eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Exploring 'minimal metabolism' as a central concept for understanding the origins of life and developing systems biology

Nino Lauber

A thesis presented within the Doctoral Programme of
Molecular Biology and Biomedicine

supervised by
Prof. Kepa Ruiz-Mirazo & Prof. Christoph Flamm

Universidad del País Vaso/Euskal Herriko Unibersitatea

14.07.2023

Contents

1 Introduction **25**

1.1 History of origins of life research 25

1.2 A systems approach to origins of life research 29

1.3 Thesis structure 32

1.4 Main hypotheses and objectives 33

I From Metabolism to Minimal Metabolism **35**

2 Metabolism **37**

2.1 Metabolic Networks 37

2.2 Metabolism in extant biology 40

2.2.1 Kinetic and variability control 40

2.2.2 Spatial and energetic control 42

3 Minimal Metabolism **45**

3.1 Previous models 45

3.2 Critical review of metabolism models 51

3.3 Tentative scheme for minimal metabolism 53

4 Methodological challenges **57**

4.1 In vitro, experimental challenges 57

4.2 In silico challenges 59

4.3 Proposal for a combined workflow 60

II Complex Reaction Networks and Simple Boundary Con-

ditions **63**

5 Prebiotic Carbon Fixation **65**
5.1 Autotrophy at the origins of life 66
5.2 Carbon fixation pathways 68
5.3 Autocatalysis 72
5.3.1 Catalytic and autocatalytic reaction cycles 72
5.3.2 Autocatalytic reaction networks 76

6 Expansion of Chemical Reaction Spaces **81**
6.1 Rule based network expansion 82
6.1.1 Formalization of molecules 82
6.1.2 Formalization of chemical reactions 83
6.1.3 Formalization of chemistries 85
6.1.4 Derivation graphs 87
6.2 Sub-graph search 91
6.3 Thermochemistry 95

7 Exploring the rTCA Chemistry **101**
7.1 Formulation of an rTCA grammar 101
7.1.1 Base activator grammar 103
7.1.2 Relaxed activator grammar 105
7.1.3 Activator independent grammar 106
7.2 Strategies and DG expansions 107
7.3 AC cycle search: main results 109
7.3.1 Network properties 110
7.3.2 Cycle energies 115

III Simple Reaction Networks under Complex Boundary Conditions **121**

8 Membraneless Compartments as Protocells **123**
8.1 Compartments and proto-cells 123
8.2 Macromolecular crowding 125
8.3 Statistical mechanic of LLPS 128

9 Lattice Model Simulations **133**
9.1 Simulation of LLPS 133
9.1.1 Initializing the lattice 134
9.1.2 Equilibrating the lattice 135
9.2 Reaction-diffusion model 139

10 Modelling Phase Separation of Catalysts **145**
10.1 Cascade reaction 146
10.2 Influence of LLPS on reaction time 148
10.2.1 Dependence on substrate-catalyst affinity 149
10.2.2 Dependence on the pathway length 150
10.2.3 Dependence on the homogeneity of the condensate . . 153
10.3 Influence of LLPS on reaction kinetics 154
10.3.1 Reaction kinetics in a homogeneous PS system 155
10.3.2 Phase separation process suppresses feedback-induced oscillations 157

IV Discussion & Conclusion **161**

11 Discussion **163**
11.1 rTCA space exploration 163
11.1.1 Analysis of results 164
11.1.2 Limitations 166
11.1.3 Outlook 168
11.2 Lattice model 169
11.2.1 Analysis of results 170
11.2.2 Limitations 172
11.2.3 Outlook 173
11.3 General outlook & perspective 175

12 Conclusion **179**

Bibliography **180**

Appendices **203**

A Graph Rewrite Rules **203**

B Lattice Model Calculations **211**

Acknowledgements

First and foremost I want to thank my supervisors Kepa Ruiz-Mirazo and Christoph Flamm for accepting me as their PhD student and, during this time, being available for questions and discussion, but also for help and advice, in general. I also want to thank my colleagues from the IAS research group and the Theoretical/Protocell Biophysics Laboratory, for the fruitful exchange of ideas, as well as the Department of Philosophy and the Biofisika Institute, for providing a nice and welcoming work environment. Furthermore, I want to thank my colleagues from the ProtoMet project for giving me the opportunity to work in an international, interdisciplinary research project. On these lines, I especially want to thank T-Y Dora Tang and Joseph Moran for hosting me in their respective research groups and, through this, giving me the opportunity to gain further experience. In relation to the lattice model project, I want to thank my colleague Daniele De Martino who not only initialized this project but also played a substantial role in managing and overseeing its subsequent development. I furthermore want to thank Ondřej Ticháček whose work during his visit to our institute brought a major contribution to this project and my colleague Krishnadev Narayanankutty for helping out and contributing during the final stage of the project. In relation to the MØD project, I want to thank Phillip Honegger for providing me with the code for his autocatalysis search tool, as well as being in general available for questions and explanations concerning its usage, etc. On a more personal note, I want to thank Patrick Braun for his initial help with the MØD software and his constant feedback and support, later on, during the writing phase of this thesis. I also want to thank Bernat Corominas-Murtra, who basically kickstarted this whole adventure when he introduced me to Kepa. Last but not least, I want to thank my family: my parents Johann Lauber and Ingrid Lauber-Pils for their continuous supporting without which I would not have made it that far and my two older sisters, Jana Lauber and Simone Lauber who have always been an inspiration for me to keep pushing forward, no matter what.

Abstract

Origins of life (OL) research, as the name suggests, is concerned with the question of how cellular organisms could have develop, starting from complex chemical reaction systems, i.e. the transition from geo-chemistry to bio-chemistry. This vast and interdisciplinary research field has been dominated for many years by rather reductionistic approaches, that were mainly searching for a primordial type of biopolymer. However, in recent years, a shift towards a more *systems* oriented view of life and its origins has taken place, spurred by the emergence of fields like 'systems biology' and 'systems chemistry'. Rather than focusing on a single type of biomolecule or biochemical reaction pathway, these new approaches target the way in which molecular components and transformation processes are organized within biological (and proto-biological) systems. On these lines, the motivation behind this dissertation is to introduce the idea of 'minimal metabolism' (MM) as a heuristic construct that lies at the interface between chemistry and biology. So this theoretical proposal should not be taken as a final or complete characterization of metabolism, but as a particular contemporary way of thinking about it to encourage further research. With that aim, an attempt is made at modelling such a minimal metabolic system. This is achieved through the implementation of two different types of computational models, each focusing on a certain aspect associated with MM. The results of these two modelling approaches are then compared and critically discussed. Finally, a general outlook and perspective for future research on MM, and OL in general, is given.

Chapter 1

Introduction

The question concerning how life initially formed on Earth is perhaps one of the oldest and most fundamental questions that preoccupied humanity (next to questions about the origin of the universe or the meaning and purpose of life, etc.). In general, it is known that all nowadays existing species developed from common ancestors through evolution by natural selection, branching into more and more diversified organisms. Therefore, there must have been something like a last universal common ancestor (LUCA) that lies at the root of this phylogenetic tree (see Fig. 1.1). Furthermore, through various kinds of indirect evidences (like geological- or paleobiological data) it is possible to date the appearance of LUCA to roughly 3.5 billion years ago. Thus, the current state of research provides a solid general framework to support that life developed on Earth, although the exact details are still unknown[1]. As such, there is still no consensus on what type of organism LUCA was (if it was a full fledged organism to begin with). Moreover, the exact processes of how LUCA developed (i.e. what came before it) are still debated. These are in principle the central questions in origins of life (OL) research.

1.1 History of origins of life research

Given it's fundamentality, it is no surprise that, across all cultures, the first explanations for the question about the origin of life were religiously motivated and therefore involved the divine creation of life by some higher being

[1] Nick Lane put it rather dramatic in his book by stating that: "There is a black hole at the heart of biology." [Lane 2015]

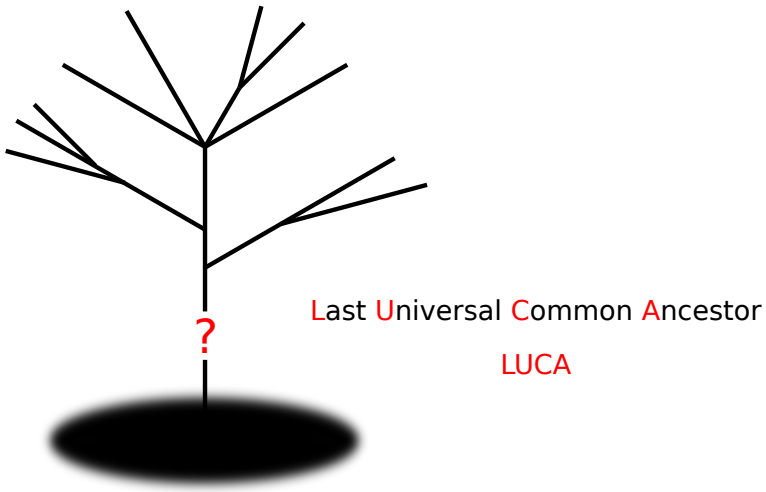


Figure 1.1: The problem of the origin of life (OL): it is clear that there must have been something like a last universal common ancestor (LUCA) that probably was a population of prokaryotes from which all current life derived. However, the exact characteristics of LUCA, as well as how it developed (i.e. what lies beyond it), are still a mystery.

(or beings). Early attempts at finding a more naturalistic explanation for this question lead to two contrasting theories. The first one (which can be traced all the way back to Aristotle) is the theory of 'spontaneous generation', i.e. the idea that life (or life matter) continuously forms from dead matter (like insects that grow in rotten food, etc.). The other theory assumes that there is a fundamental difference between living matter and dead matter and as such life is "eternal" (some people believed that there was something like a vital force inherent in life matter). The scientific advancements of the 19th century brought an end to these theories, most famously by the experiments of Pasteur and Wöhler (see Fig. 1.2). While the former demonstrated that life matter cannot generate from dead matter, the latter showed that there is furthermore no difference between them. Around the same time, Charles Darwin published his seminal work "On the Origin of Species" [Darwin 1859]. Even if it was not concerned with the origin of life per se[2], this work did introduce the concept of evolution by descent with mutation and natural selection which, as outlined above, naturally leads to the conclusion that there must have been a singular origin of life in the form

[2]There is nevertheless the famous letter from Darwin to Joseph Hooker where he speculates about the possibility of life forming in a warm little pond [Darwin 1871].

of LUCA.

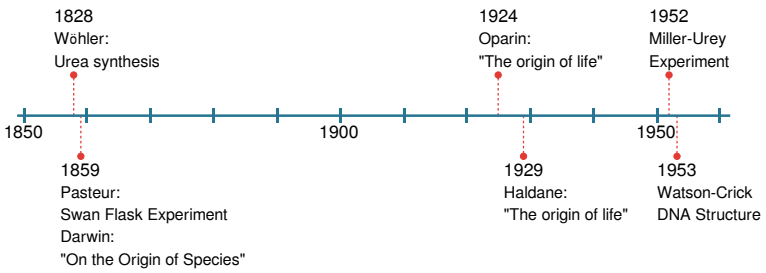


Figure 1.2: Timeline giving a short overview of the most important theories and experiments related to OL research that were conducted/formulated between the 19th and 20th century.

The first half of the 20th century finally saw the advent of the first "modern" theory about the origin of life formulated independently by Alexander Oparin and J. B. S. Haldane, which is thereafter referred to as Oparin-Haldane hypothesis [Oparin 1924; Haldane 1929]. They both assumed that the atmosphere of the early Earth mostly contained inorganic carbon in the form of methane (CH_4) as well as ammonia (NH_4) and water (in the form of steam), whereas oxygen was assumed to be absent in the atmosphere (it was assumed that it was mostly bound to minerals). Electrical discharges from lightning together with UV-radiation from the sun caused the compounds in this reducing atmosphere to react and form more complex organic molecules that then rained down into the oceans of the Early earth. Through this process, a "soup"[3] enriched with complex organic compounds was created. Those molecules were then broken down again into smaller compounds, through simple fermentation reactions, releasing energy in the process and thus building the basis for the first simple metabolic pathways [Haldane 1929]. Furthermore, Oparin even envisioned the formation of primitive cells by describing how some of these more complex organic compounds could form so called 'coacervates'[4] which could then house these reactions [Oparin 1924, 1936]. All in all, both, Oparin and Haldane, posed the first hypothesis on how organic compounds could have originated from inorganic ones, proposing furthermore that the origin of life was not a sudden transition but rather a long process of increasing complexity that could be studied

[3]Haldane himself used the term "hot dilute soup" [Haldane 1929]. In general, the Oparin-Haldane hypothesis is often colloquially referred to as "primordial soup" theory.

[4]A term invented by Bungenberg de Jong to describe polymers which in an aqueous solution do not mix but rather form colloids [Bungenberg de Jong and Kruyt 1929].

scientifically.

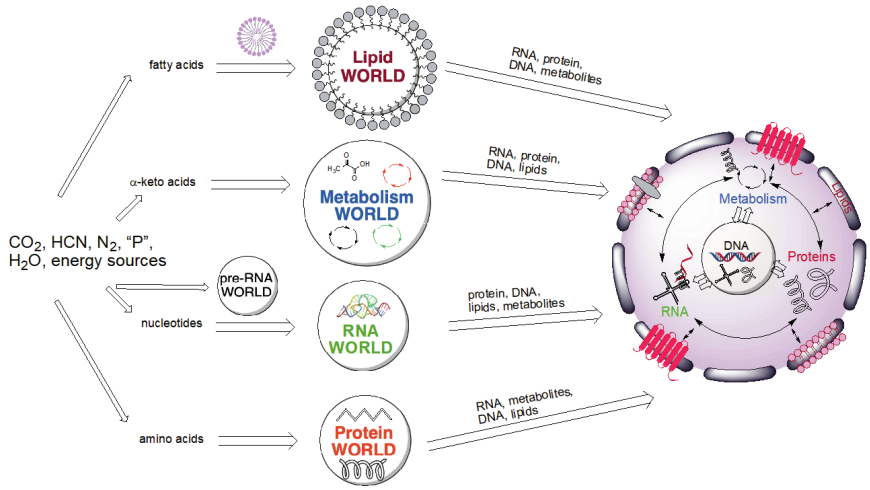


Figure 1.3: Different examples of "me-first" theories. Starting from simple precursors (left), any first appearance of one of the biomolecules or biochemical processes (center) is sufficient to lead to life as we know it (right). Figure taken from [Krishnamurthy 2020].

The second half of the 20th century brought another major push for OL research in the form of two major experimental discoveries (see again Fig. 1.2). The first was the now famous Miller-Urey experiment conducted by Stanley Miller under the supervision of Harold Urey. Creating in a test tube atmospheric conditions similar to those proposed by Oparin and Haldane, they showed that one can obtain simple amino acids, among other organic molecules, thorough electrical discharges, which simulated atmospheric lightning [Miller 1953, 1955]. While the validity of the atmospheric conditions assumed for the experiment are highly debated today, it was nevertheless an important proof of concept which demonstrated that, as Oparin and Haldane theorized, complex organic molecules can in fact be synthesized from a mixture of simple inorganic compounds under assumed prebiotic conditions. The second major finding, which had an even more dramatic effect on OL research, was the discovery of the structure of DNA by James Watson and Francis Crick[5] [Watson and Crick 1953]. The discovery of this "molecule of life" kickstarted the overall success of molecular biology. More and more discoveries about the nature and structure of the organic compounds involved

[5]Together with Rosalin Franklin, even though she never got proper credit for her contribution.

in biochemistry were made (e.g. polypeptides, phospholipids, etc.) and thus as all the processes associated with life (like reproduction, metabolism, a cell membrane) were explainable through the interaction of these complex biomolecules.

However this also lead to a certain *top-down* reductionistic approach to the origin of life. It was assumed that life at it's origins was already very similar to life as we know it and thus one should focus on the molecules or processes related to biochemistry in living systems (nucleic acids, proteins, lipids, etc.) and simply project "backwards in time" and try to find a plausible origin for them under what one would assume were prebiotic conditions. As such "classical" OL theories emerged like: DNA/RNA or replication first[6]; proteins or metabolism first, lipids or compartments first. These types of approaches may be referred to as "me-first" hypothesis [Krishnamurthy 2020], where it is assumed that once the appropriate bio-polymer or biochemical process deemed central to live has "arrived" the rest simply "falls into place" by itself (see also Fig.1.3).

1.2 A systems approach to origins of life research

Around the end of the 20th and the beginning of the 21st century a new trend in science began to take hold. Fuelled by the advancements made in the development of computers and the availability of large amounts of data, several scientists started to question the reductionism within their respective fields and instead "embraced" the complexity underlying the systems they studied. This new 'systems science' gathered under the paradigm that "the whole is more than the sum of it's parts", which had been already advanced by 'complex systems science' a few decades before [Anderson 1972], and became an established approach for origins-of-life research by ca. 2010 [Ruiz-Mirazo et al. 2014]. In other words, instead of trying to describe a system by reducing it to the average behaviour of it's individual components, the scientists realized that interactions among these components can form complex networks of relations which play a significant role in determining the overall behaviour of the system. Therefore, if two systems are formed by

[6]In particular, the discovery of ribozymes by Thomas Cech and Sidney Altman (together with their colleagues) [Kruger et al. 1982; Guerrier-Takada et al. 1983] led to the development of the RNA-world hypothesis that has had a tremendous impact in the field of OL research for many years.

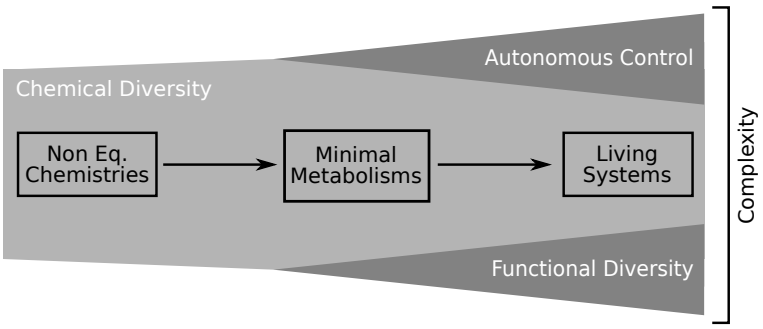


Figure 1.4: Minimal metabolisms would stand at the interface between non-equilibrium complex chemistries and biological systems. As major prebiotic transitions unfold, the complexity of the corresponding phenomena should increase, together with the capacity to develop autonomous control mechanisms. Whereas chemical diversity decreases (light grey area), in the sense that only a subset of all possible molecular compounds/types of reactions is exploited by living systems, functional diversity increases (darker grey area), in the sense that these systems manage to generate and couple together a wider (eventually, an open-ended) variety of interdependent components and transformation processes. Figure taken from [Lauber et al. 2021].

the same types of components, they can still be radically different due to a difference in their corresponding interaction networks.

A major discovery of this kind was made in the form of gene regulatory networks which led to the emerge of 'systems biology' [Kitano 2002a,b]. This new scientific discipline, trying to connect all the knowledge gathered about biomolecular mechanisms with cell properties and behaviour, combines in vitro experiments with in silico computational, theoretical models to address topics like: cellular dynamics and regulation; information processing and signalling networks in organisms; metabolic pathways; etc. [Westerhoff and Palsson 2004]. The rise and success of systems biology furthermore influenced a similar development within the field of chemistry. Together with insights from supramolecular chemistry, scientists began to move away from a purely synthetic chemistry and the study of isolated reactions in pure solutions, towards studying complex mixtures of multiple components where different reactions concurrently occur. Thus the field of 'systems chemistry' emerged [von Kiedrowski 2005; Stankiewicz and Eckardt 2006; Ludlow and Otto 2008; Ashkenasy et al. 2017]. Like systems biology, the idea is to combine in-vitro and in-silico techniques to study phenomena like: molecular self-replication; chiral symmetry breaking; molecular self-assembling and self-organizing; etc.

[de la Escosura et al. 2015].

The overall emergence of these non-reductionist systems views in both biology and chemistry has furthermore led to a new way of approaching the problem concerning the question about the origin of life through the field of 'prebiotic systems chemistry' [Ruiz-Mirazo et al. 2014; Kroiss et al. 2019]. In contrast to the more top down motivated approaches, it is assumes a more *bottom up* idea, i.e. that the chemical processes at the origin of life might have been rather different to what is observed in extant biology today. They most probably involved complex (but compared to extant biology simple) and heterogeneous chemical reaction pathways which then continuously developed and optimized, exploring multiple possibilities, thus increasing in complexity until they finally converged into life as we know it (see also Fig. 1.4). Thus the question that should be asked in connection with the origin of life is not "What came first?" but rather "How did things/processes get organized?"

It is important to realize, though, that this organization does not happen spontaneously (i.e. it is not a "downhill process"). It requires a thermodynamic effort which involves the coupling and internal management of matter as well as energy resources that both come from the local environment of the system. In other words, it is not just the generation of 'self-organizing' patterns under non-equilibrium conditions that matters, but also additional mechanisms in order to *autonomously* manage the flow of matter and energy through the system [Ruiz-Mirazo and Moreno 2004]. Therefore, the main idea which is going to be considered in this thesis project is that at the centre of this transition from non-equilibrium chemistries to living systems one should observe the development of a so called 'minimal metabolism' (MM). In general, it has already been theorized that a vital stepping stone in OL had to be some sort of proto metabolism that was able to run without enzymes [Dyson 1985; de Duve 1991; Shapiro 2007; Ralser 2018]. The aim is thus to characterize MM by providing a tentative scheme, as well as trying to investigate whether it is possible to realize such a minimal metabolic system. The latter task is going to be approached here mainly through the use of computational, theoretical models.

1.3 Thesis structure

The remainder of this thesis is structured in three major parts. The first part is aimed at characterizing the overall concept of MM. In order to this, first metabolism as it is realized within extant biology is examined, which is done in Ch. 2. Afterwards, in Ch. 3, previous models of metabolism are compared and through this the main ingredients for a minimal metabolic system are gathered, which allow to put forward a tentative scheme for MM, where both processes of transformation and "self-made" boundary conditions are given important theoretical weight. Finally, in Ch. 4 the methodological challenges of implementing and investigating such a minimal metabolism (both in vivo and in silico) are discussed. The second major part is then aimed at modelling a MM by implementing complex reaction networks under simple boundary conditions. Ch. 5, gives some context and motivation about the type of reaction system that is going to be studied, whereas the methods for this implementation are outlined in Ch. 6 and the main results then given in Ch. 7. The third major part takes a somewhat different approach to model MM, focusing on the complexity that boundary conditions could also bring about, even when the underlying chemistry is much simpler than what was explored in the previous (second) part. Therefore, following a similar scheme, Ch. 8 will serve to justify the type of boundary condition under exploration, while the methods for this implementation are outlined in Ch. 9 and the main results finally provided in Ch.10. Last but not least, in Ch. 11 the overall results from both modelling approaches are discussed, including their inherent limitations as well as possibilities for further research, followed by a general outlook and future perspectives. The dissertation is completed in Ch. 12 with a list of main conclusions, summarizing the central results of this PhD project.

1.4 Main hypotheses and objectives

The two central hypotheses that will be investigated within this dissertation are:

- Hypothesis 1: The origin of life (OL) was not a process where a population of biomolecules (be them RNA/DNA, proteins or lipids) evolved in parallel through natural selection to give rise, subsequently, to the first proto-cellular organisms and, later on, to a population similar to LUCA (Last Universal Common Ancestor).

- Hypothesis 2: Biological phenomena emerge as a combination of physical and chemical factors that enable the formation of metabolic systems. These metabolic systems were initially much simpler than current 'genetically-instructed metabolisms'.

By investigating these hypotheses, the following main objectives will be accomplished:

- Objective 1: carry out a conceptual analysis of metabolism in order to formulate a tentative scheme for minimal metabolism (MM) and characterize the methodological challenges of implementing and investigating MMs.

- Objective 2: capture the chemical reaction aspect of MM within a computer model that implements relatively complex reaction networks under simple boundary conditions.

- Objective 3: capture the importance of non-trivial physical boundary conditions for MM elaborating a computer model that implements a simple reaction network under constraints related to phase separation phenomena.

Part I

From Metabolism to Minimal Metabolism

Chapter 2

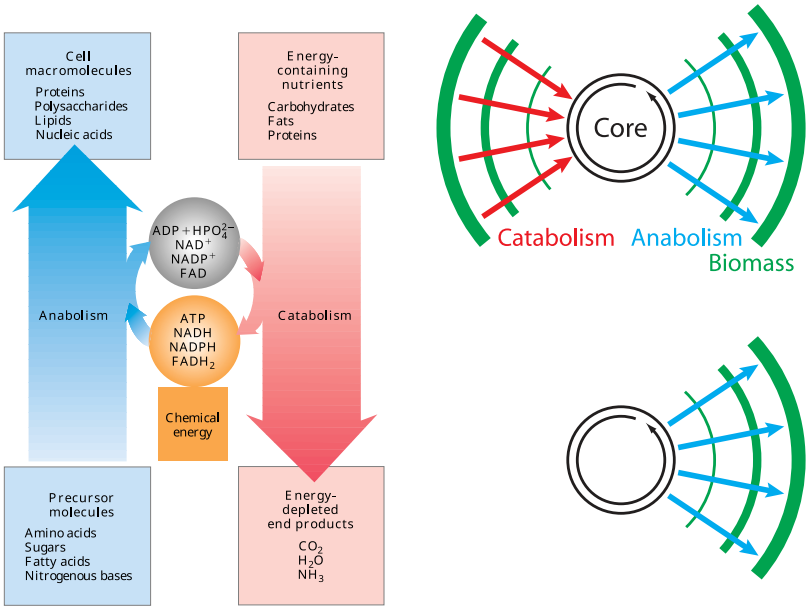
Metabolism

Before further elaborating the concept of minimal metabolism, a short characterisation of extant metabolism is necessary. A somewhat "naive" notion of it is to describe metabolism as the set of chemical reactions that occur within an organism to keep it alive. In general, it is true that metabolism plays a central role in the energy production and self maintenance of an organism. Arren Bar-Even, an expert in systems and synthetic biology, put it quite nicely by saying that: "Metabolism is the kernel of life" [Erb 2020][1]. However, in practice a metabolism is more than just a set of chemical reactions. In fact, extant metabolism can be characterized as complex network of chemical reactions which is furthermore embedded in a "hyper complex" organization of biomolecular structures, reflecting a deep coupling between processes of synthesis and different layers of control and regulation of those same processes.

2.1 Metabolic Networks

In general, metabolic pathways like glycolysis, gluconeogenesis, the tricarboxylic acid (TCA) cycle, the Calvin cycle, etc., can be classified into two main categories: catabolic pathways that break down large organic molecules into smaller compounds, releasing energy in the process (e.g. glycolysis, or the TCA cycle); anabolic pathways that require energy to build large organic molecules from smaller compounds (e.g. gluconeogenesis, or the Calvin cy-

[1] This is a rather fitting analogy, as in computer science a kernel refers to the most central software component within an operating system, controlling the most important functions and in general serves as an interface between hardware and software.



(a) **(b)**

Figure 2.1: **(a)** Schematic depiction of the interplay between anabolism and catabolism. Figure taken from [Nelson and Cox 2017]. **(b)** Top: typical "bowtie" motif of the heterotrophic mode of metabolism. Bottom: motif of the autotrophic mode of metabolism. Figure adapted from [Braakman and Smith 2013].

cle). The metabolism of an organism can then be further classified as either being heterotrophic or autotrophic, depending on the primary carbon source and on how anabolic and catabolic pathways get organized thereafter.

Heterotrophic organisms (e.g. animals or fungi) take in large and complex carbon molecules (nutrients) from the environment, which are then broken down via catabolic reactions into smaller molecules, with the released energy used to convert ADP to ATP as well as NAD(P)$^+$ and FAD to NAD(P)H and $FADH_2$ [Nelson and Cox 2017]. The stored energy and redox potential is subsequently used by anabolic pathways in order to build up macromolecules like proteins, sugars, lipids, nucleic acids, etc. which can then be further used in other biochemical processes (see Fig. 2.1a). As shown in Fig. 2.1b, this usually results in the characteristic bow tie motif associated with metabolism [Csete and Doyle 2004]. Autotrophic organisms (e.g. plants or algae), on the other hand, start from small, inorganic carbon molecules. Together with energy from the environment, these are then used to directly build up biomass via anabolic reactions (see Fig. 2.1b). Some of this biomass can also

serve as input for catabolic processes again, thus autotrophic organisms are basically self-sufficient (one can say that autotrophs "make their own food"). Autotrophy and heterotrophy tend to be idiosyncratic modes of metabolism, characteristic of each species, even if some organisms (mixotrophs) can also switch between them, depending on the environmental circumstances (i.e., the availability of a certain type of carbon source) [Braakman and Smith 2013].

As shown in Fig. 2.2, the metabolic pathways of an organism therefore form a complex network of coupled chemical reactions, some of which belong to catabolism and some to anabolism. All in all, the central function fulfilled by these complex reaction networks is to create the energy as well as the fundamental building blocks that can be subsequently used in other biochemical processes. Therefore, in line with the notion put forward by Arren Bar-Even, one can view metabolism as *the core chemistry of life*[2] [Lauber et al. 2021].

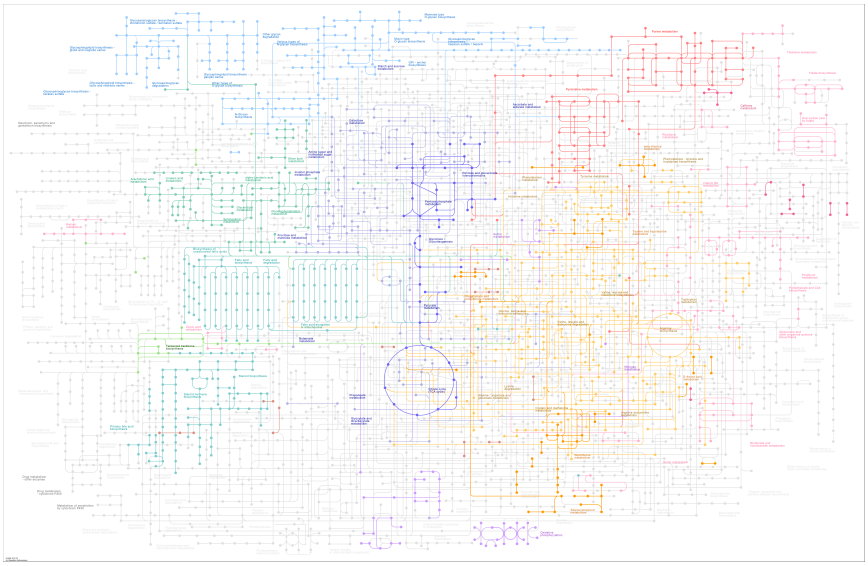


Figure 2.2: Map depicting all known metabolic pathways and their connection/-coupling with each other. Further highlighted is the metabolism of humans. Figure taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [Kanehisa and Goto 2000; KEGG].

[2]Although in the context of this project, metabolism is not going to be understood *synonymous to* life but rather as *fundamental for* life.

2.2 Metabolism in extant biology

The complex chemical reaction networks constituting metabolism alone would not manage to sustain themselves. They need to be further controlled and regulated within the complex biomolecular context provided by the organism. In general, within extant biology, different mechanisms of control and regulation are realized.

2.2.1 Kinetic and variability control

Under normal conditions the majority of reactions that occur as part of current metabolic pathways require a larger activation energy and therefore would, in principle, proceed rather slowly. If that was the case, the intermediates of a metabolic reaction could in the meantime undergo other types of reactions that might require a smaller activation energy (i.e. that happen with a faster rate). These "parasitic" side reactions would thus take away material that is otherwise needed for the main metabolic reactions to take place. This is prevented, however, through catalysts (substances like metals, acids or bases, organocatalysts – but also mineral surfaces, perhaps, in prebiotic times) that can lower the amount of energy necessary to start a chemical reaction, therefore speeding it up. In extant biology, these catalysts typically consist of a special class of proteins, referred to as enzymes. Each enzyme then catalyses a corresponding reaction step within the chemical reaction network associated with a certain metabolic pathway[3]. Therefore, through regulating the kinetics within a metabolic reaction network, enzymes effectively control the flux of material that passes through the network. This regulatory function that enzymes play within metabolism can be understood as 'kinetic control' [Ruiz-Mirazo et al. 2017], where macromolecular structures have direct constraining effects on the underlying transformation processes. Thus, enzymes can be considered as 'first-order' constraints.

However, enzymes are rather complex molecules. As mentioned above, their building blocks (amino acids) as well as the energy required for their assembly come from the core metabolic pathways. Yet, these amino acids also need to be assembled in a certain order such that the resulting pep-

[3]Thus most enzymes are 'specific catalysts' that catalyse only one reaction. By contrast, many inorganic catalysts like metals can catalyse multiple different reactions at the same time.

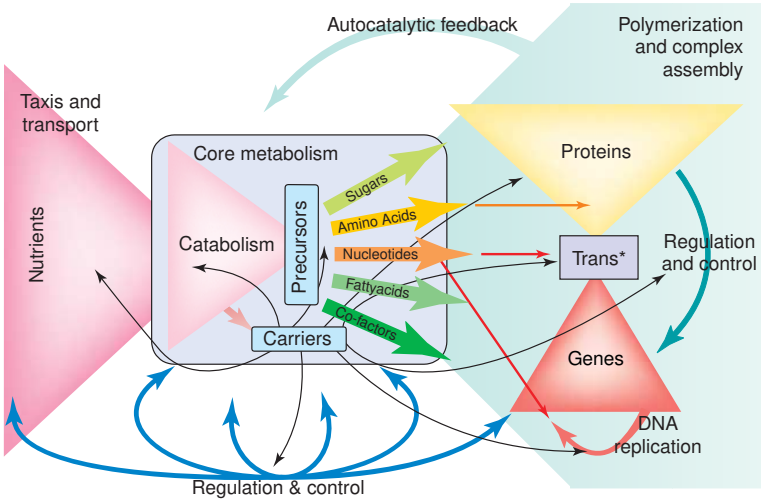


Figure 2.3: Schematic depiction of a genetically-instructed metabolism (GIM) and its different layers of feedback between synthesis, control and regulation. Figure taken from [Csete and Doyle 2004].

tide chain can fold into a protein. This order of assembly is, in general, encoded in sequences of nucleic acids on the DNA, i.e. genes. These genes then get transcribed and translated into proteins/enzymes through different kinds of RNAs[4]. The possibility for variation in these gene sequences is the basis of natural selection and therefore darwinian evolution which gives the potential for the development of more complex enzymes. Therefore this encoding of amino acid sequences through genes exerts a form of 'variability control' [Ruiz-Mirazo et al. 2017] on the enzymes and as a consequence on the metabolism as a whole (creating a level of meta-constraints). Due to this overall dependency on genes, one can say that all known metabolisms are 'genetically-instructed metabolisms' (GIMs). As shown in Fig. 2.3, the above mentioned processes of translation and transcription, as well as the replication of DNA/RNA, are regulated by other types of specific proteins, whose structure is again encoded by genes. Furthermore, as with proteins, the building blocks for the DNA/RNA (nucleic acids) are synthesized by the core metabolic pathways too. Thus, any GIM is a hyper-complex system of coupled synthesis and control relations that involves different layers of constraints and meta-constraints[5].

[4]This interplay between DNA, RNA and Proteins is commonly referred to as the 'central dogma of molecular biology' [Nelson and Cox 2017].

[5]The knowledge to describe this important aspect of extant metabolism is the product

There are research groups investigating minimal forms of GIMs, like the *Mycoplasma* construct studied by Craig Venter *et al.* [Hutchison et al. 2016; Breuer et al. 2019] , where the minimal genome that could be the basis for a metabolism of a minimal cell is explored. The study of this Mycoplasm has definitely led to valuable insights in the research of metabolism and synthetic biology. While this top-down approach to explore how far the genome of a minimal cell can be scaled down has lead to valuable insights about a possible "end point" of the origins of life (OL) process, it nevertheless provides rather little information about abiogenesis itself. By contrast, in the upcoming chapter of this work a more bottom-up motivated conception of minimal metabolism is going to be presented as an attempt at formulating a different kind of metabolic system, at an intermediate stage between chemistry and biology, which does not explicitly necessitates genes, and can therefore be more applicable and informative for OL research.

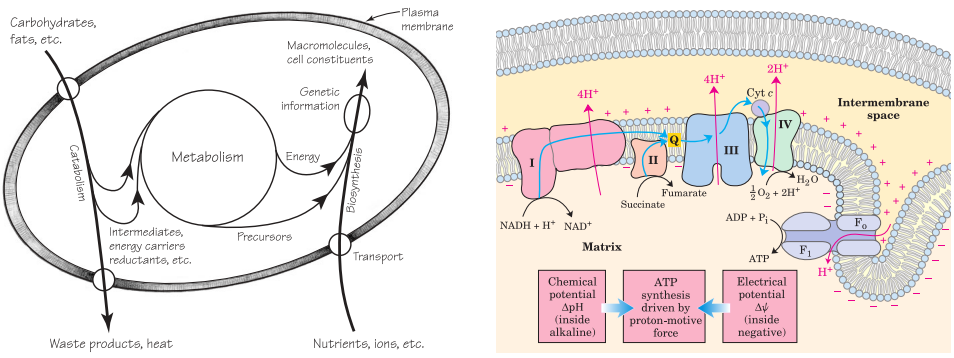
2.2.2 Spatial and energetic control

In general, all of the different chemical compounds and intermediates that partake in the above mentioned metabolic reaction processes and biochemical pathways would simply dilute away if left in a larger environment (e.g. the archean ocean). This is why extant metabolism happens within cells that contain their molecular components through a special membrane which is primarily made up of phospholipids[6], creating a spatial boundary that establishes a clear distinction between what is inside inside and outside of the cell. As shown in Fig. 2.4a, the inflow of nutrients and other compounds for the metabolism as well as the outflow of waste products is regulated by other types of specific proteins. These membrane proteins act as active/-passive carriers, pores, energy transductors, etc. that manage and modulate the flow of matter and energy through the system. By preventing the dilution of molecular compounds as well as regulating the in- and outflow, this cellular compartment realizes a 'spatial control' [Ruiz-Mirazo et al. 2017] on the metabolism within. Thus one has an additional type of control to

of the advances made in molecular biology during the 20th century. However, as mentioned in the introductory chapter, these advances also led to a more reductionistic approach in biology as well as the research for the origins of life.

[6]Molecules that have a polar head group that is hydrophilic and an apolar tail group that is hydrophobic (or lipophilic). Due to this amphiphilic nature, when put into an aqueous solution, these compounds tend to form bilayers.

the above discussed kinetic control of enzymes and the variability control of genes. Again, the respective components to create this spatial boundary (lipids for the membrane, amino acids for the membrane proteins) are also synthesised by the core metabolic pathways. In other words, one can see yet another level of coupling between synthesis and control.



(a) **(b)**

Figure 2.4: **(a)** Simple scheme of a cell depicting the spatial boundary realized by the cell membrane as well as the in- and outflow of material enabled by specialized membrane proteins. Figure taken from [Harold 2001]. **(b)** Chemiosmotic model of ATP hydrolysis: proteins in the inner mitochondrial membrane, activate by e^- from NADH, pump H^+ ions into the intermembrane space, creating a proton gradient that drives the ATP synthesis. Figure taken from [Nelson and Cox 2017].

Finally, all the different metabolic reactions, in general do not happen in the same place within a cell. Some of them occur in the cellular medium, the cytoplasm, while others happen in certain sub-compartments called organelles that additionally can have their own types of membranes, i.e. spatial boundaries. Some reactions even take place between or within such organelle membranes. An example of this is shown in Fig. 2.4b. Across such membrane spaces one can often find electrical and/or chemical gradients (in Fig. 2.4b realized through a gradient of H^+ ions). By changing the electrochemical environment, these gradients can facilitate certain thermodynamically unfavourable reactions (similar as enzymes who facilitate kinetically unfavourable reactions). Thus one has again a regulation of the flow of matter through the reaction network, creating a fourth form of control on extant metabolism: 'energetic control' [Ruiz-Mirazo et al. 2017]. As discussed in [Ruiz-Mirazo et al. 2017] and as will be further elaborated in the next chapter, extant metabolisms are not simply characterised by one of these four types of control (kinetic, spatial, energetic or variability control) in isolation

but rather by a complex combination of them.

All in all, one can see that the spatial organisation of the cell plays an important role in any known form of metabolism. This was furthermore coined by Franklin Harold and Peter Mitchell in their concept of 'vectorial metabolism' [Harold 1991; Mitchell 1991]. Vectorial refers here to having an orientation in space. In other words, the biochemical processes are oriented in the cell as they have a spatial direction: nutrients go in, they get processed, transported to different parts of the cell, further processed and then leave the cell again (usually through a different mechanism). This aspect is so fundamental that even minimalised expressions of metabolism, like the ones that are going to be proposed as intermediate systems between chemistry and biology, should include it. Following those lines, a major part of this work is going to be about the spatial organization in a simple chemical reaction model and how it might influence its kinetics.

Chapter 3

Minimal Metabolism

With extant metabolism outlined in the previous chapter, the stage is now set to give a more in depth analysis of the concept of minimal metabolism (MM). This will be done by reviewing past attempts at formulating models that represent a "minimal metabolic organization", analysing their advantages and disadvantages, as well as extracting in the process the main aspects of metabolism. These are then going to be the bases to formulate a tentative scheme of MM.

3.1 Previous models

The models that are going to be presented within this section were all conceived in the second half of the 20th century. In principle they were formulated in opposition to the more reductionistic notions of molecular biology that, as mentioned in the introductory chapter, were predominant at that time. As such, all the authors that proposed these "classical" models had a conception of a biological system as a dynamically organized set of molecules. The following pages give a brief overview, how that organization is characterized within these models[1].

***(M,R)*-Systems:** this model was formulated by Robert Rosen [Rosen 1958a,b, 1971] and is centred around the following problem: in general, the metabolism of an organism consists of enzymatic reactions where each en-

[1] While the Hypercycle model from Mafred Eigen and Peter Schuster [Eigen and Schuster 1979] is often included among this classical collection of models, it will not be included here, as for the purpose of this text it is not considered a model of metabolism but rather of the co-evolution of replicators.

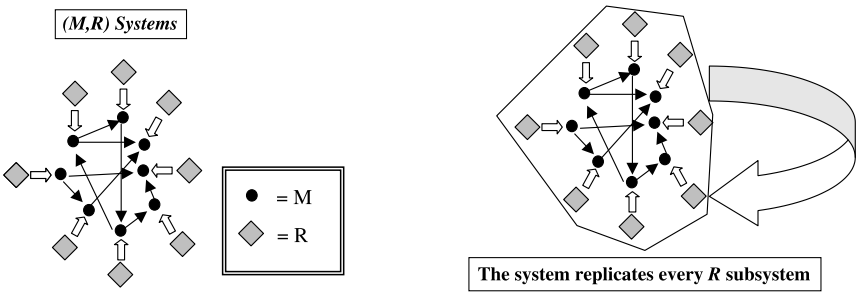


Figure 3.1: Simplified representation of an (M, R)-system. Left panel: round nodes represent enzymes where two nodes are connected by directed black arrows if ones output is the others input. Each component of this M-system needs a corresponding R-system (diamond nodes) with the repair actions represented as white arrows. Right panel: M- and $R-$systems together repair each $R-$system, forming the (M, R)-system. Figure adapted from [Letelier et al. 2003].

zyme, through its catalysing action, transforms a substrate into a product (this was introduced as kinetic control in the previous chapter). The complete collection of enzymes, which are usually connected through shared in- and outputs, is labelled M-system [Rosen 1958a,b] (see also Fig. 3.1). However, these enzymes can degrade (or dilute) over time. As shown in Fig. 3.1, each enzyme type therefore requires a system of reactions, labelled R-system, to replace (or repair) them, in order to maintain their concentration. However, the reactions of each R-system itself need the catalysing action of specific enzymes which can again degrade/dilute and thus need to be replaced/repaired as well. As further shown in Fig. 3.1, Rosen's model attempts to avoid this infinite regress of repair relations by introducing a concept called 'closure to efficient causation': the metabolism and the repair systems together form the system of reactions to replace/repair the catalysts of the repair system, thus forming the titular (M, R)-system.

