate change: Working Group III Contribution to the fourth assessment report of the IPCC*, volume 4. Cambridge University Press.

Midgley, G. and Thomas, C. B. (1987). Selectivity of radical formation in the reaction of carbonyl compounds with manganese (III) acetate. *J Chem Soc, Perkin Trans 2*, 16(8):1103–1108.

Moity, L., Molinier, V., Benazzouz, A., Barone, R., Marion, P., and Aubry, J.-M. (2014). In silico design of bio-based commodity chemicals: application to itaconic acid based solvents. *Green Chem*, 16:146–160.

Morgan, H. (1965). The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. *J Chem Doc*, 5(2):107–113.

Mori, K. (2008). Synthesis of the (5S,9R)-isomer of 5, 9-dimethylpentadecane, the major component of the female sex pheromone of the coffee leaf miner moth, Leucoptera coffeella. *Tetrahedron-Asymmetr*, 19(7):857–861.

Naik, S., Goud, V. V., Rout, P. K., and Dalai, A. K. (2010). Production of first and second generation biofuels: a comprehensive review. *Renew Sust Energ Rev*, 14(2):578–597.

OBoyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J Cheminf*, 3(1):1–14.

OECD/FAO (2011). OECD-FAO agricultural outlook 2011-2020. Technical report, OECD Publishing and FAO.

Oh, J., Dash, S., and Lee, H. (2011). Selective conversion of glycerol to 1, 3-propanediol using Pt-sulfated zirconia. *Green Chem*, 13(8):2004–2007.

Okabe, M., Lies, D., Kanamasa, S., and Park, E. Y. (2009). Biotechnological production of itaconic acid and its biosynthesis in Aspergillus terreus. *Appl Microbiol Biot*, 84(4):597–606.

Olah, G. A., Fung, A. P., and Malhotra, R. (1981). Synthetic methods and reactions; 991. Preparation of cyclic ethers over superacidic perfluorinated resinsulfonic acid (Nafion-H) catalyst. *Synth*, 1981(06):474–476.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends in biotechnology*, 22(8):400–405.

Post, W. M., Peng, T.-H., Emanuel, W. R., King, A. W., Dale, V. H., DeAngelis, D. L., et al. (1990). The global carbon cycle. *Am Sci*, 78(4):310–326.

Prickett, S. and Mavrovouniotis, M. (1997a). Construction of complex reaction systems–I. Reaction description language. *Comput Chem Eng*, 21(11):1219–1235.

Prickett, S. and Mavrovouniotis, M. (1997b). Construction of complex reaction systems–II. Molecule manipulation and reaction application algorithms. *Comput Chem Eng*, 21(11):1237–1254.

Prickett, S. and Mavrovouniotis, M. (1997c). Construction of complex reaction systems–III. An example: Alkylation of olefins. *Comput Chem Eng*, 21(12):1325–1337.

Protocol, K. (1997). United Nations framework convention on climate change. Technical report, United Nations.

Quann, R. J. and Jaffe, S. B. (1992). Structure-oriented lumping: Describing the chemistry of complex hydrocarbon mixtures. *Ind Eng Chem Res*, 31(11):2483–2497.

Rangarajan, S., Bhan, A., and Daoutidis, P. (2010). Rule-based generation of thermochemical routes to biomass conversion. *Ind Eng Chem Res*, 49(21):10459–10470.

Rangarajan, S., Bhan, A., and Daoutidis, P. (2012a). Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Comput Chem Eng*, 46:141–152.

Rangarajan, S., Bhan, A., and Daoutidis, P. (2012b). Language-oriented rule-based reaction network generation and analysis: Description of RING. *Comput Chem Eng*, 45:114–123.

Ratkiewicz, A. and Truong, T. N. (2003). Application of chemical graph theory for automated mechanism generation. *J Chem Inf Comput Sci*, 43(1):36–44.

Ratkiewicz, A. and Truong, T. N. (2006). Automated mechanism generation: From symbolic calculation to complex chemistry. *Int J Quantum Chem*, 106(1):244–255.

Robert, C., de Montigny, F., and Thomas, C. M. (2011). Tandem synthesis of alternating polyesters from renewable resources. *Nat Commun*, 2:586.

Rowley, R., Wilding, W., Oscarson, J., Zundel, N., Marshall, T., Daubert, T., and Danner, R. (2003). DIPPR data compilation of pure compound properties.

Ruth, M. (2011). Hydrogen production cost estimate using biomass gasification. Technical report, National Renewable Energy Laboratory.

Saling, P., Kicherer, A., Dittrich-Krämer, B., Wittlinger, R., Zombik, W., Schmidt, I., Schrott, W., and Schmidt, S. (2002). Eco-efficiency analysis by BASF: The method. *Int J Life Cycle Assess*, 7(4):203–218.