Autopoiesis: Francisco Varela and Humberto Maturana developed a general concept they called 'autopoiesis' [Maturana and Varela 1973]. They defined an autopoietic system as: "a network of production of components which: (i) recursively participate in the same network of production of components which produce them and (ii) realize this network of productions as a unity in the space in which the components exist" [Varela et al. 1974]. A simple example of this concept is depicted in Fig. 3.2a, where nutrient molecules A get transformed through a metabolic process (under the influence of catalysts) to monomers S. These monomers then form a polymer chain that

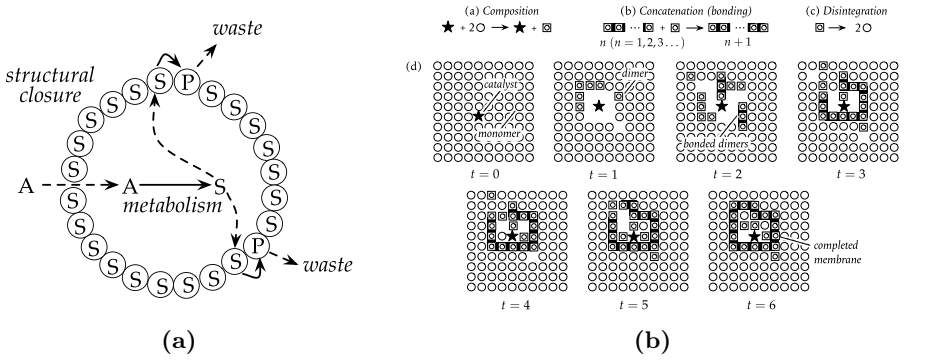


Figure 3.2: **(a)** Scheme depicting a simple example of an autopoietic system. **(b)** Simple computer model to simulate an autopoietic system using a 2-dimensional regular lattice. Circles represent substrates, squared circles represent monomers and stars represents catalysts. The different types of reactions (transformation of substrates to monomers, concatenation, etc.) are implemented through update rules on each cell. Figures taken from [Cornish-Bowden and Cárdenas 2020].

may close in on itself, thus creating a spatial boundary (akin to a membrane) that encloses the nutrients, monomers and catalysts of the metabolic process. While the monomers and catalysts cannot pass the membrane created by the polymer, it is assumed to be permeable for the substrate, allowing a constant influx of it. However, as single monomers also may degrade to waste P, this can cause ruptures in the membrane that need to be repaired with new monomers coming again from the metabolic process. Therefore, a spatial boundary is created and maintained by a system of reactions it encloses, actively sustaining them in space. Even though autopoiesis is a more general concept, a tesselation computer model of the above outlined scheme, depicted in Fig. 3.2b, was implemented in collaboration with Roberto Uribe [Varela et al. 1974] which was further extended by Barry McMullin [McMullin and Varela 1997].

Chemoton: this model was formulated by Tibor Gánti in an attempt to come up with a concept for a minimal synthetic cell [Gánti 1971, 1975]. As shown in Fig. 3.3, it consists of three coupled reaction cycles. There is a "metabolic cycle" that converts incoming food molecules X_A into waste product Y. At the same time, this cycle produces monomers V' that feed into an "information cycle" in order to replicate a template polymer pV_n. Both cycles furthermore form the precursors T' and R, respectively, which are then combined to form another monomer T that gets incorporated into a membrane polymer, T_m, enclosing the whole system. Once the template

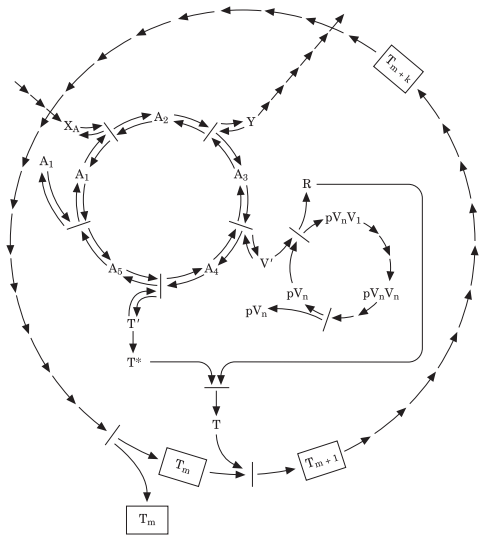


Figure 3.3: Scheme for the chemoton model depicting the three coupled reaction systems: metabolic cycle, information cycle and membrane cycle. Figure taken from [Gánti 2002].

polymer has been fully replicated and the membrane polymer has exactly doubled in size, the system splits into two equal systems. In this way, the replication of the membrane polymer forms the third reaction cycle (a "membrane cycle")[2]. The process then continues in each system anew. The system as a whole is also kept out of equilibrium through a constant influx of X_A and outflux of Y through the membrane, both of which are irreversible. A stochastic simulation of this model was later implemented by Tom Lenaerts and colleagues [Van Segbroeck et al. 2009].

Autocatalytic (AC) sets: The initial model was formulated by Stuart Kauffman [Farmer et al. 1986; Kauffman 1986] and then later extended and generalized by Vim Hordijk and Mike Steel [Steel 2000; Hordijk and Steel 2004]. The main idea is to describe the emergence of self organization in the form of a co-catalytic activity between oligopeptides that would eventually lead to life. The base model, shown in Fig. 3.4, assumes an initial food set of monomers, here labelled a and b, that can undergo consecutive concatenation reactions, creating a set of simple polymers (representing oligopeptides). Each polymer can undergo cleavage reactions, creating in turn smaller polymers and/or monomers. Through this a reaction network of oligopeptides

[2]The original chemoton model, however, did not include this membrane cycle but only involved a coupling between the metabolic- and the information cycle.

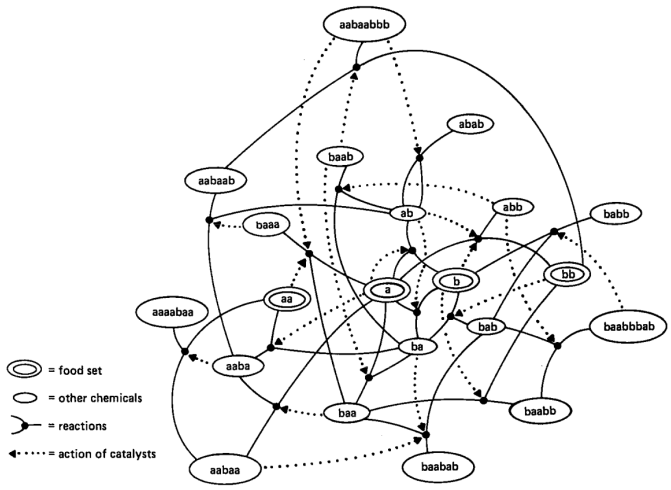


Figure 3.4: A simple example of an autocatalytic network of peptides. Figure taken from [Farmer et al. 1986].

is realized. It is assumed that each reaction in the network is catalysed by a molecule from within its set of reactants (i.e. the set of polymers) with a fixed probability P. Under these assumptions, an autocatalytic (AC) set (or network) is achieved when the following condition is met: for every molecule at least one reaction that leads to its formation is catalysed by another member of the set and connected sequences of catalysed reactions lead from a maintained food set of monomers to each member of the set [Kauffman 1986; Hordijk and Steel 2004]. It has been shown that this catalytic closure emerges by itself if, given the catalytic probability P, the system is sufficiently large i.e. contains a large enough number of molecules [Kauffman 1986; Hordijk and Steel 2004].

Algorithmic chemistry: this model was formulated by Walther Fontana and Leo Buss [Fontana and Buss 1994a,b, 1996] with the main idea to describe the emergence of different levels of self organization within networks of chemical reactions. Using the framework of λ-calculus (a generalization of mathematical functions) molecules are abstracted as functions, i.e. λ-expressions, while chemical reactions are abstracted as compositions of functions. The idea is that, in general, molecules can participate in certain reactions, due to the bonds and atoms they are made of. In a similar way, functions perform certain mathematical operations, due to the variables and symbols they contain. Furthermore, a pair of molecules can react to create a new molecule (with new capabilities to partake in reactions) while two

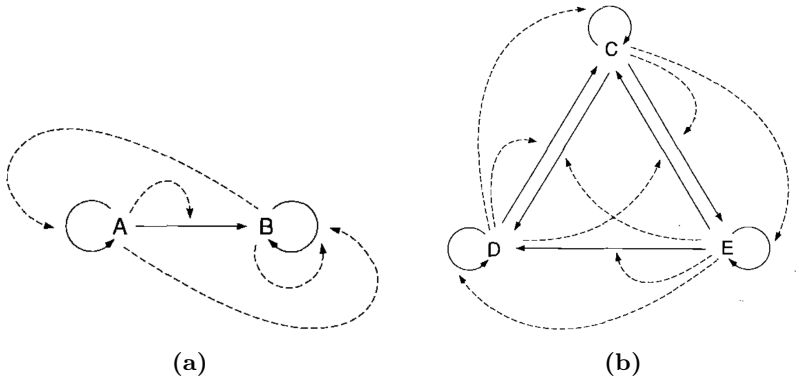


Figure 3.5: Two examples of a self maintaining, or level 0 ($L0$) organisations. The letters A, B, C, D represent different λ-expressions. Different arrow types represent the composition relation between two expressions (e.g. $E \circ C = D$ is represented as a solid arrow from C to D pointed at by a dotted arrow from E). **(a)** Simple L0 organization consisting of two expressions. **(b)** More complex L0 organization. Figures are taken from [Fontana and Buss 1994a].

functions can be composed to a new function (with new operations to perform). A "flow reactor" is then implemented in a computer model where, starting form an initial set, pairs of λ-expressions are randomly combined to potentially create new expressions for the set, until a steady state is reached [Fontana and Buss 1994a]. Within such a steady state, a certain dominant type of organisation (a certain "motif of relationships" between λ-expressions) usually emerges. Fig 3.5 shows two examples of a so called 'self maintaining' organisation, where the individual expressions produce and maintain each other. Systems of this type are the simplest forms of organization that are observed in the model and, as such, they are labelled as level 0 (or $L0$) organizations [Fontana and Buss 1994b]. Depending on the initial conditions and other constraints put on the simulation, systems of higher complexity (i.e. higher-level organizations) can be reached. These include 'meta-organizations' where various lower-level organizations themselves interact between each other [Fontana and Buss 1994b].

Class	Models
Minimal life	(M, R)-systems, Autopoiesis, Chemoton
Self organization leading to life	AC-sets, Algorithmic Chemistry

Table 3.1: Categorization of the above discussed models.

All of the models outlined above view metabolism as an organization,

either in space (autopoiesis and chemoton) or in time (AC-sets, (M, R)-systems,algorithmic chemistry). Therefore, there are apparent similarities between them. As shown in Table 3.1, it is possible to group (M, R)-systems, autopoiesis and the chemoton model together as models that attempt to define a "minimal living system"[3]. The AC-sets, on the other hand, as well as the algorithmic chemistry simulations can be viewed as models that attempt to describe the emergence of self organizing systems that would eventually open the way to life.

3.2 Critical review of metabolism models

Now that an overview of the most important models of metabolic organization was given, this section adopts a more critical view, analysing their advantages and disadvantages (or shortcomings) with the aim to gather, through that discussion, the "key ingredients" of any metabolism. For starters, a unifying feature inherent to all of these classical models is a notion of *non-reducibility*. All of them involve, each in its own way, coupled sets of diverse molecules and transformation processes that cannot be scaled down to a particular compound and/or reaction (after all, these models were formulated from the antipodes of the reductionistic tendencies of molecular biology).

Furthermore, all models take into account a *functional relationship* between a set of transformation processes and their products. More precisely, there is a constraining (or controlling) effect that the products exert backwards onto the reaction processes that synthesize them. As reviewed in the previous chapter, this coupling between synthesis and control is realized in extant biology either through macromolecules (e.g. proteins, DNA/RNA) or supramolecular structures (e.g. lipid membranes) that operate on top of a network of chemical reactions which, at the same time, produces their building blocks. As such, each classical model includes this core idea in a different manner: within $(M, R)-$systems enzymes are at the same time products as well as controllers of reactions [Rosen 1971]; in the autopoietic model a spatial boundary (i.e. a compartment) is produced and maintained by the reaction network it encloses and, through this, the boundary sustains the network in space [Varela et al. 1974]; in the chemoton model a template

[3]The connection between (M, R)-systems and autopoiesis has already been analysed in detail by Juan Letelier and colleagues [Letelier et al. 2003].

polymer controls the size of a metabolic subsystem by means of stoichiometric couplings, in turn proving precursors for the growth of the former [Gánti 1971]; the autocatlytic sets are made up of oligomers that catalyse their own complex network of synthesis/cleavage reactions [Kauffman 1986]; in the algorithmic chemistry model $\lambda-$ expressions develop into complex organizations whose maintenance relies on a hierarchical relationship with other $\lambda-$expressions from the set [Fontana and Buss 1994a].

Upon closer inspection, however, one can find significant shortcomings within all of these models. Some of them rely too much on abstract concepts, like: closure to efficient causation in the case of $(M, R)-$systems, or organizational closure in the case of autopoietic systems. Other models make too many simplifications, like: assuming perfect stoichiometric couplings between reactions in the chemoton model, or assuming a homogeneous catalysis probability among oligomers within autocatalytic sets. Or they reduce chemical reactions to operations of functional expressions, as done for the algorithmic chemistry model. In other words, none of the classical models takes properly into account that metabolic pathways, in general, do not happen in such abstract and/or idealized environments. As discussed in the previous chapter, any extant metabolism takes place in non-trivial and heterogeneous conditions, as it involves a large diversity of compounds partaking in complex chemical reaction networks/media. All of these processes furthermore operate under dissipative, far-from equilibrium thermodynamic conditions (i.e. real metabolism is rather "messy").

Additionally, even though these models do a good job at capturing the previously mentioned coupling between production processes and constraints, they do so by only focusing on a single type type of controlling mechanism: e.g., spatial control in the autopoietic model [Varela et al. 1974]; stoichiometric control via a biopolymer in the chemoton model [Gánti 1971]; kinetic control in the (M, R) systems [Rosen 1971]; etc. Nevertheless, within extant metabolism there are various types of control mechanisms (namely spatial-, energetic-, kinetic control- and variability control) operating simultaneously and in rather complex combinations. Therefore, as argued in [Ruiz-Mirazo et al. 2017], metabolism may have emerged under less restrictive or reductionist conditions, in which a diverse combination of multiple types of control mechanisms concur, forming an *irreducible core of different control mechanisms.*

Last but not least, all of the above mentioned models are purely theoretical models, almost totally detached from experimental data/observations[4]. As already argued in [Lauber et al. 2021] (and as will be better justified in the next chapter), research of metabolic systems as well as research on the origins of life in general should involve close cross-talk between experimental research and theoretical models through integration of their respective results. In this way both types of approaches can provide constant feedback to each other. All in all, the following key ingredients of metabolism can be summarized:

- it is a non reducible systems construct
- it involves a large diversity of compounds under open and far-from-equilibrium thermodynamic conditions
- it displays constructive self-maintenance in the form of a functional bootstrapping between synthesis and control

Furthermore, advances in the modelling of metabolism should go hand in hand with experimental approaches.

3.3 Tentative scheme for minimal metabolism

By taking into account the main aspects of metabolism defined in the previous section, a tentative scheme that captures the minimal requirements on a system to be called a metabolism can be formulated. As depicted in Fig. 3.6, the central premise of this proposal is the key concept of a functional bootstrapping between synthesis and control. Accordingly, two main operational levels may be distinguished in the system: one represents transformation processes, while the other is related to constraints (or boundary conditions). Both levels, which are shown in Fig 3.6, would be populated by a divers set of components, reflecting the irreducibility of the proposal. As such, the "lower level" (i.e. the processes) would be populated by simple molecules that interact with each other via chemical reactions, thus forming a complex reaction network. The "upper level" (i.e. the constraints), on the other hand, contains more complex molecules and supra-molecular structures

[4]There have been attempts at experimentally implementing autopoiteic systems [Walde et al. 1994b] as well as autocatalytic sets [Lee et al. 1996; Kauffman 1996]. Even though this led to interesting results, their implication and scope remained rather limited.

(e.g. polymer chains, compartments, etc.), which are products of the lower level reactions and at the same time exert control on these processes. Such a feedback loop between synthesis and control (i.e. a system that synthesizes it's own boundary conditions that maintain itself) is exactly what is meant with the term 'functional bootstrapping'.

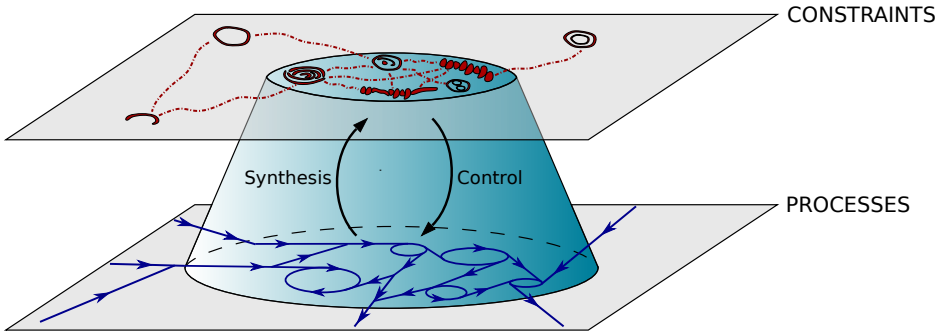


Figure 3.6: Graphical representation of the functional bootstrapping between processes and constraints. The shape and volume of the depicted truncated cone is not relevant, as it is a metaphorical illustration for the main idea of non-reducibility: the top and bottom bases cannot shrink to a point (i.e. it remains truncated). Figure taken from [Lauber et al. 2021].

The main novelty of this scheme is that this coupling between synthesis and control cannot be reduced down to a set of reactions that is coupled to only one type of constraint (as it has been done by the classical models). There rather needs to be a variety of components and interactions at each level of description. This is depicted in Fig. 3.6, as a graphical metaphor, with the drawing of a truncated cone between both layers. The idea is thus to present a less reductionistic picture of metabolism that acknowledges its intrinsic complexity.

Nevertheless, the scheme presented here involves a strong simplification, compared to the hyper-complexity of a genetically instructed metabolism (GIM), described in the previous chapter, with its many layers of different constraints and meta-constraints. Minimal metabolism (MM) is therefore not understood as something that can be found in biology, or through top-down simplifications like it has been done by Craig Venter *et al.* [Breuer et al. 2019]. It is rather a heuristic construct that stands on its own, halfway between complex chemistry and extant biology (see also Fig 1.4 of the introductory chapter). It also gives the opportunity to treat metabolism in terms of coupled reactions without the apparent need for genes and complex

enzymes.

In fact, in recent years attempts have been made at experimentally implementing metabolic pathways (or plausible precursors of them) under assumed prebiotic conditions and in the absence of enzymes ([Keller et al. 2014, 2016; Coggins and Powner 2017; Muchowska et al. 2017, 2019; Springsteen et al. 2018; Stubbs et al. 2020]). Nevertheless, despite the undoubtedly high impact of these works, they cannot be considered as examples of an in vitro realization of a MM, as it is understood here, since they only involve chemical reactions in solution. Thus, they are lacking the above outlined functional integration of these reactions into a system of synthesis and control. Nevertheless, they do demonstrate the overall plausibility of a proto-metabolism predating enzymes and genes. Thus one can (and should) build onto these experimental insights. However, future research and exploration of the origins of such a proto metabolism (or the origins of life in general) should involve the study of complex chemical mixtures that bring about irreducible combinations of molecules and supramolecular structures that exert different combinations of controlling effects [Ruiz-Mirazo et al. 2017] on the reaction processes.

Chapter 4

Methodological challenges

The tentative scheme for minimal metabolism (MM) presented in the previous chapter is a general concept that was proposed to have in mind when conducting research on the origins of metabolism or the origins of life in general. However it also shows that any candidate for such a MM would be a complex, dynamic system. Thus, the research of MM, both in vitro and in silico, has to deal with several methodological challenges resulting from these complexities. Within this chapter, some of these challenges are brought to the fore. After a first reflection on them (since they will be revisited in the last part of the dissertation, in the chapter concerning discussion & outlook), a possible work flow is formulated to address the main issues and move forward in the research on minimal metabolism and origins of life (OL) in the future.

4.1 In vitro, experimental challenges

As outlined in the previous chapters, a main aspect of minimal metabolism (or metabolism in general) is that it involves networks of chemical reactions that are prevented from reaching chemical/thermodynamic equilibrium through a continuos inflow of substrates and outflow of reaction products. Accordingly, the research for a plausible proto-metabolic systems should aim at studying reactions under such out-of-equilibrium conditions. However, keeping a reaction mixture in an out-of-equilibrium state and, in particular, developing tools of analysis of what is going on dynamically in such a system, is not an easy experimental task. This is why, in practice, most

reaction mixtures tend to be studied at equilibrium or as they relax towards an equilibrium configuration.

Another key feature of MM is that it involves a large diversity of compounds, which can interact in complex ways. Therefore one should try to study a diverse and complex reaction system where all the individual reaction steps happen within the same vessel, i.e. studying reactions as they occur in *one pot*. But trying to study reaction mixtures that involve many components is also experimentally very challenging, as there might be a lot of intermediate compounds (and thus, reactions that are difficult to predict/detect), or compounds that could cause unwanted side reactions. In general, the larger the number of different compounds involved in the reaction network, the harder it is to keep track of all the transformations that happen in the vessel, as many of them are taking place in parallel. Therefore, the system is often split up into the corresponding single reaction steps that are then studied individually under the desired conditions, i.e. within *multiple pots*. While this is easier to characterize, it also reduces the scope of the experiment, i.e. the window of opportunities for emergent chemical phenomena.

Moreover, the chemical reactions involved in metabolism often occur in a heterogeneous media that involve different environments (i.e. different physical and/or chemical conditions) with reactions potentially occurring at the interfaces between or across two environments (recall the concept of vectorial metabolism from Ch. 2). Recreating such a heterogeneous reaction medium in the lab may not be difficult to prepare with the knowledge gained so far about colloidal systems or lipid technologies, but the analysis becomes, again, much more cumbersome. Furthermore running a chemical reaction that is coupled with an interface forming system (like a vesicle) can be difficult, as experimental conditions necessary for the formation and maintenance of such interfaces can be different than the conditions necessary to run the reaction.

Last but not least, all the processes involved in metabolism do not necessarily occur under the same time scales. The above mentioned formation of interfaces and other supramolecular structures, or the characteristic lifetimes of macromolecular compounds and their corresponding mechanisms (like the catalytic action of enzymes) usually involve significantly larger time scales than the reactions coupled to them (or rather, controlled by them). Furthermore, on the level of the reaction processes themselves, a separation of

timescales can occur as some reaction pathway might be faster compared to the kinetics of another pathway. This creates an additional methodological challenge when trying to experientially recreate a system of coupled processes (not the coupling between reaction pathways or between reactions and supramolecular structures) as some reactions might already have finished before the other reactions have started or a certain structure has formed.

To be fair, all of the above outlined methodological challenges are not so relevant for most branches of chemistry. For instance, in some analytical chemistry experiment it might be perfectly sufficient to run a single reaction step in a pure solution and analysing the products once the reaction has equilibrated. But for prebiotic chemistry and the study of potential proto-metabolic pathways these challenges can be regarded as rather fundamental and therefore pose serious obstacles that future generations of chemists should try to overcome. In summary, all of these methodological challenges underline again that metabolism is a rather complex system and that the coupling between synthesis and control involves careful matching and synchronization mechanisms. More precisely, any dynamic organization of material compounds and transformation processes, in order to run coherently, requires their spatial and temporal coordination.

4.2 In silico challenges

As it was mentioned in the introductory chapter of this dissertation, major drivers in the advancement of systems sciences were the improvements made in the use computational resources. In fact, some of the above mentioned experimental challenges can, in principle be addressed by theoretical, computational models. Examples include the possibility to simulate out-of equilibrium phenomena using systems of ordinary differential equations (ODEs) or the usage of machine learning techniques to interpret and analyse large amounts of experimental data.

However, there are also some shortcomings associated with these types of approaches. For starters, implementing a computer simulation (like the formulation of a set of ODEs) requires some a priori knowledge of the system. In other words, one already needs to know which reactions are probably going to occur within the system, which is not always possible in a heterogeneous reaction medium hosting a diversity of compounds/reactions (like it seems

to be the case with metabolism). Furthermore, classical ODE models cannot include the often concurrent dynamics (i.e. chemical reactions and processes that happen in parallel) of such complex reaction systems.

This, in general, translates to a similar problem with time-scales as it was mentioned for the experimental approaches. Simulating a system where processes happen at different time-scales is not easy and in practice one often needs to make a choice on which process the focus is put and which processes are going to be ignored, as they either happen too fast to be noticeable, or too slow to be relevant. This means that it is generally difficult to simulate a system that involves complex constraints that are sometimes even the result of the reactive system itself. As such, many simulation approaches simplify the influence of constraints as scalar parameters that are then manipulated.

Again, to be fair, these methodological challenges are not necessarily a problem for most applications of computer simulations or data analysis techniques. However, when trying to simulate a minimal metabolic system one should nevertheless attempt to capture the overall complexity that is underlying such a system (compare the critiques of the classical models of metabolism in Ch. 3).

4.3 Proposal for a combined workflow

Some of the previously outlined methodological challenges for the research of minimal metabolism (associated with both, in-vitro as well as in-silico approaches) may well be quite difficult to overcome. However, as mentioned above, some of the problems that may arise in experiments can potentially be addressed by using computational resources (like analysing large amounts of data). At the same time, the knowledge coming from experimental results can be used as parameter inputs for simulations etc. In other words, the problem of the origin of life is not going to be solved just by doing experiments or computer simulations. What is rather needed is a close interaction between theoretical, computational models and experimental approaches.

An overview of how such an in-silico/in-vitro "work cycle" should operate in practice is depicted in Fig .4.1[1]. The proposed cycle involves subsequent rounds of design, performance, and interpretation of experiments or

[1] This graph was originally inspired by [Amilburu et al. 2020] (in particular, Figure 1 and the contents of section 4, in there.)

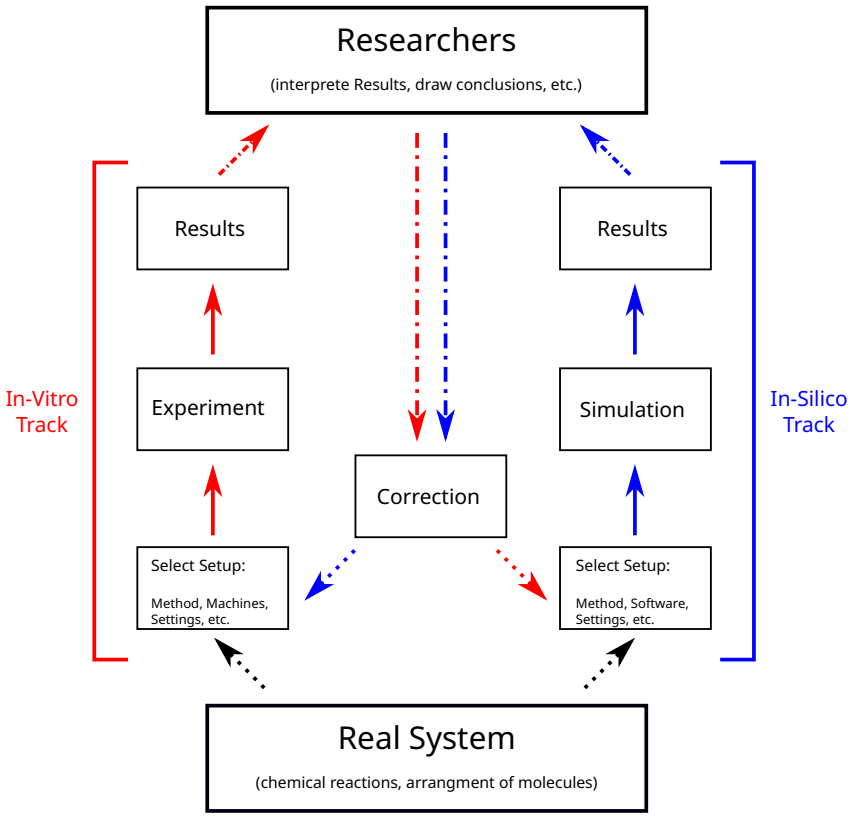


Figure 4.1: Tentative scheme of the interaction between in-silico and in-vitro approaches. At the bottom lies a certain natural phenomenon that is intended to be investigated. One can then either follow an in-vitro approach: selecting a setup and conducting an experiment; or follow an in-silico approach: selecting a software platform and conducting a simulation. Both approaches produce complementary results that help the corresponding researchers in their interpretation of the initial phenomenon and in the design of subsequent research steps (introducing corrections to either the experimental and/or the computational setup). Explanation of arrow relations: dotted arrows indicate influences, normal arrows indicate productions, dash-dotted arrows indicate interpretations. Figure taken from [Lauber et al. 2021].

simulations eventually leading to a recursive loop between both types of approaches (reminiscent of a 'figure eight'). For example, before immediately attempting to fix or redesign an experimental setup due to newly obtained data and insights, a potentially more productive strategy would be to take some more time and make use of this data/insights as the input for a closely related computational, theoretical model (e.g. constraining certain parameters, etc.). The results from this simulation can then be used to put together a more complete/adequate interpretation of the system under investigation and refine the initial experiment (e.g. the setup, settings, etc.) to redo the measurements. The process would repeat from there, performing as many cross-iterations as required. The outcome of all this is a cyclic process in which experimental work advances hand in hand with theoretical research, each providing constant feedback to one another, thus enhancing both understanding and productivity.

It is furthermore important to note, as well, that the scheme of Fig. 4.1 should in general not be conducted by a single person but through teams of researchers from different disciplines (using either in-silico or in-vitro methods) each engaging in close collaboration and exchanging valuable views and ideas. Thus the proposed workflow should also serve as an opportunity for these researchers to push each other to move out of the "comfort zone" of their respective field and try to address new relevant aspects of the phenomena they usually deal with. In any case, this workflow is simply a proposal, not to be taken as a strict paradigm that all future origins of life research should adhere to. It is rather an overall suggestion to encourage interdisciplinarity in OL research, with all the potential benefits that it could bring about. In fact, in the chapters that follow, a number of computational tools are going to be offered to start working in this direction, but they are just a first step in the cycle. While they have been motivated by current empirical knowledge from the OL research field, it will be obvious that a tighter connection with experimental data can still be pursued, as it will be commented in the final part of the dissertation.

Part II

Complex Reaction Networks and Simple Boundary Conditions

Chapter 5

Prebiotic Carbon Fixation

As outlined in the introductory chapter, the main idea of this thesis is to investigate the concept of minimal metabolism (MM) through computational, theoretical models. However, the scheme for MM that was motivated in Ch. 3 (see also Fig. 3.6) in principle involves *complex reaction networks* that are coupled with *complex boundary conditions*. Modelling and investigating such a system, as also discussed in Ch. 4, is in general not an easy task. Therefore, the second major part of this thesis is going to be about modelling complex chemical reaction networks[1] under *simple boundary conditions*. Thus a "step back" is going to be taken, so to speak, from trying to capture the full scale complexity of a minimal metabolic system, by putting the focus on the chemical reaction aspect of minimal metabolism (i.e., by considering first the lower plane of Fig. 3.6). In this context, to make things more concrete, the idea is to study the chemical reaction space associated with carbon fixation pathways and to determine if such a type of chemical reaction network could run under prebiotic conditions. This assumption is part of the general idea that the first proto-metabolic systems might have been autotrophic.

[1] The terms 'reaction network' and 'chemical reaction network' will be used interchangeably. While this is not a problem within the scope of this thesis, in general there is a fundamental difference between these two terms, as it was nicely discussed here [Müller et al. 2022].

5.1 Autotrophy at the origins of life

Most theories about the emergence of metabolism assume that it was, at first, heterotrophic. As shown in Fig. 5.1a, the overall idea of a heterotrophic proto-metabolism assumes that large organic molecules were already available in an early earth environment and were then broken down by simple catabolic reactions, supplying the energy for the anabolic build up of biomass (compare also the characteristic bow tie motif of heterotrophy from Fig. 2.1b). Potential source for these organic compounds range from "internal sources" like volcanic activities or atmospheric reactions to "external sources" like impacts of carbon containing meteorites. The most prominent example in line with this assumption is the Oparin-Haldane hypothesis, mentioned in the introductory chapter, whose starting point consists of a prebiotic soup enriched in organic compounds (resulting from reactions in a reducing atmosphere) that were used by fermentation pathways [Oparin 1924; Haldane 1929]. Other theories proposing a heterotrophic OL include the idea of retrograde evolution by Norman Horowitz [Horowitz 1945, 1965].

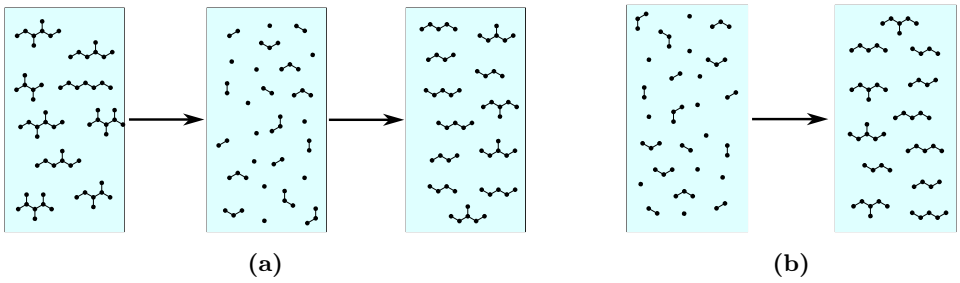


Figure 5.1: Comparison between auto- and heterotrophic OL scenarios. **(a)** In a heterotrophic scenario, large and complex molecules coming from abiotic processes are broken down to smaller compounds (e.g. through simple fermentation pathways) which are then used to build up larger compounds again. **(b)** In an autotrophic scenario, small inorganic carbon molecules are directly build up to form larger compounds.

A contrasting theory for the origins of metabolism assumes that it was autotrophic from its very beginning. As shown in Fig. 5.1b, the overall idea of an autotrophic proto-metabolism assumes that the early earth environment only contained small, inorganic carbon compounds which were then, together with energy from the environment, used to directly build up biomass via anabolic reactions (compare also the characteristic motif of autotrophy from Fig. 2.1b). One of the first ideas for such an autotrophic OL was proposed by

Günther Wächtershäuser [Wächtershäuser 1988, 1990]. He considered that the atmosphere of the early earth only contained CO_2 as well as hydrogen and sulphur. These simple compounds were then used in chemical reactions taking place on the surface of rocks within hydrothermal vents, using the energy from these vents, as well as iron as catalyst (thus his theory is often referred to as the 'iron-sulphur world'[2]). As a chemical reaction pathway, occurring in this hydrothermal vent scenario, Wächtershäuser suggested the rTCA cycle (a variant of the TCA cycle found in some bacteria that runs in reverse). A similar idea concerning the rTCA has also been suggested by Harol J. Morowitz [Morowitz 1992; Morowitz et al. 2000; Smith and Morowitz 2004].

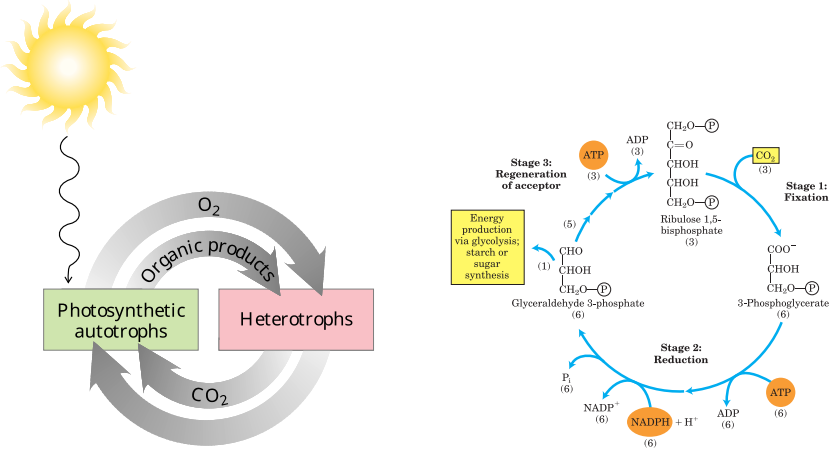
The appeal for a heterotrophic OL usually lies in the simplicity of the reaction pathways that are associated with it as well as their wide spread within extant biology (e.g., fermentation), while reactions related to autotrophy are regarded as more complex [Peretó 2005, 2012]. However, there are several counterarguments that one can make against a heterotrophic OL. First of all, heterotrophic pathways appear simple from a viewpoint of a top-down projection. In reality, preliminary versions of heterotrophy/autotrophy could involve a similar number of reaction steps, enzymes catalysing them, etc. Furthermore, the feasibility of a primordial soup enriched in organic matter (like in the Oparin-Haldane hypothesis) is also questionable. In fact, Wächtershäuser himself showed that such a scenario is rather unlikely from a thermodynamic point of view [Wächtershäuser 1988]. Additionally, heterotrophic scenarios typically invoke a reducing atmosphere (i.e. containing compounds like CH_4,NH_4, etc.), whereas recent studies suggest that the atmospheric composition of the early earth was more neutral, and richer in CO_2 (see for example [Rubey 1951, 1955]). Last but not least, autotrophy is, sooner than later, "unavoidable". Even if there was a prebiotic soup of organic compounds, which are taken to fuel metabolism, it will eventually be depleted and all that is left available to run biosynthetic pathways and produce biomass are small carbon compounds. All in all, autotrophy constitutes a perfectly plausible and justifiable candidate for the origins of metabolism, which is why the focus of this part of the dissertation is to investigate chemi-

[2]Due to Wächtershäusers stark opposition towards the primordial soup theory and the fact that his proposal involves reactions on a surface, it has also been referred to as "primordial pizza" hypothesis.

cal reaction networks that are associated with that specific metabolic mode.

5.2 Carbon fixation pathways

Autotrophy within extant biology is, in principle, realized through a process called 'carbon fixation': biochemical pathways which, as the name suggests, convert small, inorganic carbon compounds like CO_2 or, in some cases, HCO_3 into larger organic molecules, i.e. biomass. Therefore, assuming that autotrophy was at the origin of life effectively means that the first metabolic pathways most likely involved such carbon fixation reactions.



(a) **(b)**

Figure 5.2: Carbon fixation in extant biology. **(a)** Biosphere level cycling of CO_2, O_2 and biomass due to the interplay between autotrophic and heterotrophic organisms. **(b)** General reaction scheme for the Calvin cycle converting CO_2 into glyceraldehyde 3-phosphate (G3P). Figures taken from [Nelson and Cox 2017].

As shown in Fig. 5.2a, most carbon fixation within the biosphere is nowadays done by plants and algae (as well as some bacteria) through photosynthesis, where atmospheric CO_2 together with water and energy from sunlight (hence the more specific term photoautotrophy) are converted into O_2 and biomass. This biomass then serves as nutrients for heterotrophic organisms, which use the O_2 to break them down again, releasing CO_2 in the process and, thus, closing the global carbon cycle. In general, the overall process of photosynthesis uses several different biochemical reaction pathways, with the actual carbon fixation happening in the so called reductive pentose phosphate cycle, or Calvin-Benson cycle (see Fig. 5.2b). However, as important

as the Calvin-Benson cycle is today, it is probably a late stage development[3], because it is a rather complex pathway, involving several reactions that require large amounts of energy, together with a set of specific enzymes. Thus a simpler reaction pathway would be a preferable candidate for prebiotic carbon fixation.

One scenario, proposed by Bill Martin and Mike Russell, assumes that a non enzymatic version of the Wood-Ljungdal (WL) pathway [Ljungdahl and Wood 1965] could have operated within the pores of alkaline hydrothermal vents [Martin and Russell 2003; Russell and Martin 2004; Martin and Russell 2007]. As shown in Fig. 5.3, the WL-pathway (or reductive acetyl-CoA pathway) is a linear reaction pathway that transforms CO_2 into acetate (or acetyl-CoA). It is occurring in acetogenic bacteria [Fuchs 1986] as well as in methanogenic archaea [Berg et al. 2010], which makes it an appealing candidate for prebiotic carbon fixation, serving as a potential "connector" between these two domains of prokaryotes. Additionally, the proposed scenario of alkaline vents provides a plausible explanation for the existence and development of proton gradients across cell membranes [Lane 2015], which serves, as discussed in Ch. 2, a vital function within extant metabolism. Furthermore, the pathway product, acetyl-CoA, is also a central compound of metabolism serving as connection between many metabolic pathways (e.g., fatty acids are generally synthesized from acetyl-CoA, thus the WL-pathway could serve as a potential entry point into a prebiotic fatty acid production). However, it has yet to be demonstrated that the WL-pathway can truly operate under the proposed prebiotic conditions in the absence of enzymes and co-factors etc.

A different type of carbon fixation pathway (shown in Fig. 5.4a) is the reductive tricarboxylic acid (rTCA) cycle, or Arnon-Buchanan cycle, that certain bacteria carry out [Evans et al. 1966; Buchanan and Arnon 1990]. As the name suggests, this circular reaction pathway basically is the TCA cycle running in the reverse (or reductive) direction. Thus, it converts CO_2 into acetyl-CoA and oxaloacetate (instead of oxidising acetyl-CoA and releasing CO_2 in the process). As mentioned above, the rTCA is a candidate for primordial carbon fixation in scenarios proposed by Wächetershäuser

[3]Many theories suggest that the development of photosynthesis and the Calvin-Benson cycle led to the oxygen rich atmosphere of today which then further enabled the development of more complex heterotrophic organisms eventually leading to a biosphere as depicted in Fig. 5.2a. This is usually referred to as the great oxidation event (GOE).

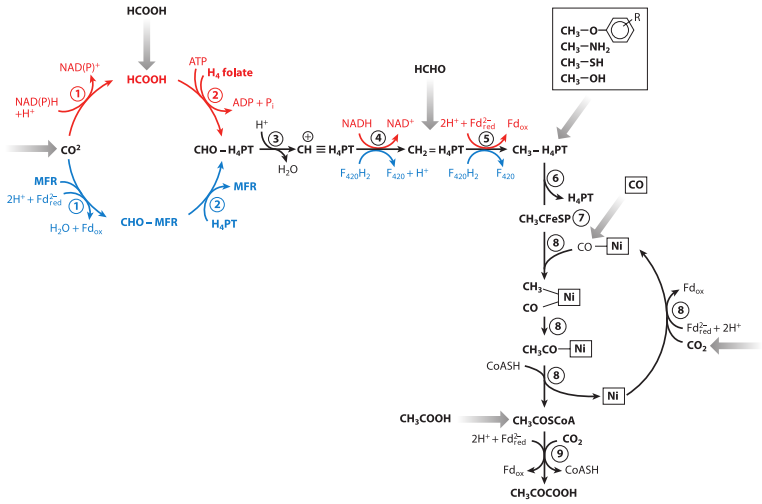


Figure 5.3: General reaction scheme of the Wood-Ljungdal (WL) pathway. Red arrows indicate the pathway variant found in acetogenic bacteria while blue arrows indicate the variant found in methanogenic archaea. Figure taken from [Fuchs 2011].

[Wächtershäuser 1988, 1990] and Morowitz [Morowitz 1992; Morowitz et al. 2000; Smith and Morowitz 2004]. Recent studies have shown that parts of the rTCA cycle can in fact run without enzymes, using different transition metals as catalysts [Muchowska et al. 2017]. Other research groups follow a somewhat different strategy: taking the rTCA cycle as a reference, they do not try to directly implement it, but rather explore other pathways in its "chemical vicinity" as potential precursors, which could be a more reasonable bottom-up scenario [Stubbs et al. 2020]. In any case, the general idea is that the TCA cycle initially ran in the reverse direction, fixing atmospheric CO_2, that was probably abundant in the early earth atmosphere (a recent study also showed that high levels of CO_2 favour the reductive directions of the TCA cycle [Steffens et al. 2021]), and building up biomass. Later on, with rising levels of O_2 in the atmosphere, the reaction cycle then switched into the oxidative direction that is widespread in extant biology. Therefore, the rTCA found in bacteria today would be a remnant of this time (i.e., a "metabolic fossil").

Furthermore, the rTCA cycle shares several reactions with three other carbon fixation pathways that occur in procaryotes: the 3-hydroxypropionate (3HP) bicycle [Zarzycki et al. 2009], the 4-hydroxybutyrate (4HB) cycle [Huber et al. 2008] and the 3-hydroxypropionate/4-hydroxybutyrate

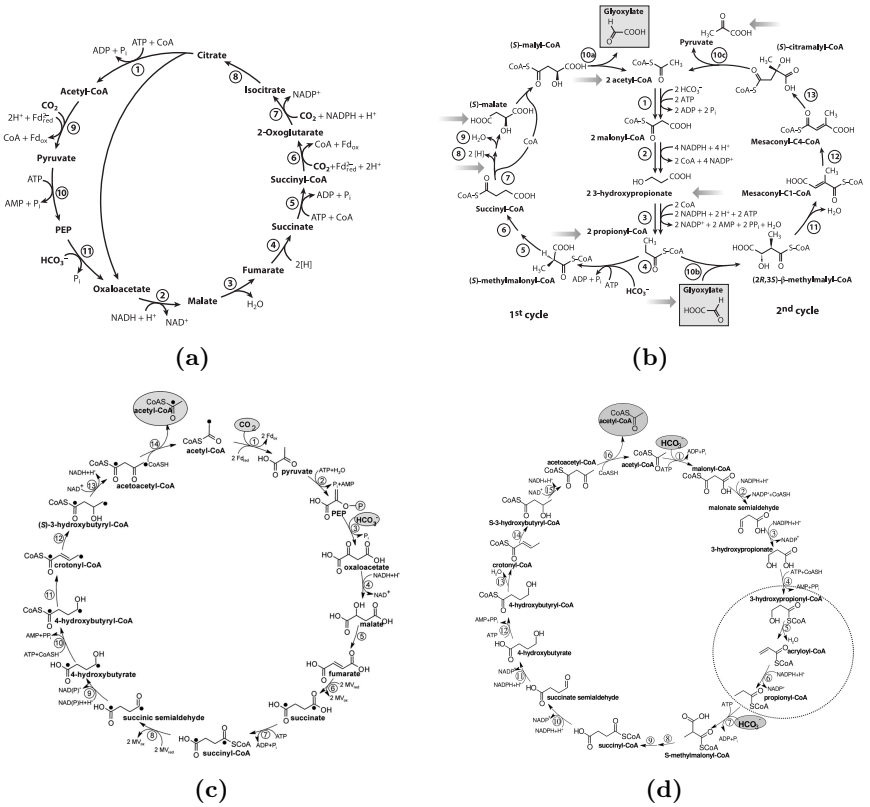


Figure 5.4: General reaction schemes for: **(a)** the rTCA cycle, **(b)** the 3HP bicycle, **(c)** the 4HB cycle, **(d)** the 3HP/4HB cycle. Figures taken from [Fuchs 2011], [Zarzycki et al. 2009], [Huber et al. 2008] and [Teufel et al. 2009] respectively.