Sanders, J., Scott, E., Weusthuis, R., and Mooibroek, H. (2007). Bio-Refinery as the bio-inspired process to bulk chemicals. *Macromol Biosci*, 7(2):105–117.

Sari, R. and Soytas, U. (2007). The growth of income and energy consumption in six developing countries. *Energy Policy*, 35(2):889–898.

Sarkar, N., Ghosh, S. K., Bannerjee, S., and Aikat, K. (2012). Bioethanol production from agricultural wastes: An overview. *Renewable Energy*, 37(1):19–27.

Schilling, C. H., Letscher, D., and Palsson, B. Ø. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248.

Scott-Cato, M. (2009). *Green economics: An introduction to theory, policy, and practice.* Earthscan.

Services, G. S. (2013). Indigo: Universal cheminformatics toolkit. `http://ggasoftware.com/opensource/indigo`. [Online, accessed 27.04.201].

Song, J. (2004). *Building robust chemical reaction mechanisms: Next generation of automatic model construction software.* PhD thesis, Massachusetts Institute of Technology.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nat*, 420(6912):190–193.

Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. *Metab Eng*, 1(1):1–11.

Terzer, M. and Stelling, J. (2008). Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235.

Todeschini, R. and Consonni, V. (2008). *Handbook of molecular descriptors*. Wiley. com.

Tomlin, A. S., Turányi, T., and Pilling, M. J. (1997). Mathematical tools for the construction, investigation and reduction of combustion mechanisms. *Compr Chem Kinet*, 35:293–437.

Ugi, I., Bauer, J., Brandt, J., Friedrich, J., Gasteiger, J., Jochum, C., and Schubert, W. (1979). New applications of computers in chemistry. *Angew Chem Int Edit*, 18(2):111–123.

Varma, A. and Palsson, B. O. (1994). Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/technology*, 12:994–998.

Victoria Villeda, J., Dahmen, M., Hechinger, M., Voll, A., and Marquardt, W. (2012). Towards model-based design of biofuel value chains. *Curr Opin Chem Eng*, 1:465–471.

Viêgas, C. V., Hachemi, I., Freitas, S. P., Mäki-Arvela, P., Aho, A., Hemming, J., Smeds, A., Heinmaa, I., Fontes, F. B., da Silva Pereira, D. C., et al. (2015). A route to produce renewable diesel from algae: Synthesis and characterization of biodiesel via in situ transesterification of chlorella alga and its catalytic deoxygenation to renewable diesel. *Fuel*, 155:144–154.

Voll, A. (2013). *Model-based screening of reaction pathways for biorenewables processing*. PhD thesis, RWTH Aachen University.

Voll, A. and Marquardt, W. (2012a). Benchmarking of next-generation biofuels from a process perspective. *Biofuel Bioprod Bior*, 6(3):292–301.

Voll, A. and Marquardt, W. (2012b). Reaction network flux analysis: Optimization-based evaluation of reaction pathways for biorenewables processing. *AIChE J*, 58(6):1788–1801.

vom Stein, T., Grande, P. M., Kayser, H., Sibilla, F., Leitner, W., and de María, P. D. (2011). From biomass to feedstock: One-step fractionation of lignocellulose components

by the selective organic acid-catalyzed depolymerization of hemicellulose in a biphasic system. *Green Chem*, 13(7):1772–1777.

Wagner, C. and Urbanczik, R. (2005). The geometry of the flux cone of a metabolic network. *Biophys J*, 89(6):3837–3845.

Waidmann, C. R., Pierpont, A. W., Batista, E. R., Gordon, J. C., Martin, R. L., West, R. M., Wu, R., et al. (2013). Functional group dependence of the acid catalyzed ring opening of biomass derived furan rings: an experimental and theoretical study. *Catal Sci Tech*, 3(1):106–115.

Wang, M. Q. (1999). GREET 1.5-transportation fuel-cycle model-Vol. 1: methodology, development, use, and results. Technical report, Argonne National Lab., IL (US).

Warth, V., Battin-Leclerc, F., Fournet, R., Glaude, P.-A., Côme, G.-M., and Scacchi, G. (2000). Computer based generation of reaction mechanisms for gas-phase oxidation. *Comput Chem*, 24(5):541–560.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28(1):31–36.

Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci*, 29(2):97–101.

Werpy, T., Petersen, G., Aden, A., Bozell, J., Holladay, J., White, J., Manheim, A., Eliot, D., Lasure, L., and Jones, S. (2004). Top value added chemicals from biomass. Volume 1-Results of screening for potential candidates from sugars and synthesis gas. Technical report, DTIC Document.

Wimmer, T., Busse, S., Hamacher, A., and der RWTH Aachen, P. (2010). Erste Ergebnisse der Exzellenzinitiative. *RWTH-Themen / Berichte aus der Rheinisch-Westfälischen Technischen Hochschule Aachen*, 1:14–25.