(3HP/4HB) cycle[4] [Teufel et al. 2009] (shown in Fig. 5.4b, Fig. 5.4c and Fig. 5.4d respectively). Eric Smith referred to these four cycles as the *carbon fixation core* [Braakman and Smith 2013]. In this context, the focus is going to be put on the rTCA cycle: more precisely, on the chemical reaction space associated with the chemistry of the rTCA cycle, which is going to be examined in more detail, trying to find viable pre-biotically plausible "alternative" pathways that might have been precursors to the extant rTCA (the exact computational methods used for the task will be explained in the next chapter). However, before delving into that, it is necessary to consider the fact that the rTCA cycle, as well as the other extant carbon fixation

[4]All the six reaction pathways mentioned in this section (Calvin cycle, WL-pathway, rTCA cycle, 3HP bicycle, 4HB cycle and 3HP/4HB cycle) are the only known carbon fixation pathways in extant biology. However there are also studies of synthetic carbon fixation pathways, both in silico [Bar-Even et al. 2010] and in vitro [Schwander et al. 2016].

pathways (except for the WL-pathway) are autocatalytic cycles, since this property has important implications on a chemical reaction network and may turn out to be an important feature for prebiotic carbon fixation.

5.3 Autocatalysis

In Ch. 3, the concept of autocatalysis, or more precisely (reflexively) autocatalytic sets, was reviewed as one of the "classical" models of metabolism. There a rather formal definition was given, where a set of molecules and reactions is considered autocatalytic, if, starting from a set of food molecules, each molecule in the set is created by at least one reaction from it and, in turn, each reaction is catalysed by at least one molecule from the set [Hordijk and Steel 2004; Kauffman 1986]. It was found that in a system of randomly catalysed reactions with a fixed catalysing probability, such an autocatalytic set would spontaneously emerge, provided that the system becomes sufficiently large [Hordijk and Steel 2004; Kauffman 1986].

Within this section, the idea of autocatalysis is going to be revisited in a more chemically "instructive" or "practical" way, i.e. analysing what effect does an autocatalytic reaction system have on the concentration of its products. Additionally, the topology of chemical reaction networks that are autocatalytic is going to be examined in more detail.

5.3.1 Catalytic and autocatalytic reaction cycles

For a chemical reaction a catalyst usually refers to a certain molecular species (e.g. an enzyme or a certain inorganic compound) that lowers the activation energy necessary to initialize it, thus effectively speeding up the corresponding reaction rate. This catalysing action can, in practice, also enable reactions that would not be feasible (i.e., that would take too long to occur) under normal conditions, given their high activation energy. While the educts of a reaction get transformed into the products, the catalyst either remains unchanged or gets replenished at the end (i.e. effectively remaining unchanged). Thus one can write the reaction equation for a simple, catalytic reaction as:

$$\mathrm{F} + \mathrm{A} \xrightarrow{\mathrm{k}} \mathrm{B} + \mathrm{A} \tag{5.1}$$

where the species A acts as a catalyst for the production of B from F (which can be viewed as a food species). As A gets replenished, it is on both sides of the reaction equation. There are various ways to visualise the catalytic effect of A on the reaction (see Fig. 5.5a). The overall reaction of Eq. 5.1 is in principle a cyclical reaction, as shown in Fig. 5.5c, where each "turn" of the cycle creates a product B and replenishes the catalyst A, enabling further reactions. One can thus generally speak of 'catalytic reaction cycles'. A special case of catalytic action would be the following example of a simple reaction equation:

$$\mathrm{F} + \mathrm{A} \xrightarrow{\mathrm{k}} 2\,\mathrm{A} \tag{5.2}$$

where A acts as a catalyst for *its own* production from F. In this way, Eq. 5.2 represents an elementary 'autocatalytic reaction', depicted graphically in Fig. 5.5b. Like in the general case of a catalytic reaction, Eq. 5.2 can be considered a cyclical process. Yet, as shown in Fig. 5.5d, by contrast to a catalytic cycle, each turn of an 'autocatalytic cycle' creates two copies of A, i.e. not only does the initial catalyst maintain in the system but also an additional copy of it is produced[5]. This is the reason why sometimes also speaks of autocatalysis has frequently been interpreted as a mechanism for 'self-replication' [Peretó 2012], even if the idea of self-replication tends to imply 'sequence conservation', as well (particularly when the involved molecules are polymers).

Using mass action kinetics, the reaction dynamics associated with a simple, catalytic cycle (given by Eq. 5.1) can be converted into a set of ordinary differential equations (ODEs):

$$\begin{aligned} \frac{\mathrm{d}\,[\mathrm{F}]}{\mathrm{d}t} &= -k\ [\mathrm{F}]\ [\mathrm{A}] \\ \frac{\mathrm{d}\,[\mathrm{B}]}{\mathrm{d}t} &= k\ [\mathrm{F}]\ [\mathrm{A}] \\ \frac{\mathrm{d}\,[\mathrm{A}]}{\mathrm{d}t} &= 0 \end{aligned} \tag{5.3}$$

Solving these ODEs, expressions for $[\mathrm{F}]\,(t), [\mathrm{A}]\,(t), [\mathrm{B}]\,(t)$, are obtained, which are the changes of concentration over time for species F, A, B respectively. Fig. 5.6a shows the trajectory for $[\mathrm{B}]\,(t)$, i.e. for the product species B of Eq 5.1. One can see that in a catalytic cycle, the concentration of the

[5]Reaction cycles (catalytic or autocatalytic) often involve multiple reactions and intermediates, until the (auto)-catalyst A gets replenished and the cycle is eventually closed.

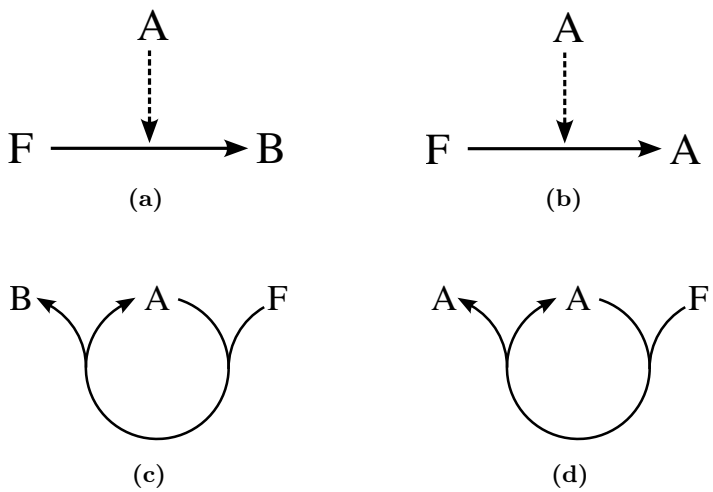


Figure 5.5: Different ways to depict **(a)**,**(c)** a simple catalytic reaction/cycle or **(b)**,**(d)** an autocatalytic reaction/cycle. Full arrows indicate reaction products while dashed arrows indicate catalytic actions.

reaction product increases linearly, at first, until all the food compounds F are converted and a steady state is reached, resulting in a logarithmic profile. Following a similar procedure, the reaction dynamics for an an elementary autocatalytic cycle (given by Eq. 5.2) can be converted into a set of ODEs:

$$\begin{aligned} \frac{\mathrm{d}\,[\mathrm{F}]}{\mathrm{d}t} &= -k\ [\mathrm{F}]\ [\mathrm{A}] \\ \frac{\mathrm{d}\,[\mathrm{A}]}{\mathrm{d}t} &= k\ [\mathrm{F}]\ [\mathrm{A}] \end{aligned} \tag{5.4}$$

These can again be solved, to obtain expressions for $[\mathrm{F}]\,(t)$ and $[\mathrm{A}]\,(t)$. Fig. 5.6b shows the trajectory of $[\mathrm{A}]\,(t)$, i.e. the product species A of Eq. 5.2. In contrast to the situation observed for a catalytic cycle, the concentration of the reaction product in an autocatalytic cycle, after a short "lag phase", increases *exponentially* (see also Fig. 5.6c for a direct comparison of the first time-steps) [Schuster 2019]. Again, once all the food compounds F are converted, a steady state is reached, resulting in a sigmoidal profile. Thus in practice, when running a chemical reaction, the observation of such an "S-shaped" curve for the change of concentration of the reaction product indicates that the investigated reaction system involves an autocatalytic cycle [Bissette and Fletcher 2013].

In the context of an autotrophic OL, a prebiotic carbon fixation pathway

that is autocatalytic would thus imply an exponential increase in biomass. Additionally, if this exponential increase is faster than potential degradation reactions and/or the dilution of the reaction product, a high concentration of increasingly elaborate carbon compounds could be maintained. This is one of the main reasons why the rTCA cycle constitutes a strong candidate for prebiotic carbon fixation (together with other autocatalytic reaction cycles that probably played a major role at the origin of life[6]). Searching for AC-cycles in a larger system of chemical reactions, therefore, seems quite critical, which is why it is going to become one of the focus of interest in this work. Nevertheless, identifying such cycles within a network of chemical reactions requires a more formal definition of autocatalysis, which is going to be approached in the following.

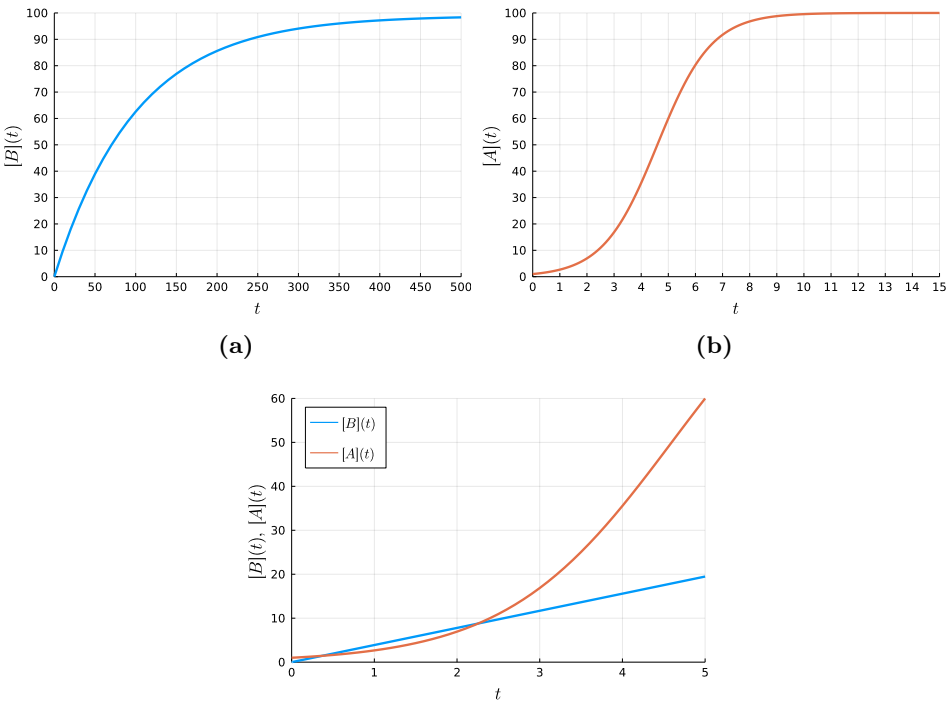


Figure 5.6: Different trajectories for the product species of Eq. 5.1 or Eq. 5.2. **(a)** Trajectory of the catalytic species [B] (t). **(b)** Trajectory of the autocatalytic species [A] (t). **(c)** Comparison of catalytic and autocatalytic trajectories.

[6]However, unlike other authors, e.g. [Xavier et al. 2020], the appearance of AC-cycles is seen here as necessary but not sufficient for life.

5.3.2 Autocatalytic reaction networks

The previously discussed catalytic and autocatalytic cycles were elementary examples that involved only one reaction. However, in practice one must deal with a system of coupled reactions that involve multiple compounds resulting in a network of chemical reactions. A good approach to study the properties of such chemical reaction networks (CRNs) is to try and formalize them. In general, a chemical reaction describes the transformation of a set of educts to a set of products, i.e. the molecules of the left hand side (LHS) and the right hand side (RHS) of the reaction equation respectively. Thus, as shown in Fig. 5.7, a system of reactions can be formalized as a 'directed, bipartite, multi graph' [Newman 2010][7] where one type of vertices represent the molecules while the other type represents the reactions. The direction of the edges (i.e. arrows), indicates whether molecules go into a reaction as educts, or come out of a reaction as products, with the multiplicity of the edges indicating the stoichiometry of a reaction. The examples in Fig. 5.7 also show that the (auto)-catalytic effect discussed above is in general not the action of a single molecule, but rather a set of species (in this case A_1,A_2 and A_3) that collectively convert a set of food compounds (F_1 and F_2) into a product compound (B or A_1) [Andersen et al. 2021].

Formalizing a CRN as a bipartite graph allows to investigate the topological properties of an autocatalytic cycle or, more precisely, an 'autocatalytic network'. In other words, it allows to identify what kind of network topology, or what type of motive, is characteristic for a a chemical reaction system that displays autocatalytic behaviour. Blokhuis *et al.* distinguished and classified five different types of network motifs, depicted in Fig. 5.8, that can lead to autocatalysis.

Comparison between these types of autocatalytic motifs, interestingly, reveals some common features among them. More precisely, one can see that each motif consists of a reaction cycle with at least one 'fork reaction' (highlighted in red in Fig. 5.8) that creates two products from a single educt. While in the type I motif these products are two equal copies of a compound (which is the autocatalytic species, similar to the example in Fig. 5.7b), in the types II-IV the products are different. What makes these cycles then

[7] A 'graph' is a mathematical object that is used to represent a network. The nodes of the network being a set of 'vertices' and the links being a set of 'edges'. A 'bi-partite graph' in this context has two types of vertices.

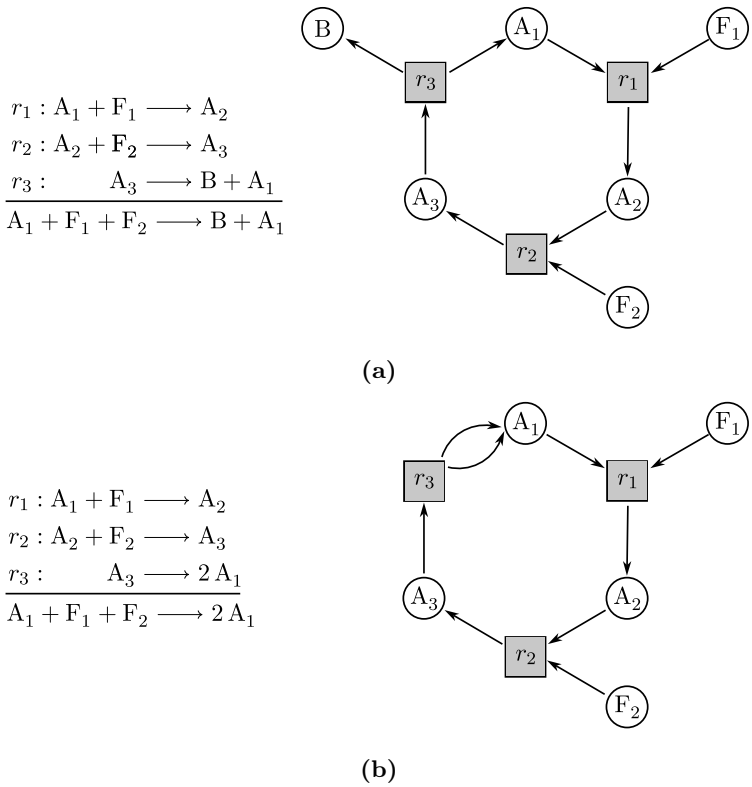


Figure 5.7: Examples for **(a)** a set of catalytic reactions and **(b)** a set of autocatalytic reactions as well as their respective representations as directed, bi-partite, multi graphs. Round vertices represent compounds while square vertices represent reactions.

autocatalytic (in contrast to the example of Fig. 5.7a) is the fact that while one of the products remains in the cycle, the other product "feeds back" into the cycle through a different reaction later on. Accordingly, these fork reactions create a "side branch" (or sub-cycle) to the main reaction cycle and the compound where the side branch reunites with the main cycle is the autocatalytic species (highlighted in blue in Fig. 5.8). While the types II and III have only one fork reaction and thus two reaction cycles, type IV has two fork reactions with three reaction cycles and type V has three fork reactions with four reaction cycles. In practice, the fact that an autocatalytic reaction cycle always contains at least one fork reaction and, subsequently, a merge point allows to formulate a constructive algorithm to search for such cycles within a larger system of reactions, which will be outlined in the next chapter.

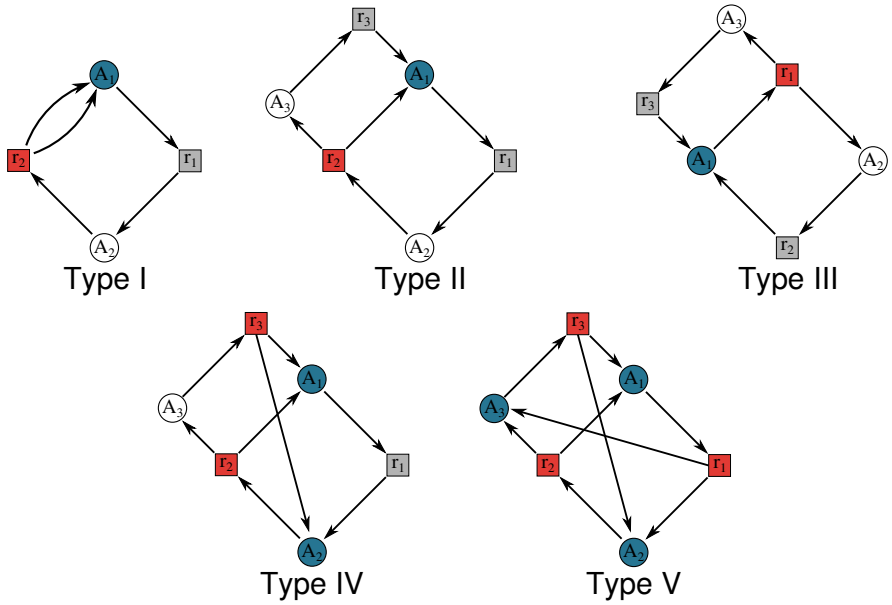


Figure 5.8: Minimal examples for the five possible motifs of CRNs that can display autocatalytic behaviour. As indicted above, round vertices represent compounds while square vertices represent reactions. Fork reactions are further highlighted in red, while autocatalysts are highlighted in blue. In general, edges from a compound vertex to a reaction vertex can be expanded to include additional intermediary compounds and reactions (but not the other way round). Figure adapted from [Blokhuis et al. 2020].

In connection to the five types of autocatalytic motifs (compare again Fig. 5.8), examples for the types I and II can be found in extant biology, more specifically among the above discussed carbon fixation pathways that are autocatalytic (compare Fig. 5.2b and Fig. 5.4). While both versions of the 4HB cycle can be classified as type I cycles, the rTCA cycle as well as the 3HP cycle and the Calvin cycle are examples of type II cycles. Additional abiotic examples for the type I and type II motif can be found in the Formose reaction [Blokhuis et al. 2020]. For the type III only chemical examples can be found, like the auto amplification of cystamine discovered by Semenov *et al.* [Semenov et al. 2016]. Finally for both, type IV and type V, no examples (neither naturally occurring nor synthetic) have been found so far (however, they can be regarded as variations of the type II cycle with one or two additional fork reactions).

Last but not least, it is quite relevant to mention that autocatalysis within this section was mostly regarded in the context of chemical reactions

–i.e., self reinforcing reaction cycles. However, autocatalysis can be seen as a general concept of self reinforcement or self replication (like the self replication of a template polymer or a compartment – see [Peretó 2012]). Furthermore, autocatalysis does not need to involve cycles that are formed through the coupling of chemical reactions. It is also possible to have an autocatalytic coupling between a reaction and some macromolecular structure (i.e. a boundary condition) or the coupling of linear pathways across interfaces (as it has already been argued in the literature – see [Bissette and Fletcher 2013; Blokhuis et al. 2020]). Nevertheless, within this dissertation autocatalysis is going to be investigated in its "classical" (strictly chemical) form, i.e. as a self reinforcing coupling within chemical reaction networks.

Chapter 6

Expansion of Chemical Reaction Spaces

"Classically", molecular reaction mixtures are modelled by formulating sets of ODEs using mean field approaches, like mass action kinetics or Michaelis-Menten kinetics, and then solving these differential equations to obtain trajectories that show the change of concentration of the molecules in the system. It is then analysed how the change of parameters like the reaction rate constants or the initial concentrations of compounds influence the obtained time traces. In other words, the microscopic behaviour of the system (the collision and reaction of molecules in the system) is modelled through macroscopic variables and parameters. This approach might work for modelling simple mixtures that involve a small selection of different molecular species and whose reactive mechanisms are already well known. However, when studying more complex reaction mixtures that contain a large and diverse set of molecules this macroscopic modelling approach fails to capture the microscopic dynamics of concurrently occurring reactions, with the corresponding loss of relevant information.

One way to address this difficulty and model such complex reaction mixtures is by expanding the underlying network of chemical reactions, which allows to, potentially, study the whole possibility space of chemical reactions that may occur within the system. There have been previous attempts by Goldford *et al.* at expanding and investigating reaction networks (e.g. [Goldford and Segrè 2018; Goldford et al. 2019] or more recently [Goldford et al. 2022]), but their modelling approach is more akin to searching for certain

pathways within an a priori defined reaction network (typically drawn from KEGG-type data bases). A more bottom-up network expansion, however, can be achieved through a *rule based* network expansion, where chemical reactions are formalized as general rules which are then iteratively applied to a set of starting molecules, thus expanding a reaction network (see for example: [Zubarev et al. 2015; Harris et al. 2016; Wołos et al. 2020]).

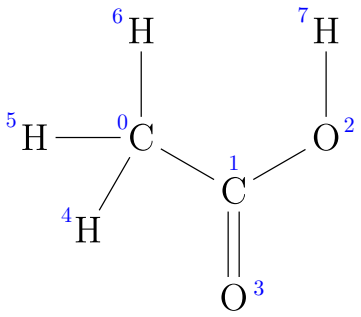
6.1 Rule based network expansion

The software tool to implement such a rule based expansion of chemical reaction networks here, is the chemistry oriented graph-rewrite framework *MedØlDatschgerl* (MØD) [Andersen et al. 2016; MØD]. As it is going to be outlined below, being a graph-based model, MØD allows a more flexible implementation of reaction rules, conceived as general graph rewrite patterns for the corresponding molecule graphs. Accordingly, changing the chemistry of the system can easily be achieved by changing these rewrite patterns. Furthermore, the graph representation of an expanded reaction network gives the opportunity to use other graph-based algorithms, including the search for specific sub-graph patterns for further analysis of the result of a reaction network expansion. Therefore, the following sections are going to give an overview about the way a chemical reaction system is formalized in MØD and how this can be used to perform a rule based expansion of the chemical reaction space.

6.1.1 Formalization of molecules

A molecule is, in principle, a collection of atoms that are connected via different types of chemical bonds. Thus, at the heart of MØD and it's graph-theoretic approach to chemistry, lies the assumption that one can represent any molecule as a 'simple, multi-labelled, undirected graph' [Andersen 2015; Andersen et al. 2016][1]. As shown in Fig. 6.1, this formalization implies that the atoms of a molecule correspond to the vertices of the appropriate 'molecule graph', where the label of each vertex is the element symbol of the respective atom (i.e. C-atoms are labeled with a letter "C", H-atoms are labeled with a letter "H", etc.). The bonds among the atoms furthermore

[1] This is actually not a new concept, see for example [Sylvester 1878].



V	E
0: "C"	(0,1): "-"
1: "C"	(0,4): "-"
2: "O"	(0,5): "-"
3: "O"	(0,6): "-"
4: "H"	(1,3): "="
5: "H"	(1,2): "-"
6: "H"	(2,7): "-"
7: "H"	

Figure 6.1: Example for the MØD implementation of an acetate molecule with the corresponding vertex-set V (i.e the collection of all atoms) and edge-set E (i.e the collection of all bonds).

correspond to the edges between the vertices, with the label of each edge indicating the type of bond (e.g. single bonds are labeled with a "-" symbol, while double bonds are labeled with a "=" symbol[2]). All the vertices and edges in such a molecule graph, as well as their respective labels, are collected in the vertex-set V and edge-set E respectively (see Fig. 6.1). Therefore, any molecule can, in principle, be completely described by the pair (V, E)[3]. In practice, molecule graphs are loaded into the MØD-framework using SMILES (Simplified Molecular-Input Line-Entry System) strings [Weininger 1988], but other formats such as GML (Graph Modelling Language) [Himsolt 1997] files are also possible.

6.1.2 Formalization of chemical reactions

A chemical reaction, in turn, can be generally conceived as the breaking or changing of existing bonds and/or the formation of new bonds between one or more molecules. Since molecules in MØD are formalized as graphs with their bonds represented as labelled edges, this implies that a reaction would be the deletion or relabelling of existing edges and/or the addition of new edges, within one or several molecule graphs. Furthermore, as some reactions can involve a change of the charge of an atom, this also includes the

[2]For completeness, a triple bond would be labeled with a "#" symbol, while an aromatic bond would be labeled with a ":" symbol. However, all the molecules considered in this work are free of triple or aromatic bonds.

[3]Note that this is only a 2D-representation of a molecule. It therefore does not include additional information about its spatial orientation, like it's stereochemistry. However, for the purpose of this project these types of molecule properties are not relevant.

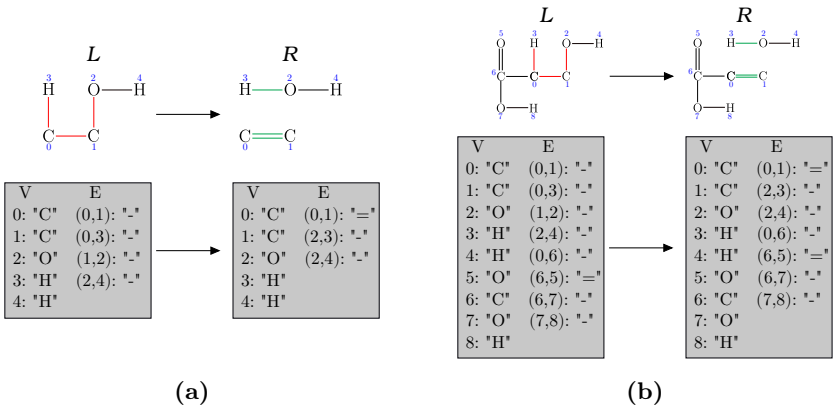


Figure 6.2: Two examples for the MØD implementation of a de-hydration reaction with the corresponding transformation of the vertices and edges. **(a)** General reaction rule with minimal context. **(b)** More specific reaction rule with more (or stricter) context.

relabelling of vertices. Therefore, within the graph-theoretical framework of MØD, a chemical reaction is interpreted as a transformation, or rewriting, of molecule graphs. However, the same type of chemical reaction can, in general, occur among different molecules. Therefore it is necessary that a graph transformation that represents a certain reaction is not only applied to one specific molecule graph (or a set of graphs), but rather that it is applied to any molecule graph (or a set of graphs) that may undergo this specific reaction. In other words, a reaction can be understood as a general rule for the transformation, or rewriting, of a molecule graph. Thus, within the MØD framework, chemical reactions are formalized as 'graph rewrite rules' [Andersen 2015; Andersen et al. 2016].

On these lines, as shown in the example of Fig. 6.2, a reaction rule, in general, defines the transformation from a graph L to a graph R which reflects the bond changes within a molecule between the left-hand side (LHS) and the right-hand side (RHS) of a reaction. Fig. 6.3 then shows that for any molecule graph (or educt graph) G, where L is found as a sub-graph, the respective rule is applied and the vertices and edges are transformed accordingly. In other words, the sub-graph of G that matches L is rewritten to a sub-graph that matches R, resulting in a new molecule graph (or product graph) H[4]. Like in the case of a real chemical reaction, note that, in general,

[4]More specifically, the way these graph rewrite-rules are implemented in MØD is by using a concept from category theory called a 'double pushout' formalism [Andersen 2015].

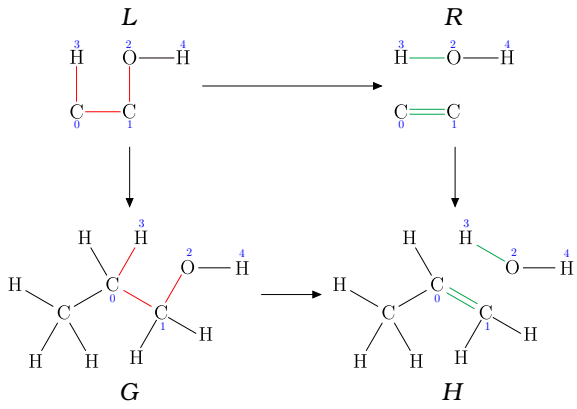


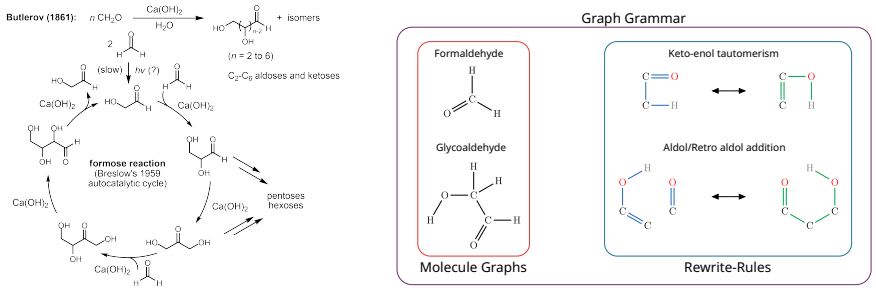
Figure 6.3: Application of the rule defined in Fig. 6.2a. Note that the rule defined by Fig. 6.2b could not be applied to the molecule graph G due to it's stricter context.

the educts and/or products can be more than one molecule, so the graphs and sub-graphs of this transformation can have more than one component.

As one can also see in Fig. 6.2, within the transition between L and R some vertices and edges remain unchanged by the transformation, even if they still need to be matched from a graph L onto a graph G. These structures that remain unchanged but are nevertheless necessary in order for the reaction rule to be applied, are referred to as 'context' which can be used to capture the fact that a chemical reaction usually occurs on specific functional groups and/or require that certain functional groups be present within the molecule. In the example given in Fig. 6.2 for a dehydration reaction, the context of the rewrite rule, as it is defined in Fig. 6.2a, makes sure that the reaction only occurs at an $-$OH group (rather than at an arbitrary $-$OR group). By contrast, the context for the rewrite rule, as it is defined in Fig. 6.2b, constrains the reaction to occur at an $-$OH group that is in a β-position to an acid group (i.e., two carbon atoms away from an acid group).

6.1.3 Formalization of chemistries

When putting together a reaction mixture, i.e. a selection of educt molecules, they will proceed to react with each other according to constraints given by the laws of thermodynamics (or physics in general) as well as kinetics. More specifically, factors like temperature, pH, redox potential, etc. determine the thermodynamic landscape that establishes whether certain reactions are en-



(a) **(b)**

Figure 6.4: Example of the MØD implementation for the chemistry of the formose reaction. **(a)** Reaction scheme for the standard formose reaction (or Breslow cycle)[5]. **(b)** Graph grammar associated with the chemistry of the formose reaction.

ergetically feasible or not. Furthermore, the concentrations of the molecules as well as the activation energies of the reactions determine the reaction kinetics, i.e. how fast a certain reaction occurs and if there are faster, parasitic reactions. Finally, the types of molecules themselves can play an important role, as certain molecules are more prone to react in a specific way under certain conditions (e.g. aldol-reactions tend happen under more basic conditions). In other words, not all reactions among a certain set of compounds are possible. All these determining factors can be summarized as the type of chemistry that is defined by the system at hand.

However, MØD, as a graph theoretical framework does not immediately include these thermodynamic and kinetic constraints. In other words, the graph-rewrite rules, which formalize chemical reactions, will always be applied, no matter if the corresponding reaction is kinetically and/or thermodynamically feasible. Therefore, within MØD, chemistries are formalized as a so called 'graph grammars'[6]. As shown in Fig. 6.4, such a graph grammar consists of a set of molecule graphs $\mathcal{G}$, representing the educt molecules initially present within the system, together with a set of rewrite rules $\mathcal{P}$. It is then through the choice of $\mathcal{P}$ that some of the constraints set by the chemistry that one intends to model are included, to make things more realistic. First, the addition of a specific context when defining a reaction rule can already amount to taking into account some thermodynamic aspects by

[5] *formose catalytic cycle* by Wikimedia-user Alsosaid1987, licensed under CC BY-SA 4.0

[6] This is because one can connect the context dependent graph-rewrite rules with a context-sensitive grammar from formal language models [Andersen 2015].

preventing the occurrence of a reaction at a molecule with an unfavourable geometry (compare again the examples in Fig. 6.2). Second, the selection of rules itself acts as a constraint, since it allows to focus on the relevant reactions that can realistically occur within the system, while actively ignoring certain reactions that could in theory happen but not in practice, because they are not thermodynamically and/or kinetically feasible.

A graph grammar with its set of starting compounds $\mathcal{G}$ and its set of reaction rules $\mathcal{P}$ is then the basis to do an expansion of the chemical reaction space. In practice, when implementing the graph grammar related to a certain type of chemistry, some a priori knowledge about it is needed, i.e. what molecules are typically involved and what reactions usually occur, in order to formulate $\mathcal{G}$ and $\mathcal{P}$. The formose reaction, portrayed in Fig. 6.4a, is a good example in which a well studied chemistry helps formulate the corresponding graph grammar (see Fig. 6.4b) in a rather straight forward manner. Furthermore, the a priori definition of a graph grammar also allows to explore the effects of modifying the chemistry of a certain system. This can be done by including compounds and/or reactions that are not typically associated with the chemistry under study or by excluding certain compounds and/or reactions. In this way, one can observe how the expanded network resulting from the modified grammar changes and make new predictions/interpretations of experimental data.

6.1.4 Derivation graphs

With the chemistry of a system formalized in the form of a graph grammar, one can continue to expand the corresponding chemical reaction space using the MØD framework. This expansion will be done iteratively according to the following algorithm:

(i) start with a set of molecule graphs $\mathcal{G}_0$ and a set of graph rewrite rules $\mathcal{P}$

(ii) apply the rules $p \in \mathcal{P}$ to the matching molecule graphs $g \in \mathcal{G}_0$, creating new molecule graphs g'

(iii) add all newly discovered molecule graphs g' to the set $\mathcal{G}_0$, creating a new set $\mathcal{G}_1 \supseteq \mathcal{G}_0$

(iv) repeat the expansion steps, starting from $\mathcal{G}_1$

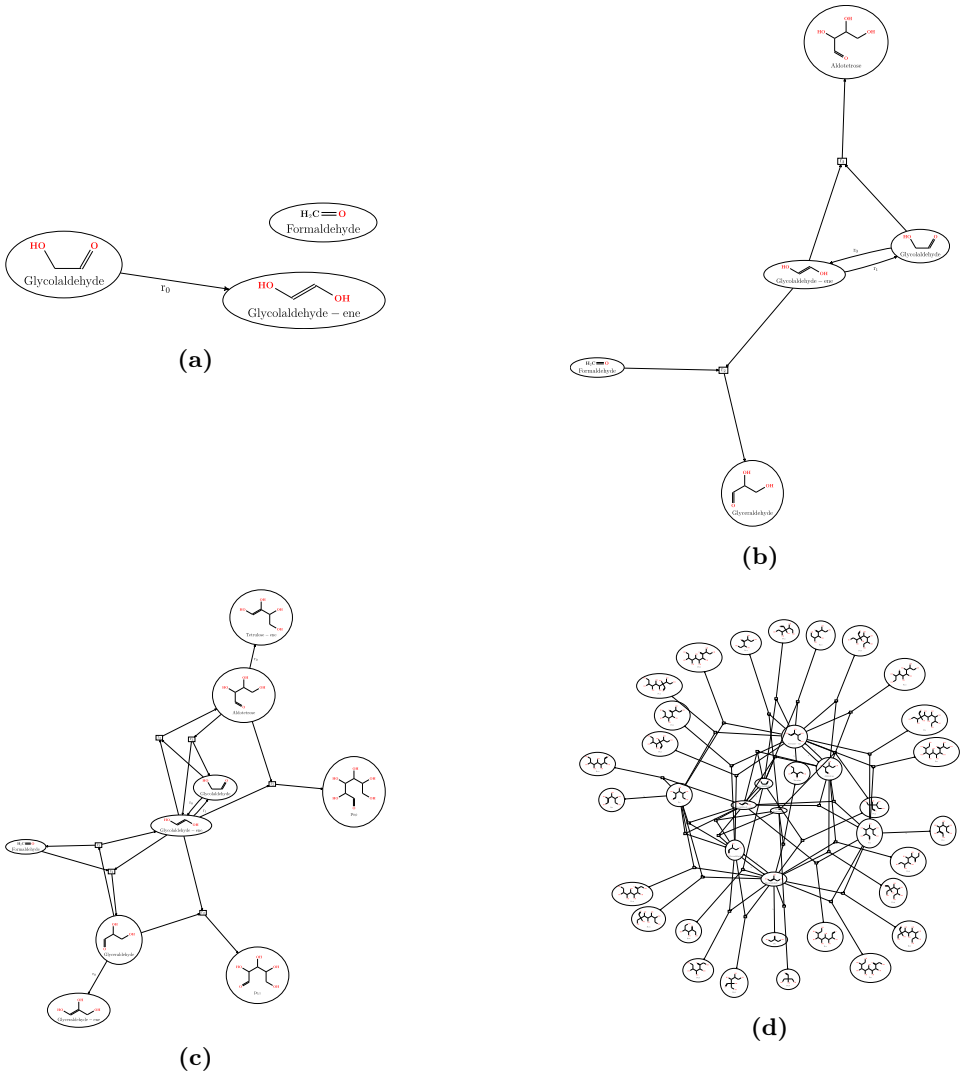


Figure 6.5: Iterative expansion of the derivation graph associated with the formose grammar (from Fig. 6.4) for: **(a)** one repetition, **(b)** two repetitions, **(c)** three repetitions, **(d)** four repetitions.

Fig. 6.5 shows the example of such a network expansion for the formose grammar, as it was defined above (see Fig. 6.4b). The expanded chemical reaction network is formalized as a directed, bipartite, multi graph[7], where, as outlined in the previous chapter, one type of vertices represent the molecules, while the other type represents the reactions, with the arrows representing the direction as well as the stoichiometry of each reaction. In general, within the MØD framework, this type of iteratively expanded reaction network is referred to as a 'derivation graph' (DG).

The expansion algorithm should, in principle, continue until no new compounds can be found through the application of the rewrite rules, i.e. $\mathcal{G}_{i+1} = \mathcal{G}_i$, and the expansion therefore converges. However, depending on the reaction rules defined in the graph grammar, the expansion could potentially never converge (e.g. a grammar containing condensation reactions that continuously creates new compounds). Furthermore, as the example in Fig. 6.5 also shows, the DG resulting from an expansion may become rather dense, due to something called 'combinatorial explosion'. In those cases it is necessary to constrain the DG expansion.

This can already be achieved on the level of the graph grammar, to some extent, through the selection and definition of the corresponding reaction rules. Defining a stricter context for some rewrite rule can, for example, limit their applicability, preventing a potential combinatorial explosion. Thus the formulation of a realistic graph grammar does not only include thermodynamic or kinetic constraints, as outlined above, but also serves the purpose of limiting the combinatorics of a DG expansion. To further constrain the the DG expansion, one can define a so called 'strategy', which can be interpreted as something like a set of general instructions for a DG expansion [Andersen et al. 2014]. Although there are various types of expansion strategies, however not all of the, are of importance for this project. The simplest one, usually called a 'repetition strategy', has no constraints and only applies the above outlined algorithm for a fixed amount of iterations (the DGs in Fig. 6.5 were obtained with such an expansion strategy). Furthermore, one can also impose constraints on the application of rewrite rules with so called 'predicate strategies'. This is a way to "prune" the DG, i.e. getting rid of combinatorics and helping to achieve a convergence of the expansion.

[7]More specifically a chemical reaction network is formalized within MØD as a directed hyper-graph[Andersen 2015].

One type of constraint is to limit the types of molecules the rules are applied on (called a left predicate for limiting the LHS of a reaction). This can for example be used as another way of preventing reaction rules to be applied to "unfavourable" molecules (similar to the context of a rule). Another possibility is to constrain the products of the rule applications (called a right predicate for limiting the RHS of a reaction). Quite common constraints on a reaction product involve: the definition of a maximum molecule size (i.e. the total nr of atoms), a maximum length of the carbon chain (i.e. the number of C-atoms), or a maximum molecule mass.

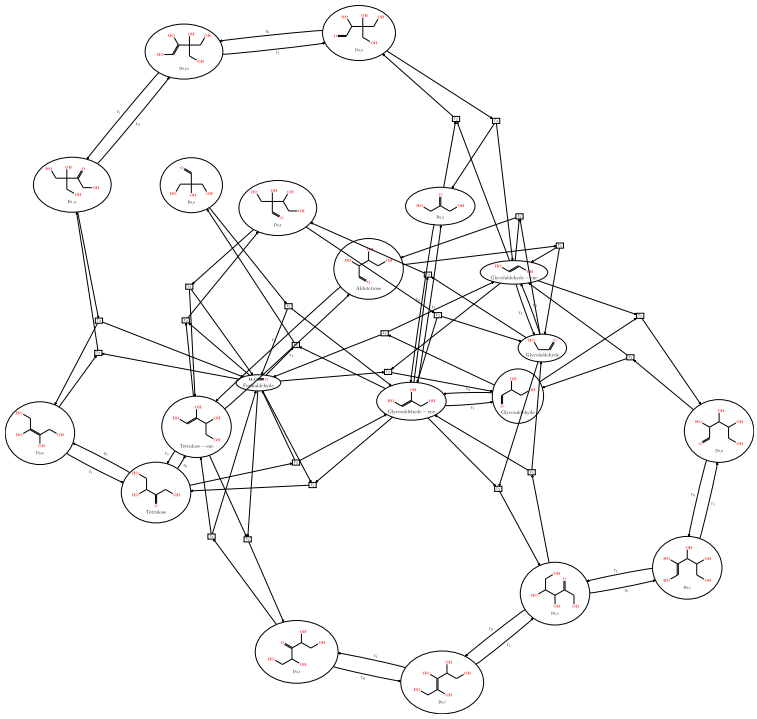


Figure 6.6: Constrained expansion of the derivation graph associated with the formose grammar using an expansion-strategy that limits the molecules to only 6 C-atoms.

If necessary, or convenient, different strategies can also be combined [Andersen et al. 2014]. The most common combination that is of interest for this project is the combination of a repetition strategy with a right predicate strategy, i.e. iteratively expand a DG but with certain constraints on the products of the rule application. Fig. 6.6 shows an example of such a strategy for the formose grammar.

To summarize, the overall reaction space expansion with MØD happens in several steps. First the molecules and reactions associated with the chemistry under study are formalized and collected as a graph grammar. Then an expansion strategy is implemented, representing a set of constraints imposed on the reaction chemistry. Finally the chemical reaction space is expanded in the form of a so called derivation graph (DG). However, as mentioned above, the MØD framework does not include thermodynamic and/or kinetic constraints by default. Therefore, a DG only represents a network of chemical reactions that could in principle happen within a certain chemistry, given the a priori choice of molecules, reactions, constraints, etc. (i.e. a possibility space), which will be sufficient for the purpose of this project. Perhaps a good analogy for a DG is that of a hiking map which shows all the possible paths that one can take from a certain start point. Some of these routes will be easier, some will be harder (one needs to invest more stamina/energy) and some, though in principle available, turn out to be impossible to take, due to constraints imposed by the topography of the landscape. In addition, there are certain types of routes whose topology is similar: linear routes that lead from a star point to an end point, or circular trails that will lead back to the starting point. A central idea of this project is to search for special types of routes in a DG, which will be elaborated further in the next section.

6.2 Sub-graph search

With a chemical reaction space expanded, one can proceed to further investigate its properties. One possible property is the existence of certain types of reaction pathways within the overall reaction space. As outlined in the previous chapter, the reaction paths of major interest for this project are autocatalytic cycles. In practice, there have been previous attempts to formulate methods for finding autocatalytic cycles within a larger system of chemical reactions, like the search algorithm proposed by Hordijk *et al.* [Hordijk and Steel 2004; Hordijk et al. 2011]. However, their approach is mostly centred around the set theoretic definition of autocatalysis (compare also Ch. 3) and is a rather coarse grained approach that does not take into account the full chemical reaction network underlying the system, effectively loosing important information about the stoichiometry of the reactions, etc. By contrast, the graph theoretical approach of MØD allows to formalize large

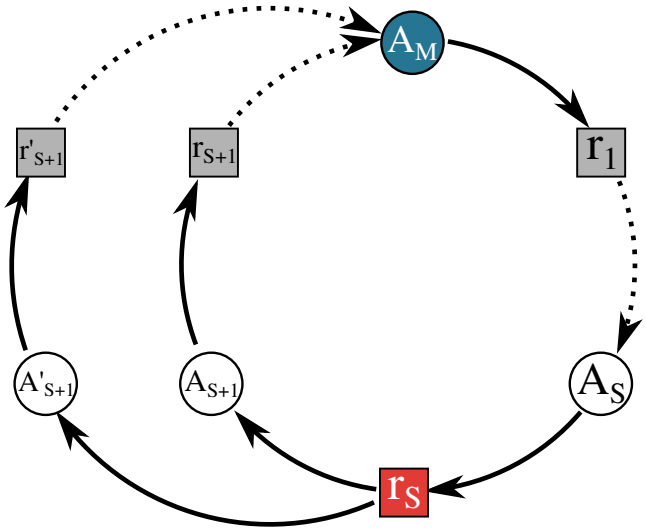


Figure 6.7: General scheme for the topology of an AC-cycle. In principle there is always a fission reaction r_S, were the cycle splits into two branches, as well as a merge point, a focal species A_M, where these two branches merge again.

systems of chemical reactions as bipartite graphs (or bi-partite networks) on which it is then possible to conduct searches for specific types of sub-graphs that form an autocatalytic cycle of reactions.

As outlined in the previous chapter, autocatalytic cycles can be classified into one of five characteristic motives of chemical reaction networks (see again Fig. 5.8). While these are distinct network motives, they share a common feature: namely, the existence of at least one fork reaction and a subsequent merge point. More precisely, it is possible to formulate a general network topology that is underlying an autocatalytic cycle, which is shown in Fig. 6.7. As such, an autocatalytic network can be formalised as an overall reaction that is cyclical, where a compound A_M, after a sequence of reactions gets replenished. However, at some point of the reaction cycle there is a reaction vertex r_S where an educt A_S is split into two products A_{S+1} and A'_{S+1}. While A_{S+1} continues to react on the main branch of the reaction cycle (eventually being transformed back into A_M), A'_{S+1} undergoes other reactions thus creating a side branch. Eventually an intermediate from the side branch will undergo a reaction whose product is again A_M, thus merging the side- and the main branch (a similar generalisation method has been applied in [Arya et al. 2022]). In general r_S and A_M are referred to as 'split

vertex' and 'merge vertex' respectively, with A_M being the autocatalytic species (or 'focal species'). The comparison between Fig. 5.8 and Fig. 6.7 shows that this generalisation in effect contains all five autocatalytic motifs and is given by the relationship between A_{S+1}, A'_{S+1} and A_M, i.e.:

- type I: $A_{S+1} = A'_{S+1} = A_M$
- type II: $A_{S+1} = A_M \neq A'_{S+1}$ or $A'_{S+1} = A_M \neq A_{S+1}$
- type III: $A_M \neq A_{S+1} \neq A'_{S+1}$

The types IV and V are, as discussed above, technically variations of type II with one or two additional split vertices, respectively. Therefore they are automatically included when the underlying type II cycle is found. All in all, searching for autocatalytic cycles within a DG, effectively amounts to searching for sub-graphs that have a given topology (displayed in Fig. 6.7) i.e. reaction cycles that have the corresponding split- and merge vertices.

In practice this is going to be done using a software package (kindly provided by Phillip Honegger [Honegger 2023]) that implements a sub-graph search algorithm for this type of network motif, which will be applied to a DG obtained with MØD. The autocatalytic reaction pathways that are found this way are therefore part of a larger chemical reaction space within which they are embedded. So it is important to realise that the compounds belonging to an autocatalytic cycle can potentially be involved in additional side reactions which can lead to cross-links within the cycle. More precisely, these cross-links are reactions where a cycle compound gets transformed into a different cycle compound, effectively creating a "shortcut" within the autocatalytic cycle that would therefore destroy it. Given that these cross-links are in general not found by searching for autocatalytic motifs, as defined in Fig. 6.7, further verification of the results of this sub-graph search is needed. In practice this is done by applying a simple flow verification on each potential autocatalytic cycle that is found [Honegger 2023].

Therefore, the search for autocatalytic cycles will be carried out in two stages: first all the sub-graphs that match the pattern of Fig. 6.7 are identified, resulting in candidates for sub-graphs that are potentially autocatalytic. Then a flow-verification is done, in order to filter out any false positives. Fig. 6.8 shows some examples of autocatalytic cycles found within the DG expansion of the formose grammar.

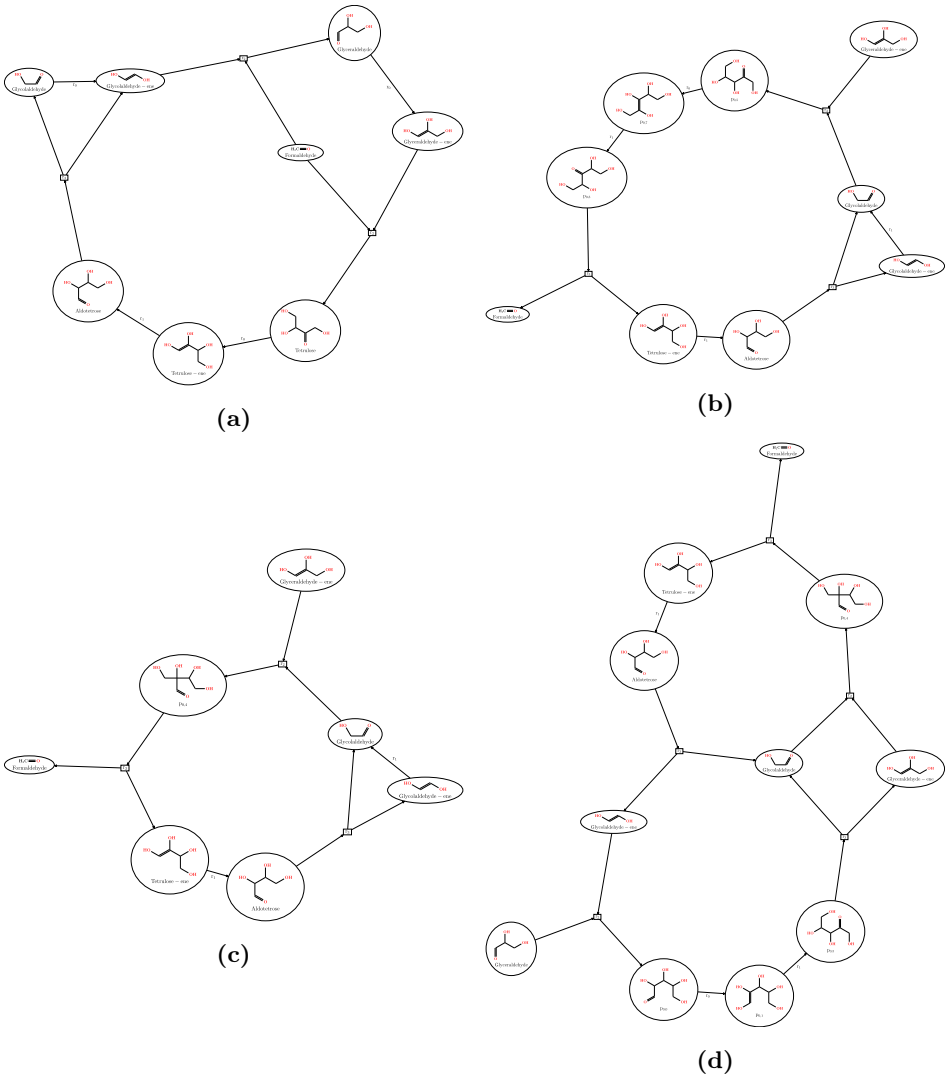


Figure 6.8: Examples for AC-networks found in the formose DG. **(a)** Standard formose reaction (i.e. the Butlerov reaction). **(b)**, **(c)** and **(d)** are "alternative" formose reactions, where glycolaldehyde (or its equivalent enol-form) gets autocatalytically produced but through different combinations of reactions.

The inclusion of additional constraints is furthermore required, to be more effective in this kind of search, since large and dense DGs can lead to an explosion of potentially autocatalytic cycles (including rather long cycles that are in general not very plausible). In practice, it is necessary to define a maximum number, L_{max}, of reactions between the split point r_S and the merge point A_M (and vice versa). This way the length of the main reaction cycle as well as the side branch is limited. Moreover, certain compounds that are involved in a lot of reactions (e.g. water, H_2, CO_2) can create strongly connected hub vertices in any DG. Some of these compounds (typically water), can also be involved as both educts and products (for instance, in a chemistry with many hydration-, dehydration reactions). In such a case, the above defined algorithm to localize autocatalytic cycles would find many solutions with these molecules as merge points (which is not very useful: a reaction that autocatalytically produces water is not very interesting). Thus, it is important to define a set of compounds $\mathcal{G}_e$ which are excluded as potential merge points (i.e., autocatalysts) of the cycles. As a result, L_{max} and $\mathcal{G}_e$ are the main parameters to constrain the autocatalysis search tool.

6.3 Thermochemistry

Taking into consideration that MØD does not include by default any information about the kinetics or thermodynamics of the modelled/searched reaction systems, it is important to complement the software with methods that make sure that the solutions it provides are chemically feasible or not. One way to test the feasibility of a certain reaction pathway is to have a look at the overall reaction energy associated with it, i.e. the energy necessary to run the whole reaction process. In general, given a chemical reaction of the form:

$$\mathrm{A} + \mathrm{B} \underset{\mathrm{k}_-}{\overset{\mathrm{k}_+}{\rightleftharpoons}} \mathrm{C} + \mathrm{D} \tag{6.1}$$

i.e. a reversible transformation of a set of educts to a set of products, with k_+ and k_- being the reaction rates of the forward and backward reactions[8], the reaction energy (i.e. the change in the Gibbs free energy) can be computed

[8]From a thermodynamic point of view, reactions are always considered to be reversible. However, one needs to define a "standard direction" in which the reaction is assumed to occur, i.e. the forward direction.

as:

$$\Delta_r G^\circ = -RT \ln K_{eq} \tag{6.2}$$

where T is the temperature of the system, R is the universal gas constant and K_{eq} is the equilibrium constant, defined as:

$$K_{eq} = \frac{k_+}{k_-} = \frac{[\mathrm{C}]\,[\mathrm{D}]}{[\mathrm{A}]\,[\mathrm{B}]} \tag{6.3}$$

with $[\mathrm{A}]\,,[\mathrm{B}]\,,[\mathrm{C}]$ and $[\mathrm{D}]$ being the concentration of the respective species. As shown in Fig. 6.9, the value of $\Delta_r G^\circ$ determines the "natural direction" for the reaction of Eq. 6.1. A value of $\Delta_r G^\circ < 0$ indicates that the forward reaction is exergonic, i.e it releases free energy. The reaction therefore proceeds naturally (or spontaneously) in the forward direction. A value of $\Delta_r G^\circ < 0$, on the other hand, indicates that the forward reaction is endergonic, i.e it requires energy to run. The reaction, thus, will not spontaneously proceed in the forward direction (indicated by the longer forward arrow in Fig. 6.9a). But at the same time the backward reaction is exergonic. Therefore, the reaction naturally proceeds in the backward direction (indicated by the longer backward arrow in Fig. 6.9b). All in all, knowledge about the value of $\Delta_r G^\circ$ gives important insights into whether a reaction can naturally proceed in the forward (or standard) direction, i.e. whether it is thermodynamically feasible[9].

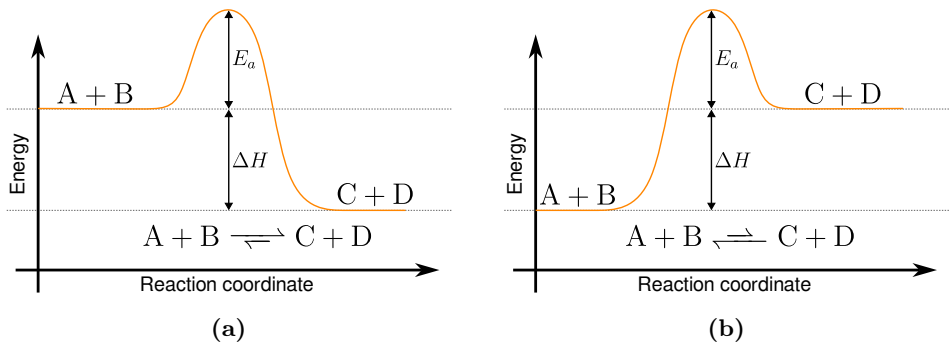


Figure 6.9: Energy landscape for the reaction of Eq. 6.1 if the forward direction is: **(a)** an exergonic reaction, or **(b)** an endergonic reaction.

[9]Note that thermodynamics only fixes the energy landscape of the reaction but not the activation energy (i.e. energy barrier) of a reaction and the resulting kinetics. Therefore, even if a reaction is thermodynamically feasible, it might still not occur (at least not within a reasonable amount of time) due to a high kinetic barrier.

For a reaction pathway that consists of multiple reactions, $\Delta_r G^\circ$ represents the overall energy of the pathway and is thus calculated as the sum of the individual reaction energies $\Delta_r G_i^\circ$:

$$\Delta_r G^\circ = \sum_i \Delta_r G_i^\circ \tag{6.4}$$

Correspondingly, $\Delta_r G^\circ < 0$ indicates that the pathway as a whole is releasing energy and can thus occur spontaneously, i.e. the pathway is thermodynamically feasible. $\Delta_r G^\circ > 0$ on the other hand, indicates that the pathway as a whole is requiring energy in order to run and therefore will not occur spontaneously, i.e. the pathway is not thermodynamically feasible.

However, calculating $\Delta_r G_i^\circ$ for a reaction is not necessarily an easy task. In practice one often relies on experimental calculations of the equilibrium constant K_{eq} from educt and product concentrations in combination with Eq.6.3. For known biochemical reactions, these calculations are then usually collected in databases that can be queried (like the KEGG database [Kanehisa and Goto 2000; KEGG]). Other, more theoretical/computational approaches rely on the use of quantum chemistry methods (examples can be found here: [Jinich et al. 2018, 2020]). Yet the use of databases is limited to known and well studied biochemical reactions and is thus not very suitable for a network expansion approach, which potentially creates reactions and/or compounds that are not included in these databases. Quantum chemistry methods, on the other hand, might be able to calculate energies of arbitrary reactions and compounds but often rely on heavy, computationally expensive calculations that become rather unstable for relatively large molecules. Furthermore, some software packages that implement quantum chemistry calculations are not particularly transparent concerning the exact methods implemented, so the software package becomes a "black box" that provides the results.

Nevertheless, as seen in Fig. 6.9, in principle one can also get the energy of a chemical reaction through the following calculation:

$$\Delta_r G^\circ = \sum_{p \in P} s_p \Delta_f G_p^\circ - \sum_{e \in E} s_e \Delta_f G_e^\circ = \Delta_f G_{prod}^\circ - \Delta_f G_{ed}^\circ \tag{6.5}$$

which is the difference between $\Delta_f G_p^\circ$, the energies of formation of the products involved in the reaction, and $\Delta_f G_e^\circ$, the energies of formation of the

educts involved in the reaction. In other words, $\Delta_r G_i^\circ$ can be obtained as the energy difference between the molecules before and after the reaction.

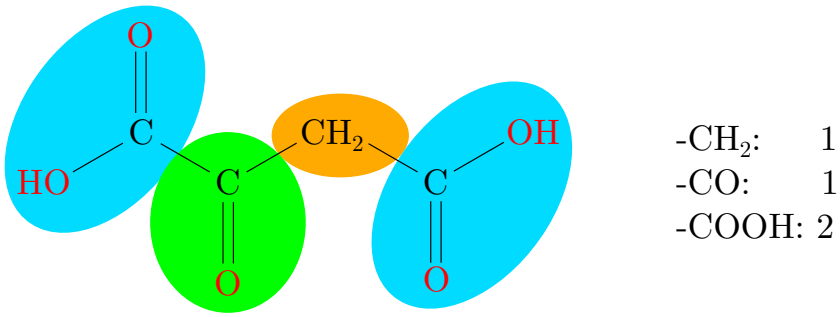


Figure 6.10: Example for the decomposition of an oxaloacetate molecule into it's main functional groups.

In general, values for the formation energies $\Delta_f G^\circ$ can be obtained again from experimental databases, like the Alberty database [Alberty 2003]. However, this again limits one to compounds that are already known to be part of biochemical reaction pathways. A different approach called 'group contribution' method was initially proposed by Mavrovouniotis [Mavrovouniotis et al. 1988; Mavrovouniotis 1990] and later further developed by Hatzimanikatis *et al.* [Jankowski et al. 2008]. The basic idea behind classical group contribution is that any molecule can be decomposed into a collection of functional groups (e.g. acid groups, $-$OH groups etc., see Fig. 6.10,). It is then assumed that the overall energy of the molecule is the sum of the individual energy contributions of these functional groups:

$$\Delta_f G_i^\circ = \sum_{g \in \mathcal{G}} n_g \; \Delta_f G_g^\circ \tag{6.6}$$

with $\Delta_f G_g^\circ$ being the energy of a certain functional group and n_g the number of its occurrence in the molecule. In other words, the overall energy of the molecule is approximated through the energy contribution of its functional groups. The advantage of this method is that, while the energy of a certain complex molecule might not be known, the energy contributions of its functional groups are usually known.

In recent years an extension to the classical group contribution method has been devised by Noor *et al.*, which they refer to as 'component contribution' method [Noor et al. 2013]. Within their approach, the reaction

energies are calculated by combining results from databases with estimations obtained by decomposing a molecule with group contribution, to get formation energies. The main idea is that while some molecules are easy to decompose (like in the example of Fig. 6.10), the group structures of more complex biomolecules may be much less obvious (since there are often overlapping functional groups). For these molecules there exist, however, good experimental estimations (derived from standard databases) that can be used instead.

This component contribution method has been condensed in a software tool called *eQuilibrator* [eQuilibrator] [Flamholz et al. 2012; Beber et al. 2022], which is available as an open source software tool (in the form of a python package). An advantage of the eQulibrator package is that it allows to compute $\Delta_r G^\circ$ for different pH values. With a slight modification, it is also possible to calculate $\Delta_r G^\circ$ for a selected redox potential. As outlined in [Goldford et al. 2019], given a value of $\Delta_r G^\circ$ obtained for a certain redox potential E_0[10] one can calculate the reaction energy $\Delta_{r'} G^\circ$ for a different selection of a redox potential E_0^+ using the following correction formula:

$$\Delta_{r'} G^\circ = \Delta_r G^\circ + nF\,(E_0^+ - E_0) \tag{6.7}$$

with n being the number of transferred electrons (usually $n = 2$) and F being the Faraday constant.

Given these advantages, the eQuilibrator software package is going ot be used here in order to compute the reaction energies of the solutions obtained from the search for autocatalytic cycles under different conditions (i.e. different selections of pH and redox potential E_0^+)[11]. The main idea is to determine which of those solutions are, under the assumed conditions, thermodynamically viable, and which are not but could potentially run if there was an external energy input.

[10] Redox reactions are usually of the form $X_{ox} + YH_2 \rightleftharpoons X_{red} + Y^{2+}$, where YH_2/Y^{2+} is the redox couple with the corresponding redox potential E_0 (in extant biology this could be: $NAD(P)H/H^+ - NAD(P)^+$ or $FADH_2 - FAD^{2+}$). In a prebiotic setting the role of these redox couples could be fulfilled by certain transition metals (e.g. iron).

[11] As the equilibrator also uses its own database of molecules one needs to use the additional eQuilibrator assets package [eQuilibrator Assets] to generate compounds.

Chapter 7

Exploring the rTCA Chemistry

As advanced in previous chapters, the main focus of this part of the thesis project is to investigate the plausibility of pre-biotic carbon fixation pathways. This is going to be achieved by exploring the chemical reaction space (CRS) associated with the chemistry of the rTCA cycle, investigating whether other autocatalytic (AC) carbon fixation pathways of similar nature are viable and could thus be candidates for a prebiotic precursor pathway. The CRS will be expanded through a rule-based approach, implemented through the software package MØD, expanding derivation graphs (DGs) on which various simple boundary conditions will be then imposed. Therefore, in the following sections different types of rTCA grammars (the 'base activator grammar', the 'relaxed activator grammar' and the 'activator independent grammar') are going to be outlined, reflecting different boundary conditions and constraining degrees. The resulting DGs as well as the properties of the obtained AC cycles will finally be compared.

7.1 Formulation of an rTCA grammar

The first step in order to expand the chemical reaction space associated with a certain type of chemistry (see previous chapter) is to define the corresponding graph grammar. This grammar includes the initial compounds assumed to be present in the system (formalized as molecule graphs), together with the main types of reactions that can realistically occur within the system (formalized as graph rewrite rules). A selection of small compounds that only

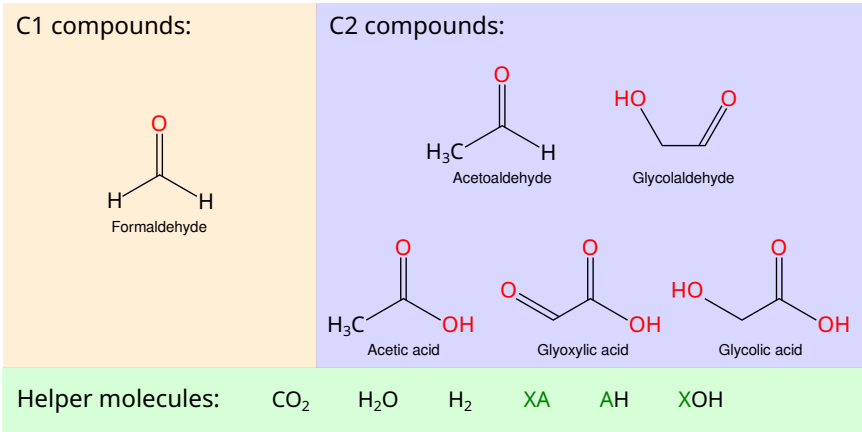


Figure 7.1: The set of initial molecules used for the different rTCA graph grammars.

contain one or two C-atoms and are pre-biotically plausible were chosen[1] as initial molecules (see Fig. 7.1). Furthermore, a group of "helper molecules" like CO_2, H_2O, H_2 was defined, which are the molecules that are needed as additional educts for many reactions (e.g. H_2O for hydration reactions).

Comparing the reaction steps of the rTCA cycle and the other autocatalytic carbon fixation pathways known in biology (like the 3HP- or the 4HB-cycle, etc.), one can see that there are certain 'classes' of reactions that keep repeating throughout these cycles. More precisely, there are: carboxylation reactions that bind CO_2, or condensation reactions that grow the carbon chain; redox reactions and (de)-hydration reactions that change the redox state of a carbon chain or bind/unbind an $-OH$ group; cleavage reactions that split the carbon chain, serving as the forking reactions that can lead to an autocatalytic cycle (autocatalytic closure). These reaction types, as well as their main function within the reaction pathway can be collected and classified, as shown in Tab. 7.1. Thus any grammar for expanding the reaction space of the rTCA cycle (or carbon fixation cycles in general) needs to at least include reaction rules according to these three classes. However, differences in the rewrite rules (e.g. selection of rules, strictness of the rules contexts, etc.) will change the density and composition of the DG (i.e. the expanded chemical reaction space) and the potential substructures (i.e. reaction paths) that can be found. Taking this into account, different types of

[1] A furhter motivation for the choice of these molecules is that they are the smallest possible molecules to which the reaction rules, which will be defined below, can be applied.

grammars will be defined and their outcomes compared.

Class	Function	Examples
1	grow the carbon chain	carboxylation reactions
2	modify the carbon chain	redox reactions, (de)-hydration reactions
3	split the carbon chain, autocatalytic closure	retro aldol reaction

Table 7.1: Main reaction classes and their functions for the rTCA cycle as well as autocatalytic carbon fixation pathways in general.

7.1.1 Base activator grammar

Closer inspection of the rTCA cycle and other carbon fixation pathways reveals that there are some reactions that involve the binding and subsequent unbinding of certain compounds called 'coenzymes'. More specifically, the coenzyme involved in the reactions of the carbon fixation pathways (and many other metabolic pathways) is coenzyme-A (CoA). In general, the role of these coenzymes is similar to regular enzymes in the sense that, by binding to a compound, they enable certain reactions that otherwise would be kinetically unfavourable. Fig. 7.2 shows the example of a reductive carboxylation of an acid. This reaction would not occur naturally on an acid group and therefore it requires the binding of CoA before it can proceed, unbinding CoA in the process. Therefore, in a certain way, this coenzyme binding "activates" the compound, enabling further reactions.

The MØD framework allows to implement both molecules and reaction rules that contain arbitrary atom labels, so it is possible to generalize this binding of CoA (or any other coenzyme in general). As shown in Fig. 7.2, this is done by formalizing the reaction as the binding of an arbitrary "activator" compound (labelled "A") that is supplied bt an arbitrary "carrier" compound (labelled "X"). Using this arbitrary activator compound relieves one from the necessity to define and implement a prebiotically plausible candidate for such a coenzyme (as they are rather complex molecules). It is simply assumed that such an activating compound (however it might look like) is present within the reaction chemistry. If a plausible prebiotic candidate is later found or specified, this can be easily implemented by substituting the activator with the corresponding compound. In practice, this means

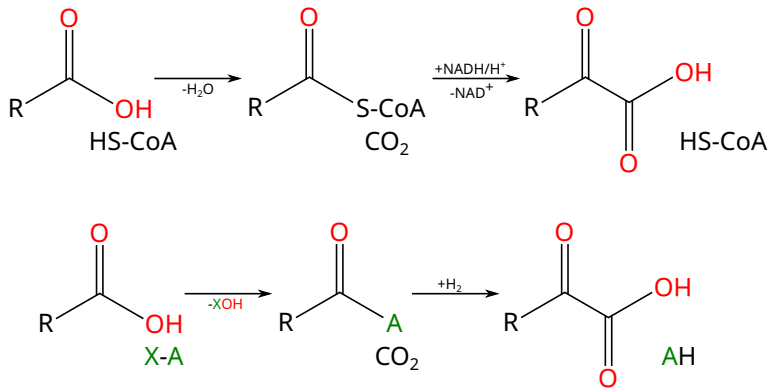


Figure 7.2: Top reaction: coenzyme-A dependence of a reductive carboxylation reaction. Bottom reaction: substitution of the coenzyme with a general activator compound A and a carrier compound X.

that there is an additional class of reactions (referred to as Class 0) that one must take into account in order to formalize the reaction grammar of the rTCA, namely the binding of an activator compound (additionally, the activator and carrier compounds have to be included as helper molecules too, see again Fig. 7.1).

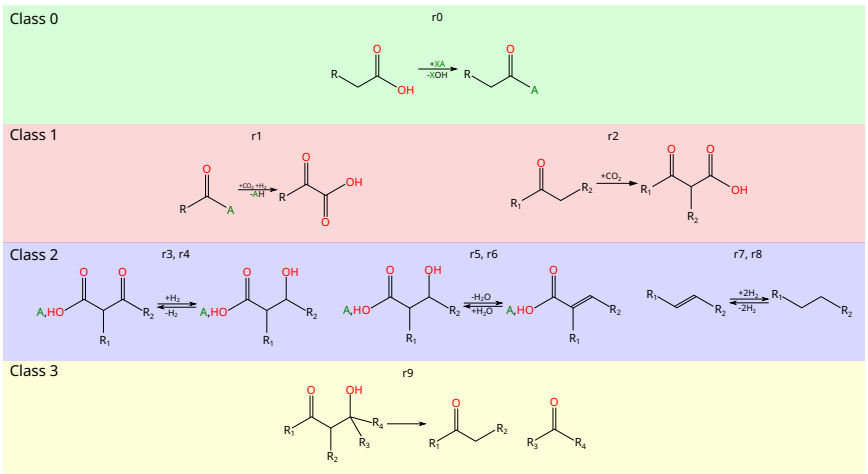


Figure 7.3: Main reaction classes and their respective selection of reaction rules for the base, activator depended grammar of the rTCA chemistry.

Taking all of the main reaction classes and their associated reactions into consideration, the set of rewrite rules, necessary to implement a reaction grammar for the chemistry rTCA cycle, can be defined. A collection of these reaction rules, as well as the constraints put on them is shown in Fig. 7.3

(a more detailed description of the rewrite rules, together with the choices of constraints can be found in Appendix A). While there are some reaction rules, like the rule $r1$ (see Fig. 7.3), that directly depend on the existence of an activator, it is important to realize that there is also an indirect activator dependence of some rules (see Fig. 7.4). In other words, the presence of an activator compound implies a simple boundary condition on the overall reaction grammar, because it directly or indirectly enables certain chemical reactions. In general, due to this overall activator dependence, the graph grammar consisting of the rewrite rules of Fig. 7.3, together with the initial molecules from Fig. 7.1, is going to be referred to as the 'base activator grammar'.

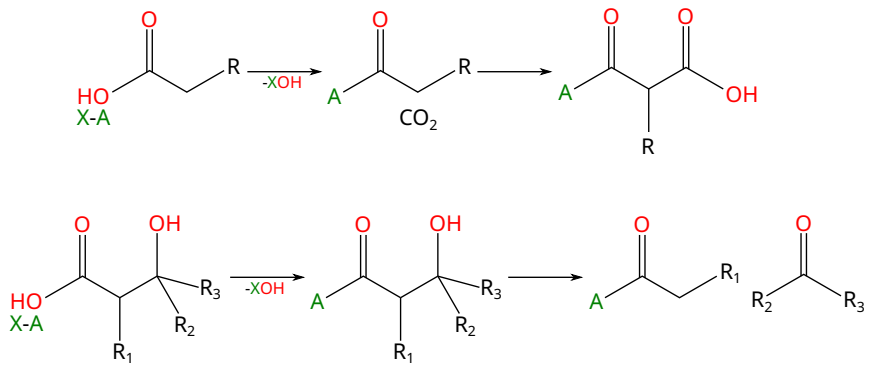


Figure 7.4: Indirect activator dependence. Top reaction: a carboxylation reaction would in general not work on a carbon atom α-to an acid group. However, if the acid gets activated, the carboxylation reaction on the α-carbon is enabled. Bottom reaction: similar a retro-aldol reaction would not work if the OH-group is β-to an acid. But by activating the acid, the retro-aldol reaction is enabled.

7.1.2 Relaxed activator grammar

The rather strict graph grammar defined above is on the one hand necessary, to prevent a combinatorial explosion when expanding the reaction space of the rTCA. On the other hand, if a grammar is too constrained, the resulting DG could become too sparse (i.e. it would contain too few molecules and/or reactions) which does not lead, normally, to very interesting results. Alternatively, one could take the set of rules from a stricter grammar, like the one defined in Fig. 7.3, and "soften up" some of the reaction rules. However, small modifications in the rule set can significantly affect the expansion algorithm, so this relaxation should be done very selectively, in order to avoid

ending up with a combinatorial explosion again, and ensuring chemical coherence (i.e. making sure the rewrite rules represent chemical reaction that can realistically occur).

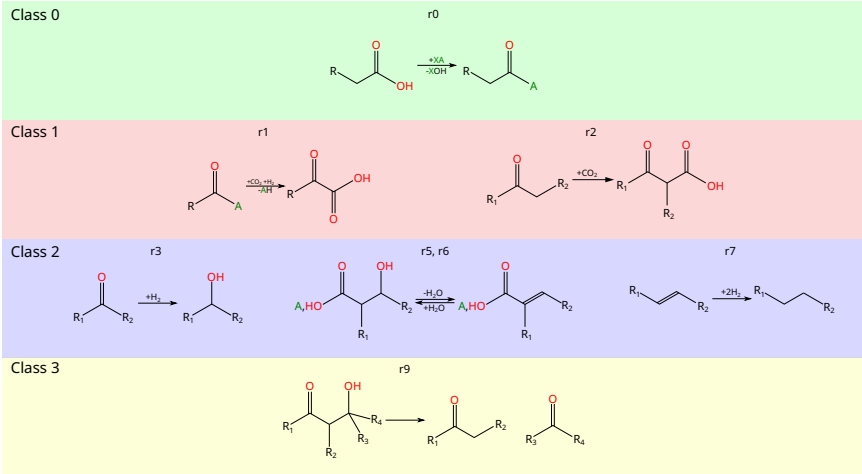


Figure 7.5: Main reaction classes and their respective selection of reaction rules for the relaxed activator grammar.

One reaction rule from Fig. 7.3 that can be relaxed rather easily is rule $r3$, i.e. the reduction of a carbonyl group. In the chemistry under analysis, this is limited to carbonyl groups that are in a β-position to a carboxy group (i.e. an acid), keeping some symmetry with the (de)-hydration reactions $r5$ and $r6$. However, it is not necessary to define the reaction rule that strict. As depicted in Fig. 7.5, the rewrite rule can be relaxed to be applied to any aldehyde or ketone. To avoid a potential combinatorial explosion during the DG expansion, the rules $r4$ and $r8$ (i.e. the oxidation reactions) are removed, which effectively means that only a reductive chemistry is studied. The remainder of the rules from Fig. 7.3 will be left unchanged. The graph grammar consisting of this modified, activator depended reaction rule set, together again with the initial molecules (as defined in Fig. 7.1) will be referred to as the 'relaxed activator grammar' (the this way defined grammar is similar to the set of reaction rules used in [Zubarev et al. 2015], but with the activator dependency added).

7.1.3 Activator independent grammar

The last grammar version that is going to be investigated is in this chapter is one that does not depend on the activating species. This is implemented

by modifying the set of reaction rules defined in Fig. 7.3 rather drastically, by removing all rules that explicitly depend on the activator group: i.e., $r0$, the binding of the activator groups (effectively removing Class 0), and $r1$, the activator dependent, reductive carboxylation. Additionally, for the rules $r3, r4$ and $r5, r6$ the option of occurring at a molecule with an activator groups is removed from the respective contexts. Finally, a reaction rule for an aldol addition, $r10$, is added to the class 1 of the rule set (constrained to only bind compounds with one or two C-atoms to an existing carbon chain). This is done as a way to compensate for the loss of $r1$, ensuring that there are reactions that can grow the carbon chain. The resulting set of rewrite rules is shown in Fig. 7.6, and the corresponding graph grammar will be referred to as the 'activator independent grammar'. The set of initial molecules for this grammar is again the one defined in Fig. 7.1, the only exception being that in this case the activator species is removed.

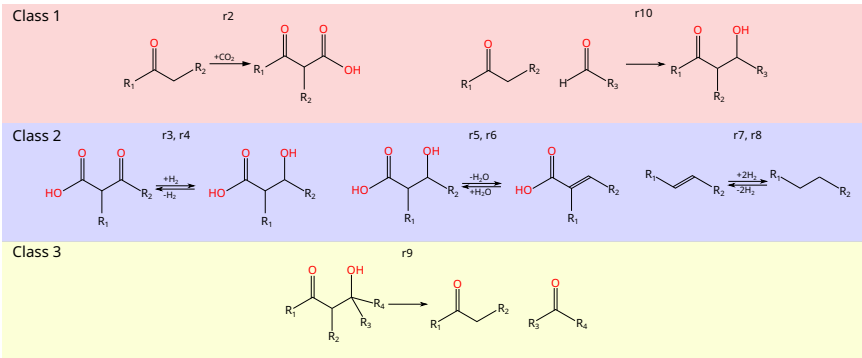


Figure 7.6: Main reaction classes and their respective selection of reaction rules for the activator independent grammar of the rTCA chemistry.

7.2 Strategies and DG expansions

With a selection of graph grammars defined, one can proceed to expand the corresponding DGs. As outlined in the previous chapter, in order to do a reaction space expansion with MØD, an expansion strategy needs to be defined in such a way that a combinatorial explosion is prevented and convergence is guaranteed. For simplicity, all the expansions for the different grammars defined above were done with the same type of strategy: a combination of

a repetition strategy, with a 'label strategy' and a 'mass strategy'[2]. More precisely, all molecules that are allowed to be obtained by iteratively applying the rewrite rules defined in the graph grammar are constrained to have (at max) one activator group (under the assumption that two or more activator groups would render the complex chemically unstable). Furthermore, a maximum mass limit is imposed for all obtained molecules. This is again done due to stability considerations: i.e., larger and heavier molecules are usually more unstable (and often not energetically favourable, either). The upper bound for the molecule mass was set to 200amu, which is reasonable, given that most biomolecules are around this mass[3].

Grammar type	**Nr. Molecules**	**Nr. Reactions**
Base activator	171	353
Relaxed activator	818	1592
Activator independent	409	1061

Table 7.2: Results of the DG expansion for different graph grammars using a combination of a label- and a mass-strategy.

All the expansions were then carried out for as many repetitions as necessary until convergence was reached and no new molecules were discovered by the expansion algorithm. At the end of the ordinary expansion a final step was performed called a 'closure'. The expansion algorithm often fails to connect new compounds to already existing ones. But, obviously, the possibility space should contain all the connecting reactions. Therefore the closure step takes all the compounds in a DG and applies all the reaction rules to them, with the constraint of allowing only products that are already in the set. The DG expansion results for all of the above defined grammars are shown in Tab. 7.2, where the measured number of molecules (i.e. the number of molecule vertices in the graph), as well as the number of possible reactions (i.e. the number reaction vertices, or hyper-edges), are reported.

[2]The rule for the aldol addition can also be applied at the same molecule graph, which would lead to a cyclic compound. As some carbon rings (three- or four-membered rings) are energetically unfavourable and cyclic compounds are generally not of interest for this work, an additional right predicate is used for the activator independent grammar that filters out such molecules.

[3]The mass of a potential activator group was ignored here. While these activating compounds are usually relatively large and complex molecules, the focus was on the on-cycle species.

7.3 AC cycle search: main results

Once a chemical reaction space is expanded in the form of a DG, like it was done above, one can proceed to do a search for autocatalytic cycles using the topological definition for autocatalysis of Fig. 6.7. To prevent a combinatorial explosion of results, the search is further constrained by defining a maximum length L_{max} of the main reaction cycle and a set of molecules $\mathcal{G}_e$ that are not allowed as on-cycle species. For both the base grammar and the relaxed grammar the same search constraints were set to allow direct comparison and analyse the difference in the results in terms of the rigidity/flexibility of the reaction grammar. Therefore, the set of ignored molecules was defined as all the helper molecules from Fig. 7.1 (recall that reaction cycles that autocatalytically produce water are of little interest). The maximum length of the main reaction cycle was set to 11 reactions, which is the number of reactions in the main cycle of the rTCA cycle (compare also Fig. 5.4a, adding the reactions of the side branch the rTCA has 13 reactions in total). This is to ensure that the results of the AC cycle search can be directly contrasted with the actual rTCA cycle.

Grammar type	L_{max}	$\mathcal{G}_e$	**Nr. AC Cycles**
Base activator	11	Helpers	138
Relaxed activator	11	Helpers	8483
Activator independent	6	Helpers + C1-/C2-mols	1145
	5	Helpers	3144

Table 7.3: Constraints and results for the AC cycle search performed on the different grammar versions.

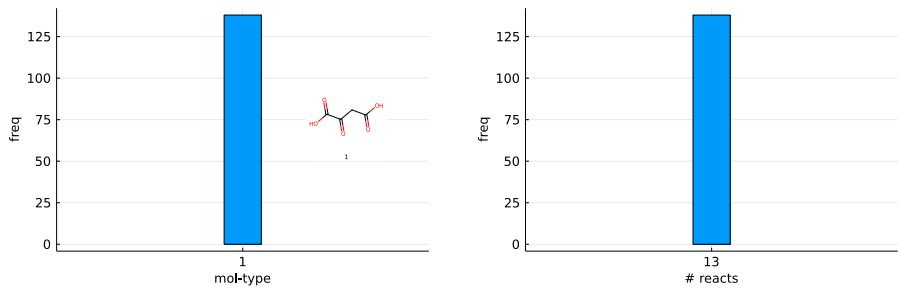
For the activator independent grammar the constraints for the AC cycle search were set a bit differently. Given that the set of rewrite rules (see Fig. 7.6) for this graph grammar includes an aldol-addition reaction that is restricted to only bin molecules with one or two C-atoms as a reaction that grows the carbon chain these C1 or C_2 can, in principle, be regarded as helper molecules too. Therefore, two searches with different constraints were performed on this grammar: one, more relaxed, in which only the "typical" helper molecules (i.e. CO_2, H_2, H_2O, etc.) were excluded; and the other, more strict, additionally excluding molecules containing one or two C-atoms. Additionally, for the strict AC-search the maximum length was

set to 6 reactions, while for the relaxed search it was set to 5 reactions[4].

Tab. 7.3 reports the search constraints for the different grammars, as well as the results of the AC-search (i.e. the amount of autocatalytic cycles that are found within the corresponding DG). In the following sections below, different properties of the obtained autocatalytic cycles are going to be examined and compared.

7.3.1 Network properties

The first types of properties that are going to be examined at this point regard the reaction networks that correspond to the autocatalytic cycles found in the DGs. Given the topological definition for autocatalytic networks in Fig. 6.7 one should begin by identifying what molecules are the focal species (i.e. the autocatalyst, of the found autocatalytic cycles) and in how many cycles do they occur as such. Furthermore, even if the number of reactions that constitute the main cycle of the autocatalytic network is limited (due to the search constraints) this does not mean that the total number of reactions (i.e. the main cycle together with the side branch) remains fixed. Therefore, another network property of immediate interest is the total amount of reactions per autocatalytic cycle, together with the number of cycles with a certain amount of reactions.



(a) **(b)**

Figure 7.7: Network related properties of the AC cycles found within the DG of the base activator grammar. **(a)** Different focal species (drawn and enumerated in the box) and their corresponding frequency of occurrence. **(b)** Frequency of occurrence for the total number of reactions of each AC cycle.

[4]As this activator independent grammar would not be able to reproduce the rTCA cycle it is not necessary to directly compare the results of the AC-search to it. Moreover, due to the nature of the resulting DGs, the AC-searches become rather combinatorial, so a more strict cycle length constraints are necessary.

Fig. 7.7 shows these results for the AC-search performed on the base activator grammar. As one can see, the molecule oxaloacetate is the focal species for all the obtained cycles, each (invariably) containing 13 reactions in total. In other words, the 128 autocatalytic cycles that are found within the chemical reaction space associated with this graph grammar are in principle functionally equivalent to the rTCA cycle, sharing a similar topology. A more informative (or intuitive way) of displaying these results is gathered in Fig. 7.8, which consists in a cumulative graph of all the autocatalytic reaction networks found, i.e. the overlap of all the networks with the multiplicity of an edge given by its weight. The number of molecules and reactions associated with this "condensed autocatalytic core" are furthermore reported in Tab. 7.4. One can therefore conclude that all the obtained autocatalytic cycles are basically variations of the rTCA cycle: they initially follow its reaction pathway but "take detours" at some points, where they go through different cycle intermediates, before "returning" to the original rTCA pathway. In any case, all the alternative pathways eventually converge at the focal species oxaloacetate. As such, the types of reactions occurring within these 128 AC cycles are in principle the same as the one occurring in the rTCA cycle, they simply proceed in a different order. In general, these results suggests a certain stability in the topology of the reaction network associated with the rTCA cycle. They also underline the centrality of compounds like succinate, acetate or oxaloacetate (through which all obtained cycles seem to pass) within extant biology.