Wipke, W., Braun, H., Smith, G., Choplin, F., and Sieber, W. (1977). SECS-Simulation and evaluation of chemical synthesis: Strategy and planning. In *ACS Symp. Ser*, volume 61, pages 97–125. ACS Publications.

Wittmann, C. and Heinzle, E. (2001). Modeling and Experimental Design for Metabolic Flux Analysis of Lysine-Producing Corynebacteria by Mass Spectrometry. *Metab Eng*, 3(2):173–191.

Wong, H.-W., Li, X., Swihart, M. T., and Broadbelt, L. J. (2004). Detailed kinetic modeling of silicon nanoparticle formation chemistry via automated mechanism generation. *J Phys Chem A*, 108(46):10122–10132.

Xie, X., Wang, M., and Han, J. (2011). Assessment of fuel-cycle energy use and greenhouse gas emissions for Fischer-Tropsch diesel from coal and cellulosic biomass. *Environmental science & technology*, 45(7):3047–3053.

Yamashita, M., Kato, Y., and Suemitsu, R. (1980). Selective reduction of. ALPHA.,. BETA.-unsaturated carbonyl compounds by sodium hydrotelluride. *Chem Lett*, 7(7):847–848.

Yan, X., Inderwildi, O. R., and King, D. A. (2010). Biofuels and synthetic fuels in the US and China: A review of well-to-wheel energy use and greenhouse gas emissions with the impact of land-use change. *Energy Environ Sci*, 3(2):190–197.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., et al. (2011). Metabolic engineering of Escherichia coli for direct production of 1, 4-butanediol. *Nat Chem Biol*, 7(7):445–452.

Zakzeski, J., Bruijnincx, P. C., Jongerius, A. L., and Weckhuysen, B. M. (2010). The catalytic valorization of lignin for the production of renewable chemicals. *Chem Rev*, 110(6):3552–3599.

Zhao, Q., Curran, D. P., Malacria, M., Fensterbank, L., Goddard, J.-P., and Lacôte, E. (2011). N-Heterocyclic Carbene-Catalyzed Hydrosilylation of Styryl and Propargylic Alcohols with Dihydrosilanes. *Chem-Eur J*, 17(36):9911–9914.

Online-Buchshop für Ingenieure

VDI nachrichten

BUCHSHOP

Online-Shops

Fachliteratur und mehr - jetzt bequem online recherchieren & bestellen unter: www.vdi-nachrichten.com/Der-Shop-im-Ueberblick

Täglich aktualisiert: Neuerscheinungen VDI-Schriftenreihen

Im Buchshop von vdi-nachrichten.com finden Ingenieure und Techniker ein speziell auf sie zugeschnittenes, umfassendes Literaturangebot.

Mit der komfortablen Schnellsuche werden Sie in den VDI-Schriftenreihen und im Verzeichnis lieferbarer Bücher unter 1.000.000 Titeln garantiert fündig.

Im Buchshop stehen für Sie bereit:

VDI-Berichte und die Reihe **Kunststofftechnik:**

Berichte nationaler und internationaler technischer Fachtagungen der VDI-Fachgliederungen

Fortschritt-Berichte VDI:

Dissertationen, Habilitationen und Forschungsberichte aus sämtlichen ingenieurwissenschaftlichen Fachrichtungen

Newsletter „Neuerscheinungen":

Kostenfreie Infos zu aktuellen Titeln der VDI-Schriftenreihen bequem per E-Mail

Autoren-Service:

Umfassende Betreuung bei der Veröffentlichung Ihrer Arbeit in der Reihe Fortschritt-Berichte VDI

Buch- und Medien-Service:

Beschaffung aller am Markt verfügbaren Zeitschriften, Zeitungen, Fortsetzungsreihen, Handbücher, Technische Regelwerke, elektronische Medien und vieles mehr – einzeln oder im Abo und mit weltweitem Lieferservice

Die Reihen der Fortschritt-Berichte VDI:

1 Konstruktionstechnik/Maschinenelemente
2 Fertigungstechnik
3 Verfahrenstechnik
4 Bauingenieurwesen
5 Grund- und Werkstoffe/Kunststoffe
6 Energietechnik
7 Strömungstechnik
8 Mess-, Steuerungs- und Regelungstechnik
9 Elektronik/Mikro- und Nanotechnik
10 Informatik/Kommunikation
11 Schwingungstechnik
12 Verkehrstechnik/Fahrzeugtechnik
13 Fördertechnik/Logistik
14 Landtechnik/Lebensmitteltechnik
15 Umwelttechnik
16 Technik und Wirtschaft
17 Biotechnik/Medizintechnik
18 Mechanik/Bruchmechanik
19 Wärmetechnik/Kältetechnik
20 Rechnerunterstützte Verfahren (CAD, CAM, CAE CAQ, CIM ...)
21 Elektrotechnik
22 Mensch-Maschine-Systeme
23 Technische Gebäudeausrüstung

ISBN 978-3-18-395003-4