In contrast, Fig. 7.9 shows the network properties for the AC-search performed on the relaxed activator grammar. One can immediately see that not only does this grammar result in more autocatalytic cycles (see Tab.7.3) but there is also more variation within the found cycles. This includes both the distribution of focal species (see Fig. 7.9a) and the distribution of total reaction numbers for each cycle (see Fig. 7.9b). Again the cumulative network shown in Fig. 7.10 compiles this data in a more informative way: while the variations of the rTCA cycle surveyed within the base grammar DG are retained (lower left part of Fig. 7.8), a larger portion of remarkably different autocatalytic cycles that are found too. The size of this autocatalytic core is also larger compared to the base grammar DG (see Tab. 7.4). All in all, one can see that a variation in the graph grammar not only increases the resulting DG (i.e. the chemical reaction space) but furthermore increases

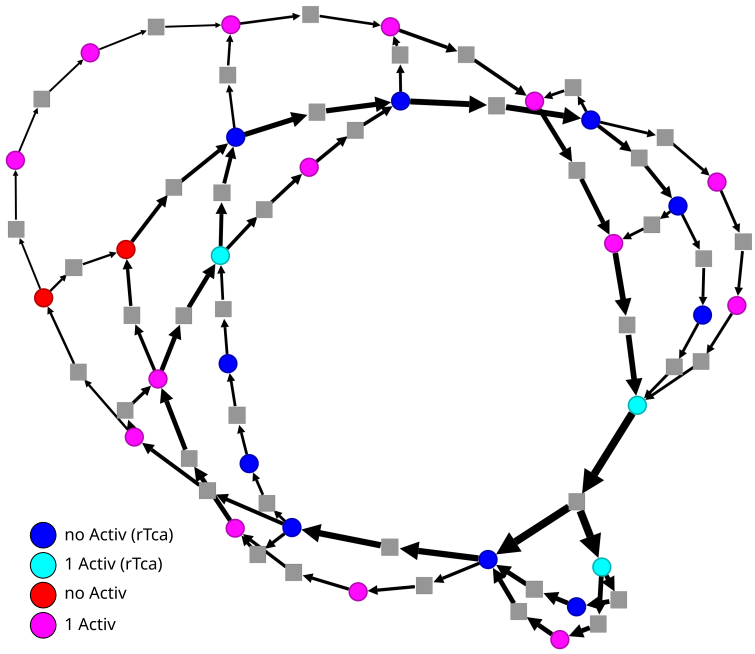
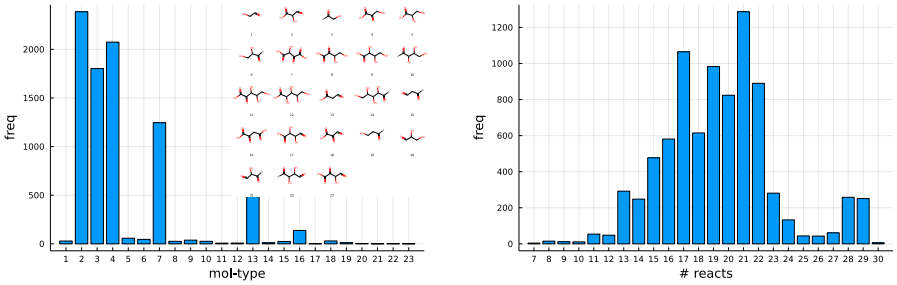


Figure 7.8: Cumulative network of all AC cycles found within the DG of the base activator grammar. Grey, square nodes are reactions and round coloured nodes are compounds. Colour coding of the compounds (shown in the legend) indicates whether compounds are part of the rTCA cycle or not and if they contain an activator group or not. Thickness of an arrow indicates the frequency of occurrence.

the amount of potential autocatalytic cycles, as well as the variation of their network properties. This suggests the possibility for the existence of autocatalytic carbon fixation cycles that, while ruining under a similar type of chemistry, are rather different than the rTCA cycle (whether these pathways are also more effective will be checked in the upcoming section when their cycle energies are compared).

Finally one can look at the properties of the autocatalytic networks found in the activator independent grammar. As outlined above, two different AC-searches were made on the resulting DG, one with more strict search constraints and one with more relaxed constraints. Fig. 7.11 reports these findings, showing that, there are less autocatalytic species, overall, in the AC cycles found with the strict search parameters, compared to the ones found with the relaxed search. In addition, the autocatalytic species for the cycles found with the strict search are larger and more complex than the cycle found with the more relaxed search. However, in the latter case



(a) **(b)**

Figure 7.9: Network related properties of the AC cycles found within the DG of the relaxed activator grammar. **(a)** Different focal species (drawn and enumerated in the box) and their corresponding frequency of occurrence. **(b)** Frequency of occurrence for the total number of reactions of each AC cycle.

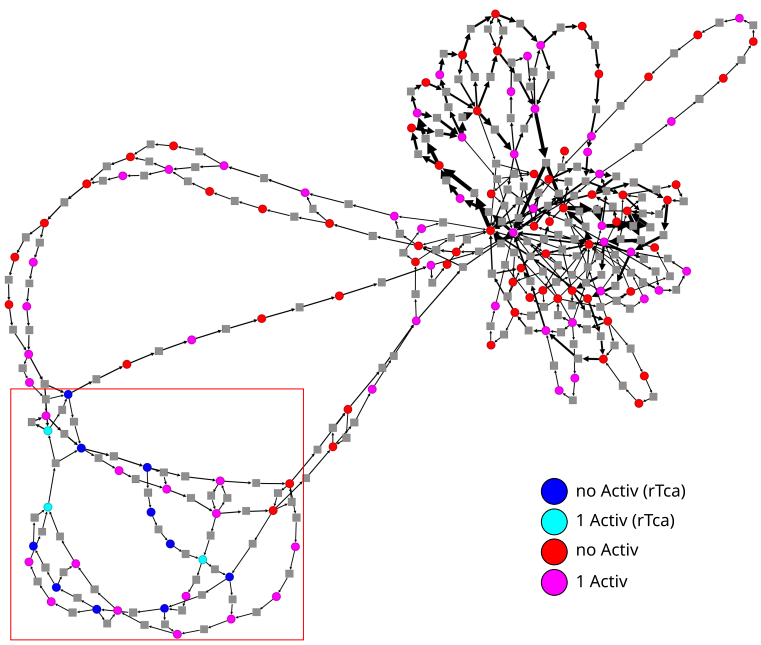


Figure 7.10: Cumulative network of all AC cycles found within the DG of the relaxed activator grammar. Nodes and arrows are drawn as in Fig. 7.8. Red box: the AC cycles found within the relaxed activator grammar (i.e. the cumulative graph of Fig 7.10) is retained as a sub-graph.

the vast majority of cycles has the same type of molecule as auto-catalyst, namely acetaldehyde (CH_3CHO, i.e. a C2-compound which would have been filtered out in the strict search). Nonetheless, in both DG surveys there is a broader distribution in the amount of reactions occurring in each AC cycle

(see Fig. 7.11a and Fig. 7.11c). Thus, while for the relaxed AC search almost all of the identified AC cycles have the same type of molecule as autocatalytic species, they still have different numbers of total reactions, i.e. they may not be functionally equivalent (as it was the case for the results of the AC cycle conducted on the base grammar DG).

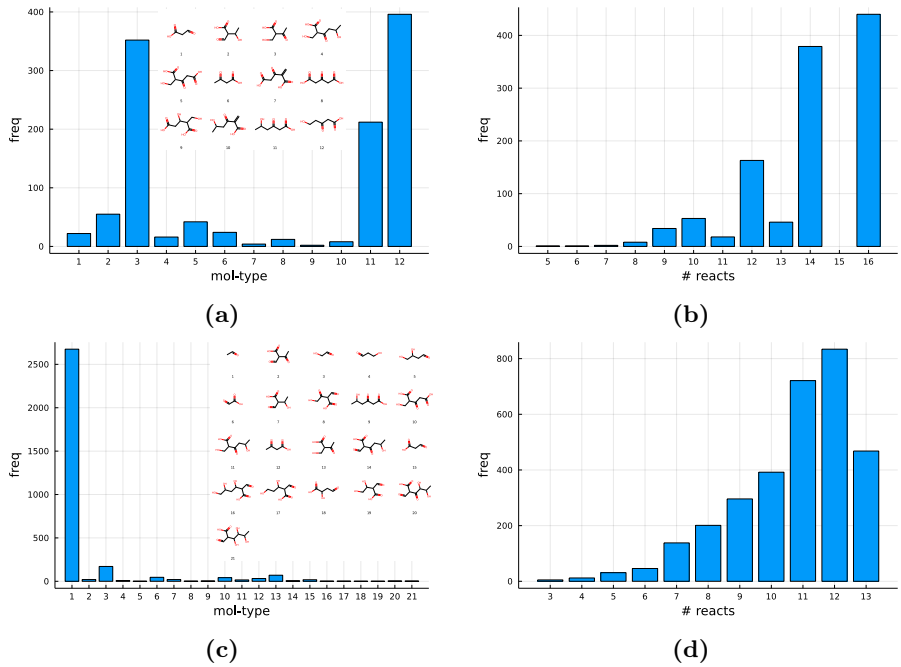
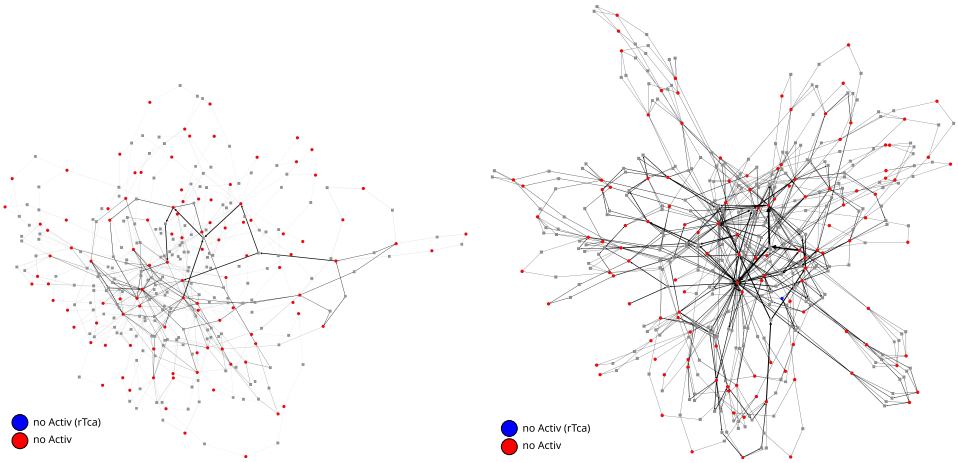


Figure 7.11: Network related properties of the AC cycles found within the DG of the activator free grammar. **(a,b)** Constrained AC search: **(a)** distribution of focal species, **(b)** distribution of cycle lengths. **(c,d)** Relaxed AC search: **(c)** distribution of focal species, **(d)** distribution of cycle lengths.

Last but not least, the cumulative autocatalytic cores can be compared again, like it is reported in Fig. 7.12. Non surprisingly, the cumulative network for the more relaxed AC search is larger and denser, compared with the stricter AC search (compare also Tab 7.4). A remarkable degree of similarity is, nevertheless, manifest between both cumulative networks: the AC core resulting from the stricter search is technically contained in the core found with the more relaxed search, since relaxing the search constraints does not prevent the re-generation of possibilities found through the stricter search. In any case, the results show that in principle it is possible to formulate a chemistry that is independent from the necessity of an activator compound but still leads to the appearance of autocatalytic cycles (however, the loss of

direct CO_2 binding has to be "substituted" through condensation reactions).



(a) **(b)**

Figure 7.12: Cumulative network of all AC cycles found within the DG of the activator free grammar. **(a)** Constrained AC cycle search. **(b)** Relaxed AC cycle search. Nodes and arrows are drawn as in Fig. 7.8.

Grammar type	Nr. Molecules	Nr. Reactions
Base activator	35	42
Relaxed activator	139	216
Activator independent 1	110	208
Activator independent 2	128	305

Table 7.4: Number of molecules and reactions for the cumulative DGs of each AC search.

7.3.2 Cycle energies

Up until this point, the activator group A was treated as an arbitrary compound, i.e. a placeholder. In order to compute the thermochemical properties of the obtained AC cycles, this activator group is now going to be substituted with a concrete type of molecule: a prebiotically plausible placeholder for a coenzyme, i.e. a simpler molecule that could have served the same function under the conditions of an early earth environment. On these lines, the activator group was chosen to be a methanethiol compound (CH_3SH), which creates a thioester bond, like CoA, and whose prebiotic plausibility is well supported [Whicher et al. 2018].

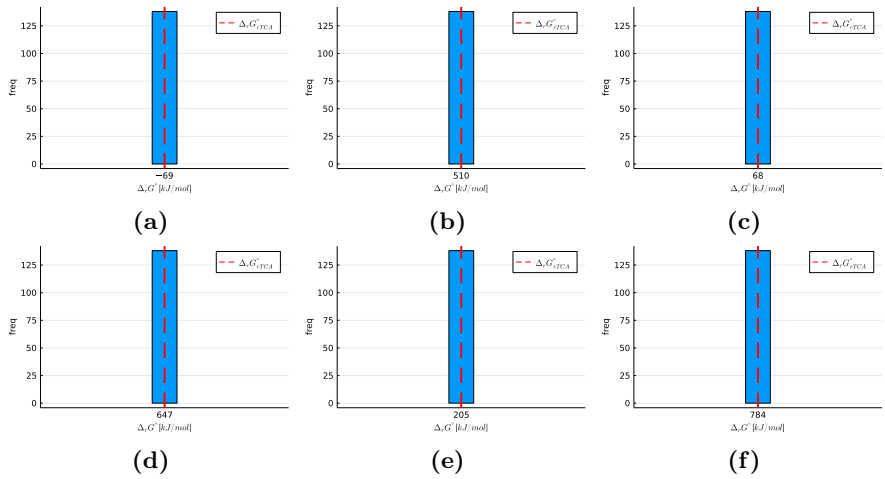


Figure 7.13: Distribution of $\Delta_r G^\circ$ values for the AC cycles obtained with the base activator grammar (and activator substituted with CH_3SH). Energies were calculated for: **(a)** pH=4, $E_0^+ = -0.3\,\text{mV}$, **(b)** pH=4, $E_0^+ = 0.3\,\text{mV}$, **(c)** pH=7, $E_0^+ = -0.3\,\text{mV}$, **(d)** pH=7, $E_0^+ = 0.3\,\text{mV}$, **(e)** pH=10, $E_0^+ = -0.3\,\text{mV}$, **(f)** pH=10, $E_0^+ = 0.3\,\text{mV}$. Red dashed line represents $\Delta_r G^0_{\text{rTCA}}$ calculated for the rTCA cycle under the respective conditions.

As outlined in Ch. 6, in order to calculate the reaction energies for the AC cycles under observation, the component contribution method implemented in the eQuilibrator software will be used. This software package, among other things, allows to calculate the energies under different pHs, as well as different redox potentials (using a correction term – see Eq. 6.7). Therefore, the reaction energies $\Delta_r G^\circ$ for all the AC cycles reported above were calculated for pH values reflecting an acidic, neutral as well as a basic environment (i.e. pH=4, 7, 10) and a redox potential[5] reflecting reducing or oxidising conditions (i.e. $E_0^+ = -0.3\,\text{mV}$, $0.3\,\text{mV}$).

Fig. 7.13 shows the distribution of cycle energies for all AC cycles found within the base activator DG, calculated for all choices of pH and redox potential. Not surprisingly, all cycles have the same overall reaction energy which, is also the same as the reaction energy obtained for the rTCA cycle under these conditions. However, it is important to notice that only under acidic conditions and a negative (i.e. reductive) redox potential (shown in Fig. 7.13a), the AC cycles (including the rTCA cycle itself) have a negative reaction energy, i.e. they can run spontaneously.

[5]In general the reaction energies were calculated using the initial redox potential E_0 associated with the half reaction $H_2 \rightleftharpoons 2\,H^+$.

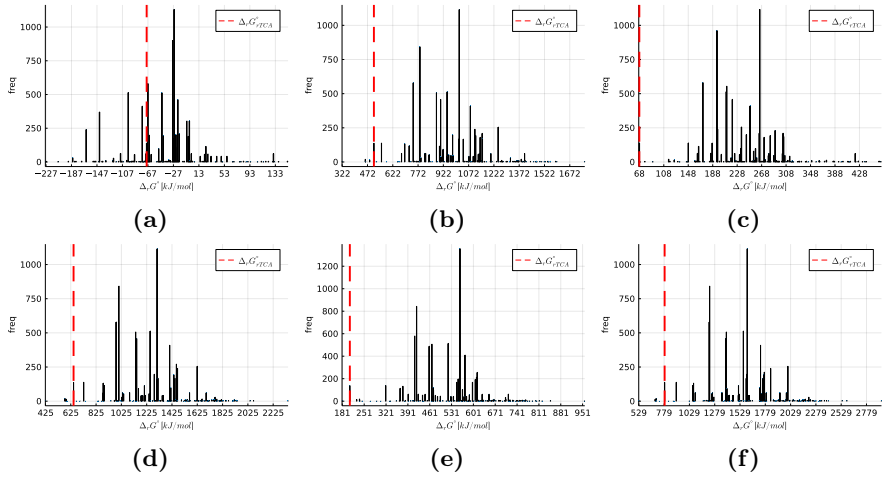


Figure 7.14: Distribution of $\Delta_r G^\circ$ values for the AC cycles obtained with the relaxed activator grammar (and activator substituted with CH_3SH). Energies were calculated for: **(a)** pH=4, $E_0^+ = -0.3\,\text{mV}$, **(b)** pH=4, $E_0^+ = 0.3\,\text{mV}$, **(c)** pH=7, $E_0^+ = -0.3\,\text{mV}$, **(d)** pH=7, $E_0^+ = 0.3\,\text{mV}$, **(e)** pH=10, $E_0^+ = -0.3\,\text{mV}$, **(f)** pH=10, $E_0^+ = 0.3\,\text{mV}$. Red dashed line represents $\Delta_r G^0_{\text{rTCA}}$ calculated for the rTCA cycle under the respective conditions.

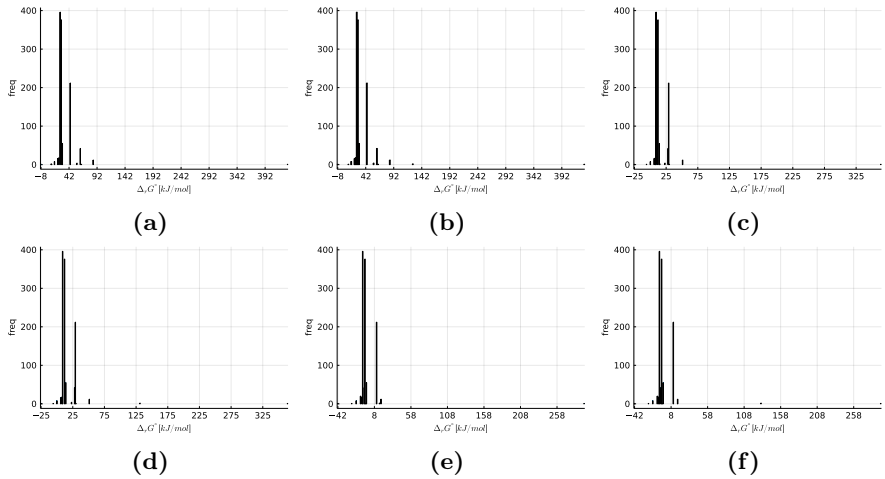


Figure 7.15: Distribution of $\Delta_r G^\circ$ values for the AC cycles obtained with the activator independent grammar and strict AC search. Energies were calculated for: **(a)** pH=4, $E_0^+ = -0.3\,\text{mV}$, **(b)** pH=4, $E_0^+ = 0.3\,\text{mV}$, **(c)** pH=7, $E_0^+ = -0.3\,\text{mV}$, **(d)** pH=7, $E_0^+ = 0.3\,\text{mV}$, **(e)** pH=10, $E_0^+ = -0.3\,\text{mV}$, **(f)** pH=10, $E_0^+ = 0.3\,\text{mV}$.

In turn, Fig. 7.14 shows the distribution of cycle energies for all AC cycles found within the relaxed activator DG, calculated for all choices of pH and redox potential. One can immediately see that there is a broader distribution of different reaction energies. Interestingly, under acidic conditions and a

negative redox potential (shown in Fig. 7.14a), several AC cycles involve a lower overall $\Delta_r G^\circ$ values than those obtained for the rTCA cycle under the same conditions (while for the remaining parameter choices, basically all found cycles have a higher energy). This suggests the possibility that under these environmental conditions it is possible to run autocatalytic carbon fixation pathways that are rather different than the extant rTCA but need comparably less energy making them more likely to occur. Therefore, these pathways are potential candidates for a prebiotic precursor to the rTCA cycle and as such could be the basis for further in-vitro explorations.

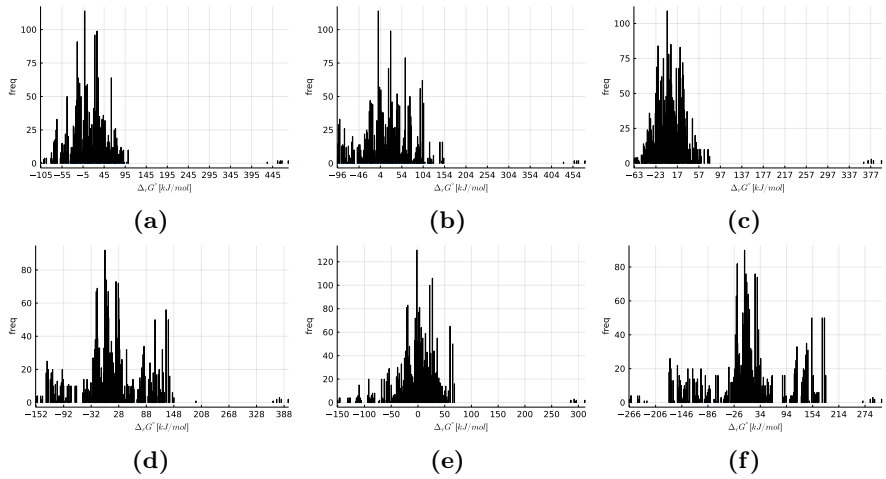


Figure 7.16: Distribution of $\Delta_r G^\circ$ values for the AC cycles obtained with the activator independent grammar and relaxed AC search. Energies were calculated for: **(a)** pH=4, $E_0^+ = -0.3\,\text{mV}$, **(b)** pH=4, $E_0^+ = 0.3\,\text{mV}$, **(c)** pH=7, $E_0^+ = -0.3\,\text{mV}$, **(d)** pH=7, $E_0^+ = 0.3\,\text{mV}$, **(e)** pH=10, $E_0^+ = -0.3\,\text{mV}$, **(f)** pH=10, $E_0^+ = 0.3\,\text{mV}$.

Last but not least, Fig. 7.15 and Fig. 7.16 show the distributions of cycle energies for the AC cycles found within the activator independent DG (again, using a strict as well as a relaxed AC search), calculated for all choices of pH and redox potential. As it could be expected, the distribution of energies for the cycles obtained with the relaxed AC search (shown in Fig. 7.16) is much larger than with the strict AC search. Furthermore, one can see that in both cases one can find AC cycles with a negative $\Delta_r G^\circ$. However, by contrast to the previous results obtained for the activator dependent chemistries, these also include the energies calculated under basic and/or oxidising conditions. This suggests that even though most of these pathways are not carbon fixating per se (as they mostly involve condensation reactions

rather than the direct binding of CO_2 via carboxylation), they can run under a broader range of environmental conditions. Therefore, they could still be precursors to the rTCA, which ran under a different chemistry (a chemistry that is independent from any activating compounds and also less dependent on a certain redox potential or pH) and, since they are still autocatalytic, lead to an exponential increase in relatively larger carbon compounds. These compounds, together with a possible change in environmental conditions, could have been the basis for direct carbon fixation through the rTCA cycle or a more closely related precursor (such a scenario, where direct carbon fixation is a late stage development is also revisited in the discussion chapter) In any case, as with the above obtained rTCA alternatives, further in vitro studies of these pathways could lead to more insights on these lines.

Part III

Simple Reaction Networks under Complex Boundary Conditions

Chapter 8

Membraneless Compartments as Protocells

The aim of the previous chapters was to explore theoretically how complex reaction networks can be modulated by relatively simple boundary conditions. This was done as an approximation to the full complexity of a minimal metabolism (MM), which (according to the scheme proposed in Ch. 3 – see Fig. 3.6) involves complex reaction networks operating under complex boundary conditions, or constraints. Now the complementary approximation is going to be taken, modelling *simple reaction networks* under complex boundary conditions. In other words, the situation is going to be "inverted" (compared to the approach in the previous chapters) with special attention now being put on the effects that non-trivial (i.e., dynamic, non-holonomic) boundary conditions may have even on relatively simple chemistries (so the upper plane of Fig. 3.6 is now considered). More specifically, the focus is going to be put on how variable spatial boundary conditions [Ruiz-Mirazo et al. 2017] in the form of compartments influence chemical reactions and, in particular, their kinetics [Serrano-Luginbühl et al. 2018]. Through this the significance of the formation of proto-cellular compartments for the origin of life is going to be investigated.

8.1 Compartments and proto-cells

One of the "classical" theories about the origin of life, which was briefly mentioned in the introductory chapter, is the 'compartments first' hypoth-

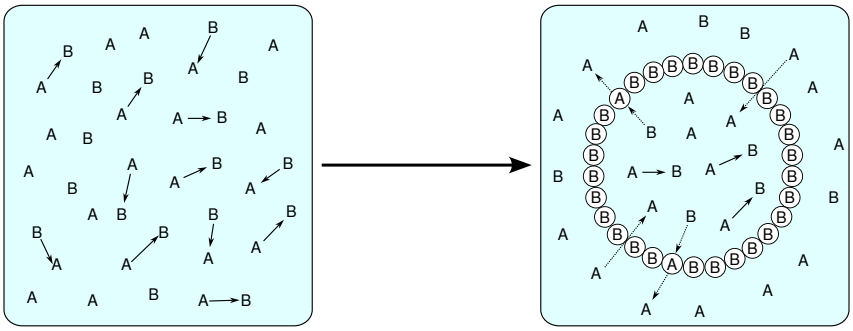


Figure 8.1: Simple scheme for a system of precursor compounds A that react to building blocks B (left panel) which can form a proto-cellular compartment (right panel). Full arrows indicate reactions, dotted arrows indicate processes of diffusion and replacement.

esis. The main idea is, as shown in Fig. 8.1, that chemical reactions, which were occurring on the early earth, converted simple precursors into building blocks[1] that self assembled into proto-cellular compartments. However, these building blocks can also degrade and potentially destroy the compartment. Therefore, they need to be replaced with new building blocks that come from the reaction inside the compartment. In this context, a complementary relationship may be established: the compartment maintains those reactions by preventing the reactants involved from diluting away and, the same time, the reactions repair and maintain the compartment. At its core, this idea of a coupling between reactions and a spatial boundary was presented in the models of autopoiesis and the chemoton, discussed in Ch. 3 (compare also Fig. 8.1 with Fig. 3.2a). The reactions inside the compartment could actually produce an excess of building blocks, which would push it to grow and divide, creating a diverse population of proto-cells. As further shown in Fig. 8.2, the individual cells within this population would then continue to grow and divide, with the potential to encapsulate additional reactions and increase their overall complexity.

While the general idea of this thesis project is to pursue a more systems chemistry motivated approach towards the problem of OL that moves beyond classical "me-first" theories [Krishnamurthy 2020], it is still undeniable that

[1]These building blocks can be relatively simple, like fatty acids or other hydrocarbon species that are found in meteorites. In case more complex organic compounds were required, the reactions that synthesised them could have been autotrophic pathways like the ones discussed in the previous chapters.

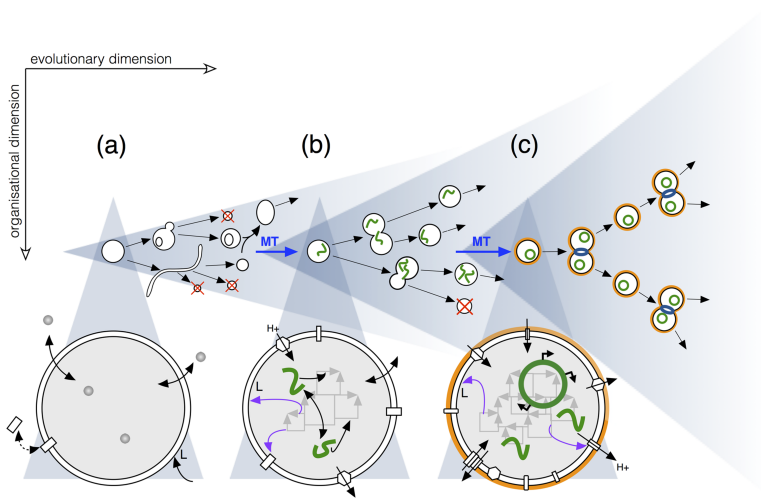


Figure 8.2: Scheme for proto-cell evolution due to growth and division within a population of proto-cellular compartments. Figure taken from [Shirt-Ediss et al. 2017].

compartments play a major role in the control and regulation of metabolism. As discussed in Ch. 2, in extant biology cell membranes not only establish the spatial boundaries of the system but also contribute to energetic control by enabling electrochemical gradients across them [Ruiz-Mirazo et al. 2017]. Furthermore, all known organisms are composed of cells, or as formulated by Harold Morowitz: "All life is cellular in nature" [Morowitz 1992]. Therefore, the formation of a proto-cellular compartment, even if it did not occur in primordial isolation, or in a direct causal sequence of events, as suggested by classical compartments first theories, must have been a major step in the origin of life. In that context, the focus of the current part of this thesis project is to model how spatial constraints can potentially influence simple chemical reaction networks.

8.2 Macromolecular crowding

As mentioned in the general introduction of this dissertation, one of the first theories of proto-cells was formulated by Alexander Oparin with his proposal of coacervates as the first type of compartments capable of housing chemical reactions [Oparin 1924, 1936]. In principle, coacervates are colloids formed by polymers in an aqueous solution as a result of a phenomenon called 'liquid-liquid phase separation' (LLPS), shown in Fig. 8.3a. This phenomenon is

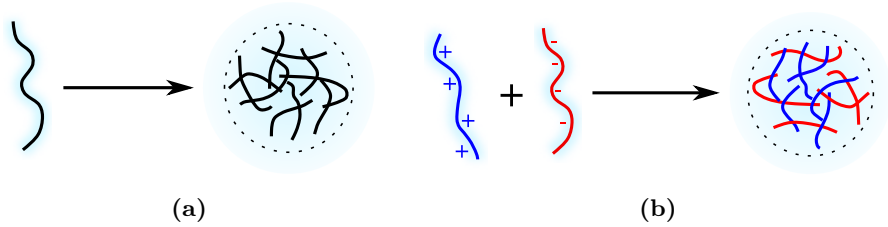


Figure 8.3: Different types of coacervates can form in polymer solutions. **(a)** Uncharged polymers undergo LLPS due to van der Waals interactions forming simple coacervates. **(b)** Complex coacervates involves two differently charged polyions.

caused by electrostatic interactions between the polymers due to either local charge densities or, as shown in Fig. 8.3b, ionic charges (the latter case is often called 'complex coacervates'[2] [Overbeek and Voorn 1957]). Although polymer droplets have no rigid membrane (unlike cells) they form a soft interface [Serrano-Luginbühl et al. 2018] where compounds can diffuse in and out. This is why coacervates are also referred to as 'membrane-less compartments' (MLCs).

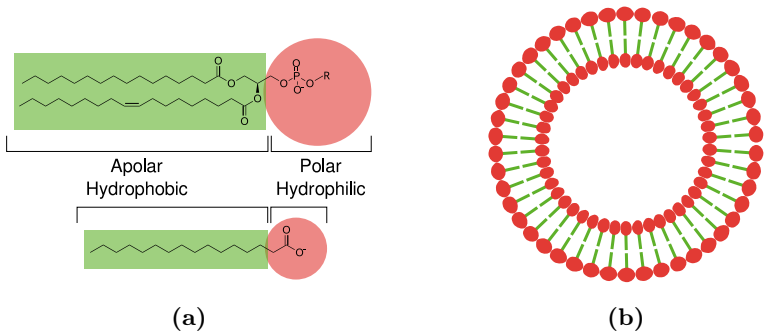


Figure 8.4: **(a)** Simple depiction of an amphiphilic molecule with its ploar (hydrophilic) head-group and it's apolar (lipophilic, hydrophobic) tail-group. **(b)** A collection of amphiphiles in an aqueous solution can self assemble into vesicles.

Nevertheless, the hypothesis of coacervates as proto-cells was left aside for many years, due to the development of an alternative scenario that could also be explored experimentally. This development was based on the discovery that phospholipids and other so called 'amphiphiles' (like fatty acids), when dissolved in aqueous solution, can self assemble into bilayers that lead to spherical structures called 'liposomes' [Bangham and Horne 1964] or 'vesi-

[2]Nevertheless, for the remainder of the text only the the term 'coacervates' will be used.

cles' (see Fig. 8.4). Since cell membranes are made of phospholipid bilayers (as discussed in Ch. 2) this quickly led to vesicles being favoured as models for proto-cells. Among other pioneers, David Deamer contributed, with his work over the years, to establish the 'lipid world' hypothesis, proposing a prebiotic origin of lipid/amphiphilic molecules that then continued to form vesicular compartments [Hargreaves et al. 1977; Deamer and Oró 1980; Deamer 1985, 1986, 1998; Segré et al. 2001]. Additional support for vesicles as proto-cells came from Pier Luigi Luisi, together with Peter Walde and other colleagues, who showed that vesicles formed by simple fatty acids can undergo growth and division (which is necessary for proto-cell evolution, as depicted in Fig. 8.2) [Walde et al. 1994b; Chen and Walde 2010; Luisi 2016]. All in all, the the appeal of vesicles lies in their relative stability, simplicity and their capability of spontaneous self assembly as well as in their apparent similarity to cells [Walde et al. 1994a].

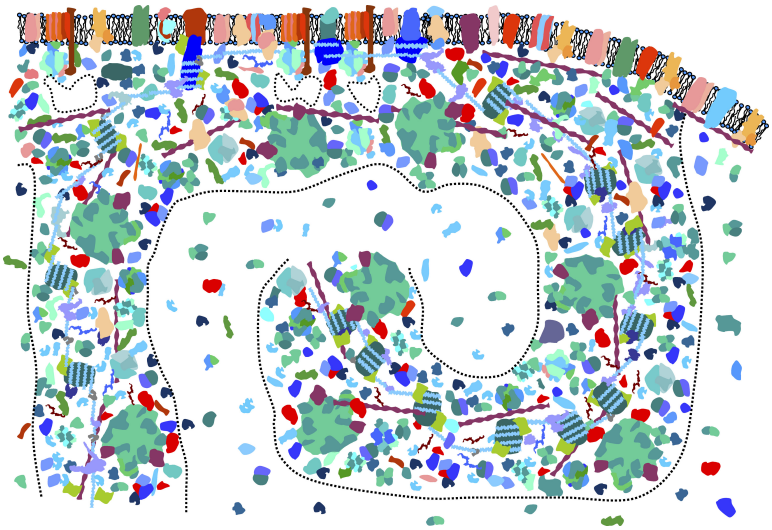


Figure 8.5: Scheme depicting macromolecular crowding and formation of biomolecular condensates within a cell. Figure taken from [Spitzer 2011].

However, while vesicles are usually made of one or two types of amphiphilic compounds, cell-membranes are much more complex, since they contain many different types of phospholipids, as well as proteins that serve as channels, etc. Furthermore, vesicles are rather "empty" as they mostly contain water (although they can be prepared to encapsulate different chemical or bio-chemical species [Walde and Ichikawa 2001]). The cytoplasm within cells, on the other hand, is densly crowded, a multitude of biopoly-

mers and many other biomolecules (see Fig. 8.5). This is nowadays described as 'macromolecular crowding' [Zhou et al. 2008; Spitzer 2011] and it has furthermore been shown that the formation of so called 'biomolecular condensates' due to LLPS within cells might play a vital role in certain cell functions[3] [Shin and Brangwynne 2017; Mitrea and Kriwacki 2016; Banani et al. 2017; Boeynaems et al. 2018]. Some authors even take a rather radical point of view, claiming that the cell membrane serves no fundamental purpose and the main function of the cell is realized through the condensates and gels formed by the cytoplasm [Pollack 2001; Trevors and Pollack 2005].

In any case, all this has lead to a sort of "rediscovery" of coacervates as candidates for proto-cells (see for example [Donau et al. 2020]). Some authors even try to combine both types of compartments, i.e. coacervates within vesicles, as plausible models for proto-cells [Tang et al. 2014; Love et al. 2020]. Here the focus will be put specifically on LLPS and how it could create boundary conditions that influence the kinetics of simple reactions. This will be done through a computational model that simulates the LLPS within heterogeneous solutions. As this model will be based on concepts from statistical mechanics, some knowledge about the physics of LLPS will be outlined next, in the final section of this chapter.

8.3 Statistical mechanic of LLPS

When putting together two different liquids (e.g. dissolving ink in water), they will typically proceed to form a well-mixed (WM) solution, consisting of a single phase that homogeneously contains both liquid components (see Fig. 8.6). However, there are some compounds (like oil or polymers) that, when put together with a solvent (e.g., water again) behave like liquid solutes but do not mix so well. Instead, they form a phase-separated (PS) solution, consisting of two phases: one phase that contains the solvent and another phase that contains the solute (see Fig. 8.6). As both phases are liquid one usually speaks of 'liquid-liquid phase separation' (LLPS). The phase separating behaviour of such a solution can be described by the free energy F of the system, which has the following general expression:

$$F = U - TS = F_{int} + F_S \tag{8.1}$$

[3]In some texts also the term 'membrane-less organelles' (MLOs) is used.

where U is the internal energy due to electrostatic interactions (attraction, repulsion, etc.) between the solvent and solute molecules, S is the entropy of the system and T the temperature. One can therefore express the free energy in Eq. 8.1 as the sum of an interaction term F_{int} and an entropic term F_S. A system of this kind will always tend to go to an equilibrium configuration of minimal energy. If there are no interactions (or very small ones) between the components, Eq. 8.1 is dominated by the entropic term and its equilibrium configuration is a configuration of maximum entropy, i.e a WM-state. However, if there are more significant interactions between the components (e.g. attraction between the solutes, or repulsion between solutes and solvent), Eq. 8.1 is dominated by the interaction term. A WM-state would then lead to an energetically less favourable configuration, which is why the system stays in a PS-state. In other words, depending on whether F_{int} or F_S dominates Eq. 8.2 (i.e. which term is larger), the system will either go to a WM- or a PS-state (see again Fig. 8.6).

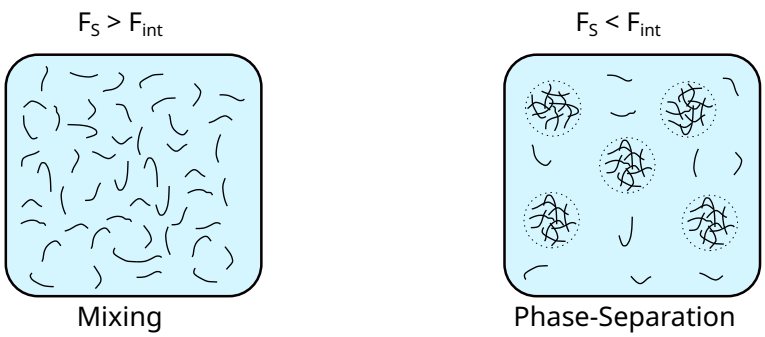


Figure 8.6: LLPS within an aqueous solution containing one type of polymer species. If $F_S > F_{int}$ the systems stays well mixes (left panel), if $F_S < F_{int}$ the system phase separates (right panel).

Both the internal energy and the entropy are functions that depend on parameters like the temperature of the system, the concentration (or the volume fraction) of the components, as well as the interaction strength between the molecules. Therefore, whether a system phase separates or not depends on various factors. One can derive a more detailed expression for the free energy of the system using the so called Florry-Huggins model [Flory 1942; Huggins 1942], resulting in the following formula:

$$\frac{F(\phi_0, \phi_1)}{N k_B T} = \underbrace{\phi_0 \ln \phi_0 + \phi_1 \ln \phi_1}_{F_S} + \underbrace{\chi_{01} \phi_0 \, \phi_1}_{F_{int}} \tag{8.2}$$

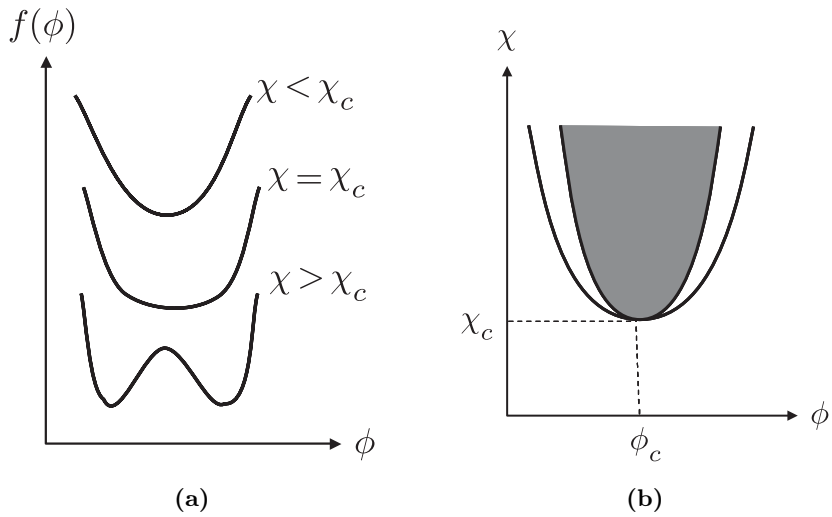


Figure 8.7: (a) Free energy as a function of ϕ for different values of the interaction parameter χ. **(b)** Phase diagram for a binary mixture: grey area indicates the region of phase separation, the area between both curves indicates regions of metastability. Figures taken from [Doi 2013].

where k_B is the Boltzmann constant an N is the total number of components in the system. χ_{01} represents the interaction between solute and solvent and is, in general, a complex function dependent on temperature but also on other factors, like pH or the concentration of counter-ions in the system. The terms ϕ_0, ϕ_1 are the volume fractions of the solvent and solute, respectively (representing their corresponding concentrations). In such cases, given that there is a single solute species, it is possible to express $\phi_1 = \phi = 1 - \phi_0$, Eq. 8.2 can thus be simplified to:

$$\frac{F(\phi)}{Nk_BT} = (1-\phi)\ln(1-\phi) + \phi\ln\phi + \chi_{01}\phi(1-\phi) \tag{8.3}$$

This expression is also called the regular solution model for a binary (two component) system.

Eq. 8.3 allows to derive phase diagram of the system that indicates whether, for a given parameter choice, a system is phase separated or not[4]. In general, there is a critical value χ_C Below which the free energy (as a function of the volume fraction ϕ) is concave, indicating that there is no phase separation (see Fig. 8.7a). Above this critical value, however, the free energy

[4]Similar to a phase diagram that indicates if a material is solid, liquid, etc.

function has a convex part, which means indicating that for any choice of ϕ within this convex region phase separation will occur. As a result, the phase diagram can be drawn, showing whether the system phase separates or stays well mixed for each pair of values for ϕ, χ (see Fig. 8.7b).

This approach can be generalized to more complex mixtures, since the formation of biomolecular condensates generally involves a heterogenous solution that consists of many components which interact in multiple ways. On those lines, the regular solution model of Eq. 8.3 can be transformed into a multi component system:

$$\frac{F(\phi_0, \ldots, \phi_N)}{N k_B T} = \sum_{i=0}^{N} \phi_i \ln \phi_i + \frac{1}{2} \sum_{i=0}^{N} \sum_{j=0}^{N} \chi_{ij} \phi_i \phi_j \tag{8.4}$$

A mean field model like this has been used to study LLPS in biological systems, [Jacobs and Frenkel 2017; Jacobs 2021]. However, as argued recently in [Lauber et al. 2023a], there are some shortcomings of the regular solution model (not just for multi component systems but also for binary systems). In particular, when modelling a system that involves both attractive and repulsive interactions (for instance, in the complex coacervation of poly-ions), Eq. 8.4 fails to reproduce the experimentally obtained phase diagrams. Taking an alternative approach (i.e., starting from the partition function and using cavity methods), it was shown that a mean field model based on the Bethe-Peierls approximation (which is also used in modelling spin-glasses) not only improved the predictions made by the regular solution model but also managed to reproduce phase diagrams for systems where the regular solution model failed [Lauber et al. 2023a]. It was furthermore shown that a similar behaviour can be achieved by a microscopic simulation, based on discretizing the system through a regular lattice formalism. In the upcoming chapter a similar approach is going to applied in order to study how the formation of a heterogeneous condensate of catalysts can have important effects on the kinetics of a simple chemical reaction system.

Chapter 9

Lattice Model Simulations

To simulate liquid-liquid phase separation (LLPS) phenomena within a heterogeneous solution and the influence that these might have on a simple reaction network, a lattice model based simulation will be implemented. This strategy has been adopted from previous work demonstrating that such coarse grained numerical simulations are better at capturing the complex microscopic dynamics of phase separating systems in comparison to classical macroscopic mean field models [Lauber et al. 2023a,b]. Indeed, lattice models offer a more explicit spatial framework to deal with phenomena like crowding or phase separation. Additionally, they can be combined with a microscopic, statistically coherent formulation of the chemical processes to be explored. Examples for this can be found in [Glotzer et al. 1994; Zhdanov 2000; Berry 2002; Schnell and Turner 2004].

9.1 Simulation of LLPS

The lattice model that proposed here is based on the classical Ising-model used to simulate magnetization in statistical physics [Ising 1925]. More precisely, it is a variation of the Ising model (similar to the Cellular Potts model [Graner and Glazier 1992] or the lattice-gas model [Frisch et al. 1986]), an approach that has already been taken up to simulate the phase separating behaviour of biomolecular condensates [Jacobs and Frenkel 2017; Jacobs 2021]. As depicted in Fig. 9.1, in order to simulate LLPS within a mixture of solutes, space is discretized into a regular lattice with N sites[1] where

[1] In general a solution would be represented by a three dimensional lattice. However, since most of the upcoming modelling work is done in two dimensions (and it is also easier

each lattice site-i is in a state $\sigma_i = \mu \in \mathbb{N}$, representing the type of compound occupying the respective site, with $\mu = 0$ standing for the solvent and $\mu \neq 0$ for a solute[2]. Assuming there are L different solutes within the solution, this means that $\sigma_i = 0, \ldots, L$ (i.e., there $L+1$ different compounds in general). A certain lattice configuration is then given by the state vector $\vec{\sigma} = (\sigma_1, \ldots, \sigma_N)$.

9.1.1 Initializing the lattice

If the total amount of type μ compounds on the lattice is denoted as N_μ, then the following relation holds: $\sum_{\mu=0}^{L} N_\mu = N$, since it is assumed that the lattice is completely filled, i.e. each lattice site is either occupied by a solvent or a type of solute. The corresponding volume fraction ϕ_μ of type μ compounds on the lattice can therefore be obtained as:

$$\phi_\mu = \frac{N_\mu}{N} \ , \qquad \sum_{\mu=0}^{L} \phi_\mu = 1 \tag{9.1}$$

In turn, the volume fraction ϕ_0 of the solvent can be obtained as:

$$\phi_0 = 1 - \phi_{tot} \ , \qquad \phi_{tot} = \sum_{\mu=1}^{L} \phi_\mu \tag{9.2}$$

with ϕ_{tot} being the total volume fraction of solutes on the lattice.

In the context of the lattice model simulation, the volume fraction ϕ_μ corresponds to $P(\sigma_i = \mu)$, the probability of a lattice site i being in the state μ, i.e. being occupied by a molecule of type μ. Thus, to initialize the lattice for the LLPS simulation, for each site-i the state σ_i is chosen randomly from the set $\mu \in \{0 \ldots, L\}$ with a probability $p_\mu = \phi_\mu$. Due to Eq. 9.1 these probabilities follow the the corresponding discrete probability distribution:

$$P(\sigma_i = \mu) = p_\mu \ , \quad \sum_{\mu=0}^{L} p_\mu = 1 \tag{9.3}$$

The volume fractions ϕ_μ on the lattice are therefore necessary simulation

to illustrate), all figures within this and the following chapter show 2D lattices.

[2]The amount of molecules of a realistic solution is typically much bigger than the amount of sites on a standard lattice, so it would be more adequate to say that a lattice site represents a patch of compounds rather than a single molecule.

parameters that are needed to initialize the lattice (however, in practice only the volume fractions of the solutes are needed, since the solvent volume fraction remains deducible from Eq. 9.2).

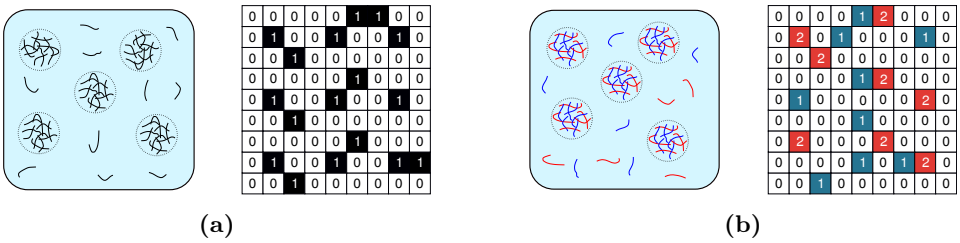


Figure 9.1: Examples for the discretization of a solution as a two dimensional $d \times d$ square lattice. **(a)** Mixture containing one solute and a solvent (binary mixture), i.e. $\sigma_i = 0, 1$. Black squares represent the solute and white squares represent the solvent. **(b)** Mixture containing two solutes and a solvent (ternary mixture), i.e. $\sigma_i = 0, 1, 2$. Blue squares represent solute-1 and red squares represent solute-2 (white squares again represent the solvent).

9.1.2 Equilibrating the lattice

Once the lattice is initialized, one can proceed to model the phase separation behaviour of the solutes that are represented on the lattice. As outlined in the previous chapter, the occurrence of LLPS in a solution is due to attractive or repulsive interactions among the components of the system, which leads to a change in the energy of the system. Within this microscopic model, the energy for a certain lattice configuration $\vec{\sigma}$ is then given by the hamiltonian:

$$H(\vec{\sigma}) = -\sum_{\langle i,j \rangle} J(\sigma_i, \sigma_j) \tag{9.4}$$

where brackets in the sum stand for neighbouring lattice sites (i.e. only nearest neighbour interactions are considered) and the function $J(\mu, \nu)$ determines the strength of interaction between type μ and ν compounds. Attraction or repulsion forces are assumed to be symmetric and constant (i.e., no variation in interaction strengths), so $J(\mu, \nu)$ can be represented by the entries $J_{\mu\nu}$ of a symmetric matrix $\mathbf{J}$, i.e. $J_{\mu\nu} = J_{\nu\mu}$. Given the sign convention in Eq. 9.4, $J_{\mu\nu} > 0$ corresponds to attraction, while $J_{\mu\nu} < 0$ corresponds to repulsion. For the examples of a binary lattice (shown in Fig. 9.1a) or a ternary lattice (shown in Fig. 9.1b), the respective interaction matrices take

the forms:

$$\mathbf{J}_{bin} = \begin{pmatrix} J_{00} & J_{01} \\ J_{01} & J_{11} \end{pmatrix} , \qquad \mathbf{J}_{tern} = \begin{pmatrix} J_{00} & J_{01} & J_{02} \\ J_{01} & J_{11} & J_{12} \\ J_{02} & J_{12} & J_{22} \end{pmatrix}$$

Thus, in general, if there are L solutes on the lattice, this means that $\mathbf{J}$ is a $(L+1) \times (L+1)$ matrix. The entries $J_{\mu\nu}$ are then additional simulation parameters that need to be set.

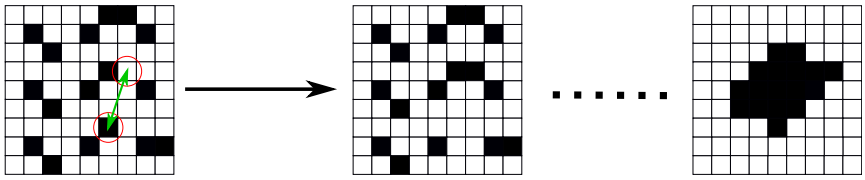


Figure 9.2: Depiction of the Metropolis Monte Carlo algorithm utilised to simulate LLPS: the state of two lattice sites is swapped and the swap accepted if it leads to an energetically more favourable lattice configuration. Subsequent swap moves will eventually lead to an equilibrium configuration.

If it is not disturbed, the system of solvent and solutes will tend towards an equilibrium configuration $\vec{\sigma}_{eq}$, which is also the configuration of minimal energy, as it was explained in the previous chapter. Therefore, depending on the strength of interactions among the compounds (and if they are attractive or repulsive), together with other factors like temperature, pH, etc., the equilibrium configuration of the system is either a well-mixed system (WM) or a phase-separated one (PS). This equilibrium configuration can be reached on the lattice, by implementing a Metropolis Monte Carlo algorithm that uses Kawasaki dynamics [Kawasaki 1966] (see Fig. 9.2), which ensures that the number of compounds of each type stays constant. The steps of this algorithm are then:

(i) Given the current lattice configuration $\vec{\sigma}$, select two random sites i, j in the respective states σ_i, σ_j.

(ii) Swap the states of the two lattice sites: $\sigma_i \mapsto \sigma_i' = \sigma_j$ and $\sigma_j \mapsto \sigma_j' = \sigma_i$, creating a new lattice configuration $\vec{\sigma}'$.

(iii) Compute the energy difference between the new and the old configurations: $\Delta H = H(\vec{\sigma}') - H(\vec{\sigma})$

(iv) Accept new configuration with probability $P_{swap} = \min(1, e^{-\beta\Delta H})$

The definition of the acceptance probability P_{swap} makes sure that configurations that lead to a lower energy are always accepted, while configurations that lead to a higher energy are only accepted with a probability that follows a Boltzmann-Gibbs distribution (where the parameter $\beta = 1/T$ represents the inverse of the temperature T of the system, a common parameter choice in statistical mechanics models). In other words, P_{swap} represents the thermal fluctuations within the system: choosing a small value for β, which corresponds to a higher temperature, will lead to a greater probability of accepting energetically unfavourable configurations (i.e. more statistical noise in the system). By contrast, choosing a high value for β, which corresponds to a smaller temperature, will cause the system to "freeze". Thus, β is the final simulation parameter that is needed to run the LLPS lattice model simulation (all simulation parameters are collected in Tab. 9.1).

Parameter	**Name**	**Function**
ϕ_μ	Volume fraction	determines the amount of μ molecules on the lattice
$J_{\mu\nu}$	Interaction matrix entries	determines the strength and type of interaction between compounds μ, ν
β	Inverse temperature	determines the temperature of the system

Table 9.1: Collection of the lattice model parameters.

As shown in Fig. 9.3 for the example of a binary system (as well as in Fig. **??** for the example of a ternary system), the equilibrium configurations obtained through microscopic numerical simulation are in good agreement with the phase separating behaviour predicted by the corresponding phase diagrams of the systems (which were obtained through the mean field model defined in [Lauber et al. 2023a], where these results were also tested and proven more thoroughly). Comparison between Fig. 9.3 and Fig. **??** also shows that going from a binary to a ternary system leads to a more complex phase separating behaviour. A binary system generally separates into two phases (a solvent phase and a solute phase). Instead, in a ternary system, depending on the interactions between the components (i.e. the entries of matrix $\mathbf{J}$) there can be various combinations of phases (well characterised

in [Lauber et al. 2023a]). For instance, one can find a two-phase system with a solvent phase and a phase containing both solutes (associative phase separation), a three-phase system with a solvent phase and two separate solute phases (segregative phase separation) or a two-phase system with one separate solute phase and the other one mixed with the solvent (counter-ionic de-mixing).

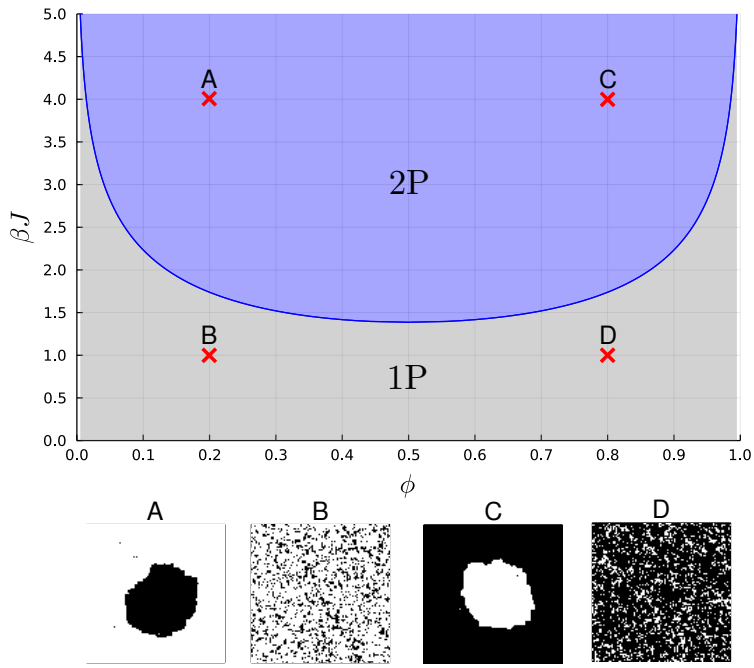
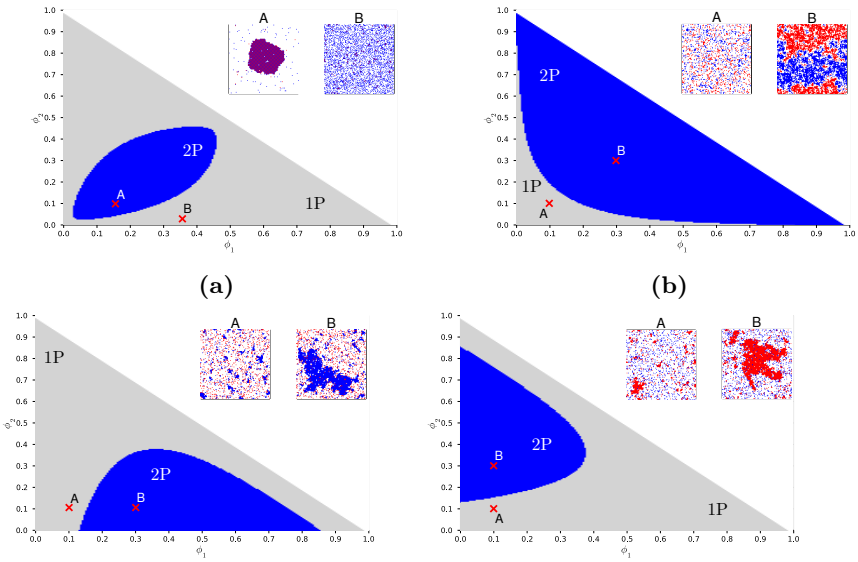


Figure 9.3: Phase diagram for a binary mixture: blue area indicates regions of phase separation (i.e. two phases), grey area indicates regions of mixing (i.e. one phase). Phase diagram obtained with the mean-field method of [Lauber et al. 2023a]. Insets show the corresponding equilibrium configurations obtained with the microscopic lattice model simulation for the corresponding parameter choices.

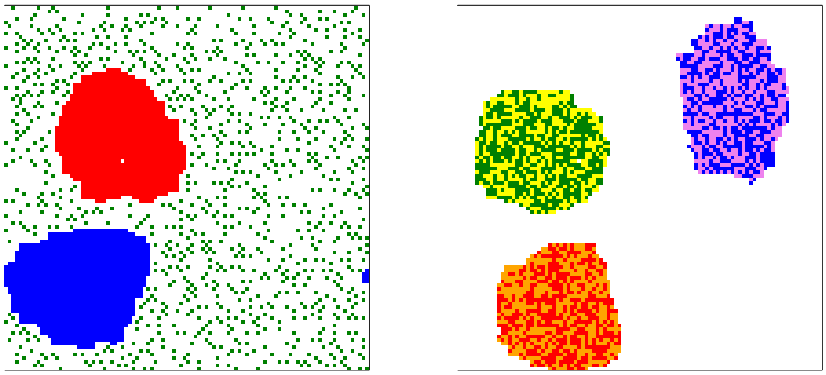
In general, having multiple solutes in the system leads to more complex phase separating behaviour (see Fig.9.5) because there are more possible combinations of phases (this has also been discussed in the literature [Harmon et al. 2017; Dignon et al. 2020; Fare et al. 2021]). Nevertheless, the microscopic lattice model defined here is in general able to capture, sometimes even quantitatively, the phase separating behaviour of more complex and heterogeneous mixtures. It is therefore a valid choice in simulating the LLPS of biomolecular condensates and the potential effect that these have on a simple reaction system.



Phase diagrams for a ternary mixture depicting: **(a)** associative de-mixing **(b)** associative de-mixing **(c)**,**(d)** counter-ionic de-mixing. Blue areas indicates regions of phase separation (i.e. two or three phases), grey areas indicates regions of mixing (i.e. one phase). Phase diagrams obtained with the mean-field method of [Lauber et al. 2023a]. Insets show the corresponding equilibrium configurations obtained with the microscopic lattice model simulation for the corresponding parameter choices.

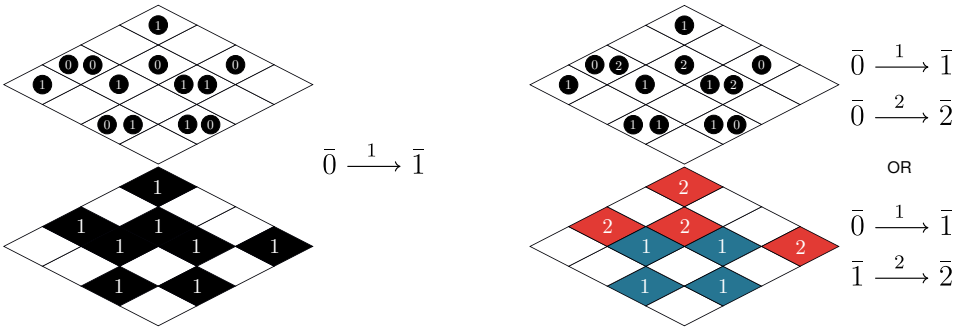
9.2 Reaction-diffusion model

The lattice model presented above will be extended by including different types of smaller substrates, with the aim of modelling the effect of LLPS on a simple reaction system. As such, these substrates will be allowed to diffuse through the system and undergo reactions that transform one type of substrate into another type. The catalysts of these reactions are, precisely, the phase separating solutes which are assumed to be more complex molecular structures (e.g., oligopetides). Within the simulation this is achieved, as shown in Fig. 9.6, by extending the initial lattice containing the solutes (or catalysts in this context) with a second layer that contains the substrates. These two layers will be referred to as 'substrate lattice' and 'catalyst lattice'. As also shown in Fig. 9.6, while each site on the catalyst lattice is occupied by one type of compound, each site on the substrate lattice can be occupied by multiple compounds of various types. This is to reflect that the



(a) **(b)**

Figure 9.5: Two examples for a more complex phase separation behaviour. **(a)** A system containing three solutes and a solvent (i.e., four components in total) forming a three-phase system: two phases, each containing a single type of solute (red and blue droplets) and a solvent phase mixed with the remaining solute (green). **(b)** A system containing five solutes and a solvent (i.e., six components in total) forming a four-phase system: three phases, each containing two types of solutes (red-orange, green-yellow and blue-purple droplets) and a solvent phase.



(a) **(b)**

Figure 9.6: Examples for the extension of the lattice model with a reaction-diffusion model for simple, catalysed reactions. **(a)** Binary mixture and the corresponding substrate lattice as well as the reaction happening in the system. **(b)** Ternary mixture and the corresponding substrate lattice as well as the two possible reaction systems that can occur in this system.

catalysts are much bigger than the substrates and thus take up more space, while the substrates will "percolate" the lattice.

Each type of substrate is denoted[3] as $\bar{\alpha} \in \mathbb{N}$ and can undergo a reaction of the form: $\bar{\alpha} \xrightarrow{\mu} \bar{\beta}$. In other words, the solute of type μ catalyses the

[3] As within the simulation the different types of substrates, like the solvents, are represented by a natural number, an overline is added in the text to distinguish them.

reaction that transforms a substrate of type $\bar{\alpha}$ to a substrate of type $\bar{\beta}$. Each catalysed reaction is assumed to be unique, thus each solute catalyses one specific transformation (see the example for a ternary system in Fig. 9.6b). Effectively this means that if there are L solutes/catalysts, there are L reactions and therefore $L+1$ substrates, i.e. $\bar{\alpha} = \bar{0}, \ldots, \bar{L}$. To initialize this reaction-diffusion model, for each substrate $\bar{\alpha}$, $C_{\bar{\alpha}}$ copies are placed at random sites of the substrate lattice, with $C_{\bar{\alpha}}$ becoming an additional set of parameter values for the simulation.

The model also includes the possibility of attraction or repulsion forces occurring between the substrates, on the one hand, and the catalysts or the solvent, on the other hand. Like in the previous section (describing phase separation), this interaction causes a change in the energy of the system. More precisely, the energy contribution for a substrate $\bar{\alpha}$ being above a lattice site occupied by a compound μ is defined as:

$$E(\bar{\alpha}, \mu) = -I_{\bar{\alpha}\mu} \tag{9.5}$$

where $I_{\bar{\alpha}\mu}$ are the entries of a second interaction matrix, $\mathbf{I}$ that determines the interaction strength between a substrate $\bar{\alpha}$ and a compound μ. Given the sign convention of Eq. 9.5 (similar to Eq. 9.4), $I_{\bar{\alpha}\mu} > 0$ corresponds to attraction, while $I_{\bar{\alpha}\mu} < 0$ corresponds to repulsion. For the examples of a binary or a ternary lattice (see Fig. 9.6a or Fig. 9.6b respectively), the corresponding interaction matrices take the forms:

$$\mathbf{I}_{bin} = \begin{pmatrix} I_{\bar{0}0} & I_{\bar{0}1} \\ I_{\bar{1}0} & I_{\bar{1}1} \end{pmatrix} , \qquad \mathbf{I}_{tern} = \begin{pmatrix} I_{\bar{0}0} & I_{\bar{0}1} & I_{\bar{0}2} \\ I_{\bar{1}0} & I_{\bar{1}1} & I_{\bar{1}2} \\ I_{\bar{2}0} & I_{\bar{2}1} & I_{\bar{2}2} \end{pmatrix}$$

Thus, if there are L catalysts on the respective lattice layer, this means that $\mathbf{I}$ is a $(L+1) \times (L+1)$ matrix. However, in contrast to $\mathbf{J}$, matrix $\mathbf{I}$ is not symmetric[4]. The entries $I_{\bar{\alpha}\mu}$ are again additional simulation parameters that need to be taken into account in any simulation.

On quite realistic, physico-chemical grounds, the reaction and diffusion of the substrates are assumed to happen on a faster time scale compared

[4]Even tough attraction and repulsion are in general symmetric interactions, $E(\bar{\alpha}, \mu)$ explicitly captures the energy contribution of a substrate being "on top of" a certain compound. In other words, matrix $\mathbf{I}$ is not symmetric due to the way the simulations (and the corresponding code) are implemented.

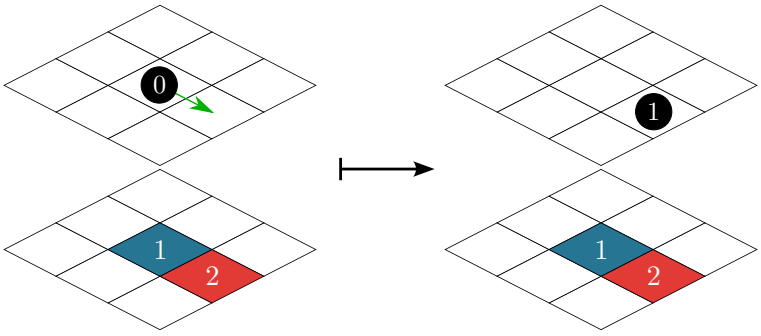


Figure 9.7: Depiction of the random walk algorithm utilised to simulate the reaction-diffusion of the substrates: a randomly selected substrate is moved to a neighbouring lattice site (also randomly selected). If the substrates lands on a catalysts, it gets transformed according to the corresponding reaction (if existent).

to the LLPS of the catalysts. Thus the simulation of those faster processes will be conducted on an equilibrium configuration $\vec{\sigma}_{eq}$ of the catalyst lattice, obtained through the LLPS simulation, which runs separately. The diffusive dynamics of the substrates will be modelled as a standard 'Brownian motion' phenomenon: namely, as a simple lattice random walk on the respective layer. In practice, the interaction between the substrates and the catalyst (or the solvent) amounts to a bias in the random walk across the lattice. A Metropolis Monte Carlo type algorithm (similar to the one used for simulating LLPS, described in the previous section) is used, as depicted in Fig. 9.7. The steps of this algorithm are then:

(i) Given an equilibrium configuration of the catalyst lattice $\vec{\sigma}_{eq}$ and a distribution of substrates on the substrate lattice, select a random site i in the state $\sigma_i = \mu$ with non-zero amount of compounds on the substrate lattice.

(ii) Select a random compound $\bar{\alpha}$ from the substrate lattice and from the lattice sites neighbouring site i, select a random site j in the state $\sigma_j = \nu$.

(iii) Move $\bar{\alpha}$ to lattice site j and compute the energy difference between $\bar{\alpha}$ being above site j or above site i: $\Delta E = E(\bar{\alpha}, \nu) - E(\bar{\alpha}, \mu)$

(iv) Accept the move with probability: $P_{move} = \min(1, e^{-\beta \Delta E})$

(v) If the move is accepted and ν catalyses a reaction $\bar{\alpha} \xrightarrow{\nu} \bar{\beta}$, change compound $\bar{\alpha}$ to $\bar{\beta}$.

The last step models the reaction process and assumes that reactions are irreversible and there is perfect catalysis, i.e. every time a substrate encounters an adequate catalyst it gets immediately transformed to the corresponding product substrate, which starts its random walk from the position of the catalyst that has led to it. As with the LLPS simulation, P_{move} represents the thermal fluctuations within the system due to its temperature (which is again set by the inverse temperature β, a key parameter to run the reaction-diffusion simulation, too). All the parameters for this model are collected in Tab. 9.2

Parameter	**Name**	**Function**
$C_{\bar{\alpha}}$	Substrate concentration	determines the amount of $\bar{\alpha}$ substrates on the lattice
$I_{\bar{\alpha}\mu}$	Interaction matrix entries	determines the strength and type of interaction between substrate $\bar{\alpha}$ and compound μ
β	Inverse temperature	determines the temperature of the system

Table 9.2: Collection of the reaction-diffusion model parameters.

Chapter 10

Modelling Phase Separation of Catalysts

As outlined in the previous chapters, the main focus of this part of the thesis project is to investigate the influence that the formation of condensates through liquid-liquid phase separation (LLPS) can have on a reaction system. More specifically, given recently collected evidences in cell physiology concerning the phase separating behaviour of various types of proteins and the key role that this may play in their corresponding functions [Boeynaems et al. 2018], the work here is going to be focused on the formation of biomolecular condensate among a group of catalysts and the influence that this could have on the underlying reaction kinetics. The molecules with potential to aggregate (typically, due to weak attractive interactions among them) can be thought to be enzymes/proteins, as it is usually the case in extant metabolic pathways, but also some other organocatalyst, like an oligopeptide or an RNA-oligonucleotide, which would be more relevant in the context of a proto-cell or under prebiotic conditions. The LLPS of these biomolecules is simulated with a lattice model and then extended with a reaction diffusion model to explore the effects on their catalytic action on a linear pathway. The influence of the LLPS of the catalysts on this reaction sequence is analysed in depth, both considering the key parameters of the simulation model and the corresponding effects, in particular, on the global reaction time (i.e., the pathway kinetics).

10.1 Cascade reaction

The reaction system that is going to be studied is a linear pathway where an initial substrate S_0 is transformed through consecutive steps and intermediary substrates S_α into a final product P. Each reaction step is catalysed by a specific biomolecule E_μ, i.e. $\mathrm{S}_{\mu-1} \xrightarrow{\mathrm{E}_\mu} \mathrm{S}_\mu$. The overall reaction pathway is then:

$$\mathrm{S}_0 \xrightarrow{\mathrm{E}_1} \mathrm{S}_1 \xrightarrow{\mathrm{E}_2} \mathrm{S}_2 \ldots \mathrm{S}_{\mathrm{L}-1} \xrightarrow{\mathrm{E}_\mathrm{L}} \mathrm{P} \tag{10.1}$$

where L denotes the number of catalysts, effectively corresponding to the length of the reaction pathway. It is assumed that these L catalysts have a mild attraction to each other (typically, through multivalent interactions) and therefore they can undergo phase-separation. For the sake of simplicity, the model and simulations reported below will be 2-dimensional, but without losing generality (see the discussion part). In fact, some metabolic processes, like the final steps in the production of lipids (e.g., the Kennedy pathway) take place on the cytoplasmic membrane (a 2D environment). Likewise, this could also be the case for prebiotic systems with simpler proto-cellular bilayers.

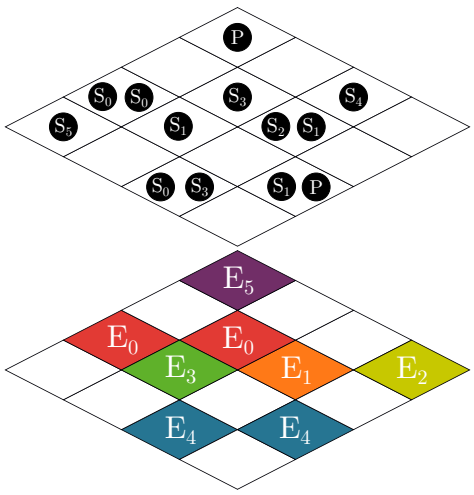
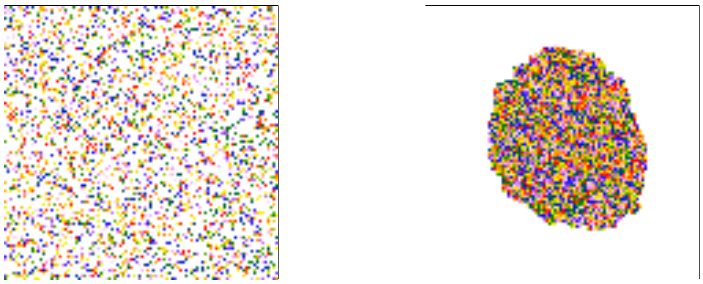


Figure 10.1: Overview of the implemented lattice model. Biomolecules E_i occupy single squares on the bottom grid (the catalyst-lattice), where different colours stand for different catalysts (white squares represent the solvent). Multiple substrates S_α can occupy a single square on the top grid (the substrate-lattice).

As depicted in Fig. 10.1, the LLPS of these catalysts is simulated using the general lattice model described in the previous chapter, with some

modifications. The system is discretized into a 2-dimensional, $d \times d$ regular lattice, where each site-i is in a state $\sigma_i = 0, 1 \ldots L$, representing the type of catalyst occupying it, with 0 representing the solvent. The volume or, rather, the surface fraction for each compound present on the lattice is given by ϕ_μ. All catalysts E_μ are assumed to be present in equal proportion: i.e. $\phi_\mu = \frac{\phi_{tot}}{L}$, $\forall \mu \neq 0$, with $\phi_{tot} = \sum_{\mu=1}^{L} \phi_\mu$ being the total surface fraction of catalysts on the lattice. The energy of a certain configuration of catalysts on the lattice is given by Eq. 9.4, with the interactions among the compounds captured by the symmetric matrix $\mathbf{J}$. For simplicity neither solvent-solvent nor solvent-catalyst interactions are taken into account: i.e., $J_{00} = J_{\mu 0} = 0$. As stated above, all interactions among the catalysts are assumed to be attractive and, for further simplicity, of the same strength (homogeneous attraction), i.e. $J_{\mu\nu} = J > 0$, $\forall \mu, \nu \neq 0$ (given the sign convention defined in Eq. 9.4). This general assumption will only be slightly modified below when discussing the problem of the homogeneity of the condensate.



(a) **(b)**

Figure 10.2: Typical equilibrium configurations from simulations with lattice size $d = 100$, pathway length $L = 20$, surface fraction $\phi_{tot} = 0.3$, interaction strength $J = 1.0$ and inverse temperature of: **(a)** $\beta = 0.01$ leading to a well-mixed (WM) system and **(b)** $\beta = 5.0$ leading to a phase separated (PS) system.

It is further assumed that there is a mild affinity between the substrates and the catalysts. Therefore, as further shown in Fig. 10.1, the diffusion of the substrates (including their interaction with the catalysts) is simulated using the reaction diffusion model described in the previous chapter, again with some modifications. The lattice containing the catalysts E_μ is extended through a layer containing the corresponding substrates S_α, with the reaction diffusion process modelled as a random walk on this substrate lattice. This random walk is biased due to the substrate-catalyst interaction, captured by matrix $\mathbf{I}$ (compare again Eq. 9.5). Possible interactions between

the substrates and the solvent will be disregarded: i.e. $I_0 = 0$. Only attractive interactions between the substrates and the catalysts will be considered and, again for further simplicity, it is assumed that this attraction is homogeneous, i.e. $I = I > 0$; $8 \quad 8 \quad 6 \quad 0$ (again given the sign convention de ned in Eq. 9.5). The relevant simulation parameters for both the LLPS simulation and the reaction di usion are collected in Tab. 10.1 (this is actually a combination of the parameters of Tab. 9.1 and Tab. 9.2, with some additional parameters, given the speci cities of the system under analysis here).

Parameter	Name	Function
d	System size	determines the size of the square lattice
L	Path length	determines the length of the reaction path (i.e. nr. of catalysts)
tot	Total surface fraction	determines the overall amount of catalysts
J	Catalyst-catalyst interaction	determines the strength of attraction among catalysts
C	Substrate concentration	determines the amount of S substrates on the lattice
I	Substrate catalyst interaction	determines the strength of attraction between substrates and catalysts
	Inverse temperature	determines the temperature of the system

Table 10.1: Collection of the overall model parameters.

In the following sections the main ndings of the investigation carried out are outlined, comparing the situation when the catalysts are well-mixed (WM-system, shown in Fig. 10.2a), which is, in general, a deeply rooted, classical assumption in biochemistry studies, and when they phase separate into a surface droplet (PS-system, shown in Fig. 10.2b).

10.2 In uence of LLPS on reaction time

One way to quantify the behaviour of the reaction pathway is to compute the mean reaction-time $\bar{t}_R$, which is the average time for an initial substrate S_0

(ii) a more comprehensive account of the spatial constraints involved (not only LLPS-related ones).

Part IV

Discussion & Conclusion

Chapter 11

Discussion

To recapitulate, the central idea of this thesis was to investigate how the formation of a system with a minimal metabolism (MM) could have been a major stepping stone on the transition from complex chemistries to simple biology, i.e. how metabolism, in its most elementary expression, might have played a central role in the origin of life (OL). In order to do this, first a tentative scheme was given in Ch. 3 that establishes MM as a functional coupling between complex chemical reaction networks (CRNs) an complex boundary conditions (BCs), where the former synthesise the latter, while at the same time the latter control and regulate the former. Then the aim was to model such a system. However, as it is rather challenging to capture the full complexity of MM, two complementary approaches were pursued: on the one hand, complex CRNs under simple BCs, and on the other hand simple CRNs under complex BCs were studied. Both approaches involved implementing computer models that yielded interesting results which are going to be further analysed, together with their limitations, in the following sections.

11.1 rTCA space exploration

As outlined in Ch. 5, the particular case study chosen to explore complex CRNs under simple BCs was the a pre-biotic carbon fixation scenario, supporting the hypothesis of an autotrophic OL. More precisely, the idea was to study the rTCA cycle using a rule based modelling approach (through the graph grammar formalism of the software package MØD) in order to

expand and investigate the chemical reaction space (CRS) of the chemistry associated with the rTCA. The overall goal was then to search for reaction pathways within this CRS which are functionally similar to the extant rTCA cycle, i.e. autocatalytic (AC) cycles that bind inorganic carbon molecules, but could be more plausible, prebiotically speaking. To this aim, several CRS were expanded using different reaction grammars, which yielded different results.

11.1.1 Analysis of results

Globally speaking, the results show that it is in fact possible to find "alternative" chemical reaction pathways that are similar to the rTCA cycle. However, the identified AC-cycles and their relationship with the rTCA, depend heavily on the graph grammar that is used to expand the CRS, i.e. on how the the reaction rules and constraints are implemented.

For instance, a rather strict reaction grammar yielded 138 AC pathways that are not just similar to the rTCA cycle, but they are basically "variations" of it. More precisely, they all involve the same reactions as the ones occurring within the rTCA cycle, although the order in which they happen varies. Therefore, all these cycles have the same type of topology (in principle they are type II AC-cycles, following the classification of Fig. 5.8), with the same number of reactions and the same focal species or cycle catalyst: oxaloacetate. Furthermore, all the cycles obtained in this way consume the same amount of energy (which was also observed across all environmental parameter settings). As it was illustrated in Fig. 7.8, all these pathways basically follow the extant rTCA but at some point take a different path, i.e. diverge from the "main cycle". Interestingly, when these different rTCA-alternatives are compared, several key molecules seem to appear in all of them as common reaction intermediaries: namely, succinate, acetate or the aforementioned oxaloacetate. This basically reflects the importance of these compounds in extant metabolism. In addition the question of why life "chose" the rTCA cycle naturally comes to ones mind: was it the most effective pathway, or was it pure chance? This apparent variability also suggests the possibility that the rTCA initially did not run as a single pathway, but as a collection of all these pathways operating concurrently. Only later, with the advent of specific enzymes favouring the catalysis of certain reaction steps over some others, did the extant pathway become eventually established.

Modifying and relaxing the reaction grammar a bit leads to a more diverse suite of rTCA-alternatives. Despite the large overlap (see Fig. 7.10), several of the AC-cycles that were found within the resulting CRS present different topologies, total number of reactions and focal species. The variation in the total amount of energy each cycle consumes is also more prominent (with further changes in the energy distributions observed upon changing the environmental parameters). Nevertheless, a majority of the obtained AC-cycles follow a similar pattern, where the chain of carbon atoms is grown through a consecutive application of the following reactions: an activator binding to the chain at the acid group, a reductive carboxylation of this the activated acid and finally a reduction of the resulting α-keto acid. The last step enables additional activator-bindings (as it is in general assumed that the activating compound cannot bind to an alpha-keto acid) and thus the pattern can repeat. In other words, the molecules within these AC-cycles are grown like putting beads on a string.

Moreover, the results show that a reaction grammar which contains condensation reactions (e.g. aldol-additions) but no activator dependent reactions can yield AC-cycles that, while not functionally equivalent to the rTCA, do not require any kind of activating compound in order to run. In that case, the growth of the carbon chains (which is necessary for an AC-cycle) is primarily achieved through these condensation reactions, rather than carboxylation reactions (even redox neutral carboxylations can still occur to some extent). Thus, the resulting reaction pathways are not necessarily carbon fixation pathways, in the sense that they bind small, inorganic carbon compounds (like CO_2). In fact, most of the molecules involved within the resulting AC-cycles (both those constituting the cycle and the ones incorporated from the outside) are relatively larger organic molecules, so they would need to come from some other sources.

In principle, this suggests two possible scenarios. In the first one autocatalysis and the direct binding of inorganic carbon was already feasible from the very beginning, due to the existence of some compound (like a simple thioester) serving as an activating species for certain reactions. In this way, large organic molecules could have been grown directly within AC-cycles (i.e. from the bottom up) and the rTCA cycle was probably just a very effective way of achieving this. The other possibility is that autocatalysis and direct carbon fixation was more of a late stage development,

with the formation of intermediate-size molecules occurring probably over a longer period of time in some non autocatalytic processes (e.g. some geological processes). These molecules would have begun, then, to constitute AC-cycles through condensation reactions, which were not carbon fixating yet but contributed to produce themselves in high copy numbers, increasing their overall concentrations. This process would have led to the build-up of some more complex molecules functioning as activator compounds for additional AC-cycles through the direct binding of CO_2, etc., leading to the emergence of the rTCA.

Regarding the global cycle energies involved, the results show that all activator dependent chemistries (strict and relaxed) are endergonic. Only under more acidic conditions (pH=4) and/or with a high reduction potential ($E_0 = -0.3$mV) would these cycles run exergonically. However, this is not so far off from the assumed conditions for pre-biotic carbon fixation, where a high CO_2 level in the atmosphere is proposed, which would probably lead to rather acidic conditions within the archaean oceans. Furthermore, the *reductive* TCA cycle, as the name suggests, involves mostly redox reactions that occur in the reductive direction. Thus it makes sense that variations of this pathway need a more reducing redox potential, in order to run. Nevertheless, for the activator independent chemistry one can find several AC-cycles that would also run spontaneously under neutral or basic conditions and/or more oxidative redox potential (as shown in Fig. 7.15 and Fig. 7.16).

11.1.2 Limitations

Despite the potential interest of these results (particularly to open new research lines in the future – see section 11.1.3 below), it is convenient to reflect on some limitations that they may present. For starters, the MØD software package that was used to do the CRS expansions, even if it is a powerful, rule-based tool, depends quite a lot on a priori choices, i.e.: the choice of a graph grammar (including the selection of the initial molecules and the reaction rules), or the choice of an expansion strategy (e.g., the choice of a size or a mass "cut-off", etc.). While these selections or constraints are intended to prevent a combinatorial explosion during the CRS expansion, they can also lead to the problem that one will only observe the molecules and reactions that were "chosen to be observed". In other words, the rules, initial conditions, etc. are formulated to obtain the desired results (like the

above discussed activator dependency, which one can argue to be, to a certain extent, explicitly written into the reaction rules). Furthermore, if the constraints are too strictly formulated, the risk is that no interesting results are obtained (in the most extreme case, no CRS can be expanded). Therefore, as with basically all types of modelling approaches, one must reach a compromise: on the one hand it is necessary to constrain the model to prevent things from becoming too complex, but at the same time one should be careful not to "suffocate" it, so it still shows some complex and therefore interesting behaviour[1].

A similar argument can be made for the search of autocatalytic (AC) cycles. Here the a priori choices that need to be made are the maximum cycle length as well as the set of molecules that are prohibited from being on-cycle compounds. Again, the main motivation behind these choices is to prevent a combinatorial explosion of potential AC-cycles, most of which might be trivial cycles (e.g. cycles that autocatalytically produce water), or cycles that involve too many reactions to be considered feasible (since they would be "destroyed" by parasitic reactions before the cycle closes). However, these constraints can again considerably bias the AC-cycle search, leading to solutions that are similar to a desired reaction path (for instance, when the maximum length of the cycle is limited to 11 reactions – which is the amount of reactions that the main branch of rTCA cycle actually has). As with the CRS expansion, a compromise must be reached here: if the AC-search is too constrained, no interesting cycles (or no cycles at all) will appear. At the same time, if the search constraints are set too loose, too many cycles will probably be found and the results will need to be post-processed in order to filter out unrealistic or uninteresting cycles.

Finally, there are also certain limitations concerning the calculations of the thermochemical properties of the AC-cycles. Although the eQuilibrator software is a good tool to get quick results for energies of formation, $\Delta_f G^\circ$, or energies of reactions, $\Delta_r G^\circ$, these are often only "rough" estimations. The group contribution method (which is the main method implemented in the eQuilibrator software) is based on a linear regression assumption, i.e. the idea that the energy of formation of a molecule is essentially the sum of the energies of the groups that constitute it. But this is not necessarily

[1] There is a common saying (often attributed to Albert Einstein) that any model should be as complex as necessary, but as simple as possible.

true, in particular for more complex molecules where functional groups can interact in non-linear ways (e.g. due to the de-localisation of electrons, etc.). This linear assumption may also lead to the problem that the same type of reaction, even if it takes place on molecules that are significantly different, could still lead to the same values of $\Delta_r G^\circ$. This is due to the fact that the group contribution method only accounts for the difference in the energy contribution of the functional groups that change during the reaction while, the remaining functional groups of the molecules remain unchanged[2]. Even if the eQuilibrator software tries to avoid this, to some extent, by using energy values from databases rather than decomposing a molecule, whenever those values are not available the problem is bound to persist. Furthermore, the decomposition into functional groups is not always obvious, which is again typically the case for more complex organic molecules (e.g., molecules that display different functional groups around their C-atoms). Last, the thermochemical properties also depend on the choice of parameters (pH, redox potential, etc.) which again could lead to a certain bias in the results.

11.1.3 Outlook

There is no doubt that the results of the modelling approach pursued in this part of the dissertation leave plenty of room for future projects. First of all, one could continue to explore other chemistries, in combination with other types of expansion strategies. The focus within this project was put on the rTCA cycle because, as argued in Ch. 5, it generally seems like a plausible candidate for prebiotic carbon fixation. Nevertheless, one could still continue to implement and explore the chemistries of the other autocatalytic carbon fixation pathways (e.g., the 3HB- or the 4HP-cycle, which were mentioned in the same chapter). However, the study of prebiotic carbon fixation should not necessarily be limited to autocatalytic cycles. While they do have some important properties (like the exponential increase of reaction products), one should also consider studying chemistries that could lead to the combination of autocatalytic cycles with other types of pathway topologies: e.g., linear pathways that feed into the cycle (a scenario like this was, for example, suggested by Eric Smith when speaking about the importance of coupling the

[2] An example would be the reduction of a carbonyl group that involves the change of a C=O group to a COH group while all the other functional groups that might exist in the reacting molecule remain unchanged.

Wood-Ljungdahl pathway with the rTCA cycle [Braakman and Smith 2012, 2013]). Additionally, the possibility of a precursor carbon fixation pathway that was significantly different from extant pathways should be further explored too. Although this was already tried here, to a certain extent, through the activator independent grammar, other options are available.

Another possibility for future projects could also be the exploration of different activator compounds and how the thermochemical properties might change. Within this project the focus was put on thioesters, because they are plausible prebiotic compounds and their presence as coenzymes in extant biochemistry indicates their potential as precursor energy currencies [de Duve 1991]. Since energy calculations were carried out via group the contribution method, only one type of thioester (i.e. methanethiol) was actually studied (other thioesters would have led to similar or even the same results). Nevertheless, one could still look into activator compounds based on different types of molecules, like phosphorous or nitrogen, etc.

Moreover, in combination with the exploration of other activator compounds, one could also try to do more precise energy calculations. This should include a denser variation of environmental parameters (i.e. the pH and the redox potential E_0) together with studying how the viability of the AC-cycles identified through the search might change for different types of activators under different conditions (in line with the approach of [Goldford et al. 2019], for instance). The implementation of more relaxed grammars should also be helpful, in parallel with this (e.g., one could explore the question: would a grammar that allows for more oxidation reactions also lead to additional AC-cycles that are viable under a more oxidative redox potential?).

11.2 Lattice model

As outlined in Ch. 8, the complementary "facilitated approach" to investigate the emergence of a minimal metabolism was the study of relatively simple CRNs under more complex BCs. Among the different types of constraints (i.e., control mechanisms operating on basic transformation processes) that could be chosen, the attention was turned to proto-cellular compartments capable of housing reactions associated with a proto-metabolism. More precisely, the idea was to investigate the formation of heterogeneous droplets

through liquid-liquid phase separation (LLPS) and their potential influence on a linear reaction pathway, testing in this way the potential prebiotic relevance of membrane-less comparments (MLC). In order to analyse such a linear pathway of catalysed reactions, where the catalysts of the individual reaction steps can undergo LLPS, a lattice model simulation was implemented. The aim was to study how the behaviour of the reaction pathway is affected when the catalysts phase separate into a droplet (PS-system) or when they stay well-mixed (WM-system).

11.2.1 Analysis of results

In general, the outcome of the simulations and modelling work performed shows that the phase separation of the catalysts into a droplet can have a positive effect on the reaction pathway. This is in particular reflected as a reduction of the time it takes a precursor substrate $S_{\bar{0}}$ to complete the full reaction pathway (i.e. to be converted to the final product P): in other words, the phase separation of the catalysts can "speed up" the overall reaction time t_R[3]. However, this effect depends on several factors. First of all, there needs to be a minimal chemical affinity between the substrates and the catalysts. In this way the substrates, once once encapsulated within the droplet, are prevented from leaving: i.e., their affinity to the catalyst, even if this is relatively mild, makes it energetically more favourable that they stay inside the droplet (i.e. the catalyst droplet effectively "entraps" the substrates). One can describe the situation as if the LLPS of the catalysts provides a "micro environment" of higher catalyst concentration for the substrates to proceed the chemical transformation.

Second, the reaction cascade needs to be sufficiently long. This is due to the fact that it takes some time for the initial substrate $S_{\bar{0}}$ to reach the catalyst droplet, which causes a delay of the first reaction step. This "hitting time" t_H needs to be compensated. As a result, there is a critical path length L_C above which the phase separation of the catalysts leads to a decrease in t_R compared to a WM-system (i.e. the formation of a catalyst droplet "pays off"), which is perfectly illustrated by the crossover point of the two parabolas in Fig. 10.4. However, this critical length is not scale invariant. The critical length depends on the overall size of the system,

[3] In practice, t_R is measured as the time it takes substrate $S_{\bar{0}}$ encounter all the catalysts E_μ in the corresponding order of the reaction pathway.

where a larger system leads to a larger value of L_C and vice versa (assuming a fixed volume/surface fraction of catalysts in the system – see Fig. 10.4). That is to say, the LLPS of the catalysts tends to be more beneficial for the reaction in a smaller system. This means that, even for a reaction cascade with only a few steps, if the system size is kept small (without increasing the volume/surface fraction of the catalysts) phase separation can still have an effect of speeding up the reaction time.

Last but not least, the degree of homogeneity of the catalyst droplet also plays a role: if there are patches of similar catalysts inside the condensate counter effects or, perhaps, further scale effects may occur. This is because a substrate might "get stuck" within a patch of catalysts, taking some extra time to diffuse to the next patch. As shown in Fig. 10.5a, if the attraction among the catalysts varies, allowing for the formation of a patchy droplet, one can observe an optimal temperature at which the reaction time reaches a minimum. Below that critical temperature t_R increases again. Therefore, the temperature should be low enough for the catalysts, in general, to start phase separating into a droplet, but at the same time not too low, to prevent the formation of patches of specific catalysts (depending on the strength of their corresponding attractive interactions) within the droplet. When the temperature is further decreased, those patches or inhomogeneities start to form and become a source of time delays for the chemical pathway to be completed.

These main results were all obtained by averaging the reaction times of a single substrate, diffusing on the lattice, but further simulations showed that the concurrent reaction diffusion of multiple substrate particles across the lattice keeps in good agreement with the predictions from a mean-field mass action kinetics model (compare Fig. 10.6 and Fig. 10.7). The same approach can be extended to simulate an open system (i.e., a system including inflow and outflow of initial substrates and products, respectively) with a mechanism of feedback inhibition where products can hinder the inflow of substrates (as it is often the case in enzymatic reactions). In such a system oscillations in the concentration of the substrate species are expected [Hunding 1974] as a result of the interplay between in-/out-flow of products and the feedback inhibition mechanism. However, this only happens in a WM-system. When phase separation comes to play it perturbs the oscillatory behaviour, even suppressing it, as confirmed by the microscopic simulations

(see Fig. 10.8). The absence of this type of behaviour in most metabolic pathways (despite the presence of the feedback mechanism) could, therefore, be explained in terms of the dynamics of biomolecular condensates.

In summary, the results of this modelling approach show that LLPS phenomena can have a beneficial effect on a simple, linear network of chemical reactions. From this interesting finding, and taking into consideration that aggregative processes of this kind could have taken place under many different conditions (and involving a wide range of organic compounds) on the early Earth, one can move a step further and suggest the hypothesis that multicomponent MLCs are plausible proto-cellular system. This is not a totally novel idea, but most other theories that propose MLCs as proto-cell models (i.e. theories of coacervates) only involve droplets of one or two components. Although sooner or later a more dynamic and multifunctional spatial boundary (probably in the form of a cellular membrane) will be necessary, it is still fair to assume that MLCs could have played a crucial role in that transition. In fact, as it has been already suggested by other research groups, the formation of MLCs through LLPS in a heterogeneous mixture, together with the self organization of amphiphiles into vesicles are not incompatible. They could have occurred concurrently, with the first proto-cells developing through the enclosure of such a droplet inside a large unilamellar vesicle (an idea that has been experimentally explored, too [Tang et al. 2014; Love et al. 2020]).

11.2.2 Limitations

Like with any other modelling approach, the way of accounting, in this thesis work, for the interplay between spatial constraints (linked to phase separation phenomena) and chemical transformations presents a set of limitations. Despite its advantages over classical mean-field approximations to the problem, the lattice model implemented here is still a course grained model. As such, the molecules within the system, which are generally considered to be polymer-like macromolecules that are often not symmetric and have different local charge densities, are strongly simplified, reducing them to occupied lattice sites with fixed interaction parameters. Therefore, the model cannot simulate how individual molecules may take different spatial configurations or reorient themselves (due to rotations etc.), which might be important to capture their dynamic behaviour (e.g., potential changes in molecule-

molecule interactions due to the asymmetric distribution of charges, or similar molecular specificities that could be relevant). Furthermore, space is modelled as a regular grid, so all movements of the molecules in the system can only happen in integer steps within the neighbourhood of a lattice site, i.e. the diffusive dynamics is also strongly constricted.

Several additional assumptions and simplification about the properties of the molecules within the system, as well as about their interactions and their reaction dynamics, were made. For instance, the reactions within the pathway were assumed to be irreversible and occurring immediately, without any delay, once a substrate encounters the corresponding catalyst (i.e., perfect catalysis). Besides, all interactions, catalyst-catalyst and substrate-catalyst, were taken to be attractive and symmetric. On top of that, no interactions with the solvent were considered, neither for the catalysts nor for the substrates. Lastly, even though the homogeneity of the attraction between the catalysts was eventually removed, the attraction between substrates and catalysts remained constant for all simulations reported. In any case, the motivation behind these simplifications was again to avoid implementing a system that becomes too complex to handle while still enough complexity is retained to observe interesting and meaningful results.

11.2.3 Outlook

The results obtained with the lattice model also open up new avenues for future research projects. First and foremost, the actual model implementation used a 2-dimensional lattice, without loss of generality in the main results obtained. In a first approximation, the focus was to study phase separation occurring on organic surfaces, like lipid membranes, or involving organic compounds aggregating in bidimensional configurations. Nevertheless, one could expand the results and generalize the model to three dimensions. This would allow to directly model LLPS phenomena taking place in an open solution environment, including the formation of spherical droplets. In any case, one should keep in mind that such a transition from a 2D- to a 3D-system will increase the amount of computational resources requested to run simulations. Furthermore, a diffusion process modelled as a random walk in three dimensions does not cover, in general, the whole space (while this is

guaranteed in two dimensions) [Pólya 1921][4].

Another aspect that can be expanded and much more carefully explored in future research work has to do with non-equilibrium conditions and the openness that the last version of the model encompassed. Situations where substrates and reaction products can freely diffuse in and out of the system allow to try other types of mechanisms beyond the feedback inhibition case that was implemented here. For instance, feed-forward loops, promotion instead of inhibition, or extending these mechanisms again to other substrates of the reaction pathway, or combinations of both, with various other substrates of the pathway getting involved. Alternatively, the product of the reaction pathway could be a compound that becomes part of the actual environment where the reactions take place (e.g., a lipid, like in the Kennedy pathway), changing its properties (like its fluidity or permeability). All these options would typically lead to a richer reaction diffusion dynamics that can be studied under different conditions in which phase separation can occur.

Additionally, the type of reaction network to be studied can be modified, of course, beyond the linear reaction cascade that was implemented here. Although the overall size of the network is limited to some extent (i.e. L catalysts imply a reaction pathway with L reactions), one could nevertheless study different types of pathway topologies. A particular case of interest could be a linear reaction cascade (similar to the one that was originally studied) that involves a forking point, where it splits into two parallel pathways. This new situation could then be combined with a modified catalyst interaction matrix $\mathbf{J}$, where only the catalysts related to one branch would attract each other leading, to a phase separation and the formation of a droplet. The idea here would be to analyse whether the LLPS of the catalysts of this branch and, in turn, the expected speeding up of the corresponding reactions can enhance the flux of material through it and effectively take away material from the other branch. Another option would be to investigate cyclic topologies (autocatalytic or not). The kinetic effect of speeding up the reaction dynamics due to the formation of a catalyst droplet, as it was demonstrated here, could be especially interesting when AC-cycles are under examination. In that case the exponential growth of the reaction product could be increased even further (which would be of particular interest, for

[4] This was also summarized in a more comical way by the mathematician Shizuo Kakutani: "A drunk man will find his way home, but a drunk bird may get lost forever".

example, in a prebiotic carbon fixation setting, see Ch. 5). Along these lines, it would also be interesting to investigate whether it is possible to create an autocatalytic cycle from two initially linear pathways, where one occurs outside a condensate while the other occurs inside, by coupling them through a diffusion process across the interface of the droplet (as a concrete example of the general mechanism suggested in [Blokhuis et al. 2020]).

Last but not least, one could also simply revert some of the simplifications introduced in the lattice model and study how this might influence the initially observed behaviour (in particular, the effects on the reaction time t_R). This could include: inhomogeneous substrate-catalyst interactions or interactions among the different substrates; interactions of the solvent with the catalysts and/or the substrates; or, more widely, asymmetric interactions or even repulsion (for either the substrates or the catalysts).

11.3 General outlook & perspective

The main idea behind all the modelling and simulation work carried out in this PhD was to investigate a system where a chemical reaction network produces one or various boundary conditions (more precisely, 'constraints') that take control of the reaction network itself, given that this functional coupling between synthesis and control is a central part of the proposed scheme for minimal metabolism. However, in the end only the latter (control) aspect of this coupling was implemented: the expansion of chemical reaction spaces was regulated by an activator compound that was externally supplied, while within the lattice model simulation the reaction was simulated in a system that was already at equilibrium and not endogenously synthesized, either. A difficulty underlying this choice has to do with the fact that the transformation processes of chemical reactions as well as the formation of boundary conditions (like the phase separation of droplets) do not usually occur on the same time scales and, therefore, they are not easy to implement in a computer model. Accordingly, those two aspects of the problem were either decoupled (like in the lattice model simulation), or only the control dynamics of a boundary condition/constraint was studied (like in the CRS expansion).

In consequence, future investigations should aim at implementing a proper coupling between a chemical reaction network and a set of boundary con-

ditions of its own making, i.e. at properly modelling the coupling between synthesis and control. On these lines, the expansion of chemical reaction spaces, for example, could be modified in such a way that the activator compound is a direct product of the reaction rules within the graph grammar (instead of assuming that it is available a priori – e.g., being externally supplied). In this way one could simulate a CRS expansion that is initially limited, but once the activator gets produced (or "discovered" by the expansion algorithm) it retroactively enables recombinations of molecules that were previously not possible, thus leading to a richer network expansion, i.e. the activator compound directly acts as an enabling constraint [Ruiz-Mirazo 2020] that gets produced from within the system. The lattice model, in turn, could be improved by considering liquid-liquid phase separation of the catalysts as a process that happens concurrently with the reaction diffusion dynamics that they constrain, i.e. trying to model the unfolding of both processes on the same time scale. In this way one could, for example, model a system where the formation of a condensate (or its disaggregation) influences a simple reaction pathway, while at the same time, the products of this pathway are the building blocks that form the condensate[5]. This could be combined with the opening of the system to the inflow and outflow of substrates and products, as well as with the implementation of different mechanisms of feedback inhibition.

The threshold of complexity that must be reached (with regard to the underlying transformation processes, or to the endogenously produced constraints – or to both since they must be deeply correlated) for observing the emergence of a minimal metabolism is a big open question that cannot be solved just with theoretical work. The decomposition of the problem that has been proposed here is rather tentative, and should be taken in relative terms. One could simplify further and still try to capture some sort of 'functional bootstrapping' between synthesis and control. However, there is a risk of oversimplification in this context. Although the strategy should remain 'bottom-up' (from physics and chemistry towards biology), getting started from slightly higher levels of complexity could pay back. A possible way to achieve this would be a continuation of what has been here: i.e., maintaining

[5]To some extend this has been explored in the form of dissipative self-assembly, both theoretically [Das et al. 2021] and experimentally [van Rossum et al. 2017; Donau et al. 2020].

a relatively complex CRN under simple BCs (or the other way around), then include a proper coupling between synthesis and control, as suggested above, and proceed to gradually increase the complexity of the part of the system that was initially kept simple.

In any case, these theoretical approximations, including the potential refinements to the work presented here, should definitely be combined with attempts to materialize minimal metabolic systems in vitro. Trying all these ideas in a test tube (a 'one-pot' reactor where all these different components and interactions must be realized) will surely modify the assumptions and conceptual premises from which this work started. However, as it was discussed in Ch. 4, there are several methodological challenges that need to be overcome for such a task: capturing the non-equilibrium nature of MM, dealing with the compositional diversity of the mixture and the heterogeneity of the reaction medium, etc. Some steps in that direction have already been made in recent years. For instance, in [Muchowska et al. 2017, 2019] or [Springsteen et al. 2018; Stubbs et al. 2020] reactions that are associated with the TCA/rTCA cycle or plausible precursor pathways were run without enzymes; or in [Ralser 2018] a proper screening of the reactions of the TCA cycle under different possible prebiotic conditions was made. However, even if these works did lead to interesting results, they also involved strong simplifications: like, observing the system at equilibrium (or as it is relaxing towards it); conducting reactions in multiple independent steps, rather than doing a one-pot type of chemistry; avoiding rich organic interfaces or lipid suspensions in aqueous solution; etc.

In fact, as argued in Ch. 4, closer collaboration between in silico and in vitro approaches would be highly beneficial, where experimental strategies are developed hand-in-hand with theoretical, computational models and vice versa. This workflow (which was visualized in Fig. 4.1) should promote rich and fruitful exchanges between different groups of investigators, coming from diverse backgrounds, sharing ideas and projected intuitions, but also pushing each other out of their respective comfort zones. Some research groups can even afford to do that by themselves, with remarkable results (e.g.: [Wołos et al. 2020]). Transdisciplinary collaborations should be pursued to investigate, as well, how minimal metabolisms could evolve towards genetically instructed metabolisms as they are realized in living organisms. This would require a better understanding of the origin of regulation and in-

formation mechanisms (i.e., constraints operating on other constraints – i.e., higher-order controls) in biological systems, which would contribute, in turn, to elaborate a more encompassing account of the full process of abiogenesis. For such a challenging task insights coming from 'protocell physiology' and 'protocell evolution' should be adequately integrated. In the words of Markus Ralser: "understanding the origins of metabolism is not only important to gain insights into how the first biomolecules formed but also to understand how cells obtained the properties which shape cellular evolution, in which conditions of life could persist, and how biological systems function and develop" [Ralser 2018].

Chapter 12

Conclusion

- A prebiotic systems chemistry approach to the problem of the origin of life (OL) is much more promising than classical approximations, which focus on single types of molecules and their synthesis under prebiotic conditions. However, this approach requires that material and thermodynamic aspects involved in the system construction (i.e., metabolic aspects) are adequately addressed.

- Extant living organisms can be viewed as a 'genetically-instructed metabolisms' (GIMs) where a deep functional coupling between complex transformation processes and multiple, hierarchically organized layers of boundary conditions and material constraints is realized. As such, these boundary conditions are the result of those transformation processes and, at the same time, control them.

- From that conception, 'minimal metabolism' (MM) is proposed as a tentative and relatively simplified theoretical scheme to approach the emergence of living cells. The core idea behind it is that a metabolism not only involves a complex network of physical and chemical transformations, but also a set of boundary conditions/constraints, where the former synthesizes the latter while, at the same time, the latter controls and regulates the former. Furthermore, this functional coupling between the material components and dynamics belonging to those two levels cannot be reduced to a one-to-one relation because both must be intrinsically diverse.

- The proposed scheme also aims at contributing to the development of a

more systems theoretical research program about the OL by providing a set of criteria through which the most elementary form of metabolism can be distinguished from a chemical reaction network.

- Two strategies were pursued in order to model and computationally simulate MM, trying to establish what could be that minimal diversity in the processes/constraints involved. The first focused on implementing a relatively complex reaction network under simple boundary conditions, while the second focused on implementing a simple reaction network under relatively complex boundary conditions. Both yielded interesting results that may foster future experimental research.

- The results of the first approach showed that potential pathways for prebiotic carbon fixation depend on the existence of a certain activator compound, which enables certain key reaction processes, and therefore should be the target of endogenous synthesis by those pathways.

- The results of the second approach showed how the formation of a heterogeneous condensate of catalysts (which results in a new spatial arrangement of them) can, under certain circumstances, speed up the time required to finish a simple cascade reaction, therefore changing the traditional principle of 'mass action kinetics' (which is only applicable to well-mixed systems).

- Future research of minimal metabolisms should go beyond the simplifying assumptions made in this PhD, which mainly focused on control aspects, and try to address the functional coupling between synthesis and control mechanisms. For that task, nevertheless, a stronger interbreeding between in silico methods (the main contribution here) and in vitro approaches should be pursued. Ultimately, this transdisciplinary effort should contribute to a better understanding of the transition from physics and chemistry towards "proto-biology" (through MMs) and, later on, when evolution comes to stage, towards full-fledged biology (i.e., GIMs).

Bibliography

R. A. Alberty. *Thermodynamics of biochemical reactions*. John Wiley & Sons, Hoboken (NJ), 2003.

A. Amilburu, Á. Moreno, and K. Ruiz-Mirazo. Definitions of life as epistemic tools that reflect and foster the advance of biological knowledge. *Synthese*, 2020. doi: 10.1007/s11229-020-02736-7. (Published online).

J. L. Andersen. *Analysis of Generative Chemistries*. PhD thesis, Universität Leipzig, 2015.

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Generic strategies for chemical space exploration. *Int. J. Comput. Biol. Drug Des.*, 7(2/3): 225–258, 2014. doi: 10.1504/IJCBDD.2014.061649. URL `https://doi.org/10.1504/IJCBDD.2014.061649`.

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. A software package for chemically inspired graph transformation. In R. Echahed and M. Minas, editors, *Graph Transformation: 9th International Conference, ICGT 2016, in Memory of Hartmut Ehrig, Held as Part of STAF 2016, Vienna, Austria, July 5-6, 2016, Proceedings*, pages 73–88, Cham, 2016. Springer International Publishing.

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Defining Autocatalysis in Chemical Reaction Networks. *arXiv:2107.03086*, 2021. doi: 10.48550/arxiv.2107.03086.

P. W. Anderson. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393.

A. Arya, J. Ray, S. Sharma, R. C. Simbron, A. Lozano, H. B. Smith, J. L. Andersen, H. Chen, M. Meringer, and H. J. Cleaves II. An open source computational workflow for the discovery of autocatalytic networks in abiotic reactions. *Chem. Sci.*, 13(17):4838–4853, 2022. doi: 10.1039/D2SC00256F.

G. Ashkenasy, T. M. Hermans, S. Otto, and A. F. Taylor. Systems chemistry. *Chem. Soc. Rev.*, 46(9):2543–2554, 2017. doi: 10.1039/C7CS00117G.

S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, 18(5):285–298, 2017. doi: 10.1038/nrm.2017.7.

A. D. Bangham and R. W. Horne. Negative staining of phospholipids and their structural modification by surface-active agents as observed in the electron microscope. *J. Mol. Biol.*, 8(5):660–668, 1964. doi: 10.1016/S0022-2836(64)80115-7.

A. Bar-Even, E. Noor, N. E. Lewis, and R. Milo. Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. U.S.A.*, 107 (19):8889–8894, 2010. doi: 10.1073/pnas.0907176107.

M. E. Beber, M. G. Gollub, D. Mozaffari, K. M. Shebek, A. I. Flamholz, R. Milo, and E. Noor. eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Res.*, 50(D1):D603–D609, 2022. doi: 10.1093/nar/gkab1106.

I. A. Berg, D. Kockelkorn, W. H. Ramos-Vera, R. F. Say, J. Zarzycki, M. Hügler, B. E. Alber, and G. Fuchs. Autotrophic carbon fixation in archaea. *Nat. Rev. Microbiol.*, 8(6):447–460, 2010. doi: 10.1038/nrmicro2365.

H. Berry. Monte carlo simulations of enzyme reactions in two dimensions: fractal kinetics and spatial segregation. *Biophys. J.*, 83(4):1891–1901, 2002. doi: 10.1016/S0006-3495(02)73953-2.

A. J. Bissette and S. P. Fletcher. Mechanisms of Autocatalysis. *Angew. Chem. Int. Ed.*, 52(49):12800–12826, 2013. doi: 10.1002/anie.201303822.

A. Blokhuis, D. Lacoste, and P. Nghe. Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. U.S.A*, 117(41):25230–25236, 2020. doi: 10.1073/pnas.2013527117.

S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. Van Den Bosch, P. Tompa, and M. Fuxreiter. Protein phase separation: a new phase in cell biology. *Trends Cell Biol.*, 28(6):420–435, 2018. doi: 10.1016/j.tcb.2018.02.004.

R. Braakman and E. Smith. The emergence and early evolution of biological carbon-fixation. *PLOS Comput. Biol.*, 8(4):e1002455, 2012. doi: 10.1371/journal.pcbi.1002455.

R. Braakman and E. Smith. The compositional and evolutionary logic of metabolism. *Phys. Biol.*, 10(1):011001, 2013. doi: 10.1088/1478-3975/10/1/011001.

M. Breuer, T. M. Earnest, C. Merryman, K. S. Wise, L. Sun, M. R. Lynott, C. A. Hutchison, H. O. Smith, J. D. Lapek, D. J. Gonzalez, V. de Crécy-Lagard, D. Haas, A. D. Hanson, P. Labhsetwar, J. I. Glass, and Z. Luthey-Schulten. Essential metabolism for a minimal cell. *eLife*, 8:e36842, 2019. doi: 10.7554/eLife.36842.

B. B. Buchanan and D. I. Arnon. A reverse KREBS cycle in photosynthesis: consensus at last. *Photosynth. Res.*, 24:47–53, 1990. doi: 10.1007/bf00032643.

H. G. Bungenberg de Jong and H. R. Kruyt. Coacervation (Partial miscibility in colloid systems). *Proc. K. Ned. Akad. Wet.*, 32:849–856, 1929.

I. A. Chen and P. Walde. From self-assembled vesicles to protocells. *Cold Spring Harb. Perspect. Biol.*, 2(7):a002170, 2010. doi: 10.1101/cshperspect.a002170.

A. J. Coggins and M. W. Powner. Prebiotic synthesis of phosphoenol pyruvate by α-phosphorylation-controlled triose glycolysis. *Nat. Chem.*, 9(4):310–317, 2017. doi: 10.1038/nchem.2624.

A. Cornish-Bowden and M. L. Cárdenas. Contrasting theories of life: Historical context, current theories. in search of an ideal theory. *BioSystems*, 188:104063, 2020. doi: 10.1016/j.biosystems.2019.104063.

M. Csete and J. Doyle. Bow ties, metabolism and disease. *Trends Biotechnol.*, 22(9):446–450, 2004. doi: 10.1016/j.tibtech.2004.07.007.

C. R. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

C. R. Darwin. Darwin correspondence project: Letter no. 7471. `https://www.darwinproject.ac.uk/letter/DCP-LETT-7471.xml`, 1871.

K. Das, L. Gabrielli, and L. J. Prins. Chemically-fueled self-assembly in biology and chemistry. *Angew. Chem. Int. Ed.*, 60(37):20120–20143, 2021. doi: 10.1002/anie.202100274.

C. de Duve. *Blueprint for a cell: The nature and origin of life*. N. Patterson, Burlington (NC), 1991.

A. de la Escosura, C. Briones, and K. Ruiz-Mirazo. The systems perspective at the crossroads between chemistry and biology. *J. Theor. Biol.*, 381: 11–22, 2015. doi: 10.1016/j.jtbi.2015.04.036.

D. W. Deamer. Boundary structures are formed by organic components of the murchison carbonaceous chondrite. *Nature*, 317(6040):792–794, 1985. doi: 10.1038/317792a0.

D. W. Deamer. Role of amphiphilic compounds in the evolution of membrane structure on the early earth. *Orig. Life Evol. Biosph.*, 17(1):3–25, 1986. doi: 10.1007/BF01809809.

D. W. Deamer. Membrane compartments in prebiotic evolution. In A. Brack, editor, *The Molecular Origins of Life: Assembling Pieces of the Puzzle*, pages 189–205. Cambridge University Press, Cambridge, 1998.

D. W. Deamer and J. Oró. Role of lipids in prebiotic structures. *BioSystems*, 12(3-4):167–175, 1980. doi: 10.1016/0303-2647(80)90014-3.

G. L. Dignon, R. B. Best, and J. Mittal. Biomolecular phase separation: from molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.*, 71:53–75, 2020. doi: 10.1146/annurev-physchem-071819-113553.

M. Doi. *Soft Matter Physics*. Oxford University Press, Oxford, 2013. doi: 10.1093/acprof:oso/9780199652952.001.0001.

C. Donau, F. Späth, M. Sosson, B. A. K. Kriebisch, F. Schnitter, M. Tena-Solsona, H.-S. Kang, E. Salibi, M. Sattler, H. Mutschler, and J. Boekhoven. Active coacervate droplets as a model for membraneless organelles and protocells. *Nat. Commun.*, 11(1):1–10, 2020. doi: 10.1038/s41467-020-18815-9.

F. J. Dyson. *Origins of life*. Cambridge University Press, Cambridge, 1985.

M. Eigen and P. Schuster. *The Hypercycle: A Principle of Natural Self-Organization*. Springer, Berlin, Heidelberg, 1979.

eQuilibrator. `https://equilibrator.weizmann.ac.il`.

eQuilibrator Assets. `https://gitlab.com/equilibrator/equilibrator-assets`.

T. Erb. Obituary for Arren Bar-Even. `https://www.mpimp-golm.mpg.de/2520342/obituary-arren-bar-even?page=2`, 2020.

M. C. Evans, B. B. Buchanan, and D. I. Arnon. A new ferredoxin-dependent carbon reduction cycle in a photosynthetic bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, 55(4):928–934, 1966. doi: 10.1073/pnas.55.4.928.

C. M. Fare, A. Villani, L. E. Drake, and J. Shorter. Higher-order organization of biomolecular condensates. *Open Biol.*, 11(6):210137, 2021. doi: 10.1098/rsob.210137.

J. D. Farmer, S. A. Kauffman, and N. Packard. Autocatalytic replication of polymers. *Physica D*, 22(1-3):50–67, 1986. doi: 10.1016/0167-2789(86)90233-2.

A. I. Flamholz, E. Noor, A. Bar-Even, and R. Milo. eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res.*, 40(D1):D770–D775, 2012. doi: 10.1093/nar/gkr874.

P. J. Flory. Thermodynamics of High Polymer Solutions. *J. Chem. Phys.*, 10(1):51–61, 1942. doi: 10.1063/1.1723621. URL `https://doi.org/10.1063/1.1723621`.

W. Fontana and L. W. Buss. The arrival of the fittest: Toward a theory of biological organization. *Bull. Math. Biol.*, 56(1):1–64, 1994a. doi: 10.1007/BF02458289.

W. Fontana and L. W. Buss. What would be conserved if "the tape were played twice"? *Proc. Natl. Acad. Sci. U.S.A.*, 91(2):757–761, 1994b. doi: 10.1073/pnas.91.2.757.

W. Fontana and L. W. Buss. The barrier of objects: From dynamical systems to bounded organizations. In J. Casti and A. Karlqvist, editors, *Boundaries and Barriers*, pages 56–116. Addison-Wesley, Reading MA, 1996.

U. Frisch, B. Hasslacher, and Y. Pomeau. Lattice-Gas Automata for the Navier-Stokes Equation. *Phys. Rev. Lett.*, 56:1505–1508, 1986. doi: 10.1103/PhysRevLett.56.1505.

G. Fuchs. CO_2 fixation in acetogenic bacteria: variations on a theme. *FEMS Microbiol. Rev.*, 2(3):181–213, 1986. doi: 10.1111/j.1574-6968.1986.tb01859.x.

G. Fuchs. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.*, 65:631–658, 2011. doi: 10.1146/annurev-micro-090110-102801.

C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, Berlin, fourth edition, 2009.

S. C. Glotzer, D. Stauffer, and N. Jan. Monte carlo simulations of phase separation in chemically reactive binary mixtures. *Phys. Rev. Lett.*, 72 (26):4109, 1994. doi: 10.1103/PhysRevLett.72.4109.

J. E. Goldford and D. Segrè. Modern views of ancient metabolic networks. *Curr. Opin. Syst. Biol.*, 8:117–124, 2018. doi: 10.1016/j.coisb.2018.01.004.

J. E. Goldford, H. Hartman, R. Marsland III, and D. Segrè. Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. *Nat. Ecol. Evol.*, 3(12):1715–1724, 2019. doi: 10.1038/s41559-019-1018-8.

J. E. Goldford, H. B. Smith, L. M. Longo, B. A. Wing, and S. E. McGlynn. Continuity between ancient geochemistry and modern metabolism enabled by non-autocatalytic purine biosynthesis. *bioRxiv*, pages 2022–10, 2022. doi: 10.1101/2022.10.07.511356.

F. Graner and J. A. Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Phys. Rev. Lett.*, 69:2013–2016, 1992. doi: 10.1103/PhysRevLett.69.2013.

C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983. doi: 10.1016/0092-8674(83)90117-4.

T. Gánti. *Az élet princípiuma*. Gondolat, Budapest, 1971. Repr. in transl.: *The Principles of Life*. Oxford University Press, Oxford, 2003.

T. Gánti. Organization of chemical reactions into dividing and metabolizing units: The Chemotons. *BioSystems*, 7(1):15–21, 1975. doi: 10.1016/0303-2647(75)90038-6.

T. Gánti. On the early evolutionary origin of biological periodicity. *Cell Biol. Int.*, 26(8):729–735, 2002. doi: 10.1006/cbir.2000.0668.

J. B. S. Haldane. The origin of life. *Rationalist Annual*, 148:3–10, 1929.

W. R. Hargreaves, S. J. Mulvihill, and D. W. Deamer. Synthesis of phospholipids and membranes in prebiotic conditions. *Nature*, 266(5597):78–80, 1977. doi: 10.1038/266078a0.

T. S. Harmon, A. S. Holehouse, and R. V. Pappu. To mix, or to demix, that is the question. *Biophys. J.*, 112(4):565, 2017. doi: 10.1016/j.bpj.2016.12.031.

F. M. Harold. Biochemical topology: from vectorial metabolism to morphogenesis. *Biosci. Rep.*, 11(6):347–385, 1991. doi: 10.1007/BF01130213.

F. M. Harold. *The Way of the Cell: Molecules, Organisms, and the Order of Life*. Oxford University Press, Oxford, 2001. ISBN 9780198030881.

L. A. Harris, J. S. Hogg, J.-J. Tapia, J. A. P. Sekar, S. Gupta, I. Korsunsky, A. Arora, D. Barua, R. P. Sheehan, and J. R. Faeder. BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics*, 32(21):3366–3368, 2016. doi: 10.1093/bioinformatics/btw469.

R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Springer US, New York (NY), 2012.

M. Himsolt. GML: A portable graph file format. Technical report, Universität Passau, 1997.

P. Honegger. Personal communication (publication forthcomming), 2023.

W. Hordijk and M. Steel. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.*, 227(4):451–461, 2004. doi: 10.1016/j.jtbi.2003.11.020.

W. Hordijk, S. A. Kauffman, and M. Steel. Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. *Int. J. Mol. Sci.*, 12(5):3085–3101, 2011. doi: 10.3390/ijms12053085.

N. H. Horowitz. On the evolution of biochemical syntheses. *Proc. Natl. Acad. Sci. U.S.A.*, 31(6):153–157, 1945. doi: 10.1073/pnas.31.6.153.

N. H. Horowitz. The evolution of biochemical syntheses–retrospect and prospect. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 15–23. Academic Press, New York (NY), 1965.

H. Huber, M. Gallenberger, U. Jahn, E. Eylert, I. A. Berg, D. Kockelkorn, W. Eisenreich, and G. Fuchs. A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation cycle in the hyperthermophilic archaeum ignicoccus hospitalis. *Proc. Natl. Acad. Sci. U.S.A.*, 105(22):7851–7856, 2008. doi: 10.1073/pnas.0801043105.

M. L. Huggins. Some Properties of Solutions of Long-chain Compounds. *J. Phys. Chem.*, 46(1):151–158, 1942. doi: 10.1021/j150415a018. URL `https://doi.org/10.1021/j150415a018`.

A. Hunding. Limit-cycles in enzyme-systems with nonlinear negative feedback. *Biophys. Struct. Mechanism*, 1(1):47–54, 1974. doi: 10.1007/BF01022559.

C. A. Hutchison, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson, and J. C. Venter. Design and synthesis of a minimal bacterial genome. *Science*, 351(6280), 2016. doi: 10.1126/science.aad6253.

E. Ising. Beitrag zur theorie des ferromagnetismus. *Z. Physik*, 31:253–258, 1925. doi: 10.1007/BF02980577.

W. M. Jacobs. Self-assembly of biomolecular condensates with shared components. *Phys. Rev. Lett.*, 126(25):258101, 2021. doi: 10.1103/PhysRevLett.126.258101.

W. M. Jacobs and D. Frenkel. Phase Transitions in Biological Systems with Many Components. *Biophys. J.*, 112(4):683–691, 2017. doi: 10.1016/j.bpj.2016.10.043.

M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.*, 95(3):1487–1499, 2008. doi: 10.1529/biophysj.107.124784.

A. Jinich, A. Flamholz, H. Ren, S.-J. Kim, B. Sanchez-Lengeling, C. A. R. Cotton, E. Noor, A. Aspuru-Guzik, and A. Bar-Even. Quantum chemistry reveals thermodynamic principles of redox biochemistry. *PLOS Comput. Biol.*, 14(10):e1006471, 2018. doi: 10.1371/journal.pcbi.1006471.

A. Jinich, B. Sanchez-Lengeling, H. Ren, J. E. Goldford, E. Noor, J. N. Sanders, D. Segrè, and A. Aspuru-Guzik. A thermodynamic atlas of carbon redox chemical space. *Proc. Natl. Acad. Sci. U.S.A.*, 117(52):32910–32918, 2020. doi: 10.1073/pnas.2005642117.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

S. A. Kauffman. Autocatalytic sets of proteins. *J. Theor. Biol.*, 119(1):1–24, 1986. doi: 10.1016/S0022-5193(86)80047-9.

S. A. Kauffman. Even peptides do it. *Nature*, 382(6591):596–497, 1996. doi: 10.1038/382496a0.

K. Kawasaki. Diffusion constants near the critical point for time-dependent ising models. i. *Phys. Rev.*, 145(1):224, 1966. doi: 10.1103/PhysRev.145.224.

KEGG. Kyoto Encyclopedia of Genes and Genomes: PATHWAY database. `https://www.genome.jp/entry/pathway+map01100`.

M. A. Keller, A. V. Turchyn, and M. Ralser. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Mol. Syst. Biol.*, 10(4):725, 2014. doi: 10.1002/msb.20145228.

M. A. Keller, A. Zylstra, C. Castro, A. V. Turchyn, J. L. Griffin, and M. Ralser. Conditional iron and ph-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Sci. Adv.*, 2(1):e1501235, 2016. doi: 10.1126/sciadv.1501235.

H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002a. doi: 10.1126/science.1069492.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002b. doi: 10.1038/nature01254.

R. Krishnamurthy. Systems Chemistry in the Chemical Origins of Life: The 18th Camel Paradigm. *J. Syst. Chem.*, 8:40–62, 2020.

D. Kroiss, G. Ashkenasy, A. B. Braunschweig, T. Tuttle, and R. V. Ulijn. Catalyst: Can Systems Chemistry Unravel the Mysteries of the Chemical Origins of Life? *Chem*, 5(8):1917–1920, 2019. doi: 10.1016/j.chempr.2019.05.003.

K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–157, 1982. doi: 10.1016/0092-8674(82)90414-7.

N. Lane. *The Vital Question: Why is life the way it is?* Profile Books, London, 2015.

N. Lauber, C. Flamm, and K. Ruiz-Mirazo. "Minimal metabolism": A key concept to investigate the origins and nature of biological systems. *BioEssays*, 43(10):2100103, 2021. doi: 10.1002/bies.202100103.

N. Lauber, O. Tichacek, R. Bose, C. Flamm, L. Leuzzi, T.-Y. D. Tang, K. Ruiz-Mirazo, and D. De Martino. Statistical mechanics of biomolecular condensates via cavity methods. *iScience*, 26(4):106300, 2023a. doi: 10.1016/j.isci.2023.106300.

N. Lauber, O. Tichacek, K. Narayanankutty, D. De Martino, and K. Ruiz-Mirazo. Modelling spatial constraints and scaling effects of catalyst phase separation on linear pathway kinetics. *arXiv:2302.05315*, 2023b. doi: 10.48550/arXiv.2302.05315.

D. H. Lee, J. R. Granja, J. A. Martinez, K. Severin, and M. R. Ghadiri. A self-replicating peptide. *Nature*, 382(6591):525–528, 1996. doi: 10.1038/382525a0.

J. C. Letelier, G. Marín, and J. Mpodozis. Autopoietic and (M,R) systems. *J. Theor. Biol.*, 222(2):261–272, 2003. doi: 10.1016/S0022-5193(03)00034-1.

L. Ljungdahl and H. G. Wood. Incorporation of C^{14} from carbon dioxide into sugar phosphates, carboxylic acids, and amino acids by clostridium thermoaceticum. *J. Bacteriol.*, 89(4):1055–1064, 1965. doi: 10.1128/jb.89.4.1055-1064.1965.

C. Love, J. Steinkühler, D. T. Gonzales, N. Yandrapalli, T. Robinson, R. Dimova, and T.-Y. D. Tang. Reversible pH-Responsive Coacervate Formation in Lipid Vesicles Activates Dormant Enzymatic Reactions. *Angew. Chem. Int. Ed.*, 59(15):5950–5957, 2020. doi: 10.1002/anie.201914893.

R. F. Ludlow and S. Otto. Systems chemistry. *Chem. Soc. Rev.*, 37(1):101–108, 2008. doi: 10.1039/b611921m.

P. L. Luisi. *The Emergence of Life: From Chemical Origins to Synthetic Biology*. Cambridge University Press, Cambridge, 2016.

M. F. Madsen, S. Danø, and P. G. Sørensen. On the mechanisms of glycolytic oscillations in yeast. *FEBS J.*, 272(11):2648–2660, 2005. doi: 10.1111/j.1742-4658.2005.04639.x.

W. Martin and M. J. Russell. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos. Trans. R. Soc. B*, 358(1429):59–85, 2003. doi: 10.1098/rstb.2002.1183.

W. Martin and M. J. Russell. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. B*, 362(1486):1887–1926, 2007. doi: 10.1098/rstb.2006.1881.

H. Maturana and F. J. Varela. *De Máquinas y Seres Vivos*. Editorial Universitaria, Santiago de Chile, 1973. Repr. in transl.: *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, Dordrecht, 1980.

M. L. Mavrovouniotis. Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.*, 36(10):1070–1082, 1990. doi: 10.1002/bit.260361013.

M. L. Mavrovouniotis, P. Bayol, T.-K. M. Lam, G. Stephanopoulos, and G. Stephanopoulos. A group contribution method for the estimation of equilibrium constants for biochemical reactions. *Biotechnol. Tech.*, 2:23–28, 1988. doi: 10.1007/BF01874203.

B. McMullin and F. J. Varela. Rediscovering Computational Autopoiesis. In P. Husbands and I. Harvey, editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 38–47, Cambridge (MA), 1997. MIT Press.

S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, 1953. doi: 10.1126/science.117.3046.528.

S. L. Miller. Production of some organic compounds under possible primitive earth conditions1. *J. Am. Chem. Soc.*, 77(9):2351–2361, 1955. doi: 10.1021/ja01614a001.

P. Mitchell. Foundations of vectorial metabolism and osmochemistry. *Biosci. Rep.*, 11(6):297–346, 1991. doi: 10.1007/BF01130212.

D. M. Mitrea and R. W. Kriwacki. Phase separation in biology; functional organization of a higher order. *Cell Commun. Signal.*, 14(1):1–20, 2016. doi: 10.1186/s12964-015-0125-7.

MØD. MedØlDatschgerl. `https://cheminf.imada.sdu.dk/mod`.

H. J. Morowitz. *Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis*. Yale University Press, New Haven (CT) and London, 1992.

H. J. Morowitz, J. D. Kostelnik, J. Yang, and G. D. Cody. The origin of intermediary metabolism. *Proc. Natl. Acad. Sci. U.S.A.*, 97(14):7704–7708, 2000. doi: 10.1073/pnas.110153997.

K. B. Muchowska, S. J. Varma, E. Chevallot-Beroux, L. Lethuillier-Karl, G. Li, and J. Moran. Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol.*, 1(11):1716–1721, 2017. doi: 10.1038/s41559-017-0311-7.

K. B. Muchowska, S. J. Varma, and J. Moran. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature*, 569(7754): 104–107, 2019. doi: 10.1038/s41586-019-1151-1.

S. Müller, C. Flamm, and P. F. Stadler. What makes a reaction network "chemical"? *J. Cheminform.*, 14(1):1–24, 2022. doi: 10.1186/s13321-022-00621-8.

D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry (7th Edition).* W.H. Freeman, New York (NY), 2017. ISBN 9781319108243.

M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, Oxford, 2010.

E. Noor, H. S. Haraldsdóttir, R. Milo, and R. M. T. Fleming. Consistent estimation of Gibbs energy using component contributions. *PLOS Comput. Biol.*, 9(7):e1003098, 2013. doi: 10.1371/journal.pcbi.1003098.

A. I. Oparin. Proiskhozhdenie zhizni. *Moskovskii Rabochii*, 1924. Repr. in transl.: The origin of life. In J. D. Bernal, editor, *The origin of life*, pages 199–234, Weidenfeld & Nicolson, London, 1967.

A. I. Oparin. *Vozniknovenie zhizni na zemle.* Izd. Akad. Nauk SSSR, Moscow, 1936. Repr. in transl.: *The Origin of Life.* MacMillan, New York (NY), 1938.

J. T. G. Overbeek and M. J. Voorn. Phase separation in polyelectrolyte solutions. Theory of complex coacervation. *J. Cell. Comp. Physiol.*, 49 (S1):7–26, 1957. doi: 10.1002/jcp.1030490404.

J. Peretó. Controversies on the origin of life. *Int. Microbiol.*, 8(1):23–31, 2005.

J. Peretó. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. *Chem. Soc. Rev.*, 41(16):5394–5403, 2012. doi: 10.1039/c2cs35054h.

G. H. Pollack. *Cells, Gels and the Engines of Life: A New, Unifying Approach to Cell Function.* Ebner & Sons, Seattle (WA), 2001.

G. Pólya. Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. *Math. Ann.*, 84(1-2):149–160, 1921. doi: 10.1007/BF01458701.

M. Ralser. An appeal to magic? The discovery of a non-enzymatic metabolism and its role in the origins of life. *Biochem J.*, 475(16):2577–2592, 2018. doi: 10.1042/BCJ20160866.

R. Rosen. A relational theory of biological systems. *Bull. Math. Biol.*, 20 (3):245–260, 1958a. doi: 10.1007/BF02478302.

R. Rosen. The representation of biological systems from the standpoint of the theory of categories. *Bull. Math. Biol.*, 20(4):317–341, 1958b. doi: 10.1007/BF02477890.

R. Rosen. Some realizations of (M, R)-systems and their interpretation. *Bull. Math. Biol.*, 33(3):303–319, 1971. doi: 10.1007/BF02476776.

W. W. Rubey. Geologic history of sea water: an attempt to state the problem. *Geol. Soc. Am. Bull.*, 62(9):1111–1148, 1951. doi: 10.1130/0016-7606(1951)62[1111:GHOSW]2.0.CO;2.

W. W. Rubey. Development of the Hydrosphere and Atmosphere, with Special Reference to Probable Composition of the Early Atmosphere. In A. Poldervaart, editor, *Crust of the Earth: A Symposium*, GSA Special Papers Volume 62. Geological Society of America, New York (NY), 1955. doi: 10.1130/SPE62-p631.

K. Ruiz-Mirazo. Boundary versus enabling conditions for the origins of life: Comment on "Mineral self-organization on a lifeless planet" by García-Ruiz et al. *Phys. Life Rev.*, 34:96–98, 2020. doi: 10.1016/j.plrev.2020.06.001.

K. Ruiz-Mirazo and A. Moreno. Basic Autonomy as a Fundamental Step in the Synthesis of Life. *Artif. Life*, 10(3):235–259, 2004. doi: 10.1162/1064546041255584.

K. Ruiz-Mirazo, C. Briones, and A. de la Escosura. Prebiotic System Chemistry: New Perspectives for the Origins of Life. *Chem. Rev.*, 114(1):285–366, 2014. doi: 10.1021/cr2004844.

K. Ruiz-Mirazo, C. Briones, and A. de la Escosura. Chemical roots of biological evolution: the origins of life as a process of development of autonomous functional systems. *Open Biol.*, 7(4):170050, 2017. doi: 10.1098/rsob.170050.

M. J. Russell and W. Martin. The rocky roots of the acetyl-CoA pathway. *Trends Biochem. Sci.*, 29(7):358–363, 2004. doi: 10.1016/j.tibs.2004.05.007.

S. Schnell and T. E. Turner. Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. *Prog. Biophys. Mol. Biol.*, 85(2-3):235–260, 2004. doi: 10.1016/j.pbiomolbio.2004.01.012.

P. Schuster. What is special about autocatalysis? *Monatsh. Chem.*, 150: 763–775, 2019. doi: 10.1007/s00706-019-02437-z.

T. Schwander, L. Schada von Borzyskowski, S. Burgener, N. S. Cortina, and T. J. Erb. A synthetic pathway for the fixation of carbon dioxide in vitro. *Science*, 354(6314):900–904, 2016. doi: 10.1126/science.aah5237.

D. Segré, D. Ben-Eli, D. W. Deamer, and D. Lancet. The lipid world. *Orig. Life Evol. Biosph.*, 31:119–145, 2001. doi: 10.1023/A:1006746807104.

E. E. Sel'Kov. Self-Oscillations in Glycolysis 1. A Simple Kinetic Model. *Eur. J. Biochem.*, 4(1):79–86, 1968. doi: 10.1111/j.1432-1033.1968.tb00175.x.

S. N. Semenov, L. J. Kraft, A. Ainla, M. Zhao, M. Baghbanzadeh, V. E. Campbell, K. Kang, J. M. Fox, and G. M. Whitesides. Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature*, 537(7622):656–660, 2016. doi: 10.1038/nature19776.

S. Serrano-Luginbühl, K. Ruiz-Mirazo, R. Ostaszewski, F. Gallou, and P. Walde. Soft and dispersed interface-rich aqueous systems that promote and guide chemical reactions. *Nat. Rev. Chem.*, 2(10):306–327, 2018. doi: 10.1038/s41570-018-0042-6.

R. Shapiro. A Simpler Origin for Life. *Sci. Am.*, 296:46–53, 2007. doi: 10.1038/scientificamerican0607-46.

Y. Shin and C. P. Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357(6357):eaaf4382, 2017. doi: 10.1126/science.aaf4382.

B. Shirt-Ediss, S. Murillo-Sánchez, and K. Ruiz-Mirazo. Framing major prebiotic transitions as stages of protocell development: three challenges for origins-of-life research. *Beilstein J. Org. Chem.*, 13:1388–1395, 2017. doi: 10.3762/bjoc.13.135.

E. Smith and H. J. Morowitz. Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. U.S.A.*, 101(36):13168–13173, 2004. doi: 10.1073/pnas.0404922101.

J. Spitzer. From water and ions to crowded biomacromolecules: in vivo structuring of a prokaryotic cell. *Microbiol. Mol. Biol. Rev.*, 75(3):491–506, 2011. doi: 10.1128/MMBR.00010-11.

G. Springsteen, J. R. Yerabolu, J. Nelson, C. J. Rhea, and R. Krishnamurthy. Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun.*, 9(91):1–8, 2018. doi: 10.1038/s41467-017-02591-0.

J. Stankiewicz and L. H. Eckardt. Chembiogenesis 2005 and systems chemistry workshop. *Angew. Chem. Int. Ed.*, 45(3):342–344, 2006. doi: 10.1002/anie.200504139.

M. Steel. The emergence of a self-catalysing structure in abstract origin-of-life models. *Appl. Math. Lett.*, 13(3):91–95, 2000. doi: 10.1016/S0893-9659(99)00191-3.

L. Steffens, E. Pettinato, T. M. Steiner, A. Mall, S. König, W. Eisenreich, and I. A. Berg. High CO_2 levels drive the TCA cycle backwards towards autotrophy. *Nature*, 592(7856):784–788, 2021. doi: 10.1038/s41586-021-03456-9.

R. T. Stubbs, M. Yadav, R. Krishnamurthy, and G. Springsteen. A plausible metal-free ancestral analogue of the Krebs cycle composed entirely of α-ketoacids. *Nat. Chem.*, 12(11):1016–1022, 2020. doi: 10.1038/s41557-020-00560-7.

J. J. Sylvester. On an Application of the New Atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics, with Three Appendices. *Am. J. Math.*, 1(1):64–104, 1878. doi: 10.2307/2369436. URL `https://doi.org/10.2307/2369436`.

T.-Y. D. Tang, C. R. C. Hak, A. J. Thompson, M. K. Kuimova, D. S. Williams, A. W. Perriman, and S. Mann. Fatty acid membrane assembly on coacervate microdroplets as a step towards a hybrid protocell model. *Nat. Chem.*, 6(6):527–533, 2014. doi: 10.1038/nchem.1921.

R. Teufel, J. W. Kung, D. Kockelkorn, B. E. Alber, and G. Fuchs. 3-hydroxypropionyl-coenzyme a dehydratase and acryloyl-coenzyme a reductase, enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in the sulfolobales. *J. Bacteriol.*, 191(14):4572–4581, 2009. doi: 10.1128/jb.00068-09.

J. T. Trevors and G. H. Pollack. Hypothesis: the origin of life in a hydrogel environment. *Prog. Biophys. Mol. Biol.*, 89(1):1–8, 2005. doi: 10.1016/j.pbiomolbio.2004.07.003.

S. A. P. van Rossum, M. Tena-Solsona, J. H. van Esch, R. Eelkema, and J. Boekhoven. Dissipative out-of-equilibrium assembly of man-made supramolecular materials. *Chem. Soc. Rev.*, 46(18):5519–5535, 2017. doi: 10.1039/C7CS00246G.

S. Van Segbroeck, A. Nowé, and T. Lenaerts. Stochastic Simulation of the Chemoton. *Artif. Life*, 15(2):213–226, 2009. doi: 10.1162/artl.2009.15.2.15203.

F. J. Varela, H. R. Maturana, and R. Uribe. Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5(4):187–196, 1974. doi: 10.1016/0303-2647(74)90031-8.

G. von Kiedrowski. Public Communication in Systems Chemistry Workshop, Venice International University, pp. 3–4, 2005.

G. Wächtershäuser. Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.*, 52(4):452–484, 1988. doi: 10.1128/mr.52.4.452-484.1988.

G. Wächtershäuser. Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. U.S.A.*, 87(1):200–204, 1990. doi: 10.1073/pnas.87.1.200.

P. Walde and S. Ichikawa. Enzymes inside lipid vesicles: preparation, reactivity and applications. *Biomol. Eng.*, 18(4):143–177, 2001. doi: 10.1016/S1389-0344(01)00088-0.

P. Walde, A. Goto, P.-A. Monnard, M. Wessicken, and P. L. Luisi. Oparin's reactions revisited: enzymic synthesis of poly(adenylic acid) in micelles and self-reproducing vesicles. *J. Am. Chem. Soc.*, 116(17):7541–7547, 1994a. doi: 10.1021/ja00096a010.

P. Walde, R. Wick, M. Fresta, A. Mangone, and P. L. Luisi. Autopoietic Self-Reproduction of Fatty Acid Vesicles. *J. Am. Chem. Soc.*, 116(26): 11649–11654, 1994b. doi: 10.1021/ja00105a004.

J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953. doi: 10.1038/171737a0.

D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, 22(10):1249–1252, 2004. doi: 10.1038/nbt1020.

A. Whicher, E. Camprubi, S. Pinna, B. Herschy, and N. Lane. Acetyl phosphate as a primordial energy currency at the origin of life. *Orig. Life Evol. Biosph.*, 48:159–179, 2018. doi: 10.1007/s11084-018-9555-8.

A. Wołos, R. Roszak, A. Żądło-Dobrowolska, W. Beker, B. Mikulak-Klucznik, G. Spólnik, M. Dygas, S. Szymkuć, and B. A. Grzybowski. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science*, 369(6511):eaaw1955, 2020. doi: 10.1126/science.aaw1955. URL `https://doi.org/10.1126/science.aaw1955`.

J. C. Xavier, W. Hordijk, S. Kauffman, M. Steel, and W. F. Martin. Autocatalytic chemical networks at the origin of metabolism. *Proc. Royal Soc. B*, 287(1922):20192377, 2020. doi: 10.1098/rspb.2019.2377.

J. Zarzycki, V. Brecht, M. Müller, and G. Fuchs. Identifying the missing steps of the autotrophic 3-hydroxypropionate co2 fixation cycle in chloroflexus aurantiacus. *Proc. Natl. Acad. Sci. U.S.A.*, 106(50):21317–21322, 2009. doi: 10.1073/pnas.0908356106.

V. P. Zhdanov. Simulation of enzymatic cellular reactions complicated by phase separation. *Physical Review E*, 63(1):011908, 2000. doi: 10.1103/PhysRevE.63.011908.

H.-X. Zhou, G. Rivas, and A. P. Minton. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.*, 37:375–397, 2008. doi: 10.1146/annurev.biophys.37.032807.125817.

D. Y. Zubarev, D. Rappoport, and A. Aspuru-Guzik. Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle. *Sci. Rep.*, 5(1):8009, 2015. doi: 10.1038/srep08009.

Appendices

Appendix A

Graph Rewrite Rules

For the expansion of chemical reaction spaces, done within Ch. 7 of the main text, a general overview of the reaction types that where implemented as graph rewrite rules for the various graph grammars as well as the corresponding reaction classes was given (compare again Fig. 7.3 or Fig. 7.6). Here a more in detailed explanation of the mechanisms that exist within MØD to constrain the application of the reaction rules is given as well as a more detailed description of the rewrite rules for the different reaction types.

A.1 Constraints on rewrite rules

In general there are two types of constraints that are used to limit and control the application of rewrite rules to the corresponding molecule graphs. First, there are so called 'adjacency constraints'. As the name suggests, they constrain for a certain vertex (i.e. atom) the labels of the vertices it is connected to (i.e. the types of the neighbouring atoms) and/or the labels of the connecting edges (i.e. the bond types). For example, one can define for a reaction of the form:

$$\mathrm{R_1C(=O)R_2 + H_2 \longrightarrow R_1C(OH)R_2}$$

that the "C" labelled vertex can only be connected to *one* vertex with an "O" label (i.e. there is only one O-atom bound to this C-atom). Another example would be to define for a reaction of the form:

$$\mathrm{R_1CCR_2 \longrightarrow R_1C{=}CR_2 + H_2}$$

that one of the "C" labelled vertices can only have *one* connecting edge with a "=" label (i.e. there is only one double bond at this C-atom). In practice, this type of constraint can be used to prevent a rule application to certain types of functional groups or to prevent rule products that are energetically unfavourable. The former example can prevent that the reaction rule is applied to an acid group, which would result in:

$$\mathrm{R_1COOH + H_2 \longrightarrow R_1C(OH)OH}$$

which is not really a realistic reaction that can occur. The latter example can prevent that the application of the reaction rule results in a molecule with two double bonds, e.g.:

$$\mathrm{R_1CC{=}CR_2 \longrightarrow R_1C{=}C{=}CR_2 + H_2}$$

which would, in general, not be a stable compound.

Furthermore, there are so called 'label constraints'. As outlined in the main text, it is possible to use custom atom labels, like "A" for the activator compound. As with standard atom labels, they can only be mapped to vertices with the same label. It is also possible to define vertex labels as variables that can only take a certain label-value from a predefined set. An example would be to define for a reaction of the form:

$$\mathrm{R_1C(=O)R_2 + H_2 \longrightarrow R_1C(OH)R_2}$$

that $\mathrm{R_1,R_2} \in$ {"C","H"}. This ensures that the reaction rule is only applied to an aldehyde or a ketone. Therefore, similar to adjacency constraints, label constraints can be used to prevent rule applications to specific functional groups or to prevent energetically unfavourable reaction products. However, in some situations they allow for more flexibility when trying to control a DG expansion (but in practice a combination of both types of constraints is used). Additionally, one can also use a special vertex label, denoted as "*". These 'wildcard' labels can be matched by a vertex with any kind of label (i.e. they are placeholders for an arbitrary atom type). For example, one can define for a reaction of the form:

$$\mathrm{*{-}C(OH){-}* \longrightarrow *{-}C(=O){-}* + H_2}$$

While it does not specify the labels of the vertices connected to the "C" labelled vertex (i.e. what other atoms are bound to the C-atom), it does specify that the edge labels are "-" (i.e. there are no further double bonds on the C-atom). Therefore, the usage of wildcard labels allows for a further method of constraining the bond types that are allowed in a reaction rule.

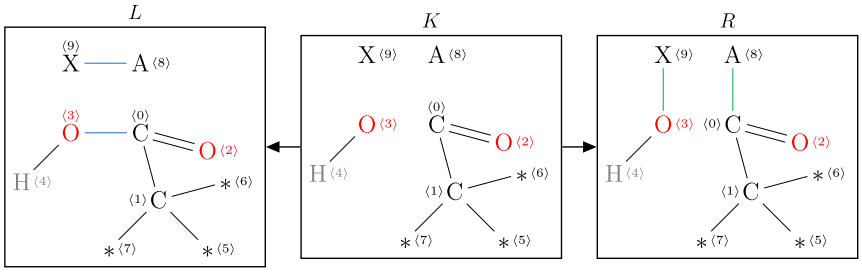
A.2 Implemented rewrite rules

In the following, the implementations of the graph rewrite rules for the different reactions of the corresponding reaction classes are given (compare again Tab. 7.1 in the main text). By contrast to the main text, the reaction rules are visualized here as double-pushout (DPO) diagrams (as this is the actual way they are implemented in MØD [Andersen 2015]). The DPO diagrams in each figure only depict the transformation made to the matching molecule graph (i.e. the addition, deletion or relabelling of edges), as well as the context etc. Further constraints that were implemented for a rewrite rule are explained in the respective figure captions.

Class 0

Activator addition

The binding of an activator compound is implemented by the following rewrite rule:



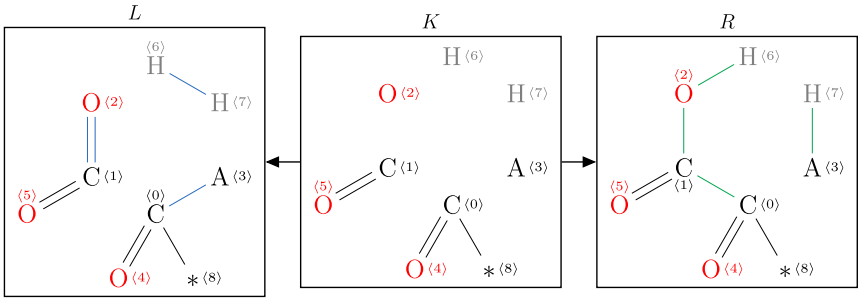
It is assumed that the reaction can only occur on acids that are not α-keto acids, which is achieved through the usage of wildcard labels on node 1, fixing all bond types at that node to single bonds (i.e. there could be no C=O group). Additionally, the rule is prevented to prevent the appearance of highly branched (and therefore unnecessarily complex) molecules, the rule is prevented form being applied to acid groups that are on a side branch of a carbon chain. This is achieved by an adjacency constraint on node 1: there

can be *at max two* neighbouring nodes with label "C".

Class 1

Reductive carboxylation

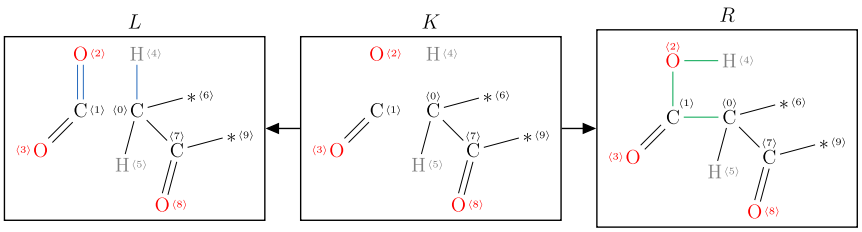
The reductive carboxylation of an activated acid is implemented by the following rewrite rule:



The wildcard label at node 0 is four is used to ensure that the valency of this C-atom is four.

Carboxylation

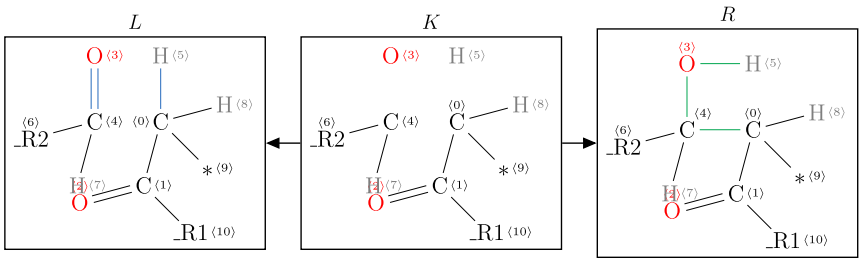
The redox neutral carboxylation is implemented by the following rewrite rule:



It is assumed that the reaction can only occur α-to a ketone (or aldehyde), resulting in the formation of a β-keto acid This is achieved by an adjacency constraint on node 7: there can be *only one* neighbouring node with label "O". The wildcard label at node 0 and node 7 is used to ensure that the valency of these C-atoms is four.

Aldol addition

The aldol addition is implemented by the following rewrite rule:

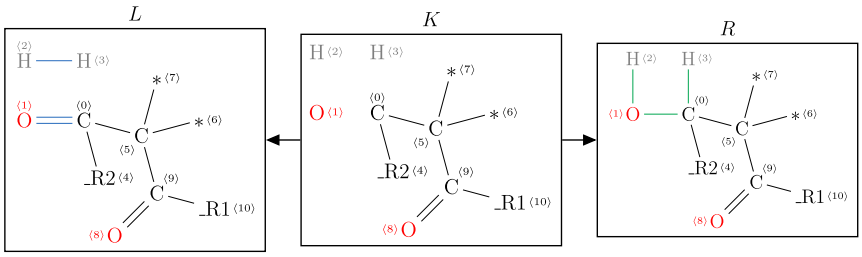


It is assumed that the reaction only occurs between two aldehydes or ketones (or an aldehyde and a ketone). This is achieved by the following label constraints: R1, R2 $\in$ {"C","H"}. Additionally, the condensation of molecules with more than two C-atoms to a carbon chain is prevented (again to avoid a combinatorial explosion of too large and complex molecules). This is achieved by an adjacency constraint on node 6: there can be *only one* neighbouring node with label "C". The wildcard label at node 0 is used to ensure that the valency of this C-atom is four.

Class 2

Carbonyl reduction (strict)

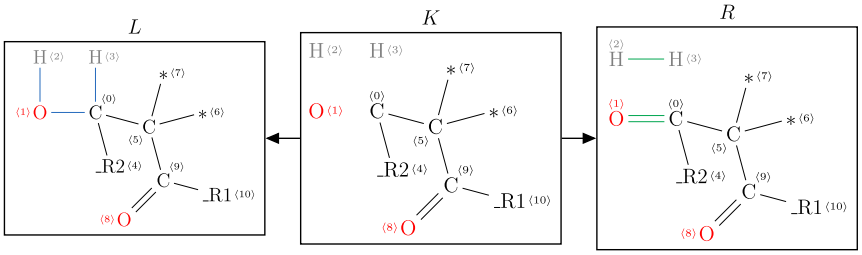
The strict version of the carbonyl reduction is implemented by the following rewrite rule:



It is assumed that the reaction only occurs on a carbonyl group that is $\beta-$to an acid (or activated β-keto acid). This is achieved by the following label constraints: R1 $\in$ {"C","H"} and R2 $\in$ {"C","H"} . The wildcard labels at node 5 are used to ensure that the valency of this C-atom is four.

Carbonyl oxidation (strict)

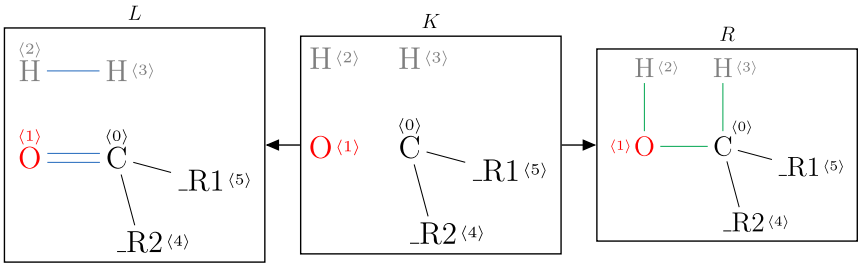
The strict version of the carbonyl oxidation is implemented by the following rewrite rule:



This rewrite rule is basically the inverse of the rule for the strict carbonyl reduction and therefore the same constraints apply to it.

Carbonyl reduction (relaxed)

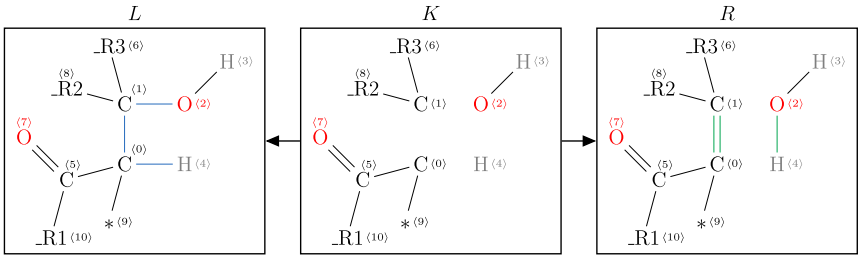
The relaxed version of the carbonyl reduction is implemented by the following rewrite rule:



It is assumed that the reaction occurs on any aldehyde or ketone. This is achieved by the following label constraints: R1, R2 $\in$ {"C","H"},

Dehydration

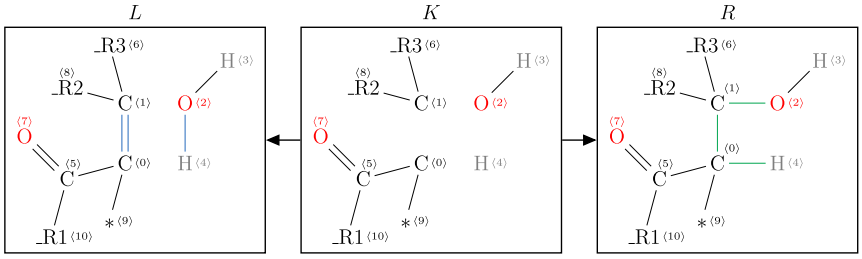
The dehydration reaction is implemented by the following rewrite rule:



It is assumed that the reaction only occurs on a hydroxyl group that is $\beta-$to an acid (or activated β-keto acid). This is achieved by the following label constraints: R1 $\in$ {"O","A"} and R2, R3 $\in$ {"C","H"}. The wildcard label at node 0 is used to ensure that the valency of this C-atom is four.

Hydration

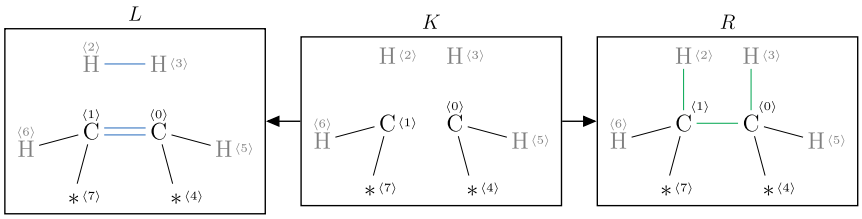
The hydration reaction is implemented by the following rewrite rule:



This rewrite rule is basically the inverse of the rule for dehydration reaction and therefore the same constraints apply to it.

Alkene reduction

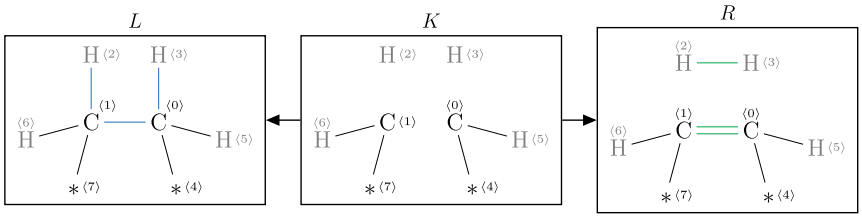
The alkene reduction is implemented by the following rewrite rule:



The wildcard labels are used to ensure that the valency of node 0 and node 1 is four.

Alkene oxidation

The alkene oxidation is implemented by the following rewrite rule:

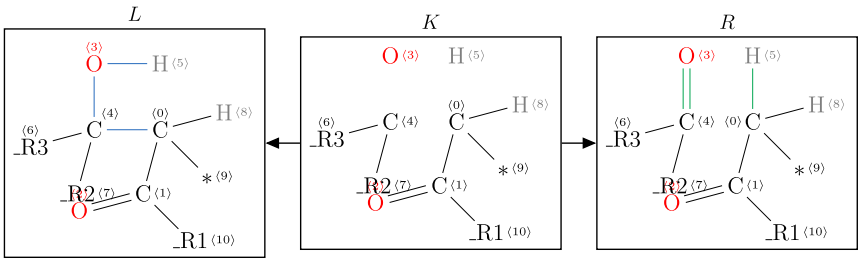


This rewrite rule is basically the inverse of the rule for alkene reduction and therefore the same constraints apply to it.

Class 3

Aldol cleavage

The aldol cleavage is implemented by the following rewrite rule:



It is assumed that the reaction only occurs on a hydroxyl group that is β-to an aldehyde or ketone (or an activated acid). This is achieved by the following label constraints: R1 $\in$ {"A","C","H"} and R2, R3 $\in$ {"C","H"}. The wildcard label is used to ensure that the valency of node 0 is four.

Appendix B

Lattice Model Calculations

For the study how the liquid-liquid phase separation (LLPS) of a mixture of catalysts may influence a linear cascade reaction, done within Ch. 10 of the main text, several formulas where used to compare the results from the microscopic lattice model with mean-field calculations. More precisely, formulas for the reaction time t_R (i.e. the time it takes to finish the reaction-cascade) as well as the change of the substrate concentration C_μ that were used. Within this part of the appendix a more detailed description how these formulas were derived is going to be given.

B.1 Estimating the reaction time

As outline in the main text, a two dimensional lattice that contains L different catalysts, is assumed. As such, L also corresponds to the length of the reaction-cascade (compare Eq. 10.1 of the main text). The total surface fraction of all catalysts is given as ϕ_{tot} and it is further assumed that each type of catalyst occupies the same surface fraction, thus the surface fraction ϕ_μ of a certain catalyst type E_μ can be obtained as: $\phi_\mu = \phi_{tot}/L$. To make an estimation for the reaction time t_R, which in the context of the microscopic simulation this corresponds to the time it takes for a random walk on the lattice to encounter all L catalysts in the correct order, two cases need to be compared: the case of the catalysts remaining well mixed system (WM-system) and the case of the catalysts phase-separating into a droplet (PS-system).

B.1.1 Well-mixed system

In a WM-system the catalysts are homogeneously distributed across the lattice. Therefore, the probability p_μ for a substrate particle $S_{\mu-1}$ to randomly jump to a lattice site, occupied by a catalyst E_μ is proportional to its concentration and, thus, to its surface fraction ϕ_μ, i.e. $p_\mu \sim x_\mu$. The mean time $\bar{t}_\mu$ for a single particle to reach a catalyst E_μ is then:

$$\bar{t}_\mu = \frac{1}{p_\mu} \sim \frac{1}{\phi_\mu} = \frac{L}{\phi_{tot}}$$

which is in principle the mean time for an individual reaction of the pathway to finish. One can assume that the mean reaction-time $\bar{t}_R$ is proportional to the sum of all the individual reaction times, i.e.

$$\bar{t}_R \sim \sum_{\mu}^{L} t_\mu = \sum_{\mu}^{L} \frac{L}{\phi_{tot}} = L\frac{L}{\phi_{tot}}$$

The mean reaction-time $\bar{t}_R^{\,wm}$ in a WM-system can thus be expressed as:

$$\bar{t}_R^{\,wm} = a\frac{L^2}{\phi_{tot}} \tag{B.1}$$

with a being a proportionality constant.

B.1.2 Phase-separated system

In a PS-system, the catalysts will eventually start forming a droplet. Thus the mean-reaction time (i.e. the time to find the catalysts in the correct order) is delayed by the time it takes an initial substrate S_0 to encounter the droplet. One can thus assume that:

$$\bar{t}_R = \bar{t}_R^{\,b} + \bar{t}_H \tag{B.2}$$

with $t_R^{\,b}$ as the reaction time within the droplet and $\bar{t}_H$ the time to reach, or hit the droplet (i.e the "hitting-time"). In other words, the problem is effectively split into two parts.

For the reaction time inside the droplet a simple assumption can be made: as the particles stay inside the droplet (it is assumed that there is an attraction between substrates and catalysts) and the catalysts are homogeneously

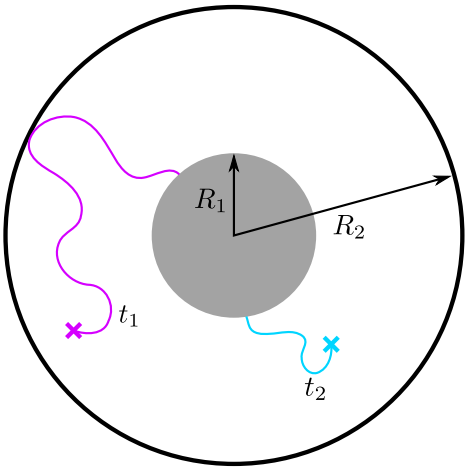


Figure B.1: To get the hitting-time $\bar{t}_h$ a spherical system is assumed with a reflective boundary-condition at R_2 and absorbing boundary-condition at R_1.

distributed inside the droplet this situation can be regarded as a well-mixed system with a total concentration of $\phi'_{tot} \sim 1$, thus Eq. B.1 can be used to obtain:

$$\bar{t}_R^{\,b} = aL^2 \tag{B.3}$$

with a being the same proportionality constant.

To get an estimation for the hitting time a more involved calculation has to be made. Moving from a discrete lattice to a continuos space, the mean time for single particle to reach the droplet $T(\vec{x})$ is a function of the position $\vec{x}$ and it verifies the Poisson Equation:

$$\nabla^2 T(\vec{x}) = -\frac{1}{2D} \tag{B.4}$$

which is the equation for the mean first exit time from a region, that can be obtained from the backward Fokker-Planck (FP) equation (a detailed explanation can be found in Chapter 6 of [Gardiner 2009], Eq. 6.6.8). The term D is the diffusion constant. In the case of the lattice model, one can in principle set $D = 1$, but it is retained here for the sake of clarity. In spherical coordinates (as shown in Fig. B.1) and with $T(r)$ only depending on the radial-coordinate, Eq.B.4 reduces to a second-order inhomogeneous

ODE:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}T\right)+\frac{1}{2D}=0$$

$$T^{''}+\frac{1}{r}T^{'}+\frac{1}{2D}=0$$

In general, the solution for an inhomogeneous ODE is the sum of the solution for the homogeneous ODE and particular solution, i.e.

$$T=T_h+T_p$$

with the homogeneous ODE being:

$$T_h^{''}+\frac{1}{r}T_h^{'}=0$$

Using a substitution method ($T_h^{'}=u$) to reduce this to a first order ODE, the homogeneous solution can be obtained as:

$$T_h=A\log r+B$$

Through a polynomial ansatz one can obtain the particular solution as:

$$T_p=-\frac{1}{8D}r^2+C$$

Thus collecting T_h and T_p results in:

$$T(r)=A\log r-\frac{r^2}{8D}+B$$

with A, B being constant factors. To obtain values for A and B the boundary conditions (BC) at R_1, R_2, outlined in Fig. B.1, can be used: $T(R_1)=0$ (absorbing BC) and $\nabla T(R_2)=0$ (reflecting BC). This way $T(r)$ can finally be obtained as:

$$\begin{aligned}T(r)&=\frac{R_2^2}{4D}\log r-\frac{r^2}{8D}+\frac{R_1^2}{8D}-\frac{R_2^2}{4D}\log R_1\\&=\frac{R_2^2}{4D}\log\frac{r}{R_1}-\frac{1}{8D}\left(r^2+R_1^2\right)\end{aligned}\tag{B.5}$$

As this is the mean time to reach the droplet starting at point r, one has to

further integrate over the whole volume to get the overall mean hitting-time:

$$\begin{aligned}\bar{t} &= \frac{1}{V}\int T(\vec{x})\mathrm{d}\vec{x} \\ &= \frac{1}{\pi R_2^2}\underbrace{\int_0^{R_1} T(r)2\pi r\mathrm{d}r}_{=0} + \frac{1}{\pi R_2^2}\int_{R_1}^{R_2} T(r)2\pi r\mathrm{d}r\end{aligned}$$

where the first integral, which corresponds to the portion of particles that are already inside the droplet, vanishes as the hitting-time is essentially zero for them (they already are inside the droplet). The second integral then corresponds to the particles that are outside the droplet. Solving this integral yields:

$$\bar{t} = \frac{R_2^2}{16D}\left[4\phi_{tot} - 3 - 2\log\phi_{tot} - \phi_{tot}^2)\right]$$

In the case of a regular lattice one can assume that $R_2 \sim d$. This finally gives the hitting-time as:

$$\bar{t}_H(\phi_{tot}, d) = \frac{d^2}{16D}(4\phi_{tot} - 3 - 2\log\phi_{tot} - \phi_{tot}^2) \tag{B.6}$$

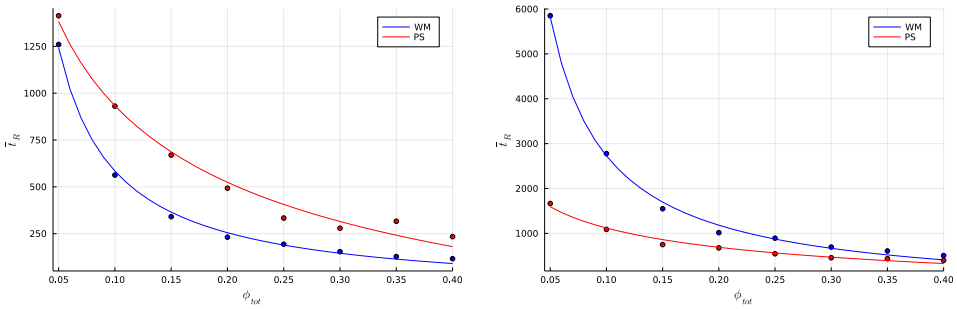
Finally, putting Eq. B.3 and Eq. B.6 together results in $\bar{t}_R^{\,ps}$, the mean reaction time in a phase-separated system:

$$\begin{aligned}\bar{t}_R^{\,ps} &= aL^2 + \frac{d^2}{16D}(4\phi_{tot} - 3 - 2\log\phi_{tot} - \phi_{tot}^2) \\ &= aL^2 + bd^2 g(\phi_{tot})\end{aligned} \tag{B.7}$$

Comparing Eq. B.1 and Eq. B.7 shows that $\bar{t}_R^{\,ps}$ is basically $\bar{t}_R^{\,wm}$ but with a different slope and an offset that is proportional to d and ϕ_{tot}. In the main text the dependence of the mean reaction time $\bar{t}_R$ on the pathway-length L, predicted by Eq. B.1 and Eq. B.7, was tested (see Fig. 10.4 in the main text). In Fig. B.2 the dependence of $\bar{t}_R$ on the surface fraction ϕ_{tot}, as predicted by the same equations, is furthermore tested.

B.1.3 Distribution of reaction times

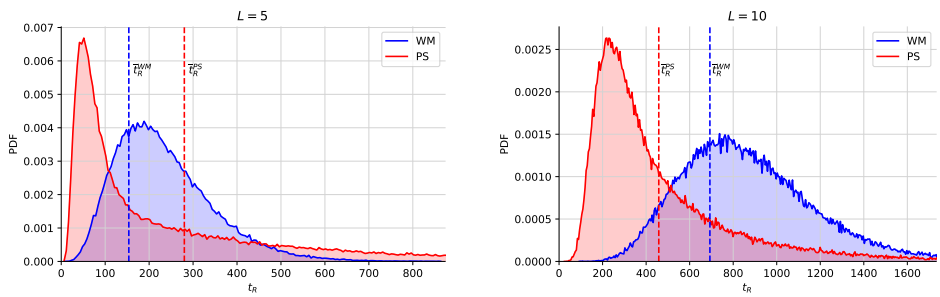
In the main text as well as in the preceding sections, only the average reaction time was discussed. However, the complete distribution of reaction times for



(a) **(b)**

Figure B.2: Average reaction time $\bar{t}_R$ as a function of the total surface fraction ϕ_{tot} ($J = 1.0, d = 50$). Simulations were carried out for **(a)** $L = 5$ and **(b)** $L = 10$, comparing WM-systems (blue dots, $\beta = 0.0$ and $I = 0.0$) with PS-systems (red dots, $\beta = 5.0$ and $I = 1.0$). Solid lines: fit of the simulation data using Eq. B.1 and Eq. B.7 for the WM- and the PS-systems, respectively.

all the substrate molecules is, in fact, significantly different across the two types of systems examined.



(a) **(b)**

Figure B.3: Distribution of total reaction times ($J = 1.0, d = 50$). Simulations were carried out for **(a)** $L = 5$ and **(b)** $L = 10$ for both WM-system (blue, $\beta = 0.0$ and $I = 0.0$) and PS-system (red, $\beta = 5.0$ and $I = 1.0$). Dashed lines show the average reaction time $\bar{t}_R$ for both cases.

As it is shown in Fig. B.3, the distribution of reaction times in the PS-system exhibits a sharp initial peak and a long exponential tail, unlike the WM-system. Substrate molecules are initiated at random locations in the system during the simulations, and in case of the PS-system, some of them start from within the condensate ($\bar{t}_h = 0$ in Eq. B.2) causing them to quickly complete all the reaction steps. This is the reason for the initial spike in the distribution of reaction times in the PS-system. The long exponential tail, on the other hand, comes from the distribution of hitting times t_H.

B.2 Mean-field model for the change of concentration

In principle the enzymatic cascade-reaction, given by Eq. 10.1 in the main text, can be formulated as an L-step sequential chemical reaction:

$$\mathrm{S_0} \xrightarrow{\mathrm{k_0}} \mathrm{S_1} \xrightarrow{\mathrm{k_1}} \mathrm{S_2} \ldots \mathrm{S_{L-1}} \xrightarrow{\mathrm{k_{L-1}}} \mathrm{P}$$

with k_μ being the reaction rate constants of the corresponding reaction steps. Using mass action kinetics, these chemical reaction equations can be converted to a set of coupled ODEs:

$$\begin{aligned} \dot{C}_0 &= -k_0 C_0 \\ \dot{C}_\alpha &= k_{\alpha-1}\left[C_{\alpha-1}\right] - k_\alpha\left[C_\alpha\right] \ , \ \forall \alpha \geq 1 \\ \dot{C}_P &= k_{L-1} C_{L-1} \end{aligned} \tag{B.8}$$

With C_α being the concentration of the substrate S_α (i.e. the amount of S_α particles on the lattice). Assuming the following initial conditions: $C_0(0) = \text{const}, C_\alpha(0) = 0 \ \forall \alpha \geq 1$ and $C_P(0) = 0$, these ODEs can be solved in order to obtained trajectories for $C_\alpha(t)$, the change of concentration of substrate S_α over time. In the following this is done for two special cases.

B.2.1 Homogeneous reaction rate constants

For the first case, homogeneous reaction rate constants are assumed, i.e. $k_\alpha = k, \forall \alpha$. In general, this corresponds to a well-mixed system in the lattice model as all the catalysts are homogeneously distributed across the lattice and, as argued above, the probabilities of encountering the catalyst necessary for a certain reaction are all equal. Therefore, Eq. B.8 can be reformulated as:

$$\begin{aligned} \dot{C}_0 &= -k C_0 \\ \dot{C}_\alpha &= k(C_{\alpha-1} - C_\alpha) \ , \ \forall \alpha \geq 1 \\ \dot{C}_P &= k C_{L-1} \end{aligned} \tag{B.9}$$

The homogeneous ODE for $C_0(t)$ can be easily solved with the above established initial conditions:

$$C_0(t) = C_0(0) e^{-kt}$$

Turning now to the inhomogeneous ODEs for $C_\alpha(t)$, one can use the method of the integrating-factor to get a recursive formula:

$$C_\alpha(t) = e^{-kt}\left[\int kC_{\alpha-1}(t)e^{kt}\ \mathrm{d}t + K\right]$$

This results in the following general expression:

$$C_\alpha(t) = k^\alpha C_0(0)\frac{t^\alpha}{\alpha!}e^{-kt}$$

This formula can be used together with the mass conservation law: $\sum_{\alpha=0}^{L-1} C_\alpha(0) + C_P(0) = \sum_{\alpha=0}^{L-1} C_\alpha(t) + C_P(t)$, and the initial conditions in order to solve the last ODE for $C_P(t)$. This way one finally obtains:

$$\begin{aligned} C_\alpha(t) &= k^\alpha C_0(0)\frac{t^\alpha}{\alpha!}e^{-kt} \\ C_P(t) &= C_0(0)\left(1 - e^{-kt}\sum_{\alpha=0}^{L-1}\frac{(kt)^\alpha}{\alpha!}\right) \end{aligned} \tag{B.10}$$

B.2.2 Homogeneous reaction rate constants with a different rate in first reaction

A slight variation of the above system would be to leave the reaction rate constants the same for all reactions except the first one, i.e. $k_0, k_\alpha = k, \forall\alpha \geq 1$. In general, this corresponds to a phase-separated system in the lattice model as all the catalysts are forming a droplet and, as argued above, the first reaction step is "delayed" by the time it takes to encounter this droplet. Therefore, Eq. B.8 can be reformulated as:

$$\begin{aligned} \dot{C}_0 &= -k_0 C_0 \\ \dot{C}_1 &= k_0 C_0 - kC_1 \\ \dot{C}_\alpha &= k(C_{\alpha-1} - C_\alpha)\ ,\ \forall\alpha \geq 2 \\ \dot{C}_P &= kC_{L-1} \end{aligned} \tag{B.11}$$

with the above established initial conditions. As before, the solution of the homogeneous ODE for $C_0(t)$ is:

$$C_0(t) = C_0(0)e^{-k_0 t}$$

and the results for the inhomogeneous ODEs for $C_\alpha(t)$ ($\alpha \geq 1$) can be obtained using the method of the integrating factor:

$$C_\alpha(t) = k_0 k^{\alpha-1} C_0(0) e^{-kt} \underbrace{\left[\frac{e^{(k-k_0)t} - \sum_{m=0}^{\alpha-1} \frac{[(k-k_0)t]^m}{m!}}{(k-k_0)^\alpha} \right]}_{=f_\alpha}$$

Again, the last ODE for $C_P(t)$ can be solved using the mass conservation law. Therefore one obtains:

$$\begin{aligned}
C_0(t) &= C_0(0) e^{-k_0 t} \\
C_\alpha(t) &= k_0 k^{\alpha-1} C_0(0) e^{-kt} \cdot f_\alpha \ , \quad \alpha \geq 1 \\
C_P(t) &= C_0(0) \left(1 - k_0 e^{-kt} \sum_{\alpha=0}^{L-1} k^{\alpha-1} \cdot f_\alpha \right) \qquad \text{(B.12)} \\
f_\alpha &= \frac{e^{(k-k_0)t} - \sum_{m=0}^{\alpha-1} \frac{[(k-k_0)t]^m}{m!}}{(k-k_0)^\alpha}
\end{aligned}$